Kullback-Leibler Divergence Estimation of Continuous Distributions

Fernando Pérez-Cruz
Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08544
Email: fp@princeton.edu

***Abstract*—We present a method for estimating the KL divergence between continuous densities and we prove it converges almost surely. Divergence estimation is typically solved estimating the densities first. Our main result shows this intermediate step is unnecessary and that the divergence can be either estimated using the empirical cdf or k-nearest-neighbour density estimation, which does not converge to the true measure for finite k. The convergence proof is based on describing the statistics of our estimator using waiting-times distributions, as the exponential or Erlang. We illustrate the proposed estimators and show how they compare to existing methods based on density estimation, and we also outline how our divergence estimators can be used for solving the two-sample problem.**

I. Introduction

The Kullback-Leibler divergence [11] measures the distance between two density distributions. This divergence is also known as information divergence and relative entropy. If the densities P and Q exist with respect to a Lebesgue measure, the Kullback-Leibler divergence is given by:

$$D(P||Q) = \int_{\mathbb{R}^d} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \geq 0. \tag{1}$$

This divergence is finite whenever P is absolutely continuous with respect to Q and it is only zero if $P = Q$.

The KL divergence is central to information theory and statistics. Mutual information measures the information one random variable contains about a related random variable and it can be computed as a special case of the KL divergence. From the mutual information we can define the entropy and differential entropy of a random variable as its self-information. KL divergence can be directly defined as the mean of the log-likelihood ratio and it is the exponent in large deviation theory. Also the two-sample problem can be naturally approach using this divergence, as its goal is to detect whether two set of samples have been drawn from the same distribution [1].

In machine learning and neuroscience the KL divergence also plays a leading role. In Bayesian machine learning, it is typically used to approximate an intractable density model. For example, expectation propagation [16] iteratively approximates an exponential family model to the desired density, minimising the inclusive KL divergence: $D(P||P_{app})$. While variational methods [15] minimize the exclusive KL divergence, $D(P_{app}||P)$, to fit the best approximation to P. Information-theoretic analysis of neural data is unavoidable given the questions neurophysiologists are interested in, see [19] for a detailed discussion on mutual information estimation in neuroscience. There are other applications in different research areas in which KL divergence estimation is used to measure the difference between two density functions. For example in [17] it is used for multimedia classification and in [8] for text classification.

In this paper, we focus on estimating the KL divergence for continuous random variables from independent and identically distributed (i.i.d.) samples. Specifically, we address the issue of estimating this divergence without estimating the densities, i.e. the density estimation used to compute the KL divergence does not converge to their measures as the number of samples tends to infinity. In a way, we follow Vapnik's advice [20] about not trying to solve an intermediate (harder) problem to estimate the quantity we are interested in. We propose a new method for estimating the KL divergence based on the empirical cumulative distribution function (cdf) and we show it converges almost surely to the actual divergence.

There are several approaches to estimate this divergence from samples for continuous random variables [21], [12], [5], [22], [13], [18], see also the references therein. Other methods concentrate on estimating the divergence for discrete random variables [19], [4], but we will not discuss them further as they lie outside the scope of this paper. Most of these approaches are based on estimating the densities first, hence ensuring the convergence of the estimator to the divergence as the number of samples tends to infinity. For example in [21] the authors propose to estimate the densities based on an data-dependent histograms with a fixed number of samples from $q(\mathbf{x})$ in each bin and in [5] the authors compute relative frequencies on data-driven partitions achieving local independence for estimating mutual information. In [12] local likelihood density estimation is used to estimate the divergence between a parametric model and the available data. In [18] the authors compute the divergence between $p(\mathbf{x})$ and $q(\mathbf{x})$ using a variational approach, in which convergence is proven ensuring that the estimate for $p(\mathbf{x})/q(\mathbf{x})$ converges to the true measure ratio. Finally, we only know of two previous approach based on k-nearest-neighbours density estimation [22], [13], in which the authors prove mean-square consistency of the divergence estimator for finite k, although this density estimate does not

converge to its measure. In [3] a good survey paper analyzes the different proposals for entropy estimation.

The rest of the paper is organized as follows. We show the proposed method for one dimensional data in Section 2 together with its proof of convergence. In Section 3, we extend our proposal to multidimensional problems. We also discuss how to extend this approach for kernels in Section 3.1, which is of relevance for solving the two-sample problem with no real-valued data, such as graphs or sequences. In Section 4, we compute the KL divergence for known and unknown density models and indicate how it can be used for solving the two-sample problem. We conclude the paper in Section 5 with some final remarks and proposed further work.

II. Divergence estimation for 1D data

We are given n i.i.d. samples from $p(x)$, $\mathcal{X} = \{x_i\}_{i=1}^n$, and m i.i.d. samples from $q(x)$, $\mathcal{X}' = \{x'_j\}_{j=1}^m$, without loss of generality we assume the samples in these sets are sorted in increasing order. Let $P(x)$ and $Q(x)$, respectively, denote the absolutely continuous cdfs of $p(x)$ and $q(x)$. The empirical cdf is given by:

$$P_e(x) = \frac{1}{n}\sum_{i=1}^{n} U(x - x_i) \tag{2}$$

where $U(x)$ is the unit-step function with $U(0) = 0.5$. We also define a continuous piece-wise linear extension to $P_e(x)$:

$$P_c(x) = \begin{cases} 0, & x < x_0 \\ a_i x + b_i, & x_{i-1} \le x < x_i \\ 1, & x_{n+1} \le x \end{cases} \tag{3}$$

where a_i and b_i are defined to ensure that $P_c(x)$ takes the same value as $P_e(x)$ at the sampled values and leads to a continuous approximation. $x_0 < \inf\{\mathcal{X}\}$ and $x_{n+1} > \sup\{\mathcal{X}\}$, their exact values are inconsequential for our estimate. Both of these empirical cdfs converges uniformly and independent of the distribution to their cdfs [20].

The proposed divergence estimator is given by:

$$\widehat{D}(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\delta P_c(x_i)}{\delta Q_c(x_i)} \tag{4}$$

$\delta P_c(x_i) = P_c(x_i) - P_c(x_i - \epsilon)$ for any $\epsilon < \min_i\{x_i - x_{i-1}\}$.

Theorem 1. *Let P and Q be absolutely continuous probability measures and assume its KL divergence is finite. Let $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\mathcal{X}' = \{x'_i\}_{i=1}^m$ be i.i.d. samples sorted in increasing order, respectively, from P and Q, then*

$$\widehat{D}(P||Q) - 1 \xrightarrow{a.s.} D(P||Q) \tag{5}$$

Proof: We can rearrange (4) as follows:

$$\widehat{D}(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P_c(x_i)/\Delta x_i}{\Delta Q_c(x'_{mi})/\Delta x'_{mi}} \tag{6}$$

where $\Delta P_c(x_i) = P_c(x_i) - P_c(x_{i-1})$, $\Delta x_i = x_i - x_{i-1}$, $\Delta x'_{mi} = \min\{x'_j | x'_j \ge x_i\} - \max\{x'_j | x'_j < x_i\}$ and $\Delta Q_c(x'_{mi}) = Q(\min\{x'_j | x'_j \ge x_i\}) - Q(\max\{x'_j | x'_j < x_i\})$. The equality holds because $P_c(x)$ and $Q_c(x)$ are piecewise linear approximations to their cdfs.

Let us rearrange (6) as follows:

$$\widehat{D}(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P(x_i)/\Delta x_i}{\Delta Q(x'_{mi})/\Delta x'_{mi}} - \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P(x_i)}{\Delta P_c(x_i)} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta Q(x'_{mi})}{\Delta Q_c(x'_{mi})} = \widehat{D}_e(P||Q) - C_1(P) + C_2(P,Q) \tag{7}$$

The first term in (7):

$$\widehat{D}_e(P||Q) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P(x_i)/\Delta x_i}{\Delta Q(x'_{mi})/\Delta x'_{mi}} \xrightarrow{a.s.} D(p||q), \tag{8}$$

because $\lim_{n\to\infty} \Delta P(x_i)/\Delta x_i = p(x_i)$ and $\lim_{n\to\infty} \Delta Q(x'_{mi})/\Delta x'_{mi} = q(x_i)$, due to $p(x)$ is absolutely continuous with respect to $q(x)$.

The second term in (7):

$$C_1(P) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\Delta P(x_i)}{\Delta P_c(x_i)} = \frac{1}{n}\sum_{i=1}^{n} \log n\Delta P(x_i) \tag{9}$$

As x_i is distributed according to $p(x)$, $P(x_i)$ is distributed according to a uniform random variable between 0 and 1. $z_i = n\Delta P(x_i)$ is the difference (waiting time) between two consecutive samples from a uniform distribution between 0 and n with one arrival per unit-time, therefore it is distributed like an unit-mean exponential random variable. Consequently

$$C_1(P) = \frac{1}{n}\sum_{i=1}^{n} \log z_i \xrightarrow{a.s.} \int_0^\infty \log z e^{-z} dz = -0.5772, \tag{10}$$

which is the negated Euler-Mascheroni constant.

The third term in (7):

$$C_2(P,Q) = \frac{1}{n}\sum_{j=1}^{m} n\Delta P_e(x'_j) \log \frac{\Delta Q(x'_j)}{\Delta Q_c(x'_j)} = \frac{1}{m}\sum_{j=1}^{m} \frac{\frac{\Delta P_e(x'_j)}{\Delta x'_j}}{\frac{\Delta Q(x'_j)}{\Delta x'_j}} m\Delta Q(x'_j) \log m\Delta Q(x'_j) \tag{11}$$

where $n\Delta P_e(x'_j)$ counts the number of samples from the set $\mathcal{X}$ between two consecutive samples from $\mathcal{X}'$. As before, $m\Delta Q(x'_j)$ is distributed like unit-mean exponential, independent from $q(x)$, and $\Delta Q(x'_j)/\Delta x'_j$ and $\Delta P_e(x'_j)/\Delta x'_j$ tend, respectively, to $q(x)$ and to $p_e(x)$, hence

$$C_2(P,Q) \xrightarrow{a.s.} \int \frac{p_e(x)}{q(x)} z \log z e^{-z} q(x) dz dx = \int_0^\infty z \log z e^{-z} dz \int_{\mathbb{R}} p_e(x) dx = 0.4228, \tag{12}$$

$p_e(x)$ is a density model, but it does not need to tend to $p(x)$ for $C_2(P,Q)$ to converge to 0.4228, i.e 1 minus the Euler-Mascheroni constant.

The three terms in (7), respectively, converge almost surely to $D(P||Q)$, -0.5772 and 0.4228, due to the strong law of large numbers, and hence so it does their sum [7]. ∎

From the last equality in (6), we can understand that we are using a data-dependent histogram, in which we put one sample in each bin, as density estimate for $p(x)$ and $q(x)$, e.g. $\widehat{p}(x_i) = 1/n\Delta x_i$, to estimate the KL divergence. In [14], the authors show that data-dependent histograms converge to their true measures when two conditions are met. The first condition states that the number of bins must grow sublinearly with the number of samples and this condition is violated by our density estimate. Hence our KL divergence estimator converges almost surely, but it is based on non-convergent density estimates.

III. Divergence estimator for vectorial data

The procedure to estimate the divergence from samples proposed in the previous section is based on the empirical cdf and it is not straightforward how it can be extended to vectorial data. But, taking a closer look at the last part of equation (6), we can reinterpret our estimator as follows, first compute nearest-neighbour estimates for $p(x)$ and $q(x)$ and then use these estimates to calculate the divergence:

$$\widehat{D}(P||Q) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\widehat{p}(x_i)}{\widehat{q}(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{m\Delta x'_{mi}}{n\Delta x_i} \quad (13)$$

where we employ the nearest-neighbour less than x_i from $\mathcal{X}$ to estimate $p(x_i)$ and the two nearest-neighbours, one less than and the other larger than x_i, from $\mathcal{X}'$ to estimate $q(x_i)$. We showed $\widehat{D}(P||Q) - 1$ converges to the KL divergence, even though $\widehat{p}(x_i)$ and $\widehat{q}(x_i)$ do not converge to their true measures, and nearest-neighbour can be readily used for multidimensional data.

The idea to use k-nearest-neighbour density estimation as an intermediate step to estimate the KL divergence was put forward in [22], [13] and follows a similar idea proposed to estimate differential entropy [9] and that has been used to estimate mutual information in [10]. In [22], [13], the authors prove mean-square consistency of their estimator for finite k, which is based on some regularity conditions imposed over the densities $p(\mathbf{x})$ and $q(\mathbf{x})$, as for finite k, nearest-neighbour density estimation does not converge to their measure. From our point of view their proof is rather technical. In this paper, we prove the almost sure convergence of this KL divergence estimator, using waiting times distributions without needing to impose additional conditions over the density models. Given a set with n i.i.d. samples from $p(\mathbf{x})$ and m i.i.d. samples from $q(\mathbf{x})$, we can estimate the $D(P||Q)$ from a k-nearest-neighbour density estimate as follows:

$$\widehat{D}_k(P||Q) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\widehat{p}_k(\mathbf{x}_i)}{\widehat{q}_k(\mathbf{x}_i)} = \frac{d}{n}\sum_{i=1}^{n}\log\frac{r_k(\mathbf{x}_i)}{s_k(\mathbf{x}_i)} + \log\frac{m}{n-1} \quad (14)$$

where

$$\widehat{p}_k(\mathbf{x}_i) = \frac{k}{(n-1)}\frac{\Gamma(d/2+1)}{\pi^{d/2}r_k(\mathbf{x}_i)^d} \quad (15)$$

$$\widehat{q}_k(\mathbf{x}_i) = \frac{k}{m}\frac{\Gamma(d/2+1)}{\pi^{d/2}s_k(\mathbf{x}_i)^d} \quad (16)$$

and $r_k(\mathbf{x}_i)$ and $s_k(\mathbf{x}_i)$ are, respectively, the Euclidean distances to the k^{th} nearest-neighbour of $\mathbf{x}_i$ in $\mathcal{X}\backslash\mathbf{x}_i$ and $\mathcal{X}'$, and $\pi^{d/2}/\Gamma(d/2+1)$ is the volume of the unit-ball in $\mathbb{R}^d$. Before proving (14) converges almost surely to $D(P||Q)$, let us show an intermediate necessary result.

Lemma 1. *Given n i.i.d. samples, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, from an absolutely continuous probability distribution P, the random variable $p(\mathbf{x})/\widehat{p}_1(\mathbf{x})$ converges in probability to a unit-mean exponential distribution for any $\mathbf{x}$ in the support of $p(\mathbf{x})$.*

Proof: Let's initially assume $p(\mathbf{x})$ is a d-dimensional uniform distribution of a given support. The set $\mathcal{S}_{\mathbf{x},R} = \{\mathbf{x}_i|\ \|\mathbf{x}_i - \mathbf{x}\|_2 \leq R\}$ contains all the samples from $\mathcal{X}$ inside the ball centred in $\mathbf{x}$ of radius R. Therefore, the samples in $\{\|\mathbf{x}_i - \mathbf{x}\|_2^d|\ \mathbf{x}_i \in \mathcal{S}_{\mathbf{x},R}\}$ are uniformly distributed between 0 and R^d, if the ball lies inside the support of $p(\mathbf{x})$. Hence the random variable $r_1(\mathbf{x})^d = \min_{\mathbf{x}_j\in\mathcal{S}_{\mathbf{x},R}}(\|\mathbf{x}_j - \mathbf{x}\|_2^d)$ is an exponential random variable, as it measures the waiting time between the origin and the first event of a uniformly spaced distribution [2]. As $p(\mathbf{x})n\pi^{d/2}/\Gamma(d/2+1)$ is the mean number of samples per unit ball centred in $\mathbf{x}$, $p(\mathbf{x})/\widehat{p}_1(\mathbf{x})$ is distributed as an exponential distribution with unit mean. This holds for all n, it is not an asymptotic result.

For non-uniform absolutely-continuous $p(\mathbf{x})$, $\mathbb{P}(r_1(\mathbf{x}) > \varepsilon) \to 0$ as $n \to \infty$ for any $\mathbf{x}$ in the support of $p(\mathbf{x})$ and every $\varepsilon > 0$. Therefore, as n tends to infinity we can consider $\mathbf{x}$ and its nearest-neighbour in $\mathcal{X}$ to come from a uniform distribution and hence $p(\mathbf{x})/\widehat{p}_1(\mathbf{x})$ converges in probability to a unit-mean exponential distribution. ■

Corolary 1. *Given n i.i.d. samples, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, from an absolutely continuous probability distribution $p(\mathbf{x})$, the random variable $p(\mathbf{x})/\widehat{p}_k(\mathbf{x})$ converges in probability to a unit-mean and $1/k$-variance gamma distribution for any $\mathbf{x}$ in the support of $p(\mathbf{x})$.*

Proof: In the previous proof, instead of measuring the waiting time to the first event, we measure the waiting time to the k^{th} event of a uniformly spaced distribution. This waiting time is distributed as an Erlang distribution or a unit-mean and $1/k$-variance gamma distribution. ■

Now we can easily proof the almost surely convergence of the KL divergence based on the k-nearest-neighbour density estimation.

Theorem 2. *Let P and Q be absolutely continuous probability measures and assume its KL divergence is finite. Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{X}' = \{\mathbf{x}'_i\}_{i=1}^m$ be i.i.d. samples, respectively, from P and Q, then*

$$\widehat{D}_k(P||Q) \xrightarrow{a.s.} D(P||Q) \quad (17)$$

Proof: We can rearrange $\widehat{D}_k(P||Q)$ in (14) as follows:

$$\widehat{D}_k(P||Q) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)} - \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(\mathbf{x}_i)}{\widehat{p}_k(\mathbf{x}_i)} + \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(\mathbf{x}_i)}{\widehat{q}_k(\mathbf{x}_i)} \quad (18)$$

The first term converges almost surely to the KL divergence between P and Q and the second and third terms converges almost surely to $\int_0^\infty z^{k-1}\log z e^{-z}dz/(k-1)!$, because the sum of random variables that converge in probability converges almost surely [7]. Finally, the sum of almost surely convergent terms also converges almost surely [7]. ■

In the proof of Theorem 2, we can see that the first element of (18) is an unbiased estimator of the divergence and, as the other two terms cancel each other out, one could think this estimator is unbiased. But this is not the case, because the convergence rates for the second and third terms to their means are not equal. For example for $k=1$, $p(\mathbf{x}_i)/\widehat{p}_1(\mathbf{x}_i)$ converges much faster to an exponential distribution than $q(\mathbf{x}_i)/\widehat{q}_1(\mathbf{x}_i)$ does, because the $\mathbf{x}_i$ samples comes from $p(\mathbf{x})$. The samples from $p(\mathbf{x})$ in the low probability region of $q(\mathbf{x})$ needs many samples from $q(\mathbf{x})$ to guarantee that its nearest-neighbour is close enough to assume that $q(\mathbf{x}_i)/\widehat{q}_1(\mathbf{x}_i)$ is distributed like an exponential. Hence, this estimator is biased and this bias depends on the distributions.

If the divergence is zero, the estimator is unbiased as the distributions $p(\mathbf{x}_i)/\widehat{p}_k(\mathbf{x}_i)$ and $q(\mathbf{x}_i)/\widehat{q}_k(\mathbf{x}_i)$ are identical. For the two-sample problem this is a very interesting result as it allows to measure the variance of our estimator for $P=Q$ and set the threshold for rejecting the null-hypothesis according to a fixed probability of type I errors (false positives).

A. Estimating KL Divergence with kernels

The KL divergence estimator in (14) can be computed using kernels. This extension allows measuring the divergence for non-real valued data, such as graphs or sequences, which could not be measured otherwise. There is only one previous proposal for solving the two-sample problem using kernels [6], which allows to compare if 2 non-real valued sets belong to the same distribution.

To compute (14) with kernels we need to measure the distance to the k^{th} nearest-neighbour to $\mathbf{x}_i$. Let's assume $\mathbf{x}_{nn_i^k}$ and $\mathbf{x}'_{nn_i^k}$ are, respectively, the k^{th} nearest-neighbour to $\mathbf{x}_i$ in $\mathcal{X}\backslash\mathbf{x}_i$ and $\mathcal{X}'$, then

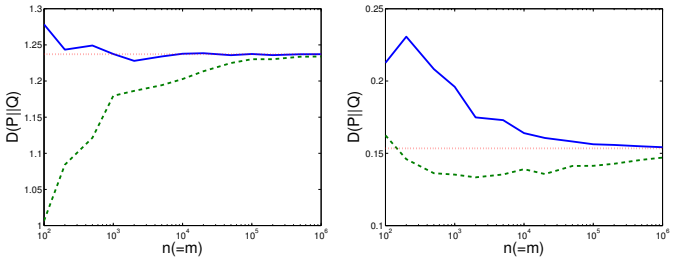
$$r_k(\mathbf{x}_i) = \sqrt{k(\mathbf{x}_i,\mathbf{x}_i) + k(\mathbf{x}_{nn_i^k},\mathbf{x}_{nn_i^k}) - 2k(\mathbf{x}_i,\mathbf{x}_{nn_i^k})} \quad (19)$$

$$s_k(\mathbf{x}_i) = \sqrt{k(\mathbf{x}_i,\mathbf{x}_i) + k(\mathbf{x}'_{nn_i^k},\mathbf{x}'_{nn_i^k}) - 2k(\mathbf{x}_i,\mathbf{x}'_{nn_i^k})} \quad (20)$$

Finally, to measure the divergence we need to set the dimension d of our feature space. For finite VC dimension kernels, as polynomial kernels, d is the VC dimension of our kernel. While for infinite VC dimension kernels we set $d=n+m-1$, as our data cannot live in a space larger than that.

IV. Experiments

We have conducted 3 series of experiments to show the performance of the proposed divergence estimators. First, we have estimated the divergence using (4), comparing a unit-mean exponential and a $\mathcal{N}(3,4)$ and two zero-mean Gaussians with variances 2 and 1, which are shown as solid lines in Figure 1. For comparison purposes, we have also plotted the divergence estimator proposed in [21] as Algorithm A with $\sqrt{m}$ as the number of bins for the density estimation, which was shown in [21] to be more accurate than the divergence estimator in [5] and one based on Partzen windows density estimation. Each curve is the mean value of 100 independent trials. We have not depicted the variance for each estimator for clarity purposes, although both are similar and tend towards zero as $1/n$. In these figure we can see that the proposed estimator is more accurate than the one in [21], as it converges faster to the true divergence as the number of samples increases.



(a) (b)

Fig. 1. Divergence estimation for a unit-mean exponential and $\mathcal{N}(3,4)$ in (a) and $\mathcal{N}(0,2)$ and $\mathcal{N}(0,1)$ in (b). The solid lines represent the estimator in (4) and the dashed lines the estimator in [21]. The dotted lines show the KL divergences.

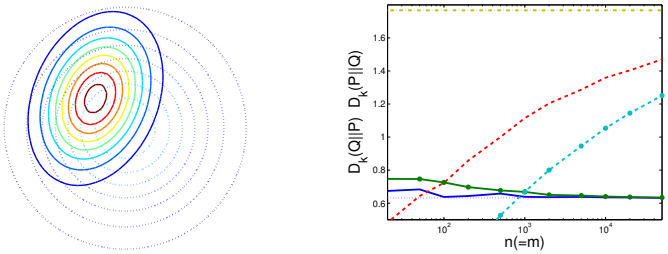
In the second experiment we measure the divergence between two 2-dimensional densities:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{0},\mathbf{I}) \quad (21)$$

$$q(\mathbf{x}) = \mathcal{N}\left(\begin{bmatrix}0.5\\-0.5\end{bmatrix}, \begin{bmatrix}0.5 & 0.1\\0.1 & 0.3\end{bmatrix}\right) \quad (22)$$

In Figure 2(a) we have contour-plotted these densities and in Figure 2(b) we have depicted $\widehat{D}_k(P||Q)$ and $\widehat{D}_k(Q||P)$ mean values for 100 independent experiments with $k=1$ and $k=10$. We can see that $\widehat{D}_k(Q||P)$ converges much faster to its divergence than $\widehat{D}_k(P||Q)$ does. This can be readily understood by looking at the density distributions in Figure 2(a). When we compute $s_k(\mathbf{x}'_i)$ for $\widehat{D}_k(Q||P)$, there is always a sample from $p(\mathbf{x})$ close by for every sample of $q(\mathbf{x})$. Hence both $p(\mathbf{x}'_i)/\widehat{p}_k(\mathbf{x}'_i)$ and $q(\mathbf{x}'_i)/\widehat{q}_k(\mathbf{x}'_i)$ converge quickly to a k-mean k-variance gamma distribution. But the converse is not true, as there is a high density region for $p(\mathbf{x})$ which is not well covered by $q(\mathbf{x})$. So $q(\mathbf{x}_i)/\widehat{q}_k(\mathbf{x}_i)$ needs many more samples to converge than $p(\mathbf{x}_i)/\widehat{p}_k(\mathbf{x}_i)$ does, which explains the strong bias in this estimator. As we increase k, we notice the divergence estimate takes longer to converge in both cases, because the tenth nearest-neighbour is further than the first and so the convergence of $p(\mathbf{x}_i)/\widehat{p}_k(\mathbf{x}_i)$ and $q(\mathbf{x}_i)/\widehat{q}_k(\mathbf{x}_i)$ to their distributions needs more samples.

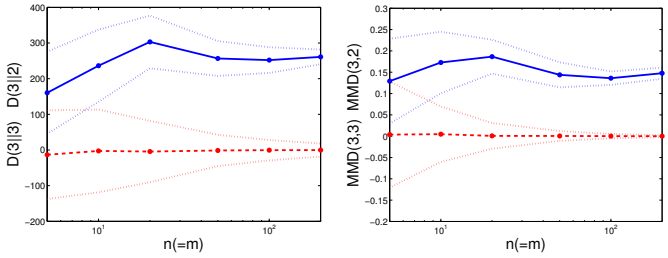
Finally, we have estimated the divergence between the three's and two's in the MNIST dataset (http://yann.lecun.com/exdb/mnist/) in a 784 dimensional space. In Figure 3a we have plotted the divergence estimator for $\widehat{D}_1(3,2)$ (solid line) and $\widehat{D}_1(3,3)$ (dashed line) mean values for 100 experiments together with their two standard deviations confidence intervals. As expected $\widehat{D}_1(3,3)$ is



(a) (b)

Fig. 2. In (a) we have plotted the contour lines of $p(\mathbf{x})$ (dashed) and of $q(\mathbf{x})$ (solid). In (b) we have plotted $\widehat{D}_k(Q||P)$ (solid) and $\widehat{D}_k(P||Q)$ (dashed), the curves with bullets represent the results for $k = 10$. The dotted and dash-dotted lines represent, respectively, $D(Q||P)$ and $D(P||Q)$.

unbiased, so it is close to zero for any sample size. $\widehat{D}_1(3,2)$ seems to level off around $260nats$, but we do not believe this is the true divergence between the three's and two's, as we need to resample from a population of around 7000 samples[1] for each digit in each experiment. But we can see that for as little as 20 samples we can clearly distinguish between these two populations. For comparison purposes we have plotted the MMD test from [6], in which a kernel method was proposed for solving the two-sample problem. We have used the code available in http://www.kyb.mpg.de/bs/people/arthur/mmd.htm and have used its bootstrap estimate for our comparisons. Although a more thorough examination is required, it seems that our divergence estimator would be perform similar to [6] for the two-sample problem without needing to chose a kernel and its hyperparameters.



(a) (b)

Fig. 3. In (a) we have plotted $\widehat{D}_1(3||2)$ (solid), $\widehat{D}_1(3||3)$ (dashed) and their ± 2 standard deviations confidence intervals (dotted). In (b) we have repeated the same plots using the MMD test from [6].

V. Conclusions and Further Work

We have proposed a divergence estimator based on the empirical cdf, which does not need to estimate the densities as an intermediate step, and we have proven its almost sure convergence to the true divergence. The extension for vectorial data coincides with a divergence estimator based on k-nearest-neighbour density estimation, which has been already proposed in [22], [13]. In this paper we prove its almost sure convergence and we do not need to impose additional conditions over the densities to ensure convergence, as need in [22], [13] to prove in mean-square convergence. We illustrated in the experimental section that the proposed estimators are more accurate than the estimators based on convergent density estimation. Finally we have also suggested this divergence estimator can be used for solving the two-sample problem, although a thorough examination of its merit as such has been left as further work.

[1]We have used all the MNIST data (training and test) for our experiments.

References

[1] N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 7 1994.

[2] K. Balakrishnan and A. P. Basu. *The Exponential Distribution: Theory, Methods and Applications*. Gordon and Breach Publishers, Amsterdam, Netherlands, 1996.

[3] J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Meulen. Nonparametric entropy estimation: An overview. *nternational Journal of the Mathematical Statistics Sciences*, pages 17–39, 1997.

[4] H. Cai, S. Kulkarni, and S. Verdú. Universal divergence estimation for finite-alphabet sources. *IEEE Trans. Information Theory*, 52(8):3456–3475, 8 2006.

[5] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Information Theory*, 45(4):1315–1321, 5 1999.

[6] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.

[7] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, UK, 3 edition, 2001.

[8] S. Mallela I. S. Dhillon and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 3 2003.

[9] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problems Inform. Transmission*, 23(2):95–101, 4 1987.

[10] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):1–16, 6 2004.

[11] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22(1):79–86, 3 1951.

[12] Y. K. Lee and B. U. Park. Estimation of kullback-leibler divergence by local likelihood. *Annals of the Institute of Statistical Mathematics*, 58(2):327–340, 6 2006.

[13] N. N. Leonenko, L. Pronzato, and V. Savani. A class of renyi information estimators for multidimensional densities. *Annals of Statistics*, 2007. Submitted.

[14] G. Lugosi and A. Nobel. Consistency of data-driven histogram methods for density estimation and classification. *Annals Statistics*, 24(2):687–706, 4 1996.

[15] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

[16] T Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[17] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. Technical Report HPL-2004-4, HP Laboratories, Cambridge, MA, 2004.

[18] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *IEEE Int. Symp. Information Theory*, Nice, France, 6 2007.

[19] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 6 2003.

[20] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[21] Q. Wang, S. Kulkarni, and S. Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Information Theory*, 51(9):3064–3074, 9 2005.

[22] Q. Wang, S. Kulkarni, and S. Verdú. A nearest-neighbor approach to estimating divergence between continuous random vectors. In *IEEE Int. Symp. Information Theory*, Seattle, USA, 7 2006.