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],

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

(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].

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

(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 $Z^n = Z_1, \ldots, Z_n$ 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 $Q$ (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 $Z^n$ and creates a partition of the space, i.e., $\pi_n(Z^n_1) \in Q$. In addition, $\Pi$ needs to satisfy a product bin condition, i.e., $\forall z^n_1 = (z_1, \ldots, z_n) \in \mathbb{R}^{d \cdot n}$ every event $A \in \pi_n(z^n_1)$ 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 $\pi_n(Z^n_1)$, second, to use again the data to estimate $P_{X,Y}$ and $P_X \times P_Y$ restricted to the sigma field $\sigma(\pi_n(Z^n_1))$ and finally, to consider the plug-in technique to get an empirical MI estimate.$^1$

$^1$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 $\mathcal{A}$.
on \((\mathbb{R}^d, \sigma(\pi_n(Z_n^i)))\) [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., \(P_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(Z_i), \forall A \in \mathcal{B}(\mathbb{R}^d),\) hence, the MI estimate is given by \(\hat{I}_n(\pi_n(Z_n^i)) = \sum_{A \in \pi_n(Z_n^i)} P_n(A) \cdot \log \frac{P_n(A \times A_2)}{P_n(A_1 \times \mathbb{R}^d) \cdot P_n(\mathbb{R}^d \times A_2)},\) \(3\),

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 \(\mathcal{L}(T)\) and \(\mathcal{L}(T)\) be the set of internal and terminal nodes of \(T\), respectively, and \(\bar{T} \subseteq T\) be the size of a tree \(T\), given by the cardinality of \(\mathcal{L}(T)\). If \(\bar{T} \subseteq T\) 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 \(\bar{T} \preceq T\).

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

\[\pi_T = \{U_t : t \in \mathcal{L}(T)\} \in \mathcal{B}(\mathbb{R}^d),\]

(4)

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

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

III. The Tree-Structured Partition Scheme

Our TSP scheme uses \(Z_n^i\) to construct a partition of \(\mathbb{R}^d\) in two consecutive stages: a growing phase and a pruning phase. For the growing stage, let \(t_0\) be the root of the tree and \(U_{t_0} = \mathbb{R}^d\). Considering \(Z_n^i = (Z_1, \ldots, Z_n)\) 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

\[\tau(t_0) = H_i(Z_1^i) = \left\{x \in \mathbb{R}^d : x(i) \leq Z^{(i/2)}(i)\right\},\]

(5)

where \(Z^{(1)}(i) < Z^{(2)}(i) < \ldots < Z^{(n)}(i)\) denotes the order statistics of the sample points \(\{Z_1, \ldots, Z_n\}\) projected in the target dimension \(i\). Using \(H_i(Z_1^i)\), \(\mathbb{R}^d\) is divided into two statistically equivalent rectangles with respect to the coordinate dimension \(i\), denoted by \(U_{l(t_o)}\) and \(U_{r(t_o)}\). Realocating 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 \(T_{bn}^{full}(Z_1^n)\), and the associated partition \(\pi_{\bar{T}_{bn}^{full}}(Z_1^n)\), where we guarantee a minimum magnitude for \(P_n\) on the events of \(\sigma(\pi_{\bar{T}_{bn}^{full}}(Z_1^n))\) 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 \(T_{bn}^{full}\) considering its dependency on \(Z_1^n\). First, we consider the following inequality: \(\forall T \ll T_{bn}^{full}, \left|\hat{I}_n(\pi_T(Z_1^n)) - I(X;Y)\right| \leq \left|\hat{I}_n(\pi_T(Z_1^n)) - I(\pi_T(Z_1^n)) + I(X;Y) - \pi_T(Z_1^n)\right|\),

(6)

where \(I(\pi_T(Z_1^n)) = \sum_{A \in \pi_T(Z_1^n)} P(A) \cdot \log \frac{P(A)}{P(A_{\bar{T}} \times \mathbb{R}^d)},\) is the KLD of the true distributions restricted to the sigma field induced by \(\pi_T(Z_1^n)\) [11]. 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,

\[T^n = \arg \min_{T \ll T_{bn}^{full}} -\hat{I}_n(\pi_T(Z_1^n)) + \phi_n(T).\]

(7)

This regularization criterion attempts to find an optimal balance in \(\{T : T \ll T_{bn}^{full}\}\) between the empirical MI (fidelity) and an indicator of complexity for \(\pi_T\) 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].
result presented in the next theorem. Let \( \tilde{I}_n(\pi_T(Z_1^n)) \equiv I_n(\pi(T_1^n)) - \phi_n(|T|) \) be the penalized EMI indicator \( \forall T \ll T_{n}^{\text{full}} \). The next result shows that \( T^n \) 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 \( l \in (0,1/3), \)
- \( (\delta_n) \approx o(1) \) and \( (1/\delta_n) \approx O(e^{l/4}) \),

then \( \forall \delta > 0 \) there exists \( N_\delta(\delta) > 0 \), such that \( \forall n > N_\delta(\delta) \) with probability \( 1 - \delta \) (with respect to \( P \)),

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

The result says two important things. On one hand, it shows that with an arbitrary high probability our penalized indicator \( \tilde{I}_n(\pi_T(Z_1^n)) \) is an underestimation of \( I(\pi_T(Z_1^n)) \), 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 \( \tilde{I}_n(\pi_T(Z_1^n)) \) 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., \( T^n = \arg\min_{T \ll T_{n}^{\text{full}}} \left| I(X; Y) - I(\pi_T(Z_1^n)) \right| + 2 \cdot \phi_n(|T|) \). 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 \( (b_n) \approx (n^{-l}) \) stated in Theorem 2, we have that \( \lim_{n \to \infty} \sup_{k \in \{1, \ldots, |T_{n}^{\text{full}}|\}} \phi_n(k) = 0 \) (the arguments presented in Section VIII-C). Consequently the oracle minimax error bound in (11) is governed by the asymptotic trend of \( \lim_{n \to \infty} \left| I(X; Y) - I(\pi_T(Z_1^n)) \right| \), associated with the approximation goodness (or the asymptotic sufficiency) of the full tree. This result formalizes this idea and proves the density free strong consistency of \( \tilde{I}_n(\pi_T(Z_1^n)) \) and \( I_n(\pi_T(Z_1^n)) \), 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 \( (\delta_n) \) satisfy the condition stated in Theorem 2, then \( \lim_{n \to \infty} \tilde{I}_n(\pi_T(Z_1^n)) = I(X;Y) \) and \( \lim_{n \to \infty} I_n(\pi_T(Z_1^n)) = I(X;Y) \), \( P \)-almost surely.

**VII. FINAL REMARK**

The conditions on \( (b_n) \approx (n^{-l}) \) to ensure that \( T^n \) induces strongly consistent estimates for \( I(X;Y) \) (Theorem 3), match the one stipulated on the full tree, i.e., \( T_{n}^{\text{full}} \), to obtain that \( \tilde{I}_n(\pi_T(Z_1^n)) \) 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 \( Z_1, Z_2, \ldots \) iid realizations with distribution \( P \) in \( \{ \mathbb{R}^d, B(\mathbb{R}^d) \} \). Then, \( \forall \epsilon > 0, \forall n, \)

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

**Proof Theorem 1:**

Concerning the first term on the RHS of (12), \( \mathcal{G}_k \). By definition \( \phi_n(k) = \frac{12}{b_n} \cdot \sqrt{n} \cdot \sqrt{(\ln(8) + \ln(n) - \ln(b_n \cdot b_n) + k \cdot (d + 1) \cdot \ln(2) + d \cdot \ln(n))} \), then considering that \( |T_{b_n}^{full}| \leq (1/b_n) \), it is simple to check that \( (b_n) \approx (n^{-i}) \) for \( i \) in \( (0,1) \). Using Taylor expansion, we get \( \lim_{n \to \infty} \frac{\phi_n}{\phi_n(k)} = 0 \).

**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 \( \phi_n(k) = \frac{12}{b_n \cdot b_n} \cdot \sqrt{n} \cdot \sqrt{(\ln(8) + \ln(n) - \ln(b_n \cdot b_n) + k \cdot (d + 1) \cdot \ln(2) + d \cdot \ln(n))} \), then considering that \( |T_{b_n}^{full}| \leq (1/b_n) \), it is simple to check that \( (b_n) \approx (n^{-i}) \) with \( i \in (0,1/3) \) and \( (1/b_n) \) being \( O(e^{-n^{1/3}}) \) are the weakest set of sufficient conditions to obtain that

\[
\lim_{n \to \infty} \sup_{k \in \mathcal{G}_k} \phi_n(k) = 0.
\]
be the $k$-typical set, well defined for all $n$ such that $k \leq T_{b_n}^{f\text{ull}}$. From Corollary 1, if $\phi_n(k) \in (0, 3)$, then $P(S^{n,k}) > 1 - b_n\delta_n$. Consequently from (18), there exists $N_c > 0$ such that $\forall k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}$ and $\forall n > N_c$, $P(S^{n,k}) > 1 - b_n\delta_n$. Hence, defining $S^n_k = \bigcap_{k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}} S^{n,k}$, we have that $P(S^n_k) > 1 - \delta_n$, $\forall n > N_c$. By definition, if $z^n_1 \in S^n_k$, then $\sup_{T \in \mathcal{G}_{b_n}^k} \left| I_n(\pi_T(z^n_1)) - I(\pi_T(z^n_1)) \right| \leq \phi_n(k), \quad \forall k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}$, which also implies that [13],

$$ \sup_{T \in \mathcal{G}_{b_n}^k} \left| I_n(\pi_T(z^n_1)) - I(\pi_T(z^n_1)) \right| \leq \phi_n(k), \quad \forall k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}. \quad (19) $$

Then for an arbitrary $z^n_1 \in S^n_k$

$$ -I_n(\pi_T(z^n_1)) = -I_n(\pi_T(z^n_1)) + \phi_n(T^n) $$

$$ \leq -I_n(\pi_T(z^n_1)) + 2 \phi_n(k), \quad \forall k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}, $$

$$ \leq -I(\pi_T(z^n_1)) + 2 \phi_n(k), \quad \forall k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}, $$

where $T^n_k = \arg \max_{T \in \mathcal{G}_{b_n}^k} I(\pi_T(z^n_1))$ is the oracle solution that maximizes the MI on $\mathcal{G}_{b_n}^k$. Also it is clear that $\forall z^n_1 \in S^n_k$, $I_n(\pi_T(z^n_1)) = I_n(\pi_T(z^n_1)) - \phi_n(T^n) \leq I(\pi_T(z^n_1)) \leq I(X;Y)$, and consequently we have that, $0 \leq I(X;Y) - I_n(\pi_T(z^n_1)) \leq \min_{k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}} (I(X;Y) - I(\pi_T(z^n_1))) + \phi_n(k). \quad (20)$

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

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 $(b_n) \approx (n^{-i})$ for some $i \in (0, 1/3)$, then

$$ \lim_{n \to \infty} I(\pi_{T_{b_n}^{f\text{ull}}}(Z^n_1)) = I(X;Y), \quad \mathbb{P}-\text{almost surely.} \quad \text{(The argument is presented in [12].)} $$

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

$$ \lim_{n \to \infty} \left| I_n(\pi_T(z^n_1)) - I(\pi_T(z^n_1)) \right| = 0 \quad \text{P-almost surely.} $$

**Lemma 3:** Under the setting of Theorem 3, if $(b_n) \approx o(n^{-i})$ with $i \in (0, 1/3), (b_n) = o(1)$ and $(1/b_n)$ is $O(e^{n^{1/3}})$, then $\forall \epsilon > 0$ there exits $N_c(\epsilon)$ such that $\forall n > N_c$ and $\forall k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}$,

$$ \mathbb{P} \left( \sup_{T \in \mathcal{G}_{b_n}^k} I(\pi_T(z^n_1)) - I(\pi_T(z^n_1)) > \epsilon \right) \leq \exp \left\{ -\frac{n}{8} \left( \frac{e b_n}{24} \right)^2 \right\} + 8(2^{d+1}n^{d+1}) \exp \left\{ -\frac{n}{8} \left( \frac{e b_n}{48} \right)^2 \right\}, $$

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

$$ \lim_{n \to \infty} I(\pi_{T_{b_n}^{f\text{ull}}}(Z^n_1)) = I(\pi_T(Z^n_1)). \quad (22) $$

**Proof of Theorem 3:** The proof comes from

$$ I(X;Y) - I_n(\pi_T(Z^n_1)) \leq I(X;Y) - I(\pi_{T_{b_n}^{f\text{ull}}}(Z^n_1)) + I(\pi_{T_{b_n}^{f\text{ull}}}(Z^n_1)) - I(\pi_T(Z^n_1)) + I(\pi_T(Z^n_1)) - I(\pi_T(Z^n_1))), $$

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 $I_n(\pi_T(Z^n_1))$ as, by definition,

$$ \lim_{n \to \infty} I_n(\pi_T(Z^n_1)) - I(\pi_T(Z^n_1)) \leq \lim_{n \to \infty} \sup_{k \in \{1, \ldots, T_{b_n}^{f\text{ull}}\}} \phi_n(k) = 0. $$

IX. Acknowledgment

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

**References**


