

Universal Estimation of Divergence for Continuous Distributions via Data-Dependent Partitions

Qing Wang, Sanjeev R. Kulkarni, Sergio Verdú
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544 USA
Email: {qingwang, kulkarni, verdu}@princeton.edu

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 i.i.d. samples from both P and Q , is based on the estimation of the Radon-Nikodym derivative $\frac{dP}{dQ}$ 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. In the simulations, we compare our estimators with the 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.

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 (Ω, \mathcal{F}) . The divergence between P and Q is defined as

$$D(P\|Q) \equiv \int_{\Omega} dP \log \frac{dP}{dQ} \quad (1)$$

when P is absolutely continuous with respect to Q , and $+\infty$ otherwise.

Divergence is an important concept in information theory, since 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 multi-media classification [2] [3]. It is also applicable as a loss function in evaluating and optimizing the performance of density estimation methods [4]. However, there has been little work done on the universal estimation of divergence between unknown distributions. The estimation of divergence between the samples drawn from unknown distributions can be used to gauge the distance between those distributions. Thus divergence estimates can be applied in clustering and in particular for deciding whether the

samples come from the same distribution by comparing the estimate to a threshold. Another application is in the efficient estimation of parameters [5]. In addition, divergence estimates can be used to determine sample sizes required to achieve given performance levels in hypothesis testing.

In the discrete case, Cai *et al.* [7] proposed new algorithms to estimate the divergence between two finite alphabet, finite memory sources. In [7], 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 asymptotically piecewise i.i.d. Experiments show that these algorithms outperform estimators based on LZ string-matching [8].

For sources with a continuous alphabet, Hero *et al.* [9] provided an entropy estimation method using the minimal spanning tree which spans a set of feature vectors. This method was generalized to divergence estimation under the assumption that the reference distribution is already known. Darbellay and Vajda [10] 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 divergence, i.e.

$$D(P\|Q) \equiv \int_{\Omega} dQ \frac{dP}{dQ} \log \frac{dP}{dQ}. \quad (2)$$

In this formula, the Radon-Nikodym derivative $\frac{dP}{dQ}$ 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 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

¹This work was supported in part by ARL MURI under Grant number DAAD19-00-1-0466, Draper Laboratory under IR&D 6002 grant DL-H-546263, and the National Science Foundation under grant CCR-0312413.

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 summary, although particular attention is given to the case of scalar observations, we 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, the algorithm can be applied at the output of the transform. More discussions on data with memory are included in [17].

II. DIVERGENCE ESTIMATION BASED ON DATA-DEPENDENT PARTITIONS

A. Algorithm A: Non-Adaptive Estimator

We begin by stating our basic estimator in the one-dimensional case. Suppose P and Q are nonatomic probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, with $D(P\|Q) < \infty$. $\{X_1, X_2, \dots, X_n\}$ and $\{Y_1, Y_2, \dots, Y_m\}$ are i.i.d. samples generated from P and Q respectively. Denote the order statistics of Y by $\{Y_{(1)}, Y_{(2)}, \dots, Y_{(m)}\}$ where $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(m)}$. The real line is partitioned into T_m empirically equivalent segments according to

$$\{I_i^m\}_{i=1, \dots, T_m} = \{(-\infty, Y_{(\ell_m)}], (Y_{(\ell_m)}, Y_{(2\ell_m)}], \dots, (Y_{(\ell_m(T_m-1))}, +\infty)\} \quad (3)$$

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

In our basic algorithm, the divergence between P and Q is estimated as

$$\hat{D}_{n,m}(P\|Q) = \sum_{i=1}^{T_m-1} \frac{k_i}{n} \log \frac{k_i/n}{\ell_m/m} + \frac{k_{T_m}}{n} \log \frac{k_{T_m}/n}{\ell_m/m + \delta_m} \quad (4)$$

where $\delta_m = (m - \ell_m T_m)/m$ is a 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 (4) can be written as

$$\hat{D}_{n,m}(P\|Q) = \sum_{i=1}^{T_m} Q_m(I_i^m) \frac{P_n(I_i^m)}{Q_m(I_i^m)} \log \frac{P_n(I_i^m)}{Q_m(I_i^m)} \quad (5)$$

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 $\frac{dP}{dQ}$, which is guaranteed to exist provided $P \ll Q$.

Furthermore, this approach can be generalized to d -dimensional data, by partitioning with statistically equivalent blocks. According to the projections of the samples

Y_1, \dots, Y_m onto the first coordinate axis, the space can be partitioned into T_m statistically equivalent cylindrical sets, where $T_m = \lfloor (m/\ell_m)^{1/d} \rfloor$. Then we 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 (5) gives an estimate of the divergence for multivariate distributions.

B. Convergence Analysis

In this section, results are provided on the strong consistency of Algorithm A.

Theorem 1 *Let P and Q be nonatomic probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Assume that the divergence between P and Q is finite. Let $\{X_1, X_2, \dots, X_n\}$ and $\{Y_1, Y_2, \dots, Y_m\}$ be i.i.d. samples generated from P and Q respectively¹. Let ℓ_m, T_m be defined as in (3). If $\ell_m, T_m \rightarrow \infty$ as $m \rightarrow \infty$, then the divergence estimator in (5) satisfies*

$$\hat{D}_{m,n}(P\|Q) \rightarrow D(P\|Q) \text{ a.s., as } n, m \rightarrow \infty \quad (6)$$

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

Definition 1 *Let (Ω, \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 $m \rightarrow \infty$ if for any probability measure μ on (Ω, \mathcal{F}) that is absolutely continuous with respect to ν , we have*

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

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 integer T .

Lemma 1 *Let Q be a nonatomic 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: $\{I_i\}_{i=1,\dots,T} =$*

$$\{(-\infty, a_1], (a_1, a_2], \dots, (a_{T-2}, a_{T-1}], (a_{T-1}, +\infty)\},$$

such that $Q(I_i) = T^{-1}$, $i = 1, 2, \dots, T$. 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 $\{I_i^m\}_{i=1,\dots,T} =$

$$\{(-\infty, a_1^m], (a_1^m, a_2^m], \dots, (a_{T-2}^m, a_{T-1}^m], (a_{T-1}^m, +\infty)\}$$

¹Recall that we are not assuming independence between $\{X_1, X_2, \dots, X_n\}$ and $\{Y_1, Y_2, \dots, Y_m\}$

be a partition such that $Q_m(I_i^m) = T^{-1}$, $i = 1, 2, \dots, T$. 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 $m \rightarrow \infty$.

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 are replaced by their empirical versions. Lemma 2 shows that the corresponding empirical probability measures satisfy similar properties when the sample size goes to infinity.

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

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \mu_n(I_i^m) = \mu(I_i) \text{ a.s., } i = 1, 2, \dots, T \quad (7)$$

Proof Sketch of Theorem 1: Define $I_i = (a_{i-1}, a_i]$ ($i = 1, 2, \dots, T_m$), where $-\infty = a_0 < a_1 < \dots < a_{T_m-1} < a_{T_m} = +\infty$ such that $Q(I_i) = T_m^{-1}$, $i = 1, 2, \dots, T_m$.

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

The estimation error can be decomposed as

$$\begin{aligned} & |\hat{D}_{m,n}(P||Q) - D(P||Q)| \\ & \leq \left| \sum_{i=1}^{T_m} P_n(I_i^m) \log \frac{P_n(I_i^m)}{Q_m(I_i^m)} - \sum_{i=1}^{T_m} P(I_i) \log \frac{P(I_i)}{Q(I_i)} \right| \\ & \quad + \left| \sum_{i=1}^{T_m} Q(I_i) \frac{P(I_i)}{Q(I_i)} \log \frac{P(I_i)}{Q(I_i)} - \int_{\Omega} dQ \frac{dP}{dQ} \log \frac{dP}{dQ} \right| \\ & = e_1 + e_2 \end{aligned} \quad (8)$$

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 ℓ_m , m and n are sufficiently large, using Lemmas 1 and 2. \square

III. SCHEMES TO IMPROVE CONVERGENCE SPEED

In the previous section, we discussed the asymptotic consistency of our divergence estimator. However, in reality, we are only provided with samples of finite sizes. An important problem is how to obtain a reliable estimate when the number of samples is limited. In this section, we propose two approaches to improve convergence rate. Section III-A shows how to choose the algorithm parameters (such as 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 details between 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 along the real line according to the estimate \hat{D}_0 with $\ell_m = \ell_0$ (e.g. $\lfloor \sqrt{m} \rfloor$). If \hat{D}_0 is low (resp. high), ℓ_m is updated by $f(\ell_0) > \ell_0$ (resp. $f(\ell_0) < \ell_0$). The final estimate is then determined according to \hat{D}_0 and the estimate with $f(\ell_0)$. 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 true divergence is 0. A drawback of this method is that it is difficult to optimize the choice of the criteria and updating functions which specify when and how ℓ_m should be adapted. Also local details might be lost if the same ℓ_m is assigned to regions where the two distributions appear to be similar and to regions where they are quite mismatched.

2) Algorithm C: Locally Adaptive Method 1

Instead of insisting on uniform partitioning of the space, this locally adaptive scheme produces a fine partition in regions where $\frac{dP}{dQ}$ is high and a coarser partition elsewhere, the rationale being that not much accuracy is required in the zones where $\frac{dP}{dQ} \log \frac{dP}{dQ}$ is low.

Let ℓ_{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 into each segment in the new partition. $\alpha > 1$ 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 this segment i empirically equiprobably into ℓ_m sub-segments. Continue this process on each sub-segment until either $k_i \leq \alpha \ell_m$ or the updated $\ell_m \leq \ell_{min}$.

The new adaptive estimate is

$$\hat{D}_{n,m}^*(P||Q) = \sum_{i=1}^{T^*} \frac{k_i}{n} \log \frac{k_i/n}{\ell_i^*/m}, \quad (9)$$

where T^* is the total number of segments, and ℓ_i^* is the number of Y s in each segment.

3) Algorithm D: Locally Adaptive Method 2

Another version of non-uniform adaptive partition is inspired by Darbellay and Vajda's (D-V) method [10]. Their idea is to first segment the space into T_0 equivalent blocks. In [10], if locally conditional similarity is achieved, i.e.

$$\sum_{I_i \in s} \frac{P_n(I_i)}{P_n(I)} \log \frac{P_n(I_i)Q_n(I)}{P_n(I)Q_n(I_i)} < \epsilon, \text{ for all } s \in S(10)$$

no further segmentation will be imposed on I_i , where S represents all testing partitioning schemes. Considering some problematic situations associated with the above method, in Algorithm D, the terminating condition is now modified as

$$\sum_{I_i \in s} P_n(I_i) \log \frac{P_n(I_i)}{Q_n(I_i)} < \epsilon, \text{ for all } s \in S \quad (11)$$

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

B. Algorithm E: Reducing the Bias

Suppose we were to use a fixed partition $\{I_i\}_{i=1,\dots,T}$ that is not data-dependent. Then the estimate bias is

$$B_{n,m} = \left\langle \sum_{i=1}^T P_n(I_i) \log \frac{P_n(I_i)}{Q_m(I_i)} \right\rangle - \sum_{i=1}^T P(I_i) \log \frac{P(I_i)}{Q(I_i)} + \sum_{i=1}^T P(I_i) \log \frac{P(I_i)}{Q(I_i)} - \int_{\Omega} dP \log \frac{dP}{dQ} \quad (12)$$

where $\langle \cdot \rangle$ denotes the expectation with respect to the joint distribution of $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$. We only consider the first difference, since the second one depends on the two underlying distributions and it involves the knowledge of the true divergence, which is to be estimated. Let $f(x, y) = x \log \frac{x}{y}$. Expanding the first summation term in the above equation at $\{(P(I_i), Q(I_i))\}_{i=1,\dots,T}$ by the Taylor series of $f(x, y)$ and assuming that $Q(I_i) = 1/T$ for each $i = 1, \dots, T$ and that $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$ are independent, the bias estimate can be approximated by

$$\hat{B}_{n,m} = \frac{T_p - 1}{2n} + \frac{T - 1}{2m} + o\left(\frac{1}{m} + \frac{1}{n}\right) \quad (13)$$

We can improve the estimator by subtracting the first two terms on the right side of (13). Experimental results show that the approximation of the bias in (13) is excellent when the true divergence is not too high.

IV. EXPERIMENTS

In Figures 1 and 3, we compare the performance of our basic estimator on scalar data with that of the plug-in estimators, which are based on kernel density estimation and the histogram estimation method [11] respectively. The distributions associated with estimates in Figure 3 are shown in Figure 2.

Figure 4 presents experiments for two-dimensional data. The solid line represents the average of the estimates based on 25 sets of samples. Figures 5 and 6 demonstrate the advantage of adaptive versions of the divergence estimator over the basic estimator (Algorithm A) for similar distributions and mismatched distributions respectively. In Figure 7, we observe that our data-driven partitioning scheme outperforms the D-V method with smaller bias. The performance of the bias-corrected version is demonstrated in Figure 8.

In conclusion, 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 D-V adaptive partitioning scheme. By subtracting a bias approximation, Algorithm E provides accurate estimates in the regime of low divergence.

Experiments on correlated data and data with memory are included in [17].

REFERENCES

- [1] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, 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," HPL-2004-4, HP Laboratories, Cambridge, MA, USA, 2004.
- [3] C. Liu and H-Y Shum, "Kullback-Leibler Boosting," *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 18-20, pp. 1-587–1-594, June 2003.
- [4] P. Hall, "On Kullback-Leibler loss and density estimation," *The Annals of Statistics*, vol. 15, no. 4, pp. 1491–1519, Dec. 1987.
- [5] P. K. Bhattacharya, "Efficient estimation of a shift parameter from grouped data," *The Annals of mathematical Statistics*, vol. 38, no. 6, pp. 1770–1787, 1967.
- [6] J. Beirlant, E. J. Dudewicz, L. Györfi and E. C. van der Meulen, "Nonparametric entropy estimation: an overview," *International Journal of Mathematical and Statistical Sciences*, 6 (1) : pp. 17–39, 1997.
- [7] H. Cai, S. R. Kulkarni and S. Verdú, "Universal divergence estimation via block sorting," submitted to *IEEE Transactions on Information Theory*, see also 2002 ISIT pp. 433.
- [8] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1270–1279, July 1993.
- [9] 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, June 1999.
- [10] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions On Information Theory*, vol. 45, no. 4, pp. 1315–1321, May 1999.
- [11] G. Lugosi and A. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *The Annals of Statistics*, vol. 24, no. 2, pp. 687–706, 1996.
- [12] I. Gijbels and J. Mielniczuk, "Asymptotic properties of kernel estimators of the Radon-Nikodym derivative with applications to discriminant analysis," *Statistica Sinica*, pp. 261–278, 1995.
- [13] S. Panzeri and A. Treves, "Analytical estimates of limited sampling biases in different information measures," *Network: Computation in Neural Systems*, pp. 87–107, 1996.
- [14] E. Çinlar, *Lecture Notes in Probability Theory, Unpublished*.
- [15] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, Vol. 15, No. 6, pp 1191-1253, June 2003.
- [16] L. Devroye, *A course in density estimation*, Boston: Birkhäuser, 1987.
- [17] Q. Wang, S. R. Kulkarni and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," to appear in *IEEE Transactions on Information Theory*.

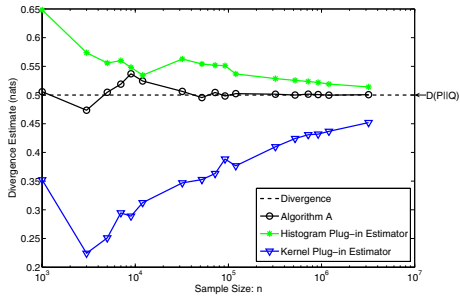


Fig. 1. $X \sim P = N(0, 1), Y \sim Q = N(1, 1); D(P||Q) = 0.5$

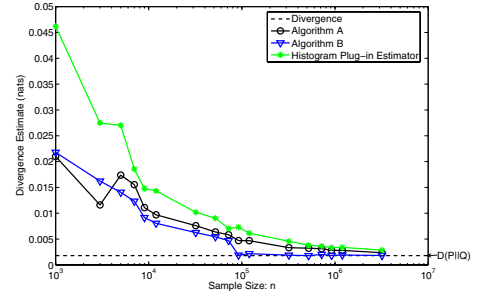


Fig. 5. $X \sim P = N(0, 1), Y \sim Q = N(0.06, 1); D(P||Q) = 0.0018$

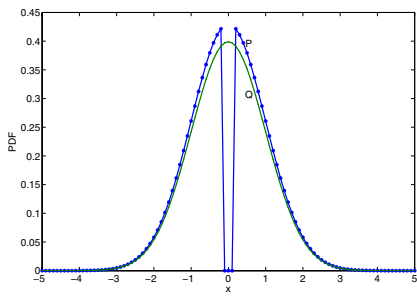


Fig. 2. $X \sim P = N(0, 1)$ with dip, $Y \sim Q = N(0, 1)$

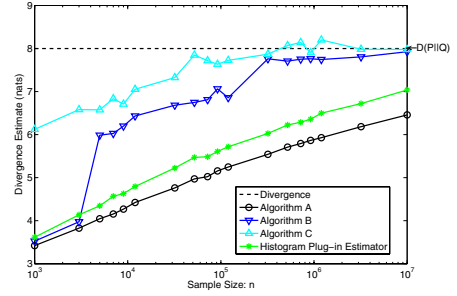


Fig. 6. $X \sim P = N(0, 1), Y \sim Q = N(4, 1); D(P||Q) = 8$

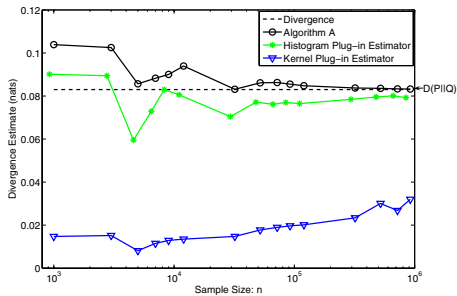


Fig. 3. $X \sim P = N(0, 1)$ with dip, $Y \sim Q = N(0, 1); D(P||Q) = 0.0830$

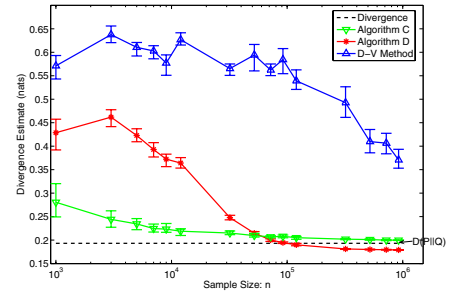


Fig. 7. $X \sim \text{Exp}(1), Y \sim \text{Exp}(2); D(P||Q) = 0.1931$

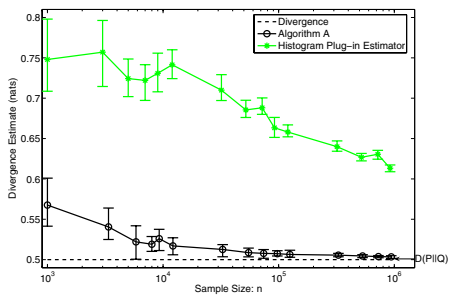


Fig. 4. $P \sim N(0, 1) \times N(0, 1), Q \sim N(1, 1) \times N(0, 1); D(P||Q) = 0.5$

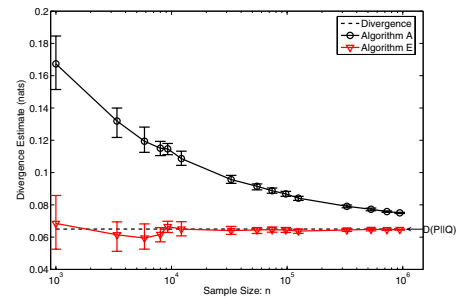


Fig. 8. $P \sim N(0, 1) \times N(0, 1), Q \sim N(0.2, 1) \times N(0.3, 1); D(P||Q) = 0.065$