entropy

MDPI

Article

Ensemble Estimation of Information Divergence †

Kevin R. Moon [1,‡], Kumar Sricharan [2], Kristjan Greenewald [3] and Alfred O. Hero III [4,*]

[1] Genetics Department and Applied Math Program, Yale University, New Haven, CT 06520, USA; kevin.moon@yale.edu
[2] Intuit Inc., Mountain View, CA 94043, USA; sricharan_kumar@intuit.com
[3] IBM Research, Cambridge, MA 02142, USA; Kristjan.H.Greenewald@ibm.com
[4] Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109, USA
* Correspondence: hero@eecs.umich.edu; Tel.: +1-734-764-0564
† This paper is an extended version of our paper published in the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1133–1137.
‡ Current address: Department of Mathematics and Statistics, Utah State University, Logan, UT 84322, USA; kevin.moon@usu.edu

Received: 29 June 2018; Accepted: 26 July 2018; Published: 27 July 2018

Abstract: Recent work has focused on the problem of nonparametric estimation of information divergence functionals between two continuous random variables. Many existing approaches require either restrictive assumptions about the density support set or difficult calculations at the support set boundary which must be known a priori. The mean squared error (MSE) convergence rate of a leave-one-out kernel density plug-in divergence functional estimator for general bounded density support sets is derived where knowledge of the support boundary, and therefore, the boundary correction is not required. The theory of optimally weighted ensemble estimation is generalized to derive a divergence estimator that achieves the parametric rate when the densities are sufficiently smooth. Guidelines for the tuning parameter selection and the asymptotic distribution of this estimator are provided. Based on the theory, an empirical estimator of Rényi-α divergence is proposed that greatly outperforms the standard kernel density plug-in estimator in terms of mean squared error, especially in high dimensions. The estimator is shown to be robust to the choice of tuning parameters. We show extensive simulation results that verify the theoretical results of our paper. Finally, we apply the proposed estimator to estimate the bounds on the Bayes error rate of a cell classification problem.

Keywords: divergence; differential entropy; nonparametric estimation; central limit theorem; convergence rates; bayes error rate

1. Introduction

Information divergences are integral functionals of two probability distributions and have many applications in the fields of information theory, statistics, signal processing, and machine learning. Some applications of divergences include estimating the decay rates of error probabilities [1], estimating bounds on the Bayes error [2–8] or the minimax error [9] for a classification problem, extending machine learning algorithms to distributional features [10–13], testing the hypothesis that two sets of samples come from the same probability distribution [14], clustering [15–17], feature selection and classification [18–20], blind source separation [21,22], image segmentation [23–25], and steganography [26]. For many more applications of divergence measures, see reference [27]. There are many information divergence families including Alpha- and Beta-divergences [28] as well as f-divergences [29,30]. In particular, the f-divergence family

includes the well-known Kullback–Leibler (KL) divergence [31], the Rényi-α divergence integral [32], the Hellinger–Bhattacharyya distance [33,34], the Chernoff-α divergence [5], the total variation distance, and the Henze–Penrose divergence [6].

Despite the many applications of divergences between continuous random variables, there are no nonparametric estimators of these functionals that achieve the parametric mean squared error (MSE) convergence rate, are simple to implement, do not require knowledge of the boundary of the density support set, and apply to a large set of divergence functionals. In this paper, we present the first information divergence estimator that achieves all of the above. Specifically, we address the problem of estimating divergence functionals when only a finite population of independent and identically distributed (i.i.d.) samples is available from the two d-dimensional distributions that are unknown, nonparametric, and smooth. Our contributions are as follows:

1. We propose the first information divergence estimator, referred to as EnDive, that is based on ensemble methods. The ensemble estimator takes a weighted average of an ensemble of weak kernel density plug-in estimators of divergence where the weights are chosen to improve the MSE convergence rate. This ensemble construction makes it very easy to implement EnDive.
2. We prove that the proposed ensemble divergence estimator achieves the optimal parametric MSE rate of $O\left(\frac{1}{N}\right)$, where N is the sample size when the densities are sufficiently smooth. In particular, EnDive achieves these rates without explicitly performing boundary correction which is required for most other estimators. Furthermore, we show that the convergence rates are uniform.
3. We prove that EnDive obeys a central limit theorem and thus, can be used to perform inference tasks on the divergence such as testing that two populations have identical distributions or constructing confidence intervals.

1.1. Related Work

Much work has focused on the problem of estimating the entropy and the information divergence of discrete random variables [1,29,35–43]. However, the estimation problem for discrete random variables differs significantly from the continuous case and thus employs different tools for both estimation and analysis.

One approach to estimating the differential entropy and information divergence of continuous random variables is to assume a parametric model for the underlying probability distributions [44–46]. However, these methods perform poorly when the parametric model does not fit the data well. Unfortunately, the structure of the underlying data distribution is unknown for many applications, and thus the chance for model misspecification is high. Thus, in many of these applications, parametric methods are insufficient, and nonparametric estimators must be used.

While several nonparametric estimators of divergence functionals between continuous random variables have been previously defined, the convergence rates are known for only a few of them. Furthermore, the asymptotic distributions of these estimators are unknown for nearly all of them. For example, Póczos and Schneider [10] established a weak consistency for a bias-corrected k- nearest neighbor (nn) estimator for Rényi-α and other divergences of a similar form where k was fixed. Li et al. [47] examined k-nn estimators of entropy and the KL divergence using hyperspherical data. Wang et al. [48] provided a k-nn based estimator for KL divergence. Plug-in histogram estimators of mutual information and divergence have been proven to be consistent [49–52]. Hero et al. [53] provided a consistent estimator for Rényi-α divergence when one of the densities is known. However none of these works studied the convergence rates or the asymptotic distribution of their estimators.

There has been recent interest in deriving convergence rates for divergence estimators for continuous data [54–60]. The rates are typically derived in terms of a smoothness condition on the densities, such as the Hölder condition [61]:

Definition 1 (Hölder Class). *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space. For $r = (r_1, \ldots, r_d)$, $r_i \in \mathbb{N}$, define $|r| = \sum_{i=1}^{d} r_i$ and $D^r = \frac{\partial^{|r|}}{\partial x_1^{r_1} \ldots \partial x_d^{r_d}}$. The Hölder class $\Sigma(s, K_H)$ of functions on $L_2(\mathcal{X})$ consists of the functions (f) that satisfy*

$$|D^r f(x) - D^r f(y)| \leq K_H \|x - y\|^{\min(s-|r|,1)},$$

for all $x, y \in \mathcal{X}$ and for all r s.t. $|r| \leq \lfloor s \rfloor$.

From Definition 1, it is clear that if a function (f) belongs to $\Sigma(s, K_H)$, then f is continuously differentiable up to order $\lfloor s \rfloor$. In this work, we show that EnDive achieves a parametric MSE convergence rate of $O(1/N)$ when $s \geq d$ and $s > \frac{d}{2}$, depending on the specific form of the divergence function.

Nguyen et al. [56] proposed an f-divergence estimator that estimates the likelihood ratio of the two densities by solving a convex optimization problem and then plugging it into the divergence formulas. The authors proved that the minimax MSE convergence rate is parametric when the likelihood ratio is a member of the bounded Hölder class $\Sigma(s, K_H)$ with $s \geq d/2$. However, this estimator is restricted to true f-divergences and may not apply to the broader class of divergence functionals that we consider here (as an example, the L_2^2 divergence is not an f-divergence). Additionally, solving the convex problem of [56] has similar computational complexity to that of training a support vector machine (SVM) (between $O(N^2)$ and $O(N^3)$), which can be demanding when N is large. In contrast, the EnDive estimator that we propose requires only the construction of simple density plug-in estimates and the solution of an offline convex optimization problem. Therefore, the most computationally demanding step in the EnDive estimator is the calculation of the density estimates, which has a computational complexity no greater than $O(N^2)$.

Singh and Póczos [58,59] provided an estimator for Rényi-α divergences as well as general density functionals that use a "mirror image" kernel density estimator. They proved that these estimators obtain an MSE convergence rate of $O\left(\frac{1}{N}\right)$ when $s \geq d$ for each of the densities. However their approach requires several computations at each boundary of the support of the densities which is difficult to implement as d gets large. Also, this computation requires knowledge of the support (specifically, the boundaries) of the densities which is unknown in most practical settings. In contrast, while our assumptions require the density support sets to be bounded and the boundaries to be smooth, knowledge of the support is not required to implement EnDive.

The "linear" and "quadratic" estimators presented by Krishnamurthy et al. [57] estimate divergence functionals that include the form $\int f_1^{\alpha}(x) f_2^{\beta}(x) d\mu(x)$ for given α and β where f_1 and f_2 are probability densities. These estimators achieve the parametric rate when $s \geq d/2$ and $s \geq d/4$ for the linear and quadratic estimators, respectively. However, the latter estimator is computationally infeasible for most functionals, and the former requires numerical integration for some divergence functionals which can be computationally difficult. Additionally, while a suitable α-β indexed sequence of divergence functionals of this form can be constructed that converge to the KL divergence, this does not guarantee convergence of the corresponding sequence of divergence estimators, as shown in reference [57]. In contrast, EnDive can be used to estimate the KL divergence directly. Other important f-divergence functionals are also excluded from this form including some that bound the Bayes error [2,4,6]. In contrast, our method applies to a large class of divergence functionals and avoids numerical integration.

Finally, Kandasamy et al. [60] proposed influence function-based estimators of distributional functionals including divergences that achieve the parametric rate when $s \geq d/2$. While this method can be applied to general functionals, the estimator requires numerical integration for some functionals. Additionally, the estimators in both Kandasamy et al. [60] and Krishnamurthy et al. [57] require an optimal kernel density estimator. This is difficult to construct when the density support is bounded as it requires difficult computations at the density support set boundary and therefore, knowledge of the density support set. In contrast, Endive does not require knowledge of the support boundary.

In addition to the MSE convergence rates, the asymptotic distribution of divergence estimators is of interest. Asymptotic normality has been established for certain divergences between a specific density estimator and the true density [62–64]. This differs from the problem we consider where we assume that both densities are unknown. The asymptotic distributions of the estimators in references[56–59] are currently unknown. Thus, it is difficult to use these estimators for hypothesis testing which is crucial in many scientific applications. Kandasamy et al. [60] derived the asymptotic distribution of their data-splitting estimator but did not prove similar results for their leave-one-out estimator. We establish a central limit theorem for EnDive which greatly enhances its applicability in scientific settings.

Our ensemble divergence estimator reduces to an ensemble entropy estimator as a special case when data from only one distribution is considered and the other density is set to a uniform measure (see reference[28] for more on the relationship between entropy and information divergence). The resultant entropy estimator differs from the ensemble entropy estimator proposed by Sricharan et al. [65] in several important ways. First, the density support set must be known for the estimator in reference [65] to perform the explicit boundary correction. In contrast, the EnDive estimator does not require any boundary correction. To show this requires a significantly different approach to prove the bias and variance rates of the EnDive estimator. Furthermore, the EnDive results apply under more general assumptions for the densities and the kernel used in the weak estimators. Finally, the central limit theorem applies to the EnDive estimator which is currently unknown for the estimator in reference [65].

We also note that Berrett et al. [66] proposed a modification of the Kozachenko and Leonenko estimator of entropy [67] that takes a weighted ensemble estimation approach. While their results require stronger assumptions for the smoothness of the densities than ours do, they did obtain the asymptotic distribution of their weighted estimator and they also showed that the asymptotic variance of the estimator is not increased by taking a weighted average. This latter point is an important selling point of the ensemble framework—we can improve the asymptotic bias of an estimator without increasing the asymptotic variance.

1.2. Organization and Notation

The paper is organized as follows. We first derive the MSE convergence rates in Section 2 for a weak divergence estimator, which is a kernel density plug-in divergence estimator. We then generalize the theory of optimally weighted ensemble entropy estimation developed in reference [65] to obtain the ensemble divergence estimator EnDive from an ensemble of weak estimators in Section 3. A central limit theorem and uniform convergence rate for the ensemble estimator are also presented in Section 3. In Section 4, we provide guidelines for selecting the tuning parameters based on experiments and the theory derived in the previous sections. We then perform experiments in Section 4 that validate the theory and establish the robustness of the proposed estimators to the tuning parameters.

Bold face type is used for random variables and random vectors. The conditional expectation given a random variable $\mathbf{Z}$ is denoted as $\mathbb{E}_{\mathbf{Z}}$. The variance of a random variable is denoted as $\mathbb{V}$, and the bias of an estimator is denoted as $\mathbb{B}$.

2. The Divergence Functional Weak Estimator

This paper focuses on estimating functionals of the form

$$G\left(f_1, f_2\right) = \int g\left(f_1(x), f_2(x)\right) f_2(x) dx,, \tag{1}$$

where $g(x, y)$ is a smooth functional, and f_1 and f_2 are smooth d-dimensional probability densities. If $g\left(f_1(x), f_2(x)\right) = g\left(\frac{f_1(x)}{f_2(x)}\right)$, g is convex, and $g(1) = 0$, then $G\left(f_1, f_2\right)$ defines the family of f-divergences. Some common divergences that belong to this family include the KL divergence ($g(t) = -\ln t$) and the total variation distance ($g(t) = |t-1|$). In this work, we consider a broader class of functionals than the f-divergences, since g is allowed to be very general.

To estimate $G(f_1, f_2)$, we first define a weak plug-in estimator based on kernel density estimators (KDEs), that is, a simple estimator that converges slowly to the true value $G(f_1, f_2)$ in terms of MSE. We then derive the bias and variance expressions for this weak estimator as a function of sample size and bandwidth. We then use the resulting bias and variance expressions to derive an ensemble estimator that takes a weighted average of weak estimators with different bandwidths and achieves superior MSE performance.

2.1. The Kernel Density Plug-in Estimator

We use a kernel density plug-in estimator of the divergence functional in (1) as the weak estimator. Assume that N_1 i.i.d. realizations $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_1}\}$ are available from f_1 and N_2 i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_{N_2}\}$ are available from f_2. Let $h_i > 0$ be the kernel bandwidth for the density estimator of f_i. For simplicity of presentation, assume that $N_1 = N_2 = N$ and $h_1 = h_2 = h$. The results for the more general case of differing sample sizes and bandwidths are given in Appendix C. Let $K(\cdot)$ be a kernel function with $\int K(x)dx = 1$ and $||K||_\infty < \infty$ where $\|K\|_\infty$ is the ℓ_∞ norm of the kernel (K). The KDEs for f_1 and f_2 are, respectively,

$$\begin{aligned}\tilde{\mathbf{f}}_{1,h}(\mathbf{X}_j) &= \frac{1}{Nh^d}\sum_{i=1}^{N} K\left(\frac{\mathbf{X}_j - \mathbf{Y}_i}{h}\right),\\ \tilde{\mathbf{f}}_{2,h}(\mathbf{X}_j) &= \frac{1}{Mh^d}\sum_{\substack{i=1\\ i\neq j}}^{N} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right),\end{aligned}$$

where $M = N - 1$. $G(f_1, f_2)$ is then approximated as

$$\tilde{\mathbf{G}}_h = \frac{1}{N}\sum_{i=1}^{N} g\left(\tilde{\mathbf{f}}_{1,h}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h}(\mathbf{X}_i)\right). \tag{2}$$

2.2. Convergence Rates

For many estimators, MSE convergence rates are typically provided in the form of upper (or sometimes lower) bounds on the bias and the variance. Therefore, only the slowest converging terms (as a function of the sample size (N)) are presented in these cases. However, to apply our generalized ensemble theory to obtain estimators that guarantee the parametric MSE rate, we required explicit expressions for the bias of the weak estimators in terms of the sample size (N) and the kernel bandwidth (h). Thus, an upper bound was insufficient for our work. Furthermore, to guarantee the parametric rate, we required explicit expressions of all bias terms that converge to zero slower than $O\left(1/\sqrt{N}\right)$.

To obtain bias expressions, we required multiple assumptions on the densities f_1 and f_2, the functional g, and the kernel K. Similar to reference [7,54,65], the principal assumptions we make were that (1) f_1, f_2, and g are smooth; (2) f_1 and f_2 have common bounded support sets $\mathcal{S}$; and (3) f_1 and f_2 are strictly lower bounded on $\mathcal{S}$. We also assume (4) that the density support set is smooth with respect to the kernel ($K(u)$). The full technical assumptions and a discussion of them are contained in Appendix A. Given these assumptions, we have the following result on the bias of $\tilde{\mathbf{G}}_h$:

Theorem 1. *For a general g, the bias of the plug-in estimator $\tilde{\mathbf{G}}_h$ is given by*

$$\mathbb{B}\left[\tilde{\mathbf{G}}_h\right] = \sum_{j=1}^{\lfloor s \rfloor} c_{10,j}h^j + c_{11}\frac{1}{Nh^d} + O\left(h^s + \frac{1}{Nh^d}\right). \tag{3}$$

To apply our generalized ensemble theory to the KDE plug-in estimator ($\tilde{\mathbf{G}}_h$), we required only an upper bound on its variance. The following variance result required much less strict assumptions than the bias results in Theorem 1:

Theorem 2. *Assume that the functional g in (1) is Lipschitz continuous in both of its arguments with the Lipschitz constant (C_g). Then, the variance of the plug-in estimator ($\tilde{\mathbf{G}}_h$) is bounded by*

$$\mathbb{V}\left[\tilde{\mathbf{G}}_h\right] \leq C_g^2 ||K||_\infty^2 \frac{11}{N}.$$

From Theorems 1 and 2, we observe that $h \to 0$ and $Nh^d \to \infty$ are required for $\tilde{\mathbf{G}}_h$ to be unbiased, while the variance of the plug-in estimator depends primarily on the sample size (N). Note that the constants depend on the densities f_1 and f_2 and their derivatives which are often unknown.

2.3. Optimal MSE Rate

From Theorem 1, the dominating terms in the bias are observed to be $\Theta(h)$ and $\Theta\left(\frac{1}{Nh^d}\right)$. If no bias correction is performed, the optimal choice of h that minimizes MSE is

$$h^* = \Theta\left(N^{\frac{-1}{d+1}}\right).$$

This results in a dominant bias term of order $\Theta\left(N^{\frac{-1}{d+1}}\right)$. Note that this differs from the standard result for the optimal KDE bandwidth for minimum MSE density estimation which is $\Theta\left(N^{-1/(d+4)}\right)$ for a symmetric uniform kernel when the boundary bias is ignored [68].

Figure 1 shows a heatmap showing the leading bias term $O(h)$ as a function of d and N when $h = N^{\frac{-1}{d+1}}$. The heatmap indicates that the bias of the plug-in estimator in (2) is small only for relatively small values of d. This is consistent with the empirical results in reference [69] which examined the MSE of multiple plug-in KDE and k-nn estimators. In the next section, we propose an ensemble estimator that achieves a superior convergence rate regardless of the dimensions (d) as long as the density is sufficiently smooth.

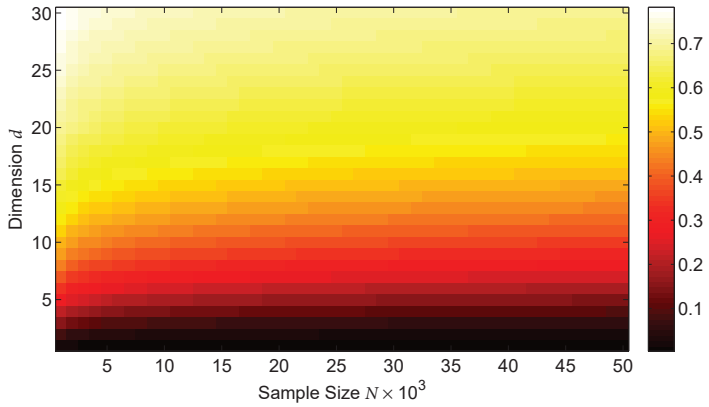

Figure 1. Heat map showing the predicted bias of the divergence functional plug-in estimator $\tilde{\mathbf{G}}_h$ based on Theorem 1 as a function of the dimensions (d) and sample size (N) when $h = N^{\frac{-1}{d+1}}$. Note that the phase transition in the bias as the dimensions (d) increase for a fixed sample size (N); the bias remains small only for relatively small values of d. The proposed weighted ensemble estimator EnDive eliminates this phase transition when the densities and the function g are sufficiently smooth.

2.4. Proof Sketches of Theorems 1 and 2

To prove the bias expressions in Theorem 1, the bias is first decomposed into two parts by adding and subtracting $g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h}(\mathbf{Z})\right)$ within the expectation creating a "bias" term and a "variance" term. Applying a Taylor series expansion on the bias and variance terms results in expressions that depend on powers of $\mathbb{B}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h}(\mathbf{Z})\right] := \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h}(\mathbf{Z}) - f_i(\mathbf{Z})$ and $\tilde{\mathbf{e}}_{i,h}(\mathbf{Z}) := \tilde{\mathbf{f}}_{i,h}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h}(\mathbf{Z})$, respectively. Within the interior of the support, moment bounds can be derived from properties of the KDEs and a Taylor series expansion of the densities. Near the boundary of the support, the smoothness assumption on the boundary $\mathcal{A}.5$ is required to obtain an expression of the bias in terms of the KDE bandwidth (h) and the sample size (N). The full proof of Theorem 1 is given in Appendix E.

The proof of the variance result takes a different approach. The proof uses the Efron–Stein inequality [70] which bounds the variance by analyzing the expected squared difference between the plug-in estimator when one sample is allowed to differ. This approach provides a bound on the variance under much less strict assumptions on the densities and the functional g than is required for Theorem 1. The full proof of Theorem 2 is given in Appendix F.

3. Weighted Ensemble Estimation

From Theorem 1 and Figure 1, we can observe that the bias of the MSE-optimal plug-in estimator $\tilde{\mathbf{G}}_h$ decreases very slowly as a function of the sample size (N) when the data dimensions (d) are not small, resulting in a large MSE. However, by applying the theory of optimally weighted ensemble estimation, we can obtain an estimator with improved performance by taking a weighted sum of an ensemble of weak estimators where the weights are chosen to significantly reduce the bias.

The ensemble of weak estimators is formed by choosing different values of the bandwidth parameter h as follows. Set $\mathcal{L} = \{l_1, \ldots, l_L\}$ to be real positive numbers that index $h(l_i)$. Thus, the parameter l indexes over different neighborhood sizes for the KDEs. Define the weight $w := \{w(l_1), \ldots, w(l_L)\}$ and $\tilde{\mathbf{G}}_w := \sum_{l\in\mathcal{L}} w(l)\tilde{\mathbf{G}}_{h(l)}$. That is, for each estimator $\tilde{\mathbf{G}}_{h(l)}$ there is a corresponding weight value ($w(l)$). The key to reducing the MSE is to choose the weight vector (w) to reduce the lower order terms in the bias while minimizing the impact of the weighted average on the variance.

3.1. Finding the Optimal Weight

The theory of optimally weighted ensemble estimation is a general theory that is applicable to any estimation problem as long as the bias and variance of the estimator can be expressed in a specific way. An early version of this theory was presented in reference [65]. We now generalize this theory so that it can be applied to a wider variety of estimation problems. Let N be the number of available samples and let $\mathcal{L} = \{l_1, \ldots, l_L\}$ be a set of index values. Given an indexed ensemble of estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l\in\mathcal{L}}$ of some parameter (E), the weighted ensemble estimator with weights $w = \{w(l_1), \ldots, w(l_L)\}$ satisfying $\sum_{l\in\mathcal{L}} w(l) = 1$ is defined as

$$\hat{\mathbf{E}}_w = \sum_{l\in\mathcal{L}} w(l)\,\hat{\mathbf{E}}_l.$$

$\hat{\mathbf{E}}_w$ is asymptotically unbiased as long as the estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l\in\mathcal{L}}$ are asymptotically unbiased. Consider the following conditions on $\left\{\hat{\mathbf{E}}_l\right\}_{l\in\mathcal{L}}$:

- $\mathcal{C}.1$ The bias is expressible as

$$\mathbb{B}\left[\hat{\mathbf{E}}_l\right] = \sum_{i\in J} c_i\psi_i(l)\phi_{i,d}(N) + O\left(\frac{1}{\sqrt{N}}\right),$$

  where c_i are constants that depend on the underlying density and are independent of N and l, $J = \{i_1, \ldots, i_I\}$ is a finite index set with $I < L$, and $\psi_i(l)$ are basis functions depending only on parameter l and not on the sample size (N).

- $\mathcal{C}.2$ The variance is expressible as

$$\mathbb{V}\left[\hat{\mathbf{E}}_l\right] = c_v\left(\frac{1}{N}\right) + o\left(\frac{1}{N}\right).$$

Theorem 3. *Assume conditions $\mathcal{C}.1$ and $\mathcal{C}.2$ hold for an ensemble of estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l\in\mathcal{L}}$. Then, there exists a weight vector (w_0) such that the MSE of the weighted ensemble estimator attains the parametric rate of convergence:*

$$\mathbb{E}\left[\left(\hat{\mathbf{E}}_{w_0} - E\right)^2\right] = O\left(\frac{1}{N}\right).$$

The weight vector (w_0) is the solution to the following convex optimization problem:

$$\begin{array}{rl} \min_w & ||w||_2 \\ \textit{subject to} & \sum_{l\in\mathcal{L}} w(l) = 1, \\ & \gamma_w(i) = \sum_{l\in\mathcal{L}} w(l)\psi_i(l) = 0,\ i \in J. \end{array} \tag{4}$$

Proof. From condition $\mathcal{C}.1$, we can write the bias of the weighted estimator as

$$\mathbb{B}\left[\hat{\mathbf{E}}_w\right] = \sum_{i\in J} c_i\gamma_w(i)\phi_{i,d}(N) + O\left(\frac{\sqrt{L}||w||_2}{\sqrt{N}}\right).$$

The variance of the weighted estimator is bounded as

$$\mathbb{V}\left[\hat{\mathbf{E}}_w\right] \leq \frac{L||w||_2^2}{N}. \tag{5}$$

The optimization problem in (4) zeroes out the lower-order bias terms and limits the ℓ_2 norm of the weight vector (w) to prevent the variance from exploding. This results in an MSE rate of $O(1/N)$ when the dimensions (d) are fixed and when L is fixed independently of the sample size (N). Furthermore, a solution to (4) is guaranteed to exist if $L > I$ and the vectors $a_i = [\psi_i(l_1), \ldots, \psi_i(l_L)]$ are linearly independent. This completes our sketch of the proof of Theorem 3. □

3.2. The EnDive Estimator

The parametric rate of $O\left(1/N\right)$ in MSE convergence can be achieved without requiring $\gamma_w(i) = 0,\ i \in J$. This can be accomplished by solving the following convex optimization problem in place of the optimization problem in Theorem 3:

$$\begin{array}{rl} \min_w & \epsilon \\ \textit{subject to} & \sum_{l\in\mathcal{L}} w(l) = 1, \\ & \left|\gamma_w(i)N^{\frac{1}{2}}\phi_{i,d}(N)\right| \leq \epsilon,\ i \in J, \\ & \|w\|_2^2 \leq \eta\epsilon, \end{array} \tag{6}$$

where the parameter η is chosen to achieve a trade-off between bias and variance. Instead of forcing $\gamma_w(i) = 0$, the relaxed optimization problem uses the weights to decrease the bias terms at a rate of $O\left(1/\sqrt{N}\right)$, yielding an MSE convergence rate of $O(1/N)$. In fact, it was shown in reference [71] that the optimization problem in (6) guarantees the parametric MSE rate as long as the conditions of Theorem 3 are satisfied and a solution to the optimization problem in (4) exists (the conditions for this existence are given in the proof of Theorem 3).

We now construct a divergence ensemble estimator from an ensemble of plug-in KDE divergence estimators. Consider first the bias result in (3) where g is general, and assume that $s \geq d$. In this case, the bias contains a $O\left(\frac{1}{h^d N}\right)$ term. To guarantee the parametric MSE rate, any remaining lower-order bias

terms in the ensemble estimator must be no slower than $O\left(1/\sqrt{N}\right)$. Let $h(l) = lN^{-1/(2d)}$ where $l \in \mathcal{L}$. Then $O\left(\frac{1}{h(l)^d N}\right) = O\left(\frac{1}{l^d \sqrt{N}}\right)$. We therefore obtain an ensemble of plug-in estimators $\left\{\tilde{\mathbf{G}}_{h(l)}\right\}_{l\in\mathcal{L}}$ and a weighted ensemble estimator $\tilde{\mathbf{G}}_w = \sum_{l\in\mathcal{L}} w(l)\tilde{\mathbf{G}}_{h(l)}$. The bias of each estimator in the ensemble satisfies the condition $\mathcal{C}.1$ with $\psi_i(l) = l^i$ and $\phi_{i,d}(N) = N^{-i/(2d)}$ for $i = 1, \ldots, d$. To obtain a uniform bound on the bias with respect to w and $\mathcal{L}$, we also include the function $\psi_{d+1}(l) = l^{-d}$ with corresponding $\phi_{d+1,d}(N) = N^{-1/2}$. The variance also satisfies the condition $\mathcal{C}.2$. The optimal weight (w_0) is found by using (6) to obtain an optimally weighted plug-in divergence functional estimator $\tilde{\mathbf{G}}_{w_0}$ with an MSE convergence rate of $O\left(\frac{1}{N}\right)$ as long as $s \geq d$ and $L \geq d$. Otherwise, if $s < d$, we can only guarantee the MSE rate up to $O\left(\frac{1}{N^{s/d}}\right)$. We refer to this estimator as the Ensemble Divergence (EnDive) estimator and denote it as $\tilde{\mathbf{G}}_{\text{EnDive}}$.

We note that for some functionals (g) (including the KL divergence and the Renyi-α divergence integral), we can modify the EnDive estimator to obtain the parametric rate under the less strict assumption that $s > d/2$. For details on this approach, see Appendix B.

3.3. Central Limit Theorem

The following theorem shows that an appropriately normalized ensemble estimator $\tilde{\mathbf{G}}_w$ converges in distribution to a normal random variable under rather general conditions. Thus, the same result applies to the EnDive estimator $\tilde{\mathbf{G}}_{\text{EnDive}}$. This enables us to perform hypothesis testing on the divergence functional which is very useful in many scientific applications. The proof is based on the Efron–Stein inequality and an application of Slutsky's Theorem (Appendix G).

Theorem 4. *Assume that the functional g is Lipschitz in both arguments with the Lipschitz constant C_g. Further assume that $h(l) = o(1)$, $N \to \infty$, and $Nh(l)^d \to \infty$ for each $l \in \mathcal{L}$. Then, for a fixed $\mathcal{L}$, the asymptotic distribution of the weighted ensemble estimator $\tilde{\mathbf{G}}_w$ is*

$$Pr\left(\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right) / \sqrt{\mathbb{V}\left[\tilde{\mathbf{G}}_w\right]} \leq t\right) \to Pr(\mathbf{S} \leq t),$$

where $\mathbf{S}$ is a standard normal random variable.

3.4. Uniform Convergence Rates

Here, we show that the optimally weighted ensemble estimators achieve the parametric MSE convergence rate uniformly. Denote the subset of $\Sigma(s, K_H)$ with densities bounded between ϵ_0 and ϵ_∞ as $\Sigma(s, K_H, \epsilon_0, \epsilon_\infty)$.

Theorem 5. *Let $\tilde{\mathbf{G}}_{EnDive}$ be the EnDive estimator of the functional*

$$G(p, q) = \int g\left(p(x), q(x)\right) q(x) dx,$$

where p and q are d-dimensional probability densities. Additionally, let $r = d$ and assume that $s > r$. Then,

$$\sup_{p,q\in\Sigma(s,K_H,\epsilon_0,\epsilon_\infty)} \mathbb{E}\left[\left(\tilde{\mathbf{G}}_{w_0} - G(p, q)\right)^2\right] \leq \frac{C}{N}, \tag{7}$$

where C is a constant.

The proof decomposes the MSE into the variance plus the square of the bias. The variance is bounded easily by using Theorem 2. To bound the bias, we show that the constants in the bias terms are continuous with respect to the densities p and q under an appropriate norm. We then show

that $\Sigma\left(s, K_H, \epsilon_0, \epsilon_\infty\right)$ is compact with respect to this norm and then apply an extreme value theorem. Details are given in Appendix H.

4. Experimental Results

In this section, we discuss the choice of tuning parameters and validate the EnDive estimator's convergence rates and the central limit theorem. We then use the EnDive estimator to estimate bounds on the Bayes error for a single-cell bone marrow data classification problem.

4.1. Tuning Parameter Selection

The optimization problem in (6) has parameters η, L, and $\mathcal{L}$. By applying (6), and the resulting MSE of the ensemble estimator is

$$O\left(\epsilon^2/N\right) + O\left(L\eta^2\epsilon^2/N\right), \tag{8}$$

where each term in the sum comes from the bias and variance, respectively. From this expression and (6), we see that the parameter η provides a tradeoff between bias and variance. Increasing η enables the norm of the weight vector to be larger. This means the feasible region for the variable w increases in size as η increases which can result in decreased bias. However, as η contributes to the variance term, increasing η may result in increased variance.

If all of the constants in (3) and an exact expression for the variance of the ensemble estimator were known, then η could be chosen to optimize this tradeoff in bias and variance and thus minimize the MSE. Since these constants are unknown, we can only choose η based on the asymptotic results. From (8), this would suggest setting $\eta = 1/\sqrt{L}$. In practice, we find that for finite sample sizes, the variance in the ensemble estimator is less than the upper bound of $L\eta^2\epsilon^2/N$. Thus, setting $\eta = 1/\sqrt{L}$ is unnecessarily restrictive. We find that, in practice, setting $\eta = 1$ works well.

Upon first glance, it appears that for fixed L, the set $\mathcal{L}$ that parameterizes the kernel widths can, in theory, be chosen by minimizing ϵ in (6) over $\mathcal{L}$ in addition to w. However, adding this constraint results in a non-convex optimization problem since w does not lie in the non-negative orthant. A parameter search over possible values for $\mathcal{L}$ is another possibility. However, this may not be practical as ϵ generally decreases as the size and spread of $\mathcal{L}$ increases. In addition, for finite sample sizes, decreasing ϵ does not always directly correspond to a decrease in MSE, as very high or very low values of $h(l)$ can lead to inaccurate density estimates, resulting in a larger MSE.

Given these limitations, we provide the following recommendations for $\mathcal{L}$. Denote the value of the minimum value of l such that $\tilde{\mathbf{f}}_{i,h(l_{min})}(\mathbf{X}_j) > 0 \; \forall i = 1, 2$ as l_{min} and the diameter of the support $\mathcal{S}$ as D. To ensure the KDEs are bounded away from zero, we require that $\min(\mathcal{L}) \geq l_{min}$. As demonstrated in Figure 2, the weights in w_0 are generally largest for the smallest values of $\mathcal{L}$. This indicates that $\min(\mathcal{L})$ should also be sufficiently larger than l_{min} to render an adequate density estimate. Similarly, $\max(\mathcal{L})$ should be sufficiently smaller than the diamter (D) as high bandwidth values can lead to high bias in the KDEs. Once these values are chosen, all other $\mathcal{L}$ values can then be chosen to be equally spaced between $\min(\mathcal{L})$ and $\max(\mathcal{L})$.

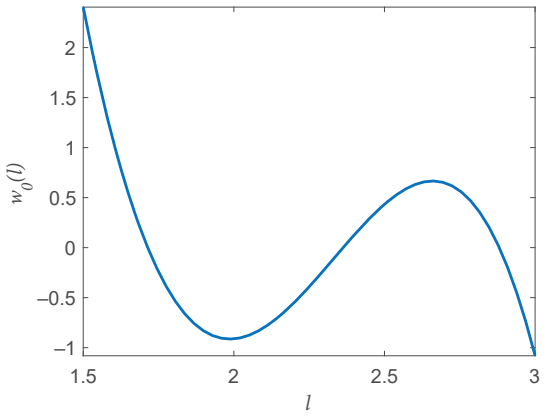

Figure 2. The optimal weights from (6) when $d = 4$, $N = 3100$, $L = 50$, and l are uniformly spaced between 1.5 and 3. The lowest values of l are given the highest weight. Thus, the minimum value of bandwidth parameters $\mathcal{L}$ should be sufficiently large to render an adequate estimate.

An efficient way to choose l_{min} and l_{max} is to select the integers k_{min} and k_{max} and compute the k_{min} and k_{max} nearest neighbor distances of all the data points. The bandwidths $h(l_{min})$ and $h(l_{max})$ can then be chosen to be the maximums of these corresponding distances. The parameters l_{min} and l_{max} can then be computed from the expression $h(l) = lN^{-\frac{1}{2d}}$. This choice ensures that a minimum of k_{min} points are within the kernel bandwidth for the density estimates at all points and that a maximum of k_{max} points are within the kernel bandwidth for the density estimates at one of the points.

Once $\min(\mathcal{L})$ and $\max(\mathcal{L})$ have been chosen, the similarity of bandwidth values $h(l)$ and basis functions $\psi_{i,d}(l)$ increases as L increases, resulting in a negligible decrease in the bias. Hence, L should be chosen to be large enough for sufficient bias but small enough so that the bandwidth values $h(l)$ are sufficiently distinct. In our experiments, we found $30 \leq L \leq 60$ to be sufficient.

4.2. Convergence Rates Validation: Rényi-α Divergence

To validate our theoretical convergence rate results, we estimated the Rényi-α divergence integral between two truncated multivariate Gaussian distributions with varying dimension and sample sizes. The densities had means of $\bar{\mu}_1 = 0.7 \times \bar{1}_d$, $\bar{\mu}_2 = 0.3 \times \bar{1}_d$ and covariance matrices of $0.4 \times I_d$, where $\bar{1}_d$ is a d-dimensional vector of ones, and I_d is a $d \times d$ identity matrix. We restricted the Gaussians to the unit cube and used $\alpha = 0.5$.

The left plots in Figure 3 show the MSE (200 trials) of the standard plug-in estimator implemented with a uniform kernel and the proposed optimally weighted estimator EnDive for various dimensions and sample sizes. The parameter set $\mathcal{L}$ was selected based on a range of k-nearest neighbor distances. The bandwidth used for the standard plug-in estimator was selected by setting $h_{fixed}(l^*) = l^* N^{-\frac{1}{d+1}}$, where l^* was chosen from $\mathcal{L}$ to minimize the MSE of the plug-in estimator. For all dimensions and sample sizes, EnDive outperformed the plug-in estimator in terms of MSE. EnDive was also less biased than the plug-in estimator and even had lower variance at smaller sample sizes (e.g., $N = 100$). This reflects the strength of ensemble estimators—the weighted sum of a set of relatively poor estimators can result in a very good estimator. Note also that for the larger values of N, the ensemble estimator MSE rates approached the theoretical rate based on the estimated log–log slope given in Table 1.

Table 1. Negative log–log slope of the EnDive mean squared error (MSE) as a function of the sample size for various dimensions. The slope was calculated beginning at N_{start}. The negative slope was closer to 1 with $N_{start} = 10^{2.375}$ than for $N_{start} = 10^2$ indicating that the asymptotic rate had not yet taken effect at $N_{start} = 10^2$.

Estimator	$d = 5$	$d = 10$	$d = 15$
$N_{start} = 10^2$	0.85	0.84	0.80
$N_{start} = 10^{2.375}$	0.96	0.96	0.95

To illustrate the difference between the problems of density estimation and divergence functional estimation, we estimated the average pointwise squared error between the KDE $\tilde{\mathbf{f}}_{1,h}$ and f_1 in the previous experiment. We used exactly the same bandwidth and kernel as the standard plug-in estimators in Figure 3 and calculated the pointwise error at 10,000 points sampled from f_1. The results are shown in Figure 4. From these results, we see that the KDEs performed worse as the dimension of the densities increased. Additionally, we observe by comparing Figures 3 and 4, the average pointwise squared error decreased at a much slower rate as a function of the sample size (N) than the MSE of the plug-in divergence estimators, especially for larger dimensions (d).

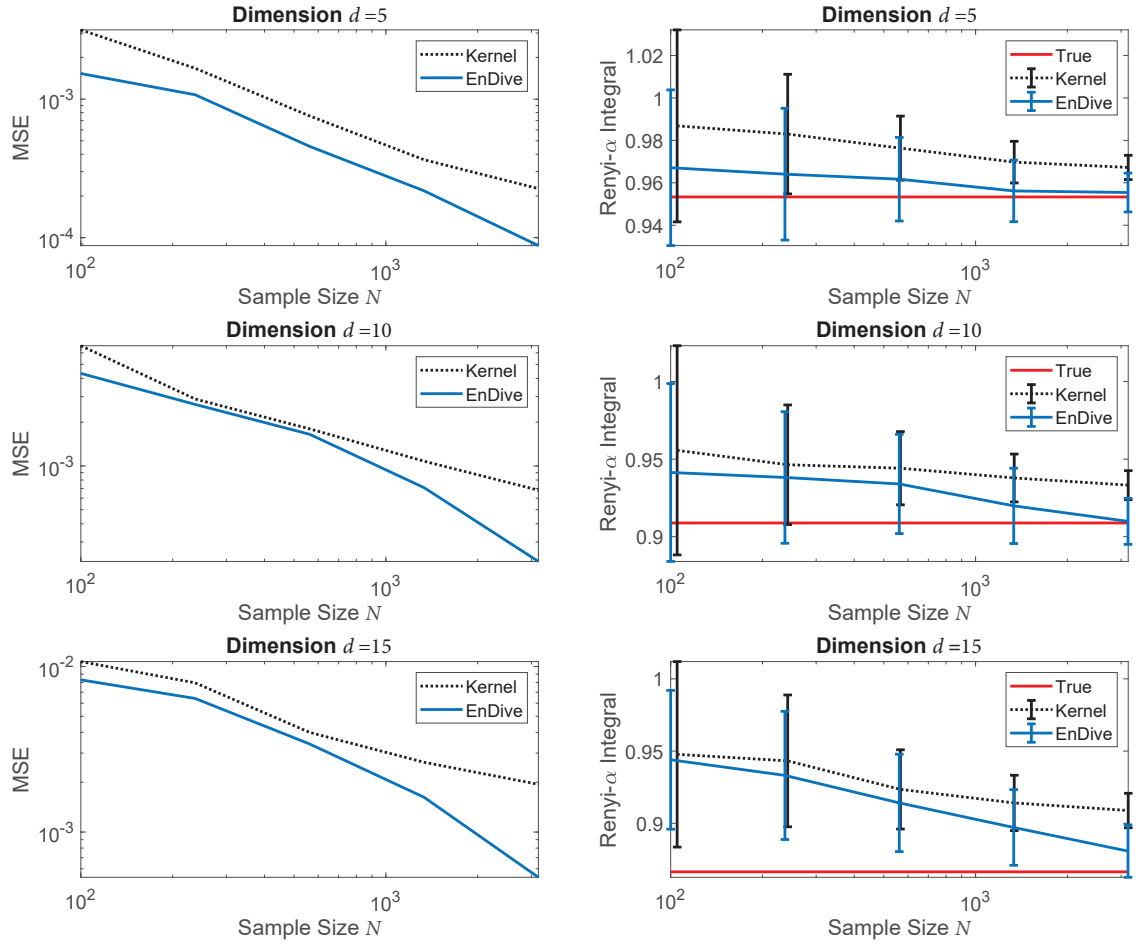

Figure 3. (**Left**) Log–log plot of MSE of the uniform kernel plug-in ("Kernel") and the optimally weighted EnDive estimator for various dimensions and sample sizes. (**Right**) Plot of the true values being estimated compared to the average values of the same estimators with standard error bars. The proposed weighted ensemble estimator approaches the theoretical rate (see Table 1), performed better than the plug-in estimator in terms of MSE and was less biased.

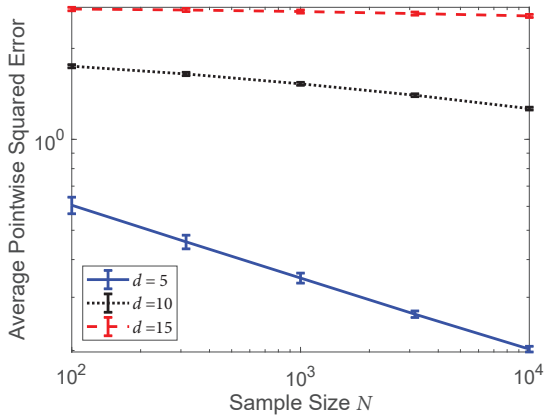

Figure 4. Log–log plot of the average pointwise squared error between the KDE $\tilde{\mathbf{f}}_{1,h}$ and f_1 for various dimensions and sample sizes using the same bandwidth and kernel as the standard plug-in estimators in Figure 3. The KDE and the density were compared at 10,000 points sampled from f_1.

Our experiments indicated that the proposed ensemble estimator is not sensitive to the tuning parameters. See reference [72] for more details.

4.3. Central Limit Theorem Validation: KL Divergence

To verify the central limit theorem of the EnDive estimator, we estimated the KL divergence between two truncated Gaussian densities, again restricted to the unit cube. We conducted two experiments where (1) the densities were different with means of $\bar{\mu}_1 = 0.7 \times \bar{1}_d$, $\bar{\mu}_2 = 0.3 \times \bar{1}_d$ and covariances of matrices $\sigma_i \times I_d$, $\sigma_1 = 0.1$, $\sigma_2 = 0.3$; and where (2) the densities were the same with means of $0.3 \times \bar{1}_d$ and covariance matrices of $0.3 \times I_d$. For both experiments, we chose $d = 6$ and four different sample sizes (N). We found that the correspondence between the quantiles of the standard normal distribution and the quantiles of the centered and scaled EnDive estimator was very high under all settings (see Table 2 and Figure 5) which validates Theorem 4.

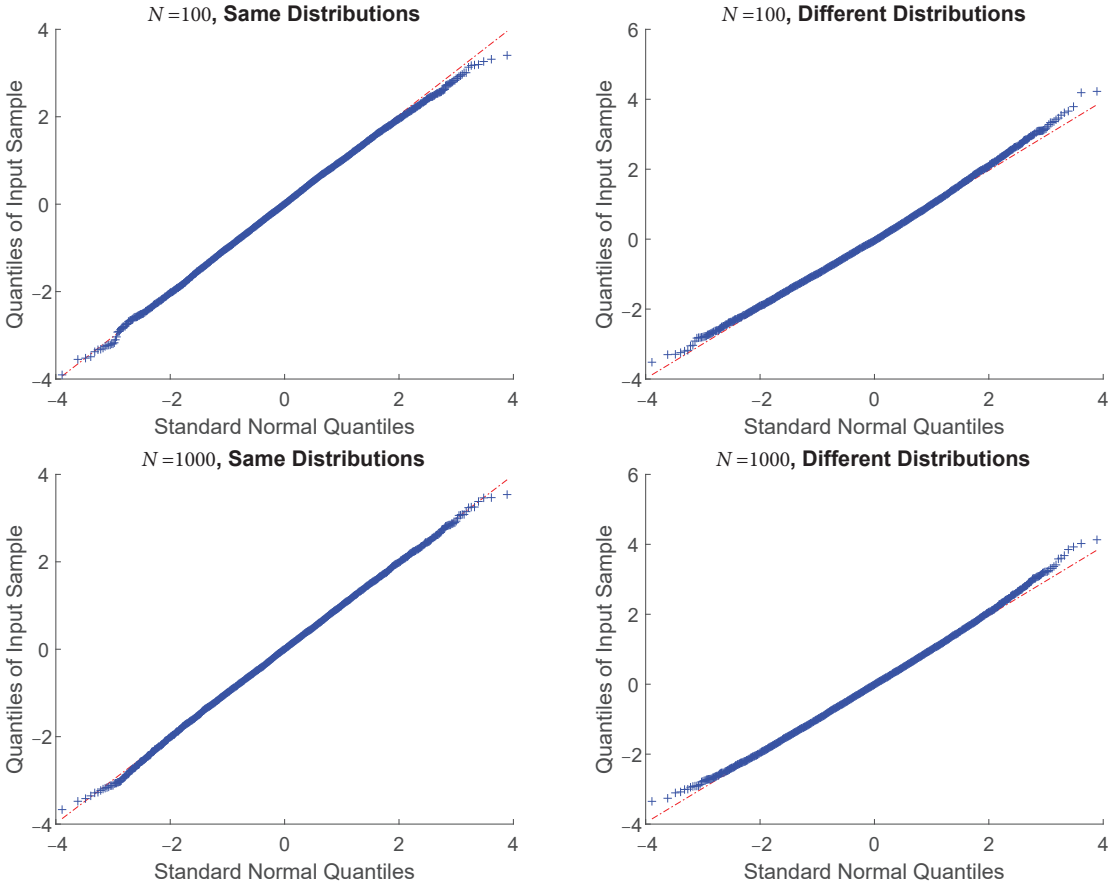

Figure 5. QQ-plots comparing the quantiles of a standard normal random variable and the quantiles of the centered and scaled EnDive estimator applied to the Kullback–Leibler (KL) divergence when the distributions were the same and different. Quantiles were computed from 10,000 trials. These plots correspond to the same experiments as in Table 2 when N = 100 and N = 1000. The correspondence between quantiles is high for all cases.

Table 2. Comparison between quantiles of a standard normal random variable and the quantiles of the centered and scaled EnDive estimator applied to the KL divergence when the distributions were the same and different. Quantiles were computed from 10,000 trials. The parameter ρ gives the correlation coefficient between the quantiles, while β is the estimated slope between the quantiles. The correspondence between quantiles was very high for all cases.

N	Same		Different	
	$1-\rho$	β	$1-\rho$	β
100	2.35×10^{-4}	1.014	9.97×10^{-4}	0.993
500	9.48×10^{-5}	1.007	5.06×10^{-4}	0.999
1000	8.27×10^{-5}	0.996	4.30×10^{-4}	0.988
5000	8.59×10^{-5}	0.995	4.47×10^{-4}	1.005

4.4. Bayes Error Rate Estimation on Single-Cell Data

Using the EnDive estimator, we estimated bounds on the Bayes error rate (BER) of a classification problem involving MARS-seq single-cell RNA-sequencing (scRNA-seq) data measured from developing mouse bone marrow cells enriched for the myeloid and erythroid lineages [73]. However, we first demonstrated the ability of EnDive to estimate the bounds on the BER of a simulated problem. In this simulation, the data were drawn from two classes where each class distribution was a $d = 10$ dimensional

Gaussian distribution with different means and the identity covariance matrix. We considered two cases, namely, the distance between the means was 1 or 3. The BER was calculated in both cases. We then estimated upper and lower bounds on the BER by estimating the Henze–Penrose (HP) divergence [4,6]. Figure 6 shows the average estimated upper and lower bounds on the BER with standard error bars for both cases. For all tested sample sizes, the BER was within one standard deviation of the estimated lower bound. The lower bound was also closer, on average, to the BER for most of the tested sample sizes (lower sample sizes with smaller distances between means were the exceptions). Generally, these resuls indicate that the true BER is relatively close to the estimated lower bound, on average.

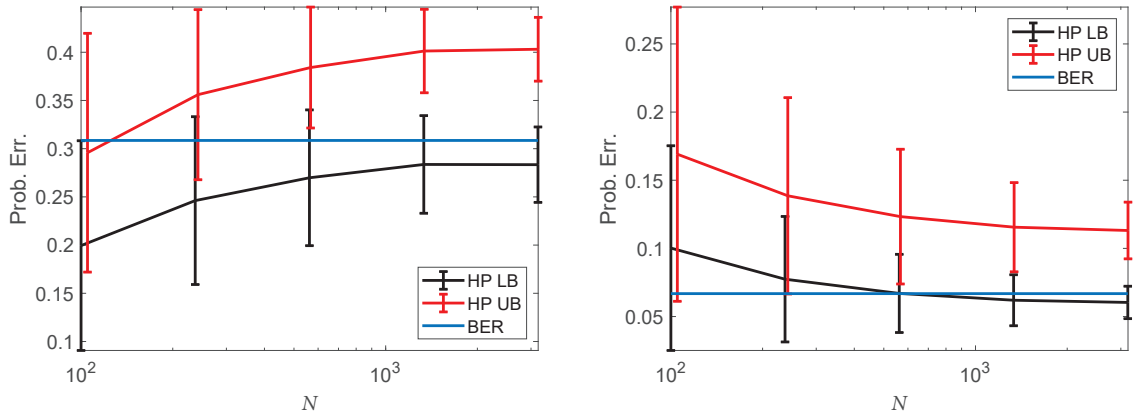

Figure 6. Estimated upper (UB) and lower bounds (LB) on the Bayes error rate (BER) based on estimating the HP divergence between two 10-dimensional Gaussian distributions with identity covariance matrices and distances between means of 1 (**left**) and 3 (**right**), respectively. Estimates were calculated using EnDive, with error bars indicating the standard deviation from 400 trials. The upper bound was closer, on average, to the true BER when N was small ($\approx$100–300) and the distance between the means was small. The lower bound was closer, on average, in all other cases.

We then estimated similar bounds on the scRNA-seq classification problem using EnDive. We considered the three most common cell types within the data: erythrocytes (eryth.), monocytes (mono.), and basophils (baso.) ($N = 1095, 559, 300$, respectively). We estimated the upper and lower bounds on the pairwise BER between these classes using different combinations of genes selected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with the hematopoietic cell lineage [74–76]. Each collection of genes contained 11–14 genes. The upper and lower bounds on the BER were estimated using the Henze–Penrose divergence [4,6]. The standard deviations of the bounds for the KEGG-based genes were estimated via 1000 bootstrap iterations. The KEGG-based bounds were compared to BER bounds obtained from 1000 random selections of 12 genes. In all cases, we compared the bounds to the performance of a quadratic discriminant analysis classifier (QDA) with 10-fold cross validation. Note that to correct for undersampling in scRNA-seq data, we first imputed the undersampled data using MAGIC [77].

All results are given in Table 3. From these results, we note that erythrocytes are relatively easy to distinguish from the other two cell types as the BER lower bounds were within nearly two standard deviations of zero when using genes associated with platelet, erythrocyte, and neutrophil development as well as a random selection of 12 genes. This is corroborated by the QDA cross-validated results which were all within two standard deviations of either the upper or lower bound for these gene sets. In contrast, the macrophage-associated genes seem to be less useful for distinguishing erythrocytes than the other gene sets.

Table 3. Misclassification rate of a quadratic discriminant analysis classifier (QDA) classifier and estimated upper bounds (UB) and lower bounds (LB) of the pairwise BER between mouse bone marrow cell types using the Henze–Penrose divergence applied to different combinations of genes selected from the KEGG pathways associated with the hematopoietic cell lineage. Results are presented as percentages in the form of mean $\pm$ standard deviation. Based on these results, erythrocytes are relatively easy to distinguish from the other two cell types using these gene sets.

	Platelets	**Erythrocytes**	**Neutrophils**	**Macrophages**	**Random**
Eryth. vs. Mono., LB	2.8 ± 1.5	1.2 ± 0.6	0.6 ± 0.6	8.5 ± 1.2	14.4 ± 8.4
Eryth. vs. Mono., UB	5.3 ± 2.9	2.4 ± 1.3	1.2 ± 1.3	15.5 ± 1.9	23.2 ± 12.3
Eryth. vs. Mono., Prob. Error	0.9	0.4	1.3	3.4	7.2 ± 5.4
Eryth. vs. Baso., LB	0.5 ± 0.6	0.05 ± 0.12	0.6 ± 0.5	5.1 ± 0.9	11.9 ± 5.5
Eryth. vs. Baso., UB	1.0 ± 1.1	0.1 ± 0.2	1.1 ± 0.9	9.6 ± 1.6	20.3 ± 8.8
Eryth. vs. Baso., Prob. Error	1.2	0.3	1.9	3.6	6.8 ± 5.0
Baso. vs. Mono., LB	31.1 ± 1.8	27.8 ± 3.1	27.1 ± 2.6	31.6 ± 1.3	32.1 ± 2.6
Baso. vs. Mono., UB	42.8 ± 1.4	39.9 ± 2.8	39.4 ± 2.4	43.2 ± 1.0	43.5 ± 1.2
Baso. vs. Mono., Prob. Error	28.8	30.9	23.9	22.4	29.7 ± 5.7

We also found that basophils are difficult to distinguish from monocytes using these gene sets. Assuming the relative abundance of each cell type is representative of the population, a trivial upper bound on the BER is $300/(300+559) \approx 0.35$ which is between all of the estimated lower and upper bounds. The QDA results were also relatively high (and may be overfitting the data in some cases based on the estimated BER bounds), suggesting that different genes should be explored for this classification problem.

5. Conclusions

We derived the MSE convergence rates for a kernel density plug-in estimator for a large class of divergence functionals. We generalized the theory of optimally weighted ensemble estimation and derived an ensemble divergence estimator EnDive that achieves the parametric rate when the densities are more than d times differentiable. The estimator we derived can be applied to general bounded density support sets and can be implemented without knowledge of the support, which is a distinct advantage over other competing estimators. We also derived the asymptotic distribution of the estimator, provided some guidelines for tuning parameter selection, and experimentally validated the theoretical convergence rates for the case of empirical estimation of the Rényi-α divergence integral. We then performed experiments to examine the estimator's robustness to the choice of tuning parameters, validated the central limit theorem for KL divergence estimation, and estimated bounds on the Bayes error rate for a single cell classification problem.

We note that based on the proof techniques employed in our work, our weighted ensemble estimators are easily extended beyond divergence estimation to more general distributional functionals which may be integral functionals of any number of probability distributions. We also show in Appendix B that EnDive can be easily modified to obtain an estimator that achieves the parametric rate when the densities are more than $d/2$ times differentiable and the functional g has a specific form that includes the Rényi and KL divergences. Future work includes extending this modification to functionals with more general forms. An important divergence of interest in this context is the Henze–Penrose divergence that we used to bound the Bayes error. Further future work will focus on extending this work on divergence estimation to k-nn based estimators where knowledge of the support is, again, not required. This will improve the computational burden, as k-nn estimators require fewer computations than standard KDEs.

Author Contributions: K.M. wrote this article primarily as part of his PhD dissertation under the supervision of A.H. and in collaboration with K.S. A.H, K.M., and K.S. edited the paper. K.S. provided the primary contribution to the proof of Theorem A1 and assisted with all other proofs. K.M. provided the primary contributions for the proofs of all other theorems and performed all other experiments. K.G. contributed to the bias proof.

Funding: This research was funded by Army Research Office (ARO) Multidisciplinary University Research Initiative (MURI) grant number W911NF-15-1-0479, National Science Foundation (NSF) grant number CCF-1217880, and a National Science Foundation (NSF) Graduate Research Fellowship to the first author under grant number F031543.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

KL	Kullback–Leibler
MSE	Mean squared error
SVM	Support vector machine
KDE	kernel density estimator
EnDive	Ensemble Divergence
BER	Bayes error rate
scRNA-seq	Single-cell RNA-sequencing
HP	Henze–Penrose

Appendix A. Bias Assumptions

Our full assumptions to prove the bias expressions for the estimator $\tilde{\mathbf{G}}_h$ were as follows:

- $(\mathcal{A}.0)$: Assume that the kernel K is symmetric, is a product kernel, and has bounded support in each dimension.
- $(\mathcal{A}.1)$: Assume there exist constants $\epsilon_0, \epsilon_\infty$, such that $0 < \epsilon_0 \leq f_i(x) \leq \epsilon_\infty < \infty$, $\forall x \in S$.
- $(\mathcal{A}.2)$: Assume that the densities are $f_i \in \Sigma(s, K_H)$ in the interior of $\mathcal{S}$ with $s \geq 2$.
- $(\mathcal{A}.3)$: Assume that g has an infinite number of mixed derivatives.
- $(\mathcal{A}.4)$: Assume that $\left|\frac{\partial^{k+l} g(x,y)}{\partial x^k \partial y^l}\right|$, $k, l = 0, 1, \ldots$ are strictly upper bounded for $\epsilon_0 \leq x, y \leq \epsilon_\infty$.
- $(\mathcal{A}.5)$: Assume the following boundary smoothness condition: Let $p_x(u) : \mathbb{R}^d \to \mathbb{R}$ be a polynomial in u of order $q \leq r = \lfloor s \rfloor$ whose coefficients are a function of x and are $r - q$ times differentiable. Then, assume that

$$\int_{x \in \mathcal{S}} \left(\int_{u: K(u) > 0,\, x + uh \notin \mathcal{S}} K(u) p_x(u) du \right)^t dx = v_t(h),$$

where $v_t(h)$ admits the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o\left(h^{r-q}\right),$$

for some constants $e_{i,q,t}$.

We focused on finite support kernels for simplicity in the proofs, although it is likely that our results extend to some infinitely supported kernels as well. We assumed relatively strong conditions on the smoothness of g in $\mathcal{A}.3$ to enable us to obtain an estimator that achieves good convergence rates without knowledge of the boundary of the support set. While this smoothness condition may seem restrictive, in practice, nearly all divergence and entropy functionals of interest satisfy this condition. Functionals of interest that do not satisfy this assumption (e.g., the total variation distance) typically have at least one point that is not differentiable which violates the assumptions of all competing estimators [54,57–60,65]. We also note that to obtain simply an upper bound on the bias for the plug-in estimator, much less restrictive assumptions on the functional g are sufficient.

Assumption $\mathcal{A}.5$ requires the boundary of the density support set to be smooth with respect to the kernel $(K(u))$ in the sense that the expectation of the area outside of $\mathcal{S}$ with respect to any random

variable u with smooth distribution is a smooth function of the bandwidth (h). Note that we do not require knowledge of the support of the unknown densities to actually implement the estimator ($\tilde{\mathbf{G}}_h$). As long as assumptions $\mathcal{A}.0$–$\mathcal{A}.5$ are satisfied, then the bias results we obtain are valid, and therefore, we can obtain the parametric rate with the EnDive estimator. This is in contrast to many other estimators of information theoretic measures such as those presented in references [59,60,65]. In these cases, the boundary of the support set must be known precisely to perform boundary correction to obtain the parametric rate, since the boundary correction is an explicit step in these estimators. In contrast, we do not need to explicitly perform a boundary correction.

It is not necessary for the boundary of $\mathcal{S}$ to have smooth contours with no edges or corners as assumption $\mathcal{A}.5$ is satisfied by the following case:

Theorem A1. *Assumption $\mathcal{A}.5$ is satisfied when $\mathcal{S} = [-1,1]^d$ and when K is the uniform rectangular kernel; that is, $K(x) = 1$ for all $x : ||x||_1 \leq 1/2$.*

The proof is given in Appendix D. The methods used to prove this can be easily extended to show that $\mathcal{A}.5$ is satisfied with the uniform rectangular kernel and other similar supports with flat surfaces and corners. Furthermore, we showed in reference [78] that $\mathcal{A}.5$ is satisfied using the uniform spherical kernel with a density support set equal to the unit cube. Note that assumption $\mathcal{A}.0$ is trivially satisfied by the uniform rectangular kernel as well. Again, this is easily extended to more complicated density support sets that have boundaries that contain flat surfaces and corners. Determining other combinations of kernels and density support sets that satisfy $\mathcal{A}.5$ is left for future work.

Densities for which assumptions $\mathcal{A}.1$–$\mathcal{A}.2$ hold include the truncated Gaussian distribution and the beta distribution on the unit cube. Functions for which assumptions $\mathcal{A}.3$–$\mathcal{A}.4$ hold include $g(x,y) = -\ln\left(\frac{x}{y}\right)$ and $g(x,y) = \left(\frac{x}{y}\right)^{\alpha}$.

Appendix B. Modified EnDive

If the functional g has a specific form, we can modify the EnDive estimator to obtain an estimator that achieves the parametric rate when $s > d/2$. Specifically, we have the following theorem:

Theorem A2. *Assume that assumptions $\mathcal{A}.0$–$\mathcal{A}.5$* hold. Furthermore, if *$g(x,y)$ has k,l-th order mixed derivatives $\frac{\partial^{k+l} g(x,y)}{\partial x^k \partial y^l}$ that depend on x,y only through $x^\alpha y^\beta$ for some $\alpha,\beta \in \mathbb{R}$, then for any positive integer $\lambda \geq 2$, the bias of $\tilde{\mathbf{G}}_h$ is*

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_h\right] &= \sum_{j=1}^{\lfloor s \rfloor} c_{10,j} h^j + \sum_{q=1}^{\lambda/2} \sum_{j=0}^{\lfloor s \rfloor} c_{11,q,j} \frac{h^j}{\left(Nh^d\right)^q} \\ &\quad + O\left(h^s + \frac{1}{\left(Nh^d\right)^{\frac{\lambda}{2}}}\right). \end{aligned} \tag{A1}$$

Divergence functionals that satisfy the mixed derivatives condition required for (A1) include the KL divergence and the Rényi-α divergence. Obtaining similar terms for other divergence functionals requires us to separate the dependence on h of the derivatives of g evaluated at $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h}(\mathbf{Z})$. This is left for future work. See Appendix E for details.

As compared to (3), there are many more terms in (A1). These terms enable us to modify the EnDive estimator to achieve the parametric MSE convergence rate when $s > d/2$ for an appropriate choice of bandwidths, whereas the terms in (3) requires $s \geq d$ to achieve the same rate. This is accomplished by letting $h(l)$ decrease at a faster rate, as follows.

Let $\delta > 0$ and $h(l) = lN^{\frac{-1}{d+\delta}}$ where $l \in \mathcal{L}$. The bias of each estimator in the resulting ensemble has terms proportional to $l^{j-dq}N^{-\frac{j+q}{d+\delta}}$, where $j, q \geq 0$ and $j + q > 0$. Then, the bias of $\tilde{\mathbf{G}}_{h(l)}$ satisfies condition $\mathcal{C}.1$ if $\phi_{j,q,d}(N) = N^{-\frac{j+q}{d+\delta}}$, $\psi_{j,q}(l) = l^{j-dq}$, and

$$J = \{\{j, q\} : 0 < j + q < (d+1)/2,\ q \in \{0, 1, 2, \ldots, \lambda/2\},\ j \in \{0, 1, 2, \ldots, \lfloor s \rfloor\}\}, \tag{A2}$$

as long as $L > |J| = I$. The variance also satisfies condition $\mathcal{C}.2$. The optimal weight (w_0) is found by using (6) to obtain an optimally weighted plug-in divergence functional estimator $\tilde{\mathbf{G}}_{w_0}$ that achieves the parametric convergence rate if $\lambda \geq d/\delta + 1$ and if $s \geq (d+\delta)/2$. Otherwise, if $s < (d+\delta)/2$, we can only guarantee the MSE rate up to $O\left(\frac{1}{N^{2s/(d+\delta)}}\right)$. We refer to this estimator as the modified EnDive estimator and denote it as $\tilde{\mathbf{G}}_{\text{Mod}}$. The ensemble estimator $\tilde{\mathbf{G}}_{\text{Mod}}$ is summarized in Algorithm A1 when $\delta = 1$.

Algorithm A1: The Modified EnDive Estimator

Input: η, L positive real numbers $\mathcal{L}$, samples $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$ from f_1, samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ from f_2, dimension d, function g, kernel K

Output: The modified EnDive estimator $\tilde{\mathbf{G}}_{\text{Mod}}$

1: Solve for w_0 using (6) with $\phi_{j,q,d}(N) = N^{-\frac{j+q}{d+1}}$ and basis functions $\psi_{j,q}(l) = l^{j-dq}$, $l \in \bar{l}$, and $\{i, j\} \in J$ defined in (A2)
2: **for all** $l \in \bar{l}$ **do**
3: $\quad h(l) \leftarrow lN^{\frac{-1}{d+1}}$
4: $\quad$ **for** $i = 1$ to N **do**
5: $\quad\quad \tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_i) \leftarrow \frac{1}{Nh(l)^d}\sum_{j=1}^{N} K\left(\frac{\mathbf{X}_i - \mathbf{Y}_j}{h(l)}\right)$, $\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_i) \leftarrow \frac{1}{(N-1)h(l)^d}\sum_{\substack{j=1 \\ j\neq i}}^{N} K\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h(l)}\right)$
6: $\quad$ **end for**
7: $\quad \tilde{\mathbf{G}}_{h(l)} \leftarrow \frac{1}{N}\sum_{i=1}^{N} g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_i)\right)$
8: **end for**
9: $\tilde{\mathbf{G}}_{\text{Mod}} \leftarrow \sum_{l\in\mathcal{L}} w_0(l)\tilde{\mathbf{G}}_{h(l)}$

The parametric rate can be achieved with $\tilde{\mathbf{G}}_{\text{Mod}}$ under less strict assumptions on the smoothness of the densities than those required for $\tilde{\mathbf{G}}_{\text{EnDive}}$. Since $\delta > 0$ can be arbitrary, it is theoretically possible to achieve the parametric rate with the modified estimator as long as $s > d/2$. This is consistent with the rate achieved by the more complex estimators proposed in reference [57]. We also note that the central limit theorem applies and that the convergence is uniform as Theorem 5 applies for $s > \lfloor (d+\delta)/2 \rfloor$ and $s \geq (d+\delta)/2$.

These rate improvements come at a cost for the number of parameters (L) required to implement the weighted ensemble estimator. If $s \geq \frac{d+\delta}{2}$, then the size of J for $\tilde{\mathbf{G}}_{\text{Mod}}$ is in the order of $\frac{d^2}{8\delta}$. This may lead to increased variance in the ensemble estimator as indicated by (5).

So far, $\tilde{\mathbf{G}}_{\text{Mod}}$ can only be applied to functionals ($g(x, y)$) with mixed derivatives of the form of $x^\alpha y^\beta$. Future work is required to extend this estimator to other functionals of interest.

Appendix C. General Results

Here we present the generalized forms of Theorems 1 and 2 where the sample sizes and bandwidths of the two datasets are allowed to differ. In this case, the KDEs are

$$\begin{aligned}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j) &= \frac{1}{N_1 h_1^d}\sum_{i=1}^{N_1} K\left(\frac{\mathbf{X}_j - \mathbf{Y}_i}{h_1}\right),\\ \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j) &= \frac{1}{M_2 h_2^d}\sum_{\substack{i=1\\ i\neq j}}^{N_2} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h_2}\right),\end{aligned}$$

where $M_2 = N_2 - 1$. $G\left(f_1, f_2\right)$ is then approximated as

$$\tilde{\mathbf{G}}_{h_1,h_2} = \frac{1}{N_2}\sum_{i=1}^{N_2} g\left(\tilde{\mathbf{f}}_{1,h_1}\left(\mathbf{X}_i\right), \tilde{\mathbf{f}}_{2,h_2}\left(\mathbf{X}_i\right)\right). \tag{A3}$$

We also generalize the bias result to the case where the kernel (K) has the order ν which means that the j-th moment of the kernel K_i defined as $\int t^j K_i(t)dt$ is zero for all $j = 1, \ldots, \nu - 1$ and $i = 1, \ldots, d$ where K_i is the kernel in the i-th coordinate. Note that symmetric product kernels have the order $\nu \geq 2$. The following theorem on the bias follows under assumptions $\mathcal{A}.0$–$\mathcal{A}.5$:

Theorem A3. *For general g, the bias of the plug-in estimator ($\tilde{\mathbf{G}}_{h_1,h_2}$) is of the form*

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] &= \sum_{j=1}^{r}\left(c_{4,1,j}h_1^j + c_{4,2,j}h_2^j\right) + \sum_{j=1}^{r}\sum_{i=1}^{r} c_{5,i,j}h_1^j h_2^i + O\left(h_1^s + h_2^s\right)\\ &\quad + c_{9,1}\frac{1}{N_1 h_1^d} + c_{9,2}\frac{1}{N_2 h_2^d} + o\left(\frac{1}{N_1 h_1^d} + \frac{1}{N_2 h_2^d}\right).\end{aligned} \tag{A4}$$

Furthermore, if $g(x,y)$ has k,l-th order mixed derivatives $\frac{\partial^{k+l} g(x,y)}{\partial x^k \partial y^l}$ that depend on x, y only through $x^\alpha y^\beta$ for some $\alpha, \beta \in \mathbb{R}$, then for any positive integer $\lambda \geq 2$, the bias is of the form

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] &= \sum_{j=1}^{r}\left(c_{4,1,j}h_1^j + c_{4,2,j}h_2^j\right) + \sum_{j=1}^{r}\sum_{i=1}^{r} c_{5,i,j}h_1^j h_2^i + O\left(h_1^s + h_2^s\right)\\ &\quad \sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\left(c_{9,1,j,m}\frac{h_1^m}{\left(N_1 h_1^d\right)^j} + c_{9,2,j,m}\frac{h_2^m}{\left(N_2 h_2^d\right)^j}\right)\\ &\quad + \sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{r} c_{9,j,i,m,n}\frac{h_1^m h_2^n}{\left(N_1 h_1^d\right)^j\left(N_2 h_2^d\right)^i}\\ &\quad + O\left(\frac{1}{\left(N_1 h_1^d\right)^{\frac{\lambda}{2}}} + \frac{1}{\left(N_2 h_2^d\right)^{\frac{\lambda}{2}}}\right).\end{aligned} \tag{A5}$$

Note that the bandwidth and sample size terms do not depend on the order of the kernel (ν). Thus, using a higher-order kernel does not provide any benefit to the convergence rates. This lack of improvement is due to the bias of the density estimators at the boundary of the density support sets. To obtain better convergence rates using higher-order kernels, boundary correction would be necessary [57,60]. In contrast, we improve the convergence rates by using a weighted ensemble that does not require boundary correction.

The variance result requires much less strict assumptions than the bias results:

Theorem A4. *Assume that the functional g in (1) is Lipschitz continuous in both of its arguments with the Lipschitz constant C_g. Then, the variance of the plug-in estimator ($\tilde{\mathbf{G}}_{h_1,h_2}$) is bounded by*

$$\mathbb{V}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] \leq C_g^2 ||K||_\infty^2 \left(\frac{10}{N_2} + \frac{N_1}{N_2^2}\right).$$

The proofs of these theorems are in Appendices E and F. Theorems 1 and 2 then follow.

Appendix D. Proof of Theorem A1 (Boundary Conditions)

Consider a uniform rectangular kernel $K(x)$ that satisfies $K(x) = 1$ for all x, such that $||x||_1 \leq 1/2$. Also, consider the family of probability densities (f) with rectangular support $\mathcal{S} = [-1,1]^d$. We prove Theorem A1 which is that that $\mathcal{S}$ satisfies the following smoothness condition ($\mathcal{A}.5$): for any polynomial $p_x(u) : \mathbb{R}^d \to \mathbb{R}$ of order $q \leq r = \lfloor s \rfloor$ with coefficients that are $r - q$ times differentiable wrt x,

$$\int_{x\in\mathcal{S}} \left(\int_{u:||u||_1\leq\frac{1}{2}, x+uh\notin\mathcal{S}} p_x(u)du\right)^t dx = v_t(h), \tag{A6}$$

where $v_t(h)$ has the expansion

$$v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o\left(h^{r-q}\right).$$

Note that the inner integral forces the xs under consideration to be boundary points via the constraint $x + uh \notin \mathcal{S}$.

Appendix D.1. Single Coordinate Boundary Point

We begin by focusing on points x that are boundary points by virtue of a single coordinate x_i, such that $x_i + u_i h \notin \mathcal{S}$. Without loss of generality, assume that $x_i + u_i h > 1$. The inner integral in (A6) can then be evaluated first with respect to (wrt) all coordinates other than i. Since all of these coordinates lie within the support, the inner integral over these coordinates will amount to integration of the polynomial $p_x(u)$ over a symmetric $d-1$ dimensional rectangular region $|u_j| \leq \frac{1}{2}$ for all $j \neq i$. This yields a function $\sum_{m=1}^q \tilde{p}_m(x) u_i^m$ where the coefficients $\tilde{p}_m(x)$ are each $r - q$ times differentiable wrt x.

With respect to the u_i coordinate, the inner integral will have limits from $\frac{1-x_i}{h}$ to $\frac{1}{2}$ for some $1 > x_i > 1 - \frac{h}{2}$. Consider the $\tilde{p}_q(x) u_i^q$ monomial term. The inner integral wrt this term yields

$$\sum_{m=1}^q \tilde{p}_m(x) \int_{\frac{1-x_i}{h}}^{\frac{1}{2}} u_i^m du_i = \sum_{m=1}^q \tilde{p}_m(x) \frac{1}{m+1} \left(\frac{1}{2^{m+1}} - \left(\frac{1-x_i}{h}\right)^{m+1}\right). \tag{A7}$$

Raising the right-hand-side of (A7) to the power of t results in an expression of the form

$$\sum_{j=0}^{qt} \check{p}_j(x) \left(\frac{1-x_i}{h}\right)^j, \tag{A8}$$

where the coefficients $\check{p}_j(x)$ are $r - q$ times differentiable wrt x. Integrating (A8) over all the coordinates in x other than x_i results in an expression of the form

$$\sum_{j=0}^{qt} \bar{p}_j(x_i) \left(\frac{1-x_i}{h}\right)^j, \tag{A9}$$

where, again, the coefficients$\bar{p}_j(x_i)$ are $r - q$ times differentiable wrt x_i. Note that since the other coordinates of x other than x_i are far away from the boundary, the coefficients $\bar{p}_j(x_i)$ are independent

of h. To evaluate the integral of (A9), consider the $r-q$ term Taylor series expansion of $\bar{p}_j(x_i)$ around $x_i = 1$. This will yield terms of the form

$$\begin{aligned}\int_{1-h/2}^{1} \frac{(1-x_i)^{j+k}}{h^k} dx_i &= -\frac{(1-x_i)^{j+k+1}}{h^k(j+k+1)}\Big|_{x_i=1-h/2}^{x_i=1} \\ &= \frac{h^{j+1}}{(j+k+1)2^{j+k+1}},\end{aligned}$$

for $0 \leq j \leq r-q$, and $0 \leq k \leq qt$. Combining terms results in the expansion $v_t(h) = \sum_{i=1}^{r-q} e_{i,q,t} h^i + o\left(h^{r-q}\right)$.

Appendix D.2. Multiple Coordinate Boundary Point

The case where multiple coordinates of point x are near the boundary is a straightforward extension of the single boundary point case, so we only sketch the main ideas here. As an example, consider the case where two of the coordinates are near the boundary. Assume for notational ease that they are x_1 and x_2 and that $x_1 + u_1 h > 1$ and $x_2 + u_2 h > 1$. The inner integral in (A6) can again be evaluated first wrt all coordinates other than 1 and 2. This yields a function $\sum_{m,j=1}^{q} \tilde{p}_{m,j}(x) u_1^m u_2^j$ where the coefficients $\tilde{p}_{m,j}(x)$ are each $r-q$ times differentiable wrt x. Integrating this wrt x_1 and x_2 and then raising the result to the power of t yields a double sum similar to (A8). Integrating this over all the coordinates in x other than x_1 and x_2 gives a double sum similar to (A9). Then, a Taylor series expansion of the coefficients and integration over x_1 and x_2 yields the result.

Appendix E. Proof of Theorem A3 (Bias)

In this appendix, we prove the bias results in Theorem A3. The bias of the base kernel density plug-in estimator $\tilde{\mathbf{G}}_{h_1,h_2}$ can be expressed as

$$\begin{aligned}\mathbb{B}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right] &= \mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) - g\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right)\right] \\ &= \mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) - g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right] \\ &\quad + \mathbb{E}\left[g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) - g\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right)\right],\end{aligned} \tag{A10}$$

where $\mathbf{Z}$ is drawn from f_2. The first term is the "variance" term, while the second is the "bias" term. We bound these terms using Taylor series expansions under the assumption that g is infinitely differentiable. The Taylor series expansion of the variance term in (A11) will depend on variance-like terms of the KDEs, while the Taylor series expansion of the bias term in (A11) will depend on the bias of the KDEs.

The Taylor series expansion of $g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)$ around $f_1(\mathbf{Z})$ and $f_2(\mathbf{Z})$ is

$$g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}), \mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left(\frac{\partial^{i+j} g(x,y)}{\partial x^i \partial y^j}\Big|_{\substack{x=f_1(\mathbf{Z}) \\ y=f_2(\mathbf{Z})}}\right)\frac{\mathbb{B}_{\mathbf{Z}}^i\left[\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\right]\mathbb{B}_{\mathbf{Z}}^j\left[\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right]}{i!j!}, \tag{A11}$$

where $\mathbb{B}_{\mathbf{Z}}^j\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right] = \left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) - f_i(\mathbf{Z})\right)^j$ is the bias of $\tilde{\mathbf{f}}_{i,h_i}$ at the point $\mathbf{Z}$ raised to the power of j. This expansion can be used to control the second term (the bias term) in (A11). To accomplish this, we require an expression for $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) - f_i(\mathbf{Z}) = \mathbb{B}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]$.

To obtain an expression for $\mathbb{B}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]$, we separately consider the cases when $\mathbf{Z}$ is in the interior of the support $\mathcal{S}$ or when $\mathbf{Z}$ is near the boundary of the support. A point $X \in \mathcal{S}$ is defined to be in the interior of $\mathcal{S}$ if for all $Y \notin \mathcal{S}$, $K\left(\frac{X-Y}{h_i}\right) = 0$. A point $X \in \mathcal{S}$ is near the boundary of the support if it is not in the interior. Denote the region in the interior and near the boundary wrt h_i as $\mathcal{S}_{I_i}$ and $\mathcal{S}_{B_i}$, respectively. We will need the following:

Lemma A1. *Let* $\mathbf{Z}$ *be a realization of the density* f_2 *independent of* $\tilde{\mathbf{f}}_{i,h_i}$ *for* $i = 1, 2$*. Assume that the densities* f_1 *and* f_2 *belong to* $\Sigma(s, L)$*. Then, for* $\mathbf{Z} \in \mathcal{S}_{I_i}$*,*

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right] = f_i(\mathbf{Z}) + \sum_{j=\nu/2}^{\lfloor s/2 \rfloor} c_{i,j}(\mathbf{Z}) h_i^{2j} + O\left(h_i^s\right). \tag{A12}$$

Proof. Obtaining the lower order terms in (A12) is a common result in kernel density estimation. However, since we also require the higher order terms, we present the proof here. Additionally, some of the results in this proof will be useful later. From the linearity of the KDE, we have that if $\mathbf{X}$ is drawn from f_i and is independent of $\mathbf{Z}$, then

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) &= \mathbb{E}_{\mathbf{Z}}\left[\frac{1}{h_i^d} K\left(\frac{\mathbf{X}-\mathbf{Z}}{h_i}\right)\right] \\
&= \int \frac{1}{h_i^d} K\left(\frac{x-\mathbf{Z}}{h_i}\right) f_i(x)dx \\
&= \int K(t)\, f_i(th_i + \mathbf{Z})dt,
\end{aligned} \tag{A13}$$

where the last step follows on from the substitution $t = \frac{x-\mathbf{Z}}{h_i}$. Since the density (f_i) belongs to $\Sigma(s, K)$, by using multi-index notation we can expand it to

$$f_i(th_i + \mathbf{Z}) = f_i(\mathbf{Z}) + \sum_{0<|\alpha|\leq\lfloor s \rfloor} \frac{D^\alpha f_i(\mathbf{Z})}{\alpha!} (th_i)^\alpha + O\left(\|th_i\|^s\right), \tag{A14}$$

where $\alpha! = \alpha_1!\alpha_2! \ldots \alpha_d!$ and $Y^\alpha = Y_1^{\alpha_1} Y_2^{\alpha_2} \ldots Y_d^{\alpha_d}$. Combining (A13) and (A14) gives

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z}) &= f_i(\mathbf{Z}) + \sum_{0<|\alpha|\leq\lfloor s \rfloor} \frac{D^\alpha f_i(\mathbf{Z})}{\alpha!} h_i^{|\alpha|} \int t^\alpha K(t)dt + O(h_i^s) \\
&= f_i(\mathbf{Z}) + \sum_{j=\nu/2}^{\lfloor s/2 \rfloor} c_{i,j}(\mathbf{Z}) h_i^{2j} + O(h_i^s),
\end{aligned}$$

where the last step follows from the fact that K is symmetric and of order ν. □

To obtain a similar result for the case when $\mathbf{Z}$ is near the boundary of $\mathcal{S}$, we use the assumption $\mathcal{A}.5$.

Lemma A2. *Let* $\gamma(x, y)$ *be an arbitrary function satisfying* $\sup_{x,y} |\gamma(x, y)| < \infty$*. Let* $\mathcal{S}$ *satisfy the boundary smoothness conditions of Assumption* $\mathcal{A}.5$*. Assume that the densities* f_1 *and* f_2 *belong to* $\Sigma(s, L)$*, and let* $\mathbf{Z}$ *be a realization of the* f_2 *density independently of* $\tilde{\mathbf{f}}_{i,h_i}$ *for* $i = 1, 2$*. Let* $h' = \min(h_1, h_2)$*. Then,*

$$\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}_{B_i}\}} \gamma\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right) \mathbb{B}_{\mathbf{Z}}^t\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]\right] = \sum_{j=1}^{r} c_{4,i,j,t} h_i^j + o\left(h_i^r\right) \tag{A15}$$

$$\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}_{B_1}\cap\mathcal{S}_{B_2}\}} \gamma\left(f_1(\mathbf{Z}), f_2(\mathbf{Z})\right) \mathbb{B}_{\mathbf{Z}}^t\left[\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\right] \mathbb{B}_{\mathbf{Z}}^q\left[\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right]\right] = \sum_{j=0}^{r-1}\sum_{i=0}^{r-1} c_{4,j,i,q,t} h_1^j h_2^i h' + o\left(\left(h'\right)^r\right) \tag{A16}$$

Proof. For a fixed X near the boundary of $\mathcal{S}$, we have

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathbf{f}}_{i,h_i}(X)\right] - f_i(X) &= \frac{1}{h_i^d}\int_{Y:Y\in\mathcal{S}} K\left(\frac{X-Y}{h_i}\right) f_i(Y)dY - f_i(X) \\ &= \left[\frac{1}{h_i^d}\int_{Y:K\left(\frac{X-Y}{h_i}\right)>0} K\left(\frac{X-Y}{h_i}\right) f_i(Y)dY - f_i(X)\right] \\ &\quad - \left[\frac{1}{h_i^d}\int_{Y:Y\notin\mathcal{S}} K\left(\frac{X-Y}{h_i}\right) f_i(Y)dY\right] \\ &= T_{1,i}(X) - T_{2,i}(X).\end{aligned}$$

Note that, in $T_{1,i}(X)$, we are extending the integral beyond the support of the f_i density. However, by using the same Taylor series expansion method as in the proof of Lemma A1, we always evaluate f_i and its derivatives at point X which is within the support of f_i. Thus, it does not matter how we define an extension of f_i since the Taylor series will remain the same. Thus, $T_{1,i}(X)$ results in an identical expression to that obtained from (A12).

For the $T_{2,i}(X)$ term, we expand it using multi-index notation as

$$\begin{aligned}T_{2,i}(X) &= \frac{1}{h_i^d}\int_{Y:Y\notin\mathcal{S}} K\left(\frac{X-Y}{h_i}\right) f_i(Y)dY \\ &= \int_{u:h_iu+X\notin\mathcal{S},K(u)>0} K(u)\, f_i(X+h_iu)du \\ &= \sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\int_{u:h_iu+X\notin\mathcal{S},K(u)>0} K(u)\, D^\alpha f_i(X)u^\alpha du + o\left(h_i^r\right).\end{aligned}$$

Recognizing that the $|\alpha|$th derivative of f_i is $r-|\alpha|$ times differentiable, we can apply assumption $\mathcal{A}.5$ to obtain the expectation of $T_{2,i}(X)$ wrt X:

$$\begin{aligned}\mathbb{E}\left[T_{2,i}(\mathbf{X})\right] &= \frac{1}{h_i^d}\int_X\int_{Y:Y\notin\mathcal{S}} K\left(\frac{X-Y}{h_i}\right) f_i(Y)dY f_2(X)dx \\ &= \sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\int_X\int_{u:h_iu+X\notin\mathcal{S},K(u)>0} K(u)\, D^\alpha f_i(X)u^\alpha du f_2(X)dX + o\left(h_i^r\right) \\ &= \sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\left[\sum_{1\le|\beta|\le r-|\alpha|} e_{\beta,r-|\alpha|}h_i^{|\beta|} + o\left(h_i^{r-|\alpha|}\right)\right] + o\left(h_i^r\right) \\ &= \sum_{j=1}^{r} e_j h_i^j + o\left(h_i^r\right).\end{aligned}$$

Similarly, we find that

$$\begin{aligned}\mathbb{E}\left[\left(T_{2,i}(\mathbf{X})\right)^t\right] &= \frac{1}{h_i^{dt}}\int_X\left(\int_{Y:Y\notin\mathcal{S}} K\left(\frac{X-Y}{h_i}\right) f_i(Y)dY\right)^t f_2(X)dx \\ &= \int_X\left(\sum_{|\alpha|\le r}\frac{h_i^{|\alpha|}}{\alpha!}\int_{u:h_iu+X\notin\mathcal{S},K(u)>0} K(u)\, D^\alpha f_i(X)u^\alpha du\right)^t f_2(X)dX \\ &= \sum_{j=1}^{r} e_{j,t}h_i^j + o\left(h_i^r\right).\end{aligned}$$

Combining these results gives

$$\begin{aligned}\mathbb{E}\left[1_{\{\mathbf{Z}\in\mathcal{S}_B\}}\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\left(\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})\right]-f_i(\mathbf{Z})\right)^t\right] &= \mathbb{E}\left[\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\left(T_{1,i}(\mathbf{Z})-T_{2,i}(\mathbf{Z})\right)^t\right]\\ &= \mathbb{E}\left[\gamma\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\sum_{j=0}^{t}\binom{t}{j}\left(T_{1,i}(\mathbf{Z})\right)^j\left(-T_{2,i}(\mathbf{Z})\right)^{t-j}\right]\\ &= \sum_{j=1}^{r}c_{4,i,j,t}h_i^j+o\left(h_i^r\right),\end{aligned}$$

where the constants are functionals of the kernel γ and the densities.

The expression in (A16) can be proved in a similar manner. □

Applying Lemmas A1 and A2 to (A11) gives

$$\mathbb{E}\left[g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\right]=\sum_{j=1}^{r}\left(c_{4,1,j}h_1^j+c_{4,2,j}h_2^j\right)+\sum_{j=0}^{r-1}\sum_{i=0}^{r-1}c_{5,i,j}h_1^jh_2^ih'+o\left(h_1^r+h_2^r\right). \quad \text{(A17)}$$

For the variance term (the first term) in (A11), the truncated Taylor series expansion of $g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)$ around $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})$ and $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})$ gives

$$g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right) = \sum_{i=0}^{\lambda}\sum_{j=0}^{\lambda}\left(\frac{\partial^{i+j}g(x,y)}{\partial x^i\partial y^j}\bigg|_{\substack{x=\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\\ y=\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})}}\right)\frac{\tilde{\mathbf{e}}_{1,h_1}^i(\mathbf{Z})\tilde{\mathbf{e}}_{2,h_2}^j(\mathbf{Z})}{i!j!}+o\left(\tilde{\mathbf{e}}_{1,h_1}^{\lambda}(\mathbf{Z})+\tilde{\mathbf{e}}_{2,h_2}^{\lambda}(\mathbf{Z})\right) \quad \text{(A18)}$$

where $\tilde{\mathbf{e}}_{i,h_i}(\mathbf{Z}) := \tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})-\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$. To control the variance term in (A11), we thus require expressions for $\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{i,h_i}^j(\mathbf{Z})\right]$.

Lemma A3. *Let* $\mathbf{Z}$ *be a realization of the* f_2 *density that is in the interior of the support and is independent of* $\tilde{\mathbf{f}}_{i,h_i}$ *for* $i=1,2$. *Let* $n(q)$ *be the set of integer divisors of* q *including* 1 *but excluding* q. *Then,*

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{i,h_i}^q(\mathbf{Z})\right] = \begin{cases}\sum_{j\in n(q)}\frac{1}{\left(N_2h_2^d\right)^{q-j}}\sum_{m=0}^{\lfloor s/2\rfloor}c_{6,i,q,j,m}(\mathbf{Z})h_i^{2m}+O\left(\frac{1}{N_i}\right), & q\geq 2\\ 0, & q=1,\end{cases} \quad \text{(A19)}$$

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{1,h_1}^q(\mathbf{Z})\tilde{\mathbf{e}}_{2,h_2}^l(\mathbf{Z})\right] = \begin{cases}\left(\sum_{i\in n(q)}\frac{1}{\left(N_1h_1^d\right)^{q-i}}\sum_{m=0}^{\lfloor s/2\rfloor}c_{6,1,q,i,m}(\mathbf{Z})h_1^{2m}\right)\times q, & l\geq 2\\ \left(\sum_{j\in n(l)}\frac{1}{\left(N_2h_2^d\right)^{l-j}}\sum_{t=0}^{\lfloor s/2\rfloor}c_{6,2,l,j,t}(\mathbf{Z})h_2^{2t}\right)+O\left(\frac{1}{N_1}+\frac{1}{N_2}\right), & \\ 0, & q=1\text{ or }l=1\end{cases} \quad \text{(A20)}$$

where $c_{6,i,q,j,m}$ *is a functional of* f_1 *and* f_2.

Proof. Define the random variable $\mathbf{V}_i(\mathbf{Z})=K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{h_2}\right)-\mathbb{E}_{\mathbf{Z}}K\left(\frac{\mathbf{X}_i-\mathbf{Z}}{h_2}\right)$. This gives

$$\begin{aligned}\tilde{\mathbf{e}}_{2,h_2}(\mathbf{Z}) &= \tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})-\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\\ &= \frac{1}{N_2h_2^d}\sum_{i=1}^{N_2}\mathbf{V}_i(\mathbf{Z}).\end{aligned}$$

Clearly, $\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i(\mathbf{Z}) = 0$. From (A13), we have for integer $j \geq 1$

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[K^j\left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2}\right)\right] &= \int K^j(t)\, f_2(th_2 + \mathbf{Z})dt \\ &= h_2^d \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,j,m}(\mathbf{Z})h_2^{2m},\end{aligned}$$

where the constants $c_{3,2,j,m}$ depend on density f_2, its derivatives, and the moments of kernel K^j. Note that since K is symmetric, the odd moments of K^j are zero for $\mathbf{Z}$ in the interior of the support. However, all even moments may now be non-zero since K^j may now be non-negative. In accordance with the binomial theorem,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\mathbf{V}_i^j(\mathbf{Z})\right] &= \sum_{k=0}^{j}\binom{j}{k}\mathbb{E}_{\mathbf{Z}}\left[K^k\left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2}\right)\right]\mathbb{E}_{\mathbf{Z}}\left[K\left(\frac{\mathbf{X}_i - \mathbf{Z}}{h_2}\right)\right]^{j-k} \\ &= \sum_{k=0}^{j}\binom{j}{k}h_2^d O\left(h_2^{d(j-k)}\right)\sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,k,m}(\mathbf{Z})h_2^{2m} \\ &= h_2^d \sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,j,m}(\mathbf{Z})h_2^{2m} + O\left(h^{2d}\right).\end{aligned}$$

We can use these expressions to simplify $\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^q(\mathbf{Z})\right]$. As an example, let $q = 2$. Then, since the $\mathbf{X}_i$s are independent,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^2(\mathbf{Z})\right] &= \frac{1}{N_2 h_2^{2d}}\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^2(\mathbf{Z}) \\ &= \frac{1}{N_2 h_2^d}\sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,2,m}(\mathbf{Z})h_2^{2m} + O\left(\frac{1}{N_2}\right).\end{aligned}$$

Similarly, we find that

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^3(\mathbf{Z})\right] &= \frac{1}{N_2^2 h_2^{3d}}\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^3(\mathbf{Z}) \\ &= \frac{1}{\left(N_2 h_2^d\right)^2}\sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,3,m}(\mathbf{Z})h_2^{2m} + o\left(\frac{1}{N_2}\right).\end{aligned}$$

For $q = 4$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^4(\mathbf{Z})\right] &= \frac{1}{N_2^3 h_2^{4d}}\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^4(\mathbf{Z}) + \frac{N_2 - 1}{N_2^3 h_2^{4d}}\left(\mathbb{E}_{\mathbf{Z}}\mathbf{V}_i^2(\mathbf{Z})\right)^2 \\ &= \frac{1}{\left(N_2 h_2^d\right)^3}\sum_{m=0}^{\lfloor s/2 \rfloor} c_{3,2,4,m}(\mathbf{Z})h_2^{2m} + \frac{1}{\left(N_2 h_2^d\right)^2}\sum_{m=0}^{\lfloor s/2 \rfloor} c_{6,2,2,m}(\mathbf{Z})h_2^{2m} + o\left(\frac{1}{N_2}\right).\end{aligned}$$

The pattern for $q \geq 2$ is then,

$$\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{2,h_2}^q(\mathbf{Z})\right] = \sum_{i \in n(q)}\frac{1}{\left(N_2 h_2^d\right)^{q-i}}\sum_{m=0}^{\lfloor s/2 \rfloor} c_{6,2,q,i,m}(\mathbf{Z})h_2^{2m} + O\left(\frac{1}{N_2}\right).$$

For any integer (q), the largest possible factor is $q/2$. Thus, for a given q, the smallest possible exponent on the $N_2 h_2^d$ term is $q/2$. This increases as q increases. A similar expression holds for $\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}_{1,h_1}^q(\mathbf{Z})\right]$, except the $\mathbf{X}_i$s are replaced with $\mathbf{Y}_i$, f_2 is replaced with f_1, and N_2 and h_2 are replaced

with N_1 and h_1, respectively, all resulting in different constants. Then, since $\tilde{\mathbf{e}}_{1,h_1}(\mathbf{Z})$ and $\tilde{\mathbf{e}}_{2,h_2}(\mathbf{Z})$ are conditionally independent given $\mathbf{Z}$,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\tilde{\mathbf{e}}^q_{1,h_1}(\mathbf{Z})\tilde{\mathbf{e}}^l_{2,h_2}(\mathbf{Z})\right] &= \left(\sum_{i\in n(q)}\frac{1}{\left(N_1h_1^d\right)^{q-i}}\sum_{m=0}^{\lfloor s/2\rfloor}c_{6,1,q,i,m}(\mathbf{Z})h_1^{2m}\right)\left(\sum_{j\in n(l)}\frac{1}{\left(N_2h_2^d\right)^{l-j}}\sum_{t=0}^{\lfloor s/2\rfloor}c_{6,2,l,j,t}(\mathbf{Z})h_2^{2t}\right)\\ &\quad +O\left(\frac{1}{N_1}+\frac{1}{N_2}\right).\end{aligned}$$

□

Applying Lemma A3 to (A18) when taking the conditional expectation given $\mathbf{Z}$ in the interior gives an expression of the form

$$\begin{aligned}&\sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\left(c_{7,1,j,m}\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\frac{h_1^{2m}}{\left(N_1h_1^d\right)^j}+c_{7,2,j,m}\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\frac{h_2^{2m}}{\left(N_2h_2^d\right)^j}\right)\\ &\quad+\sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{\lfloor s/2\rfloor}c_{7,j,i,m,n}\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\frac{h_1^{2m}h_2^{2n}}{\left(N_1h_1^d\right)^j\left(N_2h_2^d\right)^i}\\ &\quad+O\left(\frac{1}{\left(N_1h_1^d\right)^{\frac{\lambda}{2}}}+\frac{1}{\left(N_2h_2^d\right)^{\frac{\lambda}{2}}}\right).\end{aligned}\tag{A21}$$

Note that the functionals $c_{7,i,j,m}$ and $c_{7,j,i,m,n}$ depend on the derivatives of g and $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})$ which depend on h_i. To apply an ensemble estimation, we need to separate the dependence on h_i from the constants. If we use ODin1, then it is sufficient to note that in the interior of the support, $\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{i,h_i}(\mathbf{Z})=f_i(\mathbf{Z})+o(1)$ and therefore, $c\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)=c\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)+o(1)$ for some functional c. The terms in (A22) reduce to

$$c_{7,1,1,0}\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\frac{1}{N_1h_1^d}+c_{7,2,1,0}\left(f_1(\mathbf{Z}),f_2(\mathbf{Z})\right)\frac{1}{N_2h_2^d}+o\left(\frac{1}{N_1h_1^d}+\frac{1}{N_2h_2^d}\right).$$

For ODin2, we need the higher order terms. To separate the dependence on h_i from the constants, we need more information about the functional g and its derivatives. Consider a special case where the functional $g(x,y)$ has derivatives of the form of $x^\alpha y^\beta$ with $\alpha,\beta<0$. This includes the important cases of the KL divergence and the Renyi divergence. The generalized binomial theorem states that if $\binom{\alpha}{m}:=\frac{\alpha(\alpha-1)\ldots(\alpha-m+1)}{m!}$ and if q and t are real numbers with $|q|>|t|$, then for any complex number (α),

$$(q+t)^\alpha=\sum_{m=0}^{\infty}\binom{\alpha}{m}q^{\alpha-m}t^m.\tag{A22}$$

Since the densities are bounded away from zero, for sufficiently small h_i, we have $f_i(\mathbf{Z})>\left|\sum_{j=\nu/2}^{\lfloor s/2\rfloor}c_{i,j}(\mathbf{Z})h_i^{2j}+O\left(h_i^s\right)\right|$. Applying the generalized binomial theorem and Lemma A1 gives

$$\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z})\right)^\alpha=\sum_{m=0}^{\infty}\binom{\alpha}{m}f_i^{\alpha-m}(\mathbf{Z})\left(\sum_{j=\nu/2}^{\lfloor s/2\rfloor}c_{i,j}(\mathbf{Z})h_i^{2j}+O\left(h_i^s\right)\right)^m.$$

Since m is an integer, the exponents of the h_i terms are also integers. Thus, (A22) gives, in this case,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right] &= \sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\left(c_{8,1,j,m}(\mathbf{Z})\frac{h_1^{2m}}{\left(N_1h_1^d\right)^j}+c_{8,2,j,m}(\mathbf{Z})\frac{h_2^{2m}}{\left(N_2h_2^d\right)^j}\right)\\ &+\sum_{j=1}^{\lambda/2}\sum_{m=0}^{\lfloor s/2\rfloor}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{\lfloor s/2\rfloor}c_{8,j,i,m,n}(\mathbf{Z})\frac{h_1^{2m}h_2^{2n}}{\left(N_1h_1^d\right)^j\left(N_2h_2^d\right)^i} \quad \text{(A23)}\\ &+O\left(\frac{1}{\left(N_1h_1^d\right)^{\frac{\lambda}{2}}}+\frac{1}{\left(N_2h_2^d\right)^{\frac{\lambda}{2}}}+h_1^s+h_2^s\right).\end{aligned}$$

As before, the case for $\mathbf{Z}$ close to the boundary of the support is more complicated. However, by using a similar technique to the proof of Lemma A2 for $\mathbf{Z}$ at the boundary and combining with previous results, we find that for general g,

$$\mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right]=c_{9,1}\frac{1}{N_1h_1^d}+c_{9,2}\frac{1}{N_2h_2^d}+o\left(\frac{1}{N_1h_1^d}+\frac{1}{N_2h_2^d}\right). \quad \text{(A24)}$$

If $g(x,y)$ has derivatives of the form of $x^\alpha y^\beta$ with $\alpha,\beta<0$, then we can similarly obtain

$$\begin{aligned}\mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)-g\left(\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{1,h_1}(\mathbf{Z}),\mathbb{E}_{\mathbf{Z}}\tilde{\mathbf{f}}_{2,h_2}(\mathbf{Z})\right)\right] &= \sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\left(c_{9,1,j,m}\frac{h_1^m}{\left(N_1h_1^d\right)^j}+c_{9,2,j,m}\frac{h_2^m}{\left(N_2h_2^d\right)^j}\right)\\ &+\sum_{j=1}^{\lambda/2}\sum_{m=0}^{r}\sum_{i=1}^{\lambda/2}\sum_{n=0}^{r}c_{9,j,i,m,n}\frac{h_1^mh_2^n}{\left(N_1h_1^d\right)^j\left(N_2h_2^d\right)^i} \quad \text{(A25)}\\ &+O\left(\frac{1}{\left(N_1h_1^d\right)^{\frac{\lambda}{2}}}+\frac{1}{\left(N_2h_2^d\right)^{\frac{\lambda}{2}}}+h_1^s+h_2^s\right).\end{aligned}$$

Combining (A17) with either (A24) or (A26) completes the proof.

Appendix F. Proof of Theorem A4 (Variance)

To bound the variance of the plug-in estimator $\tilde{\mathbf{G}}_{h_1,h_2}$, we use the Efron–Stein inequality [70]:

Lemma A4 (Efron–Stein Inequality). *Let $\mathbf{X}_1,\ldots,\mathbf{X}_n,\mathbf{X}_1',\ldots,\mathbf{X}_n'$ be independent random variables on the space $\mathcal{S}$. Then, if $f:\mathcal{S}\times\cdots\times\mathcal{S}\to\mathbb{R}$, we have*

$$\mathbb{V}\left[f(\mathbf{X}_1,\ldots,\mathbf{X}_n)\right]\leq\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left(f(\mathbf{X}_1,\ldots,\mathbf{X}_n)-f(\mathbf{X}_1,\ldots,\mathbf{X}_i',\ldots,\mathbf{X}_n)\right)^2\right].$$

Suppose we have samples $\left\{\mathbf{X}_1,\ldots,\mathbf{X}_{N_2},\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_1}\right\}$ and $\left\{\mathbf{X}_1',\ldots,\mathbf{X}_{N_2},\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_1}\right\}$ and denote the respective estimators as $\tilde{\mathbf{G}}_{h_1,h_2}$ and $\tilde{\mathbf{G}}_{h_1,h_2}'$. We have

$$\begin{aligned}\left|\tilde{\mathbf{G}}_{h_1,h_2}-\tilde{\mathbf{G}}_{h_1,h_2}'\right| &\leq \frac{1}{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1'),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1')\right)\right|\\ &+\frac{1}{N_2}\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}'(\mathbf{X}_j)\right)\right|. \quad \text{(A26)}\end{aligned}$$

Since g is Lipschitz continuous with the constant C_g, we have

$$\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1'),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1')\right)\right| \leq C_g\left(\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1')\right|+\left|\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1')\right|\right), \quad \text{(A27)}$$

$$\begin{aligned}\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'})\right| &= \frac{1}{N_1h_1^d}\left|\sum_{i=1}^{N_1}\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)\right| \\ &\leq \frac{1}{N_1h_1^d}\sum_{i=1}^{N_1}\left|K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right| \qquad \text{(A28)} \\ \implies \mathbb{E}\left[\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'})\right|^2\right] &\leq \frac{1}{N_1h_1^{2d}}\sum_{i=1}^{N_1}\mathbb{E}\left[\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)^2\right],\end{aligned}$$

where the last step follows from Jensen's inequality. By making the substitutions $\mathbf{u}_i = \frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}$ and $\mathbf{u}_i^{'} = \frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}$, this gives

$$\begin{aligned}\frac{1}{h_1^{2d}}\mathbb{E}\left[\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)^2\right] &= \frac{1}{h^{2d}}\int\left(K\left(\frac{\mathbf{X}_1-\mathbf{Y}_i}{h_1}\right)-K\left(\frac{\mathbf{X}_1^{'}-\mathbf{Y}_i}{h_1}\right)\right)^2\times \\ &\quad f_2(\mathbf{X}_1)f_2(\mathbf{X}_1^{'})f_1(\mathbf{Y}_i)d\mathbf{X}_1d\mathbf{X}_1^{'}d\mathbf{Y}_i \\ &\leq 2||K||_\infty^2.\end{aligned}$$

Combining this with (A29) gives

$$\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'})\right|^2\right] \leq 2||K||_\infty^2.$$

Similarly,

$$\mathbb{E}\left[\left|\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)-\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right|^2\right] \leq 2||K||_\infty^2.$$

Combining these results with (A27) gives

$$\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right)\right)^2\right] \leq 8C_g^2||K||_\infty^2. \qquad \text{(A29)}$$

The second term in (A26) is controlled in a similar way. From the Lipschitz condition,

$$\begin{aligned}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|^2 &\leq C_g^2\left|\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)-\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right|^2 \\ &= \frac{C_g^2}{M_2^2h_2^{2d}}\left(K\left(\frac{\mathbf{X}_j-\mathbf{X}_1}{h}\right)-K\left(\frac{\mathbf{X}_j-\mathbf{X}_1^{'}}{h}\right)\right)^2.\end{aligned}$$

The h_2^{2d} terms are eliminated by making the substitutions of $\mathbf{u}_j = \frac{\mathbf{X}_j-\mathbf{X}_1}{h_2}$ and $\mathbf{u}_j^{'} = \frac{\mathbf{X}_j-\mathbf{X}_1^{'}}{h_2}$ within the expectation to obtain

$$\begin{aligned}\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|^2\right] &\leq \frac{2C_g^2||K||_\infty^2}{M_2^2} \qquad \text{(A30)} \\ \implies \mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\right)^2\right] &\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=2}^{N_2}\sum_{i=2}^{N_2}\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j)\right)\right|\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i),\tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_i)\right)\right|\right] \\
&\leq M_2^2\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j)\right)\right|^2\right] \qquad \text{(A31)} \\
&\leq 2C_g^2||K||_\infty^2,
\end{aligned}
$$

where we use the Cauchy Schwarz inequality to bound the expectation within each summand. Finally, applying Jensen's inequality and (A29) and (A32) gives

$$
\begin{aligned}
\mathbb{E}\left[\left|\tilde{\mathbf{G}}_{h_1,h_2}-\tilde{\mathbf{G}}'_{h_1,h_2}\right|^2\right] &\leq \frac{2}{N_2^2}\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1)\right)\right|^2\right] \\
&\quad+\frac{2}{N_2^2}\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right] \\
&\leq \frac{20C_g^2||K||_\infty^2}{N_2^2}.
\end{aligned}
$$

Now, suppose we have samples $\{\mathbf{X}_1,\ldots,\mathbf{X}_{N_2},\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_1}\}$ and $\left\{\mathbf{X}_1,\ldots,\mathbf{X}_{N_2},\mathbf{Y}'_1,\ldots,\mathbf{Y}_{N_1}\right\}$ and denote the respective estimators as $\tilde{\mathbf{G}}_{h_1,h_2}$ and $\tilde{\mathbf{G}}'_{h_1,h_2}$. Then,

$$
\begin{aligned}
\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right| &\leq C_g\left|\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j)-\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j)\right| \\
&= \frac{C_g}{N_1h_1^d}\left|K\left(\frac{\mathbf{X}_j-\mathbf{Y}_1}{h_1}\right)-K\left(\frac{\mathbf{X}_j-\mathbf{Y}'_1}{h_1}\right)\right| \\
\implies \mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right|^2\right] &\leq \frac{2C_g^2||K||_\infty^2}{N_1^2}.
\end{aligned}
$$

Thus, using a similar argument as was used to obtain (A32),

$$
\begin{aligned}
\mathbb{E}\left[\left|\tilde{\mathbf{G}}_{h_1,h_2}-\tilde{\mathbf{G}}'_{h_1,h_2}\right|^2\right] &\leq \frac{1}{N_2^2}\mathbb{E}\left[\left(\sum_{j=1}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right] \\
&\leq \frac{2C_g^2||K||_\infty^2}{N_2^2}.
\end{aligned}
$$

Applying the Efron–Stein inequality gives

$$
\mathbb{V}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right]\leq\frac{10C_g^2||K||_\infty^2}{N_2}+\frac{C_g^2||K||_\infty^2N_1}{N_2^2}.
$$

Appendix G. Proof of Theorem 4 (CLT)

We are interested in the asymptotic distribution of

$$
\begin{aligned}
\sqrt{N_2}\left(\tilde{\mathbf{G}}_{h_1,h_2}-\mathbb{E}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right]\right) &= \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)-\mathbb{E}_{\mathbf{X}_j}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right]\right) \\
&\quad+\frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\left(\mathbb{E}_{\mathbf{X}_j}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right]-\mathbb{E}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right]\right).
\end{aligned}
$$

Note that in the standard central limit theorem [79], the second term converges in distribution to a Gaussian random variable. If the first term converges in probability to a constant (specifically, 0), then we can use Slutsky's theorem [80] to find the asymptotic distribution. So, now, we focus on the first term which we denote as $\mathbf{V}_{N_2}$.

To prove convergence in probability, we use Chebyshev's inequality. Note that $\mathbb{E}\left[\mathbf{V}_{N_2}\right] = 0$. To bound the variance of $\mathbf{V}_{N_2}$, we again use the Efron–Stein inequality. Let $\mathbf{X}_1^{'}$ be drawn from f_2 and denote $\mathbf{V}_{N_2}$ and $\mathbf{V}_{N_2}^{'}$ as the sequences using $\mathbf{X}_1$ and $\mathbf{X}_1^{'}$, respectively. Then,

$$\begin{aligned}\mathbf{V}_{N_2} - \mathbf{V}_{N_2}^{'} &= \frac{1}{\sqrt{N_2}}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right) \\ &\quad - \frac{1}{\sqrt{N_2}}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right) - \mathbb{E}_{\mathbf{X}_1^{'}}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right)\right]\right) \\ &\quad + \frac{1}{\sqrt{N_2}}\sum_{j=2}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right).\end{aligned} \tag{A32}$$

Note that

$$\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right)^2\right] = \mathbb{E}\left[\mathbb{V}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right].$$

We use the Efron–Stein inequality to bound $\mathbb{V}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]$.
We do this by bounding the conditional expectation of the term

$$\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_1)\right)\right|,$$

where $\mathbf{X}_i$ is replaced with $\mathbf{X}_i^{'}$ in the KDEs for some $i \neq 1$. Using similar steps as in Appendix F, we have

$$\mathbb{E}_{\mathbf{X}_1}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_1)\right)\right|^2\right] = O\left(\frac{1}{N_2^2}\right).$$

A similar result is obtained when $\mathbf{Y}_i$ is replaced with $\mathbf{Y}_i^{'}$. Then, based on the Efron–Stein inequality, $\mathbb{V}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right] = O\left(\frac{1}{N_2} + \frac{1}{N_1}\right)$.
Therefore,

$$\mathbb{E}\left[\frac{1}{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right)^2\right] = O\left(\frac{1}{N_2^2} + \frac{1}{N_1 N_2}\right).$$

A similar result holds for the $g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1^{'}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1^{'})\right)$ terms in (A33).
For the third term in (A33),

$$\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\right)^2\right]$$

$$= \sum_{j=2}^{N_2}\sum_{i=2}^{N_2}\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_i)\right)\right|\right].$$

There are M_2 terms where $i = j$, and we have from Appendix F (see (A30)) that

$$\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}^{'}(\mathbf{X}_j)\right)\right|^2\right] \leq \frac{2C_g^2||K||_\infty^2}{M_2^2}.$$

Thus, these terms are $O\left(\frac{1}{M_2}\right)$. There are $M_2^2 - M_2$ terms when $i \neq j$. In this case, we can do four substitutions of the form $\mathbf{u}_j = \frac{\mathbf{X}_j - \mathbf{X}_1}{h_2}$ to obtain

$$\mathbb{E}\left[\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j)\right)\right| \left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_i)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_i), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_i)\right)\right|\right] \leq \frac{4C_g^2 ||K||_\infty^2 h_2^{2d}}{M_2^2}.$$

Then, since $h_2^d = o(1)$, we get

$$\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right] = o(1), \tag{A33}$$

$$\begin{aligned}
\implies \mathbb{E}\left[\left(\mathbf{V}_{N_2} - \mathbf{V}'_{N_2}\right)^2\right] &\leq \frac{3}{N_2}\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right) - \mathbb{E}_{\mathbf{X}_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_1)\right)\right]\right)^2\right] \\
&\quad + \frac{3}{N_2}\mathbb{E}\left[\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1)\right) - \mathbb{E}_{\mathbf{X}'_1}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}'_1), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}'_1)\right)\right]\right)^2\right] \\
&\quad + \frac{3}{N_2}\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}'_{2,h_2}(\mathbf{X}_j)\right)\right)\right)^2\right] \\
&= o\left(\frac{1}{N_2}\right).
\end{aligned}$$

Now, consider samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_{N_2}, \mathbf{Y}_1, \ldots, \mathbf{Y}_{N_1}\}$ and $\left\{\mathbf{X}_1, \ldots, \mathbf{X}_{N_2}, \mathbf{Y}'_1, \ldots, \mathbf{Y}_{N_1}\right\}$ and the respective sequences $\mathbf{V}_{N_2}$ and $\mathbf{V}'_{N_2}$. Then,

$$\mathbf{V}_{N_2} - \mathbf{V}'_{N_2} = \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right).$$

Using a similar argument as that used to obtain (A33), we have that if $h_1^d = o(1)$ and $N_1 \to \infty$, then

$$\mathbb{E}\left[\left(\sum_{j=2}^{N_2}\left|g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right) - g\left(\tilde{\mathbf{f}}'_{1,h_1}(\mathbf{X}_j), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X}_j)\right)\right|\right)^2\right] = o(1)$$

$$\implies \mathbb{E}\left[\left(\mathbf{V}_{N_2} - \mathbf{V}'_{N_2}\right)^2\right] = o\left(\frac{1}{N_2}\right).$$

Applying the Efron–Stein inequality gives

$$\mathbb{V}\left[\mathbf{V}_{N_2}\right] = o\left(\frac{N_2 + N_1}{N_2}\right) = o(1).$$

Thus, based on Chebyshev's inequality,

$$\Pr\left(\left|\mathbf{V}_{N_2}\right| > \epsilon\right) \leq \frac{\mathbb{V}\left[\mathbf{V}_{N_2}\right]}{\epsilon^2} = o(1),$$

and therefore, $\mathbf{V}_{N_2}$ converges to zero in probability. Based on Slutsky's theorem, $\sqrt{N_2}\left(\tilde{\mathbf{G}}_{h_1,h_2} - \mathbb{E}\left[\tilde{\mathbf{G}}_{h_1,h_2}\right]\right)$ converges in distribution to a zero mean Gaussian random variable with variance

$$\mathbb{V}\left[\mathbb{E}_{\mathbf{X}}\left[g\left(\tilde{\mathbf{f}}_{1,h_1}(\mathbf{X}), \tilde{\mathbf{f}}_{2,h_2}(\mathbf{X})\right)\right]\right],$$

where $\mathbf{X}$ is drawn from f_2.

For the weighted ensemble estimator, we wish to know the asymptotic distribution of $\sqrt{N_2}\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right)$ where $\tilde{\mathbf{G}}_w = \sum_{l\in\bar{I}} w(l)\tilde{\mathbf{G}}_{h(l)}$. We have

$$\begin{aligned}\sqrt{N_2}\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right]\right) &= \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\sum_{l\in\bar{I}} w(l)\left(g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right) - \mathbb{E}_{\mathbf{X}_j}\left[g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right)\right]\right)\\ &\quad + \frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\left(\mathbb{E}_{\mathbf{X}_j}\left[\sum_{l\in\bar{I}} w(l)g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right)\right] - \mathbb{E}\left[\sum_{l\in\bar{I}} w(l)g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right)\right]\right).\end{aligned}$$

The second term again converges in distribution to a Gaussian random variable by the central limit theorem. The mean and variance are, respectively, zero and

$$\mathbb{V}\left[\sum_{l\in\bar{I}} w(l)\mathbb{E}_{\mathbf{X}}\left[g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}),\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X})\right)\right]\right].$$

The first term is equal to

$$\begin{aligned}\sum_{l\in\bar{I}} w(l)\left(\frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\left(g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right) - \mathbb{E}_{\mathbf{X}_j}\left[g\left(\tilde{\mathbf{f}}_{1,h(l)}(\mathbf{X}_j),\tilde{\mathbf{f}}_{2,h(l)}(\mathbf{X}_j)\right)\right]\right)\right) &= \sum_{l\in\bar{I}} w(l)o_P(1)\\ &= o_P(1),\end{aligned}$$

where $o_P(1)$ denotes the convergence to zero as a probability. In the last step, we use the fact that if two random variables converge in probability to constants, then their linear combination converges in probability to the linear combination of the constants. Combining this result with Slutsky's theorem completes the proof.

Appendix H. Proof of Theorem 5 (Uniform MSE)

Since the MSE is equal to the square of the bias plus the variance, we can upper bound the left hand side of (7) with

$$\begin{aligned}\sup_{p,q\in\Sigma(s,K_H,\epsilon_0,\epsilon_\infty)} \mathbb{E}\left[\left(\tilde{\mathbf{G}}_{w_0} - G(p,q)\right)^2\right] &= \sup_{p,q\in\Sigma(s,K_H,\epsilon_0,\epsilon_\infty)}\left(\text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 + \text{Var}\left(\tilde{\mathbf{G}}_{w_0}\right)\right)\\ &\leq \sup_{p,q\in\Sigma(s,K_H,\epsilon_0,\epsilon_\infty)}\text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 + \sup_{p,q\in\Sigma(s,K_H,\epsilon_0,\epsilon_\infty)}\text{Var}\left(\tilde{\mathbf{G}}_{w_0}\right).\end{aligned}$$

From the assumptions (Lipschitz, kernel bounded, weight calculated from the relaxed optimization problem), we have

$$\begin{aligned}\sup_{p,q\in\Sigma(s,K_H,\epsilon_0,\epsilon_\infty)}\text{Var}\left(\tilde{\mathbf{G}}_{w_0}\right) &\leq \sup_{p,q\in\Sigma(s,K,\epsilon_0,\epsilon_\infty)}\frac{11C_g^2||w_0||_2^2||K||_\infty}{N}\\ &= \frac{11C_g^2||w_0||_2^2||K||_\infty}{N},\end{aligned}$$

where the last step follows on from the fact that all of the terms are independent of p and q.

For the bias, recall that if g is infinitely differentiable and if the optimal weight (w_0) is calculated using the relaxed convex optimization problem, then

$$\begin{aligned}\text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right) &= \sum_{i\in J} c_i(p,q)\epsilon N^{-1/2},\\ \implies \text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 &= \frac{\epsilon^2}{N}\left(\sum_{i\in J} c_i(p,q)\right)^2.\end{aligned} \tag{A34}$$

We use a topology argument to bound the supremum of this term. We use the Extreme Value Theorem [81]:

Theorem A5 (Extreme Value Theorem). *Let $f : X \to \mathbb{R}$ be continuous. If X is compact, then points $c, d \in X$ s.t. $f(c) \leq f(x) \leq f(d)$ exist for every $x \in X$.*

Based on this theorem, f achieves its minimum and maximum on X. Our approach is to first show that the functionals $c_i(p, q)$ are continuous wrt p and q in some appropriate norm. We then show that the space $\Sigma(s, K_H, \epsilon_0, \epsilon_\infty)$ is compact wrt this norm. The Extreme Value Theorem can then be applied to bound the supremum of (A34).

We first define the norm. Let $\alpha = s - r > 0$. We use the standard norm on the space $\Sigma(s, K_H)$ [82]:

$$\begin{aligned} ||f|| &= ||f||_{\Sigma(s,K_H)} \\ &= ||f||_{C^r} + \max_{|\beta|=r} |D^\beta f|_{C^{0,\alpha}} \end{aligned}$$

where

$$\begin{aligned} ||f||_{C^r} &= \max_{|\beta|\leq r} \sup_{x\in\mathcal{S}} |D^\beta f(x)|, \\ |f|_{C^{0,\alpha}} &= \sup_{x\neq y\in\mathcal{S}} \frac{|f(x) - f(y)|}{|x - y|^\alpha}. \end{aligned}$$

Lemma A5. *The functionals $c_i(p, q)$ are continuous wrt the norm* $\max(||p||_{C^r}, ||q||_{C^r})$.

Proof. The functionals $c_i(p, q)$ depend on terms of the form

$$c(p,q) = \int \left(\left. \frac{\partial^{i+j} g(t_1, t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1 = p(x) \\ t_2 = q(x)}} \right) D^\beta p(x) D^\gamma q(x) q(x) dx. \tag{A35}$$

It is sufficient to show that this is continuous. Let $\epsilon > 0$ and $\max\left(||p - p_0||_{C^r}, ||q - q_0||_{C^r}\right) < \delta$ where $\delta > 0$ will be chosen later. Then, by applying the triangle inequality for integration and adding and subtracting terms, we have

$$|c(p, q) - c(p_0, q_0)|$$

$$
\begin{aligned}
\leq & \int \left| \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1=p(x) \\ t_2=q(x)}} \right) D^\beta p(x) D^\gamma q(x) \left(q(x)-q_0(x)\right) \right| dx \\
& + \int \left| \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1=p(x) \\ t_2=q(x)}} \right) D^\beta p(x) q_0(x) \left(D^\gamma q(x) - D^\gamma q_0(x)\right) \right| dx \\
& + \int \left| D^\beta p_0(x) D^\gamma q_0(x) q_0(x) \left(\left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1=p(x) \\ t_2=q(x)}} \right) \right. \right. \\
& \qquad \left. \left. - \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1=p_0(x) \\ t_2=q_0(x)}} \right) \right) \right| dx \\
& + \int \left| \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1=p(x) \\ t_2=q(x)}} \right) D^\gamma q_0(x) q_0(x) \left(D^\beta p(x) - D^\beta p_0(x)\right) \right| dx.
\end{aligned} \tag{A36}
$$

Based on Assumption $\mathcal{A}.4$, the absolute value of the mixed derivatives of g is bounded on the range defined for p and q by some constant ($C_{i,j}$). Also, $q_0(x) \leq \epsilon_\infty$. Furthermore, since $D^\gamma q_0$ and $D^\beta p$ are continuous, and since $\mathcal{S} \subset \mathbb{R}^d$ is compact, then the absolute value of derivatives $D^\gamma q_0$ and $D^\beta p$ is also bounded by a constant ($\epsilon_\infty^{'}$). Let $\delta_0 > 0$. Then, since the mixed derivatives of g are continuous on the interval $[\epsilon_0, \epsilon_\infty]$, they are uniformly continuous. Therefore, we can choose a small enough δ such that (s.t.)

$$
\left| \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1=p(x) \\ t_2=q(x)}} \right) - \left(\left. \frac{\partial^{i+j} g(t_1,t_2)}{\partial t_1^i \partial t_2^j} \right|_{\substack{t_1=p_0(x) \\ t_2=q_0(x)}} \right) \right| < \delta_0. \tag{A37}
$$

Combining all of these results with (A36) gives

$$
\begin{aligned}
|c(p,q) - c(p_0,q_0)| \quad \leq \quad & \lambda(\mathcal{S}) \delta C_{ij} \epsilon_\infty^{'} \left(2 + \epsilon_\infty\right) \\
& + \lambda(\mathcal{S}) \epsilon_\infty^{'} \epsilon_\infty \left(2\delta_0 + C_{ij}\delta\right),
\end{aligned}
$$

where $\lambda(\mathcal{S})$ is the Lebesgue measure of $\mathcal{S}$. This is bounded since $\mathcal{S}$ is compact. Let $\delta_0^{'} > 0$ be s.t. if $\max\left(||p-p_0||_{C^r}, ||q-q_0||_{C^r}\right) < \delta_0^{'}$, then (A37) is less than $\frac{\epsilon}{4\lambda(\mathcal{S})\epsilon_\infty^{'}\epsilon_\infty}$. Let $\delta_1 = \frac{\epsilon}{4\lambda(\mathcal{S})C_{ij}\epsilon_\infty^{'}(1+\epsilon_\infty)}$. Then, if $\delta < \min(\delta_0^{'}, \delta_1)$,

$$
|c(p,q) - c(p_0,q_0)| < \epsilon.
$$

□

Given that each $c_i(p,q)$ is continuous, then $\left(\sum_{i \in J} c_i(p,q)\right)^2$ is also continuous wrt p and q.

We now argue that $\Sigma(s, K_H)$ is compact. First, a set is relatively compact if its closure is compact. Based on the Arzela–Ascoli theorem [83], the space $\Sigma(s, K_H)$ is relatively compact in the topology induced by the $\|\cdot\|_{\Sigma(t,K_H)}$ norm for any $t < s$. We choose $t = r$. It can then be shown that under the $\|\cdot\|_{\Sigma(r,K_H)}$ norm, $\Sigma(s, K_H)$ is complete [82]. Since $\Sigma(s, K_H)$ is contained in a metric space, it is also closed

and therefore, equal to its closure. Thus, $\Sigma(s, K_H)$ is compact. Then, since $\Sigma(s, K_H, \epsilon_0, \epsilon_\infty)$ is closed in $\Sigma(s, K_H)$, it is also compact. Therefore, since for each $p, q \in \Sigma(s, K_H, \epsilon_0, \epsilon_\infty)$, $\left(\sum_{i \in J} c_i(p, q)\right)^2 < \infty$, based on the Extreme Value Theorem, we have

$$\begin{aligned} \sup_{p,q \in \Sigma(s,K_H,\epsilon_0,\epsilon_\infty)} \text{Bias}\left(\tilde{\mathbf{G}}_{w_0}\right)^2 &= \sup_{p,q \in \Sigma(s,K_H,\epsilon_0,\epsilon_\infty)} \frac{\epsilon^2}{N} \left(\sum_{i \in J} c_i(p, q)\right)^2 \\ &= \frac{\epsilon^2}{N} C, \end{aligned}$$

where we use the fact that J is finite (see Section 3.2 or Appendix B for the set J).

References

1. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2012.
2. Avi-Itzhak, H.; Diep, T. Arbitrarily tight upper and lower bounds on the Bayesian probability of error. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 89–91. [CrossRef]
3. Hashlamoun, W.A.; Varshney, P.K.; Samarasooriya, V. A tight upper bound on the Bayesian probability of error. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 220–224. [CrossRef]
4. Moon, K.; Delouille, V.; Hero, A.O., III. Meta learning of bounds on the Bayes classifier error. In Proceedings of the 2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE), Salt Lake City, UT, USA, 9–12 August 2015; pp. 13–18.
5. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **1952**, *23*, 493–507. [CrossRef]
6. Berisha, V.; Wisler, A.; Hero, A.O., III; Spanias, A. Empirically Estimable Classification Bounds Based on a New Divergence Measure. *IEEE Trans. Signal Process.* **2016**, *64*, 580–591. [CrossRef] [PubMed]
7. Moon, K.R.; Hero, A.O., III. Multivariate f-Divergence Estimation With Confidence. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2420–2428.
8. Gliske, S.V.; Moon, K.R.; Stacey, W.C.; Hero, A.O., III. The intrinsic value of HFO features as a biomarker of epileptic activity. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.
9. Loh, P.-L. On Lower Bounds for Statistical Learning Theory. *Entropy* **2017**, *19*, 617. [CrossRef]
10. Póczos, B.; Schneider, J.G. On the estimation of alpha-divergences. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 609–617.
11. Oliva, J.; Póczos, B.; Schneider, J. Distribution to distribution regression. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1049–1057.
12. Szabó, Z.; Gretton, A.; Póczos, B.; Sriperumbudur, B. Two-stage sampled learning theory on distributions. In Proceeding of The 18th International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015.
13. Moon, K.R.; Delouille, V.; Li, J.J.; De Visscher, R.; Watson, F.; Hero, A.O., III. Image patch analysis of sunspots and active regions. II. Clustering via matrix factorization. *J. Space Weather Space Clim.* **2016**, *6*, A3. [CrossRef]
14. Moon, K.R.; Li, J.J.; Delouille, V.; De Visscher, R.; Watson, F.; Hero, A.O., III. Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis. *J. Space Weather Space Clim.* **2016**, *6*, A2. [CrossRef]
15. Dhillon, I.S.; Mallela, S.; Kumar, R. A divisive information theoretic feature clustering algorithm for text classification. *J. Mach. Learn. Res.* **2003**, *3*, 1265–1287.
16. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
17. Lewi, J.; Butera, R.; Paninski, L. Real-time adaptive information-theoretic optimization of neurophysiology experiments. In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS 2006), Vancouver, BC, Canada, 4–9 December 2006; pp. 857–864.
18. Bruzzone, L.; Roli, F.; Serpico, S.B. An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection. *IEEE Trans. Geosci. Remote Sens.* **1995**, *33*, 1318–1321. [CrossRef]

19. Guorong, X.; Peiqi, C.; Minhui, W. Bhattacharyya distance feature selection. In Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 25–29 August 1996; Volume 2, pp. 195–199.
20. Sakate, D.M.; Kashid, D.N. Variable selection via penalized minimum φ-divergence estimation in logistic regression. *J. Appl. Stat.* **2014**, *41*, 1233–1246. [CrossRef]
21. Hild, K.E.; Erdogmus, D.; Principe, J.C. Blind source separation using Renyi's mutual information. *IEEE Signal Process. Lett.* **2001**, *8*, 174–176. [CrossRef]
22. Mihoko, M.; Eguchi, S. Robust blind source separation by beta divergence. *Neural Comput.* **2002**, *14*, 1859–1886. [CrossRef] [PubMed]
23. Vemuri, B.C.; Liu, M.; Amari, S.; Nielsen, F. Total Bregman divergence and its applications to DTI analysis. *IEEE Trans. Med. Imaging* **2011**, *30*, 475–483. [CrossRef] [PubMed]
24. Hamza, A.B.; Krim, H. Image registration and segmentation by maximizing the Jensen-Rényi divergence. In Proceedings of the 4th International Workshop Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR 2003), Lisbon, Portugal, 7–9 July 2003; pp. 147–163.
25. Liu, G.; Xia, G.; Yang, W.; Xue, N. SAR image segmentation via non-local active contours. In Proceedings of the 2014 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Quebec City, QC, Canada, 13–18 July 2014; pp. 3730–3733.
26. Korzhik, V.; Fedyanin, I. Steganographic applications of the nearest-neighbor approach to Kullback-Leibler divergence estimation. In Proceedings of the 2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC), Moscow, Russia, 3–5 February 2015; pp. 133–138.
27. Basseville, M. Divergence measures for statistical data processing–An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633. [CrossRef]
28. Cichocki, A.; Amari, S. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568. [CrossRef]
29. Csiszar, I. Information-type measures of difference of probability distributions and indirect observations. *Stud. Sci. Math. Hungar.* **1967**, *2*, 299–318.
30. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1966**, *28*, 131–142.
31. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
32. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
33. Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Rein. Angew. Math.* **1909**, *136*, 210–271. (In German)
34. Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Indian J. Stat.* **1946**, *7*, 401–406.
35. Silva, J.F.; Parada, P.A. Shannon entropy convergence results in the countable infinite case. In Proceedings of the 2012 IEEE International Symposium on Information Theory Proceedings (ISIT), Cambridge, MA, USA, 1–6 July 2012; pp. 155–159.
36. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms* **2001**, *19*, 163–193. [CrossRef]
37. Valiant, G.; Valiant, P. Estimating the unseen: An n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs. In Proceedings of the 43rd Annual ACM Symposium on Theory of Computing, San Jose, CA, USA, 6–8 June 2011; pp. 685–694.
38. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2015**, *61*, 2835–2885. [CrossRef] [PubMed]
39. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2017**, *63*, 6774–6798. [CrossRef]
40. Valiant, G.; Valiant, P. The power of linear estimators. In Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS), Palm Springs, CA, USA, 22–25 October 2011; pp. 403–412.
41. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
42. Paninski, L. Estimating entropy on m bins given fewer than m samples. *IEEE Trans. Inf. Theory* **2004**, *50*, 2200–2203. [CrossRef]

43. Alba-Fernández, M.V.; Jiménez-Gamero, M.D.; Ariza-López, F.J. Minimum Penalized ϕ-Divergence Estimation under Model Misspecification. *Entropy* **2018**, *20*, 329. [CrossRef]
44. Ahmed, N.A.; Gokhale, D. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Inf. Theory* **1989**, *35*, 688–692. [CrossRef]
45. Misra, N.; Singh, H.; Demchuk, E. Estimation of the entropy of a multivariate normal distribution. *J. Multivar. Anal.* **2005**, *92*, 324–342. [CrossRef]
46. Gupta, M.; Srivastava, S. Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy* **2010**, *12*, 818–843. [CrossRef]
47. Li, S.; Mnatsakanov, R.M.; Andrew, M.E. K-nearest neighbor based consistent entropy estimation for hyperspherical distributions. *Entropy* **2011**, *13*, 650–667. [CrossRef]
48. Wang, Q.; Kulkarni, S.R.; Verdú, S. Divergence estimation for multidimensional densities via *k*-nearest-neighbor distances. *IEEE Trans. Inf. Theory* **2009**, *55*, 2392–2405. [CrossRef]
49. Darbellay, G.A.; Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inf. Theory* **1999**, *45*, 1315–1321. [CrossRef]
50. Silva, J.; Narayanan, S.S. Information divergence estimation based on data-dependent partitions. *J. Stat. Plan. Inference* **2010**, *140*, 3180–3198. [CrossRef]
51. Le, T.K. Information dependency: Strong consistency of Darbellay–Vajda partition estimators. *J. Stat. Plan. Inference* **2013**, *143*, 2089–2100. [CrossRef]
52. Wang, Q.; Kulkarni, S.R.; Verdú, S. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Inf. Theory* **2005**, *51*, 3064–3074. [CrossRef]
53. Hero, A.O., III; Ma, B.; Michel, O.; Gorman, J. Applications of entropic spanning graphs. *IEEE Signal Process. Mag.* **2002**, *19*, 85–95. [CrossRef]
54. Moon, K.R.; Hero, A.O., III. Ensemble estimation of multivariate f-divergence. In Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 356–360.
55. Moon, K.R.; Sricharan, K.; Greenewald, K.; Hero, A.O., III. Improving convergence of divergence functional ensemble estimators. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1133–1137.
56. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [CrossRef]
57. Krishnamurthy, A.; Kandasamy, K.; Poczos, B.; Wasserman, L. Nonparametric Estimation of Renyi Divergence and Friends. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 919–927.
58. Singh, S.; Póczos, B. Generalized exponential concentration inequality for Rényi divergence estimation. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), Beijing, China, 21–26 June 2014; pp. 333–341.
59. Singh, S.; Póczos, B. Exponential Concentration of a Density Functional Estimator. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 3032–3040.
60. Kandasamy, K.; Krishnamurthy, A.; Poczos, B.; Wasserman, L.; Robins, J. Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; pp. 397–405.
61. Härdle, W. *Applied Nonparametric Regression*; Cambridge University Press: Cambridge, UK, 1990.
62. Berlinet, A.; Devroye, L.; Györfi, L. Asymptotic normality of *L*1-error in density estimation. *Statistics* **1995**, *26*, 329–343. [CrossRef]
63. Berlinet, A.; Györfi, L.; Dénes, I. Asymptotic normality of relative entropy in multivariate density estimation. *Publ. l'Inst. Stat. l'Univ. Paris* **1997**, *41*, 3–27.
64. Bickel, P.J.; Rosenblatt, M. On some global measures of the deviations of density function estimates. *Ann. Stat.* **1973**, *1*, 1071–1095. [CrossRef]
65. Sricharan, K.; Wei, D.; Hero, A.O., III. Ensemble estimators for multivariate entropy estimation. *IEEE Trans. Inf. Theory* **2013**, *59*, 4374–4388. [CrossRef] [PubMed]

66. Berrett, T.B.; Samworth, R.J.; Yuan, M. Efficient multivariate entropy estimation via *k*-nearest neighbour distances. *arXiv* **2017**, arXiv:1606.00304.
67. Kozachenko, L.; Leonenko, N.N. Sample estimate of the entropy of a random vector. *Probl. Peredachi Inf.* **1987**, *23*, 9–16.
68. Hansen, B.E. (University of Wisconsin, Madison, WI, USA). Lecture Notes on Nonparametrics, 2009.
69. Budka, M.; Gabrys, B.; Musial, K. On accuracy of PDF divergence estimators and their applicability to representative data sampling. *Entropy* **2011**, *13*, 1229–1266. [CrossRef]
70. Efron, B.; Stein, C. The jackknife estimate of variance. *Ann. Stat.* **1981**, *9*, 586–596. [CrossRef]
71. Wisler, A.; Moon, K.; Berisha, V. Direct ensemble estimation of density functionals. In Proceedings of the 2018 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
72. Moon, K.R.; Sricharan, K.; Greenewald, K.; Hero, A.O., III. Nonparametric Ensemble Estimation of Distributional Functionals. *arXiv* **2016**, arXiv:1601.06884v2.
73. Paul, F.; Arkin, Y.; Giladi, A.; Jaitin, D.A.; Kenigsberg, E.; Keren-Shaul, H.; Winter, D.; Lara-Astiaso, D.; Gury, M.; Weiner, A.; et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **2015**, *163*, 1663–1677. [CrossRef] [PubMed]
74. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
75. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2015**, *44*, D457–D462. [CrossRef] [PubMed]
76. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2016**, *45*, D353–D361. [CrossRef] [PubMed]
77. Van Dijk, D.; Sharma, R.; Nainys, J.; Yim, K.; Kathail, P.; Carr, A.J.; Burdsiak, C.; Moon, K.R.; Chaffer, C.; Pattabiraman, D.; et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **2018**, *174*, 716–729. [CrossRef] [PubMed]
78. Moon, K.R.; Sricharan, K.; Hero, A.O., III. Ensemble Estimation of Distributional Functionals via *k*-Nearest Neighbors. *arXiv* **2017**, arXiv:1707.03083.
79. Durrett, R. *Probability: Theory and Examples*; Cambridge University Press: Cambridge, UK, 2010.
80. Gut, A. *Probability: A Graduate Course*; Springer: Berlin/Heidelberg, Germany, 2012.
81. Munkres, J. *Topology*; Prentice Hall: Englewood Cliffs, NJ, USA, 2000.
82. Evans, L.C. *Partial Differential Equations*; American Mathematical Society: Providence, RI, USA, 2010.
83. Gilbarg, D.; Trudinger, N.S. *Elliptic Partial Differential Equations of Second Order*; Springer: Berlin/Heidelberg, Germany, 2001.