Universal Consistency of Data-Driven Partitions for Divergence Estimation

Jorge Silva and Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory, *http://sail.usc.edu*
University of Southern California, Viterbi School of Engineering
jorgesil@usc.edu, shri@sipi.usc.edu

***Abstract*— This paper presents a general histogram based divergence estimator based on data-dependent partition. Sufficient conditions for the universal strong consistency of the data-driven divergence estimator, using Lugosi and Nobel's combinatorial notions for partition families, are presented. As a corollary this result is particularized for the emblematic case of l_m-spacing quantization scheme.**

I. Introduction

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 [1], divergence [2] or Kullback-Leibler divergence [3] 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 $P \ll Q$ [2], which makes the Radon-Nicodym (RD) derivative of P with respect to Q to be well defined in (1). Divergence is a fundamental quantity in information theory [1], also used as an indicator of the difficulty in discriminating between probabilistic models in statistical decision theory [3] and fundamental to characterize the rate function, which reflects the exponential decay of convergence of the empirical measures to their probabilities (*Sanov's theorem*), in large deviations [4].

Despite its theoretical and practical significance little work has been conducted for the universal estimation of the divergence, see [5] and references therein. In this direction, Wang *et al.* [5] recently presented a universal histogram-based divergence estimation. This work considers the RD derivative $\frac{\partial P}{\partial Q}$ by an adaptive partition scheme that approximates statistical equivalent intervals relative to the reference measure Q in (1). In this work we extend consistency results for this type of histogram-based divergence estimation considering more general properties on the adaptive partition scheme. We specify general sufficient conditions, similar to that proposed by Lugosi and Nobel [6] in the context of histogram based density estimation, for the proposed data-driven divergence estimation to be strongly consistent. As a corollary, we present sufficient conditions for the l_m-spacing quantization scheme to be strongly consistent.

A. Preliminaries

Let $\mathcal{X}$=$\mathbb{R}^d$ be a finite-dimensional Euclidian space with corresponding Borel sigma field $\mathcal{B}(\mathbb{R}^d)$. We say $\pi = \{A_1, .., A_r\}$ is a finite measurable partition 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 [6]. 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 [6]

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

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 [6]

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

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 $x \in \pi_n(x|x_1, .., x_n)$.

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

[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.

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 result presented by Lugosi and Nobel [6] is used for proving the main result presented in this work. This is a natural consequence of the celebrated *Vapnik-Chervonenkis inequality* [7], [8].

LEMMA 1: (Lugosi and Nobel [6]) 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 $X_1, .., X_n$.

The following section presents the general data-driven histogram framework for estimation of the divergence, and the main result characterizing sufficient conditions for the strong universal consistency of this estimation techniques.

II. Data-Dependent Partition for Divergence Estimation

Let P and Q be probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ absolutely continuous with respect to the Lebesgue measure, 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 $Y_1, .., Y_m$ i.i.d. realizations of random variables with values in $\mathbb{R}^d$, and distributions P and Q, respectively. Then a natural candidate for the empirical divergence is given by

$$\hat{D}_{n,m}(P||Q) = \sum_{A \in \pi_m(Y_1,..,Y_m)} P_n(A) \cdot \log \frac{P_n(A)}{Q_m(A)},$$

where P_n and Q_m respectively are the empirical distributions induced by $X_1, .., X_n$ and $Y_1, .., Y_m$, (5), in the sub-sigma field $\sigma(\pi_m(Y_1, .., Y_m)) \subset \mathcal{B}(\mathbb{R}^d)$. As suggested in [5], this construction only considers realizations associated with the reference measure Q for defining the data-dependent partition. We impose in Π the desirable condition that $Q_m(A) > 0$, $\forall A \in \pi_m(Y_1, .., Y_m)$, that ensures that $\hat{D}_{n,m}(P||Q) < \infty$. Note that $\hat{D}_{n,m}(P||Q)$ is a measurable function of $X_1, .., X_n$ and $Y_1, .., Y_m$, and consequently we are interested in studying the strong — almost surely with respect to the joint distribution of $\{X_n, n \in \mathbb{N}\}$ and $\{Y_m, m \in \mathbb{N}\}$ — universal consistency of $\hat{D}_{n,m}(P||Q)$ as m and n tend to infinity and as a function of the aforementioned notions of combinatorial complexity for Π.

Before presenting the main result let us introduce some basic definitions. For any $A \in \mathcal{B}(\mathbb{R}^d)$, we define its *diameter* by $diam(A) = \sup_{x,y \in A} ||x - y||$, where $||\cdot||$ refers to the *Euclidian norm* in $\mathbb{R}^d$. In addition, let us consider $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$ two sequences of non-negative real numbers. We say that (a_n) dominates (b_n), denoted by $(b_n) \ll (a_n)$, if there exists $C > 0$ and $k \in \mathbb{N}$ st. $b_n \leq C \cdot a_n$ for all $n \geq k$. We say that (b_n) and (a_n) are asymptotically equivalent, denoted by $(b_n) \approx (a_n)$, if there exits $C > 0$ st. $\lim_{n \to \infty} \frac{a_n}{b_n} = C$.

THEOREM 1: Let P and Q be probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ absolutely continuous with respect to the Lebesgue measure, such that $D(P||Q) < \infty$. Let $X_1, .., X_n$ and $Y_1, .., Y_m$ 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 $l \in (0, 1)$, we have that as m tends to infinity

- **a)** $m^{-l}\mathcal{M}(\mathcal{A}_m) \to 0$,
- **b)** $m^{-l} \log \Delta^*_m(\mathcal{A}_m) \to 0$,
- **c)** $\exists (k_m) \approx (m^{0.5+l/2})$ such that, $\forall m \in \mathbb{N}$, $\forall (y_1, .., y_m) \in \mathbb{R}^{d \cdot m}$, $\forall A \in \pi_m(y_1, .., y_m), Q_m(A) \geq \frac{k_m}{m}$,
- **d)** $\forall \gamma > 0, Q\left(x \in \mathbb{R}^d : diam(\pi_m(x|Y_1, .., Y_m)) > \gamma\right) \to 0$ almost surely with respect to the process distribution of $Y_1, Y_2, \cdots$,

then

$$\lim_{m \to \infty} \lim_{n \to \infty} \hat{D}_{m,n}(P||Q) = D(P||Q) \tag{6}$$

with probability one.

Proof: There are two important considerations to be taken into account in the proof. First, the asymptotic sufficient nature of the adaptive quantization framework Π, implicitly considered in **d)**, and second, the generalization ability of the learning approach, how relative frequencies converge uniformly to their respective probabilities for the estimation of the divergence, considered in **a)**, **b)** and **c)**. Let us define $Y_1^m \equiv Y_1, ..., Y_m$. The proof will be based on the following inequality:

$$\begin{aligned}
&\left|\hat{D}_{m,n}(P||Q) - D(P||Q)\right| \leq \\
&\left|\sum_{A \in \pi_m(Y_1^m)} P_n(A) \cdot \log \frac{P_n(A)}{Q_m(A)} - \sum_{A \in \pi_m(Y_1^m)} P_n(A) \cdot \log \frac{P_n(A)}{Q(A)}\right| \\
&+ \left|\sum_{A \in \pi_m(Y_1^m)} P_n(A) \cdot \log \frac{P_n(A)}{Q(A)} - \sum_{A \in \pi_m(Y_1^m)} P(A) \cdot \log \frac{P(A)}{Q(A)}\right| \\
&+ \left|\sum_{A \in \pi_m(Y_1^m)} P(A) \cdot \log \frac{P(A)}{Q(A)} - D(P||Q)\right|.
\end{aligned} \tag{7}$$

Then it is sufficient to prove that the three terms in the right side of the inequality converge to zero almost surely as n tends to infinity and as m tends to infinity. We will prove these three cases (indexed from top to bottom) separately.

Term 1: Let us consider $X_1, .., X_n$ and $Y_1, .., Y_m$, then the first term is upper bound by[2]

$$\leq \sum_{A \in \pi_m(Y_1^m)} P_n(A) \left|\log Q(A) - \log Q_m(A)\right| \tag{8}$$

$$\leq \sup_{A \in \pi_m(Y_1^m)} \left|\log Q(A) - \log Q_m(A)\right|. \tag{9}$$

[2] By construction of Π (condition **c)**), $\forall A \in \pi_m(Y_1, .., Y_m)$, $Q_m(A) > 0$. On the other hand, the event $Q_m(A) > 0$ and $Q(A) = 0$ has probability zero, more precisely $\mathbb{P}(y_1^m : Q(A) > 0, \forall A \in \pi_m(y_1^m)) = 1$, consequently the first term and upper bound in (9) are well defined with probability one.

Note that this upper bound is independent of $X_1, ..., X_n$, and then it only involves the distribution of $Y_1, .., Y_m$. The following lemma will be used to prove that (9) tends to zero almost surely as m tends to infinity.

LEMMA 2: Let $Y_1, .., Y_m$ be i.i.d. realizations of a random variable with probability measure Q in $\mathbb{R}^d$ and Π a partition scheme as presented in *Theorem* 1. If the conditions **a)**, **b)** and **c)** of *Theorem* 1 are satisfied for some $l \in (0,1)$, then

$$\lim_{m\to\infty} \sup_{A\in\pi_m(Y_1^m)} \left| \frac{Q(A)}{Q_m(A)} - 1 \right| = 0, \tag{10}$$

almost surely with respect to the process distribution of $Y_1, Y_2, \cdots$. The proof is presented in *Appendix* I.

From (10) it is simple to prove that $\lim_{m\to\infty} \sup_{A\in\pi_m(Y_1^m)} \frac{Q(A)}{Q_m(A)} = 1$ and $\lim_{m\to\infty} \sup_{A\in\pi_m(Y_1^m)} \frac{Q_m(A)}{Q(A)} = 1$ almost surely. On the other hand, we have that for all $A \in \pi_m(Y_1^m)$,

$$\left| \frac{Q_m(A)}{Q(A)} - 1 \right| \leq \frac{|Q(A) - Q_m(A)|}{Q_m(A)} \cdot \frac{Q_m(A)}{Q(A)}, \tag{11}$$

then

$$\lim_{m\to\infty} \sup_{A\in\pi_m(Y_1^m)} \left| \frac{Q_m(A)}{Q(A)} - 1 \right| = 0, \tag{12}$$

almost surely from (10) and (11). Finally, we have that $|\log(x)| \leq \max\left\{x-1, \frac{1}{x}-1\right\}$ for all $x > 0$, consequently it follows that $\forall m$,

$$\begin{aligned}
&\sup_{A\in\pi_m(Y_1^m)} \left| \log \frac{Q(A)}{Q_m(A)} \right| \leq \\
&\sup_{A\in\pi_m(Y_1^m)} \max\left\{ \left| \frac{Q(A)}{Q_m(A)} - 1 \right|, \left| \frac{Q_m(A)}{Q(A)} - 1 \right| \right\} \leq \\
&\max\left\{ \sup_{A\in\pi_m(Y_1^m)} \left| \frac{Q(A)}{Q_m(A)} - 1 \right|, \sup_{A\in\pi_m(Y_1^m)} \left| \frac{Q_m(A)}{Q(A)} - 1 \right| \right\},
\end{aligned}$$

where using (10) and (12), we have that $\lim_{m\to\infty} \sup_{A\in\pi_m(Y_1^m)} \left| \log \frac{Q(A)}{Q_m(A)} \right| = 0$ almost surely, proving the result from (9).

<u>**Term 2:**</u> (by SLLN) The second term of (7) can be upper bound by,

$$\begin{aligned}
&\left| \sum_{A\in\pi_m(Y_1^m)} P_n(A) \cdot \log P_n(A) - \sum_{A\in\pi_m(Y_1^m)} P(A) \cdot \log P(A) \right| \\
&+ \left| \sum_{A\in\pi_m(Y_1^m)} (P_n(A) - P(A)) \cdot \log \frac{1}{Q(A)} \right|
\end{aligned} \tag{13}$$

where $\left| \sum_{A\in\pi_m(Y_1^m)} (P_n(A) - P(A)) \cdot \log \frac{1}{Q(A)} \right| \leq$

$$\begin{aligned}
&2 \sup_{A\in\pi_m(Y_1^m)} \left| \log \frac{Q_m(A)}{Q(A)} \right| + \\
&\left| \sum_{A\in\pi_m(Y_1^m)} (P_n(A) - P(A)) \cdot \log \frac{1}{Q_m(A)} \right|.
\end{aligned} \tag{14}$$

Let us condition on a realization of $Y_1, .., Y_m$ and consequently we fix the measurable partition $\pi_m(Y_1^m)$. Then by the SLLN [10], we have that $\forall A \in \pi_m(Y_1^m)$,

$$\lim_{n\to\infty} P_n(A) = P(A) \tag{15}$$

almost surely with respect to the distribution of $X_1, X_2, \cdots$. Then given that $x \log x$ is a continuous real function, and $|\pi_m(Y_1^m)| < \infty$, then it is straightforward to prove that, $\lim_{n\to\infty} \sum_{A\in\pi_m(Y_1^m)} P_n(A) \cdot \log P_n(A) = \sum_{A\in\pi_m(Y_1^m)} P(A) \cdot \log P(A)$, and $\lim_{n\to\infty} \sum_{A\in\pi_m(Y_1^m)} P_n(A) \cdot \log \frac{1}{Q_m(A)} = \sum_{A\in\pi_m(Y_1^m)} P(A) \cdot \log \frac{1}{Q_m(A)}$ almost surely with respect to the distribution of $X_1, X_2, ...$ given $Y_1, ..., Y_m$. This proves that the first term in (13) and the second term of (14) converge to zero as n tends to infinity for any $m \in \mathbb{N}$ and any realization of $Y_1, ..., Y_m$. Then, it is simple to show that these two terms are zero almost surely as in addition m tends to infinity. Finally, the first term of (14) tends to zero as m tends to infinity almost surely from *Lemma* 2.

<u>**Term 3:**</u> (Approximation argument) Note that this last term only depends on the partition sequence $\pi_1(Y_1), \pi_2(Y_1, Y_2), \cdots, \pi_m(Y_1^m), \cdots$, and consequently depends strictly on the process distribution of $Y_1, Y_2, \cdots$.

By definition condition **d)** is equivalent to

$$\lim_{m\to\infty} Q\left(\bigcup_{\substack{A\in\pi_m(Y_1^m)\\ diam(A)>\gamma}} A \right) = 0, \tag{16}$$

almost surely $\forall \gamma > 0$.

Let us consider an admissible realization of the process $y_1, y_2 \cdots$, i.e. a realization where (16) holds. Let us define the measurable sequence of events $B_m = \bigcup_{\substack{A\in\pi_m(y_1^m)\\ diam(A)>\gamma}} A \in \mathcal{B}(\mathbb{R}^d)$, $\forall m \in \mathbb{N}$. From the fact that $P \ll Q$ and (16), $f_m(x) = \frac{\partial P}{\partial Q}(x) \cdot \mathbb{1}_{B_m}(x)$ tends to zero as m tends to infinity for Q-almost every $x \in \mathbb{R}^d$. Given that $f_m(x) \leq \frac{\partial P}{\partial Q}(x)$, this last one Q-integrable, the application of the *dominated convergence Theorem* implies that [10], [11]

$$\lim_{m\to\infty} \int \frac{\partial P}{\partial Q}(x) \cdot \mathbb{1}_{B_m}(x) \cdot \partial Q(x) = 0 \Leftrightarrow \lim_{m\to\infty} P(B_m) = 0.$$

This implies from (16) that

$$\lim_{m\to\infty} P\left(\bigcup_{\substack{A\in\pi_m(Y_1^m)\\ diam(A)>\gamma}} A \right) = 0. \tag{17}$$

almost surely with respect to the process distribution of $Y_1, Y_2, \cdots$, $\forall \gamma > 0$. Then the measure (P and Q) of cells of our random data-dependent partition scheme $\{\pi_m(Y_1^m) : m \in \mathbb{N}\}$, with diameter greater than an arbitrary non-zero number tends to zero almost surely as m tends to infinity. At this point we use the characterization of the divergence as the supremum with respect to finite codings or partitions of $\mathbb{R}^d$ [2], i.e.,

$$D(P||Q) = \sup_{\pi \in \mathcal{Q}} D(P_\pi||Q_\pi), \tag{18}$$

with $\mathcal{Q}$ representing the set of finite measurable partitions of $\mathbb{R}^d$, and

$$D(P_\pi||Q_\pi) = \sum_{A \in \pi} P(A) \log \frac{P(A)}{Q(A)} < \infty \tag{19}$$

because $P \ll Q$. Consequently, $D(P_{\pi_m(y_1,..,y_m)}||Q_{\pi_m(y_1,..,y_m)}) \leq D(P||Q)$, $\forall (y_1,..,y_m) \in \mathbb{R}^{d \cdot m}$ and then

$$\limsup_{m \to \infty} D(P_{\pi_m(Y_1^m)}||Q_{\pi_m(Y_1^m)}) \leq D(P||Q), \tag{20}$$

almost surely. Hence, the proof reduces to showing that the sequence of measurable partitions $\{\pi_m(Y_1^m) : m \in \mathbb{N}\} \subset \mathcal{Q}$ is almost surely sufficient to approximate $D(P||Q)$. Let us consider an arbitrary $\epsilon > 0$. Then by definition we have that $\exists \pi(\epsilon/2) \in \mathcal{Q}$, such that

$$D(P_{\pi(\epsilon/2)}||Q_{\pi(\epsilon/2)}) > D(P||Q) - \epsilon/2. \tag{21}$$

The following approximation result will be used.

LEMMA 3: Let $\pi = \{A_1,..,A_r\} \in \mathcal{Q}$ be a finite measurable partition of $\mathbb{R}^d$. If the adaptive partition scheme $\Pi = \{\pi_1, \pi_2, \cdots\}$ satisfies (16) and (17), then $\forall \delta > 0$, $\forall m \in \mathbb{N}$, $\exists \pi_m^* = \{A_{m,1},..,A_{m,r}\} \subset \sigma(\pi_m(Y_1^m))$ a finite measurable partition sequence, such that

$$\limsup_{m \to \infty} \sup_{i=1,..,r} |P(A_i) - P(A_{m,i})| < \delta \tag{22}$$

$$\limsup_{m \to \infty} \sup_{i=1,..,r} |Q(A_i) - Q(A_{m,i})| < \delta \tag{23}$$

almost surely[3].

This result shows that we can approximate arbitrarily closely the probability distribution restricted to any finite measurable partition using the partition scheme Π, under the approximation condition stipulated in **d)**. In our context, this result can be applied to $\pi(\epsilon/2)$ in (21), $\forall \epsilon > 0$.

On the other hand, given that $|\pi(\epsilon/2)| = r < \infty$ and that $x \log x$ is continuous real function, it is not difficult to show that $D(P_{\pi(\epsilon/2)}||Q_{\pi(\epsilon/2)})$ is a continuous function with respect to the total variational distance in the product space of probabilities measures on $(\mathbb{R}^d, \sigma(\pi(\epsilon/2)))$ under some additional conditions. More precisely in our problem, for $\epsilon/2$, $\exists \delta_1 > 0$ and $\delta_2 > 0$, such that if, $\sup_{i=1,..,r} \left|P^1(A_i) - P^2(A_i)\right| < \delta_1$, $\sup_{i=1,..,r} \left|Q^1(A_i) - Q^2(A_i)\right| < \delta_2$, and $P^1 \ll Q^1$, $P^2 \ll Q^2$ then,

$$\left| D(P^1_{\pi(\epsilon/2)}||Q^1_{\pi(\epsilon/2)}) - D(P^2_{\pi(\epsilon/2)}||Q^2_{\pi(\epsilon/2)}) \right| < \epsilon/2. \tag{24}$$

This last result and a direct application of *Lemma* 3 show that $\exists \pi_m^* \subset \sigma(\pi_m(Y_1^m))$, $\forall m \in \mathbb{N}$, such that

$$\liminf_{m \to \infty} D(P_{\pi_m^*}||Q_{\pi_m^*}) > D(P_{\pi(\epsilon/2)}||Q_{\pi(\epsilon/2)}) - \epsilon/2, \tag{25}$$

with probability one. Finally, note that $D(P_{\pi_m(Y_m^1)}||Q_{\pi_m(Y_1^m)}) \geq D(P_{\pi_m^*}||Q_{\pi_m^*})$, because of the fact that by construction $\pi_m^* \subset \sigma(\pi_m(Y_m^1))$ and consequently $\pi_m(Y_m^1)$ is a refinement of π_m^*, for all $m \in \mathbb{N}$ [2], then we have that

$$\begin{aligned} \liminf_{m \to \infty} D(P_{\pi_m(Y_m^1)}||Q_{\pi_m(Y_1^m)}) &> D(P_{\pi(\epsilon/2)}||Q_{\pi(\epsilon/2)}) - \epsilon/2 \\ &> D(P||Q) - \epsilon, \end{aligned} \tag{26}$$

with probability one, where the last inequality is by (21). Given that ϵ can be chosen arbitrarily small, then $\liminf_{m \to \infty} D(P_{\pi_m(Y_m^1)}||Q_{\pi_m(Y_1^m)}) \geq D(P_{\pi(\epsilon/2)}||Q_{\pi(\epsilon/2)})$ almost surely and in conjunction with (20) the result is proved. ■

REMARK *1:* Perhaps not explicit in the statement of the theorem is the natural assumption that $X_1, X_2, ...$ and $Y_1, Y_2, ...$ need to be mutually independent random sequences. This is used when invoking the SLLN in (15) which is implicitly conditioned by the random partition $\pi_m(Y_1^m)$ and consequently by $Y_1, .., Y_m$.

Note that the result presented in *Theorem* 1 can be naturally extended when $X_1, .., X_m$ is a stationary ergodic source [10], [12]. The following result states this extension.

THEOREM 2: Let us consider the same problem setting and assumptions of *Theorem* 1. If we consider instead that the random sequence $X_1, .., X_m$ is stationary and ergodic then,

$$\lim_{m \to \infty} \lim_{n \to \infty} \hat{D}_{m,n}(P||Q) = D(P||Q) \tag{27}$$

with probability one.

Proof: The same arguments for proving *Theorem* 1 can be adopted, where the proofs of **Term 1** and **Term 3** remain the same — because those terms are independent of the process distribution of $X_1, X_2, \cdots$, and the proof of the **Term 2** can be adapted by a simple application of the *Ergodic theorem* [10], [12]. ■

III. Applications

This section is devoted to show how our general strongly consistency result particularizes for an emblematic case of *statistically equivalent blocks*. This result can be related to similar results presented by Lugosi *et al.* [6] for proving how data-dependent partition schemes are strongly consistent in the $\mathbf{L}_1$ sense for the density estimation problem.

A. Statistically Equivalent Data-Dependent Partitions

Let us consider the real line $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as the target measurable space with two probability measures P and Q satisfying the conditions of *Theorem* 1. In the context of the data-dependent partition scheme for divergence estimation presented in the previous section, we consider the l_m-spacing partition scheme originally considered by Wang *et al.* [5] for the problem of divergence estimation. More precisely, let $Y_1, .., Y_m$ be the i.i.d. realizations with marginal distribution Q. The order statistics $Y^{(1)}, Y^{(2)}, .., Y^{(m)}$ is defined as the permutation of $Y_1, .., Y_m$ such that $Y^{(1)} < Y^{(2)} < \cdots < Y^{(m)}$ — this permutation exists with probability one as Q is

[3]The proof is not provided for space constraints.

absolutely continuous with respect to the Lebesgue measure. Based on this sequence, the resulting l_m-spacing quantization is given by $\pi_m(Y_1^m) = \{I_i^m : i = 1, .., T_m\}$

$$= \left\{(-\infty, Y^{(l_m)}], (Y^{(l_m)}, Y^{(2l_m)}], .., (Y^{((T_m-1)l_m)}, \infty)\right\},$$

where $T_m = \lfloor m/l_m \rfloor$ under the non-trivial case where $m > l_m$. Note that under this construction every cell of $\pi_m(Y_1^m)$ has at least l_m samples from $Y_1, .., Y_m$. The following result presents the sufficient conditions that makes this particular data-dependent divergence estimator consistent.

THEOREM 3: Let P, Q be probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ absolutely continuous with respect to the Lebesgue measure and $D(P||Q) < \infty$. Let $X_1, .., X_n$ and $Y_1, .., Y_m$ be i.i.d. realizations of P and Q respectively. Under the l_m-spacing partition scheme, if $l_m \approx m^{0.5+l/2}$ for some $l \in (1/3, 1)$, then

$$\lim_{m\to\infty} \lim_{n\to\infty} \hat{D}_{m,n}(P||Q) = D(P||Q), \tag{28}$$

with probability one.

Proof: We just need to check that under the l_m-spacing partition scheme, the conditions **a)**, **b)**, **c)** and **d)** from *Theorem* 1 are satisfied. Without loss of generality let us consider an arbitrary $l \in (1/3, 1)$. The trivial case to check is **c)**, because by construction we can consider $k_m = l_m$, $\forall m \in \mathbb{N}$, then the hypothesis of this theorem implies it. Concerning **a)**, again by construction we have that $\mathcal{M}(\mathcal{A}_m) \leq m/l_m + 1$, then $m^{-l}\mathcal{M}(\mathcal{A}_m) \leq m^{1-l}/l_m + m^{-l}$. Given that $l_m \approx m^{0.5+l/2}$ and $l \in (1/3, 1)$ it follows that

$$\lim_{m\to\infty} m^{-l}\mathcal{M}(\mathcal{A}_m) = 0. \tag{29}$$

For condition **b)**, Lugosi *et al.* [6] show that $\Delta_m^*(\mathcal{A}_m) = \binom{T_m+m}{m}$, where using that $\log\binom{s}{t} \leq s \cdot h(t/s)$ [8], with $h(x) = -x\log(x) - (1-x)\log(1-x)$ for $x \in [0,1]$ — the binary entropy function [1], it follows that $\log(\Delta_m^*(\mathcal{A}_m)) \leq 2m \cdot h\left(\frac{1}{l_m}\right)$. Consequently we have that,

$$\begin{aligned} m^{-l}\log(\Delta_m^*(\mathcal{A}_m)) \ll -\frac{m^{1-l}}{l_m}\log(1/l_m) \\ - m^{1-l}(1-1/l_m)\log(1-1/l_m). \end{aligned} \tag{30}$$

The first term on the right hand side (RHS) of (30) behaves like $m^{0.5-3/2\cdot l} \cdot \log(l_m)$, where as long as the exponent of this expression is negative (equivalent to $l > 1/3$) this sequence tends to zero as m tends to infinity considering that $(l_m) \ll (m)$. The second term on the RHS of (30) behaves asymptotically like $-m^{1-l} \cdot \log(1-1/l_m)$ which is dominated by the sequence $\frac{m^{1-l}}{l_m} \cdot \frac{1}{1-1/l_m}$ (using $\log(x) \leq x - 1$, for all $x > 0$), which tends to zero because $l_m \approx m^{0.5+l/2}$ and $l > 1/3$. Consequently from (30) $\lim_{m\to\infty} m^{-l}\log(\Delta_m^*(\mathcal{A}_m)) = 0$. Finally concerning condition **d)**, Lugosi *et al.* [6] (*Theorem 4*) proved that it is sufficient to show that $\lim_{m\to\infty} \frac{l_m}{m} = 0$, which is true in our case considering that $l < 1$. ■

REMARK *2:* In the context of statistically equivalent blocks for the real line, a universal consistency result was presented by Wang, Kulkarni and Verdú [5] under the less restrictive sufficient conditions: $l_m \to \infty$ and $l_m/m \to \infty$.

APPENDIX I
PROOF OF *Lemma* 2

Proof: Let us first note that from **c)**, $\forall m \in \mathbb{N}$, $\forall A \in \pi_m(Y_1^m)$,

$$\frac{|Q_m(A) - Q(A)|}{Q_m(A)} \leq \frac{|Q_m(A) - Q(A)|}{k_m/m}, \tag{31}$$

then we will concentrate in proving that $\sup_{A\in\pi_m(Y_1^m)} \frac{|Q_m(A)-Q(A)|}{k_m/m}$ tends to zero almost surely as $m \to \infty$. From the *Borel-Cantelli lemma*, a sufficient condition is to prove that $\forall \epsilon > 0$,

$$\sum_{m\geq 0} \mathbb{P}\left(\sup_{A\in\pi_m(Y_1^m)} |Q_m(A) - Q(A)| > \epsilon \cdot k_m/m\right) < \infty, \tag{32}$$

where $\mathbb{P}$ denotes the process distribution of the empirical process $Y_1, Y_2, \cdots$. Let us consider an arbitrary $\epsilon > 0$. From *Lemma* 1, it follows directly that

$$\begin{aligned} &\mathbb{P}\left(\sup_{A\in\pi_m(Y_1^m)} |Q_m(A) - Q(A)| > \epsilon \cdot k_m/m\right) \\ &\leq 4\Delta_{2m}^*(\mathcal{A}_m)2^{\mathcal{M}(\mathcal{A}_m)}\exp^{-\frac{(\epsilon\cdot k_m)^2}{32\cdot m}}, \end{aligned} \tag{33}$$

where using conditions **a)**, **b)** and **c)** from *Theorem* 1 we get that,

$$\begin{aligned} &\lim_{m\to\infty} \frac{1}{m^l} \cdot \log\mathbb{P}\left(\sup_{A\in\pi_m(Y_1^m)} |Q_m(A) - Q(A)| > \epsilon \cdot k_m/m\right) \\ &\leq -\lim_{m\to\infty} \epsilon^2 \cdot \frac{k_m^2}{m^{1+l}} = -\epsilon \cdot C, \end{aligned} \tag{34}$$

for some $C > 0$. Consequently the term of the summation in (32) is dominated by the sequence $(\exp^{-\epsilon\cdot C\cdot m^l\cdot})_{m\in\mathbb{N}}$, where given that $\sum_{m\in\mathbb{N}} \exp^{-\epsilon\cdot C\cdot m^l} < \infty$ for any $l \in (0,1)$, the result is proved. ■

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Interscience, New York, 1991.
[2] R. M. Gray, *Entropy and Information Theory*. Springer - Verlag, New York, 1990.
[3] S. Kullback, *Information theory and Statistics*. New York: Wiley, 1958.
[4] F. den Hollander, *Large Deviations*. American Mathematical Society, 2000.
[5] Q. Wang, S. R. Kulkarni, and S. Verdu, “Divergence estimation of continuous distributions based on data-dependent partitions,” *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.
[6] G. Lugosi and A. 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.
[7] V. 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.
[8] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
[9] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. Springer - Verlag, New York, 2001.
[10] S. Varadhan, *Probability Theory*. American Mathematical Society, 2001.
[11] P. R. Halmos, *Measure Theory*. Van Nostrand, New York, 1950.
[12] L. Breiman, *Probability*. Addison-Wesley, 1968.