Complexity-Regularized Tree-Structured Partition for Mutual Information Estimation

Jorge Silva*, *Member, IEEE* and Shrikanth Narayanan, *Fellow, IEEE*

Abstract

A new histogram-based mutual information estimator using data-driven tree-structured partitions (TSP) is presented in this work. The derived TSP is a solution to a complexity regularized empirical information maximization (EIM), with the objective of finding a good tradeoff between the known estimation and approximation errors. A distribution-free concentration inequality for this tree-structured learning problem as well as finite sample performance bounds for the proposed histogram-based solution are derived. It is shown that this solution is density-free strongly consistent and, that it provides, with an arbitrary high probability, an optimal balance between the mentioned estimation and approximation errors. Finally for the emblematic scenario of independence, $I(X;Y) = 0$, it is shown that the TSP estimate converges to zero with $O(e^{-n^{1/3}+\log\log n})$.

Index Terms

Mutual information, histogram-based estimates, data-dependent partitions, tree-structured partitions, complexity regularization, strong consistency, Vapnik and Chervonenkis inequality, minimum cost tree pruning.

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 $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $d = p + q$) and $\mathcal{B}(\mathbb{R}^d)$ denotes the *Borel sigma field*. The mutual information (MI) between X and Y can be expressed by [1], [2],

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

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 (distribution where X and Y are independent), and $D(P||Q)$ denotes the *Kullback-Leibler divergence* (KLD) or *information divergence* [2], [3],

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

J. Silva is with the Department of Electrical Engineering, University of Chile, Av. Tupper 2007 Santiago, 412-3, Room 508, Chile, Tel: 56-2-9784090, Fax: 56-2-6953881, (email: josilva@ing.uchile.cl).

S. Narayanan is with the Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California. 3740 McClilntock Avenue, Room EEB430, Los Angeles, CA 90089 2564, USA, Tel: 213-740-6432, Fax: 213-740-4651 (email: shri@sipi.usc.edu)

$I(X;Y)$ 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], [3]. In fact, $I(X;Y) = 0$ is a necessary and sufficient condition for X and Y to be independent.

Mutual information (MI) has a fundamental role in information theory and statistics [1]–[3], which justifies its large adoption in statistical learning applications [4]–[11]. A particularly crucial need for these applications is to have a distribution-free estimate of $I(X,Y)$, based on independent and identically distributed (i.i.d.) realizations of (X,Y), as the distribution in these settings is unknown. An important requirement is that the estimate has to converge to $I(X;Y)$ as the number of sample points tend to infinity with probability one (strong consistency) [12]. In this learning context, the MI estimation scenario relates with the problem of distribution (density) estimation as MI is a functional of the joint distribution of (X,Y). In this classical problem strong consistency in the L_1 sense is well known [13], in particular for histogram-based estimates [13], [14]. More recent extensions on histogram-based estimator (the *Barron-type* of estimator [15]) has considered consistency under topologically stronger notions, such as consistency in direct information divergence by Barron *et al.* [15] and Györfi *et al.* [16], χ^2-divergence and expected χ^2-divergence by Györfi *et al.* [17] and Vajda *et al.* [18] and the general family of Csiszár's ϕ-divergence by Beirlant *et al.* [19].

In the context of estimating functionals of probability distributions, the differential entropy estimation has been systematically addressed for distributions defined on a finite dimensional Euclidean space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In particular, consistency results are well known for histogram-based and kernel plug-in estimates, see Beirlant *et al.* [12] and references therein. These constructions and results extend to the case of MI estimation, since MI can be expressed in terms of differences of differential entropies [2]. However, for the important case of histogram-based estimation, which is the focus of this work, 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), and the partition is made only of a function of the amount of data and independent of how the data is distributed in the space. In contrast, non-product data-driven partitions [20]–[23] 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 [20], [21], see Figure 1. This has been shown theoretically in a number of learning problems, including density estimation, regression and classification [20], [24], [25]. Their full potential, however, remains to be studied for the estimation of MI.

In addressing this problem, Darbellay *et al.* [21] 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. Alternatively, Wang *et al.* [22], [23] and more recently Silva *et al.* [26]–[29] have studied the role of a more general family of data-driven partitions based on *partition schemes* [20], [24] in the context of MI and the KLD estimation. For MI estimation, these results are summarized in the next section.

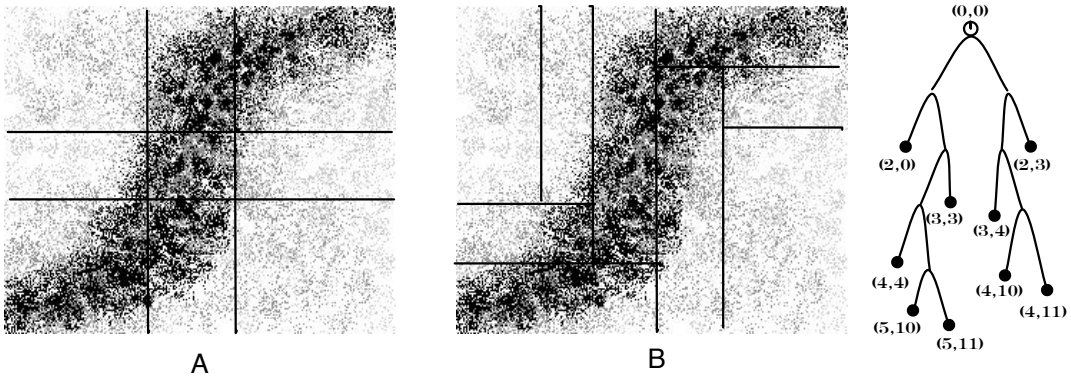

Fig. 1. **A:** A product type of partition of the space with 9 cells. Every coordinate is partitioned independently to form the full partition. **B:** Non-product data-driven tree-structured partition (TSP) with 9 cells and its binary tree representation. The space is partitioned by axis-parallel hyper-planes with a statistically equivalent splitting criterion. This splitting process is conducted inductively in a binary tree structured way, starting from the full space indexed by the node $(0, 0)$.

A. Histogram-Based MI Estimation based on Partition Schemes

For a *finite measurable partition* of $\mathbb{R}^d$, we mean a finite collection $\{A_1, .., A_k\}$ of elements in $\mathcal{B}(\mathbb{R}^d)$, such that $A_i \cap A_j = \emptyset$ if $i \neq j$ and $\bigcup_{l=1}^{k} A_l = \mathbb{R}^d$. Every element of $\{A_1, .., A_k\}$ is called a *cell*, or *partition cell*, and $|\{A_1, .., A_k\}|$ denotes its cardinality. In this context, $\mathcal{Q}$ denotes the collection of finite measurable partitions of $\mathbb{R}^d$. A *partition scheme* $\Pi = \{\pi_n(\cdot) : n \in \mathbb{N}\}$ [20] is collection of functions where, for each n, $\pi_n(\cdot)$ maps the elements in $\mathbb{R}^{d \cdot n}$ (the sequences of length n in $\mathbb{R}^d$) to $\mathcal{Q}$. In this context, we say that $\pi_n(\cdot)$ is a *partition rule of length* n [20] .

Let $Z_1^n = Z_1, .., Z_n$ be i.i.d. realizations of $Z = (X, Y)$ drawn from $P_{X,Y}$ and let us consider a partition scheme $\Pi = \{\pi_n(\cdot) : n \in \mathbb{N}\}$. In our learning scenario, $\pi_n(\cdot)$ receives the empirical data Z_1^n and creates a partition of the space $\pi_n(Z_1^n) \in \mathcal{Q}$. In addition, we impose that Π has a Cartesian product structure, in the sense that each element $A \in \pi_n(z_1^n)$ can be expressed by [21]

$$A = A_1 \times A_2, \tag{3}$$

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 on $(\mathbb{R}^d, \sigma(\pi_n(Z_1^n)))$ [28]. Concerning the phase 2, the *product bin condition* in (3) is required to estimate $P_{X,Y}$ as well as the reference measure $P_X \times P_Y$ only based on the i.i.d. realizations of the joint distribution $P_{X,Y}$ [21], [28]. More precisely, let P denote the joint distribution and P_n its empirical version, i.e.,

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_A(Z_i), \ \forall A \in \mathcal{B}(\mathbb{R}^d), \tag{4}$$

[1]Given a collection of sets $\mathcal{A}$, we denote by $\sigma(\mathcal{A})$ the smallest sigma field that contains $\mathcal{A}$ [30], [31]. When $\mathcal{A}$ is a finite partition, $\sigma(\mathcal{A})$ is the collection of elements written as unions of element of $\mathcal{A}$.

hence, the histogram-based MI estimate is given by

$$\hat{I}_n(\pi_n(Z_1^n)) = \sum_{A \in \pi_n(Z_1^n)} P_n(A) \cdot \log \frac{P_n(A)}{P_n(A_1 \times \mathbb{R}^q) \cdot P_n(\mathbb{R}^p \times A_2)}, \tag{5}$$

where $A_1 \times A_2$ denotes the product form of the event $A \in \pi_n(Z_1^n)$. Note that $\hat{I}_n(\pi_n(Z_1^n))$ can be interpreted as the KLD restricted to the sigma field $\sigma(\pi_n(Z_1^n))$, between the empirical joint distribution and its empirical product counterpart [21].

Silva *et al.* [28] particularized this construction to statistically equivalent blocks [32] and to a data-driven tree-structured partition (TSP), where conditions were shown for strong consistency. These conditions were derived from a theorem that stipulates sufficient conditions on Π to guarantee that the estimation and approximation error associated with $\hat{I}_n(\pi_n(Z_1^n))$ individually converge to zero almost-surely (a.s.). The work presented in this paper builds upon this formulation, where the learning and adaptation attributes of TSPs [24], [33]–[35] are further explored. In particular, we investigate a complexity-regularized type of learning principle [33] previously unexplored in this inference problem. With this learning criterion the idea is not only to obtain conditions under which the estimation and the approximation errors vanish asymptotically, but, with an arbitrarily high probability, provide an optimal balance between these two errors [34]–[36]. In this context, the partition is induced from Z_1^n in two stages. In the first, the space is partitioned in a binary tree-structured way using the idea of statistically equivalent splits [14], while in the second, the induced full tree is pruned back in order to find a good balance between the estimation and the approximation errors [33]–[35], [37].

Concerning the pruning stage, we address the problem of deciding the optimal TSP for the MI estimation as a complexity regularization problem. Here we have adopted the ideas of *structural risk minimization* (SRM) by Vapnik [38], [39], and *complexity regularized learning* by Barron [40], Barron and Cover [41] and others [42], where the learning principle is designed to obtain the optimal balance between an empirical fidelity indicator (the empirical MI) and a notion of learning complexity. For this last part, we have adopted the concentration inequalities by *Vapnik-Chervonenkis* [24], [39], [43], which offer closed forms for TSP [24], [35], [36]. Based on that, Theorem 1 derives an analytical expression for the penalization term (no resampling or cross validation is needed), which ends up being proportional to the square root of the size of the tree. Adopting this penalty in a complexity penalized criterion, Theorem 2 shows that the solution of this problem is able to find a nearly-optimal balance between the estimation and the approximation errors. As expected, Theorem 3 shows that our estimator is density-free strongly consistent, refining the results presented for TSP in [28]. Finally, for the important case when $I(X;Y) = 0$ (X and Y are independent), Theorem 4 shows that the proposed estimate is able to converge to zero at a rate faster than any finite polynomial order decay, i.e., $\mathbb{E}(\hat{I}_n(\pi_n(Z_1^n)))$ is $O(e^{-n^{1/3}+\log\log n})$ density-free. Concluding, we present two concrete algorithms to solve the main complexity regularization problem. These are derived from dynamic programing and offer polynomial time solutions with respect to the sampling length n.

The rest of the paper is organized as follows. Section II introduces the basic notations for TSP. Section III presents the complexity penalized tree learning formulation. Section V reports the minimax-oracle result and Section VI shows the conditions for density-free strong consistency. Section VIII presents some algorithmic solutions for the

optimal tree-pruning problem. Finally, Section IX provides concluding comments. Some of the proofs are presented in the Appendix section.

II. Binary-Trees and Tree-Structured Partitions

Let us first introduce some conventions and notations for binary trees to facilitate the description of the proposed TSP scheme. Adopting Breiman *et al.* [33] conventions, a *binary tree* T is a collection of nodes with 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 collection 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 $(T, \tau(\cdot))$ [35], 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 $t \in \mathcal{I}(T)$ (internal nodes), $\tau(t)$ corresponds to the closed halfspace that dichotomizes the cell associated with t, denoted by U_t, in two components $U_{r(t)} = U_t \cap \tau(t)$ and $U_{l(t)} = U_t \cap \tau(t)^c$. These resulting cells are associated with the left and right child of t, denoted by $r(t)$ and $l(t)$, respectively, in the case when $t \in \mathcal{I}(T)$. If we denote by t_0 the root node of T, then initializing the cell of 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, the partition indexed by T is denoted and constructed by

$$\pi_T \equiv \{U_t : t \in \mathcal{L}(T)\} \in \mathcal{Q}. \tag{6}$$

If $(T, \tau(\cdot))$ is a TSP and $\bar{T} \ll T$, then there is a unique TSP associated with $\bar{T}$ by restricting $\tau(\cdot)$ to the domain $\mathcal{I}(\bar{T})$. Note that if $\bar{T} \ll T$, π_T is a refinement of $\pi_{\bar{T}}$ by (6), 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, the partition 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., $\Pi = \{T_1, T_2, \cdots\}$. In the scope of this work, we focus on the family of TSP induced by *axis-parallel hyperplane* cuts [24], presented in Section III-A.

III. The Complexity Penalized Tree-Structured Partition Scheme

Adopting the ideas of classification and regression trees (CART) [33], we propose a scheme that uses the empirical data Z_1^n to construct a TSP of $\mathbb{R}^d$ in two consecutive stages that involve a growing and a pruning phase. In the growing phase, Z_1^n is used to iteratively split the space and create a large TSP that we denote by $T_n^{full}(Z_1^n)$. $T_n^{full}(Z_1^n)$ has, in general, few or no sample points of Z_1^n in each of its cells [33]. In this context, the deviation of P_n with respect to P on the measurable events of $T_n^{full}(Z_1^n)$, more precisely, $\sup_{A \in \pi_{T_n^{full}(Z_1^n)}} |P(A) - P_n(A)|$, is expected to be large. This motivates the second stage of pruning, where the idea is to prune-back $T_n^{full}(Z_1^n)$ in such a way that we find a good balance between a notion of estimation error (cost) and an approximation error (fidelity), both to be defined for our problem.

A. Statistically Equivalent Splitting Criterion

For the growing stage, we consider a modified version of what is known as *balanced search tree* [24, Chapter 20.3]. Here, the idea is to split the space in a non-product way (by axis-parallel hyper-planes) adopting a statistically equivalent splitting criterion. More precisely, 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 i.i.d. realizations of $Z = (X, Y)$, this scheme choses a coordinate axis of the space in a sequential order, let us say the dimension i for the first step, and then the i axis-parallel halfspace by

$$\tau(t_o) = H_i(Z_1^n) = \left\{ x \in \mathbb{R}^d : x(i) \leq Z^{(\lceil n/2 \rceil)}(i) \right\}, \tag{7}$$

where $Z^{(1)}(i) < Z^{(2)}(i) <, .., < Z^{(n)}(i)$ denotes the *order statistics* [14] obtained by a permutation of the sample points $\{Z_1, .., Z_n\}$ projected into the target dimension i. Note that this permutation exists with probability one as the i-marginal distribution of P has a density [20], [24]. Using $H_i(Z_1^n)$ in (7), $U_{to} = \mathbb{R}^d$ is divided into two rectangles $U_{l(t_o)} = U_{to} \cap H_i(Z_1^n)$ and $U_{r(t_o)} = U_{to} \cap H_i(Z_1^n)^c$, in the coordinate axis i. By construction $U_{l(t_o)}$ and $U_{r(t_o)}$ induces a partition of $\mathbb{R}^d$ with almost the same empirical mass, in fact $\left| P_n(U_{l(t_o)}) - P_n(U_{r(t_o)}) \right| \leq \frac{1}{n}$ and this quantity is zero when n is an even number. Assigning the sample points $\{Z_1, .., Z_n\}$ to their belonging cell in $\left\{ U_{l(t_o)}, U_{r(t_o)} \right\} \in \mathcal{Q}$, we can choose a new coordinate axis in the mentioned sequential order and continue with the aforementioned splitting process, independently in each of the two intermediate cells. Then we keep with this process in an inductive fashion, where as a stopping rule, we propose a criterion that finishes the refinement of the cells to guarantee a *critical number of sample points*, threshold denoted by $k_n \in \mathbb{N} \setminus \{0\}$, in each element of the resulting data-driven partition. Hence, for a given intermediate cell U_t we split U_t, by the aforementioned process, if $P_n(U_t) \geq 2k_n$, otherwise we stop the refinement process and t is, consequently, a leaf node of $T_n^{full}(Z_1^n)$. Finally, we get a full-tree $T_{b_n}^{full}(Z_1^n)$ and the associated partition $\pi_{T_{b_n}^{full}}(Z_1^n) \in \mathcal{Q}$, where a minimum magnitude for P_n on the events of $\sigma(\pi_{T_{b_n}^{full}}(Z_1^n))$ is guaranteed. More precisely,

$$P_n(A) \geq b_n, \quad \forall A \in \sigma(\pi_{T_{b_n}^{full}}(Z_1^n)), \tag{8}$$

with $b_n \equiv k_n/n \in (0, 1)$ for all $n > 0$. This stopping criterion was originally proposed by the authors for the problem of KLD estimation in [26], [27]. A graphical illustration of this process is presented in Fig. 1.

As pointed out by Darbellay *et al.* [21], this binary splitting criterion provides a non-product adaptive partition of the space with good approximation to the underlying structure of the data, Fig 1. On the other hand, the adopted *critical mass stopping criterion* is the key to derive concentration inequalities for our problem [26]–[28]. Based on this, we formulate pruning stage presented next. For the rest of the paper, the full tree will be denoted by $T_{b_n}^{full}$ considering implicit its dependency on Z_1^n.

B. Complexity-Penalized Empirical Information Maximization

In tree-structured learning [33]–[35], [44], the idea of the second stage is to prune the initial tree $T_{b_n}^{full}$ by a complexity regularized objective criterion that tries to balance the estimation and approximation errors (or the

variance-bias tradeoff). For our target problem, we consider the following inequality, $\forall T \ll T_{b_n}^{full}$,

$$\left|\hat{I}_n(\pi_T(Z_1^n)) - I(X;Y)\right| \leq \left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| + I(X;Y) - I(\pi_T(Z_1^n), \tag{9}$$

where

$$I(\pi_T(Z_1^n)) \equiv \sum_{A \in \pi_T(Z_1^n)} P(A) \cdot \log \frac{P(A)}{P(A_1 \times \mathbb{R}^q) \cdot P(\mathbb{R}^p \times A_2)}, \tag{10}$$

is the KLD of the true distributions restricted to the sigma field induced by $\pi_T(Z_1^n)$ [2]. The first term on the right hand side (RHS) of (9) characterizes the estimation error, or the difference in the MI functional between the adoption of the empirical and real measures. The second term on the RHS of (9) is non-negative and corresponds to the approximation error, which is a consequence of the well-known fact that quantization reduces the magnitude of the information divergence and consequently of the MI [1], [2].

Returning to the pruning problem, we propose the following complexity-penalized empirical information maximization criterion,

$$\hat{T}^n = \arg \min_{T \ll T_{b_n}^{full}} -\hat{I}_n(\pi_T(Z_1^n)) + \phi_n(T). \tag{11}$$

This regularization criterion attempts to find an optimal balance in $\left\{T : T \ll T_{b_n}^{full}\right\}$ between the empirical mutual information and an indicator of complexity for π_T that we denote by $\phi_n(T)$. The penalization term $\phi_n(T)$ has to reflect the estimation error $\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right|$ in (9). However, as the true distribution is unknown, we consider the approach used in classification trees of characterizing distribution-free expressions to upper bound this quantity [34]–[36]. The next section elaborates on this idea by considering the *Vapnik-Chervonenkis (VC) inequality* [24], [38], [39], [43].

IV. Concentration Results for Tree-Structured Partitions

Let us first introduce some terminologies. 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 i.i.d. realizations driven by P. Let $T_{b_n}^{full}$ be the TSP of Section III-A 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, $\forall k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}$,

$\forall n > 0$, $\forall \epsilon \in (0,3)$,

$$\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k}\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| > \epsilon\right) \leq$$
$$(n+1)^{2d}\cdot\left[\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{6}\right)^2\right\} + 2\cdot\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{12}\right)^2\right\}\right]$$
$$+ 4\cdot\left(2^{d+1}\cdot n^d\right)^k\cdot\exp\left\{-\frac{n}{32}\cdot\left(\frac{\log(1/b_n)^{-1}\cdot\epsilon}{9}\right)^2\right\}, \tag{12}$$

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

This is the main finite sample concentration inequality considered in our problem to characterize and bound the estimation error in (9). Note that the bound in (12) is distribution free and is exclusively a 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. It is important to note that this inequality is only valid for a finite range of values for the variable ϵ (details about this condition are in Section IV-A). However, this finite range is sufficient to obtain all the forthcoming results. The proof of this result is presented at the end of this section.

COROLLARY 1: Under the setting of Theorem 1, if $(b_n) \approx (n^{-l})$ for some $l \in (0, \frac{1}{2})$, then $\forall k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}$,

$$\lim_{n\to\infty}\sup_{T\in\mathcal{G}_{b_n}^k}\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| = 0 \tag{13}$$

$\mathbb{P}$-almost surely. Furthermore, if $(b_n) \approx (n^{-l})$ for some $l \in (0, \frac{1}{3})$, then with probability one with respect to $\mathbb{P}$,

$$\lim_{n\to\infty}\left|\hat{I}_n(\pi_{T_{b_n}^{full}}(Z_1^n)) - I(\pi_{T_{b_n}^{full}}(Z_1^n))\right| = 0. \tag{14}$$

(The arguments to prove these results, derived from Theorem 1, are presented in Appendix I).

Note that these two results in (13) and (14) constrain the rate of how fast $(b_n)_{n\in\mathbb{N}}$ tends to zero to ensure that the estimation error vanishes $\mathbb{P}$-almost surely. Rewriting Theorem 1, we can bound 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 then, following the ideas proposed in [34]–[36] we can construct a distribution-free expression for the estimation error.

COROLLARY 2: 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 of at least $1 - \delta$,

$$\sup_{T\in\mathcal{G}_{b_n}^k}\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| <$$
$$\frac{12}{b_n}\cdot\sqrt{\frac{8}{n}\cdot(\ln(8/\delta) + k\cdot[(d+1)\cdot\ln(2) + d\cdot\ln(n)])}. \tag{15}$$

(The proof is presented in Appendix II).

For the rest of the exposition, we denote the interval of confidence on the RHS of (15) by $\epsilon_c(n, b_n, d, \delta, k)$. It is important to mention that this result is valid for a large sampling regime ($\forall n > N(\delta, k)$) to ensure that $\epsilon_c(n, b_n, \delta, k) \in (0, 3)$, which is the domain where our concentration inequality in Theorem 1 is valid (see Appendix II for details).

A. Arguments to Prove Theorem 1

The argument considers the following consequences of the *Vapnik and Chervonenkis inequality* [24], [39], [43].

LEMMA 1: (Lugosi and Nobel [20]) Let us consider $\mathcal{G}^k$ the family of tree-structure measurable partitions of $\mathbb{R}^d$ with k cells (or terminal nodes), and $Z_1, Z_2, \cdots$ i.i.d. realizations with distribution P in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then, $\forall \epsilon > 0$, $\forall n$,

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

LEMMA 2: (Vapnik and Chervonenkis [43]) Under the setting of Lemma 1, if we instead consider $\mathcal{B}$ the family of measurable rectangle[2] of $\mathbb{R}^d$, then, $\forall \epsilon > 0$, $\forall n$,

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

Proof of Theorem 1: We use that, $\forall T \in \mathcal{G}_{b_n}^k$,

$$\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| \leq \sum_{A \in \pi_T(Z_1^n)} |P_n(A) - P(A)| \cdot 3 \cdot \log(1/b_n) + \sup_{A \in \pi_T(Z_1^n)} |\log P(A) - \log P_n(A)| + \sup_{A \in \pi_T(Z_1^n)} |\log Q(A) - \log Q_n(A)|, \quad (16)$$

this bound derived from the triangular inequality and the critical mass criterion of the full tree $T_{b_n}^{full}$. In (16), Q is a short-hand notation for the product of marginal measure, i.e., $Q(A) = P(A_1 \times \mathbb{R}^q) \cdot P(\mathbb{R}^p \times A_2)$, where any set $A \in \pi_T(Z_1^n)$ has a product form denoted by $A_1 \times A_2$. On the other hand, Q_n is a short hand for the empirical version of Q, i.e., $Q_n(A) = P_n(A_1 \times \mathbb{R}^q) \cdot P_n(\mathbb{R}^p \times A_2)$, for all $A \in \pi_T(Z_1^n)$.

Concerning the first term on the RHS of (16),

$$\mathbb{P}\left(\sup_{T \in \mathcal{G}_{b_n}^k} \sum_{A \in \pi_T(Z_1^n)} |P_n(A) - P(A)| \cdot 3 \cdot \log(1/b_n) > \epsilon\right) \leq 4 \cdot (2^{d+1} \cdot n^d)^k \exp\left\{-\frac{n}{32} \cdot \left(\frac{\log(1/b_n)^{-1} \cdot \epsilon}{3}\right)^2\right\}, \quad (17)$$

from Lemma 1 and the fact that $\mathcal{G}_{b_n}^k \subset \mathcal{G}^k$. Concerning the second term on the RHS of (16), 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 $\mathcal{S}_A = \{z_1^n : P(A) - P_n(A) > P_n(A) \cdot (e^\epsilon - 1)\} \cup \{z_1^n : P_n(A) - P(A) > P_n(A) \cdot (1 - e^{-\epsilon})\}$.

[2]The *shatter coefficient* $S_{\mathcal{B}}(n)$ associated with this family of events is bounded by $(n+1)^{2d}$ (see details in [24], [43]).

Using Taylor expansion, $\forall \epsilon \in (0,1)$, $\max\{e^{\epsilon}-1, 1-e^{-\epsilon}\} > \frac{\epsilon}{2}$, then $\forall \epsilon \in (0,1)$, $\forall n \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{P}\left(\left\{z_1^n : \sup_{T\in\mathcal{G}_{b_n}^k}\sup_{A\in\pi_T}|\log P(A) - \log P_n(A)| > \epsilon\right\}\right) \leq \\
&\mathbb{P}\left(\bigcup_{T\in\mathcal{G}_{b_n}^k}\bigcup_{A\in\pi_T}\left\{z_1^n : |P_n(A)-P(A)| > P_n(A)\cdot\frac{\epsilon}{2}\right\}\right) \leq \\
&\mathbb{P}\left(\left\{z_1^n : \sup_{T\in\mathcal{G}_{b_n}^k}\sup_{A\in\pi_T}|P_n(A)-P(A)| > b_n\cdot\frac{\epsilon}{2}\right\}\right) \leq \\
&\mathbb{P}\left(\left\{z_1^n : \sup_{A\in\mathcal{B}}|P_n(A)-P(A)| > b_n\cdot\frac{\epsilon}{2}\right\}\right) \leq \\
&(n+1)^{2d}\cdot\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{2}\right)^2\right\},
\end{aligned} \tag{18}$$

where the last two inequalities are obtained from the fact that $\forall T \ll T_{b_n}^{full}$ the cells of π_T are rectangles in $\mathcal{B}$, and Lemma 2, respectively.

Concerning the last term in the RHS of (16), by definition we have that $Q(A) = P(A_1\times\mathbb{R}^q)\cdot P(\mathbb{R}^p\times A_2)$, $\forall A \in \pi_T(Z_1^n)$. Hence,

$$\begin{aligned}
\sup_{A\in\pi_T(Z_1^n)}|\log Q(A) - \log Q_n(A)| \leq &\sup_{A\in\pi_T(Z_1^n)}|\log P(A_1\times\mathbb{R}^q) - \log P_n(A_1\times\mathbb{R}^q)| + \\
&\sup_{A\in\pi_T(Z_1^n)}|\log P(\mathbb{R}^p\times A_2) - \log P_n(\mathbb{R}^p\times A_2)|\,.
\end{aligned} \tag{19}$$

From the same inequalities shown in (18),

$$\begin{aligned}
&\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k}\sup_{A\in\pi_T}|\log P(A_1\times\mathbb{R}^q) - \log P_n(A_1\times\mathbb{R}^q)| > \frac{\epsilon}{2}\right) \\
&\leq (n+1)^{2d}\cdot\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{4}\right)^2\right\}.
\end{aligned} \tag{20}$$

The same bound in (20) is obtained for the term

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

and from (19), $\forall \epsilon \in (0,2)$,

$$\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k}\sup_{A\in\pi_T}|\log Q(A) - \log Q_n(A)| > \epsilon\right) \leq 2\cdot(n+1)^{2d}\cdot\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{4}\right)^2\right\}. \tag{21}$$

To conclude, considering the inequality in (16) and the distribution free bounds obtained for its RHS terms (in (17), (18) and (21), respectively), $\forall \epsilon \in (0,3)$, we obtain

$$\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k}\left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| > \epsilon\right) \leq$$
$$(n+1)^{2d}\left[\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{6}\right)^2\right\} + 2\cdot\exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{12}\right)^2\right\}\right]$$
$$+ 4\cdot(2^{d+1}\cdot n^d)^k \exp\left\{-\frac{n}{32}\cdot\left(\frac{\log(1/b_n)^{-1}\cdot\epsilon}{9}\right)^2\right\}. \quad (22)$$

□

V. An Oracle Result

Returning to our central problem in (11), we propose the following expression for the penalization term derived from Corollary 2, $\forall n > 0$, $\forall T \ll T_{b_n}^{full}$,

$$\phi_n(|T|) = \epsilon_c\left(n, b_n, d, \delta_n\cdot b_n, |T|\right), \quad (23)$$

for a sequence $(\delta_n)_{n\in\mathbb{N}}$ of confidence probabilities in $(0,1]$ such that (δ_n) is $o(1)$. Here $\epsilon_c\left(n, b_n, d, \delta_n\cdot b_n, |T|\right)$ is the short-hand notation for the confidence interval given in the RHS of (15) (see also (57) in the Appendix II). Hence $\phi_n(|T|)$ is obtained from the confidence interval expression derived in Corollary 2, as a way to upper bound the magnitude of the estimation error with asymptotically high probability. Note that from the inequality in Theorem 1, $\phi_n(|T|)$ exclusively depends on the size of the tree and not on its structure. 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 next. Interestingly, as in the case of classification trees [35], [36], [42], the complexity term is proportional to the square root of the tree size, i.e., $\phi_n(|T|) \propto \sqrt{|T|\log(n)/n}$ from (15) and (23).

Let

$$\tilde{I}_n(\pi_T(Z_1^n)) \equiv \hat{I}_n(\pi_T(Z_1^n)) - \phi_n(|T|), \quad (24)$$

be the penalized fidelity criterion $\forall T \ll T_{b_n}^{full}$. We can express (11) by

$$\hat{T}^n = \arg\max_{k\in\left\{1,\cdots|T_{b_n}^{full}|\right\}} \tilde{I}_n(\pi_{\hat{T}^{n,k}}(Z_1^n)), \quad (25)$$

where

$$\hat{T}^{n,k} \equiv \arg\max_{T\in\mathcal{G}_{b_n}^k} \hat{I}_n(\pi_T(Z_1^n)) \quad (26)$$

is the solution of the *empirical information maximization* (EIM) constrained to the collection of pruned trees of size k, for all $k \in \left\{1,\cdots\left|T_{b_n}^{full}\right|\right\}$. The next result shows that $\hat{T}^n$ offers a nearly optimal solution for the estimation of $I(X;Y)$ with respect to an oracle solution.

THEOREM 2: Under the problem formulation of Theorem 1, if

- $(b_n) \approx (n^{-l})$ for some $l \in (0, 1/3)$ and,
- (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 $\mathbb{P}$),

$$0 \leq I(X;Y) - \tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n)) \tag{27}$$

$$\leq \min_{T \ll T_{b_n}^{full}} \left\{ [I(X;Y) - I(\pi_T(Z_1^n))] + 2\phi_n(|T|) \right\}, \tag{28}$$

and consequently, with probability $1-\delta$ (with respect to $\mathbb{P}$),

$$-\sup_{k \in \left\{1,..,|T_{b_n}^{full}|\right\}} \phi_n(k) \leq I(X;Y) - \hat{I}_n(\pi_{\hat{T}^n}(Z_1^n))$$

$$\leq \min_{T \ll T_{b_n}^{full}} \left\{ [I(X;Y) - I(\pi_T(Z_1^n))] + 2\phi_n(|T|) \right\} - \phi_n(1). \tag{29}$$

The result says two important things. (27) 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 (23). More importantly, (28) shows that, with an arbitrary high probability, 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 our estimation error bound in (23) and the true approximation error (first term from left to right in (28)). In fact, we can see the optimization in (28) as an *oracle error bound*, in the sense that it is the performance of an ideal observer that has access to the true distribution to balance the two error sources in the learning problem, i.e., the selection of the *oracle tree*

$$T^n \equiv \arg \min_{T \ll T_{b_n}^{full}} \left\{ [I(X;Y) - I(\pi_T(Z_1^n))] + 2\phi_n(|T|) \right\}. \tag{30}$$

Overall, this result is along the lines of the related oracle results obtained in the context of complexity-based pruning schemes for classification trees [34]–[36] and concept learning using structural risk minimization [42].

It is important to emphasize that $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n)$ is nearly optimal with respect to our oracle solution T^n in (30), which used a distribution-free upper bound to quantify the estimation error. The true estimation error is unaccessible given the learning nature of the problem: and consequently, the tightness of the adopted concentration inequalities are crucial to obtain good estimation error expressions and results. This was one of the reason to consider the VC concentration inequalities as the driven tool, given its non-parametric nature and its recognized goodness to model estimation errors for tree-structured partitions in other learning settings, see for instances [20], [24], [34]–[36].

From the conditions on $(b_n)_{n\in\mathbb{N}}$ stated in Theorem 2, we have that $\lim_{n\to\infty} \sup_{k \in \left\{1,..,|T_{b_n}^{full}|\right\}} \phi_n(k) = 0$ (the argument is presented in Section V-A). Consequently the oracle error bound in (28) 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 of the full tree. In fact, the consistency of the two estimate candidates, $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n))$ and $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n))$, depends upon the analysis of $\lim_{n\to\infty} \left[I(X;Y) - I(\pi_{T_{b_n}^{full}}(Z_1^n)) \right]$. Section VI formalizes this observation and shows sufficient conditions where both $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n))$ and $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n))$ are strongly consistent estimates of $I(X;Y)$.

A. *Proof of Theorem 2*

By definition in (23)

$$\phi_n(k) = \frac{12}{b_n} \cdot \sqrt{\frac{8}{n} \cdot (\ln(8) + \ln(n) - \ln(\delta_n \cdot b_n) + k \cdot [(d+1) \cdot \ln(2) + d \cdot \ln(n)])},$$

then considering that $\left|T_{b_n}^{full}\right| \leq (1/b_n)$, 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

$$\lim_{n\to\infty} \sup_{k\in\left\{1,\cdots\left|T_{b_n}^{full}\right|\right\}} \phi_n(k) = \lim_{n\to\infty} \phi_n(\left|T_{b_n}^{full}\right|) = 0. \tag{31}$$

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 deviation $\phi_n(k)$, $\forall k \in \left\{1, \cdots \left|T_{b_n}^{full}\right|\right\}$. Let

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

be the k-typical set, well defined for all n such that $k \leq \left|T_{b_n}^{full}\right|$. From Corollary 2, if $\phi_n(k) \in (0, 3)$, then $\mathbb{P}(\mathcal{S}^{n,k}) > 1 - b_n\delta_n$. Consequently from (31), there exists $N_c > 0$ such that $\forall k \in \left\{1, \cdots, \left|T_{b_n}^{full}\right|\right\}$ and $\forall n > N_c$, $\mathbb{P}(\mathcal{S}^{n,k}) > 1 - b_n\delta_n$. Hence, defining $\mathcal{S}^n \equiv \bigcap_{k\in\left\{1,\cdots\left|T_{b_n}^{full}\right|\right\}} \mathcal{S}^{n,k}$, 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 $\sup_{T\in\mathcal{G}_{b_n}^k} \left|\hat{I}_n(\pi_T(z_1^n)) - I(\pi_T(z_1^n))\right| \leq \phi_n(k)$, $\forall k \in \left\{1, \cdots \left|T_{b_n}^{full}\right|\right\}$, which also implies that [24],

$$\left| \sup_{T\in\mathcal{G}_{b_n}^k} \hat{I}_n(\pi_T(z_1^n)) - \sup_{T\in\mathcal{G}_{b_n}^k} I(\pi_T(z_1^n)) \right| \leq \phi_n(k), \tag{32}$$

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

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

where $T_k^n \equiv \arg\max_{T\in\mathcal{G}_{b_n}^k} I(\pi_T(z_1^n))$ is the oracle solution that maximizes the MI on $\mathcal{G}_{b_n}^k$. 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,

$$0 \leq I(X;Y) - \tilde{I}_n(\pi_{\hat{T}^n}(z_1^n)) \leq \min_{k\in\left\{1,\cdots\left|T_{b_n}^{full}\right|\right\}} (I(X;Y) - I(\pi_{T_k^n}(z_1^n))) + \phi_n(k). \tag{33}$$

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

VI. Density-Free Strong Consistency

Here, we restrict to the case where P is equipped with a probability density function (pdf).

THEOREM 3: Let $\left\{T_{b_n}^{full} : n > 0\right\}$ be the TSP scheme indexed by full trees and driven by the i.i.d. process $Z_1, Z_2, \cdots$ with $Z_i \sim P$ for all $i > 0$. If P is absolutely continuous with respect to the *Lebesque* measure in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and we impose the conditions on (b_n) and (δ_n) stipulated in Theorem 2, then the MI estimates obtained from $\hat{T}^n$ satisfy:

$$\lim_{n\to\infty} \hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) = I(X;Y), \tag{34}$$

$$\lim_{n\to\infty} \tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n)) = I(X;Y), \tag{35}$$

$\mathbb{P}$-almost surely ($\mathbb{P}$-a.s.).

The proof of this theorem reduces to showing that the estimation and approximation errors in (7) converge to zero $\mathbb{P}$-almost surely. We introduce three results to bound the estimation and approximation errors. We begin with the approximation error, and for that we introduce the following definition.

Definition 1: (Darbellay *et al.* [21]) Let P be a probability measure absolutely continuous with respect to the Lebesgue measure λ in $\mathbb{R}^d$ and let $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), ...\}$ be a partition scheme driven by $Z_1, Z_2, \cdots$, i.i.d. realizations with $Z_i \sim P$. Π said to be *asymptotically sufficient* for $I(X;Y)$ if,

$$\lim_{n\to\infty} I(\pi_n(Z_1^n)) = I(X,Y)$$

$\mathbb{P}$-a.s.

LEMMA 3: (*Asymptotic sufficiency* of $T_{b_n}^{full}$) Under the setting of Theorem 3, if $(b_n) \approx (n^{-l})$ for some $l \in (0, 1/3)$, then $\left\{\pi_{T_{b_n}^{full}}(\cdot) : n \geq 0\right\}$ is asymptotically sufficient for $I(X;Y)$, i.e., $\lim_{n\to\infty} I(\pi_{T_{b_n}^{full}}(Z_1^n)) = I(X;Y)$, $\mathbb{P}$-a.s. (The proof is presented in Appendix III)

LEMMA 4: Under the setting of Theorem 3, if $(b_n) \approx (n^{-l})$ with $l \in (0, 1/3)$, then

$$\lim_{n\to\infty} \left| \hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) \right| = 0 \tag{36}$$

$\mathbb{P}$-a.s. (The proof is presented in Appendix IV)

LEMMA 5: (*Asymptotic sufficiency* of $\hat{T}^n$) Under the setting of Theorem 3, if $(b_n) \approx o(n^{-l})$ with $l \in (0, 1/3)$, $(\delta_n) = o(1)$ and $(1/\delta_n) = O(e^{n^{1/3}})$, then $\forall \epsilon > 0$ there exists $N_c(\epsilon)$ such that $\forall n > N_c$ and $\forall k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}$,

$$\mathbb{P}\left(\sup_{T \in \mathcal{G}_{b_n}^k} I(\pi_T(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) > \epsilon \right) \leq$$
$$\exp\left\{ -\frac{n}{8} \cdot \left(\frac{\epsilon \cdot b_n}{24} \right)^2 \right\} + 8 \cdot (2^{d+1} \cdot n^d)^k \exp\left\{ -\frac{n}{8} \cdot \left(\frac{\epsilon \cdot b_n}{48} \right)^2 \right\},$$

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

$$\lim_{n\to\infty} I(\pi_{T_{b_n}^{full}}(Z_1^n)) = I(\pi_{\hat{T}^n}(Z_1^n)). \tag{37}$$

(The proof is presented in Appendix V)

Proof of Theorem 3: The proof comes from the following inequality,

$$\left| I(X;Y) - \hat{I}_n(\pi_{\hat{T}^n}(Z_1^N)) \right| \leq I(X;Y) - I(\pi_{T_{b_n}^{full}}(Z_1^N)) +$$
$$I(\pi_{T_{b_n}^{full}}(Z_1^N)) - I(\pi_{\hat{T}^n}(Z_1^N)) + \left| I(\pi_{\hat{T}^n}(Z_1^N)) - \hat{I}_n(\pi_{\hat{T}^n}(Z_1^N)) \right|, \tag{38}$$

where these RHS terms tend to zero $\mathbb{P}$-almost surely from Lemma 3, Lemma 5 and Lemma 4, respectively. Finally, the same result is obtained for the regularized estimate $\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^N))$, as by definition

$$\lim_{n\to\infty} \left| \tilde{I}_n(\pi_{\hat{T}^n}(Z_1^N)) - \hat{I}_n(\pi_{\hat{T}^n}(Z_1^N)) \right| \leq \lim_{n\to\infty} \sup_{k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}} \phi_n(k) = 0.$$

□

VII. Case of Study: $I(X;Y) = 0$

An interesting scenario to study is when X and Y are independent. In this context, only the estimation error plays a role and, from Theorem 2, we can guarantee a rate of convergence of $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)$ to its true value $I(X;Y) = 0$.

THEOREM 4: Let X and Y be two independent random vectors and $Z_1, Z_2 \cdots$ be i.i.d. realizations of the distribution P. Under the assumptions of Theorem 3, and specifically considering that $(1/\delta_n) \approx (e^{n^{1/3}})$, $\mathbb{E}(\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)))$ is $O(e^{-n^{1/3}+\log\log n})$.

Theorem 4 implies that $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)$ converges to zero faster than any decreasing polynomial order $\mathbb{P}$-a.s. More formally, we have the following result.

COROLLARY 3: Under the setting of Theorem 4, $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)$ is $o(n^{-p})$ $\mathbb{P}$-a.s. for any finite $p > 0$.
(The proof is presented in Appendix VI)

A. *Proof of Theorem 4*

Let $\delta > 0$ and let us define $\phi_n(k,\delta) \equiv \epsilon_c(n, b_n, d, \delta \cdot b_n, k)$ from (15). Let

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

be the collection of δ-typical sequences associated with $\mathcal{G}_{b_n}^k$ $\forall k \in \left\{1, \cdots \left| T_{b_n}^{full} \right| \right\}$ and $\mathcal{S}_\delta^n \equiv \bigcap_{k \in \left\{1,\cdots |T_{b_n}^{full}|\right\}} \mathcal{S}^{n,k}$ be the general collection of δ-typical sequences $\forall n > 0$. In this context we can impose $\phi_n(1,\delta) = 0$, $\forall \delta > 0$, because for $k = 1$ we only have the trivial partition $\left\{\mathbb{R}^d\right\}$ in $\mathcal{G}_{b_n}^1$, and in this domain $\hat{I}_n(\pi_T(z_1^n)) = I(\pi_T(z_1^n)) = 0$, $\forall z_1^n \in \mathbb{R}^{d\cdot n}$. In fact, even with this stronger definition $\mathcal{S}_\delta^{n,1} = \mathbb{R}^{d\cdot n}$, $\forall \delta > 0$ and $\forall n > 0$.

Let $z_1^n \in \mathcal{S}_\delta^n$ be a δ-typical sequence. From the arguments presented in the proof of Theorem 2, if $\hat{T}^n$ is the solution of the complexity-penalized problem in (25), considering our more specific penalization term $\phi_n(k,\delta)$, then

$$\begin{aligned} 0 \leq & I(X;Y) - \tilde{I}_n(\pi_{\hat{T}^n}(z_1^n)) \leq \\ & \min_{k \in \left\{1,\cdots |T_{b_n}^{full}|\right\}} (I(X;Y) - I(\pi_{T_k^n}(z_1^n))) + \phi_n(k,\delta) = 0, \end{aligned}$$

where the last equality is from $I(X,Y) = 0$ and the fact that $\phi_n(1,\delta) = 0$. Consequently $\tilde{I}_n(\pi_{\hat{T}^n}(z_1^n) = 0$ and from the construction of $\mathcal{S}_\delta^n$ it is simple to show that $\left|\hat{T}^n\right| = 1$, which implies that $\hat{I}_n(\pi_{\hat{T}^n}(z_1^n) = 0$. Then, restricted to the collection of δ-typical sequences we have a zero-error empirical estimate.

On the other hand, for an arbitrary sequence $z_1^n \in \mathbb{R}^{d\cdot n}$ we have that $\hat{I}_n(\pi_{\hat{T}^n}(z_1^n)) \leq \log(1/b_n)$. This last inequality is obtained from the fact that $\hat{I}_n(\pi_{\hat{T}^n}(z_1^n))$ is bounded by the entropy of P_n restricted to $(\mathbb{R}^d, \sigma(\hat{T}^n(z_1^n)))$, which is upper bounded by the entropy of the uniform distribution with critical probability mass b_n. In addition from the Corollary 2, if $\sup_{k \in \left\{1,\cdots |T_{b_n}^{full}|\right\}} \phi_n(k,\delta) < 3$, then $\mathbb{P}(\mathcal{S}_\delta^n) \geq 1 - \delta$, which implies that

$$\mathbb{E}(\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n))) \leq \delta \cdot \log(1/b_n). \tag{39}$$

The rest of the proof reduces to finding a sequence $(\delta_n)_{n\in\mathbb{N}}$ that tends to zero at the fastest possible rate while guaranteeing

$$\lim_{n\to\infty} \sup_{k\in\left\{1,\cdots|T_{b_n}^{full}|\right\}} \phi_n(k,\delta_n) = 0, \tag{40}$$

which from (39) allows us to show that $\exists N_c > 0$ such that $\forall n > N_c$,

$$\mathbb{E}(\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n))) \leq \delta_n \cdot \log(1/b_n). \tag{41}$$

Note that in this context we have that $\phi_n(k) = \phi_n(k,\delta_n)$, then we reduce to the original complexity regularized problem in (25). Finally, from the construction of $\phi_n(k,\delta_n)$ and (15), it is simple to show that $(\delta_n) \approx (e^{-n^{1/3}})$ satisfies (40) considering that $(b_n) \approx (n^{-l})$ with $l \in (0,1/3)$, which concludes the proof. □

VIII. Algorithmic Solutions

We conclude this work by connecting the main learning-decision problem in (25) with some results and algorithmic solutions of well-understood complexity-regularized tree pruning problems [33], [37], [44]. We first rewrite (25) in the following form

$$\hat{T}^n = \arg\min_{T\ll T_{b_n}^{full}} -\rho(T) + \phi(|T|), \tag{42}$$

where $\rho(T) = \hat{I}_n(\pi_T(Z_1^n))$ and $\phi(|T|) = C_n\sqrt{|T|}$, with $C_n = 12/b_n\sqrt{((d+1)\ln(2) + d\ln(n))\frac{8}{n}}$. In this context $\rho(T)$ is a non-decreasing and additive function of the tree. It is non-decreasing in the sense that $\rho(T_2) \geq \rho(T_1)$ when $T_1 \ll T_2$ [2], [21], and it is additive (see definition in [37]) because it is written as the sum of terms indexed with the leaves of T (from (10) and (6)). Concerning the additivity, let T_v denote the branch of a tree T rooted at $v \in \mathcal{I}(T)$, and let $\Delta\rho(v|U_v)$ be the conditional MI gain obtained by partitioning U_v in terms of $\left\{U_{l(v)}, U_{r(v)}\right\}$, i.e.,

$$\Delta\rho(v|U_v) = \frac{P_n(U_{l(v)})}{P_n(U_v)} \log \frac{P_n(U_{l(v)})/P_n(U_v)}{Q_n(U_{l(v)})/Q_n(U_v)} + \frac{P_n(U_{r(v)})}{P_n(U_v)} \log \frac{P_n(U_{r(v)})/P_n(U_v)}{Q_n(U_{r(v)})/Q_n(U_v)}, \tag{43}$$

which is well defined for all $v \in \mathcal{I}(T_{full}^{b_n})$. In this last expression, Q_n is a short-hand notation for the product of marginal empirical measures, i.e., $Q_n(A) = P_n(A_1 \times \mathbb{R}^q) \cdot P_n(\mathbb{R}^p \times A_2)$, which is well characterized when its argument A has a product form $A_1 \times A_2$. In addition, $l(v)$ and $r(v)$ denote the left and right children of v, respectively. In the same way, we can define the conditional mutual information (CMI) associated to a branch of T rooted at $v \in T$ (or equivalently, the CMI gain of partitioning U_v in terms of $\{U_t : t \in \mathcal{L}(T_v)\}$) by

$$\rho(T_v|U_v) \equiv \sum_{t\in\mathcal{L}(T_v)} \frac{P_n(U_t)}{P_n(U_v)} \log \frac{P_n(U_t)/P_n(U_v)}{Q_n(U_t)/Q_n(U_v)}, \tag{44}$$

where, in particular, $\rho(T) = \rho(T_{root}|U_{root})\ \forall T \ll T_{b_n}^{full}$, with $root$ denoting the root of $T_{b_n}^{full}$. With these definitions it is simple to show that, $\forall T \ll T_{b_n}^{full}$ such that $|T| \geq 2$ and $\forall v \in \mathcal{I}(T)$,

$$\rho(T_v|U_v) = \Delta\rho(v|U_v) + \frac{P_n(U_{l(u)})}{P_n(U_v)} \cdot \rho(T_{l(v)}|U_{l(v)}) + \frac{P_n(U_{r(u)})}{P_n(U_v)} \cdot \rho(T_{r(v)}|U_{r(v)}). \tag{45}$$

A. Minimum Cost Tree Pruning Algorithm

It is known that $\hat{T}^n$ belongs to $\left\{\hat{T}^{n,k} : k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}\right\}$ [37], which is the family of *the minimum cost trees* given by

$$\hat{T}^{n,k} \equiv \arg \max_{\left\{T \ll T_{b_n}^{full} : |T|=k\right\}} \rho(T), \tag{46}$$

for all $k \in \left\{1, \cdots \left|T_{b_n}^{full}\right|\right\}$. Let $\mathbf{T}_v$ denote the brach of $T_{b_n}^{full}$ rooted at $v \in T_{b_n}^{full}$ (i.e., $\mathbf{T}_v = (T_{b_n}^{full})_v$). Then we can generalize the family of trees in (46) as follows, $\forall v \in T_{b_n}^{full}$,

$$\hat{T}_v^{n,k} \equiv \arg \max_{\{T \ll \mathbf{T}_v : |T|=k\}} \rho(T|U_v), \tag{47}$$

$\forall k \in \{1, \cdots |\mathbf{T}_v|\}$, where in particular $\hat{T}^{n,k} = \hat{T}_{root}^{n,k}$. Dynamic programing can be used to solve (47) by using the additive structure of $\rho(T)$ in (45).

PROPOSITION 1: (Scott [37]) For all $v \in \mathcal{I}(T_{b_n}^{full})$ and $\forall k \in \{2, \ldots, |\mathbf{T}_v|\}$, we have that

$$\hat{T}_v^{n,k} = \left[\left[v, \hat{T}_{l(v)}^{n,k_1^*}, \hat{T}_{r(v)}^{n,k_2^*}\right]\right], \tag{48}$$

where[3]

$$(k_1^*, k_2^*) = \arg \max_{\substack{(k_1,k_2) \in \{1,..,|\mathbf{T}_{l(v)}|\} \times \{1,..,|\mathbf{T}_{r(v)}|\} \\ k_1+k_2=k}} \left[\frac{P_n(U_{l(u)})}{P_n(U_v)} \rho(\hat{T}_{l(v)}^{n,k_1}|U_{l(v)}) + \frac{P_n(U_{r(u)})}{P_n(U_v)} \rho(\hat{T}_{r(v)}^{n,k_2}|U_{r(v)})\right]. \tag{49}$$

This result offers a bottom-up algorithm to solve $\left\{\hat{T}^{n,k} : k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}\right\}$ (a pseudo code is presented in [37]) and, by an exhaustive search on this family, a way to find $\hat{T}^n$. Bohanec *et al.* [45] showed that the computational complexity of this methodology is $O(\left|T_{b_n}^{full}\right|^2) = O(1/{b_n}^2) = O(n^{2/3})$ for our case of balanced trees.

B. Family Pruning Algorithm

An alternative solution can be derived from the analysis of the more general problem

$$\hat{T}^n(\alpha) = \arg \min_{T \ll T_{b_n}^{full}} -\rho(T) + \alpha \cdot \phi(|T|), \tag{50}$$

for all $\alpha \geq 0$, where in particular $\hat{T}^n = \hat{T}^n(1)$. Scott [37] showed the following result for the case of a *sub-additive penalty* (Definition 1 in [37]), which is our case as $\phi(|T|) \propto \sqrt{|T|}$.

THEOREM 5: (Scott [37, Theorem 2]) For the case when $\phi(T)$ is sub-additive and $\rho(T)$ is an additive and non-decreasing function of the tree, there exists $m \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}$, a strictly increasing sequence of real numbers, $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \infty$, and a nested family of trees $R_1 = T_{b_n}^{full} \gg R_2 \gg \ldots \gg R_m = \{root\}$, such that: $\forall i \in \{1, .., m\}$, $\forall \alpha \in [\alpha_{i-1}, \alpha_i)$,

$$\hat{T}^n(\alpha) = R_i. \tag{51}$$

The fact that the family of solutions $\left\{\hat{T}^n(\alpha) : \alpha \geq 0\right\}$ is nested, allows to find $\{R_1, .., R_m\}$ and $\{\alpha_1, .., \alpha_{m-1}\}$ efficiently. Algorithms to solve this problem have been proposed by Breiman *et al.* [33], Chou *et al.* [44] and more

[3]Using Scott's nomenclature [37], $[[v, T_1, T_2]]$ denotes a binary tree T with root v, and branches $T_{l(v)} = T_1$ and $T_{r(v)} = T_2$.

recently by Scott [37]. The computational complexity for our case of balanced trees is $O(\left|T_{b_n}^{full}\right| \log \left|T_{b_n}^{full}\right|) = O(1/b_n \cdot \log(1/b_n)) = O(n^{1/3} \log n)$ [33], [44].

IX. Concluding Comments

The conditions on $(b_n)_{n\in\mathbb{N}}$ to ensure that the complexity-regularized tree $\hat{T}^n$ 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 was presented by the authors in [28]. In other words, we have that the full tree obtained from the growing phase induces a strongly consistent estimate as well. At this point, it is important to highlight the adaptation character of our complexity regularized solution. This solution finds the tree's topology that offers a nearly optimal balance between our developed estimation and approximation error expressions (Theorem 2) as a function of the data and not only its size. To illustrate the idea, if the target value $I(X;Y)$ is high then we get a less conservative (or bigger) complexity regularized tree $\hat{T}^n$, than in the case of a moderate MI magnitude (in particular the case of independent random variables), this from the oracle results of Theorem 2. In contrast, the full tree solution does not allow for this tree structure adaptation to the problem. Furthermore, the fact that the size of $\hat{T}^n$ is able to adapt to the underlying magnitude of $I(X;Y)$ is what is crucial to obtain the rate of convergence result stated in Theorem 4.

Concerning Theorem 4, the rate of convergence obtained for the independent scenario suggests that $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n))$ can be used as an attractive statistic to construct a test of independence between continuous random variables of the form: decide H_0 when $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) < \tau$ (with H_0 the hypothesis of independence). Statistical tests of independence usually require the characterization of a *significant level* (probability of rejecting H_0 when it is true) using asymptotic distributions, which are only valid for the large sampling regime. An advantage of our setting is the existence of distribution-free concentration inequalities that can be used to define accurate significant levels, and consequently to construct a test with performance bounds for any finite sampling length.

Finally from the analysis of some empirical results for the case when X and Y are independent, it is observed that our solution is able to detect this condition and get a zero-error estimate (associated with the trivial partition $\pi_{\hat{T}^n}(Z_1^n) = \left\{\mathbb{R}^d\right\}$) with a finite number of samples. Consequently the empirical evidence shows that the estimate behaves better than what the theory predicts in Theorem 4. This suggests that the rate of convergence obtained in Section VII can be improved on the lines of a result that guarantees an error-free estimate for a finite number of sampling points almost surely. This conjecture is an interesting direction to explore and it is left as a future work.

X. Acknowledgment

The work of J. Silva was supported by funding from FONDECYT Grant 1110145, CONICYT-Chile. The work of S. Narayanan was supported in part by grants from the Office of Naval Research (ONR), the Army, the National Science Foundation (NSF), and DARPA. We want to thank our colleagues Pablo Navarrete, Claudio Estevez, as well as the anonymous reviewer for providing comments and suggestions that helped us to improve the quality of this work.

APPENDIX I
PROOF OF COROLLARY 1

From the distribution-free bound of $\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k}\left|\hat{I}_n(\pi_T(Z_1^n))-I(\pi_T(Z_1^n))\right|>\epsilon\right)$ in Theorem 1, $\forall\epsilon\in(0,3)$ we have two distinctive terms proportional to

$$\approx (n+1)^d\cdot\exp\left\{-n(b_n\cdot\epsilon)^2\right\}+ [(2n)^d\cdot 2]^k\cdot\exp\left\{-n\left(\log(1/b_n)^{-1}\cdot\epsilon\right)^2\right\}. \tag{52}$$

It is sufficient to check that these two last expressions are dominated by $(e^{-n^\tau})_{n\in\mathbb{N}}$, for some $\tau\in(0,1)$. As $(1/b_n)\approx(n^{-l})$ for some $l\in(0,\frac{1}{2})$ then there exits $\tau\in(0,1)$, such that $(b_n)\succeq(n^{-\frac{1-\tau}{2}})$. Working with

$$\frac{1}{n^\tau}\log\left[(n+1)^d\cdot\exp\left\{-n(b_n\cdot\epsilon)^2\right\}\right]= \frac{d\cdot\log n+1}{n^\tau}-n^{1-\tau}(b_n)^2\epsilon^2, \tag{53}$$

$\lim_{n\to\infty}\frac{1}{n^\tau}\log\left[(n+1)^d\cdot\exp\left\{-n(b_n\cdot\epsilon)^2\right\}\right]<0$ and consequently $((n+1)^d\cdot\exp\left\{-n(b_n\cdot\epsilon)^2\right\})_{n\in\mathbb{N}}\preceq(e^{-n^\tau})_{n\in\mathbb{N}}$. For the second term in (52),

$$\frac{1}{n^\tau}\log\left[[(2n)^d\cdot 2]^k\cdot\exp\left\{-n\left(\log(1/b_n)^{-1}\cdot\epsilon\right)^2\right\}\right]= \frac{k\cdot(d\log n+\log 2)}{n^\tau}-n^{1-\tau}(\log(1/b_n)^{-1}\epsilon)^2. \tag{54}$$

and considering that $\log(1/b_n)^{-1}\geq b_n$, $\forall n$ (because by definition $(1/b_n)\geq 1$), we also have that, $\lim_{n\to\infty}\frac{1}{n^\tau}\log\left[[(2n)^d\cdot 2]^k\cdot\exp\left\{-n\left(\log(1/b_n)^{-1}\cdot\epsilon\right)^2\right\}\right]<0$. Consequently, we obtain that $\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k}\left|\hat{I}_n(\pi_T(Z_1^n))-I(\pi_T(Z_1^n))\right|>\epsilon\right)$ is dominated by $(e^{-n^\tau})_{n\in\mathbb{N}}$ which from the *Borel Cantelli lemma* [30] proves the result in (13).

Concerning the second part, from Theorem 1, $\mathbb{P}\left(\left|\hat{I}_n(\pi_{T_{b_n}^{full}}(Z_1^n))-I(\pi_{T_{b_n}^{full}}(Z_1^n))\right|>\epsilon\right)$ is bounded by an expression dominated by $(n+1)^d\cdot\exp\left\{-n(b_n\cdot\epsilon)^2\right\}+[(2n)^d\cdot 2]^{\left|T_{b_n}^{full}\right|}\cdot\exp\left\{-n\left(\log(1/b_n)^{-1}\cdot\epsilon\right)^2\right\}$. As in this case $(1/b_n)\approx(n^l)$ for some $l\in(0,\frac{1}{3})\subset(0,\frac{1}{2})$, we just need to concentrate on the analysis of the second term, i.e., on getting a negative asymptotic value for

$$\frac{1}{n^\tau}\log\left[[(2n)^d\cdot 2]^{\left|T_{b_n}^{full}\right|}\cdot\exp\left\{-n\left(\log(1/b_n)^{-1}\cdot\epsilon\right)^2\right\}\right]= \frac{\left|T_{b_n}^{full}\right|\cdot(d\log n+(d+1)\log 2)}{n^\tau}-n^{1-\tau}(\log(1/b_n)^{-1}\epsilon)^2, \tag{55}$$

for some $\tau\in(0,1)$. Note that by construction $\left|T_{b_n}^{full}\right|\leq\frac{1}{b_n}$. Using this, the positive term of the RHS of (55) tends to zero if $(1/b_n)$ is $o(n^\tau/\log n)$. On the other hand, using again that $\log(1/b_n)^{-1}\geq b_n$, the second term of the RHS of (55) tends to negative infinity if $(1/b_n)$ is $o(n^{\frac{1-\tau}{2}})$. Finally, noting that $\frac{1}{3}=\max_{\tau\in(0,1)}(\tau,\frac{1-\tau}{2})$, from the assumption about (b_n) we obtain that there exists $\tau_o\in(0,\frac{1}{3})$ such that $([(2n)^d\cdot 2]^{\left|T_{b_n}^{full}\right|}\cdot\exp\left\{-n\left(\log(1/b_n)^{-1}\cdot\epsilon\right)^2\right\})_{n\in\mathbb{N}}\preceq(e^{-n^{\tau_o}})_{n\in\mathbb{N}}$, which proves (14). □

APPENDIX II
PROOF OF COROLLARY 2

Note that the distribution-free bound of Theorem 1 can be upper bounded by the following simpler expression,

$$8 \cdot \left(2^{d+1} \cdot n^d\right)^k \cdot \exp\left\{-\frac{n}{8}\left(\frac{b_n \cdot \epsilon}{12}\right)^2\right\}. \tag{56}$$

Consequently, for an arbitrary $\delta > 0$ and $k \in \mathbb{N}$, the critical ϵ such that (56) is equal to δ is

$$\epsilon_c(n, b_n, d, \delta, k) = \frac{12}{b_n} \cdot \sqrt{\frac{8}{n} \cdot (\ln(8/\delta) + k \cdot [(d+1) \cdot \ln(2) + d \cdot \ln(n)])}. \tag{57}$$

This is the probability $1 - \delta$. confidence interval of the estimation error deviant. However, this is valid as long as $\epsilon_c(n, b_n, d, \delta, k) < 3$, from Theorem 1. Here is where $(1/b_n) \approx (n^l)$ with $l \in (0, 1/3)$ comes into play, because this condition implies that $\forall \delta > 0$ and $\forall k \in \mathbb{N}$,

$$\lim_{n\to\infty} \epsilon_c(n, b_n, d, \delta, k) = 0.$$

□

APPENDIX III
PROOF OF LEMMA 3

We need to first introduce some definitions and two results.

Let us state a sufficient condition to get that a partition scheme is asymptotically sufficient for $I(X;Y)$.

THEOREM 6: (Silva and Narayanan [28, Th. 2]) Let $P_{X,Y}$ be absolutely continuous with respect to the Lebesgue measure λ in $\mathbb{R}^d$ and let $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), ...\}$ be a partition scheme driven by $Z_1, Z_2, \cdots$, i.i.d. realizations with $Z_i \sim P_{X,Y}$ for all i. If $\forall \delta > 0$,

$$\lim_{n\to\infty} P_{X,Y}\left(\left\{z \in \mathbb{R}^d : \text{diam}(\pi_n(z|Z_1^n)) > \delta\right\}\right) \to 0, \tag{58}$$

$\mathbb{P}$-almost surely (a.s.), then,

$$\lim_{n\to\infty} I(\pi_n(Z_1^n)) = I(X,Y), \quad \mathbb{P} - a.s. \tag{59}$$

Definition 2: Let T be a binary tree. For all $t \in T$ let $depth(t)$ denote the *depth* of t — the number of arcs that connect t with the root of T. In this context, let $T^{(r)}$ denote the truncated version of T, formally given by $T^{(r)} = \{t \in T : depth(t) \leq r\}$, where by construction $T^{(r)} \ll T$.

Definition 3: Let T be a binary tree, we say that T is a *balanced tree* of height r if $\forall t \in \mathcal{L}(T)$, $depth(t) = r$.

Definition 4: A TSP scheme $\Pi = \{T_1, T_2, \cdots\}$ is a *uniform balanced tree-structured scheme* (UBTSS), if each partition rule $T_n(\cdot)$ forms a balanced tree of height d_n (only function of n).

In the context of uniform balanced tree-structured scheme we have the following result.

LEMMA 6: (Silva [46]) Let $\Pi = \{T_1, T_2, \cdots\}$ be a UBTSS induced by the statistically equivalent splitting process presented in Section III. Let $(d_n)_{n\in\mathbb{N}}$ denote its height sequence, then Π satisfies the shrinking cell condition of Theorem 6, if there exists a non-negative real sequence $(q_n) \approx (n^\theta)$, for some $\theta > 0$, such that

$$\frac{n}{d_n 2^{d_n}} - \frac{q_n}{d_n} \to \infty \; and \; d_n \to \infty, \text{ as } n \text{ tends to infinity.}$$

The result derives from the ideas presented by Devroye *et al.* [24, Theorem 20.2], and the proof can be found in [46, Ch. 4, Lemma 4.3].

Proof of Lemma 3: From Theorem 6 the proof reduces to checking the shrinking cell condition in (58). If we define $r(n) = \min_{t \in \mathcal{L}(T_{b_n}^{full}(Z_1^n))} depth(t)$, by construction of $T_{b_n}^{full}$ we have that

$$\log_2 1/b_n \geq r(n) \geq \log_2 1/b_n - 1.$$

Let $\bar{d}_n \equiv \lfloor \log_2 \frac{1}{b_n} \rfloor - 1$ for all n, then we can define the UBTSS $\bar{\Pi} = \{\bar{T}_1, \bar{T}_2, \cdots\}$, by $\bar{T}_n \equiv (T_{b_n}^{full})^{(\bar{d}_n)}$ for all $n > 0$. By construction, $\bar{T}_n \ll T_{b_n}^{full}$ (and consequently $\pi_{T_{b_n}^{full}}(Z_1^n)$ is a refined version of $\pi_{\bar{T}_n}(Z_1^n)$), then the shrinking cell condition of $\bar{\Pi}$ implies the property for Π. For $\bar{\Pi}$ we can check the sufficient conditions of Lemma 6. Considering that $(b_n) \approx (n^{-l})$ for some $l \in (0, 1/3)$, then $\bar{d}_n \to \infty$ as n tends to infinity and, furthermore, we can consider an arbitrary non-negative sequence $(q_n) \approx (n^\theta)$ with $\theta \in \left(0, \frac{2}{3}\right]$, where

$$\frac{n}{\bar{d}_n 2^{\bar{d}_n}} - \frac{q_n}{\bar{d}_n} \geq \frac{n}{d_n \cdot 2^{\log_2(1/b_n)}} - \frac{q_n}{\bar{d}_n} = \frac{b_n \cdot n - q_n}{\bar{d}_n} \to \infty \tag{60}$$

as $n \to \infty$, because $(d_n) \preceq (\log_2(n))$ and (q_n) is $o(b_n \cdot n)$, which concludes the proof. □

APPENDIX IV
PROOF OF LEMMA 4

Proof: Note that

$$\begin{aligned}
&\left\{ z_1^n \in \mathbb{R}^{d \cdot n} : \left| \hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) \right| > \epsilon \right\} \subset \\
&\bigcup_{k=1}^{|T_{b_n}^{full}|} \left\{ z_1^n \in \mathbb{R}^{d \cdot n} : \left| \hat{I}_n(\pi_{\hat{T}_k^n}(Z_1^n)) - I(\pi_{\hat{T}_k^n}(Z_1^n)) \right| > \epsilon \right\} \subset \\
&\bigcup_{k=1}^{|T_{b_n}^{full}|} \left\{ z_1^n \in \mathbb{R}^{d \cdot n} : \sup_{T \in \mathcal{G}_{b_n}^k} \left| \hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n)) \right| > \epsilon \right\},
\end{aligned}$$

consequently from the estimation error expression in (56) (see Appendix II),

$$\begin{aligned}
&\mathbb{P}\left(\left| \hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) \right| > \epsilon \right) \leq \\
&\left| T_{b_n}^{full} \right| \cdot 8 \cdot \left(2^{d+1} \cdot n^d \right)^{|T_{b_n}^{full}|} \cdot \exp\left\{ -\frac{n}{8} \left(\frac{b_n \cdot \epsilon}{12} \right)^2 \right\}.
\end{aligned} \tag{61}$$

Finally using the same arguments adopted to bound the RHS expression of (55) in Corollary 1 (see details in Appendix I), we have that there exists $\tau_o \in (0, 1/3)$ such that $\forall \epsilon > 0$, $\mathbb{P}\left(\left| \hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) \right| > \epsilon \right)$ is dominated by the sequence $(e^{-n^{\tau_o}})_{n \in \mathbb{N}}$, which proves the result from the *Borel-Cantelli lemma* [30]. □

APPENDIX V
PROOF OF LEMMA 5

Proof: By triangular inequality,

$$\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} I(\pi_T(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) > \epsilon\right) \leq \mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} I(\pi_T(Z_1^n)) - \sup_{T\in\mathcal{G}_{b_n}^k} \tilde{I}_n(\pi_T(Z_1^n)) > \epsilon/2\right) + \mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} \tilde{I}_n(\pi_T(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) > \epsilon/2\right). \quad (62)$$

Without loss of generality let us consider $\epsilon < 1$. As $\lim_{n\to\infty} \sup_{k\in\{1,..,|T_{b_n}^{full}|\}} \phi_n(k) = 0$ (see proof of Theorem 2), there exists $N_c(\epsilon)$ such that $\forall n > N_c(\epsilon)$, $\sup_{k\in\{1,..,|T_{b_n}^{full}|\}} \phi_n(k) < \epsilon/4$, then for the first term in (62),

$$\begin{aligned}
&\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} I(\pi_T(Z_1^n)) - \sup_{T\in\mathcal{G}_{b_n}^k} \hat{I}_n(\pi_T(Z_1^n)) + \phi_n(k) > \epsilon/2\right) \leq \\
&\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} I(\pi_T(Z_1^n)) - \sup_{T\in\mathcal{G}_{b_n}^k} \hat{I}_n(\pi_T(Z_1^n)) > \epsilon/4\right) \leq \\
&\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} \left|I(\pi_T(Z_1^n)) - \hat{I}_n(\pi_T(Z_1^n))\right| > \epsilon/4\right) \leq \\
&8\cdot(2^{d+1}\cdot n^d)^k \exp\left\{-\frac{n}{8}\cdot\left(\frac{\epsilon\cdot b_n}{48}\right)^2\right\}. \quad (63)
\end{aligned}$$

For the second term in (62), $\forall n > N_c(\epsilon)$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} \tilde{I}_n(\pi_T(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) > \epsilon/2\right) \leq \\
&\mathbb{P}\left(\tilde{I}_n(\pi_{\hat{T}^n}(Z_1^n)) - I(\pi_{\hat{T}^n}(Z_1^n)) > \epsilon/2\right) \leq \\
&\sum_{k=1}^{|T_{b_n}^{full}|} \mathbb{P}\left(\sup_{T\in\mathcal{G}_{b_n}^k} \left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n))\right| > \epsilon/2 + \phi_n(k)\right) \leq \\
&\sum_{k=1}^{|T_{b_n}^{full}|} 8\left(2^{d+1}n^d\right)^k \exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot(\epsilon/2+\phi_n(k))}{12}\right)^2\right\} \leq \\
&\sum_{k=1}^{|T_{b_n}^{full}|} 8\left(2^{d+1}n^d\right)^k \exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{24}\right)^2\right\}\exp\left\{-\frac{n}{8}\left(\frac{b_n\phi_n(k)}{12}\right)^2\right\} \\
&\leq \sum_{k=1}^{|T_{b_n}^{full}|} \delta_n b_n \cdot \exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{24}\right)^2\right\} \\
&\leq \exp\left\{-\frac{n}{8}\left(\frac{b_n\cdot\epsilon}{24}\right)^2\right\}, \quad (64)
\end{aligned}$$

where the first inequality comes from the definition of $\hat{T}^n$, the third inequality uses the estimation error expression in (56) (Appendix II), the fourth uses that $(\epsilon/2+\phi_n(k))^2 \geq (\epsilon/2)^2 + \phi_n(k)^2$, the fifth is by the definition of $\phi_n(k)$ in (23) and the fact that $\sup_{k\in\{1,..,|T_{b_n}^{full}|\}} \phi_n(k) < \epsilon/4 < 3$, and the last uses that $\left|T_{b_n}^{full}\right| \leq (1/b_n)$ and $\delta_n \leq 1$.

Then from (62), (63) and (64) we get the inequality. Note that this inequality is valid uniformly in $k \in \left\{1, .., \left|T_{b_n}^{full}\right|\right\}$ and in particular for $k_n = \left|T_{b_n}^{full}\right| \leq 1/b_n$. Finally evaluating this distribution free bound in $k_n = (1/b_n)$ this expression is asymptotically dominated by $(e^{-n^{\tau_o}})_{n\in\mathbb{N}}$ for some $\tau_o \in (0, 1/3)$, by the same arguments adopted to bound the RHS expression of (55) in Corollary 1 (see details in Appendix I). Consequently we prove (37). □

APPENDIX VI
PROOF OF COROLLARY 3

Proof: Let us consider $\epsilon > 0$ and a sequence $(n^{-p})_{n\in\mathbb{N}}$ for some $p > 0$. To show the result we need to analyze the asymptotic decay of $\mathbb{P}(\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n) > \epsilon \cdot n^{-p})$, with $\mathbb{P}$ the probability measure of $Z_1, Z_2, \ldots$. From the *Markov's inequality* [31],

$$\begin{aligned}\mathbb{P}(\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n) > \epsilon \cdot n^{-p}) &\leq \frac{\mathbb{E}(\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)))}{n^{-p}\cdot\epsilon}, \\ &\leq \frac{n^p \cdot \log n}{e^{n^{1/3}}\cdot\epsilon}.\end{aligned}$$

Note that $\sum_{n\geq 0} \frac{n^p\cdot\log n}{e^{n^{1/3}}\cdot\epsilon} < \infty$, $\forall\epsilon > 0$, then the *Borel-Cantelli* lemma implies that $\hat{I}_n(\pi_{\hat{T}^n}(Z_1^n)$ is $o(n^{-p})$ $\mathbb{P}$-almost surely. □

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] J. W. Fisher III, M. Wainwright, E. Sudderth, and A. S. Willsky, "Statistical and information-theoretic methods for self-organization and fusion of multimodal networked sensors," *International Journal of High Performance Computing Applications*, vol. 16, no. 3, pp. 337–353, 2002.

[5] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Transactions on Image Processing*, vol. 10, no. 11, pp. 1647–1658, November 2001.

[6] 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.

[7] T. Butz and J.-P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Elsevier Signal Processing*, vol. 85, pp. 875–902, 2005.

[8] J. Kim, J. W. Fisher III, A. Yezzi, M. Cetin, and A. S. Willsky, "A nonparametric statistical method for image segmentation using information theory and curve evolution," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1486–1502, October 2005.

[9] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 512–519, July 2005.

[10] J. Silva and S. Narayanan, "Minimum probability of error signal representation," in *IEEE Workshop Machine Learning for Signal Processing*, August 2007.

[11] ——, "Discriminative wavelet packet filter bank selection for pattern recognition," *IEEE Transactions on Signal Processing*, vol. 57, no. 5, pp. 1796–1810, 2009.

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

[13] L. Devroye and L. Györfi, *Nonparametric density estimation: The L_1 view*. Wiley Interscience, New York, 1895.

[14] S. Abou-Jaoude, "Condition nécessaires et sufficantes de convergence L_1 en probabilité de l'hitograme pour une densité," *Ann. Inst. H. Poincaré*, vol. 12, pp. 213–231, 1976.

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

[16] L. Györfi and E. C. van der Meulen, "Density estimation consistent in information divergence," in *IEEE International Symposium on Information Theory*, 1994, pp. 35–35.

[17] L. Györfi, F. Liese, I. Vajda, and E. C. van der Meulen, "Distribution estimates consistent in χ^2- divergence," *Statistics*, vol. 32, no. 1, pp. 31–57, 1998.

[18] I. Vajda and E. C. van der Meulen, "Optimization of barron density estimates," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1867–1883, July 2001.

[19] A. Berlinet, I. Vajda, and E. C. van der Meulen, "About the asymptotic accuracy of Barron density estimate," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 999–1009, May 1998.

[20] 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.

[21] G. A. Darbellay and I. 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.

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

[23] ——, "Universal estimation of information measures for analog sources," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 3, pp. 265–353, 2009.

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

[25] A. B. Nobel, "Histogram regression estimation using data-dependent partitions," *The Annals of Statistics*, vol. 24, no. 3, pp. 1084–1105, 1996.

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

[27] ——, "Histogram-based estimation for the divergence revisited," in *IEEE International Symposium on Information Theory*, June 2009.

[28] ——, "Non-product data-dependent partitions for mutual information estimation: Strong consistency and applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3497–3511, July 2010.

[29] ——, "Information divergence estimation based on data-dependent partitions," *Journal of Statistical Planning and Inference*, vol. 140, no. 11, pp. 3180 – 3198, November 2010.

[30] P. R. Halmos, *Measure Theory*. Van Nostrand, New York, 1950.

[31] L. Breiman, *Probability*. Addison-Wesley, 1968.

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

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

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

[35] A. 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.

[36] C. Scott and R. D. Nowak, "Dyadic classification trees via structural risk minimization," in *in Advances in Neural Information Processing Systems*, S. Becher, S. Thrun, and K. Obermayer, Eds., vol. 15. Cambridge, MA: MIT, 2003.

[37] C. Scott, "Tree pruning with subadditive penalties," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, pp. 4518–4525, 2005.

[38] V. Vapnik, *Estimation of dependencies based on empirical Data*. Springer - Verlag, New York, 1979.

[39] ——, *Statistical Learning Theory*. John Wiley, 1998.

[40] A. R. Barron, "Logically smooth density estimation," Stanford University, Stanford, CA, Tech. Rep., 1985.

[41] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, vol. 37, pp. 1034–1054, 1991.

[42] G. Lugosi and K. Zeger, "Concept learning using complexity regularization," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 48–54, January 1996.

[43] 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.

[44] P. Chou, T. Lookabaugh, and R. Gray, "Optimal pruning with applications to tree-structure source coding and modeling," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 299–315, 1989.

[45] M. Bohanec and I. Bratko, "Trading accuracy for simplicity in decision trees," *Machine Learning*, vol. 15, pp. 223–250, 1994.

[46] J. Silva, "On optimal signal representation for statistical learning and pattern recognition," Ph.D. dissertation, University of Southern California, http://digitallibrary.usc.edu/assetserver/controller/item/etd-Silva-2450.pdf, December 2008.