Multivariate f-Divergence Estimation With Confidence

Kevin R. Moon
Department of EECS
University of Michigan
Ann Arbor, MI
krmoon@umich.edu

Alfred O. Hero III
Department of EECS
University of Michigan
Ann Arbor, MI
hero@eecs.umich.edu

Abstract

The problem of f-divergence estimation is important in the fields of machine learning, information theory, and statistics. While several nonparametric divergence estimators exist, relatively few have known convergence properties. In particular, even for those estimators whose MSE convergence rates are known, the asymptotic distributions are unknown. We establish the asymptotic normality of a recently proposed ensemble estimator of f-divergence between two distributions from a finite number of samples. This estimator has MSE convergence rate of $O\left(\frac{1}{T}\right)$, is simple to implement, and performs well in high dimensions. This theory enables us to perform divergence-based inference tasks such as testing equality of pairs of distributions based on empirical samples. We experimentally validate our theoretical results and, as an illustration, use them to empirically bound the best achievable classification error.

1 Introduction

This paper establishes the asymptotic normality of a nonparametric estimator of the f-divergence between two distributions from a finite number of samples. For many nonparametric divergence estimators the large sample consistency has already been established and the mean squared error (MSE) convergence rates are known for some. However, there are few results on the asymptotic distribution of non-parametric divergence estimators. Here we show that the asymptotic distribution is Gaussian for the class of ensemble f-divergence estimators [1], extending theory for entropy estimation [2, 3] to divergence estimation. f-divergence is a measure of the difference between distributions and is important to the fields of machine learning, information theory, and statistics [4]. The f-divergence generalizes several measures including the Kullback-Leibler (KL) [5] and Rényi-α [6] divergences. Divergence estimation is useful for empirically estimating the decay rates of error probabilities of hypothesis testing [7], extending machine learning algorithms to distributional features [8, 9], and other applications such as text/multimedia clustering [10]. Additionally, a special case of the KL divergence is mutual information which gives the capacities in data compression and channel coding [7]. Mutual information estimation has also been used in machine learning applications such as feature selection [11], fMRI data processing [12], clustering [13], and neuron classification [14]. Entropy is also a special case of divergence where one of the distributions is the uniform distribution. Entropy estimation is useful for intrinsic dimension estimation [15], texture classification and image registration [16], and many other applications.

However, one must go beyond entropy and divergence estimation in order to perform inference tasks on the divergence. An example of an inference task is detection: to test the null hypothesis that the divergence is zero, i.e., testing that the two populations have identical distributions. Prescribing a p-value on the null hypothesis requires specifying the null distribution of the divergence estimator. Another statistical inference problem is to construct a confidence interval on the divergence based on

the divergence estimator. This paper provides solutions to these inference problems by establishing large sample asymptotics on the distribution of divergence estimators. In particular we consider the asymptotic distribution of the nonparametric weighted ensemble estimator of f-divergence from [1]. This estimator estimates the f-divergence from two finite populations of i.i.d. samples drawn from some unknown, nonparametric, smooth, d-dimensional distributions. The estimator [1] achieves a MSE convergence rate of $O\left(\frac{1}{T}\right)$ where T is the sample size. See [17] for proof details.

1.1 Related Work

Estimators for some f-divergences already exist. For example, Póczos & Schneider [8] and Wang et al [18] provided consistent k-nn estimators for Rényi-α and the KL divergences, respectively. Consistency has been proven for other mutual information and divergence estimators based on plug-in histogram schemes [19, 20, 21, 22]. Hero et al [16] provided an estimator for Rényi-α divergence but assumed that one of the densities was known. However none of these works study the convergence rates of their estimators nor do they derive the asymptotic distributions.

Recent work has focused on deriving convergence rates for divergence estimators. Nguyen et al [23], Singh and Póczos [24], and Krishnamurthy et al [25] each proposed divergence estimators that achieve the parametric convergence rate ($O\left(\frac{1}{T}\right)$) under weaker conditions than those given in [1]. However, solving the convex problem of [23] can be more demanding for large sample sizes than the estimator given in [1] which depends only on simple density plug-in estimates and an offline convex optimization problem. Singh and Póczos only provide an estimator for Rényi-α divergences that requires several computations at each boundary of the support of the densities which becomes difficult to implement as d gets large. Also, this method requires knowledge of the support of the densities which may not be possible for some problems. In contrast, while the convergence results of the estimator in [1] requires the support to be bounded, knowledge of the support is not required for implementation. Finally, the estimators given in [25] estimate divergences that include functionals of the form $\int f_1^\alpha(x) f_2^\beta(x) d\mu(x)$ for given α, β. While a suitable α-β indexed sequence of divergence functionals of the form in [25] can be made to converge to the KL divergence, this does not guarantee convergence of the corresponding sequence of divergence estimates, whereas the estimator in [1] can be used to estimate the KL divergence. Also, for some divergences of the specified form, numerical integration is required for the estimators in [25], which can be computationally difficult. In any case, the asymptotic distributions of the estimators in [23, 24, 25] are currently unknown.

Asymptotic normality has been established for certain appropriately normalized divergences between a specific density estimator and the true density [26, 27, 28]. However, this differs from our setting where we assume that both densities are unknown. Under the assumption that the two densities are smooth, lower bounded, and have bounded support, we show that an appropriately normalized weighted ensemble average of kernel density plug-in estimators of f-divergence converges in distribution to the standard normal distribution. This is accomplished by constructing a sequence of interchangeable random variables and then showing (by concentration inequalities and Taylor series expansions) that the random variables and their squares are asymptotically uncorrelated. The theory developed to accomplish this can also be used to derive a central limit theorem for a weighted ensemble estimator of entropy such as the one given in [3].We verify the theory by simulation. We then apply the theory to the practical problem of empirically bounding the Bayes classification error probability between two population distributions, without having to construct estimates for these distributions or implement the Bayes classifier.

Bold face type is used in this paper for random variables and random vectors. Let f_1 and f_2 be densities and define $L(x) = \frac{f_1(x)}{f_2(x)}$. The conditional expectation given a random variable $\mathbf{Z}$ is $\mathbb{E}_{\mathbf{Z}}$.

2 The Divergence Estimator

Moon and Hero [1] focused on estimating divergences that include the form [4]

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

for a smooth, function $g(f)$. (Note that although g must be convex for (1) to be a divergence, the estimator in [1] does not require convexity.) The divergence estimator is constructed us-

ing k-nn density estimators as follows. Assume that the d-dimensional multivariate densities f_1 and f_2 have finite support $\mathcal{S} = [a, b]^d$. Assume that $T = N + M_2$ i.i.d. realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M_2}\}$ are available from the density f_2 and M_1 i.i.d. realizations $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{M_1}\}$ are available from the density f_1. Assume that $k_i \leq M_i$. Let $\rho_{2,k_2}(i)$ be the distance of the k_2th nearest neighbor of $\mathbf{X}_i$ in $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_T\}$ and let $\rho_{1,k_1}(i)$ be the distance of the k_1th nearest neighbor of $\mathbf{X}_i$ in $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{M_1}\}$. Then the k-nn density estimate is [29]

$$\hat{\mathbf{f}}_{i,k_i}(X_j) = \frac{k_i}{M_i \bar{c} \rho^d_{i,k_i}(j)},$$

where $\bar{c}$ is the volume of a d-dimensional unit ball.

To construct the plug-in divergence estimator, the data from f_2 are randomly divided into two parts $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M_2}\}$. The k-nn density estimate $\hat{\mathbf{f}}_{2,k_2}$ is calculated at the N points $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ using the M_2 realizations $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M_2}\}$. Similarly, the k-nn density estimate $\hat{\mathbf{f}}_{1,k_1}$ is calculated at the N points $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ using the M_1 realizations $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{M_1}\}$. Define $\hat{\mathbf{L}}_{k_1,k_2}(x) = \frac{\hat{\mathbf{f}}_{1,k_1}(x)}{\hat{\mathbf{f}}_{2,k_2}(x)}$. The functional $G(f_1, f_2)$ is then approximated as

$$\hat{\mathbf{G}}_{k_1,k_2} = \frac{1}{N} \sum_{i=1}^{N} g\left(\hat{\mathbf{L}}_{k_1,k_2}\left(\mathbf{X}_i\right)\right). \tag{2}$$

The principal assumptions on the densities f_1 and f_2 and the functional g are that: 1) f_1, f_2, and g are smooth; 2) f_1 and f_2 have common bounded support sets $\mathcal{S}$; 3) f_1 and f_2 are strictly lower bounded. The full assumptions $(\mathcal{A}.0) - (\mathcal{A}.5)$ are given in the supplementary material and in[17]. Moon and Hero [1] showed that under these assumptions, the MSE convergence rate of the estimator in Eq. 2 to the quantity in Eq. 1 depends exponentially on the dimension d of the densities. However, Moon and Hero also showed that an estimator with the parametric convergence rate $O(1/T)$ can be derived by applying the theory of optimally weighted ensemble estimation as follows.

Let $\bar{l} = \{l_1, \ldots, l_L\}$ be a set of index values and T the number of samples available. For an indexed ensemble of estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l \in \bar{l}}$ of the parameter E, the weighted ensemble estimator with weights $w = \{w(l_1), \ldots, w(l_L)\}$ satisfying $\sum_{l \in \bar{l}} w(l) = 1$ is defined as $\hat{\mathbf{E}}_w = \sum_{l \in \bar{l}} w(l) \hat{\mathbf{E}}_l$. The key idea to reducing MSE is that by choosing appropriate weights w, we can greatly decrease the bias in exchange for some increase in variance. Consider the following conditions on $\left\{\hat{\mathbf{E}}_l\right\}_{l \in \bar{l}}$ [3]:

- $\mathcal{C}.1$ The bias is given by

$$\text{Bias}\left(\hat{\mathbf{E}}_l\right) = \sum_{i \in J} c_i \psi_i(l) T^{-i/2d} + O\left(\frac{1}{\sqrt{T}}\right),$$

  where c_i are constants depending on the underlying density, $J = \{i_1, \ldots, i_I\}$ is a finite index set with $I < L$, $\min(J) > 0$ and $\max(J) \leq d$, and $\psi_i(l)$ are basis functions depending only on the parameter l.

- $\mathcal{C}.2$ The variance is given by

$$\text{Var}\left[\hat{\mathbf{E}}_l\right] = c_v \left(\frac{1}{T}\right) + o\left(\frac{1}{T}\right).$$

Theorem 1. *[3] Assume conditions $\mathcal{C}.1$ and $\mathcal{C}.2$ hold for an ensemble of estimators $\left\{\hat{\mathbf{E}}_l\right\}_{l \in \bar{l}}$. Then there exists a weight vector w_0 such that*

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

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

$$\begin{array}{ll} \min_w & ||w||_2 \\ subject\,to & \sum_{l \in \bar{l}} w(l) = 1, \\ & \gamma_w(i) = \sum_{l \in \bar{l}} w(l) \psi_i(l) = 0,\ i \in J. \end{array}$$

Algorithm 1 Optimally weighted ensemble divergence estimator

Input: α, η, L positive real numbers $\bar{l}$, samples $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{M_1}\}$ from f_1, samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ from f_2, dimension d, function g, $\bar{c}$

Output: The optimally weighted divergence estimator $\hat{\mathbf{G}}_{w_0}$

Solve for w_0 using Eq. 3 with basis functions $\psi_i(l) = l^{i/d}$, $l \in \bar{l}$ and $i \in \{1, \ldots, d-1\}$
$M_2 \leftarrow \alpha T$, $N \leftarrow T - M_2$
for all $l \in \bar{l}$ **do**
  $k(l) \leftarrow l\sqrt{M_2}$
  for $i = 1$ to N **do**
    $\rho_{j,k(l)}(i) \leftarrow$the distance of the $k(l)$th nearest neighbor of $\mathbf{X}_i$ in $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{M_1}\}$ and $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_T\}$ for $j = 1, 2$, respectively
    $\hat{\mathbf{f}}_{j,k(l)}(\mathbf{X}_i) \leftarrow \frac{k(l)}{M_j \bar{c} \rho^d_{j,k(l)}(i)}$ for $j = 1, 2$, $\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_i) \leftarrow \frac{\hat{\mathbf{f}}_{1,k(l)}}{\hat{\mathbf{f}}_{2,k(l)}}$
  end for
  $\hat{\mathbf{G}}_{k(l)} \leftarrow \frac{1}{N} \sum_{i=1}^{N} g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_i)\right)$
end for
$\hat{\mathbf{G}}_{w_0} \leftarrow \sum_{l \in \bar{l}} w_0(l) \hat{\mathbf{G}}_{k(l)}$

In order to achieve the rate of $O(1/T)$ it is not necessary for the weights to zero out the lower order bias terms, i.e. that $\gamma_w(i) = 0$, $i \in J$. It was shown in [3] that solving the following convex optimization problem in place of the optimization problem in Theorem 1 retains the MSE convergence rate of $O(1/T)$:

$$\begin{array}{rl} \min_w & \epsilon \\ subject\,to & \sum_{l \in \bar{l}} w(l) = 1, \\ & \left|\gamma_w(i) T^{\frac{1}{2} - \frac{i}{2d}}\right| \leq \epsilon, \ i \in J, \\ & \|w\|_2^2 \leq \eta, \end{array} \tag{3}$$

where the parameter η is chosen to 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 the rate of $O(1/\sqrt{T})$ which gives an MSE rate of $O(1/T)$.

Theorem 1 was applied in [3] to obtain an entropy estimator with convergence rate $O(1/T)$. Moon and Hero [1] similarly applied Theorem 1 to obtain a divergence estimator with the same rate in the following manner. Let $L > I = d - 1$ and choose $\bar{l} = \{l_1, \ldots, l_L\}$ to be positive real numbers. Assume that $M_1 = O(M_2)$. Let $k(l) = l\sqrt{M_2}$, $M_2 = \alpha T$ with $0 < \alpha < 1$, $\hat{\mathbf{G}}_{k(l)} := \hat{\mathbf{G}}_{k(l),k(l)}$, and $\hat{\mathbf{G}}_w := \sum_{l \in \bar{l}} w(l) \hat{\mathbf{G}}_{k(l)}$. Note that the parameter l indexes over different neighborhood sizes for the k-nn density estimates. From [1], the biases of the ensemble estimators $\left\{\hat{\mathbf{G}}_{k(l)}\right\}_{l \in \bar{l}}$ satisfy the condition $\mathcal{C}.1$ when $\psi_i(l) = l^{i/d}$ and $J = \{1, \ldots, d-1\}$. The general form of the variance of $\hat{\mathbf{G}}_{k(l)}$ also follows $\mathcal{C}.2$. The optimal weight w_0 is found by using Theorem 1 to obtain a plug-in f-divergence estimator with convergence rate of $O(1/T)$. The estimator is summarized in Algorithm 1.

3 Asymptotic Normality of the Estimator

The following theorem shows that the appropriately normalized ensemble estimator $\hat{\mathbf{G}}_w$ converges in distribution to a normal random variable.

Theorem 2. *Assume that assumptions* $(\mathcal{A}.0) - (\mathcal{A}.5)$ *hold and let* $M = O(M_1) = O(M_2)$ *and* $k(l) = l\sqrt{M}$ *with* $l \in \bar{l}$. *The asymptotic distribution of the weighted ensemble estimator* $\hat{\mathbf{G}}_w$ *is given by*

$$\lim_{M,N \to \infty} Pr\left(\frac{\hat{\mathbf{G}}_w - \mathbb{E}\left[\hat{\mathbf{G}}_w\right]}{\sqrt{Var\left[\hat{\mathbf{G}}_w\right]}} \leq t\right) = Pr(\mathbf{S} \leq t),$$

where $\mathbf{S}$ *is a standard normal random variable. Also* $\mathbb{E}\left[\hat{\mathbf{G}}_w\right] \to G(f_1, f_2)$ *and Var* $\left[\hat{\mathbf{G}}_w\right] \to 0$.

The results on the mean and variance come from [1]. The proof of the distributional convergence is outlined below and is based on constructing a sequence of interchangeable random variables $\{\mathbf{Y}_{M,i}\}_{i=1}^N$ with zero mean and unit variance. We then show that the $\mathbf{Y}_{M,i}$ are asymptotically uncorrelated and that the $\mathbf{Y}_{M,i}^2$ are asymptotically uncorrelated as $M \to \infty$. This is similar to what was done in [30] to prove a central limit theorem for a density plug-in estimator of entropy. Our analysis for the ensemble estimator of divergence is more complicated since we are dealing with a functional of two densities and a weighted ensemble of estimators. In fact, some of the equations we use to prove Theorem 2 can be used to prove a central limit theorem for a weighted ensemble of entropy estimators such as that given in [3].

3.1 Proof Sketch of Theorem 2

The full proof is included in the supplemental material. We use the following lemma from [30, 31]:

Lemma 3. *Let the random variables* $\{\mathbf{Y}_{M,i}\}_{i=1}^N$ *belong to a zero mean, unit variance, interchangeable process for all values of* M. *Assume that* $Cov(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2})$ *and* $Cov(\mathbf{Y}_{M,1}^2, \mathbf{Y}_{M,2}^2)$ *are* $O(1/M)$. *Then the random variable*

$$\mathbf{S}_{N,M} = \left(\sum_{i=1}^N \mathbf{Y}_{M,i}\right) / \sqrt{Var\left[\sum_{i=1}^N \mathbf{Y}_{M,i}\right]} \tag{4}$$

converges in distribution to a standard normal random variable.

This lemma is an extension of work by Blum et al [32] which showed that if $\{\mathbf{Z}_i; i = 1, 2, \dots\}$ is an interchangeable process with zero mean and unit variance, then $\mathbf{S}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathbf{Z}_i$ converges in distribution to a standard normal random variable if and only if $Cov\left[\mathbf{Z}_1, \mathbf{Z}_2\right] = 0$ and $Cov\left[\mathbf{Z}_1^2, \mathbf{Z}_2^2\right] = 0$. In other words, the central limit theorem holds if and only if the interchangeable process is uncorrelated and the squares are uncorrelated. Lemma 3 shows that for a correlated interchangeable process, a sufficient condition for a central limit theorem is for the interchangeable process and the squared process to be asymptotically uncorrelated with rate $O(1/M)$.

For simplicity, let $M_1 = M_2 = M$ and $\hat{\mathbf{L}}_{k(l)} := \hat{\mathbf{L}}_{k(l),k(l)}$. Define

$$\mathbf{Y}_{M,i} = \frac{\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_i)\right) - \mathbb{E}\left[\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_i)\right)\right]}{\sqrt{\text{Var}\left[\sum_{l\in\bar{l}} w(l) g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_i)\right)\right]}}.$$

Then from Eq. 4, we have that

$$\mathbf{S}_{N,M} = \left(\hat{\mathbf{G}}_w - \mathbb{E}\left[\hat{\mathbf{G}}_w\right]\right) / \sqrt{\text{Var}\left[\hat{\mathbf{G}}_w\right]}.$$

Thus it is sufficient to show from Lemma 3 that $Cov(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2})$ and $Cov(\mathbf{Y}_{M,1}^2, \mathbf{Y}_{M,2}^2)$ are $O(1/M)$. To do this, it is necessary to show that the denominator of $\mathbf{Y}_{M,i}$ converges to a nonzero constant or to zero sufficiently slowly. It is also necessary to show that the covariance of the numerator is $O(1/M)$. Therefore, to bound $Cov(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2})$, we require bounds on the quantity $Cov\left[g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_i)\right), g\left(\hat{\mathbf{L}}_{k(l')}(\mathbf{X}_j)\right)\right]$ where $l, l' \in \bar{l}$.

Define $\mathcal{M}(\mathbf{Z}) := \mathbf{Z} - \mathbb{E}\mathbf{Z}$, $\hat{\mathbf{F}}_{k(l)}(\mathbf{Z}) := \hat{\mathbf{L}}_{k(l)}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{Z})\right)$, and $\hat{\mathbf{e}}_{i,k(l)}(\mathbf{Z}) := \hat{\mathbf{f}}_{i,k(l)}(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}\hat{\mathbf{f}}_{i,k(l)}(\mathbf{Z})$. Assuming g is sufficiently smooth, a Taylor series expansion of $g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{Z})\right)$ around $\mathbb{E}_{\mathbf{Z}}\hat{\mathbf{L}}_{k(l)}(\mathbf{Z})$ gives

$$g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{Z})\right) = \sum_{i=0}^{\lambda-1} \frac{g^{(i)}\left(\mathbb{E}_{\mathbf{Z}}\hat{\mathbf{L}}_{k(l)}(\mathbf{Z})\right)}{i!} \hat{\mathbf{F}}_{k(l)}^i(\mathbf{Z}) + \frac{g^{(\lambda)}(\xi_{\mathbf{Z}})}{\lambda!} \hat{\mathbf{F}}_{k(l)}^{\lambda}(\mathbf{Z}),$$

where $\xi_{\mathbf{Z}} \in \left(\mathbb{E}_{\mathbf{Z}}\hat{\mathbf{F}}_{k(l)}(\mathbf{Z}), \hat{\mathbf{F}}_{k(l)}(\mathbf{Z})\right)$. We use this expansion to bound the covariance. The expected value of the terms containing the derivatives of g is controlled by assuming that the densities are lower bounded. By assuming the densities are sufficiently smooth, an expression for $\hat{\mathbf{F}}^q_{k(l)}(\mathbf{Z})$ in terms of powers and products of the density error terms $\hat{\mathbf{e}}_{1,k(l)}$ and $\hat{\mathbf{e}}_{2,k(l)}$ is obtained by expanding $\hat{\mathbf{L}}_{k(l)}(\mathbf{Z})$ around $\mathbb{E}_{\mathbf{Z}}\hat{\mathbf{f}}_{1,k(l)}(\mathbf{Z})$ and $\mathbb{E}_{\mathbf{Z}}\hat{\mathbf{f}}_{2,k(l)}(\mathbf{Z})$ and applying the binomial theorem. The expected value of products of these density error terms is bounded by applying concentration inequalities and conditional independence. Then the covariance between $\hat{\mathbf{F}}^q_{k(l)}(\mathbf{Z})$ terms is bounded by bounding the covariance between powers and products of the density error terms by applying Cauchy-Schwarz and other concentration inequalities. This gives the following lemma which is proved in the supplemental material.

Lemma 4. *Let $l, l' \in \bar{l}$ be fixed, $M_1 = M_2 = M$, and $k(l) = l\sqrt{M}$. Let $\gamma_1(x)$, $\gamma_2(x)$ be arbitrary functions with 1 partial derivative wrt x and $\sup_x |\gamma_i(x)| < \infty$, $i = 1, 2$ and let $1_{\{\cdot\}}$ be the indicator function. Let $\mathbf{X}_i$ and $\mathbf{X}_j$ be realizations of the density f_2 independent of $\hat{\mathbf{f}}_{1,k(l)}$, $\hat{\mathbf{f}}_{1,k(l')}$, $\hat{\mathbf{f}}_{2,k(l)}$, and $\hat{\mathbf{f}}_{2,k(l')}$ and independent of each other when $i \neq j$. Then*

$$Cov\left[\gamma_1(\mathbf{X}_i)\hat{\mathbf{F}}^q_{k(l)}(\mathbf{X}_i), \gamma_2(\mathbf{X}_j)\hat{\mathbf{F}}^r_{k(l')}(\mathbf{X}_j)\right] = \begin{cases} o(1), & i = j \\ 1_{\{q,r=1\}} c_8\left(\gamma_1(x), \gamma_2(x)\right)\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right), & i \neq j. \end{cases}$$

Note that $k(l)$ is required to grow with $\sqrt{M}$ for Lemma 4 to hold. Define $h_{l,g}(\mathbf{X}) = g\left(\mathbb{E}_{\mathbf{X}}\hat{\mathbf{L}}_{k(l)}(\mathbf{X})\right)$. Lemma 4 can then be used to show that

$$Cov\left[g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_i)\right), g\left(\hat{\mathbf{L}}_{k(l')}(\mathbf{X}_j)\right)\right] = \begin{cases} \mathbb{E}\left[\mathcal{M}\left(h_{l,g}(\mathbf{X}_i)\right)\mathcal{M}\left(h_{l',g}(\mathbf{X}_i)\right)\right] + o(1), & i = j \\ c_8\left(h_{l,g'}(x), h_{l',g'}(x)\right)\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right), & i \neq j. \end{cases}$$

For the covariance of $\mathbf{Y}^2_{M,i}$ and $\mathbf{Y}^2_{M,j}$, assume WLOG that $i = 1$ and $j = 2$. Then for l, l', j, j' we need to bound the term

$$Cov\left[\mathcal{M}\left(g\left(\hat{\mathbf{L}}_{k(l)}(\mathbf{X}_1)\right)\right)\mathcal{M}\left(g\left(\hat{\mathbf{L}}_{k(l')}(\mathbf{X}_1)\right)\right), \mathcal{M}\left(g\left(\hat{\mathbf{L}}_{k(j)}(\mathbf{X}_2)\right)\right)\mathcal{M}\left(g\left(\hat{\mathbf{L}}_{k(j')}(\mathbf{X}_2)\right)\right)\right]. \tag{5}$$

For the case where $l = l'$ and $j = j'$, we can simply apply the previous results to the functional $d(x) = \left(\mathcal{M}\left(g(x)\right)\right)^2$. For the more general case, we need to show that

$$Cov\left[\gamma_1(\mathbf{X}_1)\hat{\mathbf{F}}^s_{k(l)}(\mathbf{X}_1)\hat{\mathbf{F}}^q_{k(l')}(\mathbf{X}_1), \gamma_2(\mathbf{X}_2)\hat{\mathbf{F}}^t_{k(j)}(\mathbf{X}_2)\hat{\mathbf{F}}^r_{k(j')}(\mathbf{X}_2)\right] = O\left(\frac{1}{M}\right). \tag{6}$$

To do this, bounds are required on the covariance of up to eight distinct density error terms. Previous results can be applied by using Cauchy-Schwarz when the sum of the exponents of the density error terms is greater than or equal to 4. When the sum is equal to 3, we use the fact that $k(l) = O(k(l'))$ combined with Markov's inequality to obtain a bound of $O\left(1/M\right)$. Applying Eq. 6 to the term in Eq. 5 gives the required bound to apply Lemma 3.

3.2 Broad Implications of Theorem 2

To the best of our knowledge, Theorem 2 provides the first results on the asymptotic distribution of an f-divergence estimator with MSE convergence rate of $O\left(1/T\right)$ under the setting of a finite number of samples from two unknown, non-parametric distributions. This enables us to perform inference tasks on the class of f-divergences (defined with smooth functions g) on smooth, strictly lower bounded densities with finite support. Such tasks include hypothesis testing and constructing a confidence interval on the error exponents of the Bayes probability of error for a classification problem. This greatly increases the utility of these divergence estimators.

Although we focused on a specific divergence estimator, we suspect that our approach of showing that the components of the estimator and their squares are asymptotically uncorrelated can be adapted to derive central limit theorems for other divergence estimators that satisfy similar assumptions (smooth g, and smooth, strictly lower bounded densities with finite support). We speculate that this would be easiest for estimators that are also based on k-nearest neighbors such as in [8] and [18]. It is also possible that the approach can be adapted to other plug-in estimator approaches such as in [24] and [25]. However, the qualitatively different convex optimization approach of divergence estimation in [23] may require different methods.

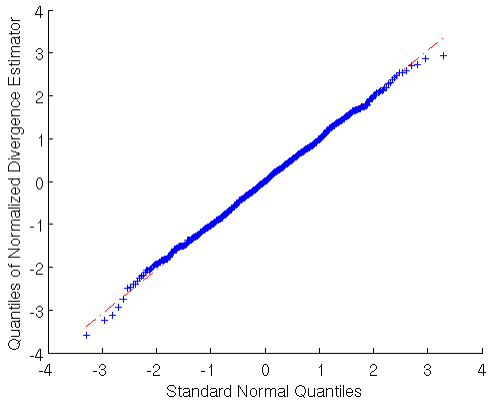

Figure 1: Q-Q plot comparing quantiles from the normalized weighted ensemble estimator of the KL divergence (vertical axis) to the quantiles from the standard normal distribution (horizontal axis). The red line shows . The linearity of the Q-Q plot points validates the central limit theorem, Theorem. 2, for the estimator.

4 Experiments

We first apply the weighted ensemble estimator of divergence to simulated data to verify the central limit theorem. We then use the estimator to obtain confidence intervals on the error exponents of the Bayes probability of error for the Iris data set from the UCI machine learning repository [33, 34].

4.1 Simulation

To verify the central limit theorem of the ensemble method, we estimated the KL divergence between two truncated normal densities restricted to the unit cube. The densities have means $\bar{\mu}_1 = 0.7 * \bar{1}_d$, $\bar{\mu}_2 = 0.3 * \bar{1}_d$ and covariance matrices $\sigma_i I_d$ where $\sigma_1 = 0.1$, $\sigma_2 = 0.3$, $\bar{1}_d$ is a d-dimensional vector of ones, and I_d is a d-dimensional identity matrix. We show the Q-Q plot of the normalized optimally weighted ensemble estimator of the KL divergence with $d = 6$ and 1000 samples from each density in Fig. 1. The linear relationship between the quantiles of the normalized estimator and the standard normal distribution validates Theorem 2.

4.2 Probability of Error Estimation

Our ensemble divergence estimator can be used to estimate a bound on the Bayes probability of error [7]. Suppose we have two classes C_1 or C_2 and a random observation x. Let the *a priori* class probabilities be $w_1 = Pr(C_1) > 0$ and $w_2 = Pr(C_2) = 1 - w_1 > 0$. Then f_1 and f_2 are the densities corresponding to the classes C_1 and C_2, respectively. The Bayes decision rule classifies x as C_1 if and only if $w_1 f_1(x) > w_2 f_2(x)$. The Bayes error P_e^* is the minimum average probability of error and is equivalent to

$$\begin{aligned} P_e^* &= \int \min\left(Pr(C_1|x), Pr(C_2|x)\right) p(x) dx \\ &= \int \min\left(w_1 f_1(x), w_2 f_2(x)\right) dx, \end{aligned} \tag{7}$$

where $p(x) = w_1 f_1(x) + w_2 f_2(x)$. For a, $b > 0$, we have

$$\min(a, b) \leq a^\alpha b^{1-\alpha}, \; \forall \alpha \in (0, 1).$$

Replacing the minimum function in Eq. 7 with this bound gives

$$P_e^* \leq w_1^\alpha w_2^{1-\alpha} c_\alpha(f_1||f_2), \tag{8}$$

where $c_\alpha(f_1||f_2) = \int f_1^\alpha(x) f_2^{1-\alpha}(x) dx$ is the Chernoff α-coefficient. The Chernoff coefficient is found by choosing the value of α that minimizes the right hand side of Eq. 8:

$$c^*(f_1||f_2) = c_{\alpha^*}(f_1||f_2) = \min_{\alpha \in (0,1)} \int f_1^\alpha(x) f_2^{1-\alpha}(x) dx.$$

Thus if $\alpha^* = \arg\min_{\alpha \in (0,1)} c_\alpha(f_1||f_2)$, an upper bound on the Bayes error is

$$P_e^* \leq w_1^{\alpha^*} w_2^{1-\alpha^*} c^*(f_1||f_2). \tag{9}$$

	Setosa-Versicolor	Setosa-Virginica	Versicolor-Virginica
Estimated Confidence Interval	$(0, 0.0013)$	$(0, 0.0002)$	$(0, 0.0726)$
QDA Misclassification Rate	0	0	0.04

Table 1: Estimated 95% confidence intervals for the bound on the pairwise Bayes error and the misclassification rate of a QDA classifier with 5-fold cross validation applied to the Iris dataset. The right endpoint of the confidence intervals is nearly zero when comparing the Setosa class to the other two classes while the right endpoint is much higher when comparing the Versicolor and Virginica classes. This is consistent with the QDA performance and the fact that the Setosa class is linearly separable from the other two classes.

Equation 9 includes the form in Eq. 1 ($g(x) = x^\alpha$). Thus we can use the optimally weighted ensemble estimator described in Sec. 2 to estimate a bound on the Bayes error. In practice, we estimate $c_\alpha(f_1||f_2)$ for multiple values of α (e.g. $0.01, 0.02, \ldots, 0.99$) and choose the minimum.

We estimated a bound on the pairwise Bayes error between the three classes (Setosa, Versicolor, and Virginica) in the Iris data set [33, 34] and used bootstrapping to calculate confidence intervals. We compared the bounds to the performance of a quadratic discriminant analysis classifier (QDA) with 5-fold cross validation. The pairwise estimated 95% confidence intervals and the misclassification rates of the QDA are given in Table 1. Note that the right endpoint of the confidence interval is less than $1/50$ when comparing the Setosa class to either of the other two classes. This is consistent with the performance of the QDA and the fact that the Setosa class is linearly separable from the other two classes. In contrast, the right endpoint of the confidence interval is higher when comparing the Versicolor and Virginica classes which are not linearly separable. This is also consistent with the QDA performance. Thus the estimated bounds provide a measure of the relative difficulty of distinguishing between the classes, even though the small number of samples for each class (50) limits the accuracy of the estimated bounds.

5 Conclusion

In this paper, we established the asymptotic normality for a weighted ensemble estimator of f-divergence using d-dimensional truncated k-nn density estimators. To the best of our knowledge, this gives the first results on the asymptotic distribution of an f-divergence estimator with MSE convergence rate of $O\left(1/T\right)$ under the setting of a finite number of samples from two unknown, non-parametric distributions. Future work includes simplifying the constants in front of the convergence rates given in [1] for certain families of distributions, deriving Berry-Esseen bounds on the rate of distributional convergence, extending the central limit theorem to other divergence estimators, and deriving the nonasymptotic distribution of the estimator.

Acknowledgments

This work was partially supported by NSF grant CCF-1217880 and a NSF Graduate Research Fellowship to the first author under Grant No. F031543.

References

[1] K. R. Moon and A. O. Hero III, "Ensemble estimation of multivariate f-divergence," in *IEEE International Symposium on Information Theory*, pp. 356–360, 2014.

[2] K. Sricharan and A. O. Hero III, "Ensemble weighted kernel estimators for multivariate entropy estimation," in *Adv. Neural Inf. Process. Syst.*, pp. 575–583, 2012.

[3] K. Sricharan, D. Wei, and A. O. Hero III, "Ensemble estimators for multivariate entropy estimation," *IEEE Trans. on Inform. Theory*, vol. 59, no. 7, pp. 4374–4388, 2013.

[4] I. Csiszar, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[5] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[6] A. Rényi, "On measures of entropy and information," in *Fourth Berkeley Sympos. on Mathematical Statistics and Probability*, pp. 547–561, 1961.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2006.

[8] B. Póczos and J. G. Schneider, "On the estimation of alpha-divergences," in *International Conference on Artificial Intelligence and Statistics*, pp. 609–617, 2011.

[9] J. B. Oliva, B. Póczos, and J. Schneider, "Distribution to distribution regression," in *International Conference on Machine Learning*, pp. 1049–1057, 2013.

[10] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, 2003.

[11] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[12] B. Chai, D. Walther, D. Beck, and L. Fei-Fei, "Exploring functional connectivities of the human brain using multivariate information analysis," in *Adv. Neural Inf. Process. Syst.*, pp. 270–278, 2009.

[13] J. Lewi, R. Butera, and L. Paninski, "Real-time adaptive information-theoretic optimization of neurophysiology experiments," in *Adv. Neural Inf. Process. Syst.*, pp. 857–864, 2006.

[14] E. Schneidman, W. Bialek, and M. J. Berry, "An information theoretic approach to the functional classification of neurons," in *Adv. Neural Inf. Process. Syst.*, pp. 197–204, 2002.

[15] K. M. Carter, R. Raich, and A. O. Hero III, "On local intrinsic dimension estimation and its applications," *Signal Processing, IEEE Transactions on*, vol. 58, no. 2, pp. 650–663, 2010.

[16] A. O. Hero III, B. Ma, O. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *Signal Processing Magazine, IEEE*, vol. 19, no. 5, pp. 85–95, 2002.

[17] K. R. Moon and A. O. Hero III, "Ensemble estimation of multivariate f-divergence," *CoRR*, vol. abs/1404.6230, 2014.

[18] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[19] G. A. Darbellay, I. Vajda, *et al.*, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Inform. Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[20] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.

[21] J. Silva and S. S. Narayanan, "Information divergence estimation based on data-dependent partitions," *Journal of Statistical Planning and Inference*, vol. 140, no. 11, pp. 3180–3198, 2010.

[22] T. K. Le, "Information dependency: Strong consistency of Darbellay–Vajda partition estimators," *Journal of Statistical Planning and Inference*, vol. 143, no. 12, pp. 2089–2100, 2013.

[23] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[24] S. Singh and B. Póczos, "Generalized exponential concentration inequality for Rényi divergence estimation," in *International Conference on Machine Learning*, pp. 333–341, 2014.

[25] A. Krishnamurthy, K. Kandasamy, B. Póczos, and L. Wasserman, "Nonparametric estimation of Rényi divergence and friends," in *International Conference on Machine Learning*, vol. 32, 2014.

[26] A. Berlinet, L. Devroye, and L. Györfi, "Asymptotic normality of L1 error in density estimation," *Statistics*, vol. 26, pp. 329–343, 1995.

[27] A. Berlinet, L. Györfi, and I. Dénes, "Asymptotic normality of relative entropy in multivariate density estimation," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 41, pp. 3–27, 1997.

[28] P. J. Bickel and M. Rosenblatt, "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, pp. 1071–1095, 1973.

[29] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *The Annals of Mathematical Statistics*, pp. 1049–1051, 1965.

[30] K. Sricharan, R. Raich, and A. O. Hero III, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4135–4159, 2012.

[31] K. Sricharan, *Neighborhood graphs for estimation of density functionals*. PhD thesis, Univ. Michigan, 2012.

[32] J. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher, "Central limit theorems for interchangeable processes," *Canad. J. Math*, vol. 10, pp. 222–229, 1958.

[33] K. Bache and M. Lichman, "UCI machine learning repository," 2013.

[34] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.