On Data-Driven Histogram-Based Estimation for Mutual Information

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— **The problem of mutual information (MI) estimation based on data-dependent partition is addressed in this work. Sufficient conditions are stipulated on a histogram-based construction to guarantee a strong consistent estimate for the MI. The practical implications of this result are in the specification of a range of design parameters for two data-driven histogram-based approaches — statistically equivalent blocks and tree-structure vector quantizations — to yield density-free strongly consistent estimates for the MI.**

I. Introduction

Consider two random variables X and Y taking values on finite dimensional Euclidean spaces, $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$ respectively, with a joint distribution denoted by $P_{X,Y}$ in $\mathbb{R}^d$ and $d = p + q$. In this case the *mutual information* (MI) between X and Y can be expressed by [1],

$$I(X;Y) = D(P_{X,Y}||P_X \times P_Y), \tag{1}$$

where $P_X \times P_Y$ is the probability measure on $\mathbb{R}^d$ induced by multiplication of the marginals of X and Y and $D(P||Q) = \int \log \frac{\partial P}{\partial Q}(x) \cdot \partial P(x)$. is the *Kullback-Leibler divergence* (KLD) [1]. MI specifies the level of statistical dependency between a pair of random variables [1], and it is fundamental to characterizing some of the most remarkable results in information theory. MI has been also adopted in other statistical decision contexts finding important applications as an indicator in feature extraction [2], in detection [3], in image registration and segmentation [4], and to characterize performance limits on pattern recognition [5], just to mention a brief spectrum of important applications.

The problem of MI estimation based on independent and identically distributed (i.i.d) realizations of (X, Y) becomes crucial as pointed out in many of these works. There is an extensive literature dealing with the related differential entropy estimation, see for example Beirlant *et al.* [6]. Most approaches are based on classical product-type quantization of the space belonging to the category of *histogram-based constructions*. These classical estimates are consistent [6] but they do not offer good approximation to the empirical distributions for small number of samples [7]. This issue translates into a significance estimation bias effect in the small sample regime (from hundreds to few thousands sample points). Motivated by that, Darbellay *et al.* [7] proposed a histogram-based estimate that partitions the space from the empirical data in a non-product way. This non-product scheme provides the flexibility of better approximating the underlying behavior of the empirical mass, however strong consistency for this setting remains an open problem [7].

In this work we present an alternative data-driven MI estimate motivated by recent work on data-driven histogram based KLD estimation [8], [9], [10], [11]. This scheme also allows an adaptive non-product partition of the space with the distinctive characteristic of using a stopping criterion based on the minimum number of sample points per quantization bin [11]. The contributions of the work are two folds: first, in the characterization of sufficient conditions to get a strongly consistent estimate for the MI, and second, the application of this result to two emblematic data-driven partition schemes. On both these constructions, we obtain a collection of density-free strongly consistent estimates for the MI. These results relate with our recent contribution on KLD estimation [10], [11]. However, the learning setting, formulation and histogram-based construction proposed here are different, and as a consequence, hitherto unexplored technical challenges are addressed.

II. Preliminaries

A. Data-Dependent Partitions

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.

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 [12]

$$\mathcal{A}_n = \left\{ \pi_n(x_1, .., x_n) : (x_1, .., x_n) \in \mathbb{R}^{d \cdot n} \right\}. \tag{2}$$

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

978-1-4244-7892-7/10/$26.00 ©2010 IEEE

B. Combinatorial Indicator of Complexity

Let $\mathcal{C} \subset \mathcal{B}(\mathbb{R}^d)$ be a collection of measurable events, and $x_1^n = (x_1, .., x_n)$ be a sequences of points in $\mathbb{R}^d$, then we can define [13],

$$\mathcal{S}(\mathcal{C}, x_1^n) = \left|\{\{x_1, x_2, .., x_n\} \cap A : A \in \mathcal{C}\}\right|, \qquad (3)$$

and the *scatter coefficient* of $\mathcal{C}$ by $S_n(\mathcal{C}) = \sup_{x_1^n \in \mathbb{R}\cdot\ltimes} \mathcal{S}(\mathcal{C}, x_1^n)$, an indicator of the richness of $\mathcal{C}$ to dichotomize sequence of points in the space, (by definition $S_n(\mathcal{C}) \leq 2^n$ [13]). This combinatorial notions were extended for collection of measurable partitions in [12]. More precisely, let $\mathcal{A}$ be a collection of measurable partitions for $\mathbb{R}^d$. The *maximum cell counts* of $\mathcal{A}$ is given by

$$\mathcal{M}(\mathcal{A}) = \sup_{\pi \in \mathcal{A}} |\pi|. \qquad (4)$$

In addition, let us consider a finite length sequence $x_1^n = (x_1, .., x_n) \in \mathbb{R}^{d \cdot n}$. 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 $\{\{x_1, .., x_n\} \cap A : A \in \pi\}$— as the number of possible partitions of $\{x_1, .., x_n\}$ induced by $\mathcal{A}$, and the *growth function* of $\mathcal{A}$ by [12],

$$\Delta_n^*(\mathcal{A}) = \max_{x_1^n \in \mathbb{R}^{d \cdot n}} \Delta(\mathcal{A}, x_1, .., x_n). \qquad (5)$$

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 $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_A(X_i)$, a probability measure defined on $(\mathbb{R}^d, \sigma(\pi_n(X_1, .., X_n)))^1$. Note that for estimating the probability distribution the i.i.d. samples are used twice: first for defining a sub-sigma field $\sigma(\pi_n(X_1, .., X_n)) \subset \mathcal{B}(\mathbb{R}^d)$ and then for characterizing the empirical distribution on it.

A fundamental statistical learning problem is being able to bound the deviation of the empirical distribution with respect to the probability for a collection of measurable events. The following result provides a general answer for this problem.

THEOREM 1: (Vapnik and Chervonenkis [13]) Let $\mathcal{C}$ be a collection of measurable events, then $\forall n \in \mathbb{N}$, $\forall \epsilon > 0$,

$$\mathbb{P}\left\{\sup_{A \in \mathcal{C}} |P_n(A) - P(A)| > \epsilon\right\} \leq \mathcal{S}_n(\mathcal{C}) \cdot \exp^{-\frac{n\epsilon^2}{8}}, \qquad (6)$$

where $\mathbb{P}$ refers to the process distribution of $X_1, X_2, \cdots$.

Lugosi and Nobel [12] proposed an extension of this inequality for a collection of measurable partitions.

LEMMA 1: (Lugosi and Nobel [12]) 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. Histogram-Based Construction

Let X and Y be random variables in $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively, with joint distribution $P_{X,Y}$ absolutely continuous with respect to the Lebesgue measure λ in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We consider a data-driven partition rule, denoted by $\pi_n(\cdot)$, that maps i.i.d. samples to the space of finite measurable partitions of $\mathbb{R}^d$. We impose the requirement that every bin induced by this partition rule has a *product form*, i.e., every measurable set $A \in \pi_n(X_n, Y_n)$ can be expressed as $A = A_1 \times A_2$, with A_1 and A_2 events on $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Then denoting by P_n the empirical joint distribution obtained from the data [14]^2, the proposed *mutual information estimate* is given by, $\hat{I}_n(X;Y) =$

$$\sum_{A \in \pi_n(X_1^n, Y_1^n)} P_n(A) \cdot \log \frac{P_n(A)}{P_n(A_1 \times \mathbb{R}^q) \cdot P_n(\mathbb{R}^p \times A_2)}, \qquad (7)$$

where $A_1 \times A_2$ denotes the product form of the event $A \in \pi_n(X_1^n, Y_1^n)$. It is important to note that the requirement that every bin in $\pi_n(X_n, Y_n)$ decomposes in a product form is strictly necessary for being able to estimate $P_{X,Y}$ as well as the reference measure $P_X \times P_Y$ just based on the i.i.d. realizations of the joint distribution $P_{X,Y}$. Also note that this product bin structure does not imply that the partition $\pi_n(X_n, Y_n)$ has a product form (ie., written as the cartesian product of quantizations of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively).

The next section addresses the problem of stipulating a set of sufficient conditions on Π that guarantee that $\hat{I}_n(X;Y)$ convergences to $I(X;Y)$ almost surely.

IV. Main Conssistency Result

Let us first introduce the following elements. For any partition rule $\pi_n(\cdot) \in \Pi$, we consider its product bin structure to define the following collection of measurable events,

$$\mathcal{C}_{[1-q]}(z_1^n) = \left\{\xi_{[1-q]}(A) : A \in \pi_n(z_1^n)\right\} \qquad (8)$$

$$\mathcal{C}_{[q+1-d]}(z_1^n) = \left\{\xi_{[q+1-d]}(A) : A \in \pi_n(z_1^n)\right\} \qquad (9)$$

with $\xi_{[1-q]}(A)$ denoting the set operator that returns the collection of projected elements of A in the range of coordinate dimensions $[1-q]$ ^{3}. Then, the following collections of events are associated to the partition rule $\pi_n(\cdot)$:

$$\mathcal{C}_{[1-q],n} = \bigcup_{z_1^n \in \mathbb{R}^{d \cdot n}} \mathcal{C}_{[1-q]}(z_1^n), \qquad (10)$$

$$\mathcal{C}_{[q+1-d],n} = \bigcup_{z_1^n \in \mathbb{R}^{d \cdot n}} \mathcal{C}_{[q+1-d]}(z_1^n). \qquad (11)$$

Considering sequences of non-negative real numbers, we say that $(a_n)_{n \in \mathbb{N}}$ dominates $(b_n)_{n \in \mathbb{N}}$, denoted by $(b_n) \preceq (a_n)$, if

[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.

[2] For every measurable set $A \subset \mathbb{R}^p$, $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_A(X_i, Y_i)$ where $\mathbb{I}_A(x)$ is the indicator function.

[3] By construction any set $A \in \pi_n(z_1^n)$ can be expressed by $A = A_1 \times A_2$, with $A_1 \in \mathbb{R}^q$ and $A_2 \in \mathbb{R}^p$, and consequently $\xi_{[1-q]}(A) = A_1$ and $\xi_{[q+1-d]}(A) = A_2$.

there exists $C > 0$ and $k \in \mathbb{N}$ such that $b_n \leq C \cdot a_n$ for all $n \geq k$. We say that $(b_n)_{n\in\mathbb{N}}$ and $(a_n)_{n\in\mathbb{N}}$ are asymptotically equivalent, denoted by $(b_n) \approx (a_n)$, if there exits $C > 0$ such that $\lim_{n\to\infty} \frac{a_n}{b_n} = C$.

THEOREM 2: Let us consider a partition scheme $\Pi = \{\pi_1(\cdot), \cdots\}$ with the mentioned product bin structure and driven by i.i.d. realizations $Z_1, Z_2, \cdots$ drawn from $P_{X,Y}$. If there exists $\tau \in (0,1)$ for which the following set of conditions are satisfied:

c.1 $\lim_{n\to\infty} \frac{1}{n^\tau} \log \mathcal{S}_n(\mathcal{C}_{[1-p],n}) = 0$, $\lim_{n\to\infty} \frac{1}{n^\tau} \log \mathcal{S}_n(\mathcal{C}_{[p+1-d],n}) = 0$,

c.2 $\lim_{n\to\infty} \frac{1}{n^\tau} \log \Delta^*_n(\mathcal{A}_n) = 0$,

c.3 $\lim_{n\to\infty} \frac{1}{n^\tau} \mathcal{M}(\mathcal{A}_n) = 0$,

c.4 $\exists\ (k_n)_{n\in\mathbb{N}}$ a sequence on non-negative numbers, with $(k_n) \approx (n^{0.5+\tau/2})$, such that, $\forall n > 0$ and $(z_1, .., z_n) \in \mathbb{R}^{d\cdot n}$,

$$\inf_{A\in\pi(z_1^n)} P_n(A) \geq \frac{k_n}{n},$$

c.5: $\forall\epsilon > 0$,

$$\lim_{n\to\infty} P_{X,Y}\left(\left\{z \in \mathbb{R}^d : diam(\pi_n(z|Z_1^n)) > \epsilon\right\}\right) \to 0,$$

$\mathbb{P}$-almost surely,

then, the estimate in (7) satisfied that

$$\lim_{n\to\infty} \hat{I}_n(X;Y) = I(X,Y) \tag{12}$$

a.s. with respect to the process distribution of $Z_1, Z_2, \cdots$.

Proof: We consider the divergence notation for the MI in (1). Then $I(X;Y) = D(P||Q)$ with P denoting the joint distribution $P_{X,Y}$ and $Q = P_X \times P_Y$. We denote by P_n and Q_n the empirical versions of P and Q, respectively, obtained from a realization of the empirical process $Z_1, .., Z_n$ and the product bin structure of $\pi_n(\cdot)$. Then the empirical MI estimate in (7) can be expressed by $D_{\pi_n(Z_1^n)}(P_n||Q_n) \equiv \sum_{A\in\pi_n(Z_1^n)} P_n(A) \cdot \log \frac{P_n(A)}{Q_n(A)}$. To prove the result we use the following inequality,

$$\begin{aligned}&\left|D_{\pi_n(Z_1^n)}(P_n||Q_n) - D(P||Q)\right| \leq \\ &\left|D_{\pi_n(Z_1^n)}(P_n||Q_n) - D_{\pi_n(Z_1^n)}(P||Q)\right| \\ &+ \left|D_{\pi_n(Z_1^n)}(P||Q) - D(P||Q)\right|.\end{aligned} \tag{13}$$

The last term in the right hand side of (13) is the approximation error. From the condition **c.5** this error converges to zero $\mathbb{P}$-a.s. as n tends to infinity (the argument is a simple extension of our results on divergence estimation [15]). Then we just need to focus on the estimation error term. From triangular inequality and its definition $\left|D_{\pi_n(Z_1^n)}(P_n||Q_n) - D_{\pi_n(Z_1^n)}(P||Q)\right| \leq$

$$\left|\sum_{A\in\pi_n(Z_1^n)} [P_n(A)\log P_n(A) - P(A)\log P(A)]\right| \tag{14}$$

$$+\left|\sum_{A\in\pi_n(Z_1^n)} [P_n(A)\log Q_n(A) - P(A)\log Q(A)]\right|. \tag{15}$$

Concerning the term in (14), it is upper bounded by $\left|\sum_{A\in\pi_n(Z_1^n)} [P_n(A) - P(A)] \log P_n(A)\right| + \left|\sum_{A\in\pi_n(Z_1^n)} [\log P_n(A) - \log P(A)]\, P(A)\right| \leq$

$$\sum_{A\in\pi_n(Z_1^n)} |P_n(A) - P(A)| \log \frac{n}{k_n} + \sup_{A\in\pi_n(Z_1^n)} |\log P(A) - \log P_n(A)|, \tag{16}$$

where this last inequality uses the fact that $P_n(A) \geq \frac{k_n}{n}$ $\forall A \in \pi_n(Z_1^n)$. Using Lugosi and Nobel inequality, Lemma 1, the probability of the first term in (16) greater than ϵ can be uniformly (density-free) bounded by,

$$\begin{aligned}&\mathbb{P}^n\left(\sum_{A\in\pi_n(Z_1^n)} |P_n(A) - P(A)| \cdot \log\frac{n}{k_n} > \epsilon\right) \\ &\leq \mathbb{P}^n\left(\sup_{\pi\in\mathcal{A}_n} \sum_{A\in\pi} |P_n(A) - P(A)| \cdot > \frac{\epsilon}{\log n/k_n}\right) \\ &\leq 4\Delta^*_{2n}(\mathcal{A}_n) 2^{\mathcal{M}(\mathcal{A}_n)} \cdot \exp\left\{-\frac{n\epsilon^2}{(\log n/k_n)^2 \cdot 32}\right\},\end{aligned} \tag{17}$$

where the exponential term $\exp\left\{-\frac{n\epsilon^2}{(\log n/k_n)^2\cdot 32}\right\} \leq \exp\left\{-\frac{n\epsilon^2}{(\log n)^2\cdot 32}\right\}$. Note that this last sequence is uniformly, in ϵ, dominated by the sequence $(\exp\{-m^{\bar{\tau}}\})_{n\in\mathbb{N}}$, $\forall\bar{\tau} \in (0,1)$. Consequently from **c.2** and **c.3**, it is simple to show that $\forall\epsilon$,

$$\limsup_{n\to\infty} \frac{1}{m^\tau} \cdot \log \mathbb{P}^n\left(\sum_{A\in\pi_n(Z_1^n)} |P_n(A) - P(A)| > \frac{\epsilon}{\log n/k_n}\right)$$

$\leq C_o$, being C_o a strictly negative constant. Finally from the fact that $\sum_{n\geq 0} \exp\{C_o \cdot m^\tau\} < \infty$ and the Borel-Cantelli Lemma, we have that $\lim_{n\to\infty} \sum_{A\in\pi_n(Z_1^n)} |P_n(A) - P(A)| \log\frac{n}{k_n} = 0$, $\mathbb{P}$-a.s. Concerning the left term in (16), $\sup_{A\in\pi_n(Z_1^n)} |\log P(A) - \log P_n(A)|$, we use the following result.

PROPOSITION *1:* (Silva and Narayanan [15]) If $\lim_{n\to\infty} \sup_{A\in\pi_n(Z_1^n)} \left|\frac{P(A)}{P_n(A)} - 1\right| = 0$, $\mathbb{P}$-a.s, then,

$$\lim_{n\to\infty} \sup_{A\in\pi_n(Z_1^n)} |\log P(A) - \log P_n(A)| = 0,\ \mathbb{P} - a.s.$$

Silva *et al.* [15] prove that under **c.2**, **c.3** and **c.4**, the sufficient condition of Proposition 1 is satisfied and consequently from (16) the term in (14) tends to zero $\mathbb{P}$-a.s.

Concerning the term in (15), we bounded it by the expression in (18), where considering the product bin structure of $\pi_n(\cdot)$, we have that $\forall A \in \pi_n(Z_1^n)$, $Q_n(A) = P_n(A_{[1-p]} \times \mathbb{R}^q) P_n(\mathbb{R}^p \times A_{[p+1-d]})$, with $A_{[1-p]}$ and $A_{[p+1d]}$ a short-hand notation for $\xi_{[1-p]}(A)$ and $\xi_{[p+1-d]}(A)$, respectively. Given the symmetric structure of the bound in (18), we focus the attention on just one of these terms, since the derivation for the other use the same arguments. Using similar derivations as those used in the set of inequalities

$$\left|\sum_{A\in\pi_n(Z_1^n)}\left[P_n(A)\log Q_n(A) - P(A)\log Q(A)\right]\right| \leq \left|\sum_{A\in\pi_n(Z_1^n)}\left[P(A)\log P(A_{[1-p]}\times\mathbb{R}^q) - P_n(A)\log P_n(A_{[1-p]}\times\mathbb{R}^q)\right]\right| + \left|\sum_{A\in\pi_n(Z_1^n)}\left[P(A)\log P(\mathbb{R}^p\times A_{[p+1-d]}) - P_n(A)\log P_n(\mathbb{R}^p\times A_{[p+1-d]})\right]\right| \quad (18)$$

(16), we have that $\left|\sum_{A\in\pi_n(Z_1^n)}\left[P(A)\log P(A_{[1-p]}\times\mathbb{R}^q) - P_n(A)\log P_n(A_{[1-p]}\times\mathbb{R}^q)\right]\right| \leq$

$$\sum_{A\in\pi_n(Z_1^n)}|P_n(A)-P(A)|\log\frac{n}{k_n} + \sup_{A\in\pi_n(Z_1^n)}\left|\log P(A_{[1-p]}\times\mathbb{R}^q) - \log P_n(A_{[1-p]}\times\mathbb{R}^q)\right|, \quad (19)$$

where we have already proved that the first term tends to zero $\mathbb{P}$-a.s as n tends to infinity. Concerning the second term in (19), from Proposition 1 it is sufficient to prove that $\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1-p]}\times\mathbb{R}^q)}{P_n(A_{[1-p]}\times\mathbb{R}^q)}-1\right| = 0$ $\mathbb{P}$-a.s. Analyzing this expression, we have that, $\forall\epsilon>0$,

$$\begin{aligned}
&\mathbb{P}^n\left(\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1-p]}\times\mathbb{R}^q)}{P_n(A_{[1-p]}\times\mathbb{R}^q)}-1\right|>\epsilon\right)\\
&\leq \mathbb{P}^n\left(\sup_{A\in\mathcal{C}_{[1-p],n}}\left|\frac{P(A_{[1-p]}\times\mathbb{R}^q)}{P_n(A_{[1-p]}\times\mathbb{R}^q)}-1\right|>\epsilon\right)\\
&\leq \mathbb{P}^n\left(\sup_{A\in\mathcal{C}_{[1-p],n}}\left|P(A_{[1-p]}\times\mathbb{R}^q)P_n(A_{[1-p]}\times\mathbb{R}^q)\right|>\frac{k_n\cdot\epsilon}{n}\right)\\
&\leq \mathcal{S}_n(\mathcal{C}_{[1-p],n})\cdot\exp\left\{-\frac{k_n^2\cdot\epsilon^2}{n\cdot 8}\right\}, \qquad (20)
\end{aligned}$$

the first inequality results from the fact that $\pi_n(Z_1^n)\subset\mathcal{C}_{[1-p],n}$, the second from $P_n(A_{[1-p]}\times\mathbb{R}^d)\geq P_n(A)\geq\frac{k_n}{n}$, $\forall A\in\pi_n(Z_1^n)$, and the last one from the distribution free version of the Vapnik-Chervonenkis inequality, Theorem 1. Finally from the fact that $(k_n)\approx(n^{0.5+\tau/2})$ and the condition **c.1**, it is simple algebra to show that,

$$\limsup_{n\to\infty}\frac{1}{n^\tau}\log\mathbb{P}^n\left(\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1-p]}\times\mathbb{R}^q)}{P_n(A_{[1-p]}\times\mathbb{R}^q)}-1\right|>\epsilon\right)$$

$<C(\epsilon)$ a constant function of ϵ that is strictly negative. Then from this and the Borel-Cantelli lemma, we have that $\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1-p]}\times\mathbb{R}^q)}{P_n(A_{[1-p]}\times\mathbb{R}^q)}-1\right| = 0$ $\mathbb{P}$-a.s, which is the last piece of result needed to prove the theorem. ■

REMARK *1:* Observing the domain of values stipulated for τ, we can see that these conditions are stronger than the one obtained for the problem of consistent density estimation in the L_1 sense [12]. In simple words, these stronger conditions has to do with the unbounded behavior of the $\log(\cdot)$ function in the neighborhood of zero — the function is not absolutely continuous in $(0,\infty)$.

Next we address the applicability of this results, in terms of how this result translates into specific design conditions when working with some specific data-dependent constructions.

V. Applications

A. Statistically Equivalent Blocks

For simplicity let us denote by $Z_1^n = (Z_1,..,Z_n)$ the i.i.d. joint samples. Following Gessaman's approach [12], this partition rule sequentially splits every coordinate of $\mathbb{R}^d$ using axis-parallel hyperplanes. More precisely, let $l_n>0$ denote the number of samples points that we ideally want to have in every bin of $\pi_n(Z_1^n)$, and let us choose a particular sequential order for the axis-coordinates, without loss of generality the standard $(1,..,d)$. With that, $T_n=\lfloor(n/l_n)^{1/d}\rfloor$ is the number of partitions to create in every coordinate. Then the inductive construction goes as follows: first, project $Z_1,..,Z_n$ into the first coordinate, which we denote by $S_1,..,S_n$. Compute the order statistics $S^{(1)},S^{(2)},..,S^{(n)}$ or the permutation of $S_1,..,S_n$ such that $S^{(1)}<S^{(2)}<\cdots<S^{(n)}$. Based on this, define the following intervals to partition the first scalar coordinate,

$$\{I_i: i=1,..,T_n\} = \left\{(-\infty,S^{(s_n)}],(S^{(s_n)},S^{(2\cdot s_n)}],..,(S^{((T_n-1)\cdot s_m)},\infty)\right\},$$

where $s_n=\lfloor n/T_n\rfloor$. Then assigning the samples of $Z_1,..,Z_n$ to the different resulting bins, i.e., $\left\{I_i\times\mathbb{R}^{d-1}: i=1,..,T_n\right\}$, we can conduct the same process in each of those bins by projecting its data into the second coordinate. Iterating this approach until the last coordinate we get the data-dependent partition $\pi_n(Z_1^n)$.

The following result can be stated whose proof reduces to checking the sufficient condition stated in *Theorem* 2.

THEOREM 3: Under the problem setting of Section III, if $(l_n)\approx(n^{0.5+\tau/2})$ for some $\tau\in(1/3,1)$. the Gessaman's partition scheme provides a density-free strongly consistent estimate for the mutual information.

Proof: Let us consider an arbitrary joint distribution $P_{X,Y}$, equipped with a density function, and $\tau\in(1/3,1)$. The trivial case to check is **c.4)**, because by construction we can consider $k_n=l_n$, $\forall n\in\mathbb{N}$, and then the hypothesis of the theorem gives the result. For **c.1)**, from the construction of $\pi_n(\cdot)$, $\mathcal{C}_{[1-p],n}$ and $\mathcal{C}_{[p+1-d],n}$ are contained in the collection of all rectangles of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, which are well known to have finite VC dimensions, which suffices to get the result [16]. Concerning **c.3)**, again by construction we have that $\mathcal{M}(\mathcal{A}_n) \leq n/l_n + 1$, then $n^{-l}\mathcal{M}(\mathcal{A}_n) \leq n^{1-\tau}/l_n + n^{-\tau}$. Given that $(l_n) \approx (n^{0.5+\tau/2})$ and $\tau \in (1/3, 1)$ it follows that,

$$\lim_{n\to\infty} n^{-\tau}\mathcal{M}(\mathcal{A}_n) = 0. \tag{21}$$

For **c.2)**, Lugosi *et al.* [12] showed that $\Delta_n^*(\mathcal{A}_n) \leq \binom{T_n+n}{n}^d$, where using that $\log\binom{s}{t} \leq s \cdot h(t/s)$ [14], with $h(x) = -x\log(x) - (1-x)\log(1-x)$ for $x \in [0,1]$ — the binary entropy function [1] and defining $\bar{T}_n \equiv \lfloor n/l_n \rfloor \geq T_n$, it follows from the arguments in [12] that,

$$n^{-\tau}\log\left(\Delta_n^*(\mathcal{A}_n)\right) \leq 2dn^{1-\tau} \cdot h\left(\frac{1}{l_n}\right) \tag{22}$$

and consequently, $\forall n \in \mathbb{N}$,

$$\begin{aligned} n^{-\tau}\log(\Delta_n^*(\mathcal{A}_n)) \leq & -\frac{2dn^{1-\tau}}{l_n}\log(1/l_n) \\ & - 2dn^{1-\tau}(1-1/l_n)\log(1-1/l_n). \end{aligned} \tag{23}$$

The first term on the right hand side (RHS) of (23) behaves like $n^{0.5-3/2\cdot\tau} \cdot \log(l_n)$, where as long as the exponent of the first term is negative (equivalent to $\tau > 1/3$) this sequence tends to zero as m tends to infinity — considering that by construction $(l_n) \preceq (n)$. The second term on the RHS of (23) behaves asymptotically like $-n^{1-\tau} \cdot \log(1 - 1/l_n)$ which is upper bounded by the sequence $\frac{n^{1-\tau}}{l_n} \cdot \frac{1}{1-1/l_n}$ — using that $\log(x) \leq x - 1$, $\forall x > 0$. This upper bound tends to zero because $(l_n) \approx (n^{0.5+\tau/2})$ and $\tau > 1/3$. Consequently from (23), $\lim_{n\to\infty} n^{-\tau}\log(\Delta_n^*(\mathcal{A}_n)) = 0$.

Finally concerning **c.5)**, Lugosi *et al.* [12] (*Theorem 4*) proved that to get this shrinking cell condition is sufficient to show that $\lim_{n\to\infty} \frac{l_n}{n} = 0$, which is the case considering that $\tau < 1$. ■

B. Tree-structured Data-Dependent Partition

Here we consider a version of a balanced search tree [14]. Let $(Z_1, .., Z_n)$ be the i.i.d. realizations of the joint distribution. This scheme choses a dimension of the space in a sequential order as the previous construction, say the dimension i for the first step, and then the i axis-parallel halfspace by

$$H_i(Z_1^n) = \left\{x \in \mathbb{R}^d : x(i) \leq Z^{(\lceil n/2 \rceil)}(i)\right\}, \tag{24}$$

where $Z^{(1)}(i) < Z^{(2)}(i) <, .., < Z^{(n)}(i)$ denotes the order statistics of the sample points $\{Z_1, .., Z_n\}$ projected in the target dimension i. Using this hyper-plane, $\mathbb{R}^d$ is divided into two statistically equivalent rectangles with respect to the coordinate dimension i, denoted by $U_{(1,0)}$ and $U_{(1,1)}$. Reallocating the sample points in $U_{(1,0)}$ and $U_{(1,1)}$, we can choose a new dimension in the mentioned sequential order and continue in an inductive fashion with this splitting process. The termination criterion is based on a stopping rule that guarantees a minimum number of sample points per cell, denoted by $k_n > 0$.

Importantly the way the data is split in terms of measurable rectangles has a binary-tree indexed structure [14], where in the iteration k of the algorithm (assuming that the stopping rule is not violated) the intermediate rectangles $U_{(k-1,l)}$ for $l \in \left\{0, .., 2^{k-1} - 1\right\}$ are partitioned in terms of their respective statistically equivalent k-axis parallel hyper-planes to create $\left\{U_{(k,2l)}, U_{(k,2l+1)} : l = 0, .., 2^{k-1} - 1\right\}$. Equivalently to Theorem 3, the following result can be stated.

THEOREM 4: If $(k_n) \approx (n^{0.5+\tau/2})$ for some $\tau \in (1/3, 1)$, $\hat{I}_n(X : Y)$ induced from the tree-structured partition rule and (7) is density-free strongly consistent.
The proof reduces to checking the conditions of *Theorem* 2. The argument can be found in [17].

VI. Acknowledgment

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

References

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

[2] Jorge Silva and Shrikanth Narayanan, "Discriminative wavelet packet filter bank selection for pattern recognition," *IEEE Transaction on Signal Processing*, vol. 57, no. 5, pp. 1796–1810, 2009.

[3] M. L. Cooper and M. I. Miller, "Information measures for object recognition accommodating signature variability," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1896–1907, August 2000.

[4] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2083–2099, December 2000.

[5] M. B. Westover and Joseph A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 299–320, January 2008.

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

[7] Georges A. Darbellay and Igor Vajda, "Estimation of the information by an adaptive partition of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[8] 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.

[9] Q. Wang, Sanjeev R. Kulkarni, and Sergio Verdú, "Universal estimation of information measures for analog universal estimation of information measures for analog sources," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 3, pp. 265–353, 2009.

[10] Jorge Silva and Shrikanth Narayanan, "Histogram-based estimation for the divergence revisited," in *IEEE International Symposium on Information Theory*, 2009.

[11] Jorge Silva and Shrikanth Narayanan, "Information divergence estimation based on data-dependent partitions," *ELSEVIER Journal of Statistical Planning and Inference*, vol. in Press, 2010.

[12] 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.

[13] 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.

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

[15] Jorge Silva and Shrikanth Narayanan, "Universal consistency of data-driven partitions for divergence estimation," in *IEEE International Symposium on Information Theory*, June 2007.

[16] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer - Verlag, New York, 2001.

[17] Jorge Silva, *On Optimal Signal Representation for Statistical Learning and Pattern Recognition*, Ph.D. thesis, University of Southern California, http://digitallibrary.usc.edu/assetserver/controller/item/etd-Silva-2450.pdf, December 2008.