

---

# 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 Poczos**  
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<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. Copyright 2015 by the authors.

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- $\alpha$  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  $\beta$  (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 Seleznev propose an estimator based on nearest neighbors and prove similar asymptotic results to ours, but they do not establish Berry-Esséen 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 practitioner’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- $\alpha$ , 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  $\theta = \int f(p, p^{(1)}, \dots, p^{(k)}) d\mu$ , where  $f$  is some known functional and  $p^{(i)}$  is the  $i$ th 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  $\mu$ . The  $L_2^2$  divergence between these two distributions, denoted throughout this paper as  $D(p, q)$  is defined as:

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

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  $\theta_p$  with  $\hat{\theta}_p$  and  $\theta_{p,q}$  with  $\hat{\theta}_{p,q}$ , given by:

$$\hat{\theta}_p = \frac{1}{n(n-1)} \sum_{i \neq j=1}^{2n} \frac{1}{h^d} K\left(\frac{X_i - X_j}{h}\right) \quad (1)$$

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

where  $K : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$  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  $q$ ,  $\theta_q$ , is estimated analogously to  $\theta_p$ , 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  $\beta$  which is the set of  $\beta$ -times differentiable functions whose  $\beta$ th derivatives have bounded  $L_1$  norm. Formally, a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  belongs to  $\mathcal{W}_1^\beta(C)$  if for all tuples of natural numbers  $r = (r_1, \dots, r_d)$  with  $\sum_j r_j \leq \beta$  we have  $\|D^r f\|_1 \leq C$ , where  $D^r = \frac{\partial^{r_1+\dots+r_d}}{\partial x_1^{r_1} \dots \partial x_d^{r_d}}$  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  $(r_1, \dots, r_d)$  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,  $\beta$ , 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  $2\beta$ , instead of order  $\beta$  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  $\beta$  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  $\beta$ -smooth densities. In particular, the parametric rate is achievable provided sufficient smoothness<sup>1</sup>. 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

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

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 bandwidth  $h$ . 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:*

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

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  $\int \hat{p}(x)^3 - (\int \hat{p}(x))^2$ . 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.

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,*

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

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

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

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:*

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

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

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 parametric 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:*

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

The result shows that  $n^{-\gamma_\star}$  lower bounds the minimax rate of convergence in squared error. Of course  $\gamma_\star = 1$  when  $\beta \geq d/4$ , 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:

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

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:

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

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:

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

where:

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

The appropriate normalization is similar to the asymptotic variance  $\sigma^2$  (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  $\sigma^2$  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:

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

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:

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

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  $\epsilon > 0$ :*

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

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

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,

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

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  $\pm 1$  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  $\epsilon$  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:

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

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 book-keeping 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  $\Lambda$  be an index set and let  $p_0, q_0, p_\lambda \forall \lambda \in \Lambda$  be densities (with corresponding distribution functions  $P_0, Q_0, P_\lambda$ ) belonging to a function space  $\Theta$ . 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  $\bar{P}^n = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} P_\lambda^n$ . If:*

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

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  $p_\lambda$  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  $p_\lambda$  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  $(0, \dots, 0) \in \mathbb{R}^d$  and  $(1, \dots, 1) \in \mathbb{R}^d$ . Note that the true  $L_2^2$  distance can be analytically computed and is  $\frac{2}{(2\sqrt{\pi})^d} (1 - e^{-d/4})$ . We use a Gaussian kernel with bandwidth  $0.5n^{-\frac{2}{5d}}$  which is the appropriate scaling if  $\beta = d$ . We use the Gaussian kernel because it is the

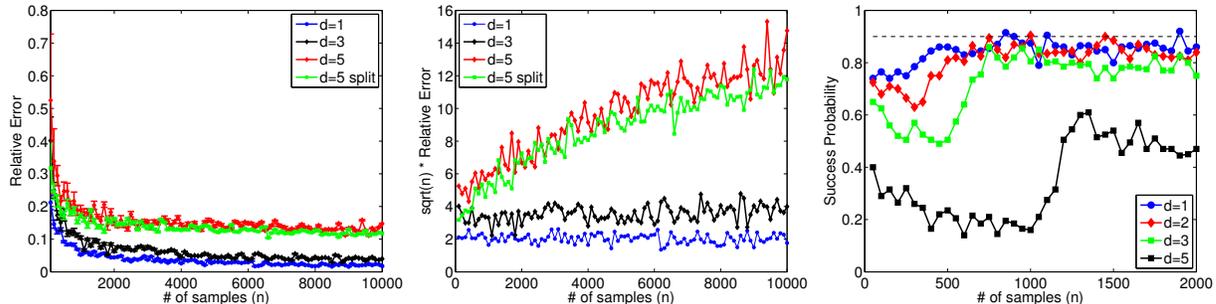


Figure 1: Simulation results showing the convergence rate of the error, rescaled convergence rate, and performance of the confidence interval (from left to right).

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  $\ell_1$  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 Titterton. 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 Seleznev. Estimation of entropy-type integral functionals. *arXiv:1209.2544*, 2012.
- [8] Gérard Kerkycharian 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 Rényi 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.

## A Proof of Theorems 3 and 4

We analyze the two estimators separately and Theorem 3 follows immediately from Theorems 10 and 11 below. For the estimator without data splitting, the result follows from below and the inequality:

$$\begin{aligned} & \mathbb{E} \left[ \left( \hat{\theta}_p + \hat{\theta}_q - 2\hat{\theta}_{pq} - \theta_p - \theta_q + 2\theta_{pq} \right)^2 \right] \\ & \leq \mathbb{E}[(\hat{\theta}_p - \theta_p)^2] + \mathbb{E}[(\hat{\theta}_q - \theta_q)^2] + \mathbb{E}[(\hat{\theta}_{pq} - \theta_{pq})^2] \\ & \quad + 2 \left| \mathbb{E}\hat{\theta}_p - \theta_p \right| \left| \mathbb{E}\hat{\theta}_q - \theta_q \right| + 4\sqrt{\mathbb{E}[(\hat{\theta}_{pq} - \theta_{pq})^2]} \left( \sqrt{\mathbb{E}[(\hat{\theta}_p - \theta_p)^2]} + \sqrt{\mathbb{E}[(\hat{\theta}_q - \theta_q)^2]} \right) \end{aligned}$$

Plugging in the bounds from Theorems 10 and 11 immediately establish the rate of convergence for the estimator without data splitting.

For the quadratic term estimators, we make a slight modification to a theorem from Gine and Nickl [5]. The only difference between our proof and theirs is in controlling the bias, where we use the bounded-variation assumption while they use a Sobolev assumption. However this has little bearing, as the bias is still of the same order, and we have the following theorem characterizing the behavior of the quadratic estimator:

**Theorem 10** (Adapted from [5]). *Under Assumption 2, we have:*

$$\left| \mathbb{E}[\hat{\theta}_p] - \theta_p \right| \leq c_b h^{2\beta} \quad \mathbb{E} \left[ (\hat{\theta}_p - \mathbb{E}[\theta_p])^2 \right] \leq c_v \left( \frac{1}{n} + \frac{1}{n^2 h^d} \right), \quad (17)$$

and when  $\beta > d/4$ :

$$\sqrt{n}(\hat{\theta}_p - \theta_p) \rightsquigarrow \mathcal{N}(0, 4 \operatorname{Var}_{x \sim p}(p(x))). \quad (18)$$

While we are not aware of any analyses of the bilinear term, it is not particularly different from the quadratic term, and we have the following theorem:

**Theorem 11.** *Under Assumption 2, we have:*

$$\left| \mathbb{E}[\hat{\theta}_{pq}] - \theta_{pq} \right| \leq c_b h^{2\beta} \quad \mathbb{E} \left[ (\hat{\theta}_{pq} - \mathbb{E}[\theta_{pq}])^2 \right] \leq c_v \left( \frac{1}{n} + \frac{1}{n^2 h^d} \right), \quad (19)$$

and when  $\beta > d/4$ :

$$\sqrt{n}(\hat{\theta}_{pq} - \theta_{pq}) \rightsquigarrow \mathcal{N}(0, \operatorname{Var}_{x \sim p}(q(x)) + \operatorname{Var}_{y \sim q}(p(y))). \quad (20)$$

*Proof of Theorem 10.* We reproduce the proof of Gine and Nickl for completeness. The bias can be bounded by:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_p] - \theta_p &= \int \int K_h(x, y) p(y) dy p(x) dx - \int p(x) p(x) dx = \int \int K_h(x, y) [p(y) - p(x)] p(x) dy dx \\ &= \int \int K(u) [p(x - uh) - p(x)] p(x) du dx = \int K(u) [(p_0 \star p)(uh) - (p_0 \star p)(0)] du, \end{aligned}$$

where  $p_0(x) = p(-x)$  and  $\star$  denotes convolution. Now by Lemma 14 below, we know that  $p_0 \star p \in \mathcal{W}_1^{2\beta}(C^2)$  and can take a Taylor expansion of order  $2\beta - 1$ . When we take such an expansion, by the properties of the kernel, all but the remainder term is annihilated and we are left with:

$$\frac{h^{2\beta}}{(2\beta)!} \sum_{r_1, \dots, r_d | \sum_i r_i = 2s} \int K(u) \Pi_i u_i^{r_i} \xi(r, uh) du \leq c_b h^{2\beta},$$

where we used the fact the function is integrable by the fact that  $\xi \in L^1$ , which in turn follows from the fact that  $p_0 \star p \in \mathcal{W}_1^{2\beta}(C^2)$  and by Taylor's remainder theorem. We are also using the compactness of  $K$  here so that we only have to integrate over  $(-1, 1)^d$  in which case all polynomial functions are also  $L_1$  integrable. This shows that the bias is  $O(h^{2\beta})$ .

Note that the main difference between our proof and that of Gine and Nickl is in the smoothness assumption, which comes into play here. Under the bounded variation assumption, we were able to argue that smoothness is additive under convolution. The same is true under the Sobolev assumption, and this property is exploited by Gine and Nickl in exactly the same way as we do here. Unfortunately, Hölder smoothness is not additive under convolution, so the more standard assumption does not provide the semiparametric rate of convergence.

As for the variance, we may write:

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

which we can split into three cases. When  $i \neq j \neq s \neq t$ , each term in the sum is exactly  $(\mathbb{E}\hat{\theta})^2$ , and this happens for  $n(n-1)(n-2)(n-3)$  terms in the sum. When one of the first indices is equal to one of the second indices we get:

$$\begin{aligned} \mathbb{E}K_h(X_i, X_j)K_h(X_i, X_t) &= \int \int \int K_h(X_i, X_j)K_h(X_i, X_t)p(X_i)p(X_j)p(X_t)dX_idX_jdX_t \\ &= \int \int \int K(u_j)K(u_t)p(X_i - u_jh)p(X_i - u_th)p(X_i)du_jdu_tdX_i \\ &\leq \|K\|_2^2 \|p\|_2^2, \end{aligned}$$

where we performed a substitution to annihilate the dependence on  $h$ . There are  $4n(n-1)(n-2)$  expressions of this form, so in total, these terms contribute:

$$\frac{1}{n} \|K\|_2^2 \|p\|_2^2.$$

Finally, the  $2n(n-1)$  terms where  $i = s, j = t$  or vice versa in total contribute:

$$\begin{aligned} \frac{2}{n(n-1)} \mathbb{E}K_h^2(X_i, X_j) &= \frac{2}{h^{2d}n(n-1)} \int K^2\left(\frac{X_i - X_j}{h}\right)p(X_i)p(X_j)dX_idX_j \\ &= \frac{2}{h^{2d}n(n-1)} \int K^2(u_j)p(X_i)p(X_i - u_jh)du_jdX_i \\ &\leq \frac{2\|K\|_2^2 \|p\|_2^2}{h^{2d}n^2}. \end{aligned}$$

Adding together these terms, establishes the variance bound in the theorem. The rate of convergence in Theorem 3 follows from plugging the definition of  $h$ , which was selected to optimize the tradeoff between bias and variance.

As for asymptotic normality, we decompose the proof into several steps.

1. Control the bias.
2. Apply Hoeffding's decomposition.
3. Control the second order term, which will be lower order.
4. Show that the first order term is close to  $P_n p - \theta$  (here  $P_n$  is the empirical measure).
5. Apply the Lindberg-Levy central limit theorem to  $P_n p - \theta$ .

As usual we have the decomposition:

$$\hat{\theta}_p - \theta_p = \underbrace{\hat{\theta}_p - \mathbb{E}\hat{\theta}_p}_{\text{Variance}} + \underbrace{\mathbb{E}\hat{\theta}_p - \theta}_{\text{Bias}}.$$

We already controlled the bias above. Specifically we know that  $\sqrt{n}(\mathbb{E}\hat{\theta}_p - \theta_p) \leq \sqrt{nh^{2\beta}} \rightarrow 0$  with our setting of  $h$  and under the assumption that  $\beta > d/4$ .

As is common in the analysis of U-statistics, we apply Hoeffding's decomposition before proceeding. That is, we write:

$$\hat{\theta}_p - \mathbb{E}\hat{\theta}_p = U_n(\pi_2 K_h) + 2P_n(\pi_1 K_h),$$

where  $U_n f = \frac{1}{n(n-1)} \sum_{i \neq j} f(X_i, X_j)$  is the U-process and  $P_n f = \frac{1}{n} \sum_i f(X_i)$  is the empirical process and:

$$\begin{aligned} (\pi_1 K_h)(X) &= \mathbb{E}_{x \sim p} K_h(x, X) - \mathbb{E}_{x, y \sim p} K_h(x, y) \\ (\pi_2 K_h)(X, Y) &= K_h(X, Y) - \mathbb{E}_{x \sim p} K_h(x, Y) - \mathbb{E}_{y \sim p} K_h(X, y) + \mathbb{E}_{x, y \sim p} K_h(x, y). \end{aligned}$$

It is easy to verify that our estimator can be decomposed in this manner. Moreover, since everything is centered, the two terms also have zero covariance. Also notice that  $\mathbb{E}_{x \sim p} K_h(x, Y) = \bar{p}(Y)$  and  $\mathbb{E}_{x, y \sim p} K_h = \int \bar{p}(x)p(x)$  where  $\bar{p}$  is the expectation of the density estimate.

We now control the second order term  $U_n(\pi_2 K_h)$  by showing convergence in quadratic mean.

$$\mathbb{E}[(U_n(\pi_2 K_h))^2] = \frac{1}{n(n-1)} \mathbb{E}[(\pi_2 K_h(X_1, X_2))^2] \leq \frac{c}{n^2 h^d} \|K\|_2^2 \|p\|_2^2.$$

The first equality follows from the fact that each term is conditionally centered, so all cross terms are zero, while the inequality is the result of performing a substitution as we have seen before. Thus  $\sqrt{n}U_n(\pi_2 K_h) \rightarrow 0$  since  $\frac{1}{nh^d} \rightarrow 0$  when  $\beta > d/4$ .

For the first order term  $P_n(\pi_1 K_h)$ , we now show that it is close to  $P_n p - \theta_p$ .

$$\mathbb{E}[(P_n(\pi_1 K_h) - (P_n p - \int p^2))^2] \leq \frac{1}{n} \mathbb{E}[(\bar{p}(X) - p(x))^2] \leq \frac{\|\bar{p} - p\|_\infty^2}{n} = \frac{ch^{2\beta}}{n},$$

so that  $\sqrt{n}P_n(\pi_1 K_h) \xrightarrow{q.m.} \sqrt{n}(P_n p - \theta_p)$  since  $h^{2\beta} \rightarrow 0$ .

Now by the Lindberg-Levy CLT, we know that:

$$\sqrt{n}(2P_n p - 2\theta_p) \rightsquigarrow \mathcal{N}(0, 4 \text{Var}_{x \sim p}(p(X))),$$

which concludes the proof of the theorem.  $\square$

We now prove Theorem 11, although the arguments are fairly similar.

*Proof of Theorem 11.* The bias is:

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

where, as before,  $p_0(x) = p(-x)$  and  $\star$  denotes convolution. So we can proceed as in the quadratic setting. Specifically, by Lemma 14, we can take a Taylor expansion of order  $2\beta + 1$ , annihilate all but the remainder term, which we know is bounded by the fact that  $p_0 \star q \in \mathcal{W}_1^{2\beta}(C^2)$ . Formally, the remainder term is:

$$\frac{h^{2\beta}}{(2\beta)!} \sum_{r_1, \dots, r_d | \sum_i r_i = 2s} \int K(u) \Pi_i u_i^{r_i} \xi(r, uh) du \leq c_b h^{2\beta},$$

where we used the fact the function is integrable by the fact that  $\xi \in L^1$ , since  $p_0 \star q \in \mathcal{W}_1^{2\beta}(C^2)$ . Thus the bias is  $O(h^{2\beta})$ .

The variance can be bounded in a similar way to the quadratic estimator:

$$\mathbb{E}[\hat{\theta}_{pq}^2] - \mathbb{E}[\theta_{pq}]^2 = \frac{1}{n^4} \sum_{i, j, s, t} \mathbb{E}[K_h(X_i, Y_j) K_h(X_s, Y_t)] - \mathbb{E}[\hat{\theta}_{pq}]^2.$$

Whenever  $i \neq s$  and  $j \neq t$  all of the terms are independent so they cancel out with the  $\mathbb{E}[\hat{\theta}_{pq}]^2$  term. This happens for  $n^2(n-1)^2$  terms.

When  $i = s, j \neq t$ , we substitute  $u_j = h^{-1}(X_i - Y_j)$  and  $u_t = h^{-1}(X_i - Y_t)$  for  $Y_j, Y_t$  to see that:

$$\begin{aligned} \frac{1}{n^4} \sum_{i,j \neq t} \mathbb{E}[K_h(X_i, Y_j)K_h(X_i, Y_t)] &= \frac{n-1}{h^{2d}n^2} \int \int \int K\left(\frac{X_i - Y_j}{h}\right)K\left(\frac{X_i - Y_t}{h}\right)p(X_i)q(Y_j)q(Y_t) \\ &= \frac{n-1}{n^2} \int \int \int K(u_j)K(u_t)p(X_i)q(X_i - u_jh)q(X_i - u_th) \\ &\leq \frac{1}{n} \|K\|_2^2 \|q\|_2^2. \end{aligned}$$

Thus, the total contribution from the terms where  $j = t, i \neq s$  is bounded by  $\frac{1}{n} \|K\|_2^2 \|p\|_2^2$ .

When  $j = t, i = s$ , we can only perform one substitution so a factor of  $h^d$  will remain. Formally:

$$\begin{aligned} \frac{1}{n^2 h^{2d}} \int \int K^2\left(\frac{X_i - Y_j}{h}\right)p(X_i)q(Y_j) &= \frac{1}{n^2 h^d} \int \int K^2(u_j)p(X_i)q(X_i - u_jh) \\ &\leq \frac{1}{n^2 h^d} \|K\|_2^2 \|p\|_2 \|q\|_2. \end{aligned}$$

Therefore, the total variance is  $O(n^{-1} + n^{-2}h^{-d})$  as in the theorem statement.

The proof of asymptotic normality of the bilinear estimator is not too different from the proof for the quadratic estimator. We can start by ignoring the bias, as when  $b \geq d/4$ , we know that  $\sqrt{n}(\mathbb{E}\hat{\theta}_{pq} - \theta_{pq}) \rightarrow 0$ . To analyze the variance term we make use of the following decomposition:

$$\begin{aligned} \hat{\theta}_{pq} - \mathbb{E}\hat{\theta}_{pq} &= \frac{1}{n^2} \sum_{ij} K_h(X_i, Y_j) + \frac{1}{n} \sum_i \bar{q}(X_i) - \frac{1}{n} \sum_i \bar{q}(X_i) + \frac{1}{n} \sum_j \bar{p}(Y_j) - \frac{1}{n} \sum_j \bar{p}(Y_j) - \mathbb{E}\hat{\theta}_{pq} \\ &= V_n(\pi_2 K_h) + P_n(\pi_{11} K_h) + Q_n(\pi_{12} K_h), \end{aligned}$$

Where:

$$\begin{aligned} (\pi_2 K_h)(X, Y) &= K_h(X, Y) - \bar{q}(X) - \bar{p}(Y) + \mathbb{E}K_h(x, y) \\ (\pi_{11}(K_h))(X) &= \bar{q}(X) - \mathbb{E}K_h(x, y) \\ (\pi_{12}(K_h))(Y) &= \bar{p}(Y) - \mathbb{E}K_h(x, y). \end{aligned}$$

Here  $P_n, Q_n$  are the empirical processes associated with the samples  $X, Y$  respectively and  $\bar{p}, \bar{q}$  are the expectations of the kernel density estimators. Also,  $V_n$  is the  $V$ -process, that is  $V_n f = \frac{1}{n^2} \sum_{i,j} f(X_i, Y_j)$ . Notice that each term is conditionally centered, which implies that each pair of terms has zero covariance. Thus we only have to look at the variances.

As before, the goal is to show that the  $V$ -process term is lower order and then to apply the Lindeberg-Levy CLT to the other two terms. Since each term is conditionally centered:

$$\mathbb{E}[(V_n(\pi_2 K_h))^2] = \frac{1}{n^2} \mathbb{E}[(\pi_2 K_h(X, Y))^2] \leq \frac{c}{n^2 h^d} \|K\|_2^2 \|p\|_2 \|q\|_2,$$

where the last step follows by performing the substitution  $u = \frac{X-Y}{h}$  in each term of the integral.

For the first order terms, we first show that they are close to  $q(X) - \mathbb{E}[q(x)]$  and  $p(Y) - \mathbb{E}[p(y)]$  so that we can apply the CLT to the latter. We will show convergence in quadratic mean.

$$\mathbb{E} \left[ (P_n(\pi_{11} K_h) - (P_n q - \int p q))^2 \right] \leq \frac{1}{n} \mathbb{E} [(\bar{q}(x) - q(x))^2] \leq \frac{\|\bar{q} - q\|_\infty^2}{n} \leq \frac{h^{2\beta}}{n},$$

which means that, under our choice of  $h$  and with  $\beta > d/4$ ,  $\sqrt{n}P_n(\pi_{11} K_h) \rightarrow^{q.m.} \sqrt{n}P_n(q - \theta_{pq})$ . Exactly the same argument shows that  $\sqrt{n}Q_n(\pi_{12} K_h) \rightarrow^{q.m.} \sqrt{n}Q_n(p - \theta_{pq})$ .

Finally, by the Lindeberg-Levy CLT, we know that:

$$\sqrt{n}(P_n q - \theta_{pq}) \rightsquigarrow \mathcal{N}(0, \text{Var}_{x \sim p}(q(x))), \quad \sqrt{n}(Q_n p - \theta_{pq}) \rightsquigarrow \mathcal{N}(0, \text{Var}_{y \sim q}(p(y))),$$

and since  $x$  and  $y$  are independent, both of these central limit theorems hold jointly. Since in our estimate for  $\hat{D}$  we have a term of the form  $2\hat{\theta}_{pq}$ , the contribution of this term to the total variance is  $4 \text{Var}(\hat{\theta}_{pq})$ . This concludes the proof.  $\square$

## B Proof of Theorem 6

In this section we fill in the missing details in the proof of Theorem 6. We will apply the Berry-Esséen inequality for multi-sample U-statistics from Chen, Goldstein and Shao [4], which we reproduce below.

In order to state the theorem we need to make several definitions. We make some simplifications to their result for ease of notation. Consider  $k$  independent sequences  $X_{j1}, \dots, X_{jn}$   $j = 1, \dots, k$  of i.i.d. random variables, all of length  $n$  (this can be relaxed). Let  $m_j \geq 1$  for each  $j$  and let  $\omega(x_{jl}, l \in [m_j], j \in [k])$  be a function that is symmetric with respect to the  $m_j$  arguments. In other words,  $\omega$  is invariant under permutation of two arguments from the same sequence. Let  $\theta = \mathbb{E}\omega(X_{jl})$ .

The multi-sample U-statistic is defined as:

$$U_n = \left\{ \prod_{j=1}^k \binom{n}{m_j}^{-1} \right\} \sum \omega(X_{jl}, j \in [k], l = i_{j1}, \dots, i_{jm_j}), \quad (21)$$

where the sum is carried out over all indices satisfying  $1 \leq i_{j1} < \dots < i_{jm} \leq n$ .

Let:

$$\sigma^2 = \mathbb{E}\omega^2(X_{jl}),$$

and for each  $j \in [k]$  define:

$$\omega_j(x) = \mathbb{E}[\omega(X_{jl}) | X_{j1} = x],$$

with:

$$\sigma_j^2 = \mathbb{E}\omega_j^2(X_{j1}).$$

Lastly, define:

$$\sigma_n^2 = \sum_{j=1}^k \frac{m_j^2}{n} \sigma_j^2.$$

We are finally ready to state the theorem:

**Theorem 12** (Theorem 10.4 of [4]). *Assume that  $\theta = 0$ ,  $\sigma^2 < \infty$ ,  $\max_{j \in [k]} \sigma_j^2 > 0$ . Then for  $2 < p \leq 3$ :*

$$\sup_{z \in \mathbb{R}} |\mathbb{P}(\sigma_n^{-1} U_n \leq z) - \Phi(z)| \leq \frac{6.1}{\sigma_n^p} \sum_{j=1}^k \frac{m_j^p}{n^{p-1}} \mathbb{E}[|\omega_j(X_{j1})|^p] + \frac{(1 + \sqrt{2})\sigma}{\sigma_n} \sum_{j=1}^k \frac{m_j^2}{n}. \quad (22)$$

As we did in our estimator, we split the data into four groups, two samples of size  $n$  from each distribution, which we will denote with superscripts, i.e.  $X_i^{(1)}$  will be the  $i$ th sample from the first group of the data from  $p$ . We can write  $\hat{D} - \mathbb{E}\hat{D}$  as a zero-mean multi-sample U-statistic with four groups where the first  $X$  and  $Y$  groups are used for  $\hat{\theta}_p - \theta_p$  and  $\hat{\theta}_q - \theta_q$  respectively, while the second two groups are used for the cross term  $\hat{\theta}_{pq} - \theta_{pq}$ .

In other words,  $\omega$  will be a function that takes 6 variables, two from the  $X^{(1)}$  group, two from the  $Y^{(1)}$  group and one each from the  $X^{(2)}$  and  $Y^{(2)}$  groups. Formally, we define:

$$\begin{aligned} \omega(x_{11}, x_{12}, y_{11}, y_{12}, x_{21}, y_{21}) \\ = K_h(x_{11}, x_{12}) - \mathbb{E}\hat{\theta}_p + K_h(y_{11}, y_{12}) - \mathbb{E}\hat{\theta}_q - 2K_h(x_{21}, y_{21}) + 2\mathbb{E}\hat{\theta}_{pq}. \end{aligned}$$

With this definition, it is clear that  $U_n = \hat{D} - \mathbb{E}\hat{D}$ .

To apply Theorem 12 on the appropriate term in the proof, we just have to bound a number of quantities involving  $\omega$ . As we will see, we will not achieve the  $n^{-1/2}$  rate because the function  $\omega$  depends on the bandwidth  $h$ , which is decreasing, so the variance  $\sigma$  is increasing. Specifically:

$$\sigma^2 = \mathbb{E}[\omega^2(X_{11}, X_{12}, Y_{11}, Y_{12}, X_{21}, Y_{21})]$$

$$= \mathbb{E}(K_h(X_{11}, X_{12}) - \mathbb{E}\hat{\theta}_p)^2 + \mathbb{E}(K_h(Y_{11}, Y_{12}) - \mathbb{E}\hat{\theta}_q)^2 + 4\mathbb{E}(K_h(X_{21}, Y_{21}) - \mathbb{E}\hat{\theta}_{pq})^2.$$

Each of the three terms can be analyzed in exactly the same way so we focus on the first term:

$$\begin{aligned} \mathbb{E}(K_h(X_{11}, X_{12}) - \mathbb{E}\hat{\theta}_p)^2 &\leq \int \int K_h^2(X_1, X_2)p(X_1)p(X_2) = \frac{1}{h^d} \int \int K(u)p(X_1 + uh)p(X_1) \\ &\leq \frac{1}{h^d} \|K\|_2^2 \|p\|_2^2, \end{aligned}$$

and the same substitution on the other two terms shows that the variance is:

$$\sigma^2 \leq \frac{1}{h^d} \|K\|_2^2 (\|p\|_2^2 + \|q\|_2^2 + 4\|p\|_2\|q\|_2). \quad (23)$$

A similar argument gives us a bound on  $\sigma_j^2$   $j = 1, \dots, 4$ . First, since the other terms are centered, we can write  $\omega_1(x) = \mathbb{E}(K_h(x, X_2)) - \mathbb{E}\hat{\theta}_p$  with similar expressions for the other terms. Then,  $\sigma_1^2$  can be simplified to:

$$\begin{aligned} \mathbb{E}\omega_1^2(X) &= \mathbb{E}(\mathbb{E}K_h(X_1, X_2))^2 - (\mathbb{E}\hat{\theta}_p)^2 \\ &\leq \int \left( \int K_h(X_1, X_2)p(X_2) \right)^2 p(X_1) = \int \left( \int K(u)p(X_1 - uh) \right)^2 p(X_1) \leq \|K\|_\infty^2. \end{aligned}$$

With exactly the same argument for the other three. Thus:

$$\sigma_n^2 = \frac{1}{n} \sum_{j=1}^4 m_j^2 \sigma_j^2 \leq \frac{10}{n} \|K\|_\infty^2. \quad (24)$$

The last thing we need is the third moments of the linearizations  $\mathbb{E}[|\omega_j(x)|^3]$ .

$$\begin{aligned} \mathbb{E} \left[ \left| \mathbb{E}K_h(X_1, X_2) - \mathbb{E}\hat{\theta}_p \right|^3 \right] &= \int \left| \int K_h(x, X_2)p(X_2) - \int \int K_h(X_1, X_2)p(X_1)p(X_2) \right|^3 p(x) \\ &= \int \left| \int K(u)p(x - uh) - \int \int K(u)p(X_1 - uh)p(X_1) \right|^3 p(x) \\ &= \int \left| \int K(u) \left( p(x - uh) - \int p(X_1 - uh)p(X_1) \right) \right|^3 p(x) \\ &\leq 8\|K\|_\infty^3 \|p\|_\infty^3 \end{aligned}$$

It is easy to verify that each of the third moments are bounded by:

$$\mathbb{E}|\omega_j|^3 \leq 8\|K\|_\infty^3 (\|p\|_\infty^3 + \|q\|_\infty^3), \forall j. \quad (25)$$

And plugging in all of these calculations into Theorem 12 shows that:

$$\begin{aligned} \sup_z |\mathbb{P}(\sqrt{n}\tilde{\sigma}_n^{-1}U_n \leq z) - \Phi(z)| &\leq \\ &\leq \frac{n^{3/2}(6.1)(18)}{n^2 10^{3/2} \|K\|_\infty^3} 8\|K\|_\infty^3 (\|p\|_\infty^3 + \|q\|_\infty^3) + \frac{\sqrt{n}(1 + \sqrt{2})}{n\sqrt{h^d}\sqrt{10}\|K\|_\infty} \|K\|_2 \sqrt{\|p\|_2^2 + \|q\|_2^2 + 4\|p\|_2\|q\|_2} \\ &\leq \frac{27}{\sqrt{n}} (\|p\|_\infty^3 + \|q\|_\infty^3) + \frac{8}{\sqrt{nh^d}} \frac{\|K\|_2}{\|K\|_\infty} \sqrt{\|p\|_2^2 + \|q\|_2^2 + 4\|p\|_2\|q\|_2}. \end{aligned}$$

This gives the bound in Equation 15.

## C Proof of Theorem 7

For completeness we introduce the construction used by Krishnamurthy et al [9]. For the remainder of the proof, we will work of  $[0, 1]^d$  and assume that  $p$  is pointwise lower bounded by  $1/\kappa_l$ , noting that a lower bound

here applies to the more general setting. For the construction, suppose we have a disjoint collection of subset  $A_1, \dots, A_m \subset [0, 1]^d$  for some parameter  $m$  with associated functions  $u_j$  that are compactly supported on  $A_j$ . Specifically assume that we have  $u_j$  satisfying:

$$\text{supp}(u_j) \subset \{x | B(x, \epsilon) \subset A_j\}, \|u_j\|_2^2 = \Omega(m^{-1}), \int_{A_j} u_j = \int_{A_j} p_0(x)u_j(x) = \int_{A_j} q_0(x)u_j(x) = 0, \|D^r u_j\|_1 \asymp m^{r/d-1}$$

The first condition ensure that the  $u_j$ s are orthogonal to each other, while the second and third will ensure separation in terms of  $L_2^2$  divergence. The last condition holds for all derivative operators with  $r \leq \beta$  and it will ensure that the densities we construct belong to the bounded variation class. The only difference between these requirements and those from [9] are the orthogonality to  $p, q$ , and the bounded-variation condition, which replaces a point-wise analog.

Deferring the question of existence of these functions, we can proceed to construct  $p_\lambda$ . Let the index set  $\Lambda = \{-1, +1\}^m$  and define the functions  $p_\lambda = p_0 + K \sum_{j=1}^m \lambda_j u_j$ , where  $K$  will be defined subsequently. A simple computation then reveals that:

$$\begin{aligned} T(p_0, q_0) - T(p_\lambda, q_0) &= \int p_0^2 - p_\lambda^2 + 2 \left[ \int p_\lambda q_0 - \int p_0 q_0 \right] \\ &= \int (p_0 - p_\lambda)(p_0 + p_\lambda) + 2 \left[ \int p_\lambda q_0 - \int p_0 q_0 \right] \\ &= K^2 \sum_{j=1}^m \|u_j\|_2^2 = \Theta(K^2) \end{aligned}$$

where we expand  $p_\lambda$  and use the orthogonality properties extensively. This gives us the desired separation.

To bound the hellinger distance, we use Theorem 1 of Birge and Massart [3] and the argument following Theorem 12 of Krishnamurthy et al [9].

**Theorem 13.** [3] Consider a set of densities  $p_0$  and  $p_\lambda = p[1 + \sum_j \lambda_j v_j(x)]$  for  $\lambda \in \Lambda = \{-1, 1\}^m$  with partition  $A_1, \dots, A_m \subset [0, 1]^d$ . Suppose that (i)  $\|v_j\|_\infty \leq 1$ , (ii)  $\|\mathbf{1}_{A_j^c} v_j\|_1 = 0$ , (iii)  $\int v_j p_0 = 0$  and (iv)  $\int v_j^2 p_0 = \alpha_j > 0$  all hold with:

$$\alpha = \sup_j \|v_j\|_\infty, s = n\alpha^2 \sup_j P_0(A_j), c = n \sup_j \alpha_j$$

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

$$h^2(P_0^n, \overline{P^n}) \leq C(\alpha, s, c) n^2 \sum_{j=1}^m \alpha_j^2 \quad (26)$$

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

The exact same bound on the hellinger distances holds for the measures  $P_0^n \times Q_0^n$  against  $\overline{P^n} \times Q^n$ . Defining  $v_j = K u_j / p_0$  then the densities we used in our construction meet the specification in the above theorem. We immediately satisfy the first three requirements and we have  $\int v_j^2 p = K^2 \int u_j^2 / p \leq K^2 \kappa_l / m \triangleq \alpha_j$ . Thus we have the hellinger bound of:

$$h^2(P_0^n \times Q_0^n, \overline{P^n} \times Q^n) \leq (1/3) n^2 \sum_{j=1}^m \alpha_j^2 \leq \frac{C n^2 K^4}{m}$$

We lastly have to make sure that the  $p_\lambda$  functions satisfy the bounded variation assumption. This follows from an application of the triangle inequality provided that  $\|D^r u_j\|_1 \leq O(m^{r/d-1})$ .

$$\|D^r p_\lambda\|_1 = \|D^r p + K \sum_{j=1}^m \lambda_j D^r u_j\|_1 \leq \|D^r p\| + K \sum_{j=1}^m \|D^r u_j\|_1 \leq \|D^r p\| + K \sum_{j=1}^m \|D^r u_j\|_1 \leq \|D^r p\| + O(K m^{r/d})$$

So as long as  $K \asymp m^{-r/d}$  and there is some wiggle room around the bounded variation assumption for  $p, p_\lambda$  will meet the bounded variation assumption.

Before we construct the  $u_j$ s, we put everything together. We must select  $K \asymp m^{-\beta/d}$  so that  $p_\lambda \in \mathcal{W}_1^\beta(C)$ , and then to make the hellinger distance  $O(1)$ , we must set  $m \asymp n^{\frac{2d}{4\beta+d}}$ . This makes  $K^2 \asymp n^{\frac{-4\beta}{4\beta+d}}$  which is precisely the lower bound on the convergence rate in absolute error.

Lastly we present the construction of the  $u_j$  functions. The construction is identical to the one used by Krishnamurthy et al [9], but we must make some modifications to ensure that bounded variation condition is satisfied. We reproduce the details here for completeness.

Let  $\{\phi_j\}_{j=1}^q$  be an orthonormal collection of functions for  $L^2([0, 1]^d)$  with  $q \geq 4$ . We can choose  $\phi_j$  to satisfy (i)  $\phi_1 = 1$ , (ii)  $\phi_j(x) = 0$  for  $x|B(x, \epsilon) \not\subset [0, 1]^d$  and (iii)  $\|D^r \phi_j\|_\infty \leq \kappa < \infty$  for all  $j$ . Certainly we can find such an orthonormal system.

Now for any pair of function  $f, g \in L^2([0, 1]^d)$ , we can find a unit-normed function in  $\tilde{w} \in \text{span}(\phi_j)$  such that  $\tilde{w} \perp \phi_1, \tilde{w} \perp f, \tilde{w} \perp g$ . If we write  $\tilde{w} = \sum_j c_j \phi_j$ , we have  $D^r \tilde{w} = \sum_j c_j D^r \phi_j$  so that  $\|D^r \tilde{w}\|_\infty \leq \kappa \sum |c_j| \leq \kappa \sqrt{q}$  since  $\tilde{w}$  is unit normed. Thus the vector  $w = \tilde{w}/(K\sqrt{q})$  has  $\ell_2$  norm equal to  $(K\sqrt{q})^{-1}$  while have  $\|D^r w\|_\infty \leq 1$  for all tuples  $r$ .

For the  $u_j$  functions, we use the partition  $A_j = \prod_{i=1}^d [j_i m^{-1/d}, (j_i + 1)m^{-1/d}]$  where  $j = (j_1, \dots, j_d)$  and  $j_i \in [m^{1/d}]$  for each  $i$ . Map  $A_j$  to  $[0, 1]^d$  and appropriately map the densities  $p, q$  from  $A_j$  to  $[0, 1]^d$ . We construct  $u_j$  by using the construction for  $w$  above on the segment of the density corresponding to  $A_j$ . In particular, let  $w_j$  be the function from above and let  $u_j = w_j(m^{1/d}(x - (j_1, \dots, j_d)))$ . With this rescaling and shift,  $u_j \in A_j$ ,  $\text{supp}(u_j) \subset \{x|B(x, \epsilon) \in A_j\}$ , and  $\int u_j^2(x) = m^{-1} \int w_j^2(x) = \Theta(1/m)$ . For the last property, by a change of variables and Hölder's inequality, we have:

$$\|D^r u_j\|_1 = \int |D^r w_j(m^{1/d}(x - (j_1, \dots, j_d)))| d\mu(x) = \frac{1}{m} \int \|m^{r/d} D^r w_j(y)\| dA_j(y) \leq m^{r/d-1}.$$

Thus these function  $u_j$  meet all of the requirements.

## D Proof of Lemma 8

Recall that the asymptotic variance of the estimator is:

$$\sigma^2 = 4 \left( \text{Var}_{X \sim p}(p(X)) + \text{Var}_{Y \sim q}(q(Y)) + \text{Var}_{X \sim p}(q(X)) + \text{Var}_{Y \sim q}(p(X)) \right),$$

and our estimator  $\hat{\sigma}^2$  is formed by simply plugging in kernel density estimates  $\hat{p}, \hat{q}$  for all occurrences of the densities. We will first bound:

$$\mathbb{E}_{X_1^n, Y_1^n} [|\sigma^2 - \hat{\sigma}^2|] = O(n^{\frac{-\beta}{2\beta+d}}),$$

and our high probability bound will follow from Markov's inequality. We will show the following bounds, and the expected  $\ell_1$  bound will follow by application of the triangle inequality. Below, let  $f, g \in \mathcal{W}_1^\beta(C)$  be any two densities; we will interchangeably substitute  $p, q$  for  $f, g$ .

$$\mathbb{E} \left[ \left| \int \hat{f}^3 - \int f^3 \right| \right] \leq O \left( h^\beta + \frac{1}{(nh^d)^{1/2}} \right) \quad (27)$$

$$\mathbb{E} \left[ \left| \left( \int \hat{f}^2 \right)^2 - \left( \int f^2 \right)^2 \right| \right] \leq O \left( h^{2\beta} + \frac{1}{\sqrt{n}} + \frac{1}{nh^{d/2}} \right) \quad (28)$$

$$\mathbb{E} \left[ \left| \int \hat{f}^2 \hat{g} - \int f^2 g \right| \right] \leq O \left( h^\beta + \frac{1}{\sqrt{nh^d}} \right) \quad (29)$$

$$\mathbb{E} \left[ \left| \left( \int \hat{f} \hat{g} \right)^2 - \left( \int f g \right)^2 \right| \right] \leq O \left( h^{2\beta} + \frac{1}{\sqrt{n}} + \frac{1}{nh^{d/2}} \right) \quad (30)$$

Before establishing the above inequalities, let us conclude the proof. The overall rate of convergence in absolute loss is  $O(h^\beta + \frac{1}{\sqrt{nh^d}})$ . TBy choosing  $h \asymp n^{\frac{-1}{2\beta+d}}$ , the rate of convergence is  $O(n^{\frac{-\beta}{2\beta+d}})$ . Finally we wrap up with an application of Markov's Inequality.

Now we turn to establishing the bounds. For Equation 27, we can write:

$$\begin{aligned} \mathbb{E} \left[ \left| \int \hat{f}^3 - \int f^3 \right| \right] &\leq \mathbb{E} \|\hat{f} - f\|_3^3 + 3\mathbb{E} \left[ \int |f(x)\hat{f}(x)(f(x) - \hat{f}(x))| d\mu(x) \right] \\ &\leq \mathbb{E} \|f - \hat{f}\|_3^3 + 3\mathbb{E} \|f - \hat{f}\|_\infty \|f\hat{f}\|_1 \\ &\leq O \left( h^{3\beta} + \frac{1}{(nh^d)^{3/2}} + h^\beta + \frac{1}{(nh^d)^{1/2}} \right). \end{aligned}$$

The first step is a fairly straightforward expansion followed by the triangle inequality while in the second step we apply Hölder's inequality. The last step follows from well known analysis on the rate of convergence of the kernel density estimator.

For Equation 28 we should actually use the  $U$ -statistic estimator for  $\theta_p$  that we have been analyzing all along. The bound above follows from Theorem 10 and the following chain of inequalities:

$$\begin{aligned} \mathbb{E} \left[ \left| \left( \int \hat{f}^2 \right)^2 - \left( \int f^2 \right)^2 \right| \right] &\leq \mathbb{E} \left[ \left( \int \hat{f}^2 - f^2 \right)^2 \right] + 2\|f\|_2^2 \mathbb{E} \left[ \left| \int \hat{f}^2 - f^2 \right| \right] \\ &\leq \mathbb{E} \left[ \left( \int \hat{f}^2 - f^2 \right)^2 \right] + C \sqrt{\mathbb{E} \left[ \left( \int \hat{f}^2 - f^2 \right)^2 \right]} \\ &\leq O \left( h^{4\beta} + \frac{1}{n} + \frac{1}{n^2 h^d} + h^{2\beta} + \frac{1}{\sqrt{n}} + \frac{1}{nh^{d/2}} \right). \end{aligned}$$

The first inequality is a result of some simple manipulations followed by the triangle inequality and the second step is Jensen's inequality. We already have a bound on the MSE of the estimator  $\hat{\theta}_p - \theta_p$  which gives us the inequality in Equation 28. Applying that bound leads to the last inequality.

The bound for Equation 30 follows from exactly the same argument with an application Theorem 11 instead of Theorem 10 in the last step. So we simply need to establish Equation 29.

$$\begin{aligned} \mathbb{E} \left[ \left| \int \hat{f}^2 \hat{g} - \int f^2 g \right| \right] &= \mathbb{E} \left[ \left| \int (\hat{f}^2 - f^2) \hat{g} \right| \right] + \mathbb{E} \left[ \left| \int f^2 (\hat{g} - g) \right| \right] \\ &\leq \mathbb{E} \|\hat{f}^2 - f^2\|_2 \|\hat{g}\|_2 + \|f^2\|_2 \|\hat{g} - g\|_2 \\ &\leq \mathbb{E} \|\hat{f}^2 - f^2\|_2 (\|\hat{g} - g\|_2 + \|g\|_2) + \|f^2\|_2 \|\hat{g} - g\|_2 \\ &\leq O \left( h^{2\beta} + \frac{1}{nh^{d/2}} + \frac{1}{\sqrt{n}} + h^\beta + \frac{1}{\sqrt{nh^d}} \right). \end{aligned}$$

Here we use that  $\|\hat{g}\|_1 = 1$  and that  $\|f^2\|_2$  and  $\|g\|_2$  are both bounded. We use the standard rate of convergence analysis of the kernel density estimator to bound  $\mathbb{E} \|\hat{g} - g\|_2 \leq O(h^\beta + (nh^d)^{-1})$ . We finally use Theorem 10 to bound  $\|\hat{f}^2 - f^2\|_2$ . Note that we are exploiting independence between the samples for  $\hat{f}$  and  $\hat{g}$  to push the expectation inside of the product in the first term. In the last line we omitted the term  $\mathbb{E} \|\hat{f}^2 - f^2\|_2 \|\hat{g} - g\|_2$  since it converges much faster than the other two terms.

To prove the second bound, we show that  $\bar{\sigma}^2$  is close to  $\sigma^2$ . We just have to look at two forms:

$$T_1 = \int \bar{p}^2(x)p(x) - \int p^3(x) \quad T_2 = \left( \int \bar{p}(x)p(x) \right)^2 - \left( \int p^2(x) \right)^2.$$

For  $T_1$  we can write:

$$T_1 = \int (\bar{p}^2(x) - p^2(x))p(x) = \int (\bar{p}(x) - p(x))(\bar{p}(x) + p(x))p(x)$$

$$\begin{aligned}
 &= \int (\bar{p}(x) - p(x))^2 p(x) + 2 \int p^2(x) (\bar{p}(x) - p(x)) \\
 &\leq \left( \sup_x |\bar{p}(x) - p(x)| \right)^2 + 2 \|p\|_2^2 \sup_x |\bar{p}(x) - p(x)| \leq O(h^{2\beta} + h^\beta),
 \end{aligned}$$

since  $p$  is  $L_2$ -integrable and the kernel density estimator has point-wise bias  $O(h^\beta)$ .

For  $T_2$  we have:

$$\begin{aligned}
 T_2 &= \left( \int (\bar{p}(x) - p(x)) p(x) \right)^2 + 2 \left( \int p^2(x) \right)^2 \left( \int (\bar{p}(x) - p(x)) p(x) \right) \\
 &\leq \left( \sup_x |\bar{p}(x) - p(x)| \right)^2 + 2 \|p\|_2^4 \sup_x \|\bar{p}(x) - p(x)\| \leq O(h^{2\beta} + h^\beta).
 \end{aligned}$$

With  $h \asymp n^{\frac{-1}{2\beta+d}}$  the additional bias incurred is:

$$\mathbb{E} |\hat{\sigma}^2 - \bar{\sigma}^2| \leq \mathbb{E} |\hat{\sigma}^2 - \sigma^2| + |\sigma^2 - \bar{\sigma}^2| \leq O(n^{\frac{-\beta}{2\beta+d}}).$$

and so  $\hat{\sigma}^2$  is an equally good estimator of  $\sigma^2$  and  $\bar{\sigma}^2$  (up to constants).

## E A Convolution Lemma

In this section we show that bounded-variation smoothness is additive under convolution.

**Lemma 14.** *If  $f, g \in \mathcal{W}_1^\beta(\mathbb{R}^d, C)$ , then  $h = f \star g \in \mathcal{W}_1^{2\beta}(\mathbb{R}^d, C^2)$ .*

*Proof.* The proof uses the fact that:

$$\frac{\partial h(x)}{\partial x} = \left( \frac{\partial f}{\partial x} \star g \right)(x)$$

which follows by pushing the derivative operator inside of the integral and continuity of  $f, g$  and their derivatives. Using the above identity, we have:

$$\frac{\partial^{2\beta} h(x)}{\partial x^{2\beta}} = \left( \frac{\partial^\beta f}{\partial x^\beta} \star \frac{\partial^\beta g}{\partial x^\beta} \right)(x),$$

or more concisely:

$$\|h^{(2\beta)}\|_1 = \|f^{(\beta)} \star g^{(\beta)}\|_1 \leq \|f^{(\beta)}\|_1 \|g^{(\beta)}\|_1 \leq C^2.$$

The first inequality is Young's inequality. This implies that  $L_1$  is closed under convolution.

It is clear, by the fact that derivatives can be distributed across the convolution that for  $k < 2\beta$ ,  $D^k h \in L^1$ . This proof strategy extends mutatis mutandis to higher dimension.  $\square$