Direct Estimation of Information Divergence Using Nearest Neighbor Ratios

Morteza Noshad¹, Kevin R. Moon², Salimeh Yasaee Sekeh¹ and Alfred O. Hero III¹,*

¹University of Michigan, Electrical Engineering and Computer Science, Ann Arbor, Michigan, U.S.A
²Yale University, Genetics and Applied Math Departments, New Haven, Connecticut, U.S.A

Abstract—We propose a direct estimation method for Rényi and f-divergence measures based on a new graph theoretical interpretation. Suppose that we are given two sample sets X and Y, respectively with N and M samples, where \( \eta := M/N \) is a constant value. Considering the k-nearest neighbor (k-NN) graph of Y in the joint data set \((X,Y)\), we show that the average powered ratio of the number of X points to the number of Y points among all k-NN points is proportional to Rényi divergence of X and Y densities. A similar method can also be used to estimate f-divergence measures. We derive bias and variance rates, and show that for the class of \( \gamma \)-Hölder smooth functions, the estimator achieves the MSE rate of \( O\left(\frac{N^{\gamma/2}/(\gamma+d)}{N}\right) \). Furthermore, by using a weighted ensemble estimation technique, for density functions with continuous and bounded derivatives of up to the order d, and some extra conditions at the support set boundary, we derive an ensemble estimator that achieves the parametric MSE rate of \( O(1/N) \). Our estimator requires no boundary correction, and remarkably, the boundary issues do not show up. Our approach is also more computationally tractable than other competing estimators, which makes them appealing in many practical applications.

I. INTRODUCTION

Shannon entropy, mutual information, and the Kullback-Leibler (KL) divergence are major information theoretic measures. Shannon entropy can measure diversity or uncertainty of samples, while KL-divergence is a measure of dissimilarity, and mutual information is a measure of dependency between two probability distributions [1]. Rényi proposed a divergence measure which generalizes KL-divergence [2]. F-divergence is another general family which is also well studied, and comprises many important divergence measures such as KL-divergence, total variation distance, and \( \alpha \)-divergence [3]. These measures have wide range of applications in information and coding theory, statistics and machine learning [1], [4], [5].

A major class of estimators for these measures is called non-parametric, for which minimal assumptions on the density functions are considered in contrast to parametric estimators. An approach used for this class is plug-in estimation, in which we find an estimate of a distribution function and then plug it in the measure function. k-Nearest Neighbor (K-NN) and Kernel Density Estimator (KDE) methods are examples of this approach. Another approach is direct estimation, in which we find a relationship between the measure function and a functional in Euclidean space. In a seminal work in 1959, Beardwood et al derived the asymptotic behavior of the weighted functional of minimal graphs such as K-NN and TSP of N i.i.d random points [6]. They showed that the sum of weighted edges of these graphs converges to the integral of a weighted density function, which can be interpreted as Rényi entropy. Since then, this work has been of great interest in signal processing and machine learning communities. Yet the extension to Rényi divergence and f-divergences has remained an open question. Moreover, among various estimators of information measures, developing accurate and computationally tractable approaches has been often a challenge. Therefore, for practical and computational reasons, direct graphical algorithms have been under attention in the literature including this work.

In this work, we propose an estimation method for Rényi and f-divergences based on a direct graph estimation method. We show that given two sample sets X and Y with respective densities of \( f_1 \) and \( f_2 \), and the k-nearest neighbor (k-NN) graph of Y in the joint data set \((X,Y)\), the average powered ratio of the number of X points to the number of Y points among all k-NN points converges to the Rényi divergence. Using this fact, we design a consistent estimator for the Rényi and f-divergences.

Unlike most distance-based divergence estimators, our proposed estimator can use non-Euclidean metrics, which makes this estimator appealing in many information theoretic and machine learning applications. Our estimator requires no boundary correction, and surprisingly, the boundary issues do not show up. This is because the proposed estimator automatically cancels the extra bias of the boundary points in the ratio of nearest neighbor points. Our approach is also more computationally tractable than other estimators, with a time complexity of \( O(kN \log N) \), required to construct the k-NN graph [7]. For example for \( k = N^{\gamma/(d+1)} \) we get the complexity of \( O(N^{(d+2)/(d+1)} \log N) \). We show that for the class of \( \gamma \)-Hölder smooth functions, the estimator achieves the MSE rate of \( O(N^{(-\gamma)/(\gamma+d)}) \). Furthermore, by using the theory of optimally weighted ensemble estimation [5], [8], for density functions with continuous and bounded derivatives of up to the order d, and some extra conditions at the support set boundary, we derive an ensemble estimator that achieves the optimal MSE rate of \( O(1/N) \), which is independent of the dimension. Finally, the current work is an important step towards extending the direct estimation method studied in [9].

*This research was partially supported by ARO grant W911NF-15-1-0479.
to more general information theoretic measures.

Several previous works have investigated the existence of a particular type of divergence measures. k-NN [10], KDE [11], and histogram [12] estimators are among the studied plug-in estimators for the f-divergence family. In general, most of these estimators suffer from several restrictions such as lack of analytic convergence rates, or high computational complexity.

Recent works have focused on the MSE convergence rates for plug-in divergence estimators, such as KDE. Singh and Póczos proposed estimators for general density functionals and Rényi divergence, based on the kernel density plug-in estimator [11], which can achieve the convergence rate of $O(1/N)$ when the densities are at least $d/2$ times differentiable. In a similar approach, Kandasamy et al proposed another KDE-based estimator for general density functionals and divergence measures, which can achieve the convergence rate of $O(1/N)$ when the densities are at least $d/2$ differentiable [13].

Moon et al proposed simple kernel density plug-in estimators using weighted ensemble methods to improve the rate [8]. The proposed estimator can achieve the convergence rate when the densities are at least $(d + 1)/2$ times differentiable. The main drawback of these estimators is handling the bias at the support set boundary. For example, using the estimators proposed in [11], [13] requires knowledge of the densities’ support set and numerous computations at the support boundary, which become complicated when the dimension increases. To circumvent this issue, Moon et al [8] assumed smoothness conditions at the support set boundary, which may not always be true in practice. In contrast, our basic estimator does not require any smoothness assumptions on the support set boundary although our ensemble estimator does. Regarding the algorithm time complexities, our estimator spends $O(kN \log N)$ time versus the time complexity of KDE-based estimators which spend $O(N^2)$ time.

A rather different method for estimating f-divergences is suggested by Nguyen et al [14], which is based on a variational representation of f-divergences that connects the estimation problem to a convex risk minimization problem. This approach achieves the parametric rate of $O(1/N)$ when the likelihood ratio is at least $d/2$ times differentiable. However, the algorithm’s time complexity is even worse than $O(N^2)$.

II. A DIRECT ESTIMATOR OF DIVERGENCE MEASURES

In this section, we first introduce the Rényi and f-divergence measures. Then we propose an estimator based on a graph theoretical interpretation, and we outline our main theoretical results, which will be proven in section III.

Consider two density functions $f_1$ and $f_2$ with support $\mathcal{M} \subseteq \mathbb{R}^d$. The Rényi divergence between $f_1$ and $f_2$ is

$$D_\alpha (f_1(x)||f_2(x)) := \frac{1}{\alpha - 1} \log \int f_1(x)^\alpha f_2(x)^{1-\alpha} dx,$$

where in the second line, $J_\alpha (f_1, f_2)$ is defined as $J_\alpha (f_1, f_2) := \mathbb{E}_{f_2} \left[ \frac{f_1(x)}{f_2(x)} \right]^\alpha$.

Another general divergence family, f-divergence, is also defined as follows [3]:

$$D_f (f_1(x)||f_2(x)) := \int g \left( \frac{f_1(x)}{f_2(x)} \right) f_2(x) dx$$

where $g$ is a smooth and convex function such that $g(1) = 0$. KL-divergence, Hellinger distance and total variation distance are particular cases of this family. Note that for our approach, we only assume that $g$ is smooth.

We assume that the densities are lower bounded by $C_2 > 0$ and upper bounded by $C_1$. Also $f_1$ and $f_2$ belong to Hölder smoothness class with parameter $\gamma$.

**Definition** Given a support $\mathcal{X} \subseteq \mathbb{R}^d$, a function $f : \mathcal{X} \to \mathbb{R}$ is called Hölder continuous with parameter $0 < \gamma \leq 1$, if there exists a positive constant $G_f$, depending on $f$, such that

$$|f(y) - f(x)| \leq G_f \|y - x\|^{\gamma},$$

for every $x \neq y \in \mathcal{X}$.

The function $g(x)$ in (2) is also assumed to be Lipschitz continuous; i.e. $g$ is Hölder continuous with $\gamma = 1$.

**Remark 1:** $\gamma$-Hölder smoothness family comprises a large class of continuous functions including continuously differentiable functions and Lipschitz continuous functions. Also note that for $\gamma > 1$, any $\gamma$-Hölder continuous function on any bounded and continuous support is constant.

**Nearest Neighbor Ratio (NNR) Estimator:** Consider the i.i.d samples $X = \{X_1, ..., X_N\}$ drawn from $f_1$ and $Y = \{Y_1, ..., Y_M\}$ drawn from $f_2$. We define the set $Z := X \cup Y$, and consider the $k$-NN points for each of the points $Y_i$ in the set $Y$, which is represented by $Q_k(Y_i)$. Let $N_i$ and $M_i$ be the number of points of the sets $X$ and $Y$ among the $k$-NN points of $Y_i$, respectively. Then an estimator for Rényi divergence is

$$\hat{D}_\alpha (X, Y) := \frac{1}{\alpha - 1} \log \left[ \frac{\eta^\alpha}{M} \sum_{i=1}^{M} \left( \frac{N_i}{M_i + 1} \right)^\alpha \right] \eta^\alpha,$$

where $\eta := M/N$. Similarly, using the alternative form in (1), we have

$$\hat{J}_\alpha (X, Y) := \eta^\alpha \frac{M}{M} \sum_{i=1}^{M} \left( \frac{N_i}{M_i + 1} \right)^\alpha.$$

Note that the estimator defined in (4) can be negative and unstable in extreme cases. To correct this, we propose the NNR estimator for Rényi divergence denoted by $\hat{D}_\alpha (X, Y)$:

$$\min \left\{ \max \left\{ \hat{D}_\alpha (X, Y), 0 \right\}, \frac{1}{1 - \alpha} \log \left( \frac{C_{\alpha}}{C_{\alpha}} \right) \right\}.$$

The NNR f-divergence estimator is defined as

$$\hat{D}_f (X, Y) := \max \left\{ \frac{1}{M} \sum_{i=1}^{M} \left( \eta^\alpha N_i (M_i + 1)^{\alpha} \right), 0 \right\},$$

for every $x \neq y \in \mathcal{X}$.
where \( \tilde{g}(x) := \max \{g(x), g(C_x/C_U)\} \).

The intuition behind the proposed estimators is that the ratio \( \frac{N}{N+1} \) can be considered an estimate of density ratios at \( Y \). Note that if the densities \( f_1 \) and \( f_2 \) are almost equal, then for each point \( Y \), \( N_i \approx M_i + 1 \), and therefore both \( \hat{D}_N(X, Y) \) and \( \tilde{D}_N(X, Y) \) tend to zero. In the following theorems we derive upper bounds on the bias and variance rates. Consider the bias and variance definitions as \( \mathbb{B}[T] = \mathbb{E}[T] - T \) and \( \mathbb{V}[T] = \mathbb{E}[T^2] - \mathbb{E}[T]^2 \), respectively, where \( T \) is an estimator of the parameter \( \theta \).

**Theorem 2.1:** The bias of NNR estimator for Rényi divergence, defined in (6), can be bounded as

\[
\mathbb{B}[\hat{D}_N(X, Y)] = O\left(\left(\frac{k}{N}\right)^{\gamma/d}\right) + O\left(\frac{1}{k}\right). \tag{8}
\]

Here \( \gamma \) is the Hölder smoothness parameter.

**Theorem 2.2:** The variance of the NNR estimator is

\[
\mathbb{V}[\hat{D}_N(X, Y)] \leq O\left(\frac{1}{N}\right) + O\left(\frac{1}{M}\right). \tag{9}
\]

**Remark 2:** The same variance bound holds true for the RV \( \hat{J}_N(X, Y) \). Also bias and variance results easily extend to the \( f \)-divergence estimator.

**Remark 3:** Note that in most cases, the \( 1/k \) term in (8) is the dominant error term, and in order to have an asymptotically unbiased NNR estimator, \( k \) should be a growing function of \( N \). The \( 1/k \) term actually comes from the error of Poissonization technique used in the proof. By equating the terms \( O(k/N)^{\gamma/d} \) and \( O(1/k) \), it turns out that for \( k_{opt} = O\left(\frac{N}{\pi^{1/2}}\right) \), we get the optimal MSE rate of \( O\left(\frac{1}{N}\right) \). The optimal choice for \( k \) can be compared to the optimum value \( k = O\left(\frac{1}{N}\right) \) in [4], where a plug-in KNN estimator is used. Also considering the computational complexity of \( O(kN\log N) \) to construct the \( k \)-NN graph [7], we see that there is a trade-off between MSE rate and complexity for different values of \( k \). In the particular case of optimal MSE, the computational complexity of this method is \( O\left(\frac{2kN}{\pi^{1/2}}\log N\right) \).

**Algorithm 1: NNR Estimator of Rényi Divergence**

**Input:** Data sets \( X = \{X_1, ..., X_N\} \), \( Y = \{Y_1, ..., Y_M\} \)

1. \( Z \leftarrow X \cup Y \)
2. For each point \( Y_i \) in \( Y \) do  
   3. \( S_i \leftarrow \{Q_1(Y_i), ..., Q_k(Y_i)\} \)
   4. \( R_i \leftarrow [S_i \cap X]/[S_i \land Y] \)
   5. \( \hat{D} \leftarrow 1/(\alpha - 1) \log [\gamma^q \sum R_i^q] / M \)

**Output:** \( \hat{D} \)

Under extra conditions on the densities and support set boundary, we can improve the bias rate by applying the ensemble theory in [5], [8]. Assume that the density functions are in the Hölder space \( \Sigma(\gamma, L) \), which consists of functions on \( X \) continuous derivatives up to order \( q = \left\lceil \gamma \right\rceil \geq d \) and the \( q \)th partial derivatives are Hölder continuous with exponent \( \gamma' = \gamma - q \). We also assume that the density derivatives up to order \( d \) vanish at the boundary. Let \( L := \{l_1, ..., l_r\} \) be a set of index values with \( l_i < c \), where \( \kappa = [c\sqrt{N}] \). Let \( k(l) := \left\lceil l\sqrt{N} \right\rceil \). The weighted ensemble estimator is defined as \( \hat{D}_w := \sum_{l \in L} w(l)\hat{D}_{k(l)} \), where \( \hat{D}_{k(l)} \) is the NNR estimator of Rényi \( \alpha \)-divergence, using the \( k(l) \)-NN graph.

**Theorem 2.3:** Let \( L > d \) and \( w_0 \) be the solution to:

\[
\min_w \|w\|_2 \quad \text{subject to} \quad \sum_{l \in L} w(l) = 1, \quad \sum_{l \in L} w(l)l^d = 0, \quad i \in N, \quad i \leq d. \tag{10}
\]

Then the MSE rate of the ensemble estimator \( \hat{D}_{w_0} \) is \( O(1/N) \).

**III. PROOF**

In this section we derive the bias terms of NNR estimator. The variance bound for NNR estimator is more straightforward and can be derived using Efron-Stein inequality. Also for proving the MSE rate of ensemble variant of the NNR estimator, we need more accurate bias rates, which is provided in the arXiv version. So, for variance and ensemble estimation proofs we refer the reader to the Appendix section of arXiv version of the paper. First, we provide a smoothness lemma for the densities. Unless stated otherwise, all proofs of lemmas are provided in the arXiv version.

**Lemma 3.1:** Suppose that the density function \( f(x) \) belongs to the Hölder smoothness class. Then if \( B(x, r) \) denotes the sphere with center \( x \) and radius \( r = \rho_k(x) \), where \( \rho_k(x) \) is defined as the \( k \)-NN distance on the point \( x \), we have the following smoothness condition:

\[
E_{\rho_k(x)} \left[ \sup_{y \in B(x, \rho_k(x))} |f(y) - f(x)| \right] \leq \epsilon_{\gamma, k}, \tag{11}
\]

where \( O((k/N)^{\gamma/d} + O(C(k))) \), and we have \( C(k) := \exp(-(3k)^{-2}) \) for a fixed \( \delta \in (2/3, 1) \).

We first state the bias proof for Rényi divergence, and then we extend the method to \( f \)-divergence. It is easier to work with \( \hat{J}_N(X, Y) \) defined in (5), instead of \( \hat{D}_N(X, Y) \). The following lemma provides the essential tool to make a relation between \( \mathbb{B}[\hat{D}] \) and \( \mathbb{B}[\hat{J}] \).

**Lemma 3.2:** Assume that \( g(x) : X \to \mathbb{R} \) is Lipschitz continuous with constant \( H_g > 0 \). If \( \hat{T} \) is a RV estimating a constant \( T \) with the bias \( \mathbb{B}[\hat{T}] \) and the variance \( \mathbb{V}[\hat{T}] \), then the bias of \( g(\hat{T}) \) can be upper bounded by

\[
\mathbb{E}\left[|g(\hat{T}) - g(T)|\right] \leq H_g \left( \sqrt{\mathbb{V}\left[\hat{T}\right]} + \mathbb{B}[\hat{T}] \right). \tag{12}
\]

An immediate consequence of this lemma is

\[
\mathbb{B}[\hat{D}_N(X, Y)] \leq C \left[ \mathbb{B}[\hat{J}_N(X, Y)] + \sqrt{\mathbb{V}[\hat{J}_N(X, Y)]} \right]. \tag{13}
\]
where \( C \) is a constant.

From theorem 2.2, \( \sqrt{\mathbb{E} [\tilde{J}_n(X,Y)]} = O(1/N) \), so we only need to bound \( \mathbb{E} [\tilde{J}_n(X,Y)] \). If \( \eta := M/N \), we have:

\[
\mathbb{E} [\tilde{J}_n(X,Y)] = \eta^\alpha \mathbb{E} \left[ \sum_{i=1}^M \left( \frac{N_i}{M_i + 1} \right)^\alpha \right]
\]

\[
= \eta^\alpha \mathbb{E} \left[ \sum_{i=1}^M \frac{N_i}{M_i + 1} \right] Y_i \quad \text{(14)}
\]

Now note that \( N_1 \) and \( M_1 \) are not independent since \( N_1 + M_1 = k \). We use the Poissonizing technique [15] [16] and assume that \( N_1 + M_1 = K \), where \( K \) is a Poisson random variable with mean \( k \). We represent the Poissonized variant of \( \tilde{J}_n(X,Y) \) by \( \mathcal{J}_n(X,Y) \), and we will show that

\[
\mathbb{E} [\mathcal{J}_n(X,Y)] = \mathbb{E} [\tilde{J}_n(X,Y)] + O(1/k).
\]

By partitioning theorem for a Poisson random variable with Bernoulli trials of probabilities \( \Pr(Q_1(Y_1) \in X) \) and \( \Pr(Q_1(Y_1) \in Y) \), we argue that \( N_1 \) and \( M_1 \) are two independent Poisson RVs. We first compute \( \Pr(Q_k(Y_1) \in X) \) and \( \Pr(Q_k(Y_1) \in Y) \) as follows:

**Lemma 3.3:** Let \( \eta := M/N \). The probability that the point \( Q_k(Y_1) \) respectively belongs to the sets \( X \) and \( Y \) is equal to

\[
\Pr(Q_k(Y_1) \in X) = \frac{f_1(Y_1)}{f_1(Y_1) + \eta f_2(Y_1)} + O(\epsilon_{\gamma,k}).
\]

\[
\Pr(Q_k(Y_1) \in Y) = \frac{\eta f_2(Y_1)}{f_1(Y_1) + \eta f_2(Y_1)} + O(\epsilon_{\gamma,k}).
\]

(15)

Using the conditional independence of \( N_1 \) and \( M_1 \) we write

\[
\mathbb{E} \left[ \frac{N_1}{M_1 + 1} \right] Y_1 = \mathbb{E} [N_1|Y_1] \mathbb{E} \left[ (M_1 + 1)^{-1} | Y_1 \right].
\]

(16)

\[
\mathbb{E} [N_1|Y_1] = \sum_{i=1}^k \Pr(Q_i(Y_1) \in X)
\]

\[
= \frac{k f_1(Y_1)}{f_1(Y_1) + \eta f_2(Y_1)} + O(k\epsilon_{\gamma,k}).
\]

(17)

Also similarly,

\[
\mathbb{E} [M_1|Y_1] = \frac{k f_2(Y_1)}{f_1(Y_1) + \eta f_2(Y_1)} + O(k\epsilon_{\gamma,k}).
\]

**Lemma 3.4:** If \( U \) is a Poisson random variable with the mean \( \lambda > 1 \), then

\[
\mathbb{E} \left[ (U + 1)^{-1} \right] = \frac{1}{\lambda} (1 - e^{-\lambda}).
\]

(18)

Using this lemma for \( M_1 \) yields

\[
\mathbb{E} \left[ (M_1 + 1)^{-1} | Y_1 \right] = \frac{k^{-1}}{\left( f_1(Y_1) + \eta f_2(Y_1) \right)^{-1}} + O\left( \frac{e^{-vk}}{k} \right),
\]

(19)

here \( v \) is some positive constant. Therefore, (16) becomes

\[
\mathbb{E} \left[ \frac{N_1}{M_1 + 1} | Y_1 \right] = \frac{f_1(Y_1)}{\eta f_2(Y_1)} + O(\epsilon_{\gamma,k}) + O(e^{-vk}).
\]

(20)

Using lemma 3.2 and theorem 2.2, we obtain

\[
\mathbb{E} \left[ \left( \frac{N_1}{M_1 + 1} \right)^\alpha | Y_1 \right] = \eta^\alpha \left( \frac{f_1(Y_1)}{f_2(Y_1)} \right)^\alpha + O(\epsilon_{\gamma,k}) + O(e^{-vk}) + O(N^{-\frac{1}{2}}).
\]

(21)

By applying an equation similar to (14), we get

\[
\mathbb{E} \left[ \mathcal{J}_n(X,Y) \right] = O(\gamma,k) + O(e^{-vk}) + O(N^{-\frac{1}{2}}).
\]

(22)

**Lemma 3.5:** De-Poissonizing \( \mathcal{J}_n(X,Y) \) adds \( O(\frac{1}{k}) \) error:

\[
\mathbb{E} \left[ \tilde{J}_n(X,Y) \right] = \mathbb{E} \left[ \mathcal{J}_n(X,Y) \right] + O(1/k).
\]

(23)

At this point the bias proof of NNR estimator for Rényi divergence is complete, and since \( O(e^{-vk}) \) and \( O(N^{-\frac{1}{2}}) \) are of higher order compared to \( O(\epsilon_{\gamma,k}) \), we obtain the final bias rate in (8). The bias proof of NNR estimator for f-divergence is similar, and by using the lemma 3.2 for \( \alpha \), we can follow the same steps to prove the bias bound. The complete proof is provided in the arXiv version.

IV. NUMERICAL RESULTS

In this section we provide numerical results to show the consistency of the proposed estimator and compare the estimation quality in terms of different parameters such as \( N \) and \( k \). In our experiments, we choose i.i.d samples for \( X \) and \( Y \) from different independent distributions such as Gaussian, truncated Gaussian and uniform functions.

The first experiment, shown in Figure 1, shows the mean estimated KL-divergence as \( N \) grows for \( k \) equal to 20, 40, 60. The divergence measure is between a 2D Gaussian RV with mean \( [0,0] \) and variance of \( 2I_2 \), and a uniform distribution with \( x, y \in [-1,1] \). For each case we repeat the experiment 100 times, and compute the mean of the estimated value and the standard deviation error bars. For small sample sizes, smaller \( k \) results in smaller bias error, which is due to the \( (\frac{1}{k})^{1/d} \) bias term. As \( N \) grows, we get larger bias for small values of \( k \), which is due to the fact that the \( (1/k) \) term dominates. If we compare the standard deviations for different values of \( k \) at \( N = 4000 \), they are almost equal, which verifies the fact that variance is independent of \( k \).
normal RVs. The RVs are 2D with means $\mu$ fixed so that the same mean and different variances ($\sigma^2_{\alpha}$) in Definition in (1). The graph shows the MSE for Rényi divergence with $\alpha = 0.5$ and different variances of $\sigma^2_1 = I_2$, $\sigma^2_2 = 3I_2$ using NNR, KDE and KNN estimators.

**V. CONCLUSION**

In this paper we proposed a direct estimation method for Rényi and f-divergence measures based on a new graph theoretical interpretation. We proved bias and variance convergence rates, and validated our results by numerical experiments. Direct estimation procedures that converge for a fixed number $k$ of nearest neighbors is a worthwhile topic for future work.

**REFERENCES**

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[2] A. Rényi, “On measures of entropy and information,” in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 547–561, University of California Press, 1961.
[3] S. M. Ali and S. D. Silvey, “A general class of coefficients of divergence of one distribution from another,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.
[4] K. R. Moorman and A. O. Hero, “Ensemble estimation of multivariate f-divergence,” in *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 356–360, IEEE, 2014.
[5] K. R. Moorman, M. Noshad, S. Y. Sekeh, and A. O. Hero III, “Information theoretic structure learning with confidence,” in *Proc IEEE Int Conf Acoust Speech Signal Process*, 2017.
[6] J. Beaudwood, J. H. Halton, and J. M. Hammersley, “The shortest path through many points,” in *Math Proc Cambridge*, vol. 55, pp. 299–327, Cambridge Univ Press, 1959.
[7] P. M. Vaidya, “An $o (n \log n)$ algorithm for the all-nearest-neighbors problem,” *Discrete & Computational Geometry*, vol. 4, no. 1, pp. 101–115, 1989.
[8] K. R. Moorman, K. Sricharan, K. Greenewald, and A. O. Hero, “Improving convergence of divergence functional ensemble estimators,” in *IEEE International Symposium Inf Theory*, pp. 1133–1137, IEEE, 2016.
[9] F. J. Y. Yukich, *Probability theory of classical Euclidean optimization problems*, 1998.
[10] B. Póczos and J. G. Schneider, “On the estimation of alpha-divergences,” in *AISTATS*, pp. 609–617, 2011.
[11] S. Singh and B. Póczos, “Exponential concentration of a density functional estimator,” in *Advances in Neural Information Processing Systems*, pp. 3032–3040, 2014.
[12] Q. Wang, S. R. Kulkarni, and S. Verdú, “Divergence estimation for multidimensional densities via nearest-neighbor distances,” *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.
[13] K. Kandasamy, A. Krishnamurthy, B. Póczos, L. Wasserman, et al., “Nonparametric Von Mises estimators for entropies, divergences and mutual informations,” in *NIPS*, pp. 397–405, 2015.
[14] X. Nguyen, M. J. Wainwright, and M. I. Jordan, “Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization,” in *NIPS*, pp. 1089–1096, 2007.
[15] A. D. Barbour, L. Holst, and S. Janson, *Poisson approximation*. Claren- don Press Oxford, 1992.
[16] P. Jacquet and W. Szpankowski, “Analytical depoisonization and its applications,” *Theor Comput Sci*, vol. 201, no. 1, pp. 1–62, 1998.