Nonparametric Estimation of Rényi Divergence and Friends

Akshay Krishnamurthy [*1], Kirthevasan Kandasamy [†2],
Barnabás Póczos [‡2], and Larry Wasserman [§3]

[1]Computer Science Department, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University
[3]Statistics Department, Carnegie Mellon University

February 13, 2014

Abstract

We consider nonparametric estimation of L_2, Renyi-α and Tsallis-α divergences of 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 analyze the rates of convergence for our estimators and show that the parametric rate of $n^{-1/2}$ is achievable 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 confirm 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 between pairs of random variables [14]. 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 [7]. It has also been successful in computer vision, where features are computed for each patch in an image and these feature vectors as modeled as independent draws from an underlying distribution [17].

[*]akshaykr@cs.cmu.edu
[†]kandasamy@cmu.edu
[‡]bapoczos@cs.cmu.edu
[§]larry@stat.cmu.edu

It is therefore crucial that we obtain accurate divergence estimates from samples drawn independently from each distribution. In the nonparametric setting, a number of authors have proposed various estimators which are provably consistent. However, the actual *rates of convergence* of these estimators and the minimax optimal rate are still unknown.

In this work, we propose three estimators for the L_2^2, Renyi-α, and Tsallis-α divergence between two continuous distributions. The estimators are based on correcting an initial plug-in estimator by estimating higher order terms in the von Mises expansion of the divergence functional. We establish the rates of convergence of our estimators under the assumption that both densities belong to a Holder 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 linear 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 [12, 19].

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 $o(n^{\frac{-4s}{4s+d}})$-rate when $s \leq d/4$ and at $o(n^{-1/2})$-rate otherwise. This establishes the optimality of our best 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 Section 3. In Section 4 we provide a proof sketch of our results. We present some simulations validating our theoretical results in Section 5 and conclude with some directions for future work in Section 6. We defer proof details to the supplementary material.

1.1 Preliminaries

Before proceeding, let us establish notational conventions and 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 continuous 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 will be drawn independently from p while the samples Y_i will be drawn independently from q. We are given n samples from each distribution. 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. **Renyi-α Divergence** [18]
$$D_\alpha(p,q) = \frac{1}{\alpha - 1} \log \left(\int p^\alpha(x) q^{1-\alpha}(x) d\mu(x) \right)$$
3. **Tsallis-α Divergence** [22]
$$T_\alpha(p,q) = \frac{1}{\alpha - 1} \left(\int p^\alpha(x) q^{1-\alpha}(x) d\mu(x) - 1 \right)$$

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 unknown terms in the divergences, and we will verify that estimators for $T(p,q)$ result in good divergence estimators.

We assume that the densities p, q belong to $\Sigma(s, L)$, the **periodic Holder 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 Holder 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:*

$$|D^r f(x) - D^r f(y)| \leq L||x - y||^{s-|r|} \tag{1}$$

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

The Holderian assumption is standard in the nonparametric literature while the periodic assumption subsumes more standard boundary smoothness conditions [11]. We also assume that the densities are bounded above and below, formally, $0 < \kappa_l \leq p(x), q(x) \leq \kappa_u < \infty$ for all $x \in [0,1]^d$.

The sine qua non of our work is the **von Mises expansion**. Given a functional T mapping distributions to the reals, the first-order von Mises expansion (in terms of the Gateaux derivative) 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, the functional 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 some differentiable function. In this case one can verify that the von Mises expansion reduces to a functional Taylor expansion on the densities:

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

We generalize this idea to functionals $T(P, Q)$ of two distributions, with $\phi(p, q) = p^\alpha q^\beta$. We will also make higher order expansions analogous to the Taylor expansion. See Chapter 20 of van der Vaart's book for an introduction to von Mises calculus [24]. We will frequently write $T(f)$ instead $T(F)$.

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 [14].

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 [5, 26, 16]. 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 Renyi-α and Tsallis-α divergences. A number of other authors have also proposed consistent estimators via the empirical CDF or histograms [25, 15]. Unfortunately the rates of convergence for all of these methods are still unknown.

Singh *et al.* [19] recently established a rate of convergence for a plug-in kernel density estimator in the divergence estimation setting. Their estimator has rate $n^{\frac{-s}{s+d}}$ when $s < d$ and $n^{-1/2}$ otherwise which matches some existing results on estimating entropy functionals [11]. 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.* [12] construct an estimator for Csiszár f-divergences (also known as Ali-Slivey divergences) via regularized M-estimation and prove a rate of convergence in the setting where the likelihood-ratio $d\mathbb{P}/d\mathbb{Q}$ belongs to a Reproducing Kernel Hilbert Space. They obtain rates of convergence depending on the complexity of this RKHS, but it is not clear how to port these results to our setting, as their assumption is incomparable to our assumption on the smoothness of the density.

A related and flourishing line of work is on estimating entropy functionals. Here, the majority of the methods are graph-based, involving either nearest neighbor graphs or spanning trees over the data [6, 10, 13, 20]. One exception is the KDE-based estimator for mutual information and joint entropy of Liu, Lafferty, and Wasserman [11]. A number of these estimators come with provable convergence rates.

While it is not clear how to port these ideas to the divergence estimation setting, it is still worth comparing rates. The estimator of Liu *et al.* converges at rate $O(n^{\frac{-s}{s+d}})$ which achieves the parametric rate when $s > d$ [11]. Similarly, Sricharan *et al.* [20] show that when $s > d$ a k-NN style estimator achieves rate $O(n^{-2/d})$ (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 still require $s > d$ orders of smoothness [21]. In contrast, our quadratic and linear estimators achieve the parametric $n^{-1/2}$ rate at lower smoothness ($s > d/4, d/2$ respectively) than theirs and enjoys 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 an optimal density estimator.

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 [2] prove a $n^{\frac{-4s}{4s+d}}$-lower bound for estimating integral functionals of a density. Hero *et al.* [6] give a matching lower bound for estimating Renyi-α entropies.

Finally, our estimators and proof techniques are based on several classical works on estimating integral functionals of a density. In this setting, the goal is to estimate a quantity $\int \phi(f(x))d\mu(x)$ given samples $X_1, \ldots, X_n \sim f$ for some known function ϕ. A series of papers study this problem and show that $n^{-1/2}$ rate of convergence is attainable if and only if $s > d/4$, which is analogous to our results [2, 9, 8, 1]. 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 Main Results

3.1 The Estimators

Recall that we are interested in estimating integral functionals of two densities, i.e. functionals of the form $T(p, q) = \int p^\alpha(x)q^\beta(x)$. As an initial attempt, if we have estimators $\hat{p}$ and $\hat{q}$ for p, q respectively, we can

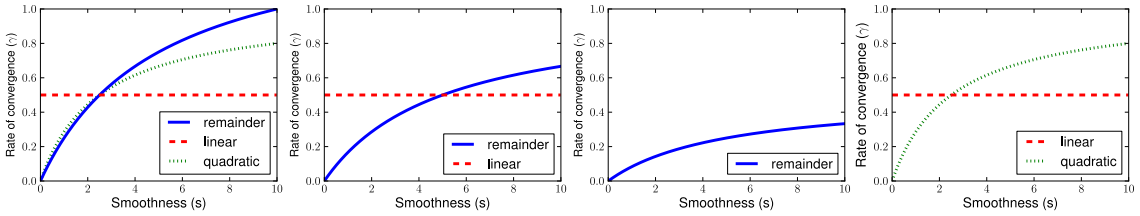

Figure 1: Rates of convergence of the estimators $\widehat{T}_{quad}, \widehat{T}_{lin}, \widehat{T}_{pl}$ along with the rate of convergence in the lower bound (Theorem 5). Plot is γ vs. smoothness s with $d = 10$, where the rate of convergence is $O(n^{-\gamma})$. The rate of convergence for each estimator 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.

use the plug-in estimator $\widehat{T}_{pl} = T(\hat{p}, \hat{q})$. Via the von Mises expansion of $T(p, q)$, the error can be upper bounded by:

$$|\widehat{T}_{pl} - T(p, q)| \leq C_1 ||\hat{p} - p||_1 + C_2 ||\hat{q} - q||_1$$

under the assumption that $p, \hat{p}, q, \hat{q}$ are all bounded from above and below. Existing results on the ℓ_1 convergence of density estimators suggests that $\widehat{T}_{pl}$ will enjoy a $O(n^{\frac{-s}{2s+d}})$ rate [3].

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

An even better estimator corrects by the both the first and second-order terms in 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.

Before proceeding, let us first mention some caveats with constructing kernel density estimators in our setting. First, except for the plug-in estimator $\widehat{T}_{pl}$, we will only train the KDEs on half of the sample. This will give us independent samples with which to estimate the remaining terms. The second caveat is that to analyze the correcting terms, we will require that the density estimators 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.

The last issue is that the standard KDE suffers high bias near the boundary of the domain. To correct this bias, we adopt the strategy used by Liu *et al.* of "mirroring" the data set over the boundaries [11]. 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. Thus this augmented estimator can be shown to match the rates of convergence in the literature [3, 23].

Below we enumerate the terms in the first and second order von Mises expansions that we will estimate or compute, conditional on the first half of the data:

$$\theta^p_{1,1} = \mathbb{E}_{X\sim p} \alpha \hat{p}^{\alpha-1}(X) \hat{q}^{\beta}(X)$$
$$\theta^q_{1,1} = \mathbb{E}_{Y\sim q} \beta \hat{p}^{\alpha}(Y) \hat{q}^{\beta-1}(Y)$$
$$\theta^p_{2,1} = \mathbb{E}_{X\sim p} \alpha(2-\alpha-\beta) \hat{p}^{\alpha-1}(X) \hat{q}^{\beta}(X)$$

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

With these definitions the zero-th, first and second order expansions of $T(p,q)$ about $T(\hat{p},\hat{q})$ are:

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

And the remainders $R_a = O(||p-\hat{p}||_a^a + ||q-\hat{q}||_a^a)$. Refer to the appendix for details of these calculations. 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 $\theta = \mathbb{E}_{X\sim f}[\psi(X)]$ where ψ is some function. A natural estimator when we are given samples $X_1^n \sim f$ is the sample mean $\hat{\theta} = \frac{1}{n}\sum_{j=1}^n \psi(X_j)$.

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

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

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

$$
\hat{\theta} = \frac{1}{n}\sum_{j=1}^n \sum_{k\in M} \left(\frac{1}{n}\sum_{i=1}^n \phi_k(X_i)\right)\phi_k(Y_j)\psi(Y_j) \tag{2}
$$

Here $M \subset D$ is chosen to tradeoff the bias and the variance of the estimator. To develop some intuition, if we knew the function f, we would simply use the sample mean $\frac{1}{n}\sum_{j=1}^n f(Y_j)\psi(Y_j)$. Unfortunately, f is unknown, so we replace it with an estimator formed by truncating the Fourier expansion of f. Specifically, we use X_1^n to compute $\hat{a}_k = \frac{1}{n}\sum_{i=1}^n \phi_k(X_i)$ and replace f with $\hat{f}(\cdot) = \sum_{k\in M}\hat{a}_k\phi_k(\cdot)$.

For the quadratic functional, a projection estimator was proposed and analyzed by Laurent [9].

$$
\hat{\theta} = \frac{2}{n(n-1)}\sum_{k\in M}\sum_{i\neq j}\phi_k(X_i)\phi_k(X_j)\psi(X_j) - \frac{1}{n(n-1)}\sum_{k,k'\in M}\sum_{i\neq j}\phi_k(X_i)\phi_{k'}(X_j)b_{k,k'}(\psi) \tag{3}
$$

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:

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

Here for $\widehat{T}_{pl}$, $\hat{p},\hat{q}$ are trained on the entire sample, while for $\widehat{T}_{lin}, \widehat{T}_{quad}$, they are trained on the first half of the sample. All of the $\hat{\theta}^{(\cdot)}_{(\cdot),(\cdot)}$ terms are constructed using the second half of the sample, using the appropriate estimator defined above.

In general, the estimators all require numeric 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 Renyi-α and Tsallis-α-divergences, the constants C_1, C_2 are identically zero, so the first term can simply be omitted from the estimator. In this setting $\widehat{T}_{lin}$ is remarkably simple, it involves training KDEs and estimating a specific linear functional of them via the sample mean. As we will show, although this estimator is not minimax optimal, it enjoys a fairly fast rate of convergence while being computationally practical. Unfortunately, even when $C_2 = 0$ as in the divergences of interest, the quadratic estimator still involves numeric integration of the terms $b_{i,i'}$. 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.

3.2 Rate of Convergence

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

Theorem 2. *Under our assumptions we have:*

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

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

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

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

For all estimators, the bandwidth and the kernel must be carefully chosen. We choose $h \asymp n^{\frac{-1}{2s+d}}$, which results in optimal kernel density estimation, in contrast to some previous work on plug-in estimators for entropy functionals [11]. For $\widehat{T}_{quad}$ we must also specify the set $M \subset D$ of basis functions. One should use $m \asymp n^{\frac{-2d}{4s+d}}$ basis elements. See Theorem 7 below for details.

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 in Theorem 5 (below). 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 Holder class. In translating back to the Holder assumption, we lose a small factor in the smoothness, since the Sobolev space only contains

the Holder space if the former is less smooth than the latter. Bickel and Ritov study kernel-based estimator for quadratic functionals that eliminates this factor [1], but the projection estimator is considerably simpler, both from a computational perspective and in terms of analysis.

For a pictorial understanding of the rates of convergence of the various estimators, 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 5, 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.

3.3 Some examples

An estimate of $T(p, q)$ can immediately be used to estimate the divergences mentioned above. Plugging the estimators into the definition of Renyi-α and Tsallis-α divergences, we immediately have the following corollary:

Corollary 3 (Estimating Renyi-α, Tsallis-α divergences)**.** *In our setting, as long as $D_\alpha(p, q) \geq c > 0$ for some constant c there exists an estimator $\hat{D}_\alpha$ for $D_\alpha(p, q)$ with convergence rate $O(n^{-1/2} + n^{\frac{-3s}{2s+d}})$. There exists an estimator $\hat{T}_\alpha$ for $T_\alpha(p, q)$ with rate $O(n^{-1/2} + n^{\frac{-3s}{2s+d}})$.*

The estimator referred to in the result is the quadratic estimator $\widehat{T}_{quad}$, plugged into the definitions of D_α and T_α. Unfortunately, since computing $\widehat{T}_{quad}$ requires numeric integration, it is somewhat impractical. On the other hand, the linear estimator $\widehat{T}_{lin}$ is computationally very simple, although its convergence rate is $O(n^{-1/2} + n^{\frac{-2s}{2s+d}})$, which is worse that that of $\widehat{T}_{quad}$.

For the L_2^2 divergence, instead of applying Theorem 2 directly, it is better to directly use the quadratic and bilinear estimators for the terms in the factorization. As a corollary of Theorem 7 below, we have:

Corollary 4 (Estimating L_2^2-divergence)**.** *Under our assumptions, there exists an estimator $\hat{L}$ for $L_2^2(p, q)$ with:*

$$\mathbb{E}_{X_1^n, Y_1^n}\left[|\hat{L} - L_2^2(p, q)|\right] = O(n^{-1/2} + n^{\frac{-4s}{4s+d}}) \tag{7}$$

The estimator referred to in the Corollary involves estimating $\int p^2$ and $\int q^2$ using an expression akin to Equation 3 and estimating $\int pq$ using Equation 2. In this case, the $b_{i,i'}$ terms in the squared-term estimator are equivalent to $\mathbf{1}[i = i']$ since the function $\psi(x) = 1$ and since $\{\phi_k\}$ is an orthonormal collection. Thus one does not have to perform numeric integration, meaning that the estimator is computationally attractive. In addition we do not have to train KDEs, removing the need for bandwidth selection, although we still must select the basis functions used in the projection.

3.4 Lower Bound

We now turn to the issue of *statistical efficiency*. Clearly $\widehat{T}_{pl}, \widehat{T}_{lin}$ 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 5. *As long as both $\alpha, \beta \neq 0, 1$, then:*

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

for $\epsilon > 0$ where $\gamma = \min\{4s/(4s+d), 1/2\}$.

It is straightforward to transform this result into a matching lower bound for Tsallis-α divergences. For Renyi-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, the above result 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 $n^{-\gamma}$ 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 $\sqrt{n}$-consistent estimator, which prevents us from achieving the $n^{-\gamma}$ 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 $n^{-\gamma}$ when $\alpha = \beta = 1$.

Comparing the bound in Theorem 5 to the rate achieved by $\widehat{T}_{quad}$ we see a gap in the non-smooth case ($s < d/4$). In this regime, our estimator converges at $n^{\frac{-3s}{2s+d}}$ which deviates from the lower bound by a polynomial factor. When estimating integral functionals of a single density, achieving the optimal rate of convergence in this regime involves further correction by the third order term in the expansion [8]. It seems as if the same ideas can be adapted to 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 [2, 9].

4 Proof Sketches

4.1 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 substantially affect the rate, which generalizes previous results to high dimension [2]. We do

not address the issue of boundary correction in our result, referring instead to Liu *et al.* and the references therein for details on how to appropriately adjust the estimator and the proofs [11].

Theorem 6. *Let f be a density that satisfies our assumption and suppose we have n samples $X_1^n \sim f$. 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 7. *Let f, g be densities belonging to $\Sigma(s, L)$ and let ψ be some bounded function. If $\theta = \int \psi(x) f^2(x)$ and $\hat{\theta}$ is given by Equation 3 where ϕ_k is the Fourier basis and:*

$$M = \{(k_1, \ldots, k_d) \in \mathbb{Z}^d || k_j| \leq \frac{1}{2} m^{1/d}\}, \;\; m \asymp n^{\frac{2d}{4s'+d}}$$

for $s' < s$. Then:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq C_1 \left(\frac{8}{n} + n^{\frac{-8s'}{4s'+d}} \left[2^{\frac{8s'}{d}} d/s + 3 \right] \right) \tag{8}$$

Where $C \leq 18||\psi||_\infty^2 \max\{1, ||f||_\infty^2\} \max\{1, L'^2\}$. Similarly if $\theta = \int \psi(x) f(x) g(x)$ and $\hat{\theta}$ is given by Equation 2 then the same setting of M, m give:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq C_2 \left(\frac{1}{n} + n^{\frac{-8s'}{4s'+d}} \left[2^{\frac{8s'}{d}} d/s + 1 \right] \right) \tag{9}$$

with $C_2 \leq 18||\psi||_\infty^2 \max\{1, ||f||_\infty ||g||_\infty\} \max\{1, L'^2\}$. In both results, $L' = dCL^2/(2\pi)^{2\lfloor s \rfloor}$ and $C = \sum_{l=0}^{\infty} 4^{l(s'-s)}$ is bounded as long as $s' < s$.

Theorem 2 follows from these results and an application of the triangle inequality.

4.2 Lower Bound

The first part of the lower bound is an application of Le Cam's method and generalizes a proof technique of Birge and Massart [2]. We begin with a specialization of a fairly standard result, which reduces the estimation problem to a simple-vs-simple hypothesis testing problem.

Lemma 8. *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) $h^2(P^n \times Q^m, \bar{G}^n \times Q^m) \leq \gamma < 2$ and (ii) $T(p, q) \geq 2\beta + T(g_\lambda, q)$ for all λ then:*

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

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

In constructing the perturbations, we partition the space $[0,1]^d$ into m cubes R_j and construct a function u_j (for each j) that is compactly supported on R_j. We then use perturbation $g_\lambda = p + K\sum_{j=1}^m \lambda_j u_j$ where K, m will be selected later. By appropriately selecting the functions, we can ensure that:

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

Ensuring Holder smoothness requires $K = O(m^{-s/d})$ and with this setting of K, making the Hellinger distance $O(1)$ requires $m = \Omega(n^{\frac{2d}{4s+d}})$. With these choices we can apply Lemma 8 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 Lemma in the appendix although related results appear in the literature [4]. 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 6 in deriving the lower bound.

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 n samples to train kernel density estimators $\hat{p}_n, \hat{q}_n$, the second n samples to compute $\hat{T}_n$. Armed with this 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 exactly a quadratic functional of the two densities. The precise estimator is $\hat{T}_n - C_2 T(\hat{p}_n, \hat{q}_n)$. The triangle inequality along with Theorem 6 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 contradict the minimax lower bound for estimating quadratic functionals of Holder smooth densities.

Some care must be taken in this argument, since the quadratic functional we are estimating is a random quantity that depends on the data. However, since the lower bound on estimating quadratic functionals holds for any bounded continuous functions, it also applies to any distribution over bounded continuous functions, which is precisely what we have here (by Theorem 6). We refer the interested reader to the appendix for the details of the proof.

5 Experiments

Our experimental study was focused on verifying our theoretical results. In particular, we are interested in examining the empirical rates of convergence of our estimators. To that effect, 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 rates and those in our theorems.

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 $-0.7, -0.75, -0.65, -0.6$. 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.

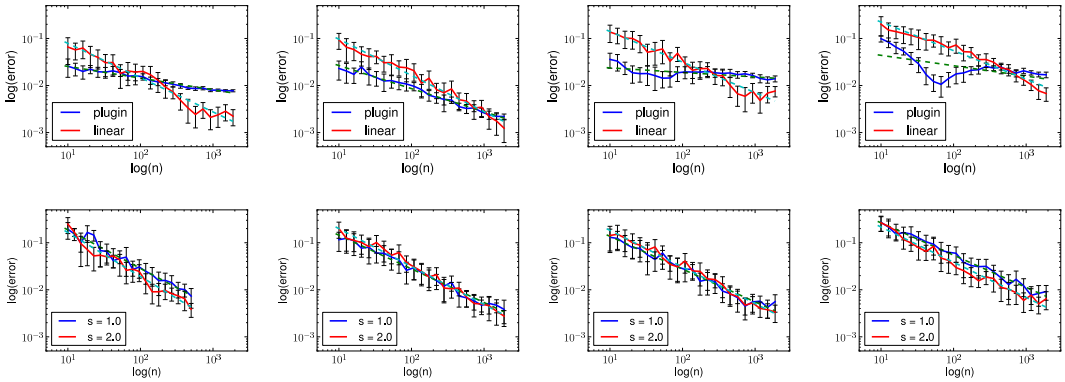

Figure 2: Top row: Rates of convergence for $\widehat{T}_{pl}, \widehat{T}_{lin}$ on a log-log scale for: left: $d = 1, s = 1$, second from left: $d = 1, s = 2$, second from right: $d = 2, s = 2$, right: $d = 2, s = 4$. Bottom Row: Left: Rate of convergence for $\widehat{T}_{quad}$ with $d = 1, s = 1.0, 2.0$. 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.

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 are converging fairly quickly and moreover we can verify that increasing the smoothness s does have some effect on the rate of convergence.

6 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? Comparing $\widehat{T}_{lin}$ to $\widehat{T}_{quad}$ in the divergence estimation setting demonstrates a computational and statistical tradeoff. On one hand $\widehat{T}_{lin}$ is quite simple but suffers statistically in comparison to $\widehat{T}_{quad}$, which is impractical computationally. 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. 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

AK is supported in part by a NSF Graduate Research Fellowship.

References

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

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

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

[4] David L. Donoho and Richard C. Liu. Geometrizing rates of convergence, ii. *The Annals of Statistics*, 1991.

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

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

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

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

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

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

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

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

[13] Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2010.

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

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

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

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

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

[19] Shashank Singh and Barnabás Póczos. Generalized exponential concentration inequality for rényi divergence estimation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

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

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

[22] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 1988.

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

[24] Aad W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

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

A Detailed Proofs of Upper Bound

Before diving into the auxiliary results of Section 4, let us first demonstrate that the remainder for the tth order von Mises expansion is $O(||p-\hat{p}||_{t+1}^{t+1}+||q-\hat{q}||_{t+1}^{t+1})$ under the assumption that $p, \hat{p}, q, \hat{q}$ are all bounded above and below. The tth order term in the von Mises expansion is.

$$\frac{1}{t!}\sum_{a=0}^{t}\binom{t}{a}\int\frac{\partial^t\hat{p}^\alpha(x)\hat{q}^\beta}{\partial\hat{p}(x)^a\partial\hat{q}(x)^{t-a}}(p(x)-\hat{p}(x))^a(q(x)-\hat{q}(x))^{t-a}dx =$$

$$\frac{1}{t!}\sum_{a=0}^{t}\binom{t}{a}\int\prod_{i=0}^{a}(\alpha-i)\prod_{i=0}^{t-a}(\beta-i)\hat{p}^{\alpha-a}(x)\hat{q}^{\beta-(t-a)}(x)(p(x)-\hat{p}(x))^a(q(x)-\hat{q}(x))^{t-a}dx$$

where $\prod_{i=0}^{0}a_i = 1$. If we are to take a $t-1$st order expansion, the remainder is of the same form as the tth term, except that the terms $\hat{p}^{\alpha-a}(x), \hat{q}^{\beta-(t-a)}(x)$ are replaced by functions $\xi_1^{\alpha-a}(x), \xi_2^{\beta-(t-a)}(x)$ for

some functions ξ_1, ξ_2 that are bounded between $p, \hat{p}$ and $q, \hat{q}$ respectively. In our setting, $p, q \in [\kappa_l, \kappa_u]$ and $\hat{p}, \hat{q} \in [\kappa_l - \epsilon, \kappa_u + \epsilon]$ so ξ_1, ξ_2 are bounded functions. With this bound, we can simplify the remainder term R_{t-1} to:

$$R_{t-1} \leq C(\alpha, \beta, \kappa_l, \kappa_u, \epsilon, t) \frac{1}{t!} \sum_{a=0}^{t} \binom{t}{a} \int |p(x) - \hat{p}(x)|^a |q(x) - \hat{q}(x)|^{t-a} dx$$

Looking at the integral pointwise, either $|p(x) - \hat{p}(x)| \leq |q(x) - \hat{q}(x)|$ in which case the expression is upper bounded by $|q(x) - \hat{q}(x)|^t$ or the opposite is true in which case it is bounded by $|p(x) - \hat{p}(x)|^t$. Either way, we can upper bound the integral by the sum. This gives:

$$R_{t-1} \leq C(\alpha, \beta, \kappa_l, \kappa_u, \epsilon, t) \frac{2^t}{t!} \left(||p - \hat{p}||_t^t + ||q - \hat{q}||_t^t \right)$$

In many cases, the constant can be worked out:

1. If $\alpha = \beta = 1$, then $R_1 = \alpha\beta$ while $R_2, \ldots, = 0$.
2. If $\alpha, \beta > 0, \alpha + \beta = 1$ as in the Renyi Divergence, $R_2 = 1$ while $R_3 = \frac{5}{6} \kappa_\epsilon^{-2} \alpha\beta$ where $\kappa_\epsilon = \min\{\kappa_l - \epsilon, (\kappa_u + \epsilon)^{-1}\}$.

The first order von Mises expansion is:

$$\begin{aligned}
T(p, q) &= T(\hat{p}, \hat{q}) + \int \frac{\partial \hat{p}^\alpha(x) \hat{q}^\beta(x)}{\partial \hat{p}(x)} (p(x) - \hat{p}(x)) + \int \frac{\partial \hat{p}^\alpha(x) \hat{q}^\beta(x)}{\partial \hat{q}(x)} (q(x) - \hat{q}(x)) + O(||p - \hat{p}||_2^2 + ||q - \hat{q}||_2^2) \\
&= T(\hat{p}, \hat{q}) + \int \alpha \hat{p}^{\alpha-1}(x) \hat{q}^\beta(x) (p(x) - \hat{p}(x)) + \int \beta \hat{p}^\alpha(x) \hat{q}^{\beta-1}(x) (q(x) - \hat{q}(x)) + O(||p - \hat{p}||_2^2 + ||q - \hat{q}||_2^2) \\
&= (1 - \alpha - \beta) T(\hat{p}, \hat{q}) + \int \alpha \hat{p}^{\alpha-1}(x) \hat{q}^\beta(x) p(x) + \int \beta \hat{p}^\alpha(x) \hat{q}^{\beta-1}(x) q(x) + O(||p - \hat{p}||_2^2 + ||q - \hat{q}||_2^2) \\
&= C_1 T(\hat{p}, \hat{q}) + \theta_{1,1}^p + \theta_{1,1}^q + R_2
\end{aligned}$$

The second order expansion is computed similarly. The three second order terms are:

$$\begin{aligned}
&\frac{1}{2} \int \alpha(\alpha - 1) \hat{p}^{\alpha-2}(x) \hat{q}^\beta(x) (p(x) - \hat{p}(x))^2 \\
&\int \alpha\beta \hat{p}^{\alpha-1}(x) \hat{q}^{\beta-1}(x) (p(x) - \hat{p}(x))(q(x) - \hat{q}(x)) \\
&\frac{1}{2} \int \beta(\beta - 1) \hat{p}^\alpha(x) \hat{q}^{\beta-1}(x) (q(x) - \hat{q}(x))^2
\end{aligned}$$

Adding these together along with the linear terms, expanding and regrouping terms we get:

$$T_2(p, q) = C_2 T(\hat{p}, \hat{q}) + \sum_{\substack{i=1,2 \\ f=p,q}} \theta_{2,i}^f + \theta_{2,2}^{p,q} + R_3$$

Let us now prove the the auxiliary results stated in Section 4

A.1 Proof of Theorem 6

The truncated kernel density estimator takes the following form: We select a parameter $\epsilon > 0$. If $\tilde{f}$ is the usual kernel density estimator for f, we set $\hat{f}(x) = \tilde{f}(x)$ if $\tilde{f}(x) \in [\kappa_l - \epsilon, \kappa_u + \epsilon]$ and otherwise we set $\hat{f}(x) = f_0(x)$ for some fixed function bounded between κ_l, κ_u.

We will require that the kernel $K : \mathbb{R}^d \to \mathbb{R}$ satisfies:

1. supp$K \in (-1, 1)^d$
2. $\int K(x)dx = 1$ and $\int \prod_i x_i^{p_i} K(x)dx = 0$ for all tuples $p = (p_1, \ldots, p_d)$ with $\sum p_i \leq \lfloor s \rfloor$.

Let us first establish the rate of convergence of $\tilde{f}$ the regular kernel density estimator in ℓ_p^p. Let us call the kernel $K : \mathbb{R}^d \to \mathbb{R}$ so that $\tilde{f}(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$. Denote by $\bar{f}(x) = \mathbb{E}[\tilde{f}(x)] = \mathbb{E}_{X \sim f}[\frac{1}{h^d} K\left(\frac{x - X}{h}\right)]$. Then:

$$\mathbb{E}[||\tilde{f} - f||_p^p] \leq 2^p \left(\mathbb{E}[||\tilde{f} - \bar{f}||_p^p] + ||\bar{f} - f||_p^p \right)$$

To bound the first term, let us write $\eta_i(x) = \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right) - \mathbb{E}_{X \sim f}[\frac{1}{h^d} K\left(\frac{x - X}{h}\right)]$. Exchanging integrals, we can look at fixed x and we have:

$$\mathbb{E}|\tilde{f}(x) - \bar{f}(x)|^p = \mathbb{E}|\frac{1}{n} \sum_{i=1}^n \eta_i(x)|^p \leq \left(\frac{1}{n^{2p}} \mathbb{E}[(\sum_{i=1}^n \eta_i(x))^{2p}] \right)^{1/2} \tag{10}$$

If we expand the expectation and drop the terms that vanish we get all terms of the form:

$$\sum_{i_1 \neq i_2 \ldots \neq i_t = 1}^n \prod_{j=1}^t \binom{p - \sum_{k=1}^{j-1} p_k}{p_j} \eta_{i_j}(x)^{p_j} = \frac{n!}{(n-t)!} \prod_{j=1}^t \binom{p - \sum_{k=1}^{j-1} p_k}{p_j} \eta_{i_j}(x)^{p_j}$$

where $1 \leq t < p$, $\sum p_j = p$ and $p_j \neq 1 \forall j$. That is, we pick a term in the polynomial with t unique variables, then assign powers p_j to each of the terms, then count the number of ways to assign those powers to those terms (which results in the binomial coefficients). Since $\mathbb{E}[\eta_j(x)] = 0$, The terms where there exists a $p_j = 1$ are all zero.

By linearity of expectation of independence, we therefore need to control $\mathbb{E}[|\eta_i(x)|^q]$ for $2 \leq q \leq p$. Applying Jensen's inequality, we get:

$$\mathbb{E}[|\eta_i(x)|^q] \leq 2^q \mathbb{E}[|\frac{1}{h^d} K\left(\frac{X - x}{h}\right)|] \leq 2^q \kappa_u h^{-(q-1)d} \int |K(u)|^q du$$

where the last expression comes from expanding the integral, performing a substitution and bounding $f(x) \leq \kappa_u$. So we can bound by $C(q, \kappa_u, K) h^{-(q-1)d}$. Plugging this into the expression above, we get:

$$\frac{n!}{(n-t)!} C(p_1^j, \kappa_u, K) h^{-pd+td} \leq n^t C'(p_1^j, \kappa_u, K) h^{-pd+td} \leq C'(p_1^j, \kappa_u, K) \frac{n^p}{h^{pd}}$$

The second inequality holds for n sufficiently large. The third inequality holds whenever $nh^d \geq 1$ which will be true for n sufficiently large, given our setting of h. To summarize, all of the terms can be upper bounded by something that is $\asymp (n^p / h^{pd})$ and there are a constant-in-p number of terms. Plugging this into Equation 10 we get

$$\mathbb{E}[||\tilde{f} - \bar{f}||_p^p] \leq C(nh^d)^{-p/2} \tag{11}$$

Remark 9. *The constant here has exponential dependence on p but we are only concerned with cases where p is a small constant (at most 4).*

As for the bias (note that x, u, t are all d-dimensional vectors here):

$$|\bar{f}(x) - f(x)| = \int \frac{1}{h^d} K\left(\frac{x-t}{h}\right) f(t)dt - f(x) = \int (f(x-uh) + f(x))K(u)du$$

Let us define $m = \lfloor s \rfloor$. Taking the $m-1$st order von Mises expansion of $f(x+uh)$ about $f(x)$ we get terms of the form:

$$= \sum_{r_1,\ldots,r_d | \sum r_i \leq m-1} \frac{1}{|r|!} D^r f(x) h^{|r|} \int \prod_i u_i^{r_i} K(u)du = 0$$

by our assumption on K. The remainder term, gives us:

$$\sum_{r_1,\ldots,r_d | \sum r_i = m} \frac{h^m}{m!} \int \xi(r,x,uh) \prod_i u_i^{r_i} K(u)du \leq \sum_{r_1,\ldots,r_d | \sum r_i = m} \frac{Lh^s}{m!} \int ||u||^{s-m} \prod_i u_i^{r_i} K(u)du$$

which is $C(m,K,d)Lh^s$. Here the function ξ is between $D^r f(x)$ and $D^r f(x-uh)$ and to reach the last expression, we use the fact that $|D^r f(x) - D^r f(x-uh)| \leq L||uh||^{s-r}$, i.e. the Holderian assumption on f. In applying the Holderian assumption, there is another term of the form $D^r f(x) \int \prod_i u_i^{r_i} K(u)du$ which is zero by the assumption on K. Equipped with this bound, we can bound the bias:

$$||\bar{f} - f||_p^p \leq C(m,K,d)L^p h^{ps} \tag{12}$$

In trading off the bias and the variance, we set $h \asymp n^{\frac{-1}{2s+d}}$ and see that the rate of convergence is $\mathbb{E}[||\tilde{f} - f||_p^p = O(n^{\frac{-ps}{2s+d}})$.

To prove Theorem 6, we just have to show that truncation does not significantly affect the rate. Fix a parameter $\epsilon > 0$ and define $S_\epsilon = \{x : \kappa_l - \epsilon \geq \tilde{f}(x) \leq \kappa_u + \epsilon\}$. We have:

$$\begin{aligned}
\mathbb{E}[||\hat{f} - f||_p^p] &= \mathbb{E}\left[\int_{S_\epsilon} |\tilde{f}(x) - f(x)|^p dx + \int_{S_\epsilon^C} |f_0(x) - f(x)|dx\right] \\
&\leq \mathbb{E}\left[||\tilde{f} - f||_p^p\right] + ||f_0 - f||_\infty^p \mathbb{E}\left[\int \mathbf{1}[x \notin S_\epsilon]dx\right] \\
&= \mathbb{E}\left[||\tilde{f} - f||_p^p\right] + ||f_0 - f||_\infty^p \int \mathbb{P}_{X_1^n}[x \notin S_\epsilon]dx
\end{aligned}$$

so we must control the probability that $x \notin S_\epsilon$. This can be done via Bernstein's inequality. First observe that the bias $|\bar{f} - f| \to 0$ with our choice of h so that for sufficiently large n, $\sup_x \bar{f}(x) - f(x) \leq \epsilon/2$. Once this happens, it is clear that $x \notin S_\epsilon$ implies that $\tilde{f}(x) - \bar{f}(x) \geq \epsilon/2$. Therefore:

$$\mathbb{P}[x \notin S_\epsilon] \leq \mathbb{P}[|\tilde{f}(x) - \bar{f}(x)| \geq \epsilon/2] = \mathbb{P}[|\frac{1}{n}\sum_i \eta_i(x)| \geq \epsilon/2] \leq 2\exp\left(\frac{-nh^d\epsilon^2/4}{\kappa_u||K||_2^2 + \frac{1}{3}||K||_\infty \epsilon}\right)$$

This last inequality is an application of Bernstein's inequality noting that $|\eta_i(x)| \leq \frac{2}{h^d}||K||_\infty$ and $\text{var}(\eta_i(x)) \leq h^{-d}\kappa_u||K||_2^2$ since:

$$\text{var}(\eta_i(x)) \leq \mathbb{E}_{X_i \sim f}\left[\frac{1}{h^{2d}} K^2(\frac{X_i - x}{h})\right] = \frac{1}{h^d}\int K^2(u) f(x+hu)du \leq h^{-d}\kappa_u||K||_2^2$$

Using our definition $h \asymp n^{\frac{-1}{2s+d}}$ and using the fact that ϵ is some constant $\mathbb{P}[x \notin S_\epsilon] \leq 2\exp(-Cn^{\frac{2s}{2s+d}})$. Plugging this bound in above, we have:

$$\mathbb{E}\left[||\hat{f} - f||_p^p\right] \leq \mathbb{E}\left[||\tilde{f} - f||_p^p\right] + 2||f_0 - f||_\infty^p \exp\left(-Cn^{\frac{2s}{2s+d}}\right) \text{vol}([0,1]^d) = O(n^{\frac{-ps}{2s+d}})$$

Since the second term goes to zero exponentially quickly in n. This proves the theorem.

A.2 Convergence Rate for Estimating Linear Functionals

It is trivial to derive the convergence rate for estimating linear functionals:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \frac{1}{n}(\mathbb{E}[\psi^2(X)] - \mathbb{E}[\psi(X)]^2) \leq 2||\psi||_\infty^2/n$$

And by Jensen's inequality, we have $\mathbb{E}[|\hat{\theta} - \theta|] \leq \sqrt{\mathbb{E}[(\hat{\theta} - \theta)^2]}$, so the rate of convergence is $\sqrt{2}||\psi||_\infty/\sqrt{n}$.

A.3 Proof of Theorem 7

For the quadratic terms, we use a result of Laurent [9]:

Theorem 10 ([9])**.** *Let X_1^n be i.i.d random variables with common density f that belongs to some Hilbert Space $L^2(d\mu)$. Let $\{\phi_i\}_{i\in D}$ be an orthonormal basis of $L^2(d\mu)$. We assume that f is uniformly bounded and belongs to the ellipsoid $\mathcal{E} = \{\sum_{i\in D} a_i\phi_i : \sum_{i\in D} |a_i^2/c_i^2| \leq 1\}$. Let ψ be bounded function and define $\theta = \int \psi(x)f(x)\mu(dx)$ and $\hat{\theta}$ as above. If we can find a subset M_n of size $|M_n| = m$ then whenever $n \geq n_0$ (some absolute constant) we have:*

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \textit{Bias}^2(\hat{\theta}) + \textit{Var}(\hat{\theta}) \leq ||\psi||_\infty^2 \sup_{i\notin M_n} |c_i|^4 + 72||\psi||_\infty^2||f||_\infty^2\left(\frac{2}{n} + \frac{m}{n^2}\right) \tag{13}$$

For the bi-linear term $\theta_{2,2}^{p,q}$ we have the following theorem:

Theorem 11. *Let X_1^n be i.i.d random variables with common density f and Y_1^n be i.i.d. with common density g. Let f, g belong to some Hilbert space $L^2(d\mu)$ and let $\{\phi_i\}_{i\in D}$ be an orthonormal basis for $L^2(d\mu)$. We assume that f, g are uniformly bounded and both belong to the ellipsoid $\mathcal{E} = \{\sum_{i\in D} a_i\phi_i : \sum_{i\in D} |a_i^2/c_i^2| \leq 1\}$. Let $\theta = \int \psi(x)f(x)g(x)\mu(dx)$ and $\hat{\theta}$ be defined by Equation 2 If we can find a subset M_n of size $|M_n| = m$ then whenever $n \geq n_0$ (some absolute constant) we have:*

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \textit{Bias}^2(\hat{\theta}) + \textit{Var}(\hat{\theta}) \leq ||\psi||_\infty^2 \sup_{i\notin M_n} |c_i|^4 + ||\psi||_\infty^2||f||_\infty||g||_\infty\left(\frac{2}{n} + \frac{m+1}{n^2}\right) \tag{14}$$

Proof. The bias is:

$$\mathbb{E}[\hat{\theta}] - \theta = \int \sum_{i\in M} \alpha_i\phi_i(x)\psi(x)g(x)dx - \int \psi(x)f(x)g(x) = \int \psi(x)\left(\mathcal{P}_M f(x) - f(x)\right)g(x)dx$$

Where $\alpha_i = \int \phi_i(x) f(x)$ and $\mathcal{P}_M f$ is the projection of f onto the basis functions in M. Define $\beta_i = \int \phi_i(x) g(x)$. If f, g live in the ellipsoid $\mathcal{E} = \{\sum a_i \phi_i | \sum |a_i|^2/|c_i|^2 \leq L\}$ then:

$$\text{Bias}^2(\hat{\theta}) = \left(\sum_{i \notin M} \alpha_i \int \psi(x) g(x) \phi_i(x) dx \right)^2 \leq ||\psi||_\infty^2 \left(\sum_{i \notin M} \alpha_i \beta_i \right)^2$$

The term inside the parenthesis can be bounded as:

$$\sum_{i \notin M} \alpha_i \beta_i \leq \frac{1}{2} \sup_{i \notin M} |c_i|^2 \sum_{i \notin M} \frac{|\alpha_i|^2 + |\beta_i|^2}{|c_i|^2} \leq L \sup_{i \notin M} |c_i|^2$$

So the bias is $\text{Bias}^2(\hat{\theta}) \leq ||\psi||_\infty^2 L^2 \sup_{i \notin M} |c_i|^4$.

As for the variance, let us define $Q(x)$ to be the m-dimensional vector of functions $\phi_i(x) - \alpha_i$ and $R(x)$ to be the m-dimensional vector of functions $\phi_i(x)\psi(x) - \int \psi \phi_i g$. Further define A, B to be the m-dimensional vectors with ith components $\alpha_i = \int \phi_i f$ and $\beta_i = \int \psi \phi_i g$ respectively. Then our estimator can alternatively be written as:

$$\hat{\theta} = \underbrace{\frac{1}{n^2} \sum_{j,k} Q(X_j)^T R(Y_k)}_{T_1} + \underbrace{\frac{1}{n} \sum_j Q(X_j)^T B}_{T_2} + \underbrace{\frac{1}{n} \sum_k A^T R(Y_k)}_{T_3} - A^T B$$

Notice that Q, R are centered functions. Since Xs are independent of the Ys, $\text{Cov}(T_2, T_3) = 0$. Since T_2 is independent of Y and $\mathbb{E}[R(Y_k)] = 0$, we see that $\text{Cov}(T_1, T_2) = 0$. Similarly, $\text{Cov}(T_1, T_3) = 0$.

Therefore,

$$\text{Var}(\hat{\theta}) = \text{Var}(T_1) + \text{Var}(T_2) + \text{Var}(T_3)$$

Let us analyze T_1. By independence,

$$\begin{aligned}
\text{Var}(T_1) &= \frac{1}{n^2} \text{Var}(Q(X_1)^T R(Y_1)) = \frac{1}{n^2} \sum_{i,i' \in M} \int \phi_i(x) \phi_{i'}(x) \phi_i(y) \phi_{i'}(y) \psi(y)^2 f(x) g(y) dx dy \\
&- \int \alpha_i \alpha_{i'} \phi_i(y) \phi_{i'}(y) \psi(y)^2 g(y) dy - \int \beta_i \beta_{i'} \phi_i(x) \phi_{i'}(x) f(x) + \alpha_i \alpha_{i'} \beta_i \beta_{i'} \\
&\leq \frac{1}{n^2} \sum_{i,i' \in M} \int \phi_i(x) \phi_{i'}(x) \phi_i(y) \phi_{i'}(y) \psi(y)^2 f(x) g(y) dx dy + \frac{1}{n^2} (\sum_i \alpha_i \beta_i)^2 \\
&= \frac{1}{n^2} \int \left(\sum_{i \in M} \phi_i(x) \phi_i(y) \right)^2 \psi(y)^2 f(x) g(y) dx dy + \frac{1}{n^2} (\sum_i \alpha_i \beta_i)^2 \\
&\leq \frac{||\psi||_\infty^2 ||f||_\infty ||g||_\infty}{n^2} \int \left(\sum_{i \in M} \phi_i(x) \phi_i(y) \right)^2 dx dy + \frac{1}{n^2} \left(\sum_i \alpha_i^2 \right) \left(\sum_i \beta_i^2 \right) \\
&\leq \frac{||\psi||_\infty^2 ||f||_\infty ||g||_\infty m}{n^2} + \frac{1}{n^2} (\int f^2)(\int g^2 \psi^2) \leq \frac{||\psi||_\infty^2 ||f||_\infty ||g||_\infty (m+1)}{n^2}
\end{aligned}$$

To arrive at the third line, notice that the cross terms are non-negative, since $\sum_{i,i'} \alpha_i \alpha_{i'} \phi_i(y) \phi_{i'}(y) = \left(\sum_i \alpha_i \phi_i(y)\right)^2$ (and analogously for the other cross term). Therefore we can simply omit them and provide

an upper bound. To go from the fourth to fifth lines, we use Holder's inequality on the first term and Cauchy-Schwarz on the right term. Notice that the expression involving $\phi_i(x)\phi_i(y)$ is positive, so we can drop the absolute values in the ℓ_1 norm term of Holder's inequality. To arrive at the fifth line, we expand out the square and use the fact that ϕ_is are orthornormal.

For T_2 again by independence we have:

$$\begin{aligned}
\text{Var}(T_2) &= \frac{1}{n}\text{Var}(Q(X_1)^T B) = \mathbb{E}[\left(\sum_{i\in M}(\phi_i(X_1)-\alpha_i)\int \psi\phi_i g\right)^2] \\
&= \sum_{i,i'\in M}\int \phi_i(x)\phi_{i'}(x)f(x)\int\psi\phi_i g\int\psi\phi_{i'}g - \int\alpha_i\psi\phi_i g\int\alpha_{i'}\psi\phi_{i'}g \\
&= \int(\sum_{i\in M}\beta_i\phi_i(x))^2 f(x) - \left(\int(\mathcal{P}_M)\psi g\right)^2 \leq \int(\mathcal{P}_M(\psi g))^2 f
\end{aligned}$$

Here the last inequality follows from the fact that $\beta_i = \int\psi\phi_i g$ is the ith fourier coefficient of ψg so $\sum_i \beta_i\phi_i$ is the projection onto M. Of course this quantity is bounded by:

$$\text{Var}(T_2) \leq \frac{1}{n}||f||_\infty\int\psi^2(x)g^2(x)dx \leq \frac{||\psi||_\infty^2||f||_\infty||g||_\infty}{n}$$

Essentially the same argument reveals that T_3 is bounded in the same way.

$$\begin{aligned}
\text{Var}(T_3) &= \frac{1}{n}\text{Var}(A^T R(Y_1)) \leq \frac{||\psi||_\infty^2}{n}\sum_{i,i'}\alpha_i\alpha_{i'}\left[\int\phi_i(y)\phi_{i'}(y)g(y)dy - \int\phi_i g\int\phi_{i'}g\right] \\
&= \frac{1}{n}||\psi||_\infty^2\left[\int(\mathcal{P}_M f)^2 g - (\int(\mathcal{P}_M f)g)^2\right] \leq \frac{||\psi||_\infty^2||f||_\infty||g||_\infty}{n}
\end{aligned}$$

So the variance of the estimator is:

$$\text{Var}(\hat{\theta}) \leq ||\psi||_\infty^2||f||_\infty||g||_\infty\left(\frac{m+1}{n^2}+\frac{2}{n}\right)$$

□

Both the quadratic and bilinear terms exhibit the same dependence on $\sup_{i\notin M_n}|c_i|, m, n$ so choosing M_n appropriately will give the rate of convergence for both terms. To establish Theorem 7 we work with the fourier basis $\{\phi_k\}_{k\in\mathbb{Z}^d}$ where $\phi_k(x) = e^{2\pi i k^T x}$ and the Sobolev class $\mathcal{W}(s,L)$ defined by:

$$\mathcal{W}(s,L) = \left\{ f = \sum_{k\in\mathbb{Z}^d} a_k\phi_k \;\middle|\; \sum_{k\in\mathbb{Z}^d}(\sum_{j=1}^d |k_j|^{2s})|a_k|^2 \leq L \right\} \tag{15}$$

In Lemma 14 we show that the class $\mathcal{W}(s',L')$ contains $\Sigma(s,L)$ as long as $s' < s$ and with appropriate choice of L'. For now let us work in $\mathcal{W}(s',L')$.

Let us choose:

$$M_n = \{k\in\mathbb{Z}^d | |k_j| \leq \frac{1}{2}m^{1/d}\},\;\; m = \left(18\frac{d}{s'}2^{4s'/d}n^{-2}\right)^{\frac{-d}{4s'+d}} \asymp n^{\frac{2d}{4s'+d}}$$

Thinking of M_n as an integer lattice with side lengths $m^{1/d}$ we see that $|M_n| = m$. Moreover $\sup_{i \notin M_n} |c_i|^4 = L^2(2/m)^{4s'/d}$. For the quadratic terms, this results in the bound:

$$\begin{aligned}\mathbb{E}[(\hat{\theta} - \theta)^2] &\leq ||\psi||_\infty^2 \left(L^2(2/m)^{4s'/d} + 72||f||_\infty^2 m/n^2 + 144||f||_\infty^2/n \right) \\ &\leq ||\psi||_\infty^2 \max\{1, ||f||_\infty^2\} \max\{L^2, 1\} \left((2/m)^{4s'/d} + 72m/n^2 + 144/n \right)\end{aligned}$$

And plugging in our definition of m followed by some algebraic simplifications, we get

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq 18||f||_\infty^2 \max\{1, ||p||_\infty^2\} \max\{L^2, 1\} \left(\frac{8}{n} + n^{\frac{-8s'}{4s'+d}} \left[2^{\frac{8s'}{d}} d/s' + 3 \right] \right)$$

For the bilinear terms, plugging into Theorem 11, we get

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq ||\psi||_\infty^2 \max\{1, ||f||_\infty ||g||_\infty\} \max\{L^2, 1\} \left((2/m)^{4s'/d} + m/n^2 + 3/n \right)$$

which when we plug in for m we get:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq ||\psi||_\infty^2 \max\{1, ||f||_\infty ||g||_\infty\} \max\{L^2, 1\} \left(3/n + n^{\frac{-8s'}{4s'+d}} \left[18 \times 2^{8s'/d} d/s' + 1 \right] \right)$$

B Proofs of Corollaries 4 and 3

The proof of Corollary 4 is immediate given the decomposition $||p - q||_2^2 = \int p^2 + \int q^2 - 2 \int pq$ and the Theorem 7.

For Corollary 3, if we use our estimator $\hat{T}$ for $T(p, q) = \int p^\alpha q^{1-\alpha}$ we can plug $\hat{T}$ into the definition of Renyi-divergence to obtain an estimator $\hat{D}_\alpha$. The rate of convergence is:

$$\mathbb{E}[|\hat{D}_\alpha - D_\alpha|] = \frac{1}{\alpha - 1} \mathbb{E}\left[\log\left(\hat{T}/T \right) \right] \leq \frac{1}{\alpha - 1} \mathbb{E}\left[\log(1 + |T - \hat{T}|/T) \right] \leq \frac{1}{\alpha - 1} c n^{-\gamma}/T(p, q)$$

where γ is the rate of convergence of our estimator. This is $O(n^{-\gamma})$ as long as $T(p, q) \geq c > 0$.

C Detailed Proofs for Lower Bound

To prove the main part of the theorem, the $\Omega(n^{\frac{-4s}{4s+d}})$ rate, we will apply Le Cam's Lemma. We decompose the proof into three parts. In the first part, we establish a variant of Le Cam's lemma that we will use. In the second part, we show how the properties established on the functions u_j, $j \in [p]$ allow us to apply Le Cam's lemma and establish the theorem. In the third part, we prove the existence of such functions u_j. We conclude this section with a proof of the $\Omega(n^{-1/2})$ when $s > d/4$.

C.1 Proof of Lemma 8

Proof. Define $\Theta_0 = \{g \in \Theta | T(g, q) \geq T(p, q)\}$ and $\Theta_1 = \{g \in \Theta | T(g, q) \leq T(p, q) - 2\beta\}$ so that all $g_\lambda \in \Theta_1$ while $p \in \Theta$. Let $\tilde{\Theta}_i = \text{conv}(\{G^n \times Q^m | g \in \Theta_i\})$ and consider the simple versus simple testing problem between $P \in \Theta_0$ and $g_\lambda \in \Theta_1$. The minimax probability of error p_e of such a test is lower

bounded by $\frac{1}{2}(1-\sqrt{h^2(P,G_\lambda)(1-h^2(P,G_\lambda))/4})$ by Theorem 2.2. of [23]. So for any test statistic ψ, taking supremum over $P \in \Theta_0, G \in \Theta_1$ we have:

$$\sup_{\theta_{0,1} \in \tilde{\Theta}_{0,1}} p_e(\psi; \theta_0, \theta_1) \geq \frac{1}{2}\left[1 - \sqrt{\gamma(1-\gamma/4)}\right]$$

since $P^n \times Q^m \in \tilde{\Theta}_0$ and $\bar{G}^n \times Q^m \in \tilde{\Theta}_1$ by convexity. The same bound holds for after taking infimum over ψ. Finally, if we make an error in the testing problem, we suffer loss at least β which results in the statement in the Lemma. □

C.2 The properties of u_j

Recall that in our proof we partition $[0,1]^d$ into m cubes $R_1, \ldots, R_m$ of side length $m^{-1/d}$. On each bin we require a function u_j such that:

$$\text{supp}(u_j) \subset \{x | B(x,\epsilon) \in R_j\},\ ||u_j||_2^2 = \Theta(m^{-1}),\ \int_{R_j} u_j = 0,\ \int_{R_j} p^{\alpha-1} q^\beta u_j = 0, ||D^r u_j||_\infty \leq m^{r/d}$$

Where the last inequality needs to hold for all tuples r with $\sum_j r_j \leq s+1$. Using these functions u_j, we construct the alternatives $g_\lambda = p + K\sum_{\lambda \in \Lambda} \lambda_j u_j \mathbf{1}_{R_j}$ for all $\lambda \in \Lambda = \{-1,1\}^m$. The third property above ensures that g_λ is a valid density.

Properties 2, 4, and 5 ensure that $T(p,q) - T(g_\lambda, q)$ is sufficiently large. Indeed, by the von Mises expansion:

$$\begin{aligned} T(p,q) - T(g_\lambda, q) &= K\alpha \sum_{j=1}^m \lambda_j \int_{R_j} p^{\alpha-1} q^\beta u_j + K^2 \alpha(\alpha-1) \sum_{j=1}^m \int_{R_j} \xi_p^{\alpha-2}(x) q^\beta(x) u_j^2(x) dx \\ &\geq c_0 K^2 \sum_{j=1}^m ||u_j||_2^2 \geq c_1 K^2 \end{aligned}$$

Here ξ is the function in the Taylor's remainder theorem, bounded between p and g_λ, both of which are bounded above and below. g_λ is bounded above and below by property 5 since $||D_0 u_j||_\infty = ||u_j||_\infty \leq 1$ which means that $g_\lambda \in [1-K, 1+K]$. K will be decreasing with n, so this quantity will certainly be bounded for n large enough. Property 2 allows us to arrive at the last line since each u_j is orthogonal to the derivative of T, so the first term in the expansion is zero. Finally Property 4 allows us to lower bound $||u_j||_2^2$.

Property 2 is also critical in ensuring that $h^2(P^n \times Q^m, \bar{G}^n \times Q^m)$ is small through the following Theorem of Birge and Massart [2].

Theorem 12 ([2]). *Consider a set of densities p and $p_\lambda = p[1 + \sum_j \lambda_j v_j(x)]$ for $\lambda \in \Lambda = \{-1,1\}^m$. Suppose that (i) $||v_j||_\infty \leq 1$ (ii) $||1_{R_j^C} v_j||_1 = 0$, (iii) $\int v_j p = 0$ and (iv) $\int v_j^2 p = \alpha_j > 0$ all hold with:*

$$\alpha = \sup_j ||v_j||_\infty,\ s = n\alpha^2 \sup_j P(R_j),\ c = n \sup_j \alpha_j$$

Define $\bar{P}_\Lambda^n = \frac{1}{|\Lambda|}\sum_{\lambda \in \Lambda} P_\lambda^n$. Then

$$h^2(P^n, \bar{P}_\Lambda^n) \leq C(\alpha, s, c) n^2 \sum_{j=1}^m \alpha_j^2$$

where $C < 1/3$ is continuous and non-decreasing with respect to each argument and $C(0,0,0) = 1/16$.

In bounding the Hellinger distance $h^2(P^n \times Q^m, \bar{G}^n \times Q^m)$ we first use the property that hellinger distance decomposes across product measures:

$$h^2(P^n \times Q^m, \bar{G}^n \times Q^m) = 2\left(1 - (1 - h^2(P^n, \bar{G}^n)/2)(1 - h^2(Q^m, Q^m)/2)\right) = h^2(P^n, \bar{G}^n)$$

If we define $v_j(x) = Ku_j(x)/p(x)$ then we have $g_\lambda = p[1 + \sum_j \lambda_j v_j]$ as needed by Theorem 12. We immediately satisfy requirements 1, 2, and 3 and we have $\int v_j^2 p = K^2 \int u_j^2/p \leq K^2\kappa_l/m = \alpha_j$. Thus in applying the theorem we have:

$$h^2(P^n \times Q^m, \bar{G}^n \times Q^m) \leq (1/3)n^2 \sum_{j=1}^{m} \alpha_j^2 \leq \frac{Cn^2K^4}{m}$$

Property 1 and 5 ensure that $g_\lambda \in \Sigma(s, L)$ via the following argument. Defining $u_\lambda = K\sum_j \lambda_j u_j$, we will first show that u_λ is holder smooth and g_λ will be holder by a final application of the triangle inequality. For u_λ, fix r with $\sum_j r_j = s$ and fix x, y. Let x_1 be the boundary point of R_j, the bin containing x along the line between x and y and let y_1 be the analogous boundary point for y.

$$\begin{aligned}
|D^r u_\lambda(x) - D^r u_\lambda(y)| &\leq |D^r u_\lambda(x) - D^r u_\lambda(x_1)| + |D^r u_\lambda(x_1) - D^r u_\lambda(y_1)| + |D^r u_\lambda(y_1) - D^r u_\lambda(y)| \\
&= |D^r u_\lambda(x) - D^r u_\lambda(x_1)| + |D^r u_\lambda(y_1) - D^r u_\lambda(y)| \\
&= \int_{\gamma(x,x_1)} \nabla D^r u_\lambda(z)dz + \int_{\gamma(y,y_1)} \nabla D^r u_\lambda(z)dz \\
&\leq K||D^{r+1}u_j||_\infty(||x - x_1||_2 + ||y - y_1||_2) \\
&\leq Km^{(r+1)/d}\left(||x - x_1||_2^{s-r}||x - x_1||_2^{1-(s-r)} + ||y - y_1||_2^{s-r}||y - y_1||_2^{1-(s-r)}\right) \\
&\leq Km^{(r+1)/d}\sqrt{d}m^{-\frac{1-(s-r)}{d}}\left(||x - x_1||_2^{s-r} + ||y - y_1||_2^{s-r}\right) \\
&\leq Km^{s/d}\sqrt{d}||x - y||_2^{s-r} \leq L||x - y||_2^{s-r}
\end{aligned}$$

The first line is an application of the triangle inequality. In the second line we use that u_λ is zero and has all derivatives equal to zero on the boundaries of the cubes R_j. This follows from the fact that u_j is not supported in the band around the border of R_j. The third line is an application of the fundamental theorem of calculus, $\gamma(x, x_1)$ is the path between x and x_1. The fourth line follows from Holder's inequality, we replace each derivative with its supremum and are left with just the path integral, which simplifies to the length of the path, i.e. $||x - x_1||_2$. In the fifth line we use the assumption $||D^r u_j||_\infty \leq m^{r/d}$ for any derivative operator with $\sum_j r_j \leq s + 1$. To arrive at the sixth line, notice that since x, x_1 are in the same box R_j, we have $||x - x_1||_2 \leq \sqrt{d}m^{-1/d}$ (there are m boxes and each one has length $m^{-1/d}$ on each side). The last line is true since x_1, y_1 are on the line segment between x, y.

In other words, g_λ is holder smooth as long as $Km^{s/d}\sqrt{d} \asymp L$, imposing the requirement that $K = O(m^{-s/d})$. So if we pick $m = n^{\frac{2d}{4s+d}}$ and $K = m^{-s/d} = n^{\frac{-2s}{4s+d}}$ we get that $g_\lambda \in \Sigma(s, L)$ as long as there is some wiggle room around p. We also get that the Hellinger distance is bounded by $O(n^2 n^{\frac{-8s}{4s+d}} n^{\frac{-2d}{4s+d}}) = O(1)$ and the distance in our metric is $n^{\frac{-4s}{4s+d}}$ as we desired. We can apply Theorem 8 and arrive at the result.

C.3 Existence of u_j

To wrap up, we need to show that we can in fact find the functions u_j. We can do this by mapping R_j to $[0, 1]^d$ and using an orthonormal system $\{\phi_j\}_{j=1}^q$ for $L^2([0, 1]^d)$ with $q \geq 3$. Suppose that ϕ_j satisfy (i)

$\phi_1 = 1$, $\phi_j(x) = 0$ for $x \notin [\epsilon, 1-\epsilon]^d$ and (iii) $||D^r \phi_j||_\infty \leq K < \infty$ for all j. Certainly we can find such an orthonormal system.

Now for any function $f \in L^2([0,1]^d)$, we can easily find a unit-normed function $\tilde{v} \in \text{span}(\{\phi_j\})$ such that $\tilde{v} \perp \phi_1$, and $\tilde{v} \perp f$. If we write $\tilde{v} = \sum_i c_i \phi_i$ we have that $D^r v = c_i D^r \phi_i$ so that $||D^r v||_\infty \leq K \sum_i |c_i| \leq K\sqrt{q}$ since $\tilde{v}$ is unit-normed. Notice that the vector $v = \tilde{v}(K\sqrt{q})^{-1}$ has upper and lower-bounded ℓ_2^2-norm while having all $||D^r v||_\infty \leq 1$.

To construct the functions u_j, map the $R_j = \Pi_{i=1}^d [j_i m^{-1/d}, (j_i+1)m^{-1/d}]$ to $[0,1]^d$ and let the function $f = p^{\alpha-1}(x)q^\beta(x)$ mapped appropriately to $[0,1]^d$. Use the function v_j constructed in the previous paragraph. In mapping back to R_j, let $u_j(x) = v_j(m^{1/d}(x - (j_1, \ldots, j_d))^T)$ so that $\int_{R_j} u_j^2(x)dx = m^{-1}\int v_j^2(x)dx = \Omega(1/m)$ and $||D^r u_j||_\infty \leq m^{r/d}$. These functions u_j meet the requirements 1-5 outlined above, allowing us to apply Le Cam's method.

C.4 An $n^{-1/2}$ Lower Bound when $s > d/4$

To obtain the $n^{-1/2}$ lower bound for the highly-smooth setting, we will reduce the problem of estimating $T(p,q)$ to that of estimating a quadratic functional of the two densities:

$$\theta(p,q) = \int a_1(x)p(x) + a_2(x)q(x) + a_3(x)p(x)q(x) + a_4(x)p^2(x) + a_5(x)q(x)d\mu(x) \tag{16}$$

for some known functions $a_i : [0,1]^d \to \mathbb{R}$, $i \in \{1, \ldots, 5\}$. We will then use the following lower bound on the rate of estimating these functionals to establish a lower bound in our problem:

Theorem 13. *Let $a_i : [0,1]^d \to \mathbb{R}, i \in \{1, \ldots, 5\}$ be continuous, bounded, non-constant functions and let $\theta(p,q)$ be as in Equation 16. Then:*

$$\liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{p,q \in \Sigma(s,L)} \mathbb{P}_{X_1^n \sim p, Y_1^n \sim q}[|\hat{\theta}_n - \theta(p,q)| \geq \epsilon n^{-1/2}] \geq c > 0 \tag{17}$$

For some constants $\epsilon, c > 0$.

Proof. We will use Le Cam's Method to establish the lower bound. Let us fix q once and for all. We will only vary p. Let $p_0(x) = 1$ and $p_1(x) = 1 + u(x)$ for some function $u(x)$ that we will select later. By Theorem 2.2 of [23] (essentially the Neyman-Pearson Lemma) if we can upper bound $KL(p_1^n \times q^n, p_0^n \times q^n)$ we have a lower bound on the probability of making an error in the simple versus simple hypothesis test between the two possible distributions when $X_1^n, \sim p_1$ and $Y_1^n \sim q$. Mathematically, define $p_{e,1}(\psi) = \mathbb{P}_{X_1^n \sim p_1, Y_1^n \sim q}[\psi(X_1^n, Y_1^n) \neq 1]$ for a test statistic ψ taking values in $\{0,1\}$. Also define $p_{e,1} = \inf_\psi p_{e,1}(\psi)$. Then Theorem 2.2 of [23] says that if $KL(p_1^n \times q^n, p_0^n \times q^n) \leq \alpha < \infty$ then

$$p_{e,1} \geq \max\left(\frac{1}{4}\exp(-\alpha), \frac{1-\sqrt{\alpha/2}}{2}\right)$$

So let us bound the KL-divergence:

$$KL(p_1^n \times q^n, p_0^n \times q^n) = nKL(p_1, p_0) = n\int (1+u(x))\log(1+u(x))dx \leq n\int u(x) + u^2(x)dx = n||u||_2^2$$

Here we used that $\int u(x) = 0$ if p_1 is to remain a density. This is one of the requirements on the function u that we will pick. If the KL-divergence is to remain bounded, we will also require that $||u||_2^2 \leq c/n$ for some constant.

If we make a mistake in the testing problem, we suffer at least $1/2|\theta(p_0,q) - \theta(p_1,q)|$ loss in the estimation problem. So we must lower bound the absolute difference between the two functional values.

$$\begin{aligned} |\theta(p_0,q) - \theta(p_1,q)| &= |\int a_1(x)u(x) + a_3(x)q(x)u(x) + 2a_4(x)u(x) + a_4(x)u^2(x)d\mu(x)| \\ &= |\int f(x)u(x) + a_4(x)u^2(x)d\mu(x)| \end{aligned}$$

where $f(x) = a_1(x) + a_3(x)q(x) + 2a_4(x)$. Suppose we had a function v such that:

$$\int v(x) = 0,\ ||v(x)||_2^2 = O(1),\ p_1 = 1 + 1/\sqrt{n}v(x) \in \Sigma(s,L),\ \int f(x)v(x) = \Omega(1)$$

Then if we use $u(x) = n^{-1/2}v(x)$ the loss we suffer is at least $c_1/\sqrt{n} - c_2/n \geq \epsilon n^{-1/2}$ for some $\epsilon > 0$ for n sufficiently large. At the same time, the KL-divergence between the two hypothesis is also $O(1)$. So we would be able to apply Le Cam's inequality.

So, we just need to find a sufficiently smooth function v with constant ℓ_2^2 norm and constant inner product with f. To do this, consider an orthonormal system $\phi_1, \ldots, \phi_q$ with $q \geq 3$ of $L^2([0,1]^d)$ such that (i) $\phi_j(x) = 1$, (ii) $f \in \text{span}(\{\phi_j\}_{j=1}^q)$ and (iii) $||D^r\phi_j||_\infty \leq K < \infty$ for all j and all tuples r with $\sum_j r_j \leq s+1$. It is always possible to construct such a system as long as f itself has bounded r-th derivatives, which is true since f itself is a continuous, bounded function over a compact domain. Let L denote the linear space spanned by $\{\phi_j\}$. Earlier we showed that if $v \in L$, then $v \in \Sigma(s,A)$ for sufficiently large constant A. So we can let v be any unit-normed function in $L' = \{v \in L | \langle v, f\rangle = c, \langle v, \phi_1\rangle = 0\}$, which is an affine space of dimension at least 1 (since $f \neq c\phi_1$).

Then $u(x) = v(x)/\sqrt{n}$ meets all of the requirements. Notice that since $v \in \Sigma(s,A)$, we have that $u \in \Sigma(s, A/\sqrt{n}) \subset \Sigma(s,L)$ for n sufficiently large. □

In what follows, the functional θ that we are trying to estimate will actually be a random quantity. However, since Theorem 13 applies to any set of five bounded continuous function $a_1, \ldots, a_5$, it actually applies to any distribution over this space of five bounded continuous functions. Mathematically, for any distribution $\mathcal{D}$ over this space of bounded continuous functions:

$$\liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{p,q\in\Sigma(s,L)} \mathbb{P}_{X_1^n\sim p, Y_1^n\sim q, (a_1,\ldots,a_5)\sim\mathcal{D}} \left[|\hat{\theta}_n(a_1^5) - \theta(a_1^5,p,q)| \geq \epsilon n^{-1/2}\right] \geq c > 0$$

Where $\theta(a_1^5, p, q)$ is given in Equation 16.

Let us use Theorem 13 to prove a lower bound for estimating $T(p,q) = \int p^\alpha q^\beta$. Suppose we had an estimator $\widehat{T}_n$ for $T(p,q)$ that converges at rate $o(n^{-1/2})$, say $\forall p, q, n, \mathbb{E}[|\widehat{T}_n - T(p,q)|] \leq c_1 n^{-1/2-\epsilon}$ for some constants $c_1, \epsilon > 0$. We will use it to construct an estimator for a quadratic functional of p, q with better-than-$\sqrt{n}$ rate, which will contradict Theorem 13.

The quadratic functional of p, q will be the terms in the second order expansion of $T(p,q)$ about $T(\hat{p}_n, \hat{q}_n)$.

Given $2n$ samples, as in our upper bound, we use the first n to construct estimators $\hat{p}_n, \hat{q}_n$ for p, q respectively. We use the second n samples to compute $\hat{T}_n$. The estimator for θ will be $\hat{\theta}_{2n} = \widehat{T}_n - C_2 T(\hat{p}_n, \hat{q}_n)$. Where we are collecting all of the terms of the form $T(\hat{p}_n, \hat{q}_n)$ together. Recall that C_2 is the coefficient for all of these terms.

The risk of the estimator is:

$$\mathbb{E}_{X_1^{2n}}[|\hat{\theta}_n - \theta|] \leq \mathbb{E}_{X_{n+1}^{2n}}[|\widehat{T}_n - T|] + \mathbb{E}_{X_1^{2n}}[|T - C_2T(\hat{p},\hat{q}) - \theta|]$$

$$\begin{aligned} &\leq c_1 n^{-1/2-\epsilon} + O(\mathbb{E}_{X_1^n}[||p-\hat{p}||_3^3 + ||q-\hat{q}||_3^3]) \\ &\leq c_1 n^{-1/2-\epsilon} + c_2 n^{\frac{-3s}{2s+d}} \end{aligned}$$

for constants $c_1, c_2 > 0$. Now if $s > d/4$, both terms are $o(n^{-1/2})$, so we have $\mathbb{E}[|\hat{\theta}_n - \theta|] = o(n^{-1/2})$. The functions $\hat{p}_n, \hat{q}_n$ are deterministic functions of X_1^n, Y_1^n, so we can think of X_1^n as encoding a distribution over functions $\hat{p}_n, \hat{q}_n$.

More formally, let $\mathcal{D}$ encode the following distribution: We drawn X_1^n, Y_1^n from p, q respectively and compute $\hat{p}_n, \hat{q}_n$. With these, the five functions $a_1, \ldots, a_5$ are:

$$\begin{aligned} a_1 &= \alpha(2-\alpha-\beta)\hat{p}_n^{\alpha-1}\hat{q}_n^{\beta} \\ a_2 &= \beta(2-\alpha-\beta)\hat{p}_n^{\alpha}\hat{q}_n^{\beta-1} \\ a_3 &= \alpha\beta\hat{p}_n^{\alpha-1}\hat{q}_n^{\beta-1} \\ a_4 &= 1/2\alpha(\alpha-1)\hat{p}_n^{\alpha-2}\hat{q}_n^{\beta} \\ a_5 &= 1/2\beta(\beta-1)\hat{p}_n^{\alpha}\hat{q}_n^{\beta-2} \end{aligned}$$

Notice that all of these functions are continuous and they can be bounded from above and below if we use the truncated kernel density estimators. Now whenever $s > d/4$:

$$\mathbb{E}_{(a_1,\ldots,a_5)\sim\mathcal{D}}\mathbb{E}_{X_1^n\sim p, Y_1^n\sim q}\left[|\hat{\theta}-\theta|\right] = \mathbb{E}_{X_1^{2n}\sim p, Y_1^{2n}\sim q}\left[|\hat{\theta}-\theta|\right] \leq cn^{-1/2-\epsilon}$$

which contradicts the lower bound. Via Markov's inequality, $\mathbb{P}_{X_1^{2n}}[|\hat{\theta}_n - \theta| \geq c_4 n^{-1/2}] \leq o(n^{-1/2})/n^{-1/2} \to 0$ which contradicts our discussion following Theorem 13. This shows that when $s > d/4$, one cannot estimate $T(p,q)$ are faster than $\sqrt{n}$ rate.

C.5 Translating to T_α and D_α

Suppose we have an estimator $\hat{S}_\alpha$ for the Tsallis-α divergence, such that for all $p, q \in \Sigma(s,l)\mathbb{E}[|\hat{S}_\alpha - S_\alpha|] \leq \epsilon_n$. We can define an estimator $\hat{T}$ for $T(p,q) = \int p^\alpha q^{1-\alpha}$ as $\hat{T} = (\alpha-1)\hat{S}_\alpha + 1$. The error between $\hat{T}$ and T is:

$$\mathbb{E}[|\hat{T} - T|] = |\alpha - 1|\mathbb{E}[|\hat{S}_\alpha - S_\alpha|] \leq |\alpha-1|\epsilon_n$$

We therefore know that $\epsilon_n = \Omega(n^{-\gamma})$ where $\gamma = \min\{\frac{4s}{4s+d}, 1/2\}$ since otherwise we would have an estimator $\hat{T}$ for $T(p,q)$ with rate $o(n^{-\gamma})$, which contradicts Theorem 5.

For D_α, we use the same proof structure, but computing the error for $\hat{T}$ is more involved. The estimator $\hat{T} = \exp\{(\alpha-1)\hat{D}\}$ has error:

$$\mathbb{E}[|\hat{T} - T|] = \mathbb{E}\left[|\exp\{(\alpha-1)\hat{D}\} - \exp\{(\alpha-1)D_\alpha\}|\right]$$

We would like to eliminate the absolute value, so we will have to consider all of the cases. If $\alpha < 1$ and $D > \hat{D}$ then the first term dominates the second so we can simply drop the absolute value sign. In this case we can use convexity of e^x to upper bound by:

$$\leq (\alpha-1)\mathbb{E}[e^{(\alpha-1)\hat{D}}(\hat{D} - D_\alpha)] = (1-\alpha)\mathbb{E}[e^{(\alpha-1)\hat{D}}(D_\alpha - \hat{D})] \leq C\epsilon_n$$

as long as D_α is bounded from below, which implies that for n large enough, $e^{(\alpha-1)\hat{D}} = O(1)$. Actually the other cases are analogous, for example if $\hat{D} > D$, then to remove the absolute value, we must swap the two terms, after which we can use convexity to arrive at the same upper bound. Thus we have shown that $\mathbb{E}[|\hat{T} - T|] = O(\epsilon_n)$ which implies that $\mathbb{E}[|\hat{D} - D|] = \Omega(n^{-\gamma})$ as claimed.

D More Auxiliary Results

Lemma 14 (Holder is contained in Sobolev)**.** *Let $f \in \Sigma(s, L)$ belong to the periodic holder class with smoothness s. Then f belongs to the sobolev ellipsoid $\mathcal{W}(s', L')$ where $\phi_k(x) = e^{2i\pi k^T x}$ is the fourier basis, $k \in \mathbb{Z}^d$, $s' < s$ and:*

$$L' = \frac{dCL^2}{(2\pi)^{2\lfloor s \rfloor}}$$

with $C = \sum_{l=0}^{\infty} 4^{l(s'-s)}$.

Proof. Let us decompose $s = r + \alpha$ where $r = \lfloor s \rfloor$ and $\alpha \in (0, 1]$. We need to bound:

$$\sum_{(k_1,\ldots,k_d)\in\mathbb{Z}^d} (\sum_{j=1}^{d} |k_j|^{2s'})|\alpha_k|^2$$

where $\alpha_k = \int f(x)\phi_k(x)dx$. This is equivalent to bounding, for each $j = [d]$, $\sum_{k\in\mathbb{Z}^d} |k_j|^{2s'}|\alpha_k|^2$ so let us fix a dimension j for now. Using repeated integration by parts and the fact that $D^{\vec{r}} f$ is period for all $\vec{r}$ with $\sum_j r_j \leq r$. we get

$$\left| \int \frac{\partial^r}{\partial x_j^r} f(x)\phi_k(x)dx \right| = |2\pi i k_j|^r | \int f(x)\phi_k(x)dx| = |2\pi i k_j|^r |\alpha_k|$$

Let us write $g(x) = \frac{\partial^r}{\partial x_j^r} f(x)$. Then since $f \in \Sigma(s, L)$, we know that g satisfies:

$$|g(x) - g(y)| \leq L||x - h||^{\alpha}$$

for all x, y. We will use this fact to bound $\sum_{k\in\mathbb{Z}^d} |k_j|^{2\alpha'}|b_k|^2$ where $b_k = \int g(x)\phi_k(x)$ and $\alpha' < \alpha$ which will give us a bound on $\sum_{k\in\mathbb{Z}^d} |k_j|^{2s'}|\alpha_k|$ via the above calculation. In particular, suppose that $\sum_{k\in ZZ^d} |k_j|^{2\alpha'}|b_k|^2 \leq \gamma_j$, then:

$$\sum_{k\in\mathbb{Z}^d} |k_j|^{2s'}|\alpha_k|^2 = \sum_{k\in\mathbb{Z}^d} |k_j|^{2r+2\alpha'}|\alpha_k^2| = |2\pi i|^{-2r} \sum_{k\in\mathbb{Z}^d} |k_j|^{2\alpha'}|b_k|^2 \leq (2\pi)^{-2r}\gamma_j$$

Notice that:

$$g(x_1, \ldots, x_j - h, \ldots, x_d) - g(x_1, \ldots, x_j + h, \ldots, x_d) = \sum_{k\in\mathbb{Z}^d} b_k e^{2i\pi k^T x} 2i \sin(2\pi k_j h)$$

This means that:

$$4 \sum_{k\in\mathbb{Z}^d} |b_k|^2 \sin^2(2\pi k_j h) = \int (g(x_1, \ldots, x_j - h, \ldots, x_d) - g(x_1, \ldots, x_j + h, \ldots, x_d))^2 dx \leq L^2|h|^{2\alpha}$$

Notice that $\sin^2(\pi/2) > \sin^2(\pi/4) \geq 1/2$ so if we pick $h = 1/(8q)$ and $k_j \in \{q, \ldots, 2q-1\} \cup \{-q, \ldots, -2q+1\}$ we can lower bound the left hand side. To be concrete, letting $S_q = \{k \in \mathbb{Z}^d | k_j \in \{q, \ldots, 2q-1\} \cup \{-q, \ldots, -2q+1\}\}$:

$$\sum_{k\in\mathbb{Z}^d} |b_k|^2|k_j|^{2\alpha'} = \sum_{l=0}^{\infty} \sum_{k\in S_{2^l}} |b_k|^2|k_j|^{2\alpha'} \leq \sum_{l=0}^{\infty} (2^{l+1})^{2\alpha'} \sum_{k\in S_{2^l}} |b_k|^2$$

But:

$$\sum_{k \in S_{2^l}} |b_k|^2 \leq 2 \sum_{k \in S_{2^l}} |b_k|^2 \sin^2(2\pi k_j (1/2^{l+3})) \leq 2 \sum_{k \in \mathbb{Z}^d} |b_k|^2 \sin^2(2\pi k_j (1/2^{l+3})) \leq \frac{L^2}{2} 2^{-2\alpha(l+3)}$$

Using this bound above, we get:

$$\sum_{k \in \mathbb{Z}^d} |b_k|^2 |k_j|^{2\alpha'} \leq \frac{L^2}{2} \frac{4^{2\alpha'}}{8^{2\alpha}} \sum_{l=0}^{\infty} 4^{l(\alpha'-\alpha)} \leq CL^2$$

whenever the series converges (as long as $\alpha' < \alpha$).

Using this as our value for γ_j and summing over the d dimensions, we get:

$$\sum_{j=1}^{d} \sum_{k \in \mathbb{Z}^d} |k_j|^{2s'} |\alpha_k| \leq d(2\pi)^{-2r} \gamma_j \leq \frac{dCL^2}{(2\pi)^{2r}}$$

□