Scalable Hash-Based Estimation of Divergence Measures

Morteza Noshad and Alfred O. Hero III
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48105, USA
{noshad,hero} @umich.edu

Abstract

We propose a scalable divergence estimation method based on hashing. Consider two continuous random variables X and Y whose densities have bounded support. We consider a particular locality sensitive random hashing, and consider the ratio of samples in each hash bin having non-zero numbers of Y samples. We prove that the weighted average of these ratios over all of the hash bins converges to f-divergences between the two samples sets. We derive the MSE rates for two families of smooth functions; the Hölder smoothness class and differentiable functions. In particular, it is proved that if the density functions have bounded derivatives up to the order d, where d is the dimension of samples, the optimal parametric MSE rate of $O(1/N)$ can be achieved. The computational complexity is shown to be $O(N)$, which is optimal. To the best of our knowledge, this is the first empirical divergence estimator that has optimal computational complexity and can achieve the optimal parametric MSE estimation rate of $O(1/N)$.

1 Introduction

Information theoretic measures such as Shannon entropy, mutual information, and the Kullback-Leibler (KL) divergence have a broad range of applications in information theory, statistics and machine learning [1–3]. When we have two or more data sets and we are interested in finding the dependence or dissimilarity between them, Shannon mutual information or KL-divergence is often used. Rényi and f-divergence measures are two well studied generalizations of KL-divergence which comprise many important divergence measures such as KL-divergence, total variation distance, and α-divergence [4, 5].

Non-parametric estimators are a major class of divergence estimators, for which minimal assumptions on the density functions are considered. Some of the non-parametric divergence estimators are based on density plug-in estimators such as k-NN [6], KDE [7], and histogram [8]. A few researchers, on the other hand, have proposed direct estimation methods such as graph theoretic nearest neighbor ratio (NNR) [9]. In general, plug-in estimation methods suffer from high computational complexity, which make them unsuitable for large scale applications.

Recent advances on non-parametric divergence estimation have been focused on the MSE convergence rates of the estimator. Singh et al in [7] proposed a plug-in KDE estimator for Rényi divergence that achieves the MSE rate of $O(1/N)$ when the densities are at least d times differentiable, and the support boundaries are sufficiently smooth. Similar plug-in KDE based estimators were proposed in [10] and [11] that can achieve the optimal MSE rate respectively for the densities that are at least $d/2$ and $d/4$ times differentiable. Moon et al proposed a weighted ensemble method to improve the MSE rate of plug-in KDE estimators [12]. The proposed estimator for f-divergence achieves the optimal MSE rate when the densities are at least $(d+1)/2$ times differentiable. They also assume stringent smoothness conditions at the support set boundary.

Noshad et al proposed a graph theoretic direct estimation method based on nearest neighbor ratios (NNR) [9]. Their estimator is simple and computationally more tractable than other competing estimators, and can achieve the optimal MSE rate of $O(1/N)$ for densities that are at least d times differentiable. Although their basic estimator does not require any smoothness assumptions on the support set boundary, the ensemble

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

estimator variant of their estimator does.

In spite of achieving the optimal theoretical MSE rate by aforementioned estimators, there remain serious. The first challenge is the high computational complexity of the estimator. Most KDE based estimators require runtime complexity of $O(N^2)$, which is not suitable for large scale applications. The NNR estimator proposed in [9] has the runtime complexity of $O(kN \log N)$, which is faster than the previous estimators. However, in [9] they require k to grow sub-linearly with N, which results in much higher complexity than linear runtime complexity. The other issue is the smoothness assumptions made on the support set boundary. Almost all previously proposed estimators assume extra smoothness conditions on the boundaries, which may not hold practical applications. For example, the method proposed in [7] assumes that the density derivatives up to order d vanish at the boundary. Also it requires numerous computations at the support boundary, which become complicated when the dimension increases. The Ensemble NNR estimator in [9] assumes that the density derivatives vanish at the boundary. To circumvent this issue, Moon et al [12] assumed smoothness conditions at the support set boundary. However, these conditions may not hold in practice.

In this paper we propose a low complexity divergence estimator that can achieve the optimal MSE rate of $O(1/N)$ for the densities with bounded derivatives of up to d. Our estimator has optimal runtime complexity of $O(N)$, which makes it an appropriate tool for large scale applications. Also in contrast to other competing estimators, our estimator does not require stringent smoothness assumptions on the support set boundary.

The structure of the proposed estimator borrows ideas from hash based methods for KNN search and graph constructions problems [13, 14], as well as from the NNR estimator proposed in [9]. The advantage of hash based methods is that they can be used to find the approximate nearest neighbor points with lower complexity as compared to the exact k-NN search methods. This suggests that fast and accurate algorithms for divergence estimation may be derived from hashing approximations of k-NN search. Noshad et al [9] consider the k-NN graph of Y in the joint data set (X, Y), and show that the average exponentiated ratio of the number of X points to the number of Y points among all k-NN points is proportional to the Rényi divergence between the X and Y densities. It turns out that for estimation of the density ratio around each point we really do not need to find the exact k-NN points, but only need sufficient local samples from X and Y around each point. By using a randomized locality sensitive hashing (LSH), we find the closest points in Euclidean space. In this manner, applying ideas from the NNR estimation and hashing techniques to KNN search problem, we obtain a more efficient divergence estimator. Consider two sample sets X and Y with a bounded density support. We use a particular two-level locality sensitive random hashing, and consider the ratio of samples in each bin with a number of Y samples. We prove that the weighted average of these ratios over all of the bins can be made to converge almost surely to f-divergences between the two samples populations. We also argue that using the ensemble estimation technique provided in [2], we can achieve the optimal parametric rate of $O(1/N)$. Furthermore, using a simple algorithm for online estimation method has $O(N)$ complexity and $O(1/N)$ convergence rate, which is the first optimal online estimator of its type.

The rest of the paper is organized as follows. In Section 2, we recall the definition of f-divergence and introduce the Hash-Based (HB) estimator. In Section 3, we provide the convergence theorems and propose the Ensemble Hash-Based (EHB) estimator. In Section 4, we propose the online version of the proposed HB and EHB estimator. In Section 5 we give proofs for the convergence results. Finally, in Section 6 we validate our theoretical results using numerical and real data experiments.

2 Hash-Based Estimation

In this section, we first introduce the f-divergence measure and propose a hash-based estimator.

Consider two density functions f_1 and f_2 with common bounded support set $\mathcal{X} \subseteq \mathbb{R}^d$.

The f-divergence is defined as follows [5].

$$D_g\left(f_1(x)||f_2(x)\right) := \int g\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x)dx = \mathbb{E}_{f_2}\left[g\left(\frac{f_1(x)}{f_2(x)}\right)\right], \quad (1)$$

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 estimation, we don't need convexity of g and $g(1) = 0$. Assume that the densities are lower bounded by $C_L > 0$ and upper bounded by C_U. Assume f_1 and f_2 belong to the Hölder smoothness class with parameter γ:

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, possibly depending on f, such that

$$|f(y) - f(x)| \leq G_f \|y - x\|^\gamma, \quad (2)$$

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

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

Remark 1 *The γ-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 γ–Hölder continuous function on any bounded and continuous support is constant.*

Hash-Based Divergence 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. Define the fraction $\eta := M/N$. We define the set $Z := X \cup Y$. We define a positive real valued constant ϵ as a user-selectable parameter of the estimator to be defined in 5. We define the hash function $H_1 : \mathbb{R}^d \to \mathbb{Z}^d$ as

$$H_1(x) = [h_1(x_1), h_1(x_2), ..., h_1(x_d)], \tag{3}$$

where x_i is the projection of x on the ith coordinate, and $h_1(x) : \mathbb{R} \to \mathbb{Z}$ is defined as

$$h_1(x) = \left\lfloor \frac{x+b}{\epsilon} \right\rfloor, \tag{4}$$

for fixed b. Let $\mathcal{F} := \{1, 2, .., F\}$, where $F := c_H N$ and c_H is a fixed real number. We define a random hash function $H_2 : \mathbb{Z}^d \to \mathcal{F}$ with a uniform density on the output and consider the combined hashing $H(x) := H_2(H_1(x))$, which maps the points in $\mathbb{R}^d$ to $\mathcal{F}$.

Consider the mappings of the sets X and Y using the hash function $H(x)$, and define the vectors $\mathcal{N}$ and $\mathcal{M}$ to respectively contain the number of collisions for each output bucket from the set $\mathcal{F}$. We represent the bins of the vectors $\mathcal{N}$ and $\mathcal{M}$ respectively by N_i and M_i, $1 \le i \le F$.

The hash based f-divergence estimator is defined as

$$\widehat{D}_g(X, Y) := \max\left\{ \frac{1}{M} \sum_{\substack{i \le F \\ M_i > 0}} M_i \widetilde{g}\left(\frac{\eta N_i}{M_i}\right), 0 \right\}, \tag{5}$$

where $\widetilde{g}(x) := \max\{g(x), g(C_L/C_U)\}$.

Note that if the densities f_1 and f_2 are almost equal, then for each point Y_i, $N_i \approx M_i$, and thus $\widehat{D}_g(X, Y)$ tends to zero, as required. Algorithm 1 shows the HB estimation procedure. We first find the sets of all hashed points in X and Y (lines 1 and 2). Then the number of collisions is counted (lines 3-5), and the divergence estimate is computed (line 6).

Similar to most of LSH structures, computing the hashing output in our estimator is of $O(1)$ complexity, and does not depend on ϵ. Thus, the computational complexity of Algorithm 1 is $O(M)$.

Algorithm 1: HB Estimator of f-Divergence

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

```
/* Find the sets of all hashed points in X
   and Y                                    */
```

1 $X' \leftarrow H(X)$.

2 $Y' \leftarrow H(Y)$.

3 **for** *each* $i \in \mathcal{F}$ **do**

```
   /* Find the number of collisions at bin i
      */
```

4 $N_i \leftarrow |X' = i|$

5 $M_i \leftarrow |Y' = i|$

6 $\widehat{D} \leftarrow \max\left\{\frac{1}{M}\sum_{M_i>0} M_i \widetilde{g}(\eta N_i/M_i), 0\right\}$,

Output : $\widehat{D}$

3 Convergence Results

In the following theorems we state upper bounds on the bias and variance rates. Let $\mathbb{B}[\hat{T}] = \mathbb{E}[\hat{T}] - T$ and $\mathbb{V}[\hat{T}] = \mathbb{E}[\hat{T}^2] - \mathbb{E}[\hat{T}]^2$, respectively represent the bias and variance of $\hat{T}$, which is an estimator of the parameter T. Then, the following provides a bound on the bias of the proposed estimator.

Theorem 3.1 *Assume that f_1 and f_2 are density functions with bounded common support set $\mathcal{X} \in \mathbb{R}^d$ and satisfying γ-Hölder smoothness. The bias of the proposed estimator for f-divergence with function g can be bounded as*

$$\mathbb{B}\Big[\widehat{D}_g(X, Y)\Big] = O(\epsilon^\gamma) + O\left(\frac{1}{N\epsilon^d}\right).$$

Remark 2 *In order for the estimator to be asymptotically unbiased, ϵ needs to be a function of N. The optimum bias rate of $O\Big(\big(\frac{1}{N}\big)^{\gamma/(\gamma+d)}\Big)$ can be achieved for $\epsilon = O\Big(\big(\frac{1}{N}\big)^{\gamma/(\gamma+d)}\Big)$.*

In the following we propose an upper bound on the variance that is independent of ϵ.

Theorem 3.2 *Let $\eta = M/N$ be fixed. The variance of the estimator* (5) *can be bounded as*

$$\mathbb{V}\Big[\widehat{D}_g(X, Y)\Big] \le O\left(\frac{1}{N}\right). \tag{6}$$

Remark 3 *The same variance bound holds for the random variable $\rho_i := \frac{N_i}{M_i}$. The bias and variance results easily extend to Rényi divergence estimation.*

We next show that, when f_1 and f_2 belong to the family of differentiable densities, we can improve the bias rate by applying the ensemble estimation approach in [3,12]. The EHB estimator is defined as follows.

Ensemble Hash-Based Estimator: Assume that the density functions have continuous derivatives up to order $q \geq d$. Let $\mathcal{T} := \{t_1, ..., t_T\}$ be a set of index values with $t_i < c$, where $c > 0$ is a constant. Let $\epsilon(t) := tN^{-1/2d}$. The weighted ensemble estimator is defined as

$$\widehat{D}_w := \sum_{t \in \mathcal{T}} w(t) \widehat{D}_{\epsilon(t)}, \tag{7}$$

where $\widehat{D}_{\epsilon(t)}$ is the hash based estimator of f-divergence, with the hashing parameter of $\epsilon(t)$. The following theorem states a sufficient condition for the weight vector w that ensures that the ensemble estimator (7) achieves an MSE rate of $O(1/N)$.

Theorem 3.3 *Let $T > d$ and w_0 be the solution to:*

$$\begin{aligned} \min_{w} \quad & \|w\|_2 \\ \textit{subject to} \quad & \sum_{t \in \mathcal{T}} w(t) = 1, \\ & \sum_{t \in \mathcal{T}} w(t) t^i = 0, i \in \mathbb{N}, i \leq d. \end{aligned} \tag{8}$$

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

4 Online Divergence Estimation

In this section we study the problem of online divergence estimation. In this setting we consider two data steams $X = \{X_1, X_2, ..., X_N\}$ and $Y = \{Y_1, Y_2, ..., Y_N\}$ with i.i.d samples, and we are interested in estimating the divergence between two data sets. The number of samples increase over time and an efficient update of the divergence estimate is desired. The time complexity of a batch update, which uses the entire update batch to compute the estimate at each time point, is $O(N)$, and it may not be so effective in cases which we need quick detection of any change in the divergence function.

Algorithm 2 updates the divergence with amortized runtime complexity of order $O(1)$. Define the sets $X^N := \{X_i\}_{i=1}^N$, $Y^N := \{Y_i\}_{i=1}^N$, the number of X and Y samples in each partition, and the divergence estimate between X^N and Y^N. Consider updating the estimator with new samples X_{N+1} and Y_{N+1}. In the first and second lines of algorithm 2, the new samples are added to the datasets and the values of N_i and M_i of the bins in which the new samples fall. We can find these bins in $O(1)$ using a simple hashing. Note that once N_i and M_i are updated, the divergence measure can be updated, but the number of bins is not increased, by Theorem 3.1, it is clear that the bias will not be reduced. Since increasing the number of bins requires recomputing the bin partitions, a brute force rebinning approach would have order $O(N)$ complexity, and it were updated N times, the total complexity would be $O(N^2)$. Here we use a trick and update the hash function only when $N + 1$ is a power of 2. In the following theorem, which is proved in appendix, we show that the MSE rate of this algorithm is order $O(1/N)$ and the total rebinngn computational complexity is order $O(N)$.

Theorem 4.1 *MSE rate of the online divergence estimator shown in Algorithm 2 is order $O(1/N)$ and the total computational complexity is order $O(N)$.*

Algorithm 2: Online Divergence Estimation

Input : $X^N := \{X_i\}_{i=1}^N, Y^N := \{Y_i\}_{i=1}^N$
$\widehat{D} = \widehat{D}\big(X^N, Y^N\big)$
(N_i, M_i)
(X_{N+1}, Y_{N+1})

1 Add X_{N+1} and Update N_k s.t $H(X_{N+1}) = k$.
2 Add Y_{N+1} and Update M_l s.t $H(Y_{N+1}) = l$.
3 If $N + 1 = 2^i$ for some i, Then
4 Update ϵ to the optimum value
5 Re-hash X and Y
6 Recompute N_i and M_i for $0 \leq i \leq F$
7 Update $\widehat{D}$

Output : $\widehat{D}$

5 Proofs

In this section we derive the bias bound for the densities in Hölder smoothness class, stated in Theorem 3.1. For the proofs of variance bound in Theorem 3.2, convergence rate of EHB estimator in Theorem 3.3, and online divergence estimator in Theorem 4.1, we refer the reader to the Appendix of the extended paper [15].

Consider the mapping of the X and Y points by the hash function H_1, and let the vectors $\{V_i\}_{i=1}^L$ represent the distinct mappings of X and Y points under H_1. Here L is the number of distinct outputs of H_1. In the following lemma we prove an upper bound on L.

Lemma 5.1 *Let $f(x)$ be a density function with bounded support $\mathbb{X} \subseteq \mathbb{R}^d$. Then if L denotes the number of distinct outputs of the hash function H_1 (defined in (3)) of i.i.d points with density $f(x)$, we have*

$$L \leq O\left(\frac{1}{\epsilon^d}\right). \tag{9}$$

Proof Let $x = [x_1, x_2, ..., x_d]$ and define $\mathcal{X}_I$ as the region defined as

$$\mathbb{X}_I := \{x | -c_X \leq x_i \leq c_X, 1 \leq i \leq d\}, \tag{10}$$

where c_X is a constant such that $\mathbb{X} \subseteq \mathbb{X}_I$.

L is clearly not greater than the total number of bins created by splitting the region $\mathbb{X}$ into partitions of volume ϵ^d. So we have

$$L \leq \frac{(2c_X)^d}{\epsilon^d}. \tag{11}$$

Proof of Theorem 3.1 Let $\{N'_i\}_{i=1}^L$ and $\{M'_j\}_{j=1}^L$ respectively denote the number of collisions of X and Y points in the bins i and j, using the hash function H_1. E_i stands for the event that there is no collision in bin i for the hash function H_2 with inputs $\{V_i\}_{i=1}^L$. We have

$$\begin{aligned} P(E_i) &= \left(1 - \frac{1}{F}\right)^L + L\left(\frac{1}{F}\right)\left(\frac{F-1}{F}\right)^{L-1} \\ &= 1 - O\left(\frac{L}{F}\right). \end{aligned} \tag{12}$$

By definition,

$$\widehat{D}_g(X,Y) := \frac{1}{M} \sum_{\substack{i \leq F \\ M_i > 0}} M_i \widetilde{g}\left(\frac{\eta N_i}{M_i}\right).$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\widehat{D}_g(X,Y)\right] &= \frac{1}{M}\mathbb{E}\left[\sum_{\substack{i \leq F \\ M_i > 0}} M_i \widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\right] \\ &= \frac{1}{M} \sum_{\substack{i \leq F \\ M_i > 0}} P(E_i)\mathbb{E}\left[M_i \widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle| E_i\right] \\ &+ \frac{1}{M} \sum_{\substack{i \leq F \\ M_i > 0}} P(\overline{E_i})\mathbb{E}\left[M_i \widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle| \overline{E_i}\right]. \end{aligned} \tag{13}$$

We represent the second term in (13) by $\mathbb{B}_H$. $\mathbb{B}_H$ has the interpretation as the bias error due to collisions in hashing. Remember that $\overline{E_i}$ is defined as the event that there is a collision at bin i for the hash function H_2 with inputs $\{V_i\}_{i=1}^L$. For proving as upper bound on $\mathbb{B}_H$, we first need to compute an upper bound on $\sum_{i=1}^L \mathbb{E}\left[M_i \middle| \overline{E_i}\right]$. This is stated in the following lemma.

Lemma 5.2 *We have*

$$\sum_{\substack{i \leq F \\ M_i > 0}} \mathbb{E}\left[M_i \middle| \overline{E_i}\right] \leq O(L) \tag{14}$$

Proof Define $\mathcal{A}_i := \{j : H_2(V_j) = i\}$. For each i we can rewrite M_i as

$$M_i = \sum_{j=1}^L \mathbb{1}_{\mathcal{A}_i}(j) M'_j. \tag{15}$$

Thus,

$$\begin{aligned} \sum_{\substack{i \leq F \\ M_i > 0}} \mathbb{E}\left[M_i \middle| \overline{E_i}\right] &= \sum_{\substack{i \leq F \\ M_i > 0}} \mathbb{E}\left[\sum_{j=1}^L \mathbb{1}_{\mathcal{A}_i}(j) M'_j \middle| \overline{E_i}\right] \\ &= \sum_{\substack{i \leq F \\ M_i > 0}} \sum_{j=1}^L M'_j \mathbb{E}\left[\mathbb{1}_{\mathcal{A}_i}(j) \middle| \overline{E_i}\right] \\ &= \sum_{\substack{i \leq F \\ M_i > 0}} \sum_{j=1}^L M'_j P\big(j \in \mathcal{A}_i | \overline{E_i}\big) \\ &= \sum_{\substack{i \leq F \\ M_i > 0}} \sum_{j=1}^L M'_j \frac{P\big(j \in \mathcal{A}_i, \overline{E_i}\big)}{P(\overline{E_i})}, \end{aligned} \tag{16}$$

where $P\big(j \in \mathcal{A}_i, \overline{E_i}\big)$ and $P(\overline{E_i})$ can be derived as

$$P\big(j \in \mathcal{A}_i, \overline{E_i}\big) = \frac{1}{F}\left(1 - \left(\frac{F-1}{F}\right)^{L-1}\right) = O\left(\frac{L}{F^2}\right), \tag{17}$$

and

$$P(\overline{E_i}) = 1 - P(E_i) = O\left(\frac{L}{F}\right). \tag{18}$$

Plugging in (17) and (18) in (16) results in

$$\begin{aligned} \sum_{\substack{i \leq F \\ M_i > 0}} \mathbb{E}\left[M_i \middle| \overline{E_i}\right] &= \sum_{\substack{i \leq F \\ M_i > 0}} \sum_{j=1}^L M'_j O\left(\frac{1}{F}\right) \\ &= \sum_{\substack{i \leq F \\ M_i > 0}} O\left(\frac{M}{F}\right) = O(L), \end{aligned} \tag{19}$$

where in the third line we use $\eta = M/N$ and $F = c_H N$. In addition, the number of the terms in the sum is upper bounded by L since L is defined as the number of distinct outputs of hashing the X and Y points. Now in the following lemma we prove a bound on $\mathbb{B}_H$.

Lemma 5.3 *Let L denote the number of distinct outputs of the hash function H_1 of the X and Y sample points. The bias of estimator* (5) *due to hashing collision can be upper bounded by*

$$\mathbb{B}_H \leq O\left(\frac{L^2}{N^2}\right) \tag{20}$$

Proof From the definition of $\mathbb{B}_H$ we can write

$$\begin{aligned}\mathbb{B}_H :&= \frac{1}{M}\sum_{\substack{i\leq F\\ M_i>0}} P(\overline{E_i})\mathbb{E}\left[M_i\widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle|\overline{E_i}\right]\\ &= \frac{P(\overline{E_1})}{M}\sum_{\substack{i\leq F\\ M_i>0}} \mathbb{E}\left[M_i\widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle|\overline{E_i}\right]\\ &\leq \frac{P(\overline{E_1})\widetilde{g}(R_{max})}{M}\sum_{\substack{i\leq F\\ M_i>0}} \mathbb{E}\left[M_i\middle|\overline{E_i}\right]\\ &= \frac{P(\overline{E_1})\widetilde{g}(R_{max})}{M}O(L)\\ &= O\left(\frac{L^2}{N^2}\right), \end{aligned} \tag{21}$$

where in the second line we used the fact that $P(\overline{E_i}) = P(\overline{E_1})$. In the third line we used the upper bound for $\tilde{g}$, and in the fourth line we used the result in equation (19).

Now we are ready to continue the proof of the bias bound in (13). Let E be defined as the event that there is no collision for the hash function H_2, and all of its outputs are distinct, that is, $E = \cap_{i=1}^{F} E_i$

(13) can be written as

$$\begin{aligned}&\mathbb{E}\left[\widehat{D}_g(X,Y)\right]\\ &= \frac{1}{M}\sum_{\substack{i\leq F\\ M_i>0}} P(E_i)\mathbb{E}\left[M_i\widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle|E_i\right] + O\left(\frac{L^2}{N^2}\right)\\ &= \frac{P(E_1)}{M}\sum_{\substack{i\leq F\\ M_i>0}} \mathbb{E}\left[M_i\widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle|E_i\right] + O\left(\frac{L^2}{N^2}\right)\\ &= \frac{P(E_1)}{M}\sum_{\substack{i\leq F\\ M_i>0}} \mathbb{E}\left[M_i\widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle|E\right] + O\left(\frac{L^2}{N^2}\right) \end{aligned} \tag{22}$$

$$= \frac{P(E_1)}{M}\mathbb{E}\left[\sum_{\substack{i\leq F\\ M_i>0}} M_i\widetilde{g}\left(\frac{\eta N_i}{M_i}\right)\middle|E\right] + O\left(\frac{L^2}{N^2}\right)$$

$$= \frac{P(E_1)}{M}\mathbb{E}\left[\sum_{i=1}^{L} M_i'\widetilde{g}\left(\frac{\eta N_i'}{M_i'}\right)\middle|E\right] + O\left(\frac{L^2}{N^2}\right) \tag{23}$$

$$= \frac{1-O(L/F)}{M}\mathbb{E}\left[\sum_{i=1}^{M}\widetilde{g}\left(\frac{\eta N_i'}{M_i'}\right)\right] + O\left(\frac{L^2}{N^2}\right) \tag{24}$$

$$= \mathbb{E}_{Y_1\sim f_2(x)}\mathbb{E}\left[\widetilde{g}\left(\frac{\eta N_1'}{M_1'}\right)\middle|Y_1\right] + O\left(\frac{L^2}{N^2}\right), \tag{25}$$

where in (22) we have used the fact that conditioned on E_i, N_i and M_i are independent of E_j for $i \neq j$. In (23) since there is no collision in H_2, M_i' and N_i' are equal to M_j and N_j for some i and j. Equation (24) is because the values M_i' and N_i' are independent of the hash function H_2 and its outputs, and finally in equation (25), we used the fact that each set N_i' and M_i' are i.i.d random variables.

At this point, assuming that the variance of $\frac{N_1'}{M_1'}$ is upper bounded by $O(1/N)$ and using (Lemma 3.2 in [9]), we only need to derive $\mathbb{E}\left[\frac{N_1'}{M_1'}\right]$, and then we can simply find the RHS in (25). Note that N_i' and M_i' are independent and have binomial distributions with the respective means of NP_i^X and MP_i^Y, where P_i^X and P_i^Y are the probabilities of mapping X and Y points with the respective densities f_0 and f_1 into bin i. Hence,

$$\mathbb{E}\left[\frac{N_1'}{M_1'}\middle|Y_1\right] = \mathbb{E}\left[N_1'|Y_1\right]\mathbb{E}\left[{M_1'}^{-1}\middle|Y_1\right]. \tag{26}$$

Let B_i denote the area for which all the points map to the same vector V_i. $\mathbb{E}\left[N_i'\right]$ can be written as:

$$\begin{aligned}\mathbb{E}\left[N_i'\right] &= N\int_{x\in B_i} f_1(x)dx\\ &= N\int_{x\in B_i} f_1(Y_i) + O(\|x-Y_i\|^\gamma)dx\\ &= N\epsilon^d f_1(Y_i) + N\int_{x\in B_i} O(\|x-Y_i\|^\gamma)dx\\ &= N\epsilon^d f_1(Y_i) + N\int_{x\in B_i+Y_i} O(\|x\|^\gamma)dx, \end{aligned} \tag{27}$$

where in the second equality we have used the definition in (2). Let define $B_i' := \frac{1}{\epsilon}B_i + \frac{1}{\epsilon}Y_i$ and

$$C_\gamma(Y_i) := \int_{x'\in B_i'} \|x'\|^\gamma dx'. \tag{28}$$

Note that $C_\gamma(Y_i)$ is a constant independent of ϵ, since the volume of B_i' is independent of ϵ. By defining $x' = x/\epsilon$ we can write

$$\int_{x\in B_i+Y_i}\|x\|^\gamma dx = \int_{x'\in B_i'}\epsilon^\gamma\|x'\|^\gamma(\epsilon^d dx') = C_\gamma(Y_i)\epsilon^{\gamma+d} \tag{29}$$

Also note that since the number of X and Y points in each bin are independent we have $\mathbb{E}\left[N_i'|Y_i\right] = \mathbb{E}\left[N_i'\right]$, and therefore

$$\mathbb{E}\left[N_i'|Y_i\right] = N\epsilon^d f_1(Y_i) + O\big(N\epsilon^{\gamma+d}C_\gamma(Y_i)\big). \tag{30}$$

Next, note that $\mathbb{E}\left[M_i'|Y_i\right]$ has a non-zero binomial distribution, for which the first order inverse moment can be written as [16]:

$$\mathbb{E}\left[{M'_i}^{-1}|Y_i\right] = \left[M\epsilon^d f_2(Y_i) + O\big(M\epsilon^{\gamma+d}C(Y_i)\big)\right]^{-1} \times\left(1+O\left(\frac{1}{M\epsilon^d f_2(Y_i)}\right)\right) = \big(M\epsilon^d f_2(Y_i)\big)^{-1}\left[1+O(\epsilon^\gamma)+O\left(\frac{1}{M\epsilon^d}\right)\right] \quad (31)$$

Thus, (26) can be simplified as

$$\mathbb{E}\left[\frac{N'_1}{M'_1}\middle|Y_1\right] = \frac{f_1(Y_1)}{\eta f_2(Y_1)} + O(\epsilon^\gamma) + O\left(\frac{1}{M\epsilon^d}\right). \quad (32)$$

We use (Lemma 3.2 in [9]) and Remark 3, and obtain

$$\mathbb{E}\left[\widetilde{g}\left(\frac{\eta N'_1}{M'_1}\right)\middle|Y_1\right] = g\left(\frac{f_1(Y_1)}{f_2(Y_1)}\right) + O(\epsilon^\gamma) + O\left(\frac{1}{M\epsilon^d}\right) + O(N^{-\frac{1}{2}}). \quad (33)$$

Finally from (25) we get

$$\begin{aligned}\mathbb{B}\Big[\widehat{D}_g(X,Y)\Big] &= O(\epsilon^\gamma) + O\left(\frac{1}{M\epsilon^d}\right) + O(N^{-\frac{1}{2}}) + O\left(\frac{L^2}{N^2}\right)\\ &= O(\epsilon^\gamma) + O\left(\frac{1}{N\epsilon^d}\right), \end{aligned} \quad (34)$$

where in the third line we have used the upper bound on L in Lemma 5.1 and the fact that $M/N = \eta$. Finally note that we can use a similar method with the same steps to prove the convergence of an estimator for Rényi divergence.

6 Discussion and Experiments

In this section we compare and contrast the advantages of the proposed estimator with competing estimators, and provide numerical results. These show the efficiency of our estimator in terms of MSE rate and computational complexity. Complementary simulated and real-data experiments are provided in the Appendix of the extended paper [15].

Table 1 summarizes the differences between the proposed optimum estimator (EHB) with other competing estimators: Ensemble NNR [9], Ensemble KDE [12] and Mirror KDE [17]. In terms of MSE rate, all of these estimators can achieve the optimal parametric MSE rate of $O(1/N)$. In terms of computational complexity, our estimator has the best runtime compared to others. The smoothness parameter required for the optimum MSE rate is stated in terms of number of required derivatives of the density functions. The proposed estimator is the first divergence estimator that requires no extra smoothness at the boundaries. It is also the first divergence estimator that is directly applicable to online settings, retaining both the accuracy and linear total runtime. Finally, similar to NNR and Ensemble KDE estimators, the proposed estimator does not require any prior knowledge of the support of the densities.

It is also worthwhile to compare the proposed hash-based estimators (HB and EHB) to the histogram plug-in estimator. While the histogram estimator performs poorly when the support set is unknown, the hash based estimator does not rely on the knowledge about the support set. There is a trade-off between bias and variance depending on the bin size parameter in histogram estimators that affects convergence rate. In hash-based estimators the variance is independent of the parameter ϵ., which results in a better performance. In the hash-based estimator, only bins for which $M_i > 0$ are used resulting in reduced memory requirements. Finally, as discussed before, the computational and space complexity of the hash-based estimator respectively grows linearly with the size of dimension. On the other hand, the histogram estimator suffers from exponential time and space complexity with respect to dimension.

Finally, handling the binning in histogram estimators for the support sets with complex contours makes histogram estimators difficult to implement, especially in high dimension. Implementation of our proposed hash-based estimator does not have this complexity since it does not depend on knowledge of the contours.

We compare the empirical performance of EHB to NNR, and the Ensemble KDE estimators. The experiments are done for two different types of f-divergence; KL-divergence and α-divergence defined in [18]. Assume that X and Y are i.i.d. samples from independent truncated Gaussian densities. Figure 1, shows the MSE estimation rate of α-divergence with $\alpha = 0.5$ of two Gaussian densities with the respective expectations of $[0,0]$ and $[0,1]$, and equal variances of $\sigma^2 = I_2$ for different numbers of samples. For each sample size we repeat the experiment 50 times, and compute the MSE of each estimator. While all of the estimators have the same asymptotic MSE rate, in practice the proposed estimator performs better. The runtime of this experiment is shown in Figure 2. The runtime experiment confirms the advantage of the EHB estimator compared to the previous estimators, in terms of computational complexity. Figure 3, shows the comparison of the estimators of KL-divergence between two truncated Gaussian densities with the respective expectations of $[0,0]$ and $[0,1]$, and equal covariance matrices of $\sigma_1^2 = \sigma_2^2 = I_2$, in terms of their mean value and %95 confidence band. The confidence band gets narrower for greater values of N, and EHB estimator has the

Table 1: Comparison of proposed estimator to Ensemble NNR [9], Ensemble KDE [12] and Mirror KDE [17]

Estimator	HB	NNR	Ensemble KDE	Mirror KDE
MSE Rate	$O(1/N)$	$O(1/N)$	$O(1/N)$	$O(1/N)$
Computational Complexity	$O(N)$	$O(kN \log N)$	$O(N^2)$	$O(N^2)$
Required Smoothness (γ)	d	d	$(d+1)/2$	$d/2$
Extra Smooth Boundaries	No	Yes	Yes	Yes
Online Estimation	Yes	No	No	No
Knowledge about Boundary	No	No	No	Yes

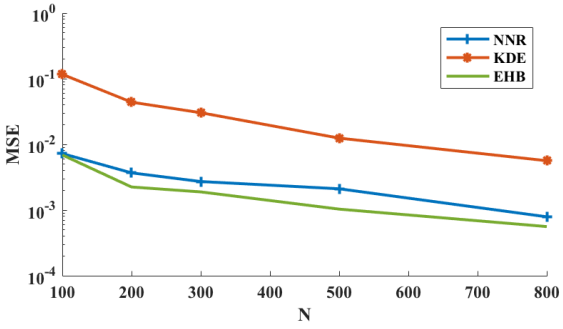

Figure 1: MSE comparison of α-divergence estimators with $\alpha = 0.5$ between two independent, mean-shifted truncated 2D Gaussian densities, versus different number of samples.

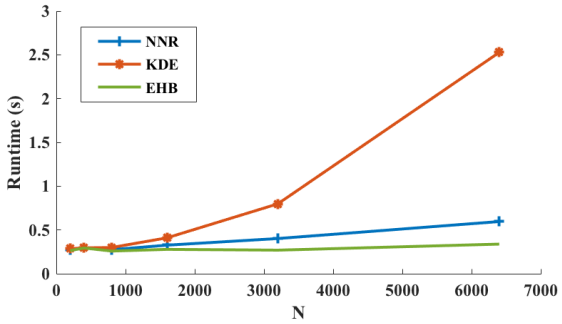

Figure 2: Runtime comparison of α-divergence with $\alpha = 0.5$ between two independent, mean-shifted truncated 2D Gaussian densities, versus different number of samples.

narrowest confidence band. In Figure 4 the MSE rates of the three α-divergence estimators are compared in dimension $d = 4$, $\alpha = 2$, for two independent truncated Gaussian densities with the expectations $\mu_1 = \mu_2$ and covariances $\sigma_1^2 = \sigma_2^2 = I_4$, versus different number of samples.

7 Conclusion

In this paper we proposed a fast hash based estimation method for f-divergence. We obtained bias and variance convergence rates, and validated our results by numerical experiments. Extending the method to hash-based mutual information estimation is a worthwhile topic for future work.

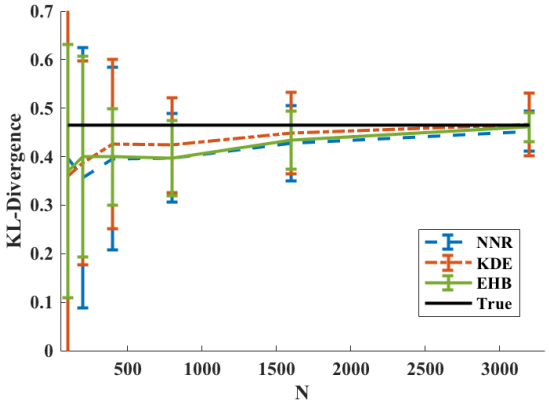

Figure 3: Comparison of the estimators of KL-divergence between two mean-shifted truncated 2D Gaussian densities, in terms of their mean value and %95 confidence band.

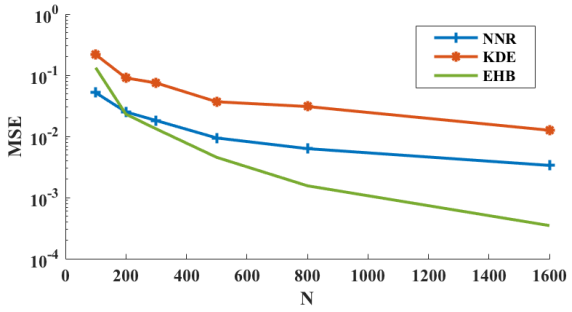

Figure 4: MSE estimation rate of α-divergence with $\alpha = 2$ between two identical truncated Gaussian densities with dimension $d = 4$, versus different number of samples.

References

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

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

[3] K. R. Moon, 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.

[4] 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.

[5] 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.

[6] 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.

[7] S. Singh and B. Póczos, "Exponential concentration of a density functional estimator," in *Advances in Neural Information Processing Systems*, pp. 3032–3040, 2014.

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

[9] M. Noshad, K. R. Moon, S. Y. Sekeh, and A. O. Hero III, "Direct estimation of information divergence using nearest neighbor ratios," *arXiv preprint arXiv:1702.05222*, 2017.

[10] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, *et al.*, "Nonparametric Von Mises estimators for entropies, divergences and mutual informations," in *NIPS*, pp. 397–405, 2015.

[11] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman, "Nonparametric estimation of renyi divergence and friends," in *International Conference on Machine Learning*, pp. 919–927, 2014.

[12] K. R. Moon, 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.

[13] Y.-m. Zhang, K. Huang, G. Geng, and C.-l. Liu, "Fast kNN graph construction with locality sensitive hashing," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 660–674, Springer, 2013.

[14] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, "Multi-probe LSH: efficient indexing for high-dimensional similarity search," in *Proceedings of the 33rd international conference on Very large data bases*, pp. 950–961, VLDB Endowment, 2007.

[15] M. Noshad and A. O. Hero III, "Scalable hash-based estimation of divergence measures," *arXiv preprint arXiv:1801.00398*, 2018.

[16] M. Znidaric, "Asymptotic expansion for inverse moments of binomial and Poisson distributions," *arXiv preprint math/0511226*, 2005.

[17] S. Singh and B. Póczos, "Generalized exponential concentration inequality for Renyi divergence estimation.," in *ICML*, pp. 333–341, 2014.

[18] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi, "Non-negative matrix factorization with α-divergence," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1433–1440, 2008.