ISIT 2009, Seoul, Korea, June 28 - July 3, 2009

Histogram-Based Estimation for the Divergence Revisited

Jorge Silva
Department of Electrical Engineering
University of Chile
josilva@ing.uchile.cl

Shrikanth S. Narayanan
Department of Electrical Engineering
University of Southern California
shri@sipi.usc.edu

Abstract— **This work revisits and extends the problem of consistent divergence estimation using data-dependent partitions. For distributions defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, the main result characterizes sufficient conditions on a data-dependent partition scheme to get a strongly consistent histogram-based estimate of the divergence.**

I. Introduction

The problem of divergence estimation in a finite dimensional Euclidean spaces is conceptually important and with implications in many statistical decision scenarios. Of particular interest is to have distribution-free estimates in a wide class of probability measures, which converge to desired theoretical values in some sense, as the number of samples points tends to infinity. The problem has been recently addressed using some classical non-parametric techniques (histogram-based and kernel-based density estimates) [1], [2], [3], where consistency was the main consideration. In the context of histogram-based constructions, which is the focus of this work, Wang *et al.* [1] proposed a histogram-based divergence estimation for probability measures defined on the real line and absolutely continuous with respect to the Lebesgue measure. The work was the first to consider an adaptive partition scheme that approximates empirical statistical equivalent intervals relative to the reference measure as a way to estimate the Radon-Nicodym (RD) derivative of the probability measures involved. Sufficient conditions on the statistically equivalent partitions were stipulated to guarantee the strong consistency. Silva *et al.* [2] took this direction one step further addressing the problem of consistency for general families of data-dependent partitions in $\mathbb{R}^d$. The main constraint of these two works is that they are valid when the limits for the sample points of the two involved distributions are taken independently to infinity, i.e., one after the other in a specific order, which limits its applicability.

In this paper we revisit our approach in [2] and extend this direction to address the more realistic scenario when the samples of both distributions jointly tend to infinity. In this new learning setting, we propose a new histogram-based learning scheme, adopting a version of *Barron* type histogram-based density estimate [10] for the RD derivative, and show a set of sufficient conditions where consistency is guaranteed.

The rest of the paper is organized as follows: Section II introduces some results and notations used for the rest of the paper. Section III states the problem and Section IV presents the main result. Finally, Section V shows an application of this result.

II. Preliminaries

A. Divergence

Let P and Q be probability measures defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ absolutely continuous with respect to the Lebesgue measure λ. The relative entropy [4], divergence [5] or Kullback-Leibler divergence [6] is given by

$$D(P||Q) = \int \log \frac{\partial P}{\partial Q}(x) \cdot \partial P(x), \tag{1}$$

where this expression is under the assumption that $D(P||Q) < \infty$ and consequently $P \ll Q$ [5], which makes the Radon-Nicodym (RD) derivative of P with respect to Q to be well defined in (1).

B. Data-Dependent Partition

We say $\pi = \{A_1, .., A_r\}$ is a finite measurable partition of $\mathcal{B}(\mathbb{R}^d)$ if: for any i, $A_i \in \mathcal{B}(\mathbb{R}^d)$; $A_i \cap A_j = \varnothing$, $i \neq j$; and $\bigcup_{i=1}^r A_i = \mathbb{R}^d$. We denote $|\pi|$ as the number of cells in π. Let $\mathcal{A}$ be a collection of measurable partitions for $\mathbb{R}^d$. The *maximum cell counts* of $\mathcal{A}$ is given by

$$\mathcal{M}(\mathcal{A}) = \sup_{\pi \in \mathcal{A}} |\pi|. \tag{2}$$

In addition, a notion of combinatorial complexity for $\mathcal{A}$ can be introduced, following Lugosi and Nobel [7]. Let us consider a finite length sequence $x_1^n = (x_1, .., x_n) \in \mathbb{R}^{d \cdot n}$, and the induced set by $\{x_1, .., x_n\}$, then we can define $\Delta(\mathcal{A}, x_1, .., x_n) = |\{\{x_1, .., x_n\} \cap \pi : \pi \in \mathcal{A}\}|$, with $\{x_1, .., x_n\} \cap \pi$ a short hand for $\{\{x_1, .., x_n\} \cap A : A \in \pi\}$. Consequently, $\Delta(\mathcal{A}, x_1, .., x_n)$ is the number of possible partitions of $\{x_1, .., x_n\}$ induced by $\mathcal{A}$, and then the *growth function* of $\mathcal{A}$ is defined by [7]

$$\Delta_n^*(\mathcal{A}) = \max_{x_1^n \in \mathbb{R}^{d \cdot n}} \Delta(\mathcal{A}, x_1, .., x_n). \tag{3}$$

A *n-sample partition rule* π_n is a mapping from $\mathbb{R}^{d \cdot n}$ to the space of finite-measurable partitions for $\mathbb{R}^d$, that we denote by $\mathcal{Q}$, where a *partition scheme* for $\mathbb{R}^d$ is a countable collection of n-sample partitions rules $\Pi = \{\pi_1, \pi_2, ...\}$. Let Π be an arbitrary partition scheme for $\mathbb{R}^d$, then for every partition rule

$\pi_n \in \Pi$ we can define its associated collection of measurable partitions by [7]

$$\mathcal{A}_n = \left\{\pi_n(x_1,..,x_n) : (x_1,..,x_n) \in \mathbb{R}^{d \cdot n}\right\}. \tag{4}$$

In this context, for a given n-sample partition rule π_n and a sequence $(x_1,..,x_n) \in \mathbb{R}^{d \cdot n}$, $\pi_n(x|x_1,..,x_n)$ denotes the mapping from any point x in $\mathbb{R}^d$ to its unique cell in $\pi_n(x_1,..,x_n)$, such that $x \in \pi_n(x|x_1,..,x_n)$.

C. Vapnik and Chervonenkis Concentration Inequalities

Let $X_1, X_2, .., X_n$ be independent identically distributed (i.i.d.) realizations of a random vector with values in $\mathbb{R}^d$, with $X \sim P$ and P a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then $\forall A \in \pi_n(X_1, X_2, .., X_n)$, we can define the empirical distribution by

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_A(X_i), \tag{5}$$

a probability measure defined on $(\mathbb{R}^d, \sigma(\pi_n(X_1,..,X_n)))$[1]. This is the abstract representation of the data-dependent partition scheme for probability estimation, where the i.i.d. samples are used twice: for defining a sub-sigma field $\sigma(\pi_n(X_1,..,X_n)) \subset \mathcal{B}(\mathbb{R}^d)$ and then again for characterizing the empirical distribution on it.

The following concentration inequality is used for proving the main result presented in this work.

LEMMA 1: (Lugosi and Nobel [7]) Let $X_1, X_2, .., X_n$ be i.i.d. realizations of a random vector X with distribution function P in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, and $\mathcal{A}$ a collection of measurable partitions for $\mathbb{R}^d$. Then $\forall n \in \mathbb{N}$, $\forall \epsilon > 0$,

$$\mathbb{P}\left(\sup_{\pi \in \mathcal{A}} \sum_{A \in \pi} |P_n(A) - P(A)| > \epsilon\right) \leq 4\Delta^*_{2n}(\mathcal{A}) 2^{\mathcal{M}(\mathcal{A})} \exp^{-\frac{n\epsilon^2}{32}}$$

where $\mathbb{P}$ denotes the distribution of the empirical process $X_1, .., X_n$.

III. Problem Statement

Here we focus on the important scenario of a finite dimensional Euclidean space $\mathbb{R}^d$, equipped with the Borel sigma field $\mathcal{B}(\mathbb{R}^d)$ and considering the Lebesgue sigma-finite measure λ as a reference. More precisely, let P and Q be probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, absolutely continuous with respect to λ (we denote the collection of Lebesgue dominated measures by $\mathcal{P}_\lambda(\mathbb{R}^d)$), such that $D(P||Q) < \infty$. Let $\Pi = \{\pi_1, \pi_2, \cdots\}$ be a partition scheme for $\mathbb{R}^d$, and let us consider $X_1, .., X_n$ and $Y_1, .., Y_n$ i.i.d. realizations of random variables with values in $\mathbb{R}^d$ and distributions P and Q, respectively.

We propose a data-driven histogram-based estimate of the divergence of the form,

$$D_{\pi_n(Y_1,..,Y_n)}(P^*_n||Q_n) \equiv \sum_{A \in \pi_n(Y_1,..,Y_n)} P^*_n(A) \cdot \log \frac{P^*_n(A)}{Q_n(A)}, \tag{6}$$

[1] $\sigma(\pi)$ denotes the smallest sigma-field that contain π, which for the case of partitions is the collection of sets that can be written as union of cells of π.

where P^*_n is a *Barron type of empirical measure* [10], given by,

$$P^*_n(A) \equiv (1 - a_n) \cdot P_n(A) + a_n \cdot Q_n(A), \tag{7}$$

$\forall A \in \sigma(\pi_n(Y_1, .., Y_n))$ with (a_n) a real sequence with values in $[0, 1]$, and P_n and Q_n the standard empirical measures in (5) induced by $X_1, .., X_n$ and $Y_1, .., Y_n$ respectively and restricted to the sub-sigma field $\sigma(\pi_n(Y_1, .., Y_n)) \subset \mathcal{B}(\mathbb{R}^d)$.

Note that $D_{\pi_n(Y_1,..,Y_n)}(P^*_n||Q_n)$ is a measurable function of $X_1, .., X_n$ and $Y_1, .., Y_n$, and consequently we are interested in studying the strong consistency of $D_{\pi_n(Y_1,..,Y_n)}(P^*_n||Q_n)$ — with respect to the joint distribution of $\{X_n, n \in \mathbb{N}\}$ and $\{Y_n, n \in \mathbb{N}\}$ — function of the aforementioned notions of combinatorial complexity for Π. The proposed construction is fundamentally based on the analysis of the following estimation-approximation error inequality,

$$\begin{aligned} &\left|D_{\pi_n(Y_1,..,Y_n)}(P^*_n||Q_n) - D(P||Q)\right| \leq \\ &\left|D_{\pi_n(Y_1,..,Y_n)}(P^*_n||Q_n) - D_{\pi_n(Y_1,..,Y_n)}(\tilde{P}_n||Q)\right| + \quad (8) \\ &\left|D_{\pi_n(Y_1,..,Y_n)}(\tilde{P}_n||Q) - D(P||Q)\right|, \quad (9) \end{aligned}$$

where $\tilde{P}_n(A) \equiv (1 - a_n) \cdot P(A) + a_n \cdot Q(A)$, $\forall A \in \pi_n(Y^n_1)$, error bound that in some way or another is presented in the consistency analysis of any histogram-based estimate [9].

The critical element to bound is the estimation error or variance in (8). For this we use two techniques. The first is due to Barron *et al.* [10] which is a smoothing technique (7) for estimating the RD derivative $\frac{\delta P(x)}{\delta Q(x)}$, which can be seen as the sufficient statistics of the problem. This smoothing technique was originally proposed for the problem of estimating probability measures consistent in direct information divergence [10], when the target probability distribution is dominated by a sigma-finite measure. The second is a condition on the partition scheme Π, where we impose that $Q_n(A) > \frac{k_n}{n}$, $\forall A \in \sigma(\pi_n(Y_1, .., Y_n))$, (k_n) denoting the critical mass for every bin. Both design sequences (a_n) and (k_n) are strictly positive and provide a way of ensuring a minimum probability mass for both P^*_n and Q_n in $(\mathbb{R}^d, \sigma(\pi_n(Y^n_1)))$, which in conjunction with the distribution free concentration inequalities presented in Section II, are the key elements to bound the estimation error in (8). On the other hand for the approximation error or bias, we have chosen the data-dependent partition as only a function of the i.i.d. realizations associated with the reference measure Q. Loosely speaking this choice of using partial information can be justified by the fact that $P \ll Q$ [2]. The next section present the statement and the proof of this result.

IV. Condition for Strong Consistency

Before presenting the main result, we introduce a generalization of the Vapnik-Chervonenkis inequality [8], [9] for the kind of mixture empirical distributions adopted in our divergence estimate, Eq. (7).

THEOREM 1: Let $\mathcal{A}$ be a collection of measurable events and P and Q be probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, Let us consider $X_1, .., X_n$ and $Y_1, .., Y_n$ i.i.d. realizations driven by

P and Q respectively and inducing the empirical distributions P_n and Q_n, respectively. Then $\forall a \in [0,1]$, $\forall \epsilon > 0$ and $\forall n > 0$,

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\mu_n^a(A) - \mu^a(A)| > \epsilon\right) \leq 8S_{2n}(\mathcal{A})\exp^{-\frac{n\epsilon^2}{32}}, \quad (10)$$

where $\mu^a(A) = (1-a)\cdot P(A) + a\cdot Q(A)$ and $\mu_n^a(A) = (1-a)\cdot P_n(A) + a\cdot Q_n(A)$ are the mixing and empirical mixing distributions, and $S_{2n}(\mathcal{A})$ denotes the scatter coefficient of $\mathcal{A}$ [9], [8].

The proof of this results is a natural consequence of the classical VC inequality. The argument is presented in Appendix I.

A corollary of this concentration inequality provides a version of Lemma 1 for the case of mixing empirical distributions.

COROLLARY *1:* (Lugosi *et al.* [7]) Under the same assumptions of Theorem 1, let $\mathcal{A}$ be instead a collection of measurable partitions for $\mathcal{X}$, then $\forall a \in [0,1]$, $\forall \epsilon > 0$ and $\forall n > 0$,

$$\mathbb{P}\left(\sup_{\pi\in\mathcal{A}} \sum_{A\in\pi} |\mu_n^a(A) - \mu^a(A)| > \epsilon\right) \leq 8\Delta_{2n}^*(\mathcal{A})2^{\mathcal{M}(\mathcal{A})}\exp^{-\frac{n\epsilon^2}{128}}. \quad (11)$$

The proof follows the same arguments proposed by Lugosi and Nobel in [7] and consequently is omitted.

A. Main Result

We have all machinery to state our main result. Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two sequences of non-negative real numbers. We say that (a_n) dominates (b_n), denoted by $(b_n) \preceq (a_n)$ (or alternatively (b_n) is $O(a_n)$), if there exists $C > 0$ and $k \in \mathbb{N}$ such that $b_n \leq C \cdot a_n$ $\forall n \geq k$. We say that $(b_n)_{n\in\mathbb{N}}$ and $(a_n)_{n\in\mathbb{N}}$ (both strictly positive) are asymptotically equivalent, denoted by $(b_n) \approx (a_n)$, if there exits $C > 0$ such that $\lim_{n\to\infty} \frac{a_n}{b_n} = C$, and on the other hand, we say that (a_n) is $o(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

THEOREM 2: Let P and Q be probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $D(P||Q) < \infty$. Let $X_1,..,X_n$ and $Y_1,..,Y_n$ be i.i.d. realizations of P and Q, respectively, and $\Pi = \{\pi_1, \pi_2, ...\}$ a partition scheme with associated sequence of measurable partitions $\mathcal{A}_1, \mathcal{A}_2, \cdots$. If for some $l \in (0,1)$, there exists $p \in (0, \frac{l}{2})$, $\tau \in (0, l-2p]$ and (k_n) a non-negative sequence, such that:

- **a)** $\lim_{n\to\infty} n^{-\tau}\mathcal{M}(\mathcal{A}_n) = 0$,
- **b)** $\lim_{n\to\infty} n^{-\tau}\log\Delta_n^*(\mathcal{A}_n) = 0$,
- **c)** $(a_n) \succeq (n^{-p})$,
- **d)** (k_n) is $o(n)$, $(k_n) \succeq (n^{0.5+l/2})$ and $\forall n \in \mathbb{N}$, $\forall (y_1,..,y_n) \in \mathcal{X}^n$, $\inf_{A\in\pi_n(y_1^n)} Q_n(A) \geq \frac{k_n}{n}$,
- **e)** $\forall \gamma > 0$,

$$\lim_{n\to\infty} Q\left(\left\{x\in\mathbb{R}^d : diam(\pi_m(x|Y_1,..,Y_n)) > \gamma\right\}\right) = 0,$$

almost surely and $\lim_{n\to\infty} a_n = 0$,

then

$$\lim_{n\to\infty} D_{\pi_n(Y_1^n)}(P_n^*||Q_n) = D(P||Q), \quad (12)$$

with probability one with respect to the joint process distribution.

There are two important considerations to be taken into account in the proof of this theorem. First, the asymptotically sufficient nature of the adaptive quantization framework Π, considered in **e)** above, and second, the generalization ability of the learning approach, how relative frequencies converge uniformly, in some sense, to their respective probabilities for the estimation of the divergence, considered by **a)**, **b)**, **c)** and **d)**. Before going to the proof, the following key concentration result will be used.

LEMMA 2: Let $X_1,..,X_n$ and $Y_1,..,Y_n$ be i.i.d. driven by P and Q respectively and Π a partition scheme as presented in *Theorem* 2. If the conditions **a)**, **b)**, **c)** and **d)** of *Theorem* 2 for some $l \in (0,1)$, $0 < p < \frac{l}{2}$ and $\tau \in (0, l-2p]$ are satisfied, then

$$\lim_{n\to\infty} \sup_{A\in\pi_n(Y_1^n)} \left|\frac{Q(A)}{Q_n(A)} - 1\right| = 0, \quad (13)$$

$$\lim_{n\to\infty} \sup_{A\in\pi_n(Y_1^n)} \left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right| = 0, \quad (14)$$

almost surely with respect to the joint process distribution. (Proof in *Appendix* II).

Proof: We consider the following estimation-approximation error bound:

$$\begin{aligned}\left| D_{\pi_n(Y_1^n)}(P_n^*||Q_n) - D(P||Q)\right| &\leq \\ \left| D_{\pi_n(Y_1^n)}(P_n^*||Q_n) - D_{\pi_n(Y_1^n)}(\tilde{P}_n||Q)\right| & \quad (15)\\ + \left| D_{\pi_n(Y_1^n)}(\tilde{P}_n||Q) - D(P||Q)\right|. & \quad (16)\end{aligned}$$

Then it is sufficient to prove that both the estimation and approximation error terms converge to zero almost surely as n tends to infinity. The result for the approximation error term in (16) can be derived from the arguments in [2] (not reported here for space considerations). For the estimation error in (15), we have that,

$$\left| D_{\pi_n(Y_1^n)}(P_n^*||Q_n) - D_{\pi_n(Y_1^n)}(\tilde{P}_n||Q)\right| \leq$$

$$\left|\sum_{A\in\pi_n(Y_1^n)} P_n^*(A)\cdot\log P_n^*(A) - \sum_{A\in\pi_n(Y_1^n)} \tilde{P}_n(A)\cdot\log\tilde{P}_n(A)\right| \quad (17)$$

$$+\left|\sum_{A\in\pi_n(Y_1^n)} \tilde{P}_n(A)\cdot\log Q(A) - \sum_{A\in\pi_n(Y_1^n)} P_n^*(A)\cdot\log Q_n(A)\right|. \quad (18)$$

Expression (17): From the construction of P_n^* on the events of $\pi_n(Y_1^n)$, the expression in (17) can be upper bounded by,

$$\log\frac{1}{a_n\cdot b_n}\cdot\sup_{\pi\in\mathcal{A}_n}\left|\sum_{A\in\pi} P_n^*(A) - \tilde{P}_n(A)\right| + \sup_{A\in\pi_n(Y_1^n)}\left|\log P_n^*(A) - \log\tilde{P}_n(A)\right|. \quad (19)$$

where $b_n \equiv \frac{k_n}{n}$ and without loss of generality we assume that $a_n < 1$ and $b_n < 1$, $\forall n > 0$. From Corollary 1 and conditions

c) and **d)**, it is simple to show that, $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \frac{\log \mathbb{P}\left(\sup_{\pi\in\mathcal{A}_n}\left|\sum_{A\in\pi} P_n^*(A) - \tilde{P}_n(A)\right| > \frac{\epsilon}{\log\frac{1}{a_n\cdot b_n}}\right)}{n} < 0, \tag{20}$$

then from *Borel-Cantelli Lemma* the first term of (19) tends to zero almost-surely. Concerning the second term in (19), from (14) it is simple to prove that $\lim_{n\to\infty}\sup_{A\in\pi_n(Y_1^n)}\frac{\tilde{P}_n(A)}{P_n^*(A)} = 1$ and $\lim_{n\to\infty}\sup_{A\in\pi_n(Y_1^n)}\frac{P_n^*(A)}{\tilde{P}_n(A)} = 1$ almost surely [2]. On the other hand, we have that $\forall A \in \pi_n(Y_1^n)$,

$$\left|\frac{P_n^*(A)}{\tilde{P}_n(A)} - 1\right| \leq \frac{\left|\tilde{P}_n(A) - P_n^*(A)\right|}{P_n^*(A)} \cdot \frac{P_n^*(A)}{\tilde{P}_n(A)}, \tag{21}$$

then

$$\lim_{n\to\infty}\sup_{A\in\pi_n(Y_1^n)}\left|\frac{P_n^*(A)}{\tilde{P}_n(A)} - 1\right| = 0, \tag{22}$$

almost surely from (13) and (21). Then considering that $|\log(x)| \leq \max\left\{x-1, \frac{1}{x}-1\right\}$ $\forall x > 0$, it follows that $\forall n$,

$$\begin{aligned}
&\sup_{A\in\pi_n(Y_1^n)}\left|\log\frac{\tilde{P}_n(A)}{P_n^*(A)}\right| \leq \\
&\sup_{A\in\pi_n(Y_1^n)}\max\left\{\left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right|, \left|\frac{P_n^*(A)}{\tilde{P}_n(A)} - 1\right|\right\} \leq \\
&\max\left\{\sup_{A\in\pi_n(Y_1^n)}\left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right|, \sup_{A\in\pi_n(Y_1^n)}\left|\frac{P_n^*(A)}{\tilde{P}_n(A)} - 1\right|\right\},
\end{aligned}$$

then from (14) and (22) we have the result.

Expression (18): On the other hand from the construction of Q_n on the events of $\pi_n(Y_1^n)$, the expression in (18) can be upper bounded by: $\log\frac{1}{b_n}\cdot\sup_{\pi\in\mathcal{A}_n}\left|\sum_{A\in\pi} P_n^*(A) - \tilde{P}_n(A)\right| + \sup_{A\in\pi_n(Y_1^n)}|\log Q_n(A) - \log Q(A)|$. The same arguments presented above with marginal variations apply to prove that the two terms of this bound tend to zero almost surely, by for this case adopting Lemma 1 and the concentration inequality in (13). ∎

V. Application to Statistically Equivalent Partitions

Let us consider the real line $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as the measurable space. and a partition scheme that dichotomizes the space in statistically equivalent intervals. More precisely, let $Y_1, .., Y_n$ be the i.i.d. realizations drawn from $Q \in \mathcal{P}_\lambda(\mathbb{R})$. The order statistics $Y^{(1)}, Y^{(2)}, .., Y^{(n)}$ is defined as the permutation of $Y_1, .., Y_n$ such that $Y^{(1)} < Y^{(2)} < \cdots < Y^{(n)}$ — this permutation exists with probability one as Q is absolutely continuous with respect to the Lebesgue measure λ. Based on this sequence, the resulting l_n-spacing partition rule is given by

$$\begin{aligned}
\pi_n(Y_1^n) &= \{I_i^n : i = 1, .., T_n\} \\
&= \left\{(-\infty, Y^{(l_n)}], (Y^{(l_n)}, Y^{(2l_n)}], .., (Y^{((T_n-1)l_n)}, \infty)\right\},
\end{aligned}$$

where $T_n = \lfloor n/l_n \rfloor$ assuming the non-trivial case where $n > l_n$. Note that under this construction every cell of $\pi_n(Y_1^n)$ has at least l_n samples from $Y_1, .., Y_n$, which match one of the design constraints of our construction (condition **d)** of Theorem 2). Then we can state the following result.

THEOREM 3: Let us consider the l_n-spacing partition scheme and the resulting histogram-based estimate from (6), with $(l_n) \approx (n^{0.5+l/2})$ and $(a_n) \approx (n^{-p})$. Then there exists a range of design parameters $\mathcal{D} = \left\{(l,p) \in \mathbb{R}^2 : l \in (0,1), p \in (0, \frac{l}{2}), 1 + 4p < 3l\right\} \neq \emptyset$ such that for any pair of probabilities P, Q in $\mathcal{P}_\lambda(\mathbb{R})$,

$$\lim_{n\to\infty} D_{\pi_n(Y_1^n)}(P_n^*||Q_n) = D(P||Q), \tag{23}$$

$\mathbb{P}$-almost surely.

Proof: The proof reduces to checking the sufficient conditions of Theorem 2. First note that **c)** and **d)** are satisfied as part of the design constraint of the estimate. Concerning **a)**, again by construction we have that $\mathcal{M}(\mathcal{A}_n) \leq n/l_n + 1$, then considering $\tau = (l-2p)$ $n^{-(l-2p)}\mathcal{M}(\mathcal{A}_n) \leq n^{-(l-2p)}/l_n + n^{-(l-2p)}$. Given that $(l_n) \approx (n^{0.5+l/2})$, $p < \frac{l}{2}$ and $1-3l < 4p$, it simple to check that,

$$\lim_{n\to\infty} n^{-(l-2p)}\mathcal{M}(\mathcal{A}_n) = 0. \tag{24}$$

For condition **b)**, Lugosi *et al.* [7] showed that $\Delta_n^*(\mathcal{A}_n) = \binom{T_n+n}{n}$, where using that $\log\binom{s}{t} \leq s\cdot h(t/s)$ [9], with $h(x) = -x\log(x) - (1-x)\log(1-x)$ for $x \in [0,1]$ — the binary entropy function [4], with the critical $\tau = l-2p > 0$ it follows that,

$$\begin{aligned}
n^{-(l-2p)}\log\left(\Delta_n^*(\mathcal{A}_n)\right) &= n^{-(l-2p)}\cdot\log\binom{T_n+n}{n} \\
&\leq n^{-(l-2p)}\cdot(n+T_n)\cdot h\left(\frac{n}{n+T_n}\right) \\
&\leq 2n^{1-(l-2p)}\cdot h\left(\frac{T_n}{n}\right) \leq 2n^{1-(l-2p)}\cdot h\left(\frac{1}{l_n}\right)
\end{aligned} \tag{25}$$

Consequently we have that,

$$\begin{aligned}
n^{-(l-2p)}\log(\Delta_n^*(\mathcal{A}_n)) \leq -&\frac{2n^{1-(l-2p)}}{l_n}\log(1/l_n) \\
&- 2n^{1-(l-2p)}(1-1/l_n)\log(1-1/l_n).
\end{aligned} \tag{26}$$

The first term on the right hand side (RHS) of (26) behaves like $O(n^{0.5(1-3l+4p)}\cdot\log(l_n))$, where as long as the exponent of the first term is negative (equivalent to $1+4p < 3l$) this sequence asymptotically tends to zero as by construction $(l_n) \preceq (n)$. The second term on the RHS of (26) behaves asymptotically like $-n^{1-(l-2p)}\cdot\log(1-1/l_n)$ which from $\log(x) \leq x-1$ is upper bounded by $(\frac{n^{1-(l-2p)}}{l_n}\cdot\frac{1}{1-1/l_n}) \approx (\frac{n^{1-(l-2p)}}{l_n})$. This last sequence tends to zero from $(l_n) \approx (m^{0.5+l/2})$ and $1+4p < 3l$. Consequently from (26), $\lim_{n\to\infty} n^{-(l-2p)}\log(\Delta_n^*(\mathcal{A}_n)) = 0$.

Finally for condition **e)**, Lugosi *et al.* [7] (*Theorem 4*) proved that it is sufficient to show that $\lim_{n\to\infty}\frac{l_n}{n} = 0$, and given that by construction (a_n) tends to zero asymmetrically we prove the theorem. ∎

VI. Final Remarks

The presented formulation offers the possibility of extending this type of histogram-based construction and results to the estimation of other information theoretic quantities — like the mutual information and more general family of Ali-Silvey divergence functionals, as well as using the rich machinery of statistical learning theory [12], [9] to explore, in this context, distribution-free rate of convergence results.

VII. Acknowledgment

This material is based upon work supported by awards of CONICYT-Chile, Fondecyt 1090138.

Appendix I
Proof of Theorem 1

Proof: It is simple to show that independent of $a \in [0,1]$,

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\mu_n^a(A) - \mu^a(A)| > \epsilon\right) \leq$$
$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| + \sup_{A\in\mathcal{A}} |Q_n(A) - Q(A)| > \epsilon\right) \tag{27}$$

where considering that $\{: \sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \leq \frac{\epsilon}{2}\} \cap \{: \sup_{A\in\mathcal{A}} |Q_n(A) - Q(A)| \leq \frac{\epsilon}{2}\}$ is contained in $\{: \sup_{A\in\mathcal{A}} |P_n(A) - P(A)| + \sup_{A\in\mathcal{A}} |Q_n(A) - Q(A)| \leq \epsilon\}$, by the union bound we obtain that,

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| + \sup_{A\in\mathcal{A}} |Q_n(A) - Q(A)| > \epsilon\right) \leq$$
$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \frac{\epsilon}{2}\right)$$
$$+\mathbb{P}\left(\sup_{A\in\mathcal{A}} |Q_n(A) - Q(A)| > \frac{\epsilon}{2}\right) \leq 8 \cdot \mathcal{S}_{2n}(\mathcal{A}) \exp^{\frac{-n\epsilon^2}{4\cdot 8}}.$$

The last inequality is because of the distribution free nature of the classical VC inequality [8], [9]. ■

Appendix II
Proof of *Lemma* 2

Proof: Let us focus on proving Eq. (14) and consequently in the probability of the following event,

$$\mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)} \left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right| > \epsilon\right)$$
$$\leq \mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)} \left|\tilde{P}_n(A) - P_n^*(A)\right| > \epsilon \cdot a_n b_n\right). \tag{28}$$

This last inequality is by the hypothesis, where $P_n^*(A) \geq a_n \cdot b_n$ $\forall A \in \pi_n(Y_1^n)$ with $b_n \equiv \frac{k_n}{n}$, $\forall n > 0$. From the VC concentration inequality for mixture distributions, Corollary 1, the probability of our target event in (28) is less than or equal to,

$$\mathbb{P}\left(\sup_{A\in\sigma(\pi_n(Y_1^n))} \left|\tilde{P}_n(A) - P_n^*(A)\right| > \epsilon \cdot a_n \cdot b_n\right)$$
$$\leq \mathbb{P}\left(\sup_{\pi\in\mathcal{A}_n} \sum_{A\in\pi} \left|\tilde{P}_n(A) - P_n^*(A)\right| > 2\epsilon \cdot a_n \cdot b_n\right)$$
$$\leq 8\Delta_{2n}^*(\mathcal{A}_n) 2^{\mathcal{M}(\mathcal{A}_n)} \exp^{-\frac{n(\epsilon\cdot a_n\cdot b_n)^2}{32}}. \tag{29}$$

Finally (29) and simple algebra can show that under the conditions **a)**, **b)**, **c)** and **d)**,

$$(n^{-\tau} \log \mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)} \left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right| > \epsilon\right))$$
$$\preceq (-n^{1-\tau}\frac{(\epsilon\cdot a_n\cdot b_n)^2}{32}) \preceq (-\epsilon^2 \cdot n^{(l-2p)-\tau}). \tag{30}$$

In particular the last relationship in (30) considers that $(a_n) \succeq (n^{-p})$ and $(b_n) \succeq (n^{l/2-0.5})$. Then it is clear that $\lim_{n\to\infty} n^{-\tau} \log \mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)} \left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right| > \epsilon\right) < 0$ or diverge to $-\infty$. This simply implies that $\exists C > 0$ such that,

$$(\mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)} \left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right| > \epsilon\right)) \preceq (\exp^{-n^\tau C}),$$

then the *Borel-Cantelli lemma* we prove the result.

The same arguments can be adopted to show Eq.(13) but in this case using the classical VC concentration inequality stated in Lemma 1. In fact weaker conditions can be stated to prove that this term converges to zero almost surely. In that sense the critical part was to bound the deviation of P_n^* with respect to $\tilde{P}_n$ in $(\mathbb{R}^d, \sigma(\pi_n(Y_1^n)))$. ■

References

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

[2] J. Silva and S. Narayanan, "Universal consistency of data-driven partitions for divergence estimation," in *IEEE International Symposium on Information Theory*. IEEE, 2007.

[3] XuanLong Nguyen, Martin Wainwright, and Michael Jordan, "Nonparametric estimation of the likelihood ratio and divergence functionals," in *IEEE International Symposium on Information Theory*. IEEE, June 2007.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, New York, 1991.

[5] R. M. Gray, *Entropy and Information Theory*, Springer - Verlag, New York, 1990.

[6] S. Kullback, *Information theory and Statistics*, New York: Wiley, 1958.

[7] G. Lugosi and Andrew B. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *The Annals of Statistics*, vol. 24, no. 2, pp. 687–706, 1996.

[8] Vladimir Vapnik and A. J. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability Apl.*, vol. 16, pp. 264–280, 1971.

[9] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, New York: Springer-Verlag, 1996.

[10] Andrew Barron, L. Gyorfi, and Edward C. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1437–1454, September 1992.

[11] M. P. Gessaman, "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," *Ann. Math. Statist.*, vol. 41, pp. 1344–1346, 1970.

[12] Vladimir Vapnik, *Statistical Learning Theory*, John Wiley, 1998.