Divergence Estimation for Multidimensional Densities Via k-Nearest-Neighbor Distances

Qing Wang, Sanjeev R. Kulkarni, *Fellow, IEEE*, and Sergio Verdú, *Fellow, IEEE*

Abstract—A new universal estimator of divergence is presented for multidimensional continuous densities based on k-nearest-neighbor (k-NN) distances. Assuming independent and identically distributed (i.i.d.) samples, the new estimator is proved to be asymptotically unbiased and mean-square consistent. In experiments with high-dimensional data, the k-NN approach generally exhibits faster convergence than previous algorithms. It is also shown that the speed of convergence of the k-NN method can be further improved by an adaptive choice of k.

Index Terms—Divergence, information measure, Kullback–Leibler, nearest-neighbor, partition, random vector, universal estimation.

I. INTRODUCTION

A. Universal Estimation of Divergence

SUPPOSE P and Q are probability distributions on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$. The divergence between P and Q is defined as [23]

$$D(P\|Q) \equiv \int_{\mathbb{R}^d} \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q} \tag{1}$$

when P is absolutely continuous with respect to Q, and $+\infty$ otherwise. If the densities of P and Q with respect to the Lebesgue measure exist, denoted by $p(x)$ and $q(x)$, respectively, with $p(x) = 0$ for P-almost every x such that $q(x) = 0$ and $0 \log \frac{0}{0} \equiv 0$, then

$$D(p\|q) \equiv \int_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x. \tag{2}$$

The key role of divergence in information theory and large deviations is well known. There has also been a growing interest in applying divergence to various fields of science and engineering for the purpose of gauging distances between distributions. In

Q. Wang is with the Credit Suisse Group, New York, NY 10010 USA (e-mail: qingwang@princeton.edu).

S. R. Kulkarni and S. Verdú are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kulkarni@princeton.edu; verdu@princeton.edu).

[2], Bhattacharya considers the problem of estimating the shift parameter θ, given samples $\{X_i\}$ and $\{Y_i\}$ generated according to density functions $p(x)$ and $p(x - \theta)$, respectively. The estimate $\hat{\theta}$ is the minimizer of the divergence between $\{X_i\}$ and $\{Y_i - \hat{\theta}\}$. Divergence also proves to be useful in neuroscience. Johnson *et al.* [18] employs divergence to quantify the difference between neural response patterns. These techniques are then applied in assessing neural codes, especially, in analyzing which portion of the response is most relevant in stimulus discrimination. In addition, divergence has been used for detecting changes or testing stationarity in Internet measurements [6], [21], and for classification purposes in speech recognition [30], image registration [15], [29], text/multimedia clustering [11], and in a general classification framework called KLBoosting [26]. Finally, since mutual information is a special case of divergence, divergence estimators can be employed to estimate mutual information. This finds application, for example, in secure wireless communications, where mutual information estimates can approximate secrecy capacity and are useful for evaluating secrecy generation algorithms [40], [41]. In biology and neuroscience, mutual information has been employed in gene clustering [4], [5], [31], [34], neuron classification [32], etc.

Despite its wide range of applications, relatively limited work has been done on the universal estimation of divergence, see [36], [38] and references therein. The traditional approach is to use histograms with equally sized bins to estimate the densities $p(x)$ and $q(x)$ and substitute the density estimates $\hat{p}(x)$ and $\hat{q}(x)$ into (2). In [38], we proposed an estimator based on data-dependent partitioning. Instead of estimating the two densities separately, this method estimates the Radon–Nikodym derivative $\mathrm{d}P/\mathrm{d}Q$ using frequency counts on a statistically equivalent partition of $\mathbb{R}^d$. The estimation bias of this method originates from two sources: finite resolutions and finite sample sizes. The basic estimator in [38] can be improved by choosing the number of partitions adaptively or by correcting the error due to finite sample sizes. Algorithm G (a self-contained description of this algorithm is provided in the Appendix) is the most advanced version which combines these two schemes. Although this algorithm is applicable to estimating divergence for multidimensional data, the computational complexity is exponential in d and the estimation accuracy deteriorates quickly as the dimension increases. The intuition is that to attain a fixed accuracy, the required number of samples grows exponentially with the dimension. However, in many applications, e.g., neural coding [37], only moderately large sizes of high-dimensional data are available. This motivates us to search for alternative approaches to efficiently estimate divergence for data in $\mathbb{R}^d$. In this paper, we present a new estimator based on k-nearest-neighbor (k-NN)

distances which bypasses the difficulties associated with partitioning in a high-dimensional space. A 1-nearest-neighbor divergence estimator is introduced in the conference version [39] of this paper.

B. Nearest Neighbor Methods

Since its inception in 1951 [13], the nearest neighbor method has been shown to be a powerful nonparametric technique for classification [3], [9], density estimation [28], and regression estimation [8], [22]. In [19], Kozachenko and Leonenko used 1-NN distances to estimate differential entropy $h(X)$ and proved the mean-square consistency of the resulting estimator for data of any dimension. Tsybakov and van der Meulen [35] considered a truncated version of the differential entropy estimator introduced in [19] and showed its $1/\sqrt{n}$-rate of convergence for a class of one-dimensional densities with unbounded support and exponentially decreasing tails. Goria *et al.* [16] extended the 1-NN differential entropy estimator to k-NN estimation, where k can be any constant. The recent work [25] considered the estimation of Rényi entropy and divergence via nearest-neighbor distances. The assumptions and the corresponding consistency results are different from those presented in this work for divergence estimation. In [37] and [20], the differential entropy based on nearest-neighbor distances is applied to estimate mutual information $I(X;Y)$, via

$$I(X;Y) = h(X) + h(Y) - h(X,Y). \tag{3}$$

Suppose we are to estimate the divergence $D(p\|q)$ between two continuous densities p and q. Let $\{X_1,\ldots,X_n\}$ and $\{Y_1,\ldots,Y_m\}$ be independent and identically distributed (i.i.d.) d-dimensional samples drawn independently from the densities p and q, respectively. Let $\hat{p}$ and $\hat{q}$ be consistent density estimates. Then, by the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{p}(X_i)}{\hat{q}(X_i)} \tag{4}$$

will give us a consistent estimate of $D(p\|q)$ provided that the density estimates $\hat{p}$ and $\hat{q}$ satisfy certain consistency conditions. Particulary, if $\hat{p} \equiv \hat{p}_k$ and $\hat{q} \equiv \hat{q}_k$ are obtained through k-NN density estimation, k should grow with the sample size such that the density estimates are guaranteed to be consistent [28]. In contrast, in this paper, we propose a k-NN divergence estimator that is consistent for any fixed k. The consistency is shown in the sense that both the bias and the variance vanish as the sample sizes increase. These consistency results are motivated by the analysis of the 1-NN differential entropy estimator introduced in [19] and the k-NN version presented in [16]. However, our emphasis is on divergence estimation which involves two sets of samples generated from the two underlying distributions.

C. k-NN Universal Divergence Estimator

Suppose p and q are continuous densities on $\mathbb{R}^d$. Let $\{X_1,\ldots,X_n\}$ and $\{Y_1,\ldots,Y_m\}$ be i.i.d. d-dimensional samples drawn independently from p and q, respectively. Let $\rho_k(i)$ be the Euclidean distance[1] between X_i and its k-NN

[1] Other distance metrics can also be used, for example, L^p-norm distance, $1 \le p \le \infty, p \neq 2$.

in $\{X_j\}_{j\neq i}$. The k-NN of x in $\{z_1,\ldots,z_n\}$ is $z_{i(k)}$ where $i(1),\ldots,i(n)$ is such that

$$\|x - z_{i(1)}\| \le \|x - z_{i(2)}\| \le \cdots \le \|x - z_{i(n)}\|.$$

The distance from X_i to its k-NN in $\{Y_j\}$ is denoted by $\nu_k(i)$. Our proposed estimator for $D(p\|q)$ is

$$\hat{D}_{n,m}(p\|q) = \frac{d}{n}\sum_{i=1}^{n}\log\frac{\nu_k(i)}{\rho_k(i)} + \log\frac{m}{n-1}. \tag{5}$$

To motivate the estimator in (5), denote the d-dimensional open Euclidean ball centered at x with radius ρ by $B(x,\rho)$. Since p is a continuous density, the closure of $B(X_i,\rho_k(i))$ contains k samples $\{X_j\}_{j\neq i}$ almost surely. The k-NN density estimate of p at X_i is

$$\hat{p}_k(X_i) = \frac{k}{n-1}\cdot\frac{1}{c_1(d)\rho_k^d(i)} \tag{6}$$

where

$$c_1(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \tag{7}$$

is the volume of the unit ball. Note that the Gamma function $\Gamma(\cdot)$ is defined as

$$\Gamma(z) = \int_0^{+\infty} t^{z-1}e^{-t}\mathrm{d}t \tag{8}$$

and satisfies

$$\Gamma(z+1) = z\Gamma(z). \tag{9}$$

Similarly, the k-NN density estimate of q evaluated at X_i is

$$\hat{q}_k(X_i) = \frac{k}{m}\cdot\frac{1}{c_1(d)\nu_k^d(i)}. \tag{10}$$

Putting together (4), (6), and (10), we arrive at (5).

D. Paper Organization

Detailed convergence analysis of the divergence estimator in (5) is given in Section II. Section III proposes a generalized version of the k-NN divergence estimator with a data-dependent choice of k and discusses strategies to improve the convergence speed. Experimental results are provided in Section IV. Appendix A describes Algorithm G which serves as a benchmark of comparison in Section IV. The remainder of the Appendix is devoted to the proofs.

II. ANALYSIS

In this section, we prove that the bias (Theorem 1) and the variance (Theorem 2) of the k-NN divergence estimator (5) vanish as sample sizes increase, provided that certain mild regularity conditions are satisfied.

Definition 1: A pair of probability density functions (p, q) is called "$\boldsymbol{\mu}$-regular" if p and q satisfy the following conditions for some $\eta > \mu$:

$$\int_{\mathbb{R}^d}\big|\log p(x)\big|^\eta p(x)\mathrm{d}x < \infty \tag{11}$$

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\big|\log\|x-y\|\big|^\eta p(x)p(y)\mathrm{d}x\mathrm{d}y < \infty \tag{12}$$

$$\int_{\mathbb{R}^d}\big|\log q(x)\big|^\eta p(x)\mathrm{d}x < \infty \tag{13}$$

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\big|\log\|x-y\|\big|^\eta p(x)q(y)\mathrm{d}x\mathrm{d}y < \infty. \tag{14}$$

Theorem 1: Suppose that the pair of probability density functions (p, q) is 1-regular. Then the divergence estimator as shown in (5) is asymptotically unbiased, i.e.,

$$\lim_{n,m \to \infty} \mathbb{E}\left[\hat{D}_{n,m}(p\|q)\right] = D(p\|q). \tag{15}$$

Theorem 2 shows that the k-NN estimator (5) is mean-square consistent. In contrast to our estimators based on partitioning [38], in this analysis, we assume that $\{X_1, \ldots, X_n\}$ is independent of $\{Y_1, \ldots, Y_m\}$ in order to establish mean-square consistency (Theorem 2) whereas this assumption is not required for showing the asymptotic unbiasedness (Theorem 1).

Theorem 2: Suppose that the pair of probability density functions (p, q) is 2-regular. Then

$$\lim_{n,m \to \infty} \mathbb{E}\left[\left(\hat{D}_{n,m}(p\|q) - D(p\|q)\right)^2\right] = 0. \tag{16}$$

III. ADAPTIVE CHOICE OF k

A. A Generalized k-NN Divergence Estimator

In Section II, k is a fixed constant independent of the data. This section presents a generalized k-NN estimator where k can be chosen differently at different sample points. The generalized k-NN divergence estimator is

$$
\begin{aligned}
\tilde{D}_{n,m}&(p\|q) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{\hat{p}_{\ell_i}(X_i)}{\hat{q}_{k_i}(X_i)} - \log\frac{\ell_i}{k_i} - \psi(k_i) + \psi(\ell_i)\right] \\
&= \frac{d}{n}\sum_{i=1}^{n}\left[\log\frac{\nu_{k_i}(i)}{\rho_{\ell_i}(i)}\right] + \frac{1}{n}\sum_{i=1}^{n}[\psi(\ell_i) - \psi(k_i)] \\
&\quad + \log\frac{m}{n-1}
\end{aligned}
\tag{17}
$$

where $\hat{p}_{\ell_i}(X_i)$ is the ℓ_i-NN density estimate of p, $\hat{q}_{k_i}(X_i)$ is the k_i-NN density estimate of q defined in (6) and (10). Compared to (5), a correction term $\frac{1}{n}\sum_{i=1}^{n}[\psi(\ell_i) - \psi(k_i)]$ is added to guarantee consistency, where $\psi(\cdot)$ is the Digamma function defined by

$$\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}. \tag{18}$$

Theorems 3 and 4 show that the divergence estimator (17) is asymptotically consistent.

Theorem 3: Suppose that the pair of probability density functions (p, q) is 1-regular. If $k_i, \ell_i < K$ almost surely for some $K < \infty$, then the divergence estimator as shown in (17) is asymptotically unbiased, i.e.,

$$\lim_{n,m \to \infty} \mathbb{E}\left[\tilde{D}_{n,m}(p\|q)\right] = D(p\|q). \tag{19}$$

Theorem 4: Suppose that the pair of probability density functions (p, q) is 2-regular. If $k_i, \ell_i < K$ almost surely for some $K < \infty$, then

$$\lim_{n,m \to \infty} \mathbb{E}\left[\left(\tilde{D}_{n,m}(p\|q) - D(p\|q)\right)^2\right] = 0. \tag{20}$$

The proofs of Theorems 3 and 4 follow the same lines as those of Theorems 1 and 2 except that we take into account the fact that the value of k depends on the given samples.

Alternatively, k can be identical for various samples but updated as a function of the sample size. As discussed in [7], [33], the choice of k trades off bias and variance: a smaller k will lead to a lower bias and a higher variance. In most cases, it is easier to further reduce the variance by taking the average of multiple runs of the estimator. If the sample sizes are large enough, we can select a larger k to decrease the variance while still guaranteeing a small bias.

Specifically, let us consider the following divergence estimator:

$$D^*_{n,m}(p\|q) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\hat{p}_n(X_i)}{\hat{q}_m(X_i)}, \tag{21}$$

where $\hat{p}_n$ is the ℓ_n-NN density estimate of p and $\hat{q}_m$ is the k_m-NN density estimate of q and $\{X_i\}_{i=1,\ldots,n}$ are i.i.d. samples generated according to p. The following theorem establishes the consistency of the divergence estimator in (21) provided certain regularity conditions.

Theorem 5: Suppose densities p and q are uniformly continuous on $\mathbb{R}^d$ and $D(p\|q) < \infty$. Let k_m and ℓ_n be positive integers satisfying

$$\frac{k_m}{m} \to 0, \quad \frac{k_m}{\log m} \to \infty \tag{22}$$

$$\frac{\ell_n}{n} \to 0, \quad \frac{\ell_n}{\log n} \to \infty. \tag{23}$$

If $\inf_{p(x)>0} p(x) > 0$ and $\inf_{q(x)>0} q(x) > 0$, then

$$\lim_{n,m \to \infty} D^*_{n,m}(p\|q) = D(p\|q), \text{ almost surely.} \tag{24}$$

B. Bias Reduction

To improve the accuracy of the divergence estimator (5), we need to reduce the estimation bias at finite sample sizes. One source of bias is the nonuniformity of the distribution near the vicinity of each sample point. For example, when the underlying multidimensional distributions are skewed [20], the k-NN divergence estimator will tend to overestimate. To tackle this problem, instead of fixing a constant k, we fix the nearest neighbor distance (denoted as $\epsilon(i)$) and count how many X samples and Y samples fall into the ball $B(X_i, \epsilon(i))$ respectively, the intuition being that the biases induced by the two density estimates can empirically cancel each other. By the generalized estimator (17), we could estimate the divergence by

$$\tilde{D}^{(1)}_{n,m}(p\|q) = \frac{1}{n}\sum_{i=1}^{n}[\psi(\ell_i) - \psi(k_i)] + \log\frac{m}{n-1} \tag{25}$$

where ℓ_i (or k_i) is the number of samples X (or samples Y) contained in the ball $B(X_i, \epsilon(i))$.

Then the problem is how to choose $\epsilon(i)$: one possibility is

$$\epsilon(i) = \max\{\rho(i), \nu(i)\}, \tag{26}$$

where

$$\rho(i) = \min_{j=1,\ldots,n, j\neq i}\|X_i - X_j\| \tag{27}$$

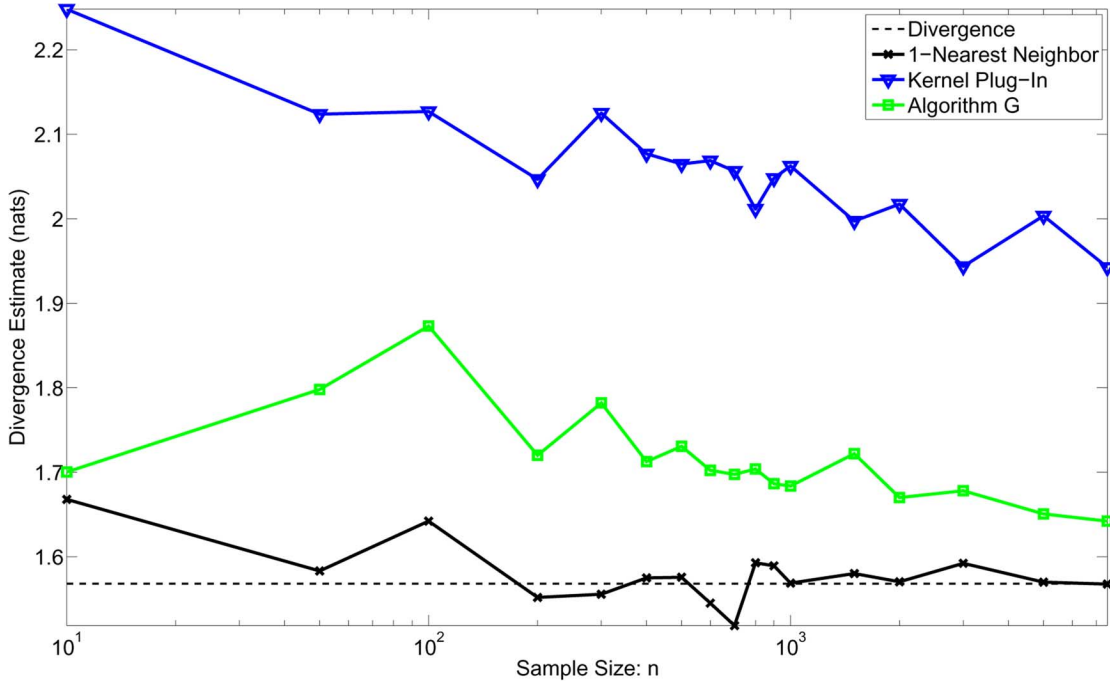$$\nu(i) = \min_{j=1,\ldots,m}\|X_i - Y_j\|. \tag{28}$$

Fig. 1. $X \sim P = \text{Exp}(1)$, $Y \sim Q = \text{Exp}(12)$, $D(P\|Q) = 1.5682$ nats. $k = 1$. In the kernel plug-in method, the divergence is estimated by $\frac{1}{n}\sum_{i=1}^{n} \log \frac{\hat{p}(X_i)}{\hat{q}(X_i)}$, where the kernel method is used for density estimation.

Note that in this case $\epsilon(i)$ may not be the exact ℓ_i-NN distance from X_i to $\{X_j\}_{i \neq j}$ or the exact k_i-NN distance from X_i to $\{Y_j\}$. To correct this error, we may keep the distance terms in the NN estimator and use the divergence estimator in (17) i.e.,

$$\tilde{D}_{n,m}^{(2)}(p\|q) = \tilde{D}_{n,m}(p\|q)$$
$$= \frac{d}{n}\sum_{i=1}^{n}\left[\log\frac{\nu_{k_i}(i)}{\rho_{\ell_i}(i)}\right] + \frac{1}{n}\sum_{i=1}^{n}\left[\psi(\ell_i) - \psi(k_i)\right]$$
$$+ \log\frac{m}{n-1}. \tag{29}$$

On the other hand, since the skewness of the distribution causes the problem, we could apply a linear transformation such that the covariance matrix of the transformed data is close to identity matrix. The idea is to calculate the sample covariance matrix $\hat{C}$ based on the data $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ via

$$\hat{C} = \frac{1}{n+m-1}\left[\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T\right.$$
$$\left. + \sum_{i=1}^{m}(Y_i - \hat{\mu})(Y_i - \hat{\mu})^T\right] \tag{30}$$

where

$$\hat{\mu} = \frac{1}{n+m}\left[\sum_{i=1}^{n}X_i + \sum_{i=1}^{m}Y_i\right] \tag{31}$$

is the sample mean. Then we multiply each centered sample $(X_i - \hat{\mu})$ (or $(Y_i - \hat{\mu})$) by the inverse of the square root of $\hat{C}$ such that the sample covariance matrix of the transformed samples is the identity matrix. Namely

$$X_i' = \left(\hat{C}\right)^{-1/2}(X_i - \hat{\mu}) \tag{32}$$
$$Y_i' = \left(\hat{C}\right)^{-1/2}(Y_i - \hat{\mu}). \tag{33}$$

TABLE I
ESTIMATION VARIANCE FOR $k = 1, 2, 4$

Sample Size	500	1000	2000
1-NN	0.0349	0.0068	0.0016
2-NN	0.0188	0.0054	0.0012
4-NN	0.0182	0.0049	0.0010

IV. EXPERIMENTS

An advantage of the k-NN divergence estimators is that they are more easily generalized and implemented for higher dimensional data than our previous algorithms via data-dependent partitions [38]. Furthermore, various algorithms [1], [14] have been proposed to speed up the k-NN search. However, the NN method becomes unreliable in a high-dimensional space due to the sparsity of the data objects. Hinneburg *et al.* [17] put forward a new notion of NN search, which does not treat all dimensions equally but uses a quality criterion to select relevant dimensions with respect to a given query.

The following experiments are performed on simulated data to compare the NN method with Algorithm G (see Appendix) from our previous work using data-dependent partitions [38]. Algorithm G combines locally adaptive partitions and finite-sample-size error corrections ([38, Algorithms C and E, respectively]). The figures show the average of 25 independent runs, and n and m are equal in the experiments. Fig. 1 shows a case with two exponential distributions. The k-NN method exhibits better convergence than Algorithm G as sample sizes increase. In general, for scalar distributions, Algorithm G suffers from relatively higher bias even when a large number of samples are available. The k-NN method has higher variance for small sample sizes, but as sample sizes increase, the variance decreases quickly. Fig. 1 is for $k = 1$; other choices of k give similar performance in terms of the average of multiple
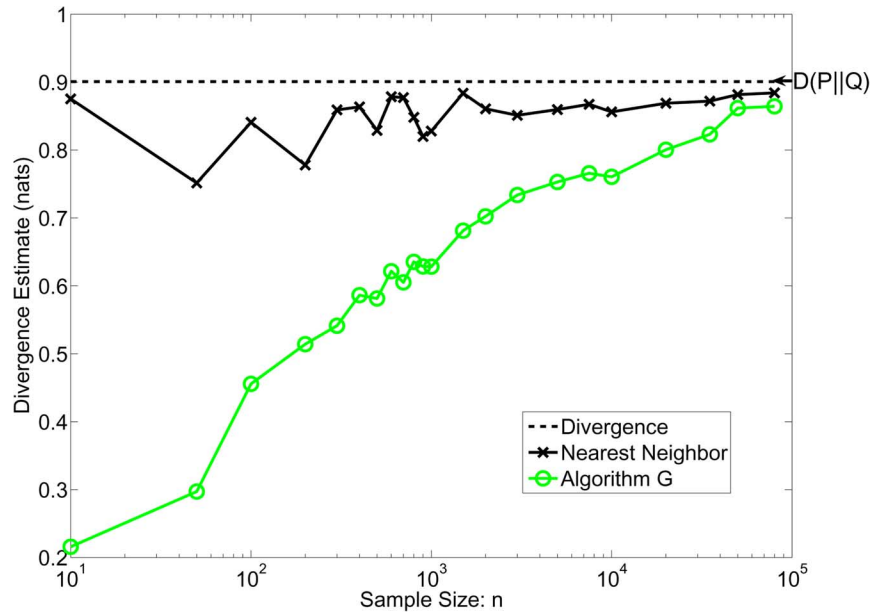
Fig. 2. $X \sim P = \text{Gaussian}(\boldsymbol{\mu}^P, \boldsymbol{C}^P), Y \sim Q = \text{Gaussian}(\boldsymbol{\mu}^Q, \boldsymbol{C}^Q); dim = 4; \boldsymbol{\mu}^P = [0.1\ 0.3\ 0.6\ 0.9]^T, \boldsymbol{\mu}_\ell^Q = [0\ 0\ 0\ 0]^T, \boldsymbol{C}_{\ell,\ell}^P = 1, \boldsymbol{C}_{\ell,s}^P = 0.5, \boldsymbol{C}_{\ell,\ell}^Q = 1, \boldsymbol{C}_{\ell,s}^Q = 0.1$, for $\ell = 1, \ldots, 4, s = 1, \ldots, 4, \ell \neq s; D(P\|Q) = 0.9009$ nats. $k = 1$.
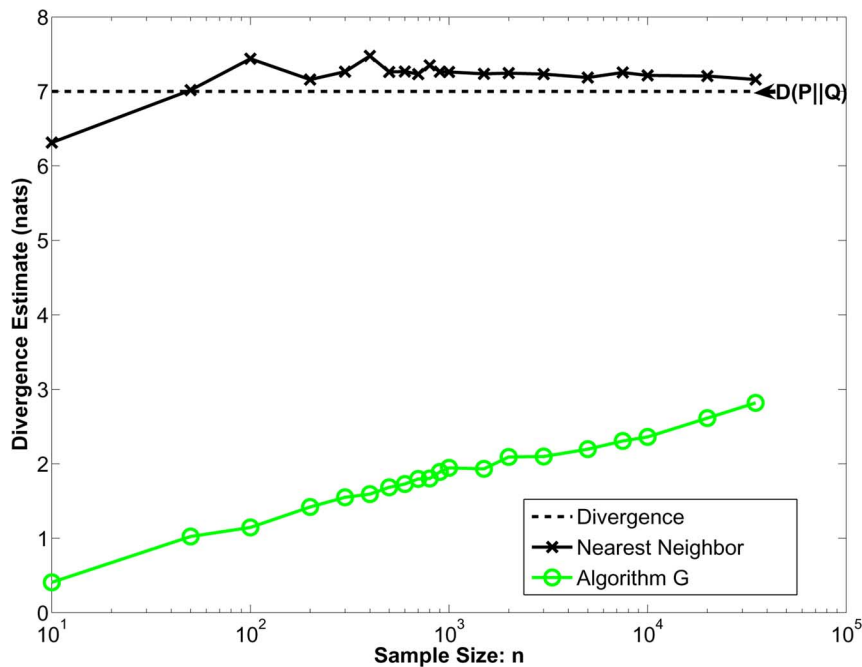

Fig. 3. $X \sim P = \text{Gaussian}(\boldsymbol{\mu}^P, \boldsymbol{C}^P), Y \sim Q = \text{Gaussian}(\boldsymbol{\mu}^Q, \boldsymbol{C}^Q); dim = 10; \boldsymbol{\mu}_\ell^P = \boldsymbol{\mu}_\ell^Q = 0, \boldsymbol{C}_{\ell,\ell}^P = 1, \boldsymbol{C}_{\ell,s}^P = 0.9, \boldsymbol{C}_{\ell,\ell}^Q = 1, \boldsymbol{C}_{\ell,s}^Q = 0.1$, for $\ell = 1, \ldots, 10, s = 1, \ldots, 10, \ell \neq s; D(P\|Q) = 6.9990$ nats. $k = 1$.

independent estimates. The estimation variance decreases with k (see Table I).

In Fig. 2, we consider a pair of four-dimensional Gaussian distributions with different means and different covariance matrices. The 1-NN estimator converges very quickly to the actual divergence whereas Algorithm G is biased and exhibits slower convergence.

In Fig. 3, both distributions are ten-dimensional Gaussian with equal means but different covariance matrices. The estimates by the 1-NN method are closer to the true value, whereas Algorithm G seriously underestimates the divergence.

In Fig. 4, we have two identical distributions in $\mathbb{R}^{20}$. The 1-NN method outperforms Algorithm G, which has a very large positive bias. Note that in the experiments on high-dimensional data, the 1-NN estimator suffers from a larger estimation variance than Algorithm G.

In the experiments, we also apply the k-NN variants (25) and (29) to the raw data (see curves corresponding to $\epsilon\ 1$ and $\epsilon\ 2$, respectively) and we apply the original k-NN estimator (5) (with $k = 1$) to the whitening filtered data (33). Results are shown in Figs. 5 and 6. It is observed that the original 1-NN method with unprocessed data is overestimating and Algorithm G also suf-
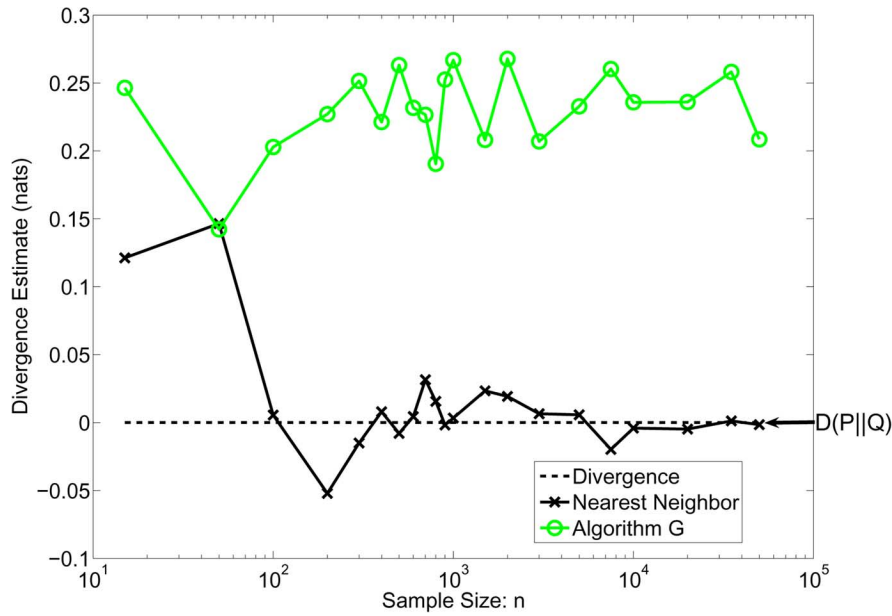
Fig. 4. $X \sim P = \mathrm{Gaussian}(\boldsymbol{\mu}^P, \boldsymbol{C}^P), Y \sim Q = \mathrm{Gaussian}(\boldsymbol{\mu}^Q, \boldsymbol{C}^Q); dim = 20; \boldsymbol{\mu}_\ell^P = \boldsymbol{\mu}_\ell^Q = 0, \boldsymbol{C}_{\ell,\ell}^P = \boldsymbol{C}_{\ell,\ell}^Q = 1, \boldsymbol{C}_{\ell,s}^P = \boldsymbol{C}_{\ell,s}^Q = 0.2,$ for $\ell = 1, \ldots, 20, s = 1, \ldots, 20, \ell \neq s; D(P\|Q) = 0$ nats. $k = 1$.
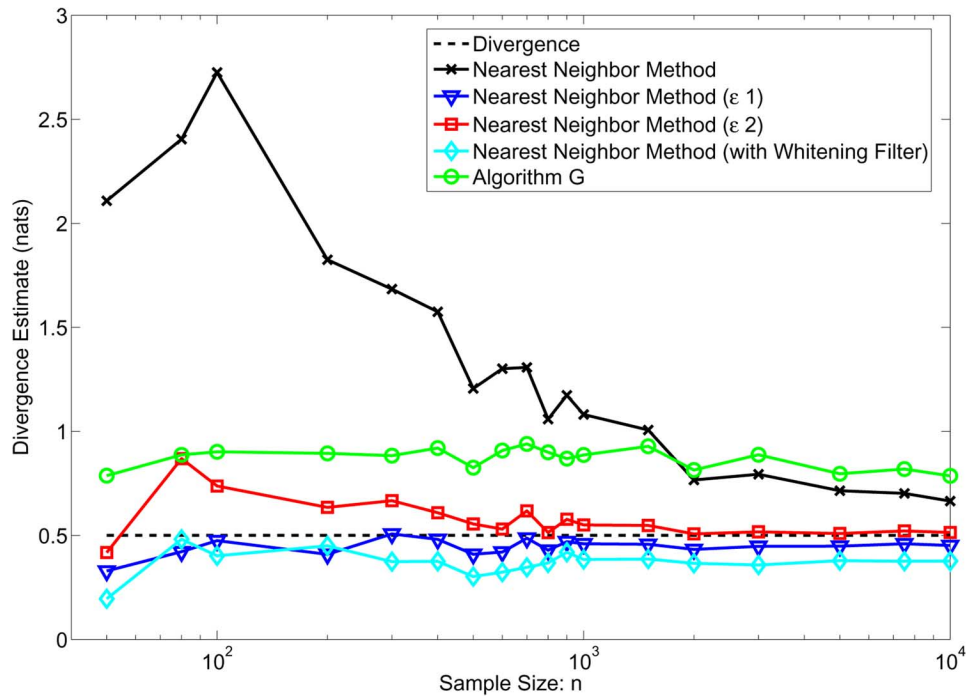
Fig. 5. $\boldsymbol{X} \sim P = \mathrm{Gaussian}(\boldsymbol{\mu}^P, \boldsymbol{C}^P), \boldsymbol{Y} \sim Q = \mathrm{Gaussian}(\boldsymbol{\mu}^Q, \boldsymbol{C}^Q); \dim = 10; \boldsymbol{\mu}_\ell^P = 0, \boldsymbol{\mu}_\ell^Q = 1, \boldsymbol{C}_{\ell,\ell}^P = 1, \boldsymbol{C}_{\ell,s}^P = 0.999, \boldsymbol{C}_{\ell,\ell}^Q = 1, \boldsymbol{C}_{\ell,s}^Q = 0.999,$ for $\ell = 1, \ldots, 10, s = 1, \ldots, 10, \ell \neq s; D(P\|Q) = 0.5005$ nats. For each sample point $X_i, \epsilon(i)$ is chosen according to (26).

fers from serious bias. The performance of (25), (29), and the whitening approach is not quite distinguishable for these examples. All of these approaches can give more accurate results for highly skewed distributions.

In summary, the divergence estimator using the k-NN distances is preferable to partitioning-based methods, especially for multidimensional distributions when the number of samples is limited.

APPENDIX

A. Algorithm G

Algorithm G is a universal divergence estimator which incorporates two methods from [38].

Let P and Q be absolutely continuous probability distributions defined on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$, with $D(P\|Q) < \infty$. $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ are i.i.d. samples
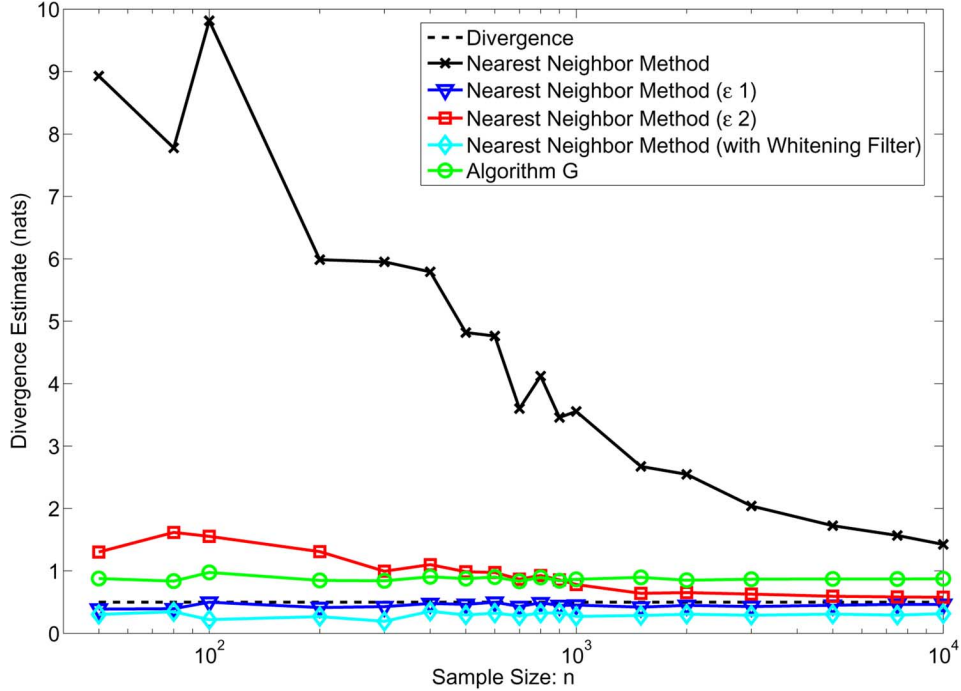
Fig. 6. $\boldsymbol{X} \sim P = \mathrm{Gaussian}(\boldsymbol{\mu}^P, \boldsymbol{C}^P), \boldsymbol{Y} \sim Q = \mathrm{Gaussian}(\boldsymbol{\mu}^Q, \boldsymbol{C}^Q); \dim = 25; \boldsymbol{\mu}_\ell^P = 0, \boldsymbol{\mu}_\ell^Q = 1, \boldsymbol{C}_{\ell,\ell}^P = 1, \boldsymbol{C}_{\ell,s}^P = 0.9999, \boldsymbol{C}_{\ell,\ell}^Q = 1, \boldsymbol{C}_{\ell,s}^Q = 0.9999,$ for $\ell = 1, \ldots, 25, s = 1, \ldots, 25, \ell \neq s; D(P\|Q) = 0.5$ nats. For each sample point $X_i, \epsilon(i)$ is chosen according to (26).

generated from P and Q, respectively. According to Algorithm A [38], the divergence is estimated by

$$\hat{D}_{m,n}(P\|Q) \triangleq \sum_{i=1}^{T_m} P_n(I_i^m) \log \frac{P_n(I_i^m)}{Q_m(I_i^m)} \qquad (34)$$

where P_n, Q_m are the empirical probability measures based on the given samples, $\{I_i^m\}_{i=1,\ldots,T_m}$ is the statistically uniform partition based on ℓ_m-spacing of samples Y (i.e., $Q(I_i^m) = 1/T_m$, for all $i = 1, \ldots, T_m - 1$).

The estimation bias of (34) can be decomposed into two terms

$$\begin{aligned}
B_{n,m} &\triangleq |\hat{D}_{m,n}(P\|Q) - D(P\|Q)| \\
&\leq \left| \sum_{i=1}^{T_m} P_n(I_i^m) \log \frac{P_n(I_i^m)}{Q_m(I_i^m)} - \sum_{i=1}^{T_m} P(I_i) \log \frac{P(I_i)}{Q(I_i)} \right| \\
&\quad + \left| \sum_{i=1}^{T_m} Q(I_i) \frac{P(I_i)}{Q(I_i)} \log \frac{P(I_i)}{Q(I_i)} - \int_\Omega \mathrm{d}Q \frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q} \right| \\
&= B_1 + B_2.
\end{aligned} \qquad (35)$$

B_1 is due to finite sample sizes, and can be approximated by

$$B_1 \approx \frac{T_p - 1}{2n} + \frac{1}{2m} \left[\tilde{\sum_i} \frac{P_n(I_i^m)}{Q_m(I_i^m)} - 1 \right] \qquad (36)$$

where T_p is the total number of cells with $P_n > 0$ and $\tilde{\sum}_i$ denotes that the sum is taken over all I_i^m with $Q_m(I_i^m) > 0$.

B_2 is due to finite resolution. The magnitude of B_2 relies on the partition as well as the underlying distributions and can be decreased by updating the partition according to the distributions. In [38], Algorithm C is proposed to reduce B_2 by locally updating the partition according to $\frac{\mathrm{d}P}{\mathrm{d}Q}$; Algorithm E compensates for finite sample sizes by subtracting an approximation of B_1. Algorithm G, which combines approaches C and E, achieves better accuracy for various distributions.

Algorithm G can be generalized to estimate divergence for distributions defined on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$. We first partition $\mathbb{R}^d$ into T_1 slabs according to projections onto the first coordinate such that each slab contains an equal number of Y samples (generated from distribution Q). Then we scan through all the T_1 slabs. For each slab, if there are many more X samples (generated from distribution P) than Y samples, we further partition that slab into T_2 subsegments according to projections on the second axis. Let k_i (or ℓ_i) denote the number of X (or Y) samples in block i. We continue this process until $\ell_i < \ell_{min}$ or $k_i < \alpha \ell_i$ ($\alpha > 1$ is a parameter) or the partitioning has gone through all the d dimensions of the data. Reduction of finite-sample bias is integrated in Algorithm G using the approximation (36), where the summation is over all the partitions produced by the algorithm.

B. Auxiliary Results

Lemma 1: [12, vol. 2, p. 251] Let ξ_n, $n = 1, 2, \ldots$ be a sequence of random variables, which converges in distribution to a random variable ξ, and suppose there exists $\epsilon > 0$ and a constant $c > 0$ such that

$$\mathbb{E}\left[|\xi_n|^{1+\epsilon}\right] < c, \quad \text{for all } n \geq 1. \qquad (37)$$

Then, for all $\alpha < 1 + \epsilon$ and all integers $r < 1 + \epsilon$

$$\mathbb{E}[|\xi_n|^\alpha] \to \mathbb{E}[|\xi|^\alpha], \quad \mathbb{E}[\xi_n^r] \to \mathbb{E}[\xi^r], \quad \text{as } n \to \infty. \qquad (38)$$

Lemma 2: Let $F(u)$ be a distribution function. Then for $\alpha \geq 1$

$$\int_0^1 \left(\log \frac{1}{u}\right)^\alpha \mathrm{d}F(u) = \alpha \int_0^1 \left(\log \frac{1}{u}\right)^{\alpha-1} u^{-1} F(u) \mathrm{d}u. \qquad (39)$$

Proof: First let us assume both integrals are finite. Integrating by parts, we get

$$\int_b^1 \left(\log\frac{1}{u}\right)^\alpha \mathrm{d}F(u)$$
$$= -F(b)\left(\log\frac{1}{b}\right)^\alpha + \alpha\int_b^1 \left(\log\frac{1}{u}\right)^{\alpha-1} u^{-1}F(u)\mathrm{d}u \quad (40)$$

where $0 < b < 1$.

The fact that if one integral diverges so does the other is easy to show. Assume the left side of (39) it converges, then

$$\lim_{b\to 0}\int_0^b \left(\log\frac{1}{u}\right)^\alpha \mathrm{d}F(u) = 0. \quad (41)$$

Since

$$\int_0^b \left(\log\frac{1}{u}\right)^\alpha \mathrm{d}F(u) \geq \left(\log\frac{1}{b}\right)^\alpha \int_0^b \left(\log\frac{1}{u}\right)^\alpha \mathrm{d}F(u)$$
$$= \left(\log\frac{1}{b}\right)^\alpha F(b) \quad (42)$$

we have

$$\lim_{b\to 0} F(b)\left(\log\frac{1}{b}\right)^\alpha = 0. \quad (43)$$

Therefore

$$\alpha\int_0^1 \left(\log\frac{1}{u}\right)^{\alpha-1} u^{-1}F(u)\mathrm{d}u$$
$$= \lim_{b\to 0}\alpha\int_b^1 \left(\log\frac{1}{u}\right)^{\alpha-1} u^{-1}F(u)\mathrm{d}u$$
$$= \lim_{b\to 0}\int_b^1 \left(\log\frac{1}{u}\right)^\alpha \mathrm{d}F(u)$$
$$= \int_0^1 \left(\log\frac{1}{u}\right)^\alpha \mathrm{d}F(u) < \infty. \quad (44)$$

If the right side of (39) converges, the equality (39) can be proved similarly. □

Lemma 3: Let $F(u)$ be a distribution function. Then for $\alpha \geq 1$

$$\int_1^\infty (\log u)^\alpha \mathrm{d}F(u) = \alpha\int_1^\infty (\log u)^{\alpha-1} u^{-1}(1 - F(u))\mathrm{d}u. \quad (45)$$

Proof: Integrating by parts, we get

$$\int_1^B (\log u)^\alpha \mathrm{d}F(u)$$
$$= -\int_1^B (\log u)^\alpha \mathrm{d}(1 - F(u))$$
$$= -(\log B)^\alpha(1 - F(B))$$
$$\quad + \alpha\int_1^B (\log u)^{\alpha-1} u^{-1}(1 - F(u))\mathrm{d}u \quad (46)$$

where $B > 1$.

By considering the limit as $B \to \infty$, (45) can be obtained similarly as in the proof of Lemma 2 □

C. Proof of Theorem 1

Rewrite $\hat{D}_{n,m}(p\|q)$ as

$$\hat{D}_{n,m}(p\|q) = \frac{1}{n}\sum_{i=1}^n [\phi_{m,k}(i) - \zeta_{n,k}(i)] \quad (47)$$

where

$$\phi_{m,k}(i) \triangleq \log\big(m\nu_k^d(i)\big), \quad \zeta_{n,k}(i) \triangleq \log\big((n-1)\rho_k^d(i)\big).$$

Thus, we have

$$\mathbb{E}\left[\hat{D}_{n,m}(p\|q)\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\phi_{m,k}(i) - \zeta_{n,k}(i)]. \quad (48)$$

Now it suffices to show that the right-hand side of (48) converges to $D(p\|q)$ as $m, n \to \infty$. Using some of the techniques in [19] and [16], we proceed to obtain $\lim_{m\to\infty}\mathbb{E}[\phi_{m,k}(i)]$. Then $\lim_{n\to\infty}\mathbb{E}[\zeta_{n,k}(i)]$ can be derived similarly and $\lim_{m\to\infty}\mathbb{E}[\phi_{m,k}(i)]$ can be obtained by the following the steps.

Step 1: Get the conditional distribution of $\exp(\phi_{m,k}(i))$ given $X_i = x$ and the limit of the distribution as $m \to \infty$.

Step 2: Show that the limit of the conditional expectation $\mathbb{E}[\phi_{m,k}(i)|X_i = x]$ as $m \to \infty$ can be obtained using the limit of the distribution of $\exp(\phi_{m,k}(i))$ conditioned on $X_i = x$.

Step 3: Calculate the limit of $\mathbb{E}[\phi_{m,k}(i)]$ using the limit of $\mathbb{E}[\phi_{m,k}(i)|X_i = x]$ as $m \to \infty$.

Step 1:

Let $G_{m,k,x}$ be the conditional distribution of $\exp(\phi_{m,k}(i))$ given $X_i = x$. Then, for almost all $x \in \mathbb{R}^d$, for all $k \in \mathbb{N}$

$$G_{m,k,x}(u) \triangleq \mathbb{P}\left[\exp(\phi_{m,k}(i)) < u | X_i = x\right]$$
$$= \mathbb{P}\left[\nu_k(i) < (u/m)^{1/d} | X_i = x\right] \quad (49)$$
$$= 1 - \sum_{s=0}^{k-1}\binom{m}{s}\left(\int_{B(x,(u/m)^{1/d})} q(y)\mathrm{d}y\right)^s$$
$$\cdot \left(1 - \int_{B(x,(u/m)^{1/d})} q(y)\mathrm{d}y\right)^{m-s},$$
$$= 1 - \sum_{s=0}^{k-1} f_s(m,u) \quad (50)$$

where $u > 0$, (49) is equivalent to the probability that there exist at least k points in $\{Y_j\}$ belonging to $B(x,(u/m)^{1/d})$, and

$$f_s(m,u) = \binom{m}{s}\left(\int_{B(x,(u/m)^{1/d})} q(y)\mathrm{d}y\right)^s$$
$$\cdot \left(1 - \int_{B(x,(u/m)^{1/d})} q(y)\mathrm{d}y\right)^{m-s}. \quad (51)$$

Note that for almost all $x \in \mathbb{R}^d$, as the radius $r \to 0$, if $q(x) \in L^1(\mathbb{R}^d)$, we have [19], [24]

$$\lim_{r\to 0}\frac{1}{\lambda(B(x,r))}\int_{B(x,r)} q(y)\mathrm{d}y = q(x) \quad (52)$$

where λ represents the Lebesgue measure. Then it can be shown that, as $m \to \infty$

$$G_{m,k,x}(u) \to G_{k,x}(u)$$
$$\triangleq 1 - \sum_{s=0}^{k-1} \frac{1}{s!} \left(c_d q(x)u\right)^s \exp\left(-c_d q(x)u\right). \quad (53)$$

Let $\xi_{m,k,x}$ be a random variable with the distribution function $G_{m,k,x}(u)$ and $\xi_{k,x}$ be a random variable with the distribution function $G_{k,x}(u)$. The corresponding density function of $G_{k,x}(u)$ is

$$\frac{dG_{k,x}(u)}{du} = \frac{1}{(k-1)!} c_d q(x) \left(c_d q(x)u\right)^{k-1} \exp\left(-c_d q(x)u\right). \quad (54)$$

Then for $q(x) > 0$

$$\mathbb{E}[\log \xi_{k,x}]$$
$$= \int_0^\infty \log u \frac{c_d q(x)}{(k-1)!} \left(c_d q(x)u\right)^{k-1} \exp\left(-c_d q(x)u\right) du$$
$$= \frac{1}{(k-1)!} \int_0^\infty \log\left[\frac{t}{c_d q(x)}\right] e^{-t} t^{k-1} dt \quad (55)$$

where the last equality is obtained by the change of the variable $t = c_d q(x)u$. To calculate the integral in (55), first note that

$$\int_0^\infty e^{-t} t^{k-1} dt = \Gamma(k) \quad (56)$$

and $\Gamma(k) = (k-1)!$ due to the property of Gamma function, which leads to

$$\frac{1}{(k-1)!} \int_0^\infty \log\left(c_d q(x)\right) e^{-t} t^{k-1} dt = \log q(x) + \log c_d. \quad (57)$$

Now we only need to consider the first integral in (56). By taking the derivative of (56) with respect to k, we obtain

$$\Gamma'(k) = \int_0^\infty e^{-t} t^{k-1} \log t \, dt. \quad (58)$$

Recall that the Digamma function satisfies

$$\psi(k) = \frac{\Gamma'(k)}{(k-1)!}, \quad \text{for integer } k. \quad (59)$$

Thus

$$\frac{1}{(k-1)!} \int_0^\infty e^{-t} t^{k-1} \log t \, dt = \psi(k). \quad (60)$$

Therefore, with (57) and (60), we have

$$\mathbb{E}[\log \xi_{k,x}] = \psi(k) - \log q(x) - \log c_d. \quad (61)$$

Step 2:
Note that $\mathbb{E}[\log \xi_{m,k,x}] = \mathbb{E}[\phi_{m,k}(i) | X_i = x]$. Hence

$$\lim_{m \to \infty} \mathbb{E}[\phi_{m,k}(i) | X_i = x] = \psi(k) - \log q(x) - \log c_d, \quad (62)$$

for any x such that

$$\lim_{m \to \infty} \mathbb{E}[\log \xi_{m,k,x}] = \mathbb{E}[\log \xi_{k,x}]. \quad (63)$$

Since we already know that $\xi_{m,k,x} \xrightarrow{\mathcal{D}} \xi_{k,x}$, then according to Lemma 2, (63) holds, if we have

$$\mathbb{E}\left[|\log \xi_{m,k,x}|^{1+\epsilon}\right] < C \quad (64)$$

for some $\epsilon > 0$ and some $C > 0$.

Now we are to prove (64). $\mathbb{E}\left[|\log \xi_{m,k,x}|^{1+\epsilon}\right]$ can be decomposed as

$$\mathbb{E}\left[|\log \xi_{m,k,x}|^{1+\epsilon}\right]$$
$$= \int_0^1 |\log u|^{1+\epsilon} dG_{m,k,x}(u) + \int_1^{+\infty} (\log u)^{1+\epsilon} dG_{m,k,x}(u). \quad (65)$$

By Lemmas 2 and 3, it is equivalent to consider the following integrals:

$$(1+\epsilon) \int_0^1 \left(\log \frac{1}{u}\right)^\epsilon u^{-1} G_{m,k,x}(u) du$$
$$+ (1+\epsilon) \int_1^{+\infty} (\log u)^\epsilon u^{-1} (1 - G_{m,k,x}(u)) du$$
$$= (1+\epsilon) \int_0^1 \left(\log \frac{1}{u}\right)^\epsilon u^{-1} G_{m,k,x}(u) du$$
$$+ (1+\epsilon) \int_1^{\sqrt{m}} (\log u)^\epsilon u^{-1} (1 - G_{m,k,x}(u)) du$$
$$+ (1+\epsilon) \int_{\sqrt{m}}^{+\infty} (\log u)^\epsilon u^{-1} (1 - G_{m,k,x}(u)) du$$
$$= I_1 + I_2 + I_3. \quad (66)$$

We estimate the integrals in (66) separately.
Note that

$$f_s(m,u) \le m^k \left(1 - \int_{B(x,(u/m)^{1/d})} q(y) dy\right)^{m-k} \quad (67)$$

thus

$$I_3 = (1+\epsilon) \int_{\sqrt{m}}^{+\infty} (\log u)^\epsilon u^{-1} \sum_{s=0}^{k-1} f_s(m,u) du$$
$$\le (1+\epsilon) k m^k \left(1 - \int_{B(x,(1/\sqrt{m})^{1/d})} q(y) dy\right)^{m-k-1}$$
$$\cdot \int_{\sqrt{m}}^{+\infty} (\log u)^\epsilon u^{-1} \left(1 - \int_{B(x,(u/m)^{1/d})} q(y) dy\right) du \quad (68)$$

where

$$m^k \left(1 - \int_{B(x,(u/m)^{1/d})} q(y) dy\right)^{m-k-1}$$
$$\le m^k \exp\left(-(m-k-1) \int_{B(x,(u/m)^{1/d})} q(y) dy\right) \quad (69)$$
$$\le m^k \exp\left(-(m-k-1)(q(x) - \delta) \cdot \lambda(B(x,(u/m)^{1/d}))\right) \quad (70)$$
$$= m^k \exp\left(-(m-k-1)(q(x) - \delta) \frac{c_d}{\sqrt{m}}\right) \quad (71)$$

and

$$\int_{\sqrt{m}}^{+\infty} (\log u)^{\epsilon} u^{-1} \left(1 - \int_{B(x,(u/m)^{1/d})} q(x) \mathrm{d}x \right) \mathrm{d}u$$

$$= \int_{1/\sqrt{m}}^{+\infty} (\log(mt))^{\epsilon} t^{-1} \left(1 - \int_{B(x,t^{1/d})} q(x) \mathrm{d}x \right) \mathrm{d}t \tag{72}$$

$$= \left(\int_{1/\sqrt{m}}^{1} + \int_{1}^{+\infty} \right) (\log(mt))^{\epsilon} t^{-1}$$

$$\cdot \left(1 - \int_{B(x,t^{1/d})} q(x) \mathrm{d}x \right) \mathrm{d}t$$

$$\leq \sqrt{m}(\log m)^{\epsilon}$$

$$+ C_{\epsilon}(\log m)^{\epsilon} \int_{1}^{+\infty} u^{-1} (1 - G_{1,1,x}(u)) \mathrm{d}u$$

$$+ C_{\epsilon} \int_{1}^{+\infty} (\log u)^{\epsilon} u^{-1} ((1 - G_{1,1,x}(u)) \mathrm{d}u \tag{73}$$

$$\leq \sqrt{m}(\log m)^{\epsilon} + B(\log m)^{\epsilon} + D. \tag{74}$$

The inequality (69) is due to the fact that

$$(1-x)^{A} \leq e^{-Ax} \tag{75}$$

and (70) holds for $\delta > 0$, $q(x) - \delta > 0$, and sufficiently large m. Equation (72) is obtained by change of variable: $t = u/m$. The inequality (73) holds because for $x, y > 0$, $\exists\, C_{\epsilon} > 0$, s.t. $(x+y)^{\epsilon} \leq C_{\epsilon} x^{\epsilon} + C_{\epsilon} y^{\epsilon}$. Equation (74) is true for some constants $B, D > 0$.

To prove that the integrals in (73) converge, it suffices to show that

$$\int_{1}^{+\infty} (\log u)^{\epsilon} u^{-1} ((1 - G_{1,1,x}(u)) \mathrm{d}u < \infty. \tag{76}$$

By Lemma 3, equivalently we need to show

$$\int_{1}^{+\infty} (\log u)^{1+\epsilon} \mathrm{d}G_{1,1,x}(u) < \infty \tag{77}$$

which is implied by the fact that

$$\int_{0}^{+\infty} |\log u|^{1+\epsilon} \mathrm{d}G_{1,1,x}(u)$$

$$= \mathbb{E} \left[|\log \xi_{1,1,x}|^{1+\epsilon} \right] \tag{78}$$

$$= \mathbb{E} \left[\left| \log \|x - y\|^{d} \right|^{1+\epsilon} \right] \tag{79}$$

$$= d^{\epsilon} \int_{\mathbb{R}^{d}} |\log \|x - y\||^{1+\epsilon} q(y) \mathrm{d}y < \infty. \tag{80}$$

Equation (80) holds since (p, q) is 1-regular and the condition (14) is satisfied for $\eta = 1 + \epsilon$.

By (71) and (74), it follows that

$$\lim_{m \to \infty} I_3 = 0. \tag{81}$$

Next consider I_2

$$I_2 = (1+\epsilon) \int_{1}^{\sqrt{m}} (\log u)^{\epsilon} u^{-1} (1 - G_{m,k,x}(u)) \mathrm{d}u$$

$$\leq (1+\epsilon) \sum_{s=0}^{k-1} \int_{1}^{\sqrt{m}} (\log u)^{\epsilon} u^{-1} \left(\sup_{\mathrm{supp}(q)} q(y) \right)^{s}$$

$$\cdot m^{s} \left(c_d \frac{u}{m} \right)^{s} \left(1 - \int_{B(x,(u/m)^{1/d})} q(y) \mathrm{d}y \right)^{m-s} \mathrm{d}u$$

$$\leq (1+\epsilon) \sum_{s=0}^{k-1} M^{s} c_d^{s} \int_{1}^{\sqrt{m}} (\log u)^{\epsilon} u^{s-1}$$

$$\cdot \exp \left(-(m-s) \int_{B(x,(u/m)^{1/d})} q(y) \mathrm{d}y \right) \mathrm{d}u \tag{82}$$

$$\leq (1+\epsilon) \sum_{s=0}^{k-1} M^{s} c_d^{s}$$

$$\cdot \int_{1}^{\sqrt{m}} (\log u)^{\epsilon} u^{s-1} \exp \left(-c_d q(x) u \right) \mathrm{d}u \tag{83}$$

$$\leq (1+\epsilon) \sum_{s=0}^{k-1} M^{s} c_d^{s} \int_{1}^{\sqrt{m}} u^{\epsilon} u^{s-1} \exp \left(-c_d q(x) u \right) \mathrm{d}u$$

$$\leq (1+\epsilon) \sum_{s=0}^{k-1} M^{s} c_d^{-s} q(x)^{-s-\epsilon} \int_{0}^{+\infty} t^{s+\epsilon-1} e^{-t} \mathrm{d}t \tag{84}$$

$$= (1+\epsilon) \sum_{s=0}^{k-1} M^{s} c_d^{-s} q(x)^{-s-\epsilon} \Gamma(s+\epsilon) < \infty \tag{85}$$

where $M = \sup_{\mathrm{supp}(q)} q(y)$ in (82), the inequality (83) holds for sufficiently large m, and (84) is obtained by change of variable $t = c_d q(x) u$.

Finally, we bound I_1

$$I_1 = (1+\epsilon) \int_{0}^{1} \left(\log \frac{1}{u} \right)^{\epsilon} u^{-1} G_{m,k,x}(u) \mathrm{d}u$$

$$\leq \int_{0}^{1} \left(\log \frac{1}{u} \right)^{\epsilon} u^{-1}$$

$$\cdot \left(1 - \left(1 - \int_{B(x,(u/m)^{1/d})} q(y) \mathrm{d}y \right)^{m} \right) \mathrm{d}u. \tag{86}$$

Since

$$\lim_{m \to \infty} \left(1 - \int_{B(x,(u/m)^{1/d})} q(y) \mathrm{d}y \right)^{m} = \exp \left(-c_d q(x) u \right) \tag{87}$$

and

$$1 - e^{-x} < x, \quad \text{for positive } x \tag{88}$$

for sufficiently large m,

$$I_1 < (1+\epsilon) \int_{0}^{1} \left(\log \frac{1}{u} \right)^{\epsilon} u^{-1} c_d q(x) u \mathrm{d}u$$

$$= (1+\epsilon) c_d q(x) \int_{0}^{1} \left(\log \frac{1}{u} \right)^{\epsilon} \mathrm{d}u$$

$$= (1+\epsilon) c_d q(x) \Gamma(1+\epsilon) < \infty. \tag{89}$$

Combining (89), (85) and (81), we have proved (64) and thus, (63) and as a consequence, the desired result (62).

Step 3: We need only to show that for $m \to \infty$

$$\mathbb{E}\left[\phi_{m,k}(i)\right] = \int_{\mathbb{R}^d} \mathbb{E}\left[\phi_{m,k}(i)|X_i = x\right]p(x)\mathrm{d}x$$

$$\to \int_{\mathbb{R}^d}(\psi(k) - \log q(x) - \log c_d)p(x)\mathrm{d}x \quad (90)$$

which follows from [27, p. 176] and the fact that

$$\limsup_{m\to\infty} \int_{\mathbb{R}^d} |\mathbb{E}\left[\phi_{m,k}(i)|X_i = x\right]|^{1+\epsilon}p(x)\mathrm{d}x$$

$$\le \int_{\mathbb{R}^d} \limsup_{m\to\infty} |\mathbb{E}\left[\phi_{m,k}(i)|X_i = x\right]|^{1+\epsilon}p(x)\mathrm{d}x \quad (91)$$

$$= \int_{\mathbb{R}^d} |\psi(k) - \log q(x) - \log c_d|^{1+\epsilon}p(x)\mathrm{d}x \quad (92)$$

$$< \infty \quad (93)$$

where (91) is by Fatou's lemma, (92) follows from (62), and the last inequality is implied by condition (13)) for $\eta = 1 + \epsilon$. The proof of (90) is completed.

Using the same approach and the conditions (11) and (12) for $\eta = 1 + \epsilon$, if follows that

$$\lim_{n\to\infty} \mathbb{E}\left[\zeta_{n,k}(i)\right] = \int_{\mathbb{R}^d}(\psi(k) - \log p(x) - \log c_d)p(x)\mathrm{d}x.$$
$$\quad (94)$$

By (48), (90) and (94), we conclude that

$$\lim_{n,m\to\infty} \mathbb{E}\left[\hat{D}_{n,m}(p\|q)\right] = D(p\|q). \qquad \square$$

D. Proof of Theorem 2

By the triangle inequality, we have

$$\sqrt{\mathbb{E}\left[\left(\hat{D}_{n,m}(p\|q) - D(p\|q)\right)^2\right]}$$

$$\le \sqrt{\mathbb{E}\left[\left(\hat{D}_{n,m}(p\|q) - \mathbb{E}\hat{D}_{n,m}(p\|q)\right)^2\right]}$$

$$+ \sqrt{\mathbb{E}\left[\left(\mathbb{E}\hat{D}_{n,m}(p\|q) - D(p\|q)\right)^2\right]}. \quad (95)$$

Theorem 1 implies that the second term on the right-hand side of (95) will vanish as n, m increase. Thus, it suffices to show that $\mathrm{Var}\left[\hat{D}_{n,m}(p\|q)\right] \to 0$ as $n, m \to \infty$. By the alternative expression (47) for $\hat{D}_{n,m}(p\|q)$, $\mathrm{Var}\left[\hat{D}_{n,m}\right]$ can be written in terms of $\zeta_{n,k}(i)$ and $\phi_{m,k}(i)$, i.e.,

$$\mathrm{Var}\left[\hat{D}_{n,m}\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^n \mathrm{Var}\left[\zeta_{n,k}(i)\right] + \sum_{i\ne j}\mathrm{Cov}\left[\zeta_{n,k}(i), \zeta_{n,k}(j)\right]\right.$$

$$+ \sum_{i=1}^n \mathrm{Var}\left[\phi_{m,k}(i)\right] + \sum_{i\ne j}\mathrm{Cov}\left[\phi_{m,k}(i), \phi_{m,k}(j)\right]$$

$$\left. - \sum_{i,j}\mathrm{Cov}\left[\zeta_{n,k}(i), \phi_{m,k}(j)\right]\right]$$

$$= \frac{1}{n}\mathrm{Var}\left[\zeta_{n,k}(i)\right] + \frac{1}{n^2}\sum_{i\ne j}\mathrm{Cov}\left[\zeta_{n,k}(i), \zeta_{n,k}(j)\right]$$

$$+ \frac{1}{n}\mathrm{Var}\left[\phi_{m,k}(i)\right] + \frac{1}{n^2}\sum_{i\ne j}\mathrm{Cov}\left[\phi_{m,k}(i), \phi_{m,k}(j)\right]$$

$$- \frac{1}{n}\mathrm{Cov}\left[\zeta_{n,k}(i), \phi_{m,k}(i)\right]$$

$$- \frac{1}{n^2}\sum_{i\ne j}\mathrm{Cov}\left[\zeta_{n,k}(i), \phi_{m,k}(j)\right]. \quad (96)$$

Let us first consider the third term on the right-hand side of the preceding equation. As in the proof of Theorem 1, for almost all x

$$\lim_{m\to\infty} \mathbb{E}\left[\phi_{m,k}^2(i)|X_i = x\right]$$

$$= \int_0^\infty \log^2 u \frac{c_d q(x)\left(c_d q(x)u\right)^{k-1}\exp\left(-c_d q(x)u\right)}{(k-1)!}\mathrm{d}u$$

$$= \frac{1}{(k-1)!}\int_0^\infty \left[\log t - \log\left(c_d q(x)\right)\right]^2 e^{-t}t^{k-1}\mathrm{d}t \quad (97)$$

where (97) follows from the change of the integration variable $t = c_d q(x)u$. Then it can be shown that $\lim_{m\to\infty}\mathrm{Var}\left[\phi_{m,k}(i)\right] \equiv C < \infty$. Therefore, the third term in (96) goes to zero as $n, m \to \infty$. Now we need to prove that as $n, m \to \infty$ and $i \ne j$

$$\mathrm{Cov}\left[\phi_{m,k}(i), \phi_{m,k}(j)\right] \to 0. \quad (98)$$

For $i \ne j$, $u, w > 0$, $x, y \in \mathbb{R}^d$, and $k \in \mathbb{N}$, we have

$$\mathbb{P}\left[\exp(\phi_{m,k}(i)) < u, \exp(\phi_{m,k}(j)) < w|\right.$$
$$X_i = x, X_j = y]$$
$$= 1 - \mathbb{P}\left[\exp(\phi_{m,k}(i)) \ge u|X_i = x, X_j = y\right]$$
$$- \mathbb{P}\left[\exp(\phi_{m,k}(j)) \ge w|X_i = x, X_j = y\right]$$
$$+ \mathbb{P}\left[\exp(\phi_{m,k}(i)) \ge u, \exp(\phi_{m,k}(j)) \ge w|\right.$$
$$X_i = x, X_j = y]. \quad (99)$$

The following can be proven by similar reasoning as in Theorem 1:

$$\mathbb{P}\left[\exp(\phi_{m,k}(i)) \ge u|X_i = x, X_j = y\right]$$
$$\stackrel{m\to\infty}{\to} 1 - \sum_{s=0}^{k-1}\frac{1}{s!}\left(c_d q(x)u\right)^s \exp\left(-c_d q(x)u\right); \quad (100)$$
$$\mathbb{P}\left[\exp(\phi_{m,k}(j)) \ge w|X_i = x, X_j = y\right]$$
$$\stackrel{m\to\infty}{\to} 1 - \sum_{t=0}^{k-1}\frac{1}{t!}\left(c_d q(y)w\right)^t \exp\left(-c_d q(y)w\right). \quad (101)$$

For the last term in (99), for m sufficiently large, we should have $B(x, (u/m)^{1/d}) \cap B(y, (w/m)^{1/d}) = \emptyset$. Therefore, when m is large enough

$$\mathbb{P}\left[\exp(\phi_{m,k}(i)) \ge u, \exp(\phi_{m,k}(j)) \ge w|\right.$$
$$X_i = x, X_j = y]$$
$$= \sum_{s=0}^{k-1}\sum_{t=0}^{k-1}\frac{m!}{s!t!(m-s-t)!}\left(\int_{B(x,(u/m)^{1/d})}q(z)\mathrm{d}z\right)^s$$
$$\cdot \left(\int_{B(y,(w/m)^{1/d})}q(z)\mathrm{d}z\right)^t$$
$$\cdot \left(1 - \int_{B(x,(u/m)^{1/d})\cup B(y,(w/m)^{1/d})}q(z)\mathrm{d}z\right)^{m-s-t}$$

$$\to \sum_{s=0}^{k-1} \sum_{t=0}^{k-1} \frac{1}{s!} \frac{1}{t!} \big(c_d q(x)u\big)^s \big(c_d q(y)w\big)^t$$
$$\cdot \exp\big(-c_d q(x)u - c_d q(y)w\big). \qquad (102)$$

Substituting (100)–(102) into (99) and taking derivatives with respect to u and w, we can obtain the conditional joint density of $\exp(\phi_{m,k}(i))$ and $\exp(\phi_{m,k}(j))$. This joint density gives

$$\lim_{m\to\infty} \mathbb{E}\left[\phi_{m,k}(i)\phi_{m,k}(j)|X_i = x, X_j = y\right]$$
$$= \int_0^\infty \int_0^\infty \log u \log w \frac{c_d q(x)}{(k-1)!} \frac{c_d q(y)}{(k-1)!} \big(c_d q(x)u\big)^{k-1}$$
$$\cdot \big(c_d q(y)w\big)^{k-1} \exp\big(-c_d q(x)u - c_d q(y)w\big) \mathrm{d}u\mathrm{d}w$$
$$= \left[\psi(k) - \log\big(q(x)c_d\big)\right] \left[\psi(k) - \log\big(q(y)c_d\big)\right]. \quad (103)$$

The passing to the limit in (103) can be justified as in Step 2 in the proof of Theorem 1 by applying the condition (14).
Then

$$\lim_{m\to\infty} \mathbb{E}\left[\phi_{m,k}(i)\phi_{m,k}(j)\right]$$
$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x)p(y)(\psi(k) - \log(q(x)c_d))$$
$$\cdot (\psi(k) - \log(q(y)c_d)) \, \mathrm{d}x\mathrm{d}y \qquad (104)$$

which can be proved by condition (13) and the argument in (93) The right-hand side of (104) is equal to

$$\lim_{m\to\infty} \mathbb{E}\left[\phi_{m,k}(i)\right] \mathbb{E}\left[\phi_{m,k}(j)\right] = \lim_{m\to\infty} \left(\mathbb{E}\left[\phi_{m,k}(i)\right]\right)^2.$$

Thus, (98) holds. Therefore, the third and fourth terms of (96) vanish as $m \to \infty$. In the same fashion, we can show that the first two terms of (96) also diminish as sample sizes increase.

Now let us consider the last two terms in the right side of (96). Since the samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ are assumed to be independent from each other, $\zeta_{n,k}(i)$ and $\phi_{m,k}(j)$ are independent given X_i and X_j (i can be equal to j). We have

$$\mathbb{E}\left[\zeta_{n,k}(i)\phi_{m,k}(j)|X_i = x, X_j = y\right]$$
$$= \mathbb{E}\left[\zeta_{n,k}(i)|X_i = x, X_j = y\right] \mathbb{E}\left[\phi_{m,k}(j)|X_j = y\right] \quad (105)$$

where

$$\lim_{n\to\infty} \mathbb{E}\left[\zeta_{n,k}(i)|X_i = x, X_j = y\right] = \lim_{n\to\infty} \mathbb{E}\left[\zeta_{n,k}(i)|X_i = x\right]$$
$$= \psi(k) - \log(p(x)c_d) \qquad (106)$$

$$\lim_{m\to\infty} \mathbb{E}\left[\phi_{m,k}(j)|X_j = y\right] = \psi(k) - \log(q(y)c_d). \qquad (107)$$

If $i \neq j$, we have

$$\lim_{n,m\to\infty} \mathbb{E}\left[\zeta_{n,k}(i)\phi_{m,k}(j)\right]$$
$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left[\psi(k) - \log(p(x)c_d)\right] \left[\psi(k) - \log(q(y)c_d)\right]$$
$$\cdot p(x)p(y)\mathrm{d}x\mathrm{d}y,$$

which is equal to $\lim_{n,m\to\infty} \mathbb{E}\left[\zeta_{n,k}(i)\right] \mathbb{E}\left[\phi_{m,k}(j)\right]$. Hence

$$\lim_{n,m\to\infty} \mathrm{Cov}\left[\zeta_{n,k}(i), \phi_{m,k}(j)\right] = 0, \quad \text{if } i \neq j. \qquad (108)$$

If $i = j$, since (p, q) is 2-regular, by conditions (11) and (13), we obtain

$$\int_{\mathbb{R}^d} |\log p(x)|^\epsilon |\log q(x)|^\epsilon p(x)\mathrm{d}x < \infty, \quad \text{for some } \epsilon > 1. \qquad (109)$$

Thus, $\lim_{n,m\to\infty} \mathbb{E}\left[\zeta_{n,k}(i)\phi_{m,k}(i)\right] < \infty$, which implies

$$\lim_{n,m\to\infty} \mathrm{Cov}\left[\zeta_{n,k}(i), \phi_{m,k}(i)\right] < \infty \qquad (110)$$

since we already have (90) and (94). Therefore, the last two terms of (96) are guaranteed to vanish as $n, m \to \infty$. $\qquad \square$

E. Proof of Theorem 5

Consider the following decomposition of the error:

$$|D_{n,m}^*(p\|q) - D(p\|q)|$$
$$\leq \left| \frac{1}{n}\sum_{i=1}^n \log\frac{\hat{p}_n(X_i)}{\hat{q}_m(X_i)} - \frac{1}{n}\sum_{i=1}^n \log\frac{p(X_i)}{q(X_i)} \right|$$
$$+ \left| \frac{1}{n}\sum_{i=1}^n \log\frac{p(X_i)}{q(X_i)} - D(p\|q) \right|$$
$$\leq \frac{1}{n}\sum_{i=1}^n |\log\hat{p}_n(X_i) - \log p(X_i)|$$
$$+ \frac{1}{n}\sum_{i=1}^n |\log\hat{q}_m(X_i) - \log q(X_i)|$$
$$+ \left| \frac{1}{n}\sum_{i=1}^n \log\frac{p(X_i)}{q(X_i)} - D(p\|q) \right|$$
$$= e_1 + e_2 + e_3. \qquad (111)$$

By the Law of Large Numbers, it follows that $e_3 \to 0$ almost surely. Hence, for $\forall \epsilon > 0$, there exists N_1 such that $e_3 < \epsilon/3$ for $\forall n > N_1$,.

By a result from [10, p. 537] and the conditions (22) and (23), it can be shown that $\hat{p}_n(x)$ and $\hat{q}_m(x)$ are uniformly strongly consistent, i.e.,

$$\lim_{n\to\infty} \sup_x |\hat{p}_n(x) - p(x)| \to 0, \text{ almost surely} \qquad (112)$$
$$\lim_{m\to\infty} \sup_x |\hat{q}_m(x) - p(x)| \to 0, \text{ almost surely.} \qquad (113)$$

Therefore, for almost every z in the support of p, for sufficiently large n

$$|\hat{p}_n(z) - p(z)| < \frac{p(z)}{2}, \text{ almost surely.} \qquad (114)$$

Now let us consider

$$|\log\hat{p}_n(X_i) - \log p(X_i)| = \frac{|\hat{p}_n(X_i) - p(X_i)|}{\theta\hat{p}_n(X_i) + (1-\theta)p(X_i)}$$
$$\leq \frac{2|\hat{p}_n(X_i) - p(X_i)|}{p(X_i)}$$
$$\leq \frac{2|\hat{p}_n(X_i) - p(X_i)|}{\inf_{p(x)>0} p(x)} \qquad (115)$$

where $\theta \in [0,1]$ and the first equality is obtained by Taylor's expansion and the first inequality is due to the fact (114). Then because of the uniform convergence of $\hat{p}_n$, we could find $N_2 > N_1$ such that $\forall\, n > N_2, 6|\hat{p}_n(x) - p(x)| < \inf_{p(x)>0} p(x)$ for all x, which implies that $e_1 < \epsilon/3$.

Similarly, it can be shown that there exists $M > 0$ such that $e_2 < \epsilon/3$ for $\forall\, m > M$. Therefore, (111) is upper-bounded by ϵ if $n > N_2, m > M$. $\qquad\square$

REFERENCES

[1] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[2] P. K. Bhattacharya, "Efficient estimation of a shift parameter from grouped data," *Ann. Math. Statist.*, vol. 38, no. 6, pp. 1770–1787, Dec. 1967.

[3] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[4] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. C. Mueller, "Gene mapping and marker clustering using Shannon's mutual information," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 3, no. 1, pp. 47–56, Jan.-Mar. 2006.

[5] Z. Dawy, F. M. Gonzaléz, J. Hagenauer, and J. C. Mueller, "Modeling and analysis of gene expression mechanisms: A communication theory approach," in *Proc. 2005 IEEE Int. Conf. Communications*, Seoul, Korea, May 2005, vol. 2, pp. 815–819.

[6] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi, "An information-theoretic approach to detecting changes in multi-dimensional data streams," in *Proc. 38th Symp. Interface of Statistics, Computing Science, and Applications (Interface '06)*, Pasadena, CA, May 2006.

[7] L. Devroye, "A course in density estimation," in *Progress in Probability and Statistics*. Boston, MA: Birkhäuser, 1987, vol. 14.

[8] L. Devroye, L. Györfi, A. Krzyżak, and G. Lugosi, "On the strong universal consistency of nearest neighbor regression function estimates," *Ann. Statist.*, vol. 22, no. 3, pp. 1371–1385, Sep. 1994.

[9] L. Devroye, L. Györfi, and G. Lugosi, *A Probablistic Theory of Pattern Recognition*. New York: Springer, 1996.

[10] L. P. Devroye and T. J. Wagner, "The strong uniform consistency of nearest neighbor density estimates," *Ann. Statist.*, vol. 5, no. 3, pp. 536–540, May 1977.

[11] I. S. Dhillon, S. Mallela, and R. Kuma, "A divisive information-theoretic feature clustering algorithm for text classification," *J. Mach. Learning Res.*, vol. 3, pp. 1265–1287, Mar. 2003.

[12] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York: Wiley, 1970.

[13] E. Fix and J. L. Hodges, "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," USAF School of Aviation Medicine, Randolph Field, TX, Tech. Rep. 4, Project Number 21-49-004, 1951.

[14] K. Fukunaga and P. M. Nerada, "A branch and bound algorithm for computing k-nearest neighbors," *IEEE Trans. Comput.*, vol. C-24, no. 7, pp. 750–753, Jul. 1975.

[15] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Proc. 9th IEEE Int. Conf. Computer Vision*, Nice, France, Oct. 2003, vol. 1, pp. 487–493.

[16] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. N. Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *J. Nonparam. Statist.*, vol. 17, no. 3, pp. 277–297, Apr. 2005.

[17] A. Hinneburg, C. C. Aggarwal, and D. A. Keim, "What is the nearest neighbor in high dimensional spaces?," in *Proc. 26th Very Large Data Base (VLDB) Conf.*, Cairo, Egypt, 2000, pp. 506–515.

[18] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri, "Information-theoretic analysis of neural coding," *J. Comput. Neurosci.*, vol. 10, no. 1, pp. 47–69, Jan. 2001.

[19] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Probl. Inf. Transm.*, vol. 23, pp. 95–101, 1987.

[20] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, 2004.

[21] B. Krishnamurthy, H. V. Madhyastha, and S. Venkatasubramanian, "On stationarity in internet measurements through an information-theoretic lens," in *Proc. 21st Int. Conf. Data Engineering Workshops*, Tokyo, Japan, Apr. 2005, pp. 1185–1185.

[22] S. R. Kulkarni, S. E. Posner, and S. Sandilya, "Data-dependent k_n-NN and kernel estimators consistent for arbitrary processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 10, pp. 2785–2788, Oct. 2002.

[23] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[24] H. Lebesgue, "Sur l'intégration des fonctions discontinues," *Ann. Ecole Norm.*, no. 27, pp. 361–450, 1910.

[25] N. Leonenko, L. Pronzato, and V. Savani, "Estimation of entropies and divergences via nearest neighbors," *Tatra Mt. Publ.*, vol. 39, pp. 265–273, 2008.

[26] C. Liu and H.-Y. Shum, "Kullback-Leibler boosting," in *Proc. 2003 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Madison, WI, Jun. 2003, vol. 1, pp. I–587.

[27] M. Loève, *Probability Theory*, 4th ed. New York: Springer-Verlag, 1977.

[28] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, no. 3, pp. 1049–1051, Jun. 1965.

[29] J. R. Mathiassen, A. Skavhaug, and K. Bø, "Texture similarity measure using Kullback-Leibler divergence between gamma distributions," in *Proc. 7th European Conf. Computer Vision-Part III*, Copenhagen, Denmark, Jun. 2002, pp. 133–147.

[30] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. J. Rubio, "A new Kullback-Leibler vad for speech recognition in noise," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 266–269, Feb. 2004.

[31] M. Sarkis, B. Goebel, Z. Dawy, J. Hagenauer, P. Hanus, and J. C. Mueller, "Gene mapping of complex diseases—A comparison of methods from statistics informnation theory, and signal processing," *IEEE Signal Process. Mag.*, vol. 24, no. 1, pp. 83–90, Jan. 2007.

[32] E. Schneidman, W. Bialek, and M. J. Berry, II, "An information theoretic approach to the functional classification of neurons," in *Advances in Neural Information Processing*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, vol. 15, pp. 197–204.

[33] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman&Hall, 1986.

[34] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, pp. S231–S240, Oct. 2002.

[35] A. B. Tsybakov and E. C. van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," *Scand. J. Statist.*, vol. 23, pp. 75–83, 1996.

[36] S. Verdú, "Universal estimation of information measures," in *Proc. 2005 IEEE Information Theory Workshop on Coding and Complexity*, Rotorua, New Zealand, Aug. 2005, pp. 232–232.

[37] J. D. Victor, "Binless strategies for estimation of information from neural data," *Phys. Rev. E*, vol. 66, 2002.

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

[39] Q. Wang, S. R. Kulkarni, and S. Verdú, "A nearest-neighbor approach to estimating divergence between continuous random vectors," in *Proc. IEEE Int. Symp. Information Theory (ISIT2006)*, Seattle, WA, Jul. 2006, pp. 242–246.

[40] R. Wilson, D. N. C. Tse, and R. Scholtz, "Channel identification: Secret sharing using reciprocity in UWB channels," *IEEE Trans. Inf. Forensics and Security*, vol. 2, pp. 364–375, Sep. 2007.

[41] C. Ye, A. Reznik, and Y. Shah, "Extracting secrecy from jointly Gaussian random variables," in *Proc. IEEE Int. Symp. Information Theory (ISIT2006)*, Seattle, WA, Jul. 2006, pp. 2593–2597.

Qing Wang was born in Shanghai, China, in 1980. She received the B.S. degree in electronics and information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2002, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 2004 and 2008, respectively.

She is currently with Credit Suisse Securities, New York City, working on quantitative trading strategies.

Sanjeev R. Kulkarni (M'91–SM'96–F'04) received the B.S. degree in mathematics, the B.S. degree in electrical engineering, the M.S. degree in mathematics from Clarkson University, Potsdam, NY, in 1983, 1984, and 1985, respectively, the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1985, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1991.

From 1985 to 1991, he was a Member of the Technical Staff at MIT Lincoln Laboratory, Lexington, MA. Since 1991, he has been with Princeton University, Princeton, NJ, where he is currently Professor of Electrical Engineering, and an affiliated faculty member in the Department of Operations Research and Financial Engineering and the Department of Philosophy. He spent January 1996 as a research fellow at the Australian National University, Canberra; 1998 with Susquehanna International Group, and Summer 2001 with Flarion Technologies. His research interests include statistical pattern recognition, nonparametric statistics, learning and adaptive systems, information theory, wireless networks, and image/video processing.

Prof. Kulkarni received an ARO Young Investigator Award in 1992, an NSF Young Investigator Award in 1994, and several teaching awards at Princeton University. He has served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY

Sergio Verdú (S'80–M'84–SM'88–F'93) received the telecommunications engineering degree from the Universitat Politècnica de Barcelona, Barcelona, Spain, in 1980 and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign in 1984.

Since 1984, he has been a member of the faculty of Princeton University, where he is the Eugene Higgins Professor of Electrical Engineering.

Prof. Verdú is the recipient of the 2007 Claude E. Shannon Award and the 2008 IEEE Richard W. Hamming Medal. He is a member of the National Academy of Engineering and was awarded a Doctorate *Honoris Causa* from the Universitat Politècnica de Catalunya in 2005. He is a recipient of several paper awards from the IEEE: the 1992 Donald Fink Paper Award, the 1998 Information Theory Outstanding Paper Award, an Information Theory Golden Jubilee Paper Award, the 2002 Leonard Abraham Prize Award, and the 2006 Joint Communications/Information Theory Paper Award. In 1998, Cambridge University Press published his book *Multiuser Detection*, for which he received the 2000 Frederick E. Terman Award from the American Society for Engineering Education. He served as President of the IEEE Information Theory Society in 1997 and as Associate Editor for Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY. He is currently Editor-in-Chief of *Foundations and Trends in Communications and Information Theory*.