A Near-Optimal (Minimax) Tree-Structured Partition for Mutual Information Estimation

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*— A novel histogram-based mutual information estimator using data-driven tree-structured partitions (TSP) is presented in this work. The TSP is the solution of a complexity regularized empirical information maximization (EIM) criterion, with the objective to find a good tradeoff between the known estimation and approximation errors. We show that this solution is density-free strongly consistent and, furthermore, it provides a near-optimal balance between the mentioned variance-bias errors.**

I. Introduction

Let X and Y be two random vectors taking values in $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$, respectively, with a joint distribution $P_{X,Y}$ defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (with $d = p+q$), and $\mathcal{B}(\mathbb{R}^d)$ denoting the *Borel sigma field*. 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 distribution on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ induced by multiplication of the marginals of X and Y and $D(P||Q)$ denotes the *Kullback-Leibler divergence* (KLD) [2], [1],

$$D(P||Q) = \int \log \frac{\partial P}{\partial Q}(x) \cdot \partial P(x). \tag{2}$$

$I(X;Y)$ is an indicator of the level of statistical dependency between X and Y, i.e., how $P_{X,Y}$ differs from $P_X \times P_Y$ in the KLD sense [2], [1], and has a fundamental role in information theory and statistics [1], [2]. This role justifies its large adoption in statistical learning applications [3], [4]. A crucial need for these applications is to have a distribution-free estimate of $I(X,Y)$, based on independent and identically distributed (iid) realizations of (X,Y), that converges to $I(X;Y)$ (almost surely) as the number of sample tends to infinity (strong consistency) [5]. The problem has been systematically addressed for distributions defined on a finite dimensional Euclidean space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where consistency is well known for histogram-based and kernel plug-in estimates (see the survey by Beirlant *et al.* [5]). For the case of histogram-based estimation, these results usually consider non-adaptive and product type of partitions of the space. In this setting, every coordinate of the space is partitioned independently to form the full partition of $\mathbb{R}^d$ (a product partition). In contrast, non-product data-driven partitions [6], [7] can approximate the nature of the empirical distribution better with few quantization bins and provide the flexibility to improve the approximation quality of histogram-based estimates [6], [7].

In addressing this problem, Darbellay and Vajda [7] proposed an histogram-based approach based on a non-product adaptive tree-structured partitions (TSP), where the inductive nature of TSP was used to dynamically increase the resolution of the quantization in areas of the space that provide higher empirical MI gains. This adaptive TSP estimate shows promising empirical evidence, although ensuring strong consistency remains an open problem [7]. Alternatively, Wang *et al.* [8], [9] and more recently Silva *et al.* [10], [11], [12] studied the role of a more general family of data-driven partitions, based on *partition schemes* [6], [13]. The work presented in this paper builds upon this formulation, where the learning and adaptation advantages of TSP schemes are further explored [14], [13], [15], [16]. In particular, we investigate a complexity-regularized type of learning principle [14], previously unexplored in this inference problem. Here we stipulate conditions under which the estimation and the approximation errors vanish asymptotically, and more importantly from a learning perspective, conditions that offer, with an arbitrary high probability, an optimal balance between these two errors.

II. Problem Setting and Notation

We start with introducing the learning problem and some required notations. Let $Z_1^n = Z_1, .., Z_n$ be iid realizations of (X,Y) drawn from $P_{X,Y}$. Let $\Pi = \{\pi_n(\cdot) : n \in \mathbb{N}\}$ be a *partition scheme* where $\pi_n(\cdot)$ is a function from $\mathbb{R}^{d \cdot n}$ (the sequences of length n in $\mathbb{R}^d$) to $\mathcal{Q}$ (the collection of finite alphabet measurable partitions of $\mathbb{R}^d$) that we call the *partition rule of length* n. $\pi_n(\cdot)$ receives the empirical data Z_1^n and creates a partition of the space, i.e., $\pi_n(Z_1^n) \in \mathcal{Q}$. In addition, Π needs to satisfy a *product bin condition*, i.e., $\forall z_1^n = (z_1, .., z_n) \in \mathbb{R}^{d \cdot n}$ every event $A \in \pi_n(z_1^n)$ is expressed by [7], $A = A_1 \times A_2$, where $A_1 \in \mathcal{B}(\mathbb{R}^p)$ and $A_2 \in \mathcal{B}(\mathbb{R}^q)$. With this, the learning-estimation process involves three phases: first, to use the empirical data to partition $\mathbb{R}^d$ by $\pi_n(Z_1^n)$, second, to use again the data to estimate $P_{X,Y}$ and $P_X \times P_Y$ restricted to the sigma field $\sigma(\pi_n(Z_1^n))$ [1], and finally, to consider the plug-in technique to get an empirical MI estimate

[1] Given a collection of sets $\mathcal{A}$, we denote by $\sigma(\mathcal{A})$ the smallest sigma field that contains $\mathcal{A}$ [17]. When $\mathcal{A}$ is a finite partition, $\sigma(\mathcal{A})$ is the collection of elements written as unions of element of $\mathcal{A}$.

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

on $(\mathbb{R}^d, \sigma(\pi_n(Z_1^n)))$ [12]. Concerning the phase 2, the product bin condition is needed to estimate $P_{X,Y}$ as well as the reference measure $P_X \times P_Y$ only based on the iid realizations of the joint distribution $P_{X,Y}$ [7], [12]. More precisely, let P denote the joint distribution and P_n its empirical version, i.e., $P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_A(Z_i)$, $\forall A \in \mathcal{B}(\mathbb{R}^d)$, hence, the MI estimate is given by $\hat{I}_n(\pi_n(Z_1^n)) =$

$$\sum_{A \in \pi_n(Z_1^n)} P_n(A) \cdot \log \frac{P_n(A)}{P_n(A_1 \times \mathbb{R}^q) \cdot P_n(\mathbb{R}^p \times A_2)}, \quad (3)$$

where $A_1 \times A_2$ denotes the product form of the event A.

A. Binary-Trees and Tree-Structured Partitions

Adopting Breiman *et al.* [14] conventions, a *binary tree* T is a collection of nodes: one node of degree 2 (the *root*), and the remaining nodes of degree 3 (*internal* nodes) or degree 1 (*leaf* or *terminal* nodes). Let $\mathcal{I}(T)$ and $\mathcal{L}(T)$ be the set of internal and terminal nodes of T, respectively, and $|T|$ be the *size* of a tree T, given by the cardinality of $\mathcal{L}(T)$. If $\bar{T} \subset T$ and $\bar{T}$ is a binary tree by itself, we say that $\bar{T}$ is a *subtree* of T and moreover, if both have the same root we say that $\bar{T}$ is a *pruned* version of T, denoted by $\bar{T} \ll T$.

A *tree-structured partition* (TSP) can be represented by a pair $(T, \tau(\cdot))$ [16], with T a binary tree and $\tau(\cdot)$ a function from T to $\mathcal{H}$, with $\mathcal{H}$ denoting the collection of closed halfspaces of the form $H = \left\{x : x^\dagger w \geq \alpha\right\}$, for some $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$. Then for any $t \in \mathcal{I}(T)$, $\tau(t)$ corresponds to the closed halfspace that dichotomizes the cell associated with t, denoted by U_t, in $U_{r(t)} = U_t \cap \tau(t)$ and $U_{l(t)} = U_t \cap \tau(t)^c$, which are the cells associated with the left and right child of t, denoted by $r(t)$ and $l(t)$, respectively. Then initializing the cell of the root node t_0 by $U_{t_0} = \mathbb{R}^d$, $\tau(\cdot) : \mathcal{I}(T) \to \mathcal{H}$ provides a way to characterize U_t, $\forall t \in T$. In particular,

$$\pi_T \equiv \{U_t : t \in \mathcal{L}(T)\} \subset \mathcal{B}(\mathbb{R}^d), \quad (4)$$

is the TSP induced by $(T, \tau(\cdot))$. Note that if $\bar{T} \ll T$ then π_T is a refinement of $\pi_{\bar{T}}$, that we denote consistently by $\pi_{\bar{T}} \ll \pi_T$. For the sake of simplicity, we will use the binary tree notation T to refer to both $(T, \tau(\cdot))$ and more frequently π_T.

Finally, a n-sample TSP rule $T_n(\cdot)$ is a function from the space of finite sequences $\mathbb{R}^{d \cdot n}$ to the space of TSP with halfspace splitting rules, and the resulting TSP partition scheme is the collection of TSP rules, i.e., $\Pi = \{T_1, T_2, \cdots\}$.

III. The Tree-Structured Partition Scheme

Our TSP scheme uses Z_1^n to construct a partition of $\mathbb{R}^d$ in two consecutive stages: a growing phase and a pruning phase.

For the growing stage, let t_o be the root of the tree and $U_{to} = \mathbb{R}^d$. Considering $Z_1^n = (Z_1, .., Z_n)$ as the iid realizations of (X, Y), this scheme choses a dimension of the space in a sequential order, let say the dimension i for the first step, and then the i axis-parallel halfspace by

$$\tau(t_o) = H_i(Z_1^n) = \left\{x \in \mathbb{R}^d : x(i) \leq Z^{(\lceil n/2 \rceil)}(i)\right\}, \quad (5)$$

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 $H_i(Z_1^n)$, $\mathbb{R}^d$ is divided into two statistically equivalent rectangles with respect to the coordinate dimension i, denoted by $U_{l(t_o)}$ and $U_{r(t_o)}$. Reallocating the sample points in $U_{l(t_o)}$ and $U_{r(t_o)}$, respectively, we can choose a new dimension in the mentioned sequential order and continue in an inductive fashion with this splitting process. As the stopping rule, we propose a criterion that finishes the refinement when a *minimum number of sample points per cell*, threshold denoted by $k_n \in \mathbb{N} \setminus \{0\}$, is reached (or violated). Hence at the end, we get a full-tree, denoted by $T_{b_n}^{full}(Z_1^n)$, and the associated partition $\pi_{T_{b_n}^{full}}(Z_1^n)$, where we guarantee a minimum magnitude for P_n on the events of $\sigma(\pi_{T_{b_n}^{full}}(Z_1^n))$ that we denote by $b_n = k_n/n \in (0, 1)$ for all $n > 0$. This full TSP is designed to have in general few points per quantization cell, where the deviation of P_n with respect to P on these events is expected to be large (estimation error). This motivates the second stage of pruning detailed next.

A. Complexity-Penalized Empirical Information Maximization

For the rest, the full tree will be denoted by $T_{b_n}^{full}$ considering implicit its dependency on Z_1^n. First, we consider the following inequality: $\forall T \ll T_{b_n}^{full}$, $\left|\hat{I}_n(\pi_T(Z_1^n)) - I(X;Y)\right| \leq$

$$\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| + I(X;Y) - \pi_T(Z_1^n), \quad (6)$$

where $I(\pi_T(Z_1^n)) \equiv \sum_{A \in \pi_T(Z_1^n)} P(A) \cdot \log \frac{P(A)}{P(A_1 \times \mathbb{R}^q) \cdot P(\mathbb{R}^p \times A_2)}$, is the KLD of the true distributions restricted to the sigma field induced by $\pi_T(Z_1^n)$ [1]. The first term of (6) characterizes the estimation error, or the difference in the MI functional between the adoption of the empirical and real measures. The second term of (6) is non-negative and corresponds to the approximation error, which is a consequence of the fact that quantization reduces the magnitude of information theoretic quantities [1]. Motivated by the well understood tradeoff between the estimation and approximation errors [13], we propose the following complexity-penalized empirical information maximization criterion,

$$\hat{T}^n = \arg \min_{T \ll T_{b_n}^{full}} -\hat{I}_n(\pi_T(Z_1^n)) + \phi_n(T). \quad (7)$$

This regularization criterion attempts to find an optimal balance in $\left\{T : T \ll T_{b_n}^{full}\right\}$ between the empirical MI (fidelity) and an indicator of complexity for π_T that we denote by $\phi_n(T)$. $\phi_n(T)$ is designed to reflect the estimation error $\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right|$ in (6). However, as the true distribution is unknown, we consider the standard approach of characterizing distribution-free expressions to upper bound this quantity [16], [15]. The next section elaborates on this idea by considering the *Vapnik-Chervonenkis inequality* [13].

IV. Concentration Inequality for Trees

Let $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$ be two sequences of non-negative real numbers. (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 $b_n \leq C \cdot a_n$, $\forall n \geq k$. $(b_n)_{n\in\mathbb{N}}$ and $(a_n)_{n\in\mathbb{N}}$ (both strictly positive) are asymptotically equivalent, denoted by $(b_n) \approx (a_n)$, if there exists $C > 0$ such that $\lim_{n\to\infty} \frac{a_n}{b_n} = C$, Finally, (b_n) is $o(a_n)$ (for $(a_n)_{n\in\mathbb{N}}$ strictly positive) if $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$.

THEOREM 1: Let P be a probability measure in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $Z_1, Z_2, \cdots$ be iid realizations driven by P. Let $T_{b_n}^{full}$ be the full TSP of the growing phase where $(b_n)_{n\in\mathbb{N}}$ is the critical empirical mass sequence. In addition, let $\mathcal{G}_{b_n}^k \equiv \left\{T \ll T_{b_n}^{full} : |T| = k\right\}$ be the family of pruned TSPs of size k induced from $T_{b_n}^{full}$. Then, $\forall k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}$, $\forall n > 0$, $\forall \epsilon \in (0,3)$, $\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} \left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| > \epsilon\right) \leq$

$$(n+1)^{2d}\left[\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{6}\right)^2\right\} + 2\cdot\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{12}\right)^2\right\}\right]$$
$$+ 4\cdot\left(2^{d+1}\cdot n^d\right)^k\cdot\exp\left\{-\frac{n}{32}\cdot\left(\frac{\log(1/b_n)^{-1}\cdot\epsilon}{9}\right)^2\right\}, \tag{8}$$

where $\mathbb{P}$ refers to the process distribution of $Z_1, Z_2, \cdots$.

Note that this bound is distribution free, valid for any finite n, and exclusively function of the size of the tree, the dimension of the space and the critical empirical mass sequence $(b_n)_{n\in\mathbb{N}}$ of our TSP construction. Concerning the deviation variable ϵ, this concentration inequality is only valid for a finite range of small values, which, however, is sufficient to obtain all the relevant forthcoming results. Rewriting Theorem 1, we could quantify the deviation of $\hat{I}_n(\pi_T(Z_1^n))$ with respect to $I(\pi_T(Z_1^n))$ in terms of an interval of confidence and with that obtain a distribution-free expression for the estimation error.

COROLLARY *1:* Under the setting of Theorem 1, if $(b_n) \approx (n^{-l})$ for some $l \in (0, \frac{1}{3})$, then $\forall \delta > 0$, $\forall k \in \mathbb{N}$, there exists $N(\delta, k) > 0$, such that $\forall n > N(\delta, k)$, with probability at least $1-\delta$, $\sup_{T\in\mathcal{G}_{b_n}^k} \left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| <$

$$\frac{12}{b_n}\cdot\sqrt{\frac{8}{n}\cdot(\ln(8/\delta) + k\cdot[(d+1)\cdot\ln(2) + d\cdot\ln(n)])}. \tag{9}$$

It is important to mention that this result is valid for a large sampling regime ($\forall n > N(\delta, k)$) to ensure that $\epsilon_c(n, b_n, \delta, k) \in (0,3)$, domain where Theorem 1 is valid (see Section VIII-C for details). For the rest of the exposition, we denote the interval of confidence on the RHS of (9) by $\epsilon_c(n, b_n, d, \delta, k)$.

V. Minimax Oracle Result

Returning to our central problem in (7), we propose the following expression for the penalization term, $\forall n > 0$, $\forall T \ll T_{b_n}^{full}$,

$$\phi_n(|T|) = \epsilon_c\left(n, b_n, d, \delta_n\cdot b_n, |T|\right), \tag{10}$$

for a sequence $(\delta_n)_{n\in\mathbb{N}}$ of confidence probabilities in $(0,1]$ such that (δ_n) is $o(1)$. Loosely speaking, the motivation of this choice is justified by the concentration results presented in Section IV, but substantiated rigorously from the oracle result presented in the next theorem. Let $\tilde{I}_n(\pi_T(Z_1^n)) \equiv \hat{I}_n(\pi_T(Z_1^n)) - \phi_n(|T|)$ be the penalized EMI indicator $\forall T \ll T_{b_n}^{full}$. The next result shows that $\hat{T}^n$ offers a near-optimal solution for the estimation of $I(X;Y)$.

THEOREM 2: Under the setting of Theorem 1, if

- $(b_n) \approx (n^{-l})$ for some $l \in (0, 1/3)$,
- (δ_n) is $o(1)$ and $(1/\delta_n)$ is $O(e^{n^{1/3}})$,

then $\forall \delta > 0$ there exists $N_c(\delta) > 0$, such that $\forall n > N_c(\delta)$ with probability $1-\delta$ (with respect to $\mathbb{P}$),

$$0 \leq I(X;Y) - \tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n)) \leq \min_{T\ll T_{b_n}^{full}} \left[I(X;Y) - I(\pi_T(Z_1^n))\right] + 2\cdot\phi_n(|T|). \tag{11}$$

The result says two important things. On one hand, it shows that with an arbitrary high probability our penalized indicator $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n))$ is an underestimation of $I(\pi_{\hat{T}^n}(Z_1^n))$, which ratifies the correctness of the penalization term in (10). On the other hand, and more importantly, it shows that the deviation of the penalized quantity $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n)$ from $I(X;Y)$ is upper bounded by an expression that reflects the optimal balance between the estimation error and the true approximation error, right hand side (RHS) of (11). Alternatively, we can see the RHS of (11) as an *oracle minimax error bound* in the sense that it is the choice, T^n, of an ideal observed that has access to the true distribution to balance the two errors of this learning problem, i.e., $T^n = \arg\min_{T\ll T_{b_n}^{full}} I(X;Y) - I(\pi_T(Z_1^n)) + 2\phi_n(|T|)$. Note that T^n offers the best performance for the worse scenario, where the two errors add constructively (the minimax solution).

From the conditions on $(b_n)_{n\in\mathbb{N}}$ stated in Theorem 2, we have that $\lim_{n\to\infty} \sup_{k\in\left\{1,..,\left|T_{b_n}^{full}\right|\right\}} \phi_n(k) = 0$ (the arguments presented in Section VIII-C). Consequently the oracle minimax error bound in (11) is governed by the asymptotic trend of $\lim_{n\to\infty}\left[I(X;Y) - I(\pi_{T_{b_n}^{full}}(Z_1^n))\right]$, associated with the approximation goodness (or the asymptotic sufficiency) of the full tree. The next result formalizes this idea and proves the *density free* strong consistency of $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n))$ and $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n))$, respectively.

VI. Density-Free Strong Consistency

THEOREM 3: Under the setting of Theorem 1, if P is absolutely continuous with respect to the *Lebesque* measure in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and (b_n) and (δ_n) satisfy the condition stated in Theorem 2, then $\lim_{n\to\infty} \hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) = I(X;Y)$ and $\lim_{n\to\infty} \tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n)) = I(X;Y)$, $\mathbb{P}$-almost surely.

VII. Final Remark

The conditions on $(b_n)_{n\in\mathbb{N}}$ to ensure that $\hat{T}^n$ induces strongly consistent estimates for $I(X;Y)$ (Theorem 3), match the one stipulated on the full tree, i.e., $T_{b_n}^{full}$, to obtain that $\hat{I}_n(\pi_{T_{b_n}^{full}}(Z_1^n))$ is strongly consistent. This last result presented by the authors in a companion manuscript [12]. At this point, it is important to highlight the adaptation character of our TSP, which a function of the data finds the tree's topology that offers a near-optimal estimation-approximation error tradeoff (Theorem 2). To illustrate the idea, if the target value $I(X;Y)$ is high we expect to get a less conservative (or bigger) complexity regularized tree $\hat{T}^n$, than in the case of a moderate MI magnitude. In contrast, the full tree solution does not allow for this tree structure adaptation to the problem.

VIII. Proofs

A. Theorem 1

LEMMA 1: (Lugosi and Nobel [6]) Let $\mathcal{G}^k$ be the family of tree-structure measurable partitions of $\mathbb{R}^d$ with k cells (or terminal nodes), and $Z_1, Z_2, \cdots$ iid realizations with distribution P in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then, $\forall \epsilon > 0$, $\forall n$,

$$\mathbb{P}\left(\sup_{\pi \in \mathcal{G}^k} \sum_{A \in \pi} |P_n(A) - P(A)| > \epsilon\right) \leq 4(2^{d+1} n^d)^k \exp\left\{\frac{-n\epsilon^2}{32}\right\}$$

LEMMA 2: (Vapnik and Chervonenkis [13]) Under the setting of Lemma 1, if we instead consider $\mathcal{B}$ the family of measurable rectangleof $\mathbb{R}^d$, then, $\forall \epsilon > 0$, $\forall n$,

$$\mathbb{P}\left(\sup_{A \in \mathcal{B}} |P_n(A) - P(A)| > \epsilon\right) \leq (n+1)^{2d} \exp\left\{-\frac{n\epsilon^2}{8}\right\}.$$

Proof Theorem 1: We use that, $\forall T \in \mathcal{G}^k_{b_n}$, $\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| \leq \sum_{A \in \pi_T(Z_1^n)} |P_n(A) - P(A)| \cdot 3 \cdot \log(1/b_n) +$

$$\sup_{A \in \pi_T(Z_1^n)} |\log P(A) - \log P_n(A)| + \sup_{A \in \pi_T(Z_1^n)} |\log Q(A) - \log Q_n(A)|, \qquad (12)$$

this derived from the triangular inequality and the critical mass criterion of the full tree $T^{full}_{b_n}$. Concerning the first term on the RHS of (12), $\mathbb{P}\left(\sup_{T \in \mathcal{G}^k_{b_n}} \sum_{A \in \pi_T(Z_1^n)} |P_n(A) - P(A)| \cdot 3 \cdot \log(1/b_n) > \epsilon\right)$

$$\leq 4 \cdot (2^{d+1} \cdot n^d)^k \exp\left\{-\frac{n}{32} \cdot \left(\frac{\log(1/b_n)^{-1} \cdot \epsilon}{3}\right)^2\right\}, \qquad (13)$$

from Lemma 1 and the fact that $\mathcal{G}^k_{b_n} \subset \mathcal{G}^k$. Concerning the second term on the RHS of (12), for an arbitrary $A \in \mathcal{B}(\mathbb{R}^d)$ let us consider the following collection of sequences $\mathcal{S}_A = \left\{z_1^n \in \mathbb{R}^{d \cdot n} : |\log P(A) - \log P_n(A)| > \epsilon\right\}$. This can be written as $\mathcal{S}_A = \{z_1^n : P(A) - P_n(A) > P_n(A) \cdot (e^\epsilon - 1)\} \cup \{z_1^n : P_n(A) - P(A) > P_n(A) \cdot (1 - e^{-\epsilon})\}$. Using Taylor expansion, $\forall \epsilon \in (0,1)$, $\max\{e^\epsilon - 1, 1 - e^{-\epsilon}\} > \frac{\epsilon}{2}$, then $\forall \epsilon \in (0,1)$, $\forall n \in \mathbb{N}$,

$$\mathbb{P}\left(\left\{z_1^n : \sup_{T \in \mathcal{G}^k_{b_n}} \sup_{A \in \pi_T} |\log P(A) - \log P_n(A)| > \epsilon\right\}\right) \leq$$

$$\mathbb{P}\left(\bigcup_{T \in \mathcal{G}^k_{b_n}} \bigcup_{A \in \pi_T} \left\{z_1^n : |P_n(A) - P(A)| > P_n(A) \cdot \frac{\epsilon}{2}\right\}\right) \leq$$

$$\mathbb{P}\left(\left\{z_1^n : \sup_{A \in \mathcal{B}} |P_n(A) - P(A)| > b_n \cdot \frac{\epsilon}{2}\right\}\right) \leq$$

$$(n+1)^{2d} \cdot \exp\left\{-\frac{n}{8}\left(\frac{b_n \cdot \epsilon}{2}\right)^2\right\}, \qquad (14)$$

where the last two inequalities are obtained from the fact that $\forall T \ll T^{full}_{b_n}$ the cells of π_T are rectangles in $\mathcal{B}$, and Lemma 2, respectively. Concerning the last term in the RHS of (12), by construction of $T^{full}_{b_n}$, $\forall T \ll T^{full}_{b_n}$, $\forall A \in \pi_T(Z_1^n)$, A has a product form , $A_1 \times A_2$, and by construction $Q(A) = P(A_1 \times \mathbb{R}^q) \cdot P(\mathbb{R}^p \times A_2)$. Hence, $\sup_{A \in \pi_T(Z_1^n)} |\log Q(A) - \log Q_n(A)| \leq$

$$\sup_{A \in \pi_T(Z_1^n)} |\log P(A_1 \times \mathbb{R}^q) - \log P_n(A_1 \times \mathbb{R}^q)| + \sup_{A \in \pi_T(Z_1^n)} |\log P(\mathbb{R}^p \times A_2) - \log P_n(\mathbb{R}^p \times A_2)|. \qquad (15)$$

From the same inequalities shown in (14),

$$\mathbb{P}\left(\sup_{T \in \mathcal{G}^k_{b_n}} \sup_{A \in \pi_T} |\log P(A_1 \times \mathbb{R}^q) - \log P_n(A_1 \times \mathbb{R}^q)| > \frac{\epsilon}{2}\right)$$

$$\leq (n+1)^{2d} \cdot \exp\left\{-\frac{n}{8}\left(\frac{b_n \cdot \epsilon}{4}\right)^2\right\}. \qquad (16)$$

The same bound in (16) is obtained for the term

$$\mathbb{P}\left(\sup_{T \in \mathcal{G}^k_{b_n}} \sup_{A \in \pi_T} |\log P(\mathbb{R}^p \times A_2) - \log P_n(\mathbb{R}^p \times A_2)| > \frac{\epsilon}{2}\right),$$

and from (15), $\forall \epsilon \in (0,2)$, $\mathbb{P}\left(\sup_{T \in \mathcal{G}^k_{b_n}} \sup_{A \in \pi_T} |\log Q(A) - \log Q_n(A)| > \epsilon\right) \leq$

$$2 \cdot (n+1)^{2d} \cdot \exp\left\{-\frac{n}{8}\left(\frac{b_n \cdot \epsilon}{4}\right)^2\right\}. \qquad (17)$$

To conclude, considering the inequality in (12) and the distribution free bounds obtained for its RHS terms (in (13), (14) and (17), respectively), we obtain (8) $\forall \epsilon \in (0,3)$. ■

B. Corollary 1

The result derives directly from Theorem 1, and it is not reported here for the space constraint.

C. Theorem 2

By definition $\phi_n(k) = \frac{12}{b_n}\sqrt{\frac{8}{n}} \cdot \sqrt{(\ln(8) + \ln(n) - \ln(\delta_n \cdot b_n) + k \cdot [(d+1) \cdot \ln(2) + d \cdot \ln(n)])}$, then considering that $\left|T^{full}_{b_n}\right| \leq (1/b_n)$, it is simple to check that $(b_n) \approx (n^{-l})$ with $l \in (0, 1/3)$ and $(1/\delta_n)$ being $O(e^{n^{1/3}})$ are the weakest set of sufficient conditions to obtain that

$$\lim_{n \to \infty} \sup_{k \in \left\{1, \cdots \left|T^{full}_{b_n}\right|\right\}} \phi_n(k) = \lim_{n \to \infty} \phi_n(\left|T^{full}_{b_n}\right|) = 0. \qquad (18)$$

This is crucial for the rest of the proof, as the inequality in Theorem 1 is valid only for $\epsilon \in (0,3)$, represented in this case by the intervals of deviations $\phi_n(k)$, $\forall k \in \left\{1, \cdots \left|T^{full}_{b_n}\right|\right\}$. Let $\mathcal{S}^{n,k} \equiv$

$$\left\{z_1^n \in \mathbb{R}^{d \cdot n} : \sup_{T \in \mathcal{G}^k_{b_n}} \left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| \leq \phi_n(k)\right\},$$

be the k-typical set, well defined for all n such that $k \leq \left|T^{full}_{b_n}\right|$. From Corollary 1, if $\phi_n(k) \in (0,3)$, then $\mathbb{P}(\mathcal{S}^{n,k}) > 1 - b_n\delta_n$. Consequently from (18), there exists $N_c > 0$ such that $\forall k \in \left\{1, \cdots, \left|T^{full}_{b_n}\right|\right\}$ and $\forall n > N_c$, $\mathbb{P}(\mathcal{S}^{n,k}) > 1 - b_n\delta_n$. Hence, defining $\mathcal{S}^n = \bigcap_{k\in\left\{1,\cdots\left|T^{full}_{b_n}\right|\right\}} \mathcal{S}^{n,k}$, we have that $\mathbb{P}(\mathcal{S}^n) > 1 - \delta_n$, $\forall n > N_c$. By definition, if $z_1^n \in \mathcal{S}^n$, then $\sup_{T\in\mathcal{G}^k_{b_n}} \left|\hat{I}_n(\pi_T(z_1^n)) - I(\pi_T(z_1^n))\right| \leq \phi_n(k)$, $\forall k \in \left\{1, \cdots \left|T^{full}_{b_n}\right|\right\}$, which also implies that [13],

$$\left|\sup_{T\in\mathcal{G}^k_{b_n}} \hat{I}_n(\pi_T(z_1^n)) - \sup_{T\in\mathcal{G}^k_{b_n}} I(\pi_T(z_1^n))\right| \leq \phi_n(k), \qquad (19)$$

$\forall k \in \left\{1, \cdots \left|T^{full}_{b_n}\right|\right\}$. Then for an arbitrary $z_1^n \in \mathcal{S}^n$

$$\begin{aligned} -\tilde{I}_n(\pi_{\hat{T}^n}(z_1^n)) &= -\hat{I}_n(\pi_{\hat{T}^n}(z_1^n)) + \phi_n(\left|\hat{T}^n\right|) \\ &\leq -\hat{I}_n(\pi_{\hat{T}^n_k}(z_1^n)) + \phi_n(k), \\ &\leq -I(\pi_{T^n_k}(z_1^n)) + 2 \cdot \phi_n(k),\ \forall k \in \left\{1, \cdots \left|T^{full}_{b_n}\right|\right\}, \end{aligned}$$

where $T^n_k \equiv \arg\max_{T\in\mathcal{G}^k_{b_n}} I(\pi_T(z_1^n))$ is the oracle solution that maximizes the MI on $\mathcal{G}^k_{b_n}$. Also it is clear that $\forall z_1^n \in \mathcal{S}^n$, $\tilde{I}_n(\pi_{\hat{T}^n}(z_1^n)) = \hat{I}_n(\pi_{\hat{T}^n}(z_1^n)) - \phi_n(\left|\hat{T}^n\right|) \leq I(\pi_{\hat{T}^n}(z_1^n)) \leq I(X;Y)$, and consequently we have that, $0 \leq I(X;Y) - \tilde{I}_n(\pi_{\hat{T}^n}(z_1^n)) \leq$

$$\min_{k\in\left\{1,\cdots\left|T^{full}_{b_n}\right|\right\}} (I(X;Y) - I(\pi_{T^n_k}(z_1^n))) + \phi_n(k). \qquad (20)$$

The argument concludes from the fact that $\mathcal{S}^n$ has probability at least $1 - \delta_n$, $\forall n > N_c$ and that (δ_n) is $o(1)$. ■

D. Sketch of the Proof of Theorem 3

We consider the following results, whose proofs are omitted for the space constraint.

PROPOSITION *1:* Under the setting of Theorem 3, if $(b_n) \approx (n^{-l})$ for some $l \in (0, 1/3)$, then

$$\lim_{n\to\infty} I(\pi_{T^{full}_{b_n}}(Z_1^n)) = I(X;Y),$$

$\mathbb{P}$-almost surely. (The argument is presented in [12].)

PROPOSITION *2:* Under the setting of Theorem 3, if $(b_n) \approx (n^{-l})$ with $l \in (0, 1/3)$, then

$$\lim_{n\to\infty} \left|\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n))\right| = 0 \qquad (21)$$

$\mathbb{P}$-almost surely.

LEMMA 3: Under the setting of Theorem 3, if $(b_n) \approx o(n^{-l})$ with $l \in (0, 1/3)$, (δ_n) is o(1) and $(1/\delta_n)$ is $O(e^{n^{1/3}})$, then $\forall \epsilon > 0$ there exits $N_c(\epsilon)$ such that $\forall n > N_c$ and $\forall k \in \left\{1, .., \left|T^{full}_{b_n}\right|\right\}$, $\mathbb{P}\left(\sup_{T\in\mathcal{G}^k_{b_n}} I(\pi_T(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) > \epsilon\right) \leq$

$$\exp\left\{-\frac{n}{8}\left(\frac{\epsilon b_n}{24}\right)^2\right\} + 8(2^{d+1}n^d)^k \exp\left\{-\frac{n}{8}\left(\frac{\epsilon b_n}{48}\right)^2\right\},$$

and consequently $\mathbb{P}$-almost everywhere,

$$\lim_{n\to\infty} I(\pi_{T^{full}_{b_n}}(Z_1^n)) = I(\pi_{\hat{T}^n}(Z_1^n)). \qquad (22)$$

Proof of Theorem 3: The proof comes from

$$\left|I(X;Y) - \hat{I}_n(\pi_{\hat{T}^n}(Z_1^N))\right| \leq I(X;Y) - I(\pi_{T^{full}_{b_n}}(Z_1^N)) +$$
$$I(\pi_{T^{full}_{b_n}}(Z_1^N)) - I(\pi_{\hat{T}^n}(Z_1^N)) + \left|I(\pi_{\hat{T}^n}(Z_1^N)) - \hat{I}_n(\pi_{\hat{T}^n}(Z_1^N))\right|,$$

where these RHS terms tend to zero $\mathbb{P}$-almost surely from Proposition 1, Lemma 3 and Proposition 2, respectively. Finally, the same result is obtained for the regularized estimate $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^N))$ as, by definition, $\lim_{n\to\infty}\left|\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^N)) - \hat{I}_n(\pi_{\hat{T}^n}(Z_1^N))\right| \leq \lim_{n\to\infty} \sup_{k\in\left\{1,..,\left|T^{full}_{b_n}\right|\right\}} \phi_n(k) = 0.$ ■

IX. Acknowledgment

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

References

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

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

[3] 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.

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

[5] 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.

[6] 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.

[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 sources," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 3, pp. 265–353, 2009.

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

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

[12] Jorge Silva and Shrikanth Narayanan, "Non-product data-dependent partitions for mutual information estimation: Strong consistency and applications," *IEEE Transactions on Signal Processing*, in Press, 2010.

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

[14] Leo Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworth, 1984.

[15] Clayton Scott and Robert D. Nowak, "Minimax-optimal classification with dyadic decision trees," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1335–1353, April 2006.

[16] Andrew B. Nobel, "Analysis of a complexity-based pruning scheme for classification tree," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2362–2368, 2002.

[17] Leo Breiman, *Probability*, Addison-Wesley, 1968.