On Estimating L_2^2 Divergence

Akshay Krishnamurthy
Computer Science Department
Carnegie Mellon University
akshaykr@cs.cmu.edu

Kirthevasan Kandasamy
Machine Learning Department
Carnegie Mellon University
kandasamy@cs.cmu.edu

Barnabás Poczós
Machine Learning Department
Carnegie Mellon University
bapoczos@cs.cmu.edu

Larry Wasserman
Statistics Department
Carnegie Mellon University
larry@stat.cmu.edu

Abstract

We give a comprehensive theoretical characterization of a nonparametric estimator for the L_2^2 divergence between two continuous distributions. We first bound the rate of convergence of our estimator, showing that it is $\sqrt{n}$-consistent provided the densities are sufficiently smooth. In this smooth regime, we then show that our estimator is asymptotically normal, construct asymptotic confidence intervals, and establish a Berry-Esséen style inequality characterizing the rate of convergence to normality. We also show that this estimator is minimax optimal.

1 INTRODUCTION

One of the most natural ways to quantify the dissimilarity between two continuous distributions is with the L_2-distance between their densities. This distance – which we typically call a divergence – allows us to translate intuition from Euclidean geometry and consequently makes the L_2-divergence particularly interpretable. Despite this appeal, we know of very few methods for estimating the L_2-divergence from data. For the estimators that do exist, we have only a limited understanding of their properties, which limits their applicability. This paper addresses this lack of understanding with a comprehensive theoretical study of an estimator for the L_2^2-divergence.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

Our analysis of the L_2^2-divergence is motivated by both practical and theoretical considerations. On the practical side, this divergence is used for discrete distributions in a variety of applications in information retrieval [14], ecology [11], and elsewhere. It therefore seems natural to consider the continuous analog, which can be used in neuroscience and astronomy applications. The L_2^2-divergence has also been used in two sample testing [1] and tasks involving machine learning on distributions [19, 20], where in particular, it has been shown to outperform several other divergence measures in anomalous image detection tasks [19]. On the theoretical front, while it is not clear *a priori* which divergence is best for a particular problem, the L_2^2 divergence contrasts with other popular divergences such as the Kullback-Leibler and the Renyi-α divergences in that it is *symmetric*, and this property is desirable in many applications.

Our estimator is the same kernel multi-sample U-statistic that has appeared numerous times in the literature [1, 5], but has lacked a complete theoretical development. Under a standard smoothness assumption, parameterized by β (formalized in the sequel), and given n samples from two densities supported over $\mathbb{R}^d$, we establish the following properties.

1. We analyze the rate of convergence in squared error, showing an $n^{\frac{-8\beta}{4\beta+d}}$ rate if $\beta < d/4$ and the parametric n^{-1} rate if $\beta \geq d/4$ (Theorem 3).

2. When $\beta > d/4$, we prove that the estimator is asymptotically normal (Theorem 4).

3. We derive a principled method for constructing a confidence interval that we justify with asymptotic arguments (Theorem 5).

4. We also prove a Berry-Esséen style inequality in the $\beta > d/4$ regime, characterizing the distance of the appropriately normalized estimator to the $\mathcal{N}(0,1)$ limit (Theorem 6).

5. Lastly, we modify an existing proof to establish a matching lower bound on the rate of convergence (Theorem 7). This shows that our estimator achieves the minimax rate.

We are not aware of such a characterization of an estimator for this divergence. Indeed, we are not aware of such a precise characterization for *any* nonparametric divergence estimators.

The most novel technical ingredient of our work is the proof of Theorem 6, where we upper bound the distance to the $\mathcal{N}(0,1)$ limit of our estimator. The challenges in this upper bound involve carefully controlling the bias in both our estimator and our estimator for its asymptotic variance so that we can appeal to classical Berry-Esséen bounds. This technical obstacle arises in many nonparametric settings, but we are not aware of any related results.

The remainder of this paper is organized as follows. After mentioning some related ideas in Section 2, we specify the estimator of interest in Section 3. In Section 4, we present the main theoretical results, deferring proofs to Section 5 and the appendix. We present some simulations confirming our results in Section 6 and conclude in Section 7 with some future directions.

2 RELATED WORK

A few other works have considered estimation of the L_2-divergence under non-parametric assumptions [1, 9, 18]. Anderson et al. propose essentially the same estimator that we analyze in this paper [1].When used for two-sample testing, they argue that one should not shrink the bandwidth with n, as it does not lend additional power to the test, while only increasing the variance. Unfortunately, this choice of bandwidth does not produce a consistent estimator. When used for estimation, they remark that one should use a bandwidth that is smaller than for density estimation, but do not pursue this idea further. By formalizing this undersmoothing argument, we achieve the parametric n^{-1} squared error rate.

Poczos et al. establish consistency of a nearest neighbor based L_2 divergence estimator, but do not address the rate of convergence or other properties [18]. Krishnamurthy et al. propose an estimator based on a truncated Fourier expansion of the densities [9]. They establish a rate of convergence that we match, but do not develop any additional properties. Similarly, Källberg and Seleznjev propose an estimator based on nearest neighbors and prove similar asymptotic results to ours, but they do not establish Berry-Esseén or minimax lower bounds [7]. In contrast to these works, our estimator and our analysis are considerably simpler, which facilitates both applicability and theoretical development.

As will become clear in the sequel, our estimator is closely related to the maximum mean discrepancy (MMD) for which we have a fairly deep understanding [6]. While the estimators are strikingly similar, they are motivated from vastly different lines of reasoning and the analysis reflects this difference. The most notable difference is that with MMD, the population quantity is *defined* by the kernel and bandwidth. That is, the choice of kernel influences not only the estimator but also the population quantity. We believe that our estimand is more interpretable as it is independent of the practioner's choices. Nevertheless, some of our results, notably the Berry-Esséen bound, can be ported to an estimate of the MMD.

There is a growing body of literature on estimation of various divergences under nonparametric assumptions. This line of work has primarily focused on Kullback-Leibler, Renyi-α, and Csiszar f-divergences [12, 16, 17]. As just one example, Nguyen et al. develop a convex program to estimate f-divergences under the assumption that the density ratio belongs to a reproducing kernel Hilbert space. Unfortunately, we have very little understanding as to which divergence is best suited to a particular problem, so it is important to have an array of estimators at our disposal.

Moreover, apart from a few examples, we do not have a complete understanding of the majority of these estimators. In particular, except for the MMD [6], we are unaware of principled methods for building confidence intervals for any of these divergences, and this renders the theoretical results somewhat irrelevant for testing and other inference problems.

Our estimator is based on a line of work studying the estimation of integral functionals of a density in the nonparametric setting [2, 3, 5, 8, 10]. These papers consider estimation of quantities of the form $\theta = \int f(p, p^{(1)}, \ldots, p^{(k)}) d\mu$, where f is some known functional and $p^{(i)}$ is the ith derivative of the density p, given a sample from p. Giné and Nickl specifically study estimation of $\int p(x)^2 d\mu$ and our work generalizes their results to the L_2^2-divergence functional [5].

Turning to lower bounds, while we are not aware of a lower bound for L_2^2-divergence estimation under nonparametric assumptions, there are many closely related results. For example, Birge and Massart [3] establish lower bounds on estimating integral functionals

of a single density, while Krishnamurthy et al. extend their proof to a class of divergences [9]. Our lower bound is based on some modifications to the proof of Krishnamurthy et al.

3 THE ESTIMATOR

Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions supported over $\mathbb{R}^d$ with Radon-Nikodym derivatives (densities) $p \triangleq d\mathbb{P}/d\mu, q \triangleq d\mathbb{Q}/d\mu$ with respect to a measure μ. The L_2^2 divergence between these two distributions, denoted throughout this paper as $D(p,q)$ is defined as:

$$D(p,q) \triangleq \int (p(x)-q(x))^2 d\mu(x)$$
$$= \underbrace{\int p^2(x)d\mu}_{\theta_p} + \underbrace{\int q^2(x)d\mu}_{\theta_q} - 2\underbrace{\int p(x)q(x)d\mu}_{\theta_{p,q}}.$$

Estimation of the first two terms in the decomposition has been extensively studied in the nonparametric statistics community [2, 3, 5, 10]. For these terms, we use the kernel-based U-statistic of Gine and Nickl [5]. For the bilinear term, $\theta_{p,q}$, we use a natural adaptation of their U-statistic to the multi-sample setting. Specifically, given samples $\{X_i\}_{i=1}^{2n} \sim p, \{Y_i\}_{i=1}^{2n} \sim q$, we estimate θ_p with $\hat{\theta}_p$ and $\theta_{p,q}$ with $\hat{\theta}_{p,q}$, given by:

$$\hat{\theta}_p = \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} \frac{1}{h^d} K\left(\frac{X_i - X_j}{h}\right) \quad (1)$$

$$\hat{\theta}_{p,q} = \frac{1}{n^2} \sum_{i,j=n+1}^{2n} \frac{1}{h^d} K\left(\frac{X_i - Y_j}{h}\right), \quad (2)$$

where $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a kernel function and $h \in \mathbb{R}_{\geq 0}$ is a bandwidth parameter. In Assumption 2 below, we prescribe some standard restrictions on the kernel and a scaling of the bandwidth.

The squared term involving q, θ_q, is estimated analogously to θ_p, and we denote the estimator $\hat{\theta}_q$. The final L_2^2-divergence estimator is simply $\hat{D}(p,q) = \hat{\theta}_p + \hat{\theta}_q - 2\hat{\theta}_{p,q}$. Notice that we have split the data so that each point X_i (respectively Y_j) is used in exactly one term. Forcing this independence will simplify our theoretical analysis without compromising the properties.

Our estimator involves data splitting, as we are using half of the sample for each of the terms in the estimand. Data splitting is a very common technique in these types of nonparametric problems (See for example [3, 15]). As we will show, data-splitting only affects the convergence rate in constant factors, but it plays a much larger role in our other results. In particular, asymptotic normality, the confidence interval and the Berry-Esséen bound *do not* hold for the estimator without data-splitting [6], so it is necessary to split the sample for most inference problems. We defer a more detailed discussion of this fact to after the statement of Theorem 4. However, applications such as machine learning on distributions [20] that do not leverage these theoretical properties may not require splitting the sample.

We also remark that the estimator can naively be computed in quadratic time. However, with a compact kernel, a number of data structures are available that lead to more efficient implementations. In particular, the dual tree algorithm of Ram et al. can be used to compute $\hat{D}$ in linear time [21].

4 THEORETICAL PROPERTIES

In this section, we highlight some of the theoretical properties enjoyed by the divergence estimator $\hat{D}$. We begin by stating the main assumptions, regarding the smoothness of the densities, properties of the kernel, and the choice of bandwidth h.

Definition 1. *We call $\mathcal{W}_1^\beta(C)$, for $\beta \in \mathbb{N}$ and $C > 0$, the **Bounded Variation class** of order β which is the set of β-times differentiable funtions whose βth derivatives have bounded L_1 norm. Formally, a function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to $\mathcal{W}_1^\beta(C)$ if for all tuples of natural numbers $r = (r_1, \ldots, r_d)$ with $\sum_j r_j \leq \beta$ we have $\|D^r f\|_1 \leq C$, where $D^r = \frac{\partial^{r_1+\ldots+r_d}}{\partial x_1^{r_1} \ldots \partial x_d^{r_d}}$ is a derivative operator.*

Assumption 2. *Assume p, q, K, and h satisfy:*

1. ***Smoothness**: The densities p, q belong to the bounded variation class $\mathcal{W}_1^\beta(C)$.*
2. ***Kernel Properties**: K is bounded, symmetric, supported on $(-1,1)^d$, and has $\int K(u)d\mu(u) = 1$. $\int \prod_i x_i^{r_i} K(x)dx = 0$ for all $(r_1, \ldots, r_d)$ with $\sum_j r_j \leq 2\beta$.*
3. ***Kernel Bandwidth**: We choose $h \asymp n^{\frac{-2}{4\beta+d}}$.*

The smoothness assumption is similar in spirit to both the Hölder and Sobolev assumptions which are more standard in the nonparametric literature. Specifically, the bounded variation assumption is the integrated analog of the Hölder assumption, which is a pointwise characterization of the function. It is also the L_1 analog of the Sobolev assumption, which requires that $||D^r f||_2^2$ is bounded.

One difference is that the class $\mathcal{W}_1^\beta$ can not be defined for non-integral smoothness, β, while both the Hölder and Sobolev classes can. While our results can be shown for the Sobolev class, working with bounded

variation class considerably simplifies the proofs as we avoid the need for any Fourier analysis. The Hölderian assumption is insufficient as Hölder smoothness is not additive under convolution, which is critical for establishing the low order bias of our estimator.

The kernel properties are now fairly standard in the literature. Notice that we require the kernel to be of order 2β, instead of order β as is required in density estimation. This will allow us to exploit additional smoothness provided by the convolution implicit in our estimators. Of course one can construct such kernels for any β using the Legendre polynomials [22]. We remark that the scaling of the kernel bandwidth is not the usual scaling used in density estimation.

We now turn to characterizing the rate of convergence of the estimator $\hat{D}$. While we build off of the analysis of Giné and Nickl, who analyze the estimator $\hat{\theta}_p$ [5], our proof has two main differences. First, since we work with a different smoothness assumption, we use a different technique to control the bias. Second, we generalize to the bilinear term $\hat{\theta}_{p,q}$, which involves some modifications. We have the following theorem:

Theorem 3. *Under Assumption 2 we have:*

$$\mathbb{E}[(\hat{D}(p,q) - D(p,q))^2] \leq \begin{cases} c_3 n^{\frac{-8\beta}{4\beta+d}} & \text{if } \beta < d/4 \\ c_4 n^{-1} & \text{if } \beta \geq d/4 \end{cases} \quad (3)$$

This bounds holds both with and without data splitting.

Notice that the rate of convergence is substantially faster than the rate of convergence for estimation of β-smooth densities. In particular, the parametric rate is achievable provided sufficient smoothness[1]. This agrees with the results on estimation of integral functionals in the statistics community [3, 5]. It also matches the rate of the orthogonal series estimator studied by Krishnamurthy et al. [9].

One takeaway from the theorem is that one should *not* use the optimal density estimation bandwidth of $n^{\frac{-1}{2\beta+d}}$ here. As we mentioned, this choice was analyzed by Anderson et al. and results in a slower convergence rate [1]. Indeed our choice of bandwidth $h \asymp n^{\frac{-2}{4\beta+d}}$ is always smaller, so we are undersmoothing the density estimate. This allows us to exploit additional smoothness provided by implicit convolution in our estimator, while the additional variance induced by undersmoothing is mitigated by integration in the estimand.

Interestingly, there seem to be two distinct approaches to estimating integral functionals. On one hand, one could plug in an undersmoothed density estimator directly into the functional. This is the approach we take here and it has also been used for other divergence estimation problems [18]. Another approach is to plug in a minimax optimal density estimator and then apply some post-hoc correction. This latter approach can be shown to achieve similar rates for divergence estimation problems [9].

The advantage of the post-hoc correction approach is that one can use cross validation on the density estimate to select the bandwith h. Cross-validating the density estimate does not work for our estimator, since our bandwidth is not optimal for density estimation. Instead, we advocate setting the bandwidth based on the median pairwise distance between the samples, which is a heuristic used in similar problems [6]. The disadvantage of the post-hoc correction approach is computational; the estimator involves numeric integration, which becomes intractable even in moderate dimension. In contrast, our estimator can be computed in quadratic time.

The next theorem establishes asymptotic normality in the smooth regime:

Theorem 4. *When $\beta > d/4$:*

$$\sqrt{n}\left(\hat{D}(p,q) - D(p,q)\right) \rightsquigarrow \mathcal{N}(0, \sigma^2),$$

where $\rightsquigarrow$ denotes convergence in distribution and:

$$\sigma^2 = \begin{cases} 4 \operatorname*{Var}_{x\sim p}(p(x)) + 4 \operatorname*{Var}_{y\sim q}(q(y)) \\ + 4 \operatorname*{Var}_{x\sim p}(q(x)) + 4 \operatorname*{Var}_{y\sim q}(p(y) \end{cases} \quad (4)$$

Note that this theorem *does not* hold without data splitting. Indeed, Gretton *et al.* [6] show that for the MMD, when $p = q$, the limiting distribution of the U-statistic estimator without data splitting is an infinite weighted sum of terms involving the squared difference of Gaussian random variables. Their argument carries through to our setting, as the only difference between their estimator and ours is that we let the bandwidth h shrink with the number of samples. For this reason, the remainder of our theoretical analysis applies only to the data-split estimator.

With this characterization of the limiting distribution, we can now turn to construction of an asymptotic confidence interval.

The most straightforward approach is to estimate the asymptotic variance and appeal to Slutsky's Theorem. We simply use a plugin estimator for the variance, which amounts to replacing all instances of p, q in Equation 4 with estimates $\hat{p}, \hat{q}$ of the densities. For example, we replace the first term with $\int \hat{p}(x)^3 - (\int \hat{p}(x)^2)^2$. We denote the resulting estimator by $\hat{\sigma}^2$, and mention that one should use a bandwidth $h \asymp n^{\frac{-1}{2\beta+d}}$ for estimating this quantity.

[1] The parametric rate is n^{-1} in squared error which implies an $n^{-1/2}$ rate in absolute error.

In Section 5 (specifically Lemma 8), we bound the rate of convergence of this estimator, and its consistency immediately gives an asymptotic confidence interval:

Theorem 5. *Let $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ be the $1-\alpha/2$th quantile of the standard normal distribution. Then,*

$$\frac{\sqrt{n}(\hat{D}(p,q) - D(p,q))}{\hat{\sigma}} \rightsquigarrow \mathcal{N}(0,1), \tag{5}$$

whenever $\beta > d/4$. Consequently,

$$\mathbb{P}\left(D \in \left[\hat{D} - \frac{z_{\alpha/2}\hat{\sigma}}{\sqrt{n}}, \hat{D} + \frac{z_{\alpha/2}\hat{\sigma}}{\sqrt{n}}\right]\right) \to 1-\alpha \tag{6}$$

which means that $[\hat{D} - \frac{z_{\alpha/2}\hat{\sigma}}{\sqrt{n}}, \hat{D} + \frac{z_{\alpha/2}\hat{\sigma}}{\sqrt{n}}]$ is an asymptotic $1-\alpha$ confidence interval for D.

While the theorem does lead to a confidence interval, it is worth asking how quickly the distribution of the self-normalizing estimator converges to a standard normal, so that one has a sense for the quality of the interval in finite sample. We therefore turn to establishing a more precise guarantee. To simplify the presentation, we assume that we have a fresh set of n samples per distribution to compute $\hat{\sigma}^2$. Thus we are given $3n$ samples per distribution in total, and we use $2n$ of them to compute $\hat{D}$ and the last set for $\hat{\sigma}^2$. As before, in computing $\hat{\sigma}^2$, we set $h \asymp n^{\frac{-1}{2\beta+d}}$.

Theorem 6. *Let $\Phi(z)$ denote the CDF of the standard normal. Under Assumption 2, there exists a constant $c_\star > 0$ such that:*

$$\sup_z \left| \mathbb{P}\left(\frac{\sqrt{n}(\hat{D}(p,q) - D(p,q))}{\hat{\sigma}} \le z \right) - \Phi(z) \right| \le \tag{7}$$

$$c_\star \left(n^{\frac{d-4\beta}{8\beta+d}} + n^{\frac{-\beta/2}{2\beta+d}} \right). \tag{8}$$

This bound is $o(1)$ as soon as $\beta > d/4$.

As an immediate consequence of the theorem, we obtain an error bound on the quality of approximation of the confidence interval in Theorem 5. We remark that one can explicitly track all of the constants in the theorem and leave the result in terms of the bandwidth h and problem dependent constants, although this is somewhat tedious. For ease of exposition we have chosen to present the asymptotic version of the theorem, focusing instead on the rate of convergence to the limiting $\mathcal{N}(0,1)$ distribution.

It is not surprising that the rate of convergence to Gaussianity is not the typical $n^{-1/2}$ rate, as it depends on the third moment of the U-statistic, which is decreasing with n. It also depends on the non-negligible bias of the estimator. However, as soon as $\beta > d/4$, it is easily verified that the bound is $o(1)$. This matches our asymptotic guarantee in Theorem 4. Of course, for smoother densities, the rate of convergence in the theorem is polynomially faster.

In addition to the practical consequences, we believe the techniques used in the proof of the theorem are fairly novel. While establishing Berry-Esséen bounds for linear and other parameteric estimators is fairly straightforward [4], this type of result is uncommon in the nonparametric literature. The main challenge is dealing with the bias and additional error introduced by estimating the variance.

Finally, let us address the question of optimality. The following theorem lower bounds the rate of convergence of any estimator for the L_2^2 divergence, when the densities belong to the bounded variation class.

Theorem 7. *With $\gamma_\star = \min\{8\beta/(4\beta+d), 1\}$ and for any $\epsilon > 0$, we have:*

$$\inf_{\hat{D}_n} \sup_{p,q \in \mathcal{W}_1^\beta(C)} \mathbb{P}_{p,q}^n \left[(\hat{D}_n - D)^2 \ge \epsilon n^{-\gamma_\star} \right] \ge c > 0 \tag{9}$$

The result shows that $n^{-\gamma_\star}$ lower bounds the minimax rate of convergence in squared error. Of course $\gamma_\star = 1$ when $\beta \ge d/4$, so the rate of convergence can be no better than the parametric rate. Comparing with Theorem 3, we see that our estimator achieves the minimax rate.

5 PROOFS

The proofs of Theorems 3 and 4 are based on modifications to the analysis of Giné and Nickl [5] so we will only sketch the ideas here. The majority of this section is devoted to proving the Berry-Esséen bound in Theorem 6, proving Theorem 5 along the way. We close the section with a sketch of the proof of Theorem 7.

5.1 Proof Sketch of Theorem 3 and 4

Theorem 3 follows from bounding the bias and the variance of the terms $\hat{\theta}_p, \hat{\theta}_q$, and $\hat{\theta}_{pq}$. The terms are quite similar and we demonstrate the ideas with $\hat{\theta}_{pq}$.

We show that the bias can be written in terms of a convolution and then use the fact that bounded-variation smoothness is additive under convolution. By a substitution, we see that the bias for $\hat{\theta}_{pq}$ is:

$$\begin{aligned}\mathbb{E}[\hat{\theta}_{pq}] - \theta_{pq} &= \int\int K(u)[p(x-uh) - p(x)]q(x)dudx \\ &= \int K(u)[(p_0 \star q)(uh) - (p_0 \star q)(0)]du,\end{aligned}$$

where $p_0(x) = p(-x)$ and $\star$ denotes convolution. Next, we use Young's inequality to show that if two functions f, g belong to $\mathcal{W}_1^\beta(C)$, then $f \star g \in \mathcal{W}_1^{2\beta}(C^2)$. Using

this inequality, we can take a Taylor expansion of order $2\beta - 1$ and use the kernel properties to annihilate all but the remainder term, which is of order $h^{2\beta}$.

To bound the variance, we expand:

$$\mathbb{E}[\hat{\theta}_p^2] = \mathbb{E}\left[\frac{1}{n^2(n-1)^2}\sum_{i\neq j, s\neq t} K_h(X_i, Y_j), K_h(X_s, Y_t)\right]$$

By analyzing each of the different scenarios (i.e. the terms where all indices are different, there is one equality, or there are two equalities), it is not hard to show that the variance is:

$$\mathrm{Var}(\hat{\theta}_p) \leq O\left(\frac{1}{n} + \frac{1}{h^d n^2}\right)$$

Equipped with these bounds, the rate of convergence follows from the bias-variance decomposition and our choice of bandwidth. For the estimator without data splitting, we first used the Cauchy-Schwarz inequality to separate the terms in the expansion of the mean-squared error and then apply the above bounds.

The proof of normality is quite technical and we just briefly comment on the steps, deferring all calculations to the appendix. We apply Hoeffding's decomposition, writing the centered estimator as the sum of a U-process and two empirical processes, one for p and one for q. The U-process converges in quadratic mean to 0 at faster than $1/\sqrt{n}$ rate, so it can be ignored. For the empirical processes, we show that they are close (in quadratic mean) to $\sqrt{n}(P_n q - \theta_{pq})$ and $\sqrt{n}(Q_n p - \theta_{pq})$, where P_n, Q_n are the empirical measures. From here, we apply the Lindberg-Levy central limit theorem to these empirical processes.

5.2 Proof of Theorem 6

The Berry-Esséen theorem can be applied to an unbiased multi-sample U-statistic, normalized by a term involving the conditional variances. Specifically, we will be able to apply the theorem to:

$$\frac{\sqrt{n}(\hat{D} - \mathbb{E}\hat{D})}{\bar{\sigma}}, \tag{10}$$

where:

$$\bar{\sigma}^2 = \begin{cases} 4\operatorname*{Var}_{x\sim p}(\bar{p}(x)) + 4\operatorname*{Var}_{y\sim q}(\bar{q}(y)) \\ + 4\operatorname*{Var}_{x\sim p}(\bar{q}(x)) + 4\operatorname*{Var}_{y\sim q}(\bar{p}(y)) \end{cases}$$

The appropriate normalization is similar to the asymptotic variance σ^2 (Equation 4) except that the densities are replaced with the mean of their kernel density estimates, i.e. $\bar{p}(x) = \int K_h(x, y)p(y)$.

We would like to establish a Berry-Esséen bound for $\sqrt{n}\hat{\sigma}^{-1}(\hat{D} - D)$, but must first make several translations to arrive at Equation 10. We achieve this with several applications of the triangle inequality and some Gaussian anti-concentration properties. We must also analyze the rate of convergence of the variance estimator $\hat{\sigma}^2$ to $\bar{\sigma}^2$ for this bound and to σ^2 for Theorem 5.

Let $F_{\hat{\sigma}}$ be the distribution of $\hat{\sigma}/\bar{\sigma}$, induced by the second half of the sample. Then we may write:

$$\begin{aligned}&\mathbb{P}\left(\frac{\sqrt{n}}{\hat{\sigma}}(\hat{D} - D) \leq z\right) \\ &= \int \mathbb{P}\left(\frac{\sqrt{n}}{\bar{\sigma}}(\hat{D} - D) \leq tz\right) dF_{\hat{\sigma}}(t),\end{aligned}$$

so that we can decompose the proximity to the standard normal CDF as:

$$\begin{aligned}&\sup_z \left|\mathbb{P}\left(\frac{\sqrt{n}}{\hat{\sigma}}(\hat{D} - D) \leq z\right) - \Phi(z)\right| \\ &\leq \sup_z \int \left|\mathbb{P}\left(\frac{\sqrt{n}}{\bar{\sigma}}(\hat{D} - D) \leq tz\right) - \Phi(tz)\right| dF_{\hat{\sigma}}(t) \\ &+ \sup_z \left|\int \Phi(tz) dF_{\hat{\sigma}}(t) - \Phi(z)\right|.\end{aligned}$$

For the first term it is quite easy to eliminate the integral by pushing the supremum inside and replacing tz with the variable being maximized. This leads to:

$$\begin{aligned}&\sup_z \left|\mathbb{P}\left(\frac{\sqrt{n}}{\bar{\sigma}}(\hat{D} - D) \leq z\right) - \Phi(z)\right| \\ &\leq \sup_z \left|\mathbb{P}\left(\frac{\sqrt{n}}{\bar{\sigma}}(\hat{D} - \mathbb{E}\hat{D}) \leq z\right) - \Phi\left(z\right)\right| \\ &+ \sup_z \left|\Phi\left(z - \frac{\sqrt{n}}{\bar{\sigma}}(\mathbb{E}\hat{D} - D)\right) - \Phi(z)\right|,\end{aligned}$$

which follows by adding and subtracting $\mathbb{E}\hat{D}$, adding and subtracting a term involving the Gaussian CDF and the bias and redefining z in the first term. The first term on the right hand side involves the expression in Equation 10 and we will apply Theorem 10.4 from Chen et al. to control it [4]. The second term can be bounded since $\mathbb{E}\hat{D} - D \asymp h^{2\beta}, \sigma = \Theta(1)$ and the Gaussian density is at most $(2\pi)^{-1/2}$. This gives:

$$\sup_z \left|\Phi\left(z - \frac{\sqrt{n}}{\sigma}(\mathbb{E}\hat{D} - D)\right) - \Phi(z)\right| \leq c_b\sqrt{n}h^{2\beta}. \tag{11}$$

Returning to the term involving the variance estimator, we will need the following lemma, which bounds the error in the variance estimate:

Lemma 8. *Under Assumption 2, but with $h \asymp n^{\frac{-1}{2\beta+d}}$, we have that for any $\epsilon > 0$:*

$$\mathbb{P}[|\hat{\sigma}^2 - \sigma^2| > \epsilon] \leq C_1 \epsilon^{-1} n^{\frac{-\beta}{2\beta+d}}, \tag{12}$$

$$\mathbb{P}[|\hat{\sigma}^2 - \bar{\sigma}^2| > \epsilon] \leq C_2 \epsilon^{-1} n^{\frac{-\beta}{2\beta+d}}. \quad (13)$$

The first part of Lemma 8 immediately gives the asymptotic confidence interval in Theorem 5, as we have a consistent estimator of the asymptotic variance. The second part is used in the Berry-Esséen bound.

Notice that since $\bar{\sigma}, \bar{\sigma}^2 = \Theta(1)$ and since $\hat{\sigma}^2 > 0$, we also have that:

$$\mathbb{P}[|\hat{\sigma} - \bar{\sigma}| > \epsilon] \leq C \epsilon^{-1} n^{\frac{-\beta}{2\beta+d}},$$

where the constant has changed slightly. Since $F_{\hat{\sigma}}$ is the CDF for $\hat{\sigma}/\sigma$ and since the difference between two Gaussian CDFs is bounded by two, we therefore have,

$$\int_{-\infty}^{1-\epsilon} \Phi(tz) - \Phi(z) dF_{\hat{\sigma}}(t) + \int_{1+\epsilon}^{\infty} \Phi(tz) - \Phi(z) dF_{\hat{\sigma}}(t) \leq C \epsilon^{-1} n^{\frac{-\beta}{2\beta+d}}.$$

So we only have to consider the situation where $1-\epsilon \leq t \leq 1+\epsilon$. The difference between the Gaussian CDF at z and $(1-\epsilon)z$ is small, since while the width of integration is growing linearly, the height of the integral is decaying exponentially. This term is maximized at ± 1 and it is $O(\epsilon)$, so that the entire term depending on the variance estimate is:

$$\left| \int \Phi(tz) dF_{\hat{\sigma}}(t) - \Phi(z) \right| \leq O\left(\epsilon + n^{\frac{-\beta}{2\beta+d}}/\epsilon \right). \quad (14)$$

Optimizing over ϵ gives a rate of $O(n^{\frac{-\beta/2}{2\beta+d}})$.

The Berry-Esséen inequality applied to the term $\sqrt{n}\sigma^{-1}(\hat{D} - \mathbb{E}\hat{D})$ reveals that:

$$\sup_z \left| \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}(\hat{D} - \mathbb{E}\hat{D}) \leq z \right) - \Phi(z) \right| \leq O\left(n^{-1/2} + \frac{1}{\sqrt{nh^d}} \right), \quad (15)$$

where all of the constants can be tracked explicitly, although they depend on the unknown densities p, q. The application of the theorem from Chen et al. requires bounding various quantities related to the moments of the U-statistic. All of these terms can be bounded using straightforward techniques and we defer these details along with some more careful bookkeeping to the appendix.

Theorem 6 follows from the application of Berry-Esséen in Equation 15, the variance bound in Equation 14, the bias bound in Equation 11 and our choice of bandwidth in Assumption 2.

5.3 Proof of Theorem 7

The proof is a modification of Theorem 2 of [9]. The idea is to reduce the estimation problem to a simple hypothesis test, and then lower bound the probability of error by appealing to the Neyman-Pearson Lemma. If the null and alternative hypotheses, which will consist of pairs of distributions, are well separated, in the sense that the L_2^2 divergence of the null hypothesis is far from the divergence of the alternative, then a lower bound on the probability of error immediately lower bounds the estimation error. This argument is formalized in the following Lemma (from [9]), which is a consequence of Theorem 2.2 of Tsybakov [22].

Lemma 9 ([9]). *Let Λ be an index set and let $p_0, q_0, p_\lambda \forall \lambda \in \Lambda$ be densities (with corresponding distribution functions P_0, Q_0, P_λ) belonging to a function space Θ. Let T be a bivariate functional defined on some subset of $\Theta \times \Theta$ which contains (p_0, q_0) and $(p_\lambda, q_0) \forall \lambda \in \Lambda$. Define $\overline{P^n} = \frac{1}{|\Lambda|}\sum_{\lambda \in \Lambda} P_\lambda^n$. If:*

$$h^2(P_0^n \times Q_0^n, \overline{P^n} \times Q_0^n) \leq \gamma < 2$$
$$T(p_0, q_0) \geq 2\beta + T(p_\lambda, q) \forall \lambda \in \Lambda$$

Then,

$$\inf_{\hat{T}_n} \sup_{p,q \in \Theta} \mathbb{P}_{p,q}^n \left[|\hat{T}_n - T(p,q)| > \beta \right] \geq c_\gamma \quad (16)$$

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

Equipped with the above lemma, we can lower bound the rate of convergence by constructing densities p_λ satisfying the bounded variation assumption, checking that they are well separated in the L_2^2 divergence sense, and bounding the Hellinger distance. We use the same construction as Krishnamurthy et al. and can therefore apply their Hellinger distance bound (which is originally from Birge and Massart [3]).

We defer verifying the bounded variation assumption and the separation in L_2^2 divergence to the appendix as the arguments are a fairly technical and require several new definitions. There, we show that the functions p_λ can be chosen to belong to $\mathcal{W}_1^\beta(C)$, have separation $\beta = n^{-\frac{4\beta}{4\beta+d}}$ (in absolute error), with $\gamma = O(1)$, resulting in the desired lower bound. The n^{-1} term in the lower bound follows from a standard application of Le Cam's method (See Krishnamurthy et al. [9]).

6 EXPERIMENTS

The results of our simulations are in Figure 1. For the first two plots, we trained our estimator on data generated from two Gaussian with means $(0, \ldots, 0) \in \mathbb{R}^d$ and $(1, \ldots, 1) \in \mathbb{R}^d$. Note that the true L_2^2 distance can be analytically computed and is $\frac{2}{(2\sqrt{\pi})^d}\left(1 - e^{-d/4}\right)$. We use a Gaussian kernel with bandwidth $0.5 n^{\frac{-2}{5d}}$ which is the appropriate scaling if $\beta = d$. We use the Gaussian kernel because it is the

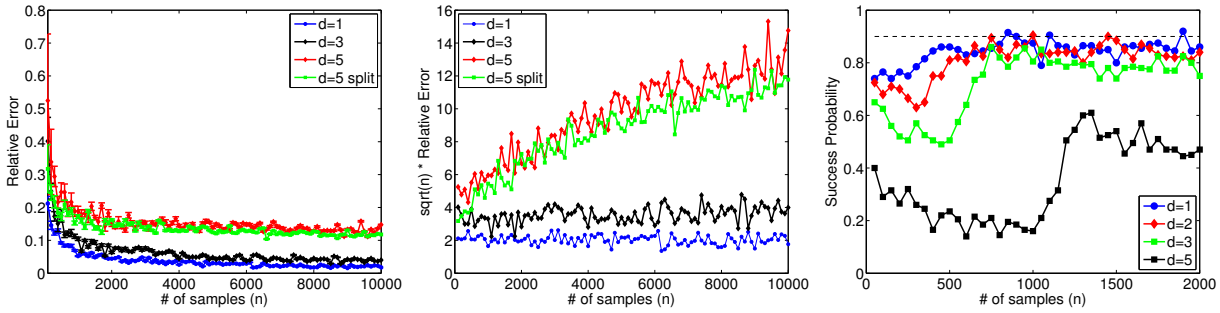

Figure 1: Simulation results showing the convergence rate of the error, rescaled convergence rate, and performance of the confidence interval (from left to right).

standard choice in practice, but notice that it does not meet all of our kernel requirements.

In the first plot, we record the relative error $\frac{|\hat{D}-D|}{D}$ of the estimator as a function of the number of samples for three different problem dimensions. We use relative error in this plot to ensure that the curves are on the same scale, as the L_2-divergence between Gaussians decreases exponentially with dimension. In the second plot, we rescale the relative error by $\sqrt{n}$. As a comparison, we also include the estimator computed without data splitting for $d = 5$ in both of these plots.

The first plot shows that the error is indeed converging to zero and that the relative error increases with dimension. In the second plot, we see that the rescaled error curves all flatten out, confirming the $n^{-1/2}$ convergence rate in the ℓ_1 metric. However, notice that both the asymptote and the sample size at which the curves flatten out is increasing with dimension. The latter suggests that, in high dimension, one needs a large number of samples before the $\sqrt{n}$-rate comes into effect. Comparing the estimators with and without data splitting, we see that data-splitting leads to only a slight degradation in performance.

In the third plot, we explore the empirical properties of our confidence interval. As before, we generate data from two Gaussian distributions, compute the confidence interval and record whether the interval traps the true parameter or not. In the figure, we plot the empirical probability that the 90% confidence interval traps the true parameter as a function of the number of samples. In low dimension, the confidence interval seems to be quite accurate as the empirical probability approaches 90%. However, even in moderate dimension, the confidence interval is less effective, as the sample size is too small for the asymptotic approximation to be accurate. This is confirmed by the previous figure, as the sample size must be quite large for the $\sqrt{n}$-asymptotics to take effect.

7 DISCUSSION

In this paper, we studied a simple estimator for the L_2^2 divergence in the nonparametric setting. We showed that the estimator achieves the parametric $\sqrt{n}$ rate of convergence as soon as the densities have $d/4$-orders of smoothness, which we showed to be optimal. We also established asymptotic normality, derived an asymptotic confidence interval, and characterized the quality of the asymptotic approximation with a Berry-Esséen style inequality. This gives a thorough characterization of the theoretical properties of this estimator.

It is worth exploring how the L_2^2 divergence estimator and other nonparametric functionals can be used algorithmically in learning problems. One challenging problem involves optimizing a nonparametric functional over a finite family of distributions in an active learning setting (for example, finding the closest distribution to a target). Here the so-called Hoeffding racing algorithm, which carefully constructs confidence intervals and focuses samples on promising distributions, has been used in the discrete setting with considerable success [13]. This algorithm relies on finite-sample confidence intervals that are absent from the nonparametrics literature, so extension to continuous distributions would require new theoretical developments.

Regarding two sample testing, an important open question is to identify which test statistic is best for a particular problem. To our knowledge, little progress has been made in this direction.

Acknowledgements

This research is supported by DOE grant DESC0011114, NSF Grants DMS-0806009, and IIS1247658, and Air Force Grant FA95500910373. AK is supported in part by a NSF Graduate Research Fellowship. AK would also like to thank Arthur Gretton and Aaditya Ramdas for feedback on an earlier version of this paper.

References

[1] Niall H. Anderson, Peter Hall, and D. Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 1994.

[2] Peter 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.

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

[4] Louis H.Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Steins Method*. Springer, 2010.

[5] Evarist Giné and Richard Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, February 2008.

[6] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, March 2012.

[7] David Källberg and Oleg Selezniev. Estimation of entropy-type integral functionals. *arXiv:1209.2544*, 2012.

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

[9] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. Nonparametric Estimation of Renyi Divergence and Friends. In *International Conference on Machine Learning*, 2014.

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

[11] Pierre Legendre and Loic FJ Legendre. *Numerical ecology*, volume 20. Elsevier, 2012.

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

[13] Po-Ling Loh and Sebastian Nowozin. Faster hoeffding racing: Bernstein races via jackknife estimates. In *Algorithmic Learning Theory*, 2013.

[14] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[15] Kevin Moon and Alfred Hero. Multivariate f-divergence estimation with confidence. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2014.

[16] 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.

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

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

[19] Barnabás Póczos, Liang Xiong, and Jeff Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Uncertainty and Artificial Intelligence*, 2011.

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

[21] Parikshit Ram, Dongryeol Lee, William B. March, and Alexander G. Gray. Linear-time algorithms for pairwise statistical problems. In *Advances in Neural Information Processing Systems*, 2009.

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