Divergence Estimation of Continuous Distributions Based on Data-Dependent Partitions

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

***Abstract*—We present a universal estimator of the divergence $D(P \,\|\, Q)$ for two arbitrary continuous distributions P and Q satisfying certain regularity conditions. This algorithm, which observes independent and identically distributed (i.i.d.) samples from both P and Q, is based on the estimation of the Radon–Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ via a data-dependent partition of the observation space. Strong convergence of this estimator is proved with an empirically equivalent segmentation of the space. This basic estimator is further improved by adaptive partitioning schemes and by bias correction. The application of the algorithms to data with memory is also investigated. In the simulations, we compare our estimators with the direct plug-in estimator and estimators based on other partitioning approaches. Experimental results show that our methods achieve the best convergence performance in most of the tested cases.**

***Index Terms*—Bias correction, data-dependent partition, divergence, Radon–Nikodym derivative, stationary and ergodic data, universal estimation of information measures.**

I. Introduction

Kullback and Leibler [1] introduced the concept of information divergence, which measures the distance between the distributions of random variables. Suppose P and Q are probability measures on a measurable space $(\Omega, \mathcal{F})$. The (Kullback–Leibler) divergence between P and Q is defined as

$$D(P \,\|\, Q) \equiv \int_{\Omega} \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q} \tag{1}$$

when P is absolutely continuous with respect to Q, and $+\infty$ otherwise. If the densities of P and Q with respect to Lebesgue measure exist, which are denoted as $p(x)$ and $q(x)$, respectively, then

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

Divergence is an important concept in information theory, since other information-theoretic quantities including entropy and mutual information may be formulated as special cases. For continuous distributions in particular, it overcomes the difficulties with differential entropy. Divergence also plays a key role in large-deviations results including the asymptotic rate of decrease of error probability in binary hypothesis testing problems. Moreover, divergence has proven to be useful in applications. For example, divergence can be employed as a similarity measure in image registration or multimedia classification [2]. It is also applicable as a loss function in evaluating and optimizing the performance of density estimation methods [3]. However, there has been relatively little work done on the universal estimation of divergence between unknown distributions. The estimation of divergence between the samples drawn from unknown distributions gauges the distance between those distributions. Divergence estimates can then be used in clustering and in particular for deciding whether the samples come from the same distribution by comparing the estimate to a threshold. Divergence estimates can also be used to determine sample sizes required to achieve given performance levels in hypothesis testing.

In the discrete case, Cai *et al.* [5], [6] proposed new algorithms to estimate the divergence between two finite alphabet, finite memory sources. In [5], the Burrows–Wheeler block-sorting transform is applied to the concatenation of two random sequences, such that the output sequences possess the convenient property of being piecewise memoryless. Experiments show that these algorithms outperform the estimators based on Lempel–Ziv (LZ) string matching [7].

For sources with a continuous alphabet, Hero *et al.* [8] provided an entropy estimation method using the minimal spanning tree which spans a set of feature vectors. In [8], this method was generalized to divergence estimation, under the assumption that the reference distribution is already known. Darbellay and Vajda [9] worked on the estimation of mutual information, namely, the divergence between the joint distribution and the product of the marginals. Their method is to approximate the mutual information directly by calculating relative frequencies on some data-driven partitions and achieving conditional local independence.

In this paper, we propose a universal divergence estimator for absolutely continuous distributions P and Q, based on independent and identically distributed (i.i.d.) samples generated from each source. The sources are allowed to be dependent and to output samples of different sizes. Our algorithm is inspired by the alternative expression for the divergence, i.e.,

$$D(P \,\|\, Q) \equiv \int_{\Omega} \mathrm{d}Q \frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q}. \tag{3}$$

In this formula, the Radon–Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ can be approximated by $\frac{\Delta P}{\Delta Q}$ as ΔQ diminishes, if P is absolutely continuous with respect to Q. Here ΔP and ΔQ denote empirical

Manuscript received November 11, 2004; revised April 27, 2005. This work was supported in part by ARL MURI under Grant DAAD19-00-1-0466, by Draper Laboratory under IR&D 6002 Grant DL-H-546263, and by the National Science Foundation under Grant CCR-0312413.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton NJ 08544 USA (e-mail: qingwang@princeton.edu; kulkarni@princeton.edu; verdu@princeton.edu).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2005.853314

probability measures of a segment in the σ-algebra $\mathcal{F}$. Our algorithm first partitions the space Ω into T contiguous segments such that each segment contains an equal number of sample points drawn from the reference measure Q (with the possible exception of one segment). Then by counting how many samples from P fall into each segment, we calculate the empirical measure ΔP of each segment. Note that ΔQ vanishes as T increases. Thus, the divergence can be estimated by the ratio between the empirical probability measures on each segment. The almost-sure convergence of this estimator is established under mild regularity conditions.

In this paper, although particular attention is given to the case of scalar observations, we also present results for the multivariate case. Furthermore, our algorithm can be used to estimate the divergence between non-i.i.d. data when the correlation structure is identical for both sources. For example, if a given invertible linear transform whitens both sources, then the algorithm can be applied at the output of the transform. We also prove that as long as the samples from the reference measure Q are i.i.d., then our divergence estimate of the marginals is strongly consistent when the data from measure P are stationary and ergodic.

The rest of the paper is organized as follows. In Section II, we present our algorithm for divergence estimation and prove strong consistency. Schemes to improve convergence speed are proposed in Section III, including a method for bias correction and data-driven partitioning schemes with an adaptive choice of the parameters. Section IV discusses the application of our algorithm to data with memory. Experimental results on randomly generated data are shown in Section V, where we compare our algorithm to the direct plug-in approach (where the two underlying densities are estimated separately and the two estimates are then substituted into (2)) and to the estimator based on the partitioning proposed by Darbellay and Vajda [9]. Simulations are also presented for correlated samples and samples with memory.

II. Divergence Estimation Via Data-Dependent Partitions

A. Algorithm A

We begin by stating our basic estimator, Algorithm A, in the one-dimensional case. Suppose P and Q are atom-free probability measures 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 generated from P and Q, respectively. Denote the order statistics of Y by $\{Y_{(1)}, Y_{(2)}, \ldots Y_{(m)}\}$ where $Y_{(1)} \le Y_{(2)} \le \ldots Y_{(m)}$. The real line is partitioned into empirically equivalent segments (called ℓ_m-spacings) according to

$$\{I_i^m\}_{i=1,\ldots,T_m} = \left\{\left(-\infty, Y_{(\ell_m)}\right], \left(Y_{(\ell_m)}, Y_{(2\ell_m)}\right], \ldots, \left(Y_{(\ell_m(T_m-1))}, +\infty\right)\right\} \quad (4)$$

where $\ell_m \in \mathbb{N} \le m$ is the number of points in each interval except possibly the last one, and $T_m = \lfloor m/\ell_m \rfloor$ is the number of intervals. For $i = 1, \ldots, T_m$, let k_i denote the number of samples from P that fall into the segment I_i^m.

In Algorithm A, the divergence between P and Q is estimated as

$$\hat{D}_{n,m}(P\,\|\,Q) = \sum_{i=1}^{T_m-1} \frac{k_i}{n} \log \frac{k_i/n}{\ell_m/m} + \frac{k_{T_m}}{n} \log \frac{k_{T_m}/n}{\ell_m/m + \delta_m} \quad (5)$$

where $\delta_m = (m - \ell_m T_m)/m$ is the correction term for the last segment.

Let P_n and Q_m be the corresponding empirical probability measures induced by the random samples X and Y, respectively. The divergence estimate (5) can be written as

$$\begin{aligned}\hat{D}_{n,m}(P\,\|\,Q) &= \sum_{i=1}^{T_m} P_n\left(I_i^m\right) \log \frac{P_n\left(I_i^m\right)}{Q_m\left(I_i^m\right)} \\ &= \sum_{i=1}^{T_m} Q_m\left(I_i^m\right) \frac{P_n\left(I_i^m\right)}{Q_m\left(I_i^m\right)} \log \frac{P_n\left(I_i^m\right)}{Q_m\left(I_i^m\right)}. \quad (6)\end{aligned}$$

In contrast to the direct plug-in method where the densities of P and Q are estimated separately with respect to the Lebesgue measure, our algorithm estimates the density of P with respect to Q, i.e., the Radon–Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$, which is guaranteed to exist provided $P \ll Q$.

Furthermore, this approach can be generalized to multidimensional data, by partitioning with statistically equivalent blocks of the space. Namely, according to the projections of the samples $Y_1, \ldots, Y_m$ onto the first coordinate axis, the space can be partitioned into T_m statistically equivalent cylindrical sets, where

$$T_m = \lfloor (m/\ell_m)^{1/d} \rfloor. \quad (7)$$

Then partition each cylindrical set along the second axis into T_m boxes, each of which contains the same number of points. Continuing in a similar fashion along the remaining axes produces T_m^d statistically equivalent rectangular cells. Based on this partition, the application of (6) gives an estimate of the divergence for multivariate distributions.

B. Convergence Analysis

In this section, we prove that Algorithm A is strongly consistent.

Theorem 1: Let P and Q be absolutely continuous probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Assume that the divergence between P and Q is finite. Let $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ be i.i.d. samples generated from P and Q, respectively.[1] Let ℓ_m, T_m be defined as in (4). If $\ell_m, T_m \to \infty$ as $m \to \infty$, then the divergence estimator in (6) satisfies

$$\hat{D}_{m,n}(P\,\|\,Q) \to D(P\,\|\,Q) \quad \text{a.s.} \quad (8)$$

as $n, m \to \infty$.

The reliability of our estimator depends not only on the proximity of the empirical probability measure to the true measure but also on the proximity of the empirically equivalent partition to the true equivalent partition. To resolve the second issue, we

[1] Recall that we are not assuming independence of $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$

first introduce the following concept, which defines the convergence of a sequence of partitions.

Definition 1: Let $(\Omega, \mathcal{F})$ be a measurable space and ν be a probability measure defined on this space. Let $\{I_1, I_2, \dots, I_T\}$ be a finite measurable partition of Ω. A sequence of partitions $\{I_1^m, I_2^m, \dots, I_T^m\}_m$ of Ω is said to converge to $\{I_1, I_2, \dots, I_T\}$ with respect to ν as $m \to \infty$ if for any probability measure μ on $(\Omega, \mathcal{F})$ that is absolutely continuous with respect to ν, we have

$$\lim_{m\to\infty} \mu\left(I_i^m\right) = \mu(I_i) \quad \text{a.s. for each } i \in \{1, 2, \dots, T\}.$$

The following result shows the convergence of the data-dependent sequence of partitions $\{I_i^m\}_{i=1,2,\dots,T_m}$ when T_m is a fixed constant T.

Lemma 1: Let Q be an absolutely continuous probability measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and $\{I_i\}_{i=1,\dots,T}$ be a fixed finite partition of the real line: $I_i = (a_{i-1}, a_i](i = 1, 2, \dots, T)$, where $-\infty = a_0 < a_1 < \cdots < a_{T-1} < a_T = +\infty$ such that

$$Q(I_i) = \frac{1}{T}, \qquad i = 1, 2, \dots, T. \tag{9}$$

Let Q_m be the empirical probability measure based on i.i.d. samples $\{Y_1, Y_2, \dots, Y_m\}$ generated from Q with $m = \ell_m T, \ell_m \in \mathbb{N}$. Let

$$\begin{aligned}\{I_i^m\}_{i=1,\dots,T} = \{(-\infty, a_1^m], (a_1^m, a_2^m], \dots,\\ \left(a_{T-2}^m, a_{T-1}^m\right], \left(a_{T-1}^m, +\infty\right)\}\end{aligned} \tag{10}$$

be a partition such that

$$Q_m\left(I_i^m\right) = \frac{1}{T}, \qquad i = 1, 2, \dots, T. \tag{11}$$

Then the sequence of partitions $\{\{I_i^m\}_{i=1,\dots,T}\}_m$ converges to the partition $\{I_i\}_{i=1,\dots,T}$ with respect to Q as $m \to \infty$.

Before proving Lemma 1, we invoke a result by Lugosi and Nobel [10] that specifies sufficient conditions on the partition of the space under which the empirical measure converges to the true measure.

Let $\mathcal{A}$ be a family of partitions of $\mathbb{R}^d$. The maximal cell count of $\mathcal{A}$ is given by

$$c(\mathcal{A}) = \sup_{\pi\in\mathcal{A}} |\pi| \tag{12}$$

where $|\pi|$ denotes the number of cells in partition π.

The complexity of $\mathcal{A}$ is measured by the growth function as described below. Fix n points in $\mathbb{R}^d$

$$x_1^n = \{x_1, \dots, x_n\}. \tag{13}$$

Let $\Delta(\mathcal{A}, x_1^n)$ be the number of distinct partitions

$$\{A_1 \cap x_1^n, \dots, A_r \cap x_1^n\}$$

of the finite set x_1^n that can be induced by partitions $\pi = \{A_1, \dots, A_r\} \in \mathcal{A}$. Define the *growth function* of $\mathcal{A}$ as

$$\Delta_n^*(\mathcal{A}) = \max_{x_1^n \in \mathbb{R}^{n\cdot d}} \Delta\left(\mathcal{A}, x_1^n\right) \tag{14}$$

which is the largest number of distinct partitions of any n-point subset of $\mathbb{R}^d$ that can be induced by the partitions in $\mathcal{A}$.

Proposition 1: (Lugosi and Nobel [10]): Let $Y_1, Y_2, \dots$ be i.i.d. random vectors in $\mathbb{R}^d$ with $Y_i \sim \mu$ and let μ_m denote the empirical probability measure based on m samples. Given a sequence of partition families $\mathcal{A}_1, \mathcal{A}_2, \dots$, if, as $m \to \infty$, a) $m^{-1}c(\mathcal{A}_m) \to 0$ and b) $m^{-1}\log\Delta_m^*(\mathcal{A}_m) \to 0$, then

$$\sup_{\pi\in\mathcal{A}_m} \sum_{A\in\pi} |\mu_m(A) - \mu(A)| \to 0 \quad \text{a.s.} \tag{15}$$

We are now ready to prove Lemma 1.

Proof of Lemma 1: First we show the convergence of the empirical measure Q_m to the true measure Q on the partition $\{I_i^m\}$. In fact, the two conditions of Proposition 1 are satisfied. Suppose $\mathcal{A}_m$ is the collection of all the partitions of $\mathbb{R}$ into T empirically equiprobable intervals based on m sample points. Then

$$c(\mathcal{A}_m) = \sup_{\pi\in\mathcal{A}_m} |\pi| = T. \tag{16}$$

Since $\ell_m = m/T \to \infty$ as $m \to \infty$, we have that

$$m^{-1}c(\mathcal{A}_m) = 1/\ell_m \to 0. \tag{17}$$

Next consider the growth function $\Delta_m^*(\mathcal{A}_m)$ which is defined as the largest number of distinct partitions of any m-point subset of $\mathbb{R}^d$ that can be induced by the partitions in $\mathcal{A}_m$. Namely

$$\Delta_m^*(\mathcal{A}_m) = \max_{Y_1^m \in \mathbb{R}^{dm}} \Delta\left(\mathcal{A}_m, Y_1^m\right). \tag{18}$$

In our algorithm, the partitioning number $\Delta_m^*(\mathcal{A}_m)$ is the number of ways that m fixed points can be partitioned by T intervals. Then

$$\Delta_m^*(\mathcal{A}_m) = \binom{m+T}{T}. \tag{19}$$

Let h be the binary entropy function, defined as

$$h(x) = -x\log(x) - (1-x)\log(1-x), \text{ for } x \in (0,1). \tag{20}$$

By the inequality $\log\binom{s}{t} \le sh(t/s)$, we obtain

$$\log\Delta_m^*(\mathcal{A}_m) \le (m+T)h\left(\frac{T}{m+T}\right) \le 2mh\left(\frac{1}{\ell_m}\right). \tag{21}$$

As $\ell_m \to \infty$, the last inequality implies that

$$\frac{1}{m}\log\Delta_m^*(\mathcal{A}_m) \to 0. \tag{22}$$

With (17) and (22), applying Proposition 1, we have that

$$\lim_{m\to\infty} \sum_{i=1}^{T} \left|Q_m\left(a_{i-1}^m, a_i^m\right] - Q\left(a_{i-1}^m, a_i^m\right]\right| = 0 \tag{23}$$

where $a_0^m = -\infty, a_T^m = +\infty$. Obviously

$$\lim_{m\to\infty} |Q_m(-\infty, a_i^m] - Q(-\infty, a_i^m]| = 0 \text{ a.s.} \tag{24}$$

for $i = 1, 2, \dots, T$, which implies that

$$\lim_{m\to\infty} Q(-\infty, a_i^m] = \frac{i}{T} \text{ a.s.} \tag{25}$$

or

$$\lim_{m\to\infty} Q\left(a_i \wedge a_i^m, a_i \vee a_i^m\right] = 0 \text{ a.s.} \tag{26}$$

for each $i = 1, 2, \dots, T$.

Thus, for any μ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ absolutely continuous with respect to Q

$$\lim_{m\to\infty} \mu\,(a_i \wedge a_i^m, a_i \vee a_i^m] = 0, \qquad \text{a.s. } i = 1, 2, \ldots, T. \tag{27}$$

Alternatively

$$\lim_{m\to\infty} \mu\,(-\infty, a_i^m] = \mu(-\infty, a_i], \qquad \text{a.s. } i = 1, 2, \ldots, T. \tag{28}$$

Therefore,

$$\begin{aligned}\lim_{m\to\infty} \mu\left(I_i^m\right) &= \lim_{m\to\infty} \left(\mu\left(-\infty, a_i^m\right] - \mu\left(-\infty, a_{i-1}^m\right]\right) \\ &= \mu(-\infty, a_i] - \mu(-\infty, a_{i-1}] \\ &= \mu(I_i), \text{ a.s.}\end{aligned} \tag{29}$$

for any $i = 1, 2, \ldots, T$. □

In Lemma 1, we consider probability measures which are absolutely continuous with respect to the reference measure. However, in our universal estimation problem, those measures are not known and thus are replaced by their empirical versions. Lemma 2 shows that the corresponding empirical probability measures also satisfy similar properties when the sample size goes to infinity.

Lemma 2: Let ν be a probability measure on $(\Omega, \mathcal{F})$ and let $\{I_1, I_2, \ldots, I_T\}$ and $\{I_1^m, I_2^m, \ldots, I_T^m\}_{m=1,2,\ldots}$ be finite measurable partitions of Ω. Let μ be an arbitrary probability measure on $(\Omega, \mathcal{F})$, which is absolutely continuous with respect to ν. Suppose μ_n is the empirical probability measure based on i.i.d. samples $\{X_1, \ldots, X_n\}$ generated from μ. If $\{I_1^m, I_2^m, \ldots, I_T^m\}_m$ converges to $\{I_1, I_2, \ldots, I_T\}$ with respect to (w.r.t.) ν as $m \to \infty$, then

$$\lim_{m\to\infty}\lim_{n\to\infty} \mu_n\left(I_i^m\right) = \mu(I_i) \quad \text{a.s. } i = 1, 2, \ldots, T. \tag{30}$$

Proof: Note that

$$\mu_n\left(I_i^m\right) = \frac{1}{n}\sum_{s=1}^{n} 1_{I_i^m}(X_s) \tag{31}$$

where

$$\mathrm{E}\left(1_{I_i^m}(X_s)\right) = \mu(I_i^m). \tag{32}$$

Therefore, by the strong law of large numbers, we have

$$\lim_{n\to\infty} \mu_n\left(I_i^m\right) = \mu\left(I_i^m\right) \quad \text{a.s.}. \tag{33}$$

Further, since the convergence of $\{\{I_i^m\}_{i=1,\ldots,T}\}_m$ implies that

$$\lim_{m\to\infty} \mu\left(I_i^m\right) = \mu(I_i) \quad \text{a.s. for each } \; i \in \{1, 2, \ldots, T\} \tag{34}$$

taking $m \to \infty$ on both sides of (33) gives (30). □

The strong consistency of our divergence estimator is proved in the following.

Proof of Theorem 1: Define

$$I_i = (a_{i-1}, a_i] \qquad (i = 1, 2, \ldots, T_m)$$

where $-\infty = a_0 < a_1 < \cdots < a_{T_m - 1} < a_{T_m} = +\infty$ such that

$$Q(I_i) = \frac{1}{T_m}, \qquad i = 1, 2, \ldots, T_m. \tag{35}$$

Namely, $\{I_i\}_{i=1,\ldots,T_m}$ is the equiprobable partition of the real line according to Q.

The estimation error can be decomposed as

$$\begin{aligned}&|\hat{D}_{m,n}(P \,\|\, Q) - D(P \,\|\, Q)| \\ &\leq \left|\sum_{i=1}^{T_m} P_n\left(I_i^m\right)\log\frac{P_n\left(I_i^m\right)}{Q_m\left(I_i^m\right)} - \sum_{i=1}^{T_m} P(I_i)\log\frac{P\left(I_i\right)}{Q(I_i)}\right| \\ &\quad + \left|\sum_{i=1}^{T_m} Q(I_i)\frac{P(I_i)}{Q(I_i)}\log\frac{P(I_i)}{Q(I_i)} - \int_\Omega \mathrm{d}Q\frac{\mathrm{d}P}{\mathrm{d}Q}\log\frac{\mathrm{d}P}{\mathrm{d}Q}\right| \\ &= e_1 + e_2.\end{aligned} \tag{36}$$

Intuitively, e_2 is the approximation error caused by numerical integration, which diminishes as T_m increases; e_1 is the estimation error caused by the difference of the statistically equivalent partitions from the true equiprobable partitions and the difference of the empirical probability on an interval from its true probability. The term e_1 can be shown to be arbitrarily small when $\ell_m, m,$ and n are sufficiently large, using Lemmas 1 and 2. The details are as follows.

Since

$$\int_\Omega \mathrm{d}Q\frac{\mathrm{d}P}{\mathrm{d}Q}\log\frac{\mathrm{d}P}{\mathrm{d}Q} < \infty$$

then for any $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that $e_2 < \epsilon/2$ as long as $\max_i Q(I_i) \leq \delta$. Recall that $Q(I_i) = 1/T_m$, so

$$e_2 < \epsilon/2 \quad \text{a.s. for any } T_m \geq 1/\delta. \tag{37}$$

Now fix a $T \geq T_m(\epsilon) = 1/\delta(\epsilon)$. With $\ell_m \to \infty$ as $m \to \infty$, Lemma 1 tells us that the sequence of partitions $\{\{I_i^m\}_{i=1,2,\ldots,T}\}_m$ converges to $\{I_i\}_{i=1,\ldots,T}$ with respect to measure Q as $m \to \infty$. Further, since $P \ll Q$, by Lemma 2, it is guaranteed that

$$\lim_{m\to\infty}\lim_{n\to\infty} P_n\left(I_i^m\right) = P(I_i) \quad \text{a.s.} \tag{38}$$

for any $i \in \{1, 2, \ldots, T\}$. Also, note that $Q_m(I_i^m) = Q(I_i) = 1/T$. Moreover, since

$$f(x) = x\log\frac{x}{1/T} \tag{39}$$

is a continuous function in x, if $f(0) \triangleq 0$, then for any $\epsilon > 0$, there exists $N_i = n_i(\epsilon, T) > 0, M_i = m_i(N_i, \epsilon, T)$, such that

$$\left|P_n\left(I_i^m\right)\log\frac{P_n\left(I_i^m\right)}{Q_m\left(I_i^m\right)} - P(I_i)\log\frac{P(I_i)}{Q(I_i)}\right| < \frac{\epsilon}{2T} \quad \text{a.s.} \tag{40}$$

for any $n \geq N_i$, $m \geq M_i$, any $i \in \{1, 2, \ldots, T\}$. Therefore,

$$e_1 \leq \sum_{i=1}^{T}\left|P_n\left(I_i^m\right)\log\frac{P_n\left(I_i^m\right)}{Q_m\left(I_i^m\right)} - P(I_i)\log\frac{P(I_i)}{Q(I_i)}\right| < \frac{\epsilon}{2} \tag{41}$$

with $n \geq \max_i N_i$ and $m \geq \max_i M_i$. Together with (37), this completes the proof. □

III. Schemes to Improve Convergence Speed

In the previous section, our divergence estimator is proved to be asymptotically consistent. However, in reality, we are only provided with samples of finite sizes. The problem is how to obtain a reliable estimate when the number of samples is limited. In this section, we propose two approaches to accelerate convergence. Section III-A shows how to choose the algorithm parameters (such as the number of segments) as a function of the data. Although as shown in Theorem 1 the universal estimator is consistent, the estimation bias only vanishes as the data size increases. In Section III-B we examine how to reduce the bias for any given sample size.

A. A Data-Driven Choice of ℓ_m

In the area of kernel density estimation, there is a large amount of literature dealing with the optimal choice of window width to achieve the least error or the fastest convergence speed. In our algorithm, ℓ_m plays a role analogous to that of window width. By finely tuning the value of ℓ_m, we can improve the accuracy of the estimation. Basically, our divergence estimator examines how differently the samples from P and Q are distributed among the segments. Note that there is a tradeoff in the choice of the number of segments: the larger the number, the better we can discriminate signature details from the distributions; the smaller the number, the more accurate are the empirical histograms.

1) Algorithm B: Global Adaptive Method: This method updates ℓ_m uniformly through the space according to the initial estimation result. Set $\ell_m = \ell_0$ (e.g., $\lfloor\sqrt{m}\rfloor$). Basically, if the estimate $\hat{D}_0$ at ℓ_0 is low (resp., high), ℓ_m is updated by $f(\ell_0) >$ (resp., $<$)ℓ_0. The estimate at the $f(\ell_0)$ will be our final estimate if $\hat{D}_0 \in (\epsilon, \log T_m)$, where ϵ is a small positive number. Otherwise, if $\hat{D}_0$ is too low (less than ϵ), we need to test whether the two distributions are homogeneous. Or if $\hat{D}_0$ is too high, it is necessary to verify whether the sample size is large enough to provide a reliable divergence estimate. The parameters associated with this algorithm include the following:

- thresholds: what is the criterion to determine whether the divergence estimate is high or low;
- updating functions: how to adapt ℓ_m in each region;
- initial value of ℓ_m.

This method is particularly suitable for detecting whether the two underlying distributions are identical, since we have found that $\hat{D}$ diminishes at a rate of roughly $\frac{1}{\ell_m}$ when the the true divergence is 0. A drawback of this method is that it is difficult to optimize the choice of the above parameters. Also, local details might be lost if the same ℓ_m is assigned to the regions where the two distributions appear to be similar and the regions where they are quite mismatched.

2) Algorithm C: Local Adaptive Method 1: Instead of insisting on uniform partitioning of the space, this local updating scheme produces a fine partition in the region where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is high and a coarser partition elsewhere, the rationale being that not much accuracy is required in the zones where $\frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q}$ is low.

Let $\ell_{\min}$ be the smallest possible number of data points from Q in each segment, which represents the finest possible partition. k_i denotes how many data points from P fall in each segment in the new partition. T^* is the total number of segments. $\alpha > 1$ is a parameter regulating whether a further partition is necessary. The procedure is to partition the space such that each segment contains $\ell_m = \ell_0$ number of sample points from Q. Scan through all the segments. In segment i, if $\ell_m > \ell_{\min}$ and $k_i > \alpha\ell_m$, update ℓ_m by $f(\ell_m)$ and again partition segment i statistically equiprobably into ℓ_m subsegments. Continue this process until either $k_i \leq \alpha\ell_m$ or the updated $\ell_m \leq \ell_{\min}$.

The new adaptive estimate is simply

$$\hat{D}^*_{n,m}(P\,\|\,Q) = \sum_{i=1}^{T^*} \frac{k_i}{n} \log \frac{k_i/n}{\ell_i^*/m} \tag{42}$$

where ℓ_i^* is the number of Y's in each segment.

3) Algorithm D: Local Adaptive Method 2: Another version of nonuniform adaptive partition is inspired by Darbellay and Vajda's method [9]. The idea is to first segment the space into T_0 equivalent blocks. For any block I, if the following condition holds:

$$\sum_{I_i \in s} \frac{P_n(I_i)}{P_n(I)} \log \frac{P_n(I_i)Q_n(I)}{P_n(I)Q_n(I_i)} < \epsilon, \qquad \text{for all } s \in S. \tag{43}$$

No further segmentation will be imposed on I, where S represents all testing partitioning schemes. However, there are two possible problematic situations with this terminating condition, which may cause bad behavior of the estimator.

- Over-fine partitioning:
$$\sum_{I_i \in s} \frac{P_n(I_i)}{P_n(I)} \log \frac{P_n(I_i)Q_n(I)}{P_n(I)Q_n(I_i)}$$
can be occasionally large and induce an unnecessary further partition. This happens when $P_n(I) < Q_n(I)$, but
$$\frac{P_n(I_i)}{P_n(I)} \log \frac{P_n(I_i)Q_n(I)}{P_n(I)Q_n(I_i)}$$
is large, for some $I_i \in s$.
- Over-coarse partitioning: Suppose $P_n(I) \gg Q_n(I)$, which means that a finer partition is needed to reveal more details. However, it might happen that
$$\frac{P_n(I_i)}{P_n(I)} \approx \frac{Q_n(I_i)}{Q_n(I)}$$
for example, when the two distributions are both uniform. In that case, the termination condition (43) will be satisfied and thus no further partition will be imposed.

Thus, instead of using local conditional similarity, in our local adaptive approach, the terminating condition is now modified as

$$\sum_{I_i \in s} P_n(I_i) \log \frac{P_n(I_i)}{Q_n(I_i)} < \epsilon, \qquad \text{for all } s \in S \tag{44}$$

which implies that no finer partition is necessary if the contribution to the divergence estimate by a further partitioning is small enough.

Simulation results are presented in Section V to compare the performance of estimators using different partitions.

B. Algorithm E: Reducing the Bias

Let $\hat{D}_{n,m}$ be the divergence estimate and the operator $< \cdot >$ denote the expectation taken with respect to the i.i.d. random variables. The systematic bias of $\hat{D}_{n,m}$ can be expressed as

$$\begin{aligned} B_{n,m} &\equiv \langle \hat{D}_{n,m} \rangle - D \\ &= \left\langle \sum_{i=1}^{T_m} P_n\left(I_i^m\right) \log \frac{P_n\left(I_i^m\right)}{Q_m\left(I_i^m\right)} \right\rangle - \int_\Omega \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}. \end{aligned} \tag{45}$$

Suppose we were to use a fixed data-independent partition $\{I_i\}_{i=1,\ldots,T}$ that satisfies $Q(I_i) = 1/T$ for each $i = 1, \ldots, T$. We also assume that the samples X and Y are independent of each other. Then the estimate bias is

$$\begin{aligned} B_{n,m} &= \left\langle \sum_{i=1}^{T} P_n(I_i) \log \frac{P_n(I_i)}{Q_m(I_i)} \right\rangle - \int_\Omega \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q} \\ &= \left\langle \sum_{i=1}^{T} P_n(I_i) \log \frac{P_n(I_i)}{Q_m(I_i)} \right\rangle - \sum_{i=1}^{T} P(I_i) \log \frac{P(I_i)}{Q(I_i)} \\ &\quad + \sum_{i=1}^{T} P(I_i) \log \frac{P(I_i)}{Q(I_i)} - \int_\Omega \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}. \end{aligned} \tag{46}$$

We only consider the first difference, since the second one depends on the two underlying distributions and it involves knowledge of the true divergence, which is to be estimated. Let $f(x,y) = x \log(x/y)$. Expanding the first summation term in the preceding equation at $\{(P(I_i), Q(I_i))\}_{i=1,\ldots,T}$ by the Taylor series of $f(x,y)$, we obtain

$$\begin{aligned} &\left\langle \sum_{i=1}^{T} P_n(I_i) \log \frac{P_n(I_i)}{Q_m(I_i)} \right\rangle - \sum_{i=1}^{T} P(I_i) \log \frac{P(I_i)}{Q(I_i)} \\ &= \left\langle \sum_{i=1}^{T} \left[(P_n(I_i) - P(I_i)) \left.\frac{\partial f}{\partial x}\right|_{x=P(I_i)} \right.\right. \\ &\qquad \left.\left. + \ (Q_m(I_i) - Q(I_i)) \left.\frac{\partial f}{\partial y}\right|_{y=Q(I_i)} \right] \right\rangle \\ &\quad + \left\langle \sum_{i=1}^{T} \left[(P_n(I_i) - P(I_i))^2 \left.\frac{\partial^2 f}{\partial x^2}\right|_{x=P(I_i)} \right.\right. \\ &\qquad + 2(P_n(I_i) - P(I_i))(Q_m(I_i) \\ &\qquad -Q(I_i)) \cdot \left.\frac{\partial^2 f}{\partial x \partial y}\right|_{x=P(I_i),y=Q(I_i)} \\ &\qquad \left.\left. +(Q_m(I_i) - Q(I_i))^2 \left.\frac{\partial^2 f}{\partial y^2}\right|_{y=Q(I_i)} \right] \right\rangle \\ &\quad + o\left(\frac{1}{m} + \frac{1}{n}\right). \end{aligned} \tag{47}$$

Note that the first expectation is zero since it is linear in the empirical probability measure. The expectation of the mixed term is also zero if the samples from P and Q are assumed to be independent. Now consider

$$(P_n(I_i) - P(I_i))^2 \left.\frac{\partial^2 f}{\partial x^2}\right|_{x=P(I_i)}$$

where $P_n(I_i)$ can be viewed as the average of n independent binomial random variables with expectation $P(I_i)$. Therefore,

$$\langle (P_n(I_i) - P(I_i))^2 \rangle = P(I_i)(1 - P(I_i))/n. \tag{48}$$

And since

$$\left.\frac{\partial^2 f}{\partial x^2}\right|_{x=P(I_i)} = 1/P(I_i) \tag{49}$$

we have that

$$\begin{aligned} &\frac{1}{2} \left\langle \widehat{\sum_i} (P_n(I_i) - P(I_i))^2 \left.\frac{\partial^2 f}{\partial x^2}\right|_{x=P(I_i)} \right\rangle \\ &\quad = \frac{1}{2} \widehat{\sum_i} \frac{P(I_i)(1 - P(I_i))}{n} \frac{1}{P(I_i)} = \frac{T_p - 1}{2n} \end{aligned} \tag{50}$$

where $\widehat{\sum}_i$ denotes that the sum is taken on all segments I_i with $P(I_i) > 0$ and T_p is the number of these segments. Observe that (50) depends on the true probability only through T_p. In the same fashion, we obtain

$$\begin{aligned} &\frac{1}{2} \left\langle \widetilde{\sum_i} (Q_m(I_i) - Q(I_i))^2 \left.\frac{\partial^2 f}{\partial y^2}\right|_{y=Q(I_i)} \right\rangle \\ &\quad = \frac{1}{2} \widetilde{\sum_i} \frac{Q(I_i)(1 - Q(I_i))}{m} \frac{P(I_i)}{Q^2(I_i)} \\ &\quad = \frac{1}{2m} \left[\widetilde{\sum_i} \sum_i \frac{P(I_i)}{Q(I_i)} - 1 \right] = \frac{T - 1}{2m} \end{aligned} \tag{51}$$

where $\widetilde{\sum}_i$ is the sum taken on all segments I_i with $Q(I_i) > 0$. Therefore, the estimate bias can be approximated as

$$\hat{B}_{n,m} = \frac{T_p - 1}{2n} + \frac{T - 1}{2m} + o\left(\frac{1}{m} + \frac{1}{n}\right). \tag{52}$$

We can improve the estimator by subtracting the first two terms on the right-hand ide of (52). Experimental results show that the approximation of the bias in (52) is excellent when the true divergence is not too high.

IV. Sources With Memory

We have elaborated on divergence estimation with i.i.d. data. If the sources have memory and they are such that a one-to-one transform is available that whitens both samples simultaneously, then our divergence estimator is readily applicable at the outputs of the transform, yielding an estimate of the divergence rate of the original sources.

When the samples from the reference measure are i.i.d. with distribution Q, but the samples with marginal P have memory, our algorithm still generates consistent estimates of the divergence between the marginals $D(P \,\|\, Q)$, instead of the divergence rate

$$\limsup_{n \to \infty} \frac{1}{n} D(P_{X_1,\ldots,X_n} \,\|\, P_{Y_1,\ldots,Y_n}). \tag{53}$$

This divergence estimate can be used to gauge the additional redundancy of Huffman symbol-by-symbol compressors that assume mismatched distributions.

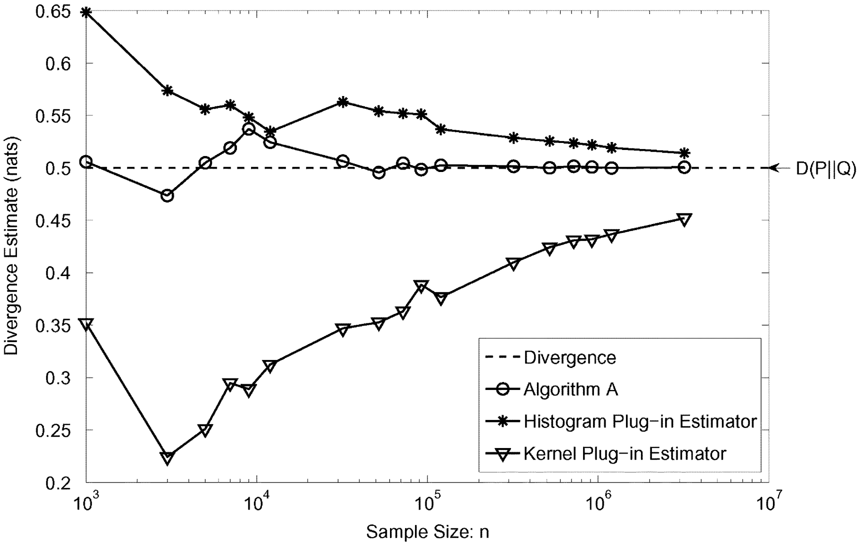

Fig. 1. $X \sim P = \mathrm{N}(0,1), Y \sim Q = \mathrm{N}(1,1); D(P \,\|\, Q) = 0.5$. In Algorithm A, $\ell_m = \lfloor \sqrt{m} \rfloor$. In the plug-in estimators, the underlying density is evaluated with $\lfloor \sqrt{n} \rfloor$ (resp., $\lfloor \sqrt{m} \rfloor$) points based on samples from P (resp., from Q.) The window width is chosen to be optimal for Gaussian distributions for the kernel density estimator. The choice of parameters is the same for each algorithm in the experiments.

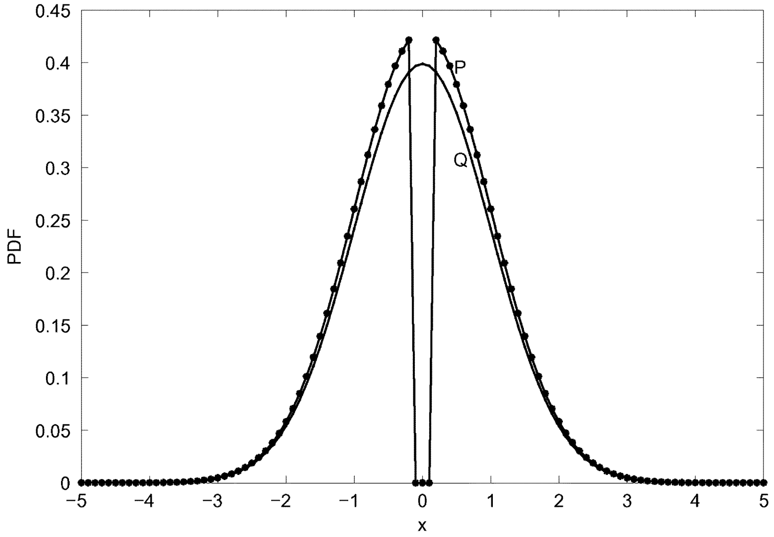

Fig. 2. $X \sim P = \mathrm{N}(0,1)$ with dip, $Y \sim Q = \mathrm{N}(0,1)$.

Let $\{X_1, \ldots, X_n\}$ be stationary ergodic samples with marginal distribution $X_i \sim P$ and $\{Y_1, \ldots, Y_m\}$ be i.i.d. samples generated according to Q. Since $\{Y_1, \ldots, Y_m\}$ are i.i.d., the convergence of the ℓ_m-spacing partitions is still valid according to Lemma 1. Then by invoking the stationarity and ergodicity of sample X, Lemma 2 can be generalized as follows.

Lemma 3: Let ν be a probability measure on $(\Omega, \mathcal{F})$ and let $\{I_1, I_2, \ldots, I_T\}$ and $\{I_1^m, I_2^m, \ldots, I_T^m\}_{m=1,2,\ldots}$ be finite measurable partitions of Ω. Let μ be an arbitrary probability measure on $(\Omega, \mathcal{F})$, which is absolutely continuous with respect to ν. Suppose μ_n is the empirical probability measure based on stationary and ergodic samples $\{X_1, \ldots, X_n\}$ with $X_i \sim \mu$. If $\{I_1^m, I_2^m, \ldots, I_T^m\}_m$ converges to $\{I_1, I_2, \ldots, I_T\}$ w.r.t. ν as $m \to \infty$, then

$$\lim_{m\to\infty} \lim_{n\to\infty} \mu_n\left(I_i^m\right) = \mu(I_i) \quad \text{a.s. } i = 1, 2, \ldots, T. \tag{54}$$

Proof: The only difference from the Proof of Lemma 2 is that (33) is obtained by using the fact that X is stationary and ergodic. □

The main result in this section is as follows.

Theorem 2: Let P and Q be absolutely continuous probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Assume that the divergence between P and Q is finite. Let $\{X_1, X_2, \ldots, X_n\}$ be stationary and ergodic samples with $X_i \sim P$ and $\{Y_1, Y_2, \ldots, Y_m\}$ be i.i.d. samples generated from Q. Let ℓ_m, T_m be defined as in (4). If $\ell_m, T_m \to \infty$ as $m \to \infty$, then the divergence estimator in (6) satisfies

$$\hat{D}_{m,n}(P \,\|\, Q) \to D(P \,\|\, Q) \quad \text{a.s.} \tag{55}$$

as $n, m \to \infty$.

The proof is the same as that of Theorem 1, except that Lemma 3 is applied instead of Lemma 2.

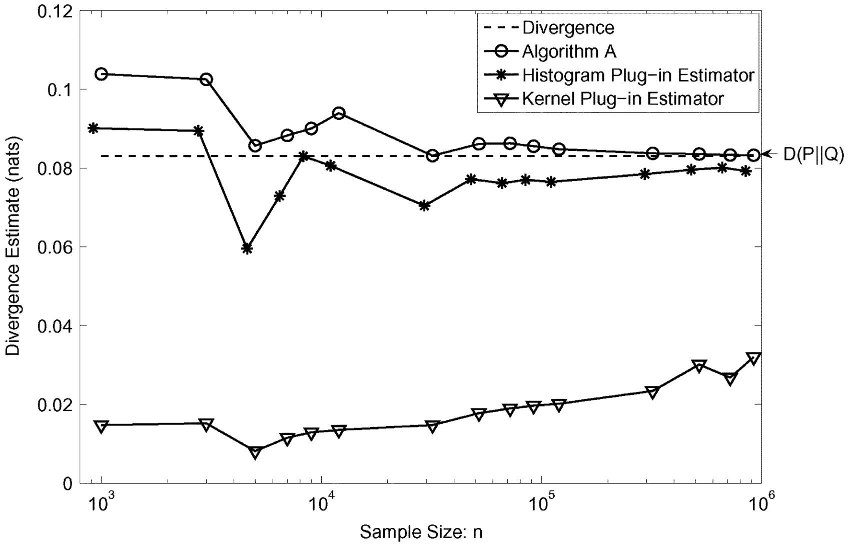

Fig. 3. $X \sim P = \mathrm{N}(0,1)$ with dip, $Y \sim Q = \mathrm{N}(0,1)$; $D(P \,\|\, Q) = 0.0830$.

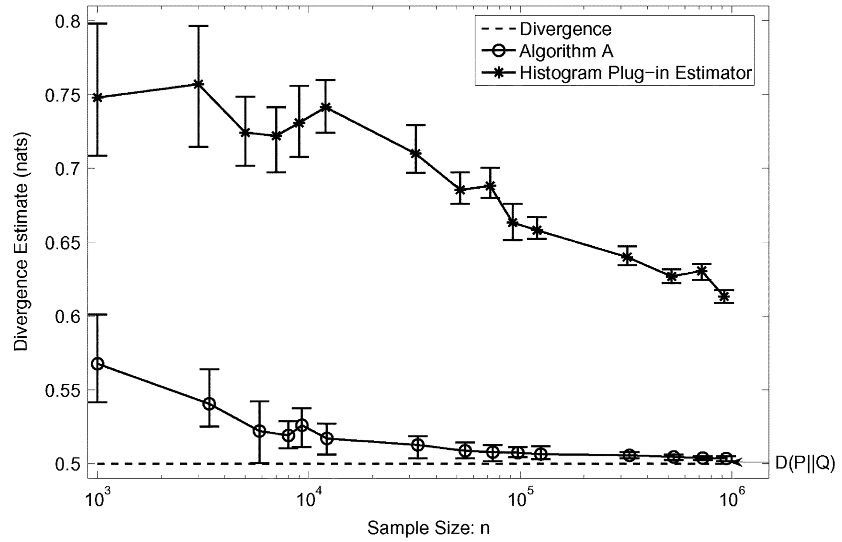

Fig. 4. $P \sim \mathrm{N}(0,1) \times \mathrm{N}(0,1), Q \sim \mathrm{N}(1,1) \times \mathrm{N}(0,1); D(P \,\|\, Q) = 0.5$.

Now let us consider a more general setting where samples from the reference measure have memory. In this case, if both samples have only short dependency, then our algorithms can be applied to samples which are far apart. Otherwise, if the samples from reference measure with marginal Q are stationary ergodic, we conjecture that it is not possible to design a consistent estimator of the divergence $D(P \,\|\, Q)$. The reasoning is that the consistency of our estimator is built upon the validity of Vapnik–Chervonenkis (VC) Inequality, where data are assumed to be i.i.d. Then the question arises whether there are similar results for stationary ergodic data. However, in the proofs in [16, pp. 269] and [17, pp. 194], the independence assumption cannot be removed. In fact, in the event that the VC inequality could be generalized to stationary ergodic data, then following the arguments in [10], we can find a density estimate which is strongly L^1-consistent given stationary ergodic data. However, as proved in [18], no density estimation procedure is weakly L^1-consistent for every stationary ergodic process with an absolutely continuous marginal distribution.

V. Experiments

In Figs. 1 and 3, we compare the performance of Algorithm A on scalar data with that of the direct plug-in estimators, which are based on kernel density estimation and the histogram estimation method [10] respectively. (Each dot in the curve represents an estimate given one set of samples.) The distributions associated with estimates in Fig. 2 are shown in Fig. 3. Note that the kernel method is not performing well when the underlying distributions are not smooth enough. Fig. 4 presents experiments for two-dimensional data. The solid line represents the average of the estimates based on 25 sets of samples. The upper/lower error bar corresponds to the mean of the differences between the estimate average and the estimates which are larger/smaller than the average. Fig. 5 shows that Algorithm A works well even when there is correlation between samples from P and Q.

Figs. 6 and 7 demonstrate the advantage of adaptive versions of the divergence estimator over the basic one for very similardistributions and quite mismatched ones, respectively. The simulations are performed on one set of samples.

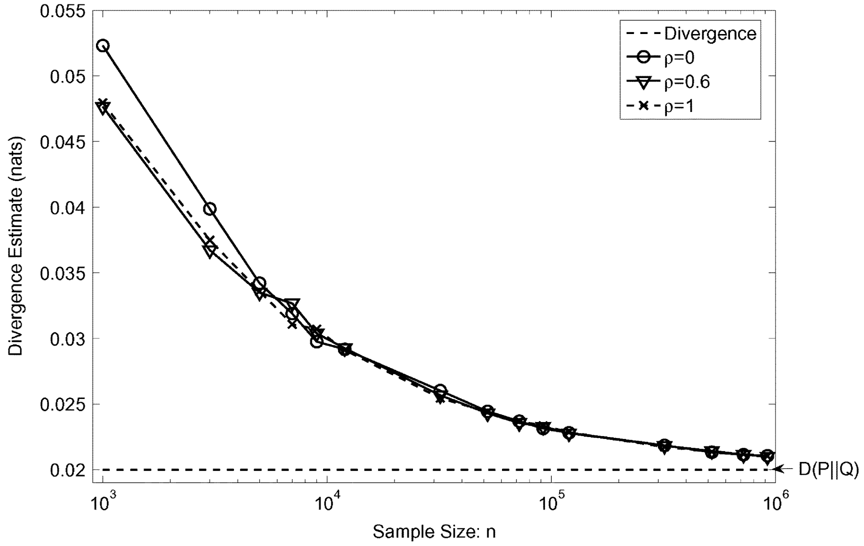

Fig. 5. $X \sim P = \mathrm{N}(0,1), Y \sim Q = \mathrm{N}(0.2,1); D(P \,\|\, Q) = 0.02$. ρ is the correlation coefficient.

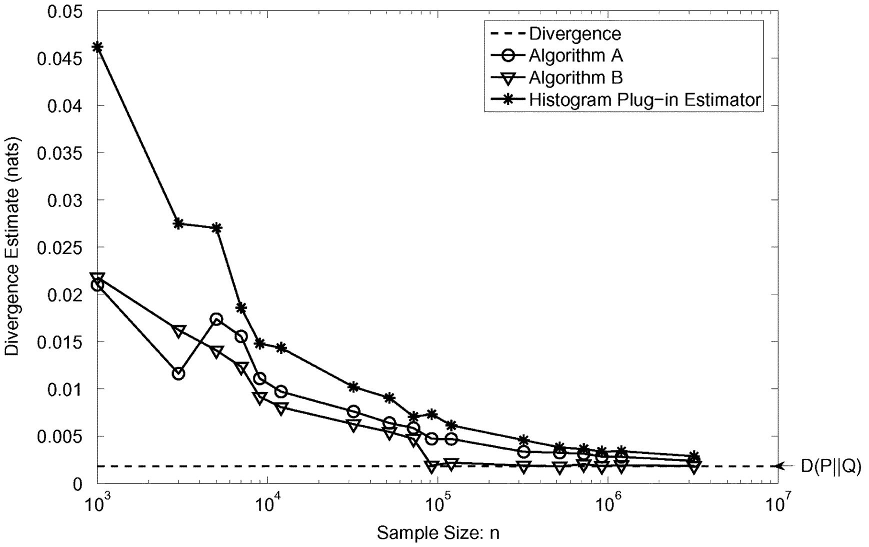

Fig. 6. $X \sim P = \mathrm{N}(0,1), Y \sim Q = \mathrm{N}(0.06,1); D(P \,\|\, Q) = 0.0018$. For Algorithm B, since the estimation result with $\ell_m = \lfloor \sqrt{m} \rfloor$ is very low, we first test the homogeneity of the two distributions. According to the test, the two distributions are not identical. Then ℓ_m is updated to $\lfloor m^{0.7} \rfloor$.

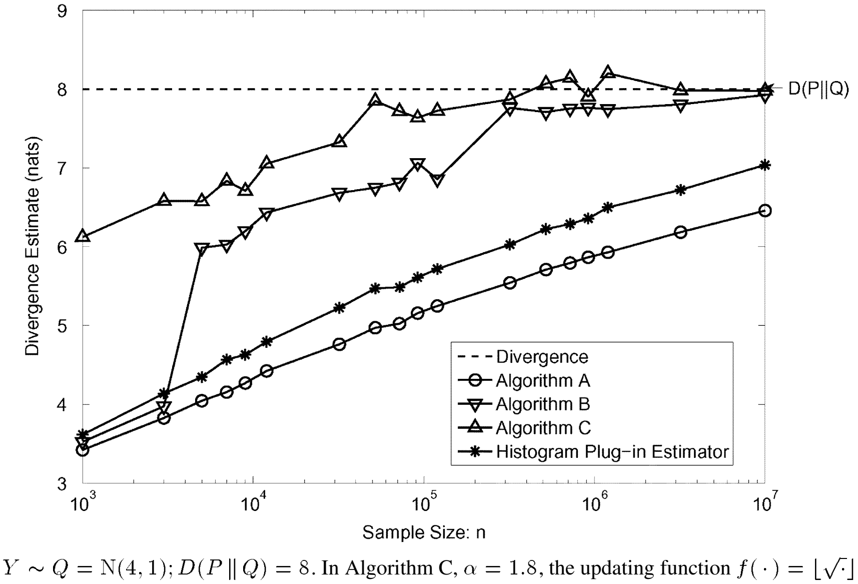

Fig. 7. $X \sim P = \mathrm{N}(0,1), Y \sim Q = \mathrm{N}(4,1); D(P \,\|\, Q) = 8$. In Algorithm C, $\alpha = 1.8$, the updating function $f(\cdot) = \lfloor \sqrt{\cdot} \rfloor$, and $\ell_{\min} = 2$.

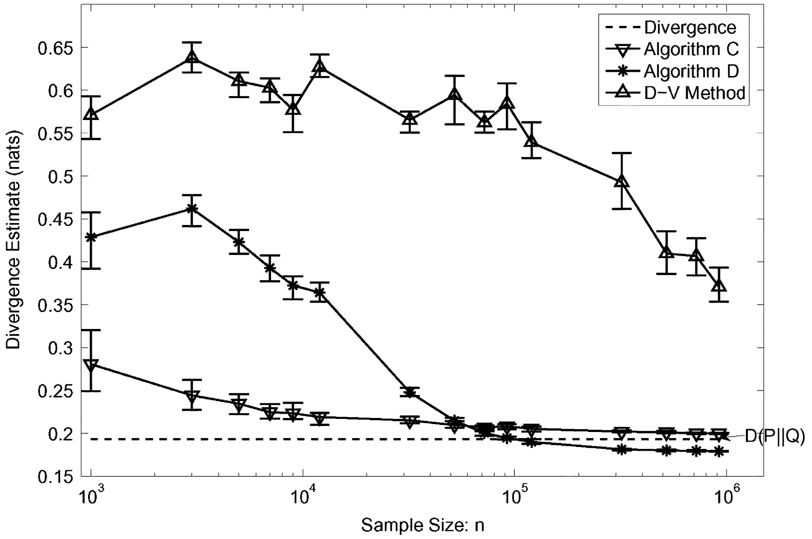

Fig. 8. $X \sim \mathrm{Exp}(1), Y \sim \mathrm{Exp}(2); D(P \,\|\, Q) = 0.1931$. For both Algorithm D and the DV method, $\epsilon = 0.001$. The terminating condition is tested on two sets of partitions, which consist of two and four subsegments, respectively.

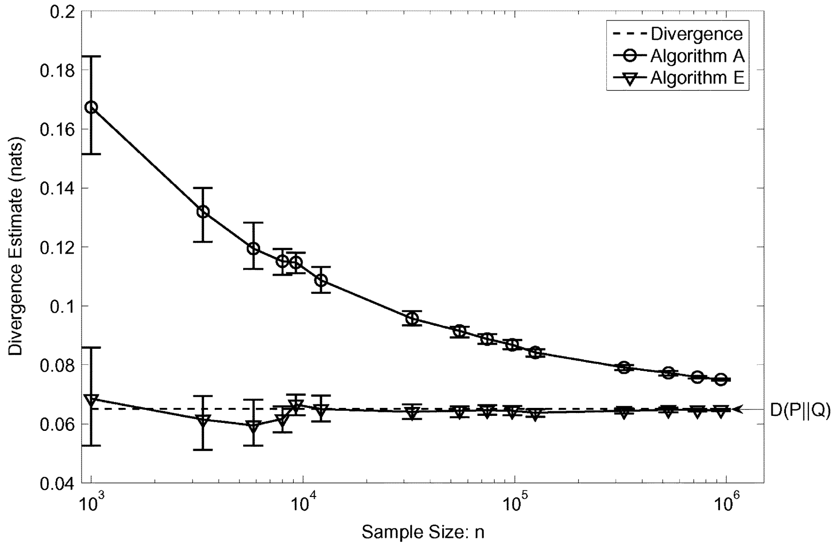

Fig. 9. $P \sim \mathrm{N}(0,1) \times \mathrm{N}(0,1), Q \sim \mathrm{N}(0.2,1) \times \mathrm{N}(0.3,1), D(P \,\|\, Q) = 0.065$.

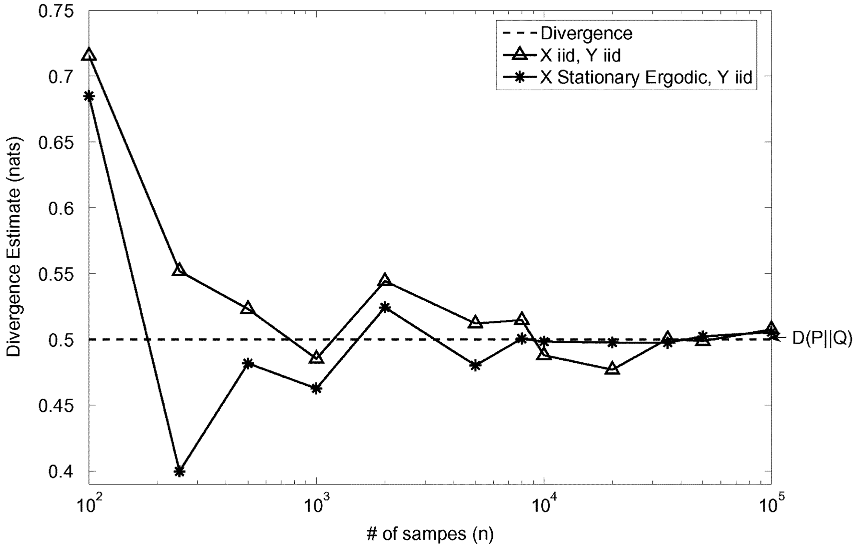

Fig. 10. $X_i \sim P = \mathrm{N}(1,1), Y_i \sim Q = \mathrm{N}(0,1); D(P \,\|\, Q) = 0.5$. The stationary process X is generated by autoregression as indicated in (56).

In Fig. 8, we observe that our data-driven partitioning scheme outperforms the Darbellay-Vajda (DV) method with much smaller bias for the same sample size.

The performance of the bias-corrected version is demonstrated in Fig. 9. We see that the subtraction of the approximated bias provides a much more reliable result.

Experiments on data with memory are shown in Fig. 10, where the stationary ergodic process $\{X_1, \dots, X_n\}$ is generated by

$$\begin{aligned} X_1 = Y_1, \quad X_i &= 0.6X_{i-1} + 0.8Y_i \quad (i \geq 2);\\ \text{then} \quad X_i &= X_i + 1, \quad i = 1, \dots, n \end{aligned} \tag{56}$$

with $Y_1, \dots, Y_n$ being i.i.d. and $Y_i \sim \mathrm{N}(0, 1)$. This figure demonstrates that our algorithm still provides good estimates when the data $\{X_1, \dots, X_n\}$ have memory.

In conclusion, our basic Algorithm A exhibits faster convergence than the direct plug-in methods. Algorithms B, C, and D, which are based on data-driven partitions, further improve the speed of convergence. In addition, numerical evidence indicates that our estimators outperform the DV adaptive partitioning scheme. By subtracting a bias approximation, Algorithm E provides accurate estimates in the regime of low divergence. We have also shown theoretically and experimentally that the algorithms are robust with respect to dependencies between the two data sources or to memory in the samples $\{X_1, \dots, X_n\}$.

References

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

[2] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," HP Laboratories, Cambridge, MA, Tech. Rep. HPL-2004-4, 2004.

[3] P. Hall, "On Kullback-Leibler loss and density estimation," *Ann. Statist.*, vol. 15, no. 4, pp. 1491–1519, Dec. 1987.

[4] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Statist. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[5] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal estimation of entropye and divergence via block sorting," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun./Jul. 2002, p. 433.

[6] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal divergenceestimation for finite-alphabet sources," *IEEE Trans. Inf. Theory*, submitted for publication.

[7] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1270–1279, Jul. 1993.

[8] A. Hero, B. Ma, and O. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," in *IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, Jun. 1999.

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

[10] G. Lugosi and A. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *Ann. Statist.*, vol. 24, no. 2, pp. 687–706, 1996.

[11] I. Gijbels and J. Mielniczuk, "Asymptotic properties of kernel estimators of the Radon-Nikodym derivative with applications to discriminant analysis," *Statistica Sinica*, vol. 5, pp. 261–278, 1995.

[12] S. Panzeri and A. Treves, "Analytical estimates of limited sampling biases in different information measures," *Network: Computation in Neural Systems*, vol. 7, pp. 87–107, 1996.

[13] E. Çinlar, *Lecture Notes in Probability Theory*. Unpublished.

[14] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[15] L. Devroye, *A Course in Density Estimation*. Boston, MA: Birkhäuser, 1987.

[16] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Its Applic.*, vol. 16, no. 2, pp. 264–280, 1971.

[17] L. Devroye, L. Györfi, and G. Lugosi, *A Probablistic Theory of Pattern Recognition*. Berlin, Germany: Springer-Verlag, 1996.

[18] T. M. Adams and A. B. Nobel, "On density estimation from ergodic processes," *Ann. Probab.*, vol. 26, no. 2, pp. 794–804, Apr. 1971.