Information Divergence Estimation based on Data-Dependent Partitions

Jorge Silva[a], Shrikanth S. Narayanan[b]

[a] *University of Chile, Department of Electrical Engineering, Av. Tupper 2007, Santiago, 412-3, Chile.*
[b] *University of Southern California, Department of Electrical Engineering, Los Angeles, CA 90089 2564, USA.*

Abstract

This work studies the problem of information divergence estimation based on data-dependent partitions. A histogram-based data-dependent estimate is proposed adopting a version of *Barron*-type histogram-based estimate. The main result is the stipulation of sufficient conditions on the partition scheme to make the estimate strongly consistent. Furthermore, when the distributions are equipped with density functions in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we obtain sufficient conditions that guarantee a density-free strongly consistent information divergence estimate. In this context, the result is presented for two emblematic partition schemes: the statistically equivalent blocks (*Gessaman*'s data-driven partition) and data-dependent tree-structured vector quantization (TSVQ).

Key words: Information divergence estimation, data-dependent partitions, Barron density estimate, Vapnik-Chervonenkis inequality, statistically equivalent blocks, tree-structured partitions.

1. Introduction

Let P and Q be probability measures defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, the finite dimensional Euclidean space equipped with the Borel sigma field, then the information divergence of P with respect to Q is expressed by (see, eg., Kullback (1958); Gray (1990)),

$$D(P\|Q) = \sup_{\pi \in Q} \sum_{A \in \pi} P(A) \cdot \log \frac{P(A)}{Q(A)}, \tag{1}$$

where Q denotes the collection of finite measurable partitions of $\mathbb{R}^d$. For this quantity to be finite, it is necessary that $P \ll Q$ (Kullback, 1958), which makes $\frac{\partial P}{\partial Q}(x)$ the *Radon-Nicodym* (RN) derivative of P with respect to Q well defined. Considering the important case when P and Q are absolutely continuous with respect to the Lebesgue measure λ, i.e., $P \ll \lambda$ and $Q \ll \lambda$, it is sometime convenient to use the following expression (see, Gray (1990)),

$$D(P\|Q) = \int_{\mathbb{R}^d} p(x) \cdot \log \frac{p(x)}{q(x)} \lambda(\partial x), \tag{2}$$

where $p(x) = \frac{\partial P}{\partial \lambda}(x)$ and $q(x) = \frac{\partial Q}{\partial \lambda}(x)$ are the density functions of P and Q, respectively. The information divergence, also known Kullback-Leibler (KL) divergence or relative entropy, is a well known fundamental quantity in statistics and information theory (Kullback, 1958; Cover and Thomas, 1991; Gray, 1990). In statistics, KL divergence expresses the average information per observation to discriminate between two probabilistic models (Kullback, 1958). In large deviations, it characterizes the rate function, which reflects the exponential decay of convergence of empirical measures to their probabilities, *Sanov's Theorem* (see, eg., den Hollander (2000)), and the rate of decay of the probability of error in a binary hypothesis testing problem, *Stein's Lemma* (see Cover and Thomas (1991)).

On the application side, mainly because of its role as a discriminative measure (Kullback, 1958), the information divergence has found wide use in statistical learning-decision problems. It has been adopted as an optimality criterion

Email addresses: jorgesil@ing.uchile.cl (Jorge Silva), shri@sipi.usc.edu (Shrikanth S. Narayanan)

for parameter re-estimation (Singer and Warmuth, 1996; Juang and Rabiner, 1985), as a similarity measure for modeling clustering and indexing (Vasconcelos, 2004b, 2000; Do and Vetterli, 2002), as an indicator to quantify the effect of estimation error in a Bayes decision approach (Vasconcelos, 2004a; Silva and Narayanan, 2009), to quantify the approximation error of vector quantization in statistical hypothesis testing (Jain et al., 2002; Poor and Tomas, 1977) and as fidelity indicator for feature selection and feature extraction (Saito and Coifman, 1994; Novovicova et al., 1996). These learning scenarios do not have access to the distributions and consequently they rely on empirical data to estimate this quantity. A standard setting considers $X_1, ..X_n$ and $Y_1, .., Y_n$ to be independent and identically distributed (i.i.d.) realizations of P and Q, respectively. Then the problem becomes finding a distribution-free function or estimator $\hat{D}(\cdot)$ from $\mathbb{R}^{d\cdot n} \times \mathbb{R}^{d\cdot n}$ to $\mathbb{R}$ such that $\hat{D}(X_1, ..X_n, Y_1, .., Y_n)$ converges to $D(P\|Q)$ almost surely as n tends to infinity (strong consistency).

In this regard, the closely related problem of differential entropy estimation has been systematically studied for distributions equipped with densities, adopting for instance non-parametric histogram-based, kernel-based and nearest-neighbor techniques. In these settings, the conditions for density-free strong consistency are well understood. An excellent review can be found in Beirlant et al. (1997) and some recent contributions in Darbellay and Vajda (1999); Paninski (2003, 2008). Another closely related problem is the non-parametric density estimation, as the KL divergence is a functional of two probability measures. In this context the classical problem of strong consistency in L_1 sense is well understood (Lugosi and Nobel, 1996; Devroye and Györfi, 1985). More recent work on non-parametric distribution estimation considers consistency under stronger notions (Györfi and van der Meulen, 1994; Barron et al., 1992; Györfi et al., 1998; Berlinet et al., 1998). In particular the seminal work of Barron et al. (1992) proposed variations of classical histogram-based density estimates to achieve consistency in two types of information divergences, motivated by the learning problem on universal lossy compression. This approach has been extended by Györfi et al. (1998) and Berlinet et al. (1998) for the problem of consistency in χ^2-divergence and in Csiszár's ϕ-divergence (where the information divergence is a particular case), respectively. Although the two aforementioned research lines have been systematically explored, to the best of our knowledge, their estimates and results do not extend directly to the consistent estimation of information divergence. The main reason is that the learning setting here is different. On the one hand, we need to consider finite samples from the two distributions, P and Q, while on the other, we need to infer the distributions from the data in a way that is appropriate to the particular nature of the divergence information functional. However because of their inherent connections, the extensions of techniques and results from distribution and differential entropy estimation to KL divergence estimation are important directions to explore.

In that spirit, there have been some recent contributions, in particular for P and Q defined in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and both absolutely continuous with respect to the Lebesgue measure λ. The first important reference in this regard is from Wang et al. (2005), who proposed a histogram-based divergence estimation based on partitioning the space in statistically equivalent intervals. Sufficient conditions on the proposed data-driven partition were stipulated to guarantee strong consistency. Silva and Narayanan (2007) took this direction a step further finding consistency conditions for a general family of data-driven partition schemes. The main limitation of these two works is that they are only valid when the sample points of P and Q are taken to infinity in a specific order, one after the other, which limits their applicability. Alternatively, Nguyen et al. (2007) proposed a variational approach to estimate the divergence (see, Gray (1990); den Hollander (2000)). Under certain approximation assumptions and smoothness condition on the likelihood-ratio, strong consistency and asymptotic rate of convergence for the proposed estimate were obtained. More recently, Wang et al. (2009) proposed nearest-neighbor techniques, where mean-square consistency was the main focus of analysis.

In this work we present contributions in the area of histogram-based information divergence estimation, in particular studying data-driven partitions schemes (Lugosi and Nobel, 1996; Nobel, 1996; Devroye et al., 1996; Darbellay and Vajda, 1999). We have significantly improved the initial findings in (Wang et al., 2005; Silva and Narayanan, 2007). We reformulate the problem, propose new estimates and results to address properly the case when the samples of P and Q jointly tend to infinity, and furthermore, report new practical implications by getting concrete density-free KL divergence estimates from previously unexplored multivariate data-driven partition schemes.

Specifically in Section 3, we present the general histogram-based estimation scheme. This scheme quantizes the space function of the data and constructs a version of the Barron-type of histogram-based density estimate (Barron et al., 1992) as a way to approximate $\frac{\delta P}{\delta Q}(x)$, which can be considered the sufficient statistics for the problem. Then assuming that $D(P\|Q) < \infty$, Theorem 4 in Section 5 characterizes sufficient conditions on the partitions scheme to make the estimate strongly consistent. This result does not require P and Q to be absolutely continuous with respect

to λ, and furthermore, it is valid for distributions defined on a general measurable space $(\mathcal{X}, \mathcal{S})$. Concerning the approximation error presented in Section 4, we adopt Csiszár's notion of *asymptotically sufficient partitions* (Csiszár, 1967, 1973), and when $P \ll \lambda$ and $Q \ll \lambda$, Theorem 2 presents a condition for this error to vanish based on a *shrinking cell* property for data-driven partitions (Lugosi and Nobel, 1996; Breiman et al., 1984; Devroye et al., 1996). For the estimation error, in Section 5, we use the *Vapnik-Chervonenkis* (VC) inequality (Vapnik and Chervonenkis, 1971; Vapnik, 1998) (see also Devroye et al. (1996); Lugosi and Nobel (1996)) and characterize a concentration result on the empirical distributions, Lemma 3, that makes this error tend to zero as n tend to infinity with probability one.

In the second part of this work, we explore applications of our main result. In Section 7 consistency is demonstrated for multivariate statistically equivalent blocks — *Gessaman's* data-dependent partition (Gessaman, 1970), while Section 8 shows equivalent results for tree-structured vector quantizations (TSVQ) (Devroye et al., 1996; Breiman et al., 1984; Nobel, 2002). Importantly in both settings, a range of pamarametric values are characterized to obtain a family of density-free consistent estimates. The main challenge faced in deriving these results, is to prove the adopted *shrinking cell* condition, which is achieved from the adaptive nature of the data-driven partition schemes. Finally, some of the proofs and derivations are organized in the appendix.

2. Preliminaries

This section provides notation and key results used for the rest of paper.

2.1. Complexity Notions for Partitions

Let $(\mathcal{X}, \mathcal{S})$ be a measurable space, and let us denote by $\mathcal{Q}$ the collection of finite measurable partitions for $\mathcal{X}$. Considering $\mathcal{A} \subset \mathcal{Q}$, the *maximum cell count* of $\mathcal{A}$ is given by (see Devroye et al. (1996))

$$\mathcal{M}(\mathcal{A}) = \sup_{\pi \in \mathcal{A}} |\pi| \,, \tag{3}$$

where $|\pi|$ denotes the number of cells of π. In addition, a notion of combinatorial complexity for $\mathcal{A}$ can be introduced (Lugosi and Nobel, 1996; Devroye et al., 1996). Let us consider a finite length sequence $x_1^n = (x_1, .., x_n) \in \mathcal{X}^n$, and the induced set by $\{x_1, .., x_n\}$, then we can define $\Delta(\mathcal{A}, x_1, .., x_n)$ as the number of different partitions of $\{x_1, .., x_n\}$ induced by the elements of $\mathcal{A}$, i.e.,

$$\Delta(\mathcal{A}, x_1, .., x_n) = |\{\{x_1, .., x_n\} \cap \pi : \pi \in \mathcal{A}\}| \,, \tag{4}$$

where $\{x_1, .., x_n\} \cap \pi$ is a short hand for $\{\{x_1, .., x_n\} \cap A : A \in \pi\}$. Then the *growth function* of $\mathcal{A}$ is given by (Lugosi and Nobel, 1996)

$$\Delta_n^*(\mathcal{A}) = \max_{x_1^n \in \mathcal{X}^n} \Delta(\mathcal{A}, x_1, .., x_n). \tag{5}$$

2.2. Partition Schemes

A *n-sample partition rule* π_n is a mapping from $\mathcal{X}^n$ to the space of finite-measurable partitions for $\mathcal{X}$, i.e., $\mathcal{Q}$, where a *partition scheme* for $\mathcal{X}$ is the countable collection of n-sample partitions rules $\Pi = \{\pi_1, \pi_2, ...\}$. Let Π be an arbitrary partition scheme for $\mathcal{X}$, then for every partition rule $\pi_n \in \Pi$ we can define its associated collection of measurable partitions (Lugosi and Nobel, 1996) by

$$\mathcal{A}_n = \{\pi_n(x_1, .., x_n) : (x_1, .., x_n) \in \mathcal{X}^n\} \,. \tag{6}$$

In this context, for a given n-sample partition rule π_n and a sequence $(x_1, .., x_n) \in \mathcal{X}^n$, $\pi_n(x|x_1, .., x_n)$ denotes the mapping from any point $x \in \mathcal{X}$ to its unique cell in $\pi_n(x_1, .., x_n)$, such that $x \in \pi_n(x|x_1, .., x_n)$.

2.3. Vapnik-Chervonenkis Inequalities

Let $X_1, X_2, .., X_n$ be i.i.d. realizations of a random variable with values in $\mathcal{X}$ and with probability measure P on $(\mathcal{X}, \mathcal{S})$. Let $\mathcal{A}$ be a collection of measurable partitions for $\mathcal{X}$, where $\forall \pi \in \mathcal{A}$, $\forall B \in \pi$ we can obtain the classical *empirical distribution* by

$$P_n(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_B(X_i), \tag{7}$$

with $\mathbb{I}_B(x)$ being the indicator function of the set B. In this context, the following concentration inequality can be stated.

LEMMA 1. *(Lugosi and Nobel, 1996)* $\forall n \in \mathbb{N}$, $\forall \epsilon > 0$*,*

$$\mathbb{P}\left(\sup_{\pi \in \mathcal{A}} \sum_{A \in \pi} |P_n(A) - P(A)| > \epsilon \right) \leq 4\Delta^*_{2n}(\mathcal{A}) 2^{\mathcal{M}(\mathcal{A})} \exp^{-\frac{n\epsilon^2}{32}},$$

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

The following is a simple extension for the case of mixture distributions.

LEMMA 2. *Let* $X_1, .., X_n$ *and* $Y_1, .., Y_n$ *be i.i.d. realizations driven by* P *and* Q *respectively in* $(\mathcal{X}, \mathcal{S})$ *and inducing the empirical distributions* P_n *and* Q_n*, respectively. Then* $\forall a \in [0, 1]$*,* $\forall \epsilon > 0$ *and* $\forall n > 0$*,*

$$\mathbb{P}\left(\sup_{\pi \in \mathcal{A}} \sum_{A \in \pi} \left| \mu^a_n(A) - \mu^a(A) \right| > \epsilon \right) \leq 8\Delta^*_{2n}(\mathcal{A}) 2^{\mathcal{M}(\mathcal{A})} \exp^{-\frac{n\epsilon^2}{128}},$$

where $\mu^a(A) = (1 - a) \cdot P(A) + a \cdot Q(A)$ *and* $\mu^a_n(A) = (1 - a) \cdot P_n(A) + a \cdot Q_n(A)$ *are the mixing and empirical mixing distributions. (Derivation presented in Appendix A).*

These results are versions of the celebrated *Vapnik-Chervonenkis inequality* (Vapnik and Chervonenkis, 1971; Vapnik, 1998). These inequalities bound the deviation of the empirical distribution with respect to the probability, in the total variational distance sense, uniformly in the collection of partitions $\mathcal{A}$. Remarkably these bounds are distribution free and functions of the aforementioned complexity notions for $\mathcal{A}$.

2.4. Asymptotic Relationships

Let $(a_n)_{n \in \mathbb{N}}$ and $(b_n)_{n \in \mathbb{N}}$ be two sequences of non-negative real numbers. We say that $(a_n)_{n \in \mathbb{N}}$ dominates $(b_n)_{n \in \mathbb{N}}$, denoted by $(b_n) \preceq (a_n)$ (or alternatively (b_n) is $O(a_n)$), if there exists $C > 0$ and $k \in \mathbb{N}$ such that $b_n \leq C \cdot a_n$ $\forall n \geq k$. We say that $(b_n)_{n \in \mathbb{N}}$ and $(a_n)_{n \in \mathbb{N}}$ (both strictly positive) are asymptotically equivalent, denoted by $(b_n) \approx (a_n)$, if there exits $C > 0$ such that $\lim_{n \to \infty} \frac{a_n}{b_n} = C$, and on the other hand, we say that (a_n) is $o(b_n)$ if $\lim_{n \to \infty} \frac{a_n}{b_n} = 0$.

3. The Data-Driven Estimator

Let P and Q be probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $D(P\|Q) < \infty$. For the learning problem let us consider $X_1, .., X_n$ and $Y_1, .., Y_n$ i.i.d. realizations of random variables in $\mathbb{R}^d$ and driven by P and Q, respectively, and let $\Pi = \{\pi_1, \pi_2, \cdots\}$ be a data-driven partition scheme for $\mathbb{R}^d$. We propose a plug-in histogram-based estimate for the information divergence of the form,

$$D_{\pi_n(Y_1,..,Y_n)}(P^*_n \| Q_n) \equiv \sum_{A \in \pi_n(Y_1,..,Y_n)} P^*_n(A) \cdot \log \frac{P^*_n(A)}{Q_n(A)}, \tag{8}$$

where P^*_n is a *Barron type of empirical measure* (Barron et al., 1992) given by,

$$P^*_n(A) \equiv (1 - a_n) \cdot P_n(A) + a_n \cdot Q_n(A), \tag{9}$$

with $(a_n)_{n\in\mathbb{N}}$ a real sequence with values in $[0,1]$, and P_n and Q_n the empirical measures in (7) induced by $X_1,..,X_n$ and $Y_1,..,Y_n$, respectively, and restricted to the sub-sigma field $\sigma(\pi_n(Y_1,..,Y_n)) \subset \mathcal{B}(\mathbb{R}^d)$[1]. Note that the role of the data-driven partition is to restrict the domain where we construct the empirical distributions, and consequently the sub-sigma field where the information divergence is defined (see, eg., Gray (1990)).

Considering that $D_{\pi_n(Y_1,..,Y_n)}(P_n^*\|Q_n)$ is a measurable function of $X_1,..,X_n$ and $Y_1,..,Y_n$, we are interested in studying the strong — with respect to empirical process $\{(X_n,Y_n), n\in\mathbb{N}\}$– consistency of $D_{\pi_n(Y_1,..,Y_n)}(P_n^*\|Q_n)$. The proposed construction was based on the analysis of the following estimation-approximation error inequality,

$$\left|D_{\pi_n(Y_1,..,Y_n)}(P_n^*\|Q_n) - D(P\|Q)\right| \leq \left|D_{\pi_n(Y_1,..,Y_n)}(P_n^*\|Q_n) - D_{\pi_n(Y_1,..,Y_n)}(\tilde{P}_n\|Q)\right| \tag{10}$$

$$+ \left|D_{\pi_n(Y_1,..,Y_n)}(\tilde{P}_n\|Q) - D(P\|Q)\right|, \tag{11}$$

where $\tilde{P}_n(A) \equiv (1-a_n)\cdot P(A) + a_n \cdot Q(A)$, $\forall A \in \pi_n(Y_1^n)$. Concerning the estimation error in (10), we use two techniques to bound the deviation of the divergence functional in (8) when considering empirical measures. The first is a condition on the partition scheme Π, where we impose that $Q_n(A) \geq \frac{k_n}{n}$, $\forall A \in \sigma(\pi_n(Y_1,..,Y_n))$, (k_n) representing the critical empirical mass. The second is due to Barron et al. (1992) which is a smoothing technique (9) for estimating the *Radon-Nicodym* derivative $\frac{\partial P}{\partial Q}(x)$ when $P \ll Q$, which is given in our setting considering that $D(P\|Q) < \infty$. Both design sequences (a_n) and (k_n) are strictly positive and provide a way of ensuring a minimum probability mass for both P_n^* and Q_n in $(\mathbb{R}^d, \sigma(\pi_n(Y_1^n)))$, which in conjunction with the distribution free concentration inequalities in Section 2.3, offer the key elements to bound the estimation error. Concerning the approximation error in (11), we have chosen the data-dependent partition as only a function of the i.i.d. realizations associated with the reference measure Q. This partial information choice is justified by the fact that $P \ll Q$ (details are presented in Section 4).

For the following sections the process distributions of $Y_1, Y_2, \cdots$ and the joint process $(X_1,Y_1),(X_2,Y_2),\cdots$ will be denoted by $\mathbb{Q}$ and $\mathbb{P}$, respectively, and their marginal probabilities restricted to finite blocks, i.e. $Y_1^n \equiv (Y_1,..,Y_n)$ and $Z_1^n \equiv ((X_1,Y_1),...,(X_n,Y_n))$, by $\mathbb{Q}^n$ and $\mathbb{P}^n$, respectively.

4. Approximation Error Analysis

In this section we study the approximation quality of data-dependent partition schemes Π for the information divergence estimation (8). This notion is strongly related with the concept of *asymptotically sufficient partition* developed by Csiszár (1973, 1967) and recent extensions presented by Vajda (2002), Liese et al. (2006) and Berlinet and Vajda (2005) (see also Liese and Vajda (1987)). The main difference here is that we are dealing with data-dependent partitions driven by an empirical process, instead of the deterministic sequence of partitions considered in these previous works. However, these notions extend naturally to our domain.

Definition 1. *Let P, Q be probability measures in $(\mathcal{X},\mathcal{S})$ and Π be a partition scheme of $\mathcal{X}$ driven by the process $Y_1, Y_2, \cdots$ with distribution $\mathbb{Q}$. Π is said to be* simultaneously (P,Q)-approximating *with respect to $\mathbb{Q}$ if, $\forall\delta > 0$ and for any measurable partition $\pi = \{A_1,..,A_r\} \in \mathcal{Q}$, there exists a sequence of finite measurable partitions $\pi_n^* = \{A_{n,1},..,A_{n,r}\} \subset \sigma(\pi_n(Y_1^n))$, such that*

$$\limsup_{n\to\infty} \sup_{i=1,..,r} \left|P(A_i) - P(A_{n,i})\right| < \delta \;\; \textit{and} \;\; \limsup_{n\to\infty} \sup_{i=1,..,r} \left|Q(A_i) - Q(A_{n,i})\right| < \delta,$$

$\mathbb{Q}$-almost surely.

Adopting this definition in our histogram-based construction in (8), we have the following theorem.

THEOREM 1. *Let P, Q be probability measures in $(\mathcal{X},\mathcal{S})$ such that $D(P\|Q) < \infty$. Let Π be a partition scheme driven by the i.i.d. realizations $Y_1, Y_2, \cdots$ of the reference measure Q. If Π is simultaneously (P,Q)-approximating with respect to $\mathbb{Q}$ (the process distribution of $Y_1, Y_2 \cdots$) and (a_n) is $o(1)$, then,*

$$\lim_{n\to\infty} \left| \sum_{A\in\pi_n(Y_1^n)} \tilde{P}_n(A)\cdot\log\frac{\tilde{P}_n(A)}{Q(A)} - D(P\|Q) \right| = 0, \tag{12}$$

$\mathbb{Q}$-almost surely. (Proof in Appendix B.)

[1] $\sigma(\pi)$ denotes de smallest sigma field containing the element of $\pi \subset \mathcal{B}(\mathbb{R}^d)$.

The verification of the (P, Q)-approximating condition for Π is in practice a difficult problem. A more concrete condition can be stated for the case of $(\mathcal{X}, \mathcal{S}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (and in general for complete and separable spaces equipped with a norm or distance), based on what is called a *shrinking cell* condition[2] for data-dependent partition schemes (see, eg., Devroye et al. (1996)). Let us first introduce the following concept. For any $A \in \mathcal{B}(\mathbb{R}^d)$, we define its *diameter* by

$$diam(A) = \sup_{x,y \in A} \|x - y\|, \tag{13}$$

where $\|\cdot\|$ refers to the Euclidian norm in $\mathbb{R}^d$.

THEOREM 2. *Let $\Pi = \{\pi_1, \cdots\}$ be a partition scheme for $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and P and Q be probability measures in $\mathcal{P}_\lambda(\mathbb{R}^d)$ (the space of distributions absolutely continuous with respect to the Lebesgue measure λ). Considering $P \ll Q$ and Π driven by the i.i.d. realizations $Y_1, Y_2, \cdots$ with $Y_i \sim Q$, the scheme is simultaneously (P, Q)-approximating with respect to $\mathbb{Q}$ if, $\forall \gamma > 0$,*

$$\lim_{n \to \infty} Q\left(\left\{x \in \mathbb{R}^d : diam(\pi_n(x|Y_1, .., Y_n)) > \gamma\right\}\right) = 0, \tag{14}$$

$\mathbb{Q}$-almost surely. (The proof is presented in Appendix C.)

Remark 1. *In the context of a deterministic sequence of partitions, Liese et al. (2006, Theorem 6) propose a condition to check their notion of asymptotically sufficiency with respect to the ϕ-divergence, Csiszár (1973, 1967), for every pair of probabilities (P, Q) dominated by the Lebesgue measure. This is based on the notion of L_∞-covering of the partition sequence with decreasing radius. This result could be naturally extended in our domain, however, the proposed shrinking cell condition in (14) is weaker than the mentioned L_∞-covering condition.*

To conclude this section, in the context of $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, if the partition scheme has a *product rectangle structure*, we can extend a result of asymptotic sufficient partition developed by Vajda (2002) (for completeness, see also Liese et al. (2006), Liese and Vajda (1987) and Berlinet and Vajda (2005)).

Definition 2. *A partition scheme $\Pi = \{\pi_1, \pi_2, \cdots\}$ is a product rectangle partition, if for all $n > 0$, for all $y_1^n \in \mathbb{R}^{d \cdot n}$, the partition rule π_n can be written as,*

$$\pi_n(y_1^n) = \pi_n^{(1)}(y_1^n) \otimes \cdots \otimes \pi_n^{(d)}(y_1^n),$$

where $\pi_n^{(j)}(y_1^n)$ is a partition rule that dichotomize $\mathbb{R}$ in terms of intervals, for all $j \in \{1, \ldots, d\}$.

Definition 3. *(Vajda (2002)) Let $\{\mathcal{P}_1, \mathcal{P}_2, \cdots\}$ be an indexed sequence of finite measurable partitions of $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and μ a sigma finite measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We say that the partition sequence is asymptotically μ-sufficient, if for all $x \in \mathbb{R}^d$*

$$\lim_{n \to \infty} \mu(\mathcal{P}_n(x)) = 0,$$

with $\mathcal{P}_n(x)$ denoting the set in $\mathcal{P}_n$ that contains x.

For our histogram-based construction and problem setting we have the following result.

THEOREM 3. *(Vajda (2002)) Under the general setting and assumptions of Theorem 2, let $Q^{(1)}, \ldots, Q^{(d)}$ denote the marginals probability measure of Q on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If Π is a product rectangle partition scheme (Definition 2) and for every $j \in \{1, \ldots, d\}$, $\left\{\pi_1^{(j)}(Y_1), \pi_2^{(j)}(Y_1^2), \cdots\right\}$ is asymptotically $Q^{(j)}$-sufficient partition of $\mathbb{R}$ (Definition 3) $\mathbb{Q}$-almost surely, (a_n) is $o(1)$, and $D(P\|Q) < \infty$, then*

$$\lim_{n \to \infty} \left| \sum_{A \in \pi_n(Y_1^n)} \tilde{P}_n(A) \cdot \log \frac{\tilde{P}_n(A)}{Q(A)} - D(P\|Q) \right| = 0, \tag{15}$$

$\mathbb{Q}$-almost surely. (Proof derives from Vajda (2002) (Theorem 5).)

[2]This shrinking cell condition was proposed by Lugosi *et al.* (Lugosi and Nobel, 1996) for controlling approximation error in histogram-based density estimation. Also Csiszár (Csiszár, 1973, 1967; Liese and Vajda, 1987) presented a similar sufficient condition for his notion of approximating sequence of partitions in separable metric spaces.

For the case of product rectangle partition scheme the sufficient conditions impose in Theorem 3 are weaker than the shrinking cell condition stipulated in Theorem 2. However, the product structural constraint provides an important limitation when we want to partition the space function of the data, as pointed out by Darbellay and Vajda (1999). In fact, (14) is the condition we use to control the approximation error in our practical settings in Section 6.

Remark 2. *The shrinking cell condition for Q in (14) implies that the same asymptotic property is satisfied for the measure P (Halmos, 1950; Varadhan, 2001). This provides the justification for our choice of data-dependent construction in (8) which is exclusively driven by i.i.d. samples of Q. In practice this choice ends up to be sufficient to obtain (14), as demonstrated for two concrete schemes later in this work.*

5. The Main Result

THEOREM 4. *Let P and Q be probability measures in $(\mathcal{X}, \mathcal{S})$ such that $D(P\|Q) < \infty$. Let $X_1, .., X_n$ and $Y_1, .., Y_n$ be i.i.d. realizations of P and Q, respectively, and $\Pi = \{\pi_1, \pi_2, ...\}$ a partition scheme with associated sequence of measurable partitions $\mathcal{A}_1, \mathcal{A}_2, \cdots$. If for some $l \in (0, 1)$, there exists $p \in (0, l/2)$, $\tau \in (0, l - 2p]$ and $(k_n)_{n\in\mathbb{N}}$ a non-negative sequence such that*

a) $(k_n) \geq (n^{0.5+l/2})$, $(a_n) \geq (n^{-p})$ *and* $(a_n) = o(1)$,

and on Π we impose that:

b) $\lim_{n\to\infty} n^{-\tau}\mathcal{M}(\mathcal{A}_n) = 0$,

c) $\lim_{n\to\infty} n^{-\tau} \log \Delta_n^*(\mathcal{A}_n) = 0$,

d) $\forall n \in \mathbb{N}$, $\forall (y_1, .., y_n) \in \mathcal{X}^n$, $\inf_{A\in\pi_n(y_1^n)} Q_n(A) \geq \frac{k_n}{n}$,

e) *and Π is simultaneously (P, Q)-approximating with respect to $\mathbb{Q}$,*

then

$$\lim_{n\to\infty} D_{\pi_n(Y_1^n)}(P_n^*\|Q_n) = D(P\|Q),$$

$\mathbb{P}$*-almost surely.*

There are two sets of conditions stipulated in this result: conditions **a)**, **b)**, **c)**, **d)** that account for the estimation error, while the asymptotically sufficient nature of Π in **e)** in conjunction with **a)** accounts for the approximation error. Concerning the estimation error, we use the following result.

LEMMA 3. *Under the learning setting and conditions of* Theorem *4 (in particular* **a)**, **b)**, **c)** *and* **d)**),

$$\lim_{n\to\infty} \sup_{A\in\pi_n(Y_1^n)} \left| \frac{Q(A)}{Q_n(A)} - 1 \right| = 0, \tag{16}$$

$$\lim_{n\to\infty} \sup_{A\in\pi_n(Y_1^n)} \left| \frac{\tilde{P}_n(A)}{P_n^*(A)} - 1 \right| = 0, \tag{17}$$

$\mathbb{P}$*-almost surely. (Derivation in Appendix D.)*

Proof: The approximation error converges to zero $\mathbb{P}$-almost surely from Theorem 1. For the estimation error in (10), we use the following inequality:

$$\left| D_{\pi_n(Y_1^n)}(P_n^*\|Q_n) - D_{\pi_n(Y_1^n)}(\tilde{P}_n\|Q) \right| \leq \left| \sum_{A\in\pi_n(Y_1^n)} P_n^*(A) \cdot \log P_n^*(A) - \sum_{A\in\pi_n(Y_1^n)} \tilde{P}_n(A) \cdot \log \tilde{P}_n(A) \right| \tag{18}$$

$$+ \left| \sum_{A\in\pi_n(Y_1^n)} \tilde{P}_n(A) \cdot \log Q(A) - \sum_{A\in\pi_n(Y_1^n)} P_n^*(A) \cdot \log Q_n(A) \right|. \tag{19}$$

The expression in the right hand side (RHS) of (18) is upper bounded by,

$$\sum_{A\in\pi_n(Y_1^n)} \left|P_n^*(A) - \tilde{P}_n(A)\right| \cdot \log \frac{1}{P_n^*(A)} + \sum_{A\in\pi_n(Y_1^n)} \left|\log P_n^*(A) - \log \tilde{P}_n(A)\right| \cdot \tilde{P}_n(A) \tag{20}$$

$$\leq \log \frac{1}{a_n \cdot b_n} \cdot \sup_{\pi\in\mathcal{A}_n} \sum_{A\in\pi} \left|P_n^*(A) - \tilde{P}_n(A)\right| + \sup_{A\in\pi_n(Y_1^n)} \left|\log P_n^*(A) - \log \tilde{P}_n(A)\right|, \tag{21}$$

where $b_n \equiv \frac{k_n}{n}$. (20) is from triangular inequality and (21) from the construction of P_n^* on $\pi_n(Y_1^n)$ ($P_n^*(A) \geq a_n \cdot b_n$, for all $A \in \pi_n(Y_1^n)$) and the fact that by definition $\pi_n(Y_1^n) \in \mathcal{A}_n$. Without loss of generality we assume that $a_n < 1$ and $b_n < 1$, $\forall n > 0$. From Lemma 2 and conditions **a)**, **b)** and **c)**, it is simple to show that, $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\log \frac{1}{a_n \cdot b_n} \cdot \sup_{\pi\in\mathcal{A}_n} \sum_{A\in\pi} \left|P_n^*(A) - \tilde{P}_n(A)\right| > \epsilon\right) < 0,$$

then from *Borel-Cantelli lemma* the first term of (21) tends to zero $\mathbb{P}$-almost surely. Concerning the second term in (21), from (17) $\lim_{n\to\infty} \sup_{A\in\pi_n(Y_1^n)} \frac{\tilde{P}_n(A)}{P_n^*(A)} = 1$ and $\lim_{n\to\infty} \sup_{A\in\pi_n(Y_1^n)} \frac{P_n^*(A)}{\tilde{P}_n(A)} = 1$ $\mathbb{P}$-almost surely. On the other hand, we have that $\forall A \in \pi_n(Y_1^n)$, $\left|\frac{P_n^*(A)}{\tilde{P}_n(A)} - 1\right| \leq \frac{\left|\tilde{P}_n(A) - P_n^*(A)\right|}{P_n^*(A)} \cdot \frac{P_n^*(A)}{\tilde{P}_n(A)}$, then $\lim_{n\to\infty} \sup_{A\in\pi_n(Y_1^n)} \left|\frac{P_n^*(A)}{\tilde{P}_n(A)} - 1\right| = 0$ $\mathbb{P}$-almost surely. Finally noting that $\forall n$,

$$\sup_{A\in\pi_n(Y_1^n)} \left|\log \frac{\tilde{P}_n(A)}{P_n^*(A)}\right| \leq \max\left\{\sup_{A\in\pi_n(Y_1^n)} \left|\frac{\tilde{P}_n(A)}{P_n^*(A)} - 1\right|, \sup_{A\in\pi_n(Y_1^n)} \left|\frac{P_n^*(A)}{\tilde{P}_n(A)} - 1\right|\right\}.$$

proves the result for (18). Similarly from the construction of Q_n on $\pi_n(Y_1^n)$, the expression in (19) is upper bounded by

$$\log \frac{1}{b_n} \cdot \sup_{\pi\in\mathcal{A}_n} \sum_{A\in\pi} \left|P_n^*(A) - \tilde{P}_n(A)\right| + \sup_{A\in\pi_n(Y_1^n)} \left|\log Q_n(A) - \log Q(A)\right|.$$

The same arguments presented for (18) apply to prove that this bound tends to zero with probability one, in this case adopting Lemma 1 and (16). □

6. Applications

In this section we address two practical questions. First, is there a partition scheme that using the histogram-based estimate in (8), provides a strongly consistent KL divergence estimator distribution-free for a family of probability measures? Second, assuming a positive answer for the previous question, what are the range of design values on these constructions that guarantee this result?

To address these questions, we study how the set of sufficient conditions presented in Theorems 2 and 4 translate into specific design conditions in the context of two specific partition schemes: non-product statistically equivalent partitions (Gessaman, 1970; Lugosi and Nobel, 1996) and tree-structure vector quantization (TSVQ) (Devroye et al., 1996; Nobel, 2002; Scott, 2005; Breiman et al., 1984). In this regard, we restrict to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and to the case when P and Q belong to $P_\lambda(\mathbb{R}^d)$ (the family of distributions absolutely continuous with respect to the *Lebesgue* measure in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$).

7. Statistically Equivalent Data-Dependent Partitions

7.1. l_n-spacing Partition Rule for $\mathbb{R}$

Let us first start with a simple scanario. Let us consider the real line $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as the measurable space and a partition scheme that dichotomizes the space in statistically equivalent intervals. This was the setting explored by Wang et al. (2005). More precisely, let $Y_1, .., Y_n$ be i.i.d. realizations drawn from $Q \in \mathcal{P}_\lambda(\mathbb{R})$. The order statistics $Y^{(1)}, .., Y^{(n)}$ are defined as the permutation of $Y_1, .., Y_n$ such that $Y^{(1)} < Y^{(2)} < \cdots < Y^{(n)}$ — this permutation exists

with probability one as $Q \ll \lambda$. Based on this sequence, the resulting l_n-spacing partition rule is given by $\pi_n(Y_1^n) = \left\{I_i^n : i = 1, .., T_n\right\} = \left\{(-\infty, Y^{(l_n)}], (Y^{(l_n)}, Y^{(2l_n)}], .., (Y^{((T_n-1)l_n)}, \infty)\right\}$, where $T_n = \lfloor n/l_n \rfloor$ assuming the non-trivial case where $n > l_n$. Note that under this construction every cell of $\pi_n(Y_1^n)$ has at least l_n samples from $Y_1, .., Y_n$, which match with the critical mass constraints of Theorem 4. Then we can state the following result.

THEOREM 5. *Adopting the l_n-spacing partition scheme for the histogram-based estimate in (8) with $(l_n) \approx (n^{0.5+l/2})$ and $(a_n) \approx (n^{-p})$, there exists a range of design parameters $\mathcal{D} = \left\{(l, p) \in \mathbb{R}^2 : l \in (0, 1), p \in (0, \frac{l}{2}), 1 + 4p < 3l\right\} \neq \emptyset$ (see Figure 1), such that for any pair P, Q in $\mathcal{P}_\lambda(\mathbb{R})$ where $D(P\|Q) < \infty$,*

$$\lim_{n\to\infty} D_{\pi_n(Y_1^n)}(P_n^*\|Q_n) = D(P\|Q).$$

$\mathbb{P}$-almost surely.

Proof: We check the sufficient conditions of Theorem 4. First note that **a)** and **d)** are satisfied by construction of the estimate. Concerning **b)**, again by construction $\mathcal{M}(\mathcal{A}_n) \leq n/l_n + 1$, then considering $\tau = (l - 2p)$, $n^{-(l-2p)}\mathcal{M}(\mathcal{A}_n) \leq n^{1-(l-2p)}/l_n + n^{-(l-2p)}$. Given that $(l_n) \approx (n^{0.5+l/2})$, $p < \frac{l}{2}$ and $1 - 3l + 4p < 0$,

$$\lim_{n\to\infty} n^{-(l-2p)}\mathcal{M}(\mathcal{A}_n) = 0. \tag{22}$$

For **c)**, Lugosi and Nobel (1996) showed that $\Delta_n^*(\mathcal{A}_n) = \binom{T_n+n}{n}$, where using that $\log\binom{s}{t} \leq s \cdot h(t/s)$ (Devroye et al., 1996), with $h(x) = -x\log(x) - (1-x)\log(1-x)$ for $x \in [0, 1]$ — the binary entropy function (see Cover and Thomas (1991)), it follows that,

$$\begin{aligned} n^{-(l-2p)}\log\left(\Delta_n^*(\mathcal{A}_n)\right) &\leq n^{-(l-2p)} \cdot (n + T_n) \cdot h\left(\frac{n}{n+T_n}\right) \\ &\leq 2n^{1-(l-2p)} \cdot h\left(\frac{T_n}{n}\right) \leq 2n^{1-(l-2p)} \cdot h\left(\frac{1}{l_n}\right) \\ &= -\frac{2n^{1-(l-2p)}}{l_n}\log(1/l_n) - 2n^{1-(l-2p)}(1 - 1/l_n)\log(1 - 1/l_n). \end{aligned} \tag{23}$$

The first term on the right hand side (RHS) of (23) behaves like $O(n^{0.5(1-3l+4p} \cdot \log(l_n))$, where from the fact that $1 + 4p > 3l$ and $(l_n) \leq (n)$ this sequence tends to zero. The second term on the RHS of (23) behaves asymptotically like $-n^{1-(l-2p)} \cdot \log(1 - 1/l_n)$, which is upper bounded by $(\frac{n^{1-(l-2p)}}{l_n} \cdot \frac{1}{1-1/l_n}) \approx (\frac{n^{1-(l-2p)}}{l_n})$ (from $\log(x) \leq x - 1$). This last sequence tends to zero as $(l_n) \approx (m^{0.5+l/2})$ and $1 + 4p < 3l$. Finally for condition **e)**, Lugosi and Nobel (1996, *Theorem 4)* proved that it is sufficient to show that $\lim_{n\to\infty} \frac{l_n}{n} = 0$, and given that by construction (a_n) is $o(1)$, we prove the theorem. □

Note that the proof reduces to checking the sufficient conditions of Theorem 4. In fact these are the restrictions that define the domain of admissible parameters $\mathcal{D}$.

Remark 3. *These conditions imply that $\lim_{n\to\infty} l_n = \infty$ and (l_n) is $o(n)$, which are the sufficient conditions presented in Lugosi and Nobel (1996) for the l_n-based histogram based density estimation to be strongly consistent in the $\mathbf{L}_1$ sense. The fact that more restriction are needed to get strong consistency for the information divergence functional agrees with recent results by Piera and Parada (2009) showing that stronger conditions on the convergence of probability measures (relative to the total variational distance (Devroye and Lugosi, 2001)) are needed to get convergence of the information divergence under certain compactly supported considerations. Furthermore, this also agrees with the results on the context of density estimation consistent in direct information divergence (Barron et al., 1992), where this notion of consistency requires stronger conditions than the classical L_1-consistency.*

7.2. Gessaman's Statistically Equivalent Partition for $\mathbb{R}^d$

For the finite dimensional Euclidean space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we consider the particular type of statistically equivalent partition proposed by Gessaman (1970). In this context, the partition rule considers $T_n = \lfloor (n/l_n)^{1/d} \rfloor$ as the number

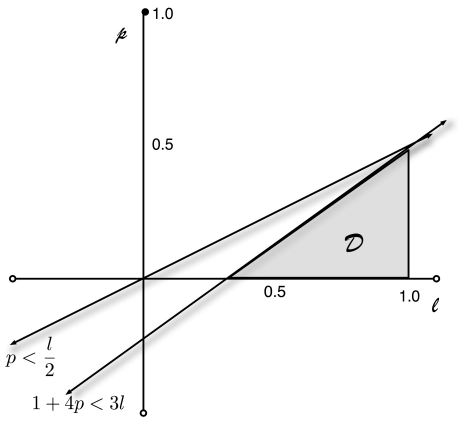

Figure 1: Range of parameters for consistent histogram-based estimators of the divergence adopting: statistically equivalent blocks and axis-parallel tree-structured partitions.

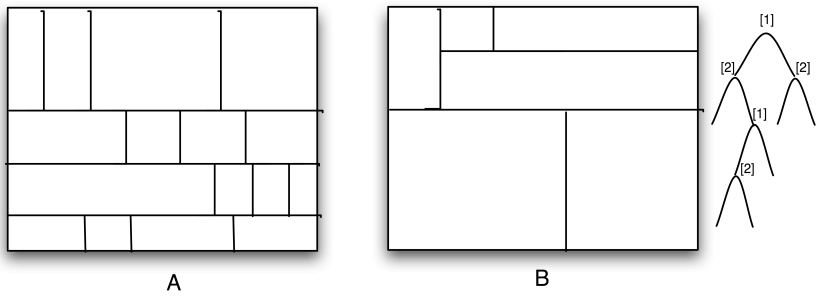

Figure 2: **A**: Example of Gessaman's statistically equivalent partition for a two dimensional bounded space. **B**: Example of a tree-structured data dependent partition and its tree-indexed structure. Each internal node has a label indicating the spatial coordinate used to split its associated rectangular set.

of axis-paralled splits to be induced in any coordinate of the space. More precisely first, the i.i.d samples $Y_1, .., Y_n$ associated with the reference measure Q are projected on to the first coordinate to create a statistically equivalent partition: with T_n cells and using axis-parallel hyper-planes perpendicular to the first coordinate. Then for any resulting rectangular cell, its respective sample points are projected on to the second coordinate and used to partition the cell in T_n statistically equivalent sets, in this case by hyper-planes perpendicular to the second coordinate. By iterating this process until the last coordinate, we have an adaptive partition scheme of exactly $(T_n)^d$ rectangular cells with at least l_n-sample points per cell, see Fig. 2 for an illustration[3].

THEOREM 6. *Adopting the Gessaman's partition scheme for the divergence estimate in (8), if* $(l_n) \approx (n^{0.5+l/2})$, $(a_n) \approx (n^{-p})$ *and the design parameters belong to* $\mathcal{D} = \left\{(l,p) \in \mathbb{R}^2 : l \in (0,1), p \in (0, \frac{l}{2}), 1 + 4p < 3l\right\}$ *(Fig. 1), then* $D_{\pi_n(Y_1^n)}(P_n^* \| Q_n)$ *is strongly consistent for any pair P and Q in* $\mathcal{P}_\lambda(\mathbb{R}^d)$ *for which* $D(P\|Q) < \infty$.

This result is an important generalization of *Theorem* 5. The proof follows similar arguments as its counterpart Theorem 5, however the technique used to prove the shrinking cell condition does not extend from the argument proposed by Lugosi and Nobel (1996, *Theorem 4)* that was adopted for the scalar case. The details of this argument and in particular the shrinking cell condition for the Gessaman's partition scheme are presented in Appendix E.

8. Tree-Structured Partition Schemes

We start with some definitions and preliminaries to facilitate the exposition of the main result in Section 8.3.

8.1. Basic Notation and Terminology

Using the conventions of Breiman et al. (1984), a *binary tree* T is a collection of nodes with only one with degree 2 (the *root* node), and the remaining nodes with degree 3 (*internal* nodes) or degree 1 (*leaf* or *terminal* nodes) [4]. Let $depth(t)$ denote the *depth* of $t \in T$ — the number of arcs that connect t with the root of T, and $\mathcal{L}(T)$ be the collection of terminal nodes of T. We define the *size* of a tree T as the cardinality of $\mathcal{L}(T)$ and denote it by $|T|$. 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 $\bar{T} \ll T$. Finally, T^r denotes the truncated version of T, formally given by $T^r = \{t \in T : depth(t) \leq r\}$ for all $r > 0$.

We will concentrate on the family of TSP induced by *hyperplane* cuts (Devroye et al., 1996). Following Nobel's conventions (Nobel, 2002, 1997), a *tree-structured partition* (TSP) can be represented by a pair $(T, \tau(\cdot))$, with T a binary tree and $\tau(\cdot)$ a function from T to $\mathcal{H}$, the collection of closed halfspaces of the form $\left\{x : x^\dagger w \geq \alpha\right\}$, with $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$. Then for any $t \in T$, $\tau(t)$ corresponds to the closed halfspace that dichotomizes the cell associated with t, denoted by U_t, in two components $U_t \cap \tau(t)$ and $U_t \cap \tau(t)^c$. These resulting cells are associated with the left and right child of t, respectively, when t is not a terminal node of T. Then initializing the cell of the root node t_0 with $U_{t_0} = \mathbb{R}^d$, $\tau(\cdot)$ provides a way to characterize U_t, $\forall t \in T$. In particular,

$$\pi(T) \equiv \{U_t : t \in \mathcal{L}(T)\} \subset \mathcal{B}(\mathbb{R}^d), \tag{24}$$

is the TSP induced by $(T, \tau(\cdot))$. Because of this construction, the cell associated with a node of depth k is a *convex polytope* of at most k faces[5] — this property will turn out to be crucial to prove consistency. 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 of $\bar{T}$. Note that if $\bar{T} \ll T$ then $\pi(T)$ is a refinement of $\pi(\bar{T})$, that we denote consistently by $\pi(\bar{T}) \ll \pi(T)$. Finally, we will use the tree notation T to refers to $(T, \tau(\cdot))$ or the partition $\pi(T)$ depending on the context.

[3] Note that the l_n-spacing partition is a particular case of Gessaman's partition scheme when $d = 1$.

[4] Formally a tree is a connected graph with no cycles. However, Breiman et al. (1984) propose a simplification where only the nodes are used to represent trees, making implicit the arcs that connect them.

[5] A polytope refers to sets induced by finite intersections of closed or open halfspaces (Devroye et al., 1996).

8.2. The Tree-Structured Data-Dependent Partitions

A n-sample TSP rule T_n is a function from the space of finite sequences $\mathbb{R}^{d\cdot n}$ to the space of TSP presented above, and the resulting partition scheme is the collection of TSP rules $\Pi = \{T_1, T_2, \cdots\}$.

Here we focus on the general family of TSP rules induced by a local splitting and stopping criteria. Let $\mathcal{U}$ be the collection of *polytopes* in $\mathbb{R}^d$ and $\mathcal{P}$ be the space of probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then a *local splitting rule* can be seen as a function $\Psi : \mathcal{U} \times \mathcal{P} \to \mathcal{H}$, that for a given cell $U \in \mathcal{U}$ and probability measure $P \in \mathcal{P}$ it defines a closed halfspace $\Psi(U, P) \in \mathcal{H}$ to partition U. Associated with Ψ we consider a *local stopping criterion*. This is a binary function $\Phi : \mathcal{U} \times \mathcal{P} \times [0, 1] \to \{0, 1\}$, which for given $U \in \mathcal{U}$ and $P \in \mathcal{P}$, indicates when to apply the local splitting criteria $\Psi(\cdot)$ on the cell U. We consider stopping rules of the form,

$$\Phi(U, P, p) = \begin{cases} 1 & \text{if } P(U \cap \Psi(U, P)) > p \text{ and } P(U \cap \Psi(U, P)^c) > p, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

for $U \in \mathcal{U}$, $P \in \mathcal{P}$ and $p \in (0, 1)$.

Finally given $Y_1^n = (Y_1, .., Y_n)$ i.i.d. realizations of the reference measure Q, the corresponding empirical distribution Q_n and a non-negative sequence $(k_n) \in \mathbb{N}^{\mathbb{N}}$, the n-sample partition rule $\pi(T_n(Y_1^n))$ is induced by

1. Initialization: $T_n = \{t_0\}$ (the root node), $U_{t_0} = \mathbb{R}^d$, $\pi(T_n) = \{U_{t_0}\}$ and $\tau_n(t_0) = \Psi(U_{t_0}, Q_n)$
2. Recursion: for all $t \in \mathcal{L}(T_n)$
   if $\Phi(U_t, Q_n, k_n/n) = 1$, then consider t_1 and t_2 as the left and right extensions of t and update as follows:
   - $T_n = T_n \cup \{t_1, t_2\}$,
   - $U_{t_1} = U_t \cap \tau_m(t)$, $U_{t_2} = U_t \cap \tau_n(t)^c$
   - $\tau_n(t_1) = \Psi(U_{t_1}, Q_n)$, $\tau_n(t_2) = \Psi(U_{t_2}, Q_n)$.
   - $\pi(T_n) = \pi(T_n) \setminus \{U_t\} \cup \{U_{t_1}, U_{t_2}\}$
3. Termination: Repeat 2), until $\Phi(U_t, Q_n, k_n/n) = 0$, $\forall t \in \mathcal{L}(T_n)$.

Note that by construction $Q_n(U_t) \geq k_n/n$, $\forall t \in \mathcal{L}(T_n(Y_1^n))$, which is consistent with condition **d)** of Theorem 4. The following result from Theorem 4 can be stated.

THEOREM 7. *Let P, Q be two probability measures in $\mathcal{P}_\lambda(\mathbb{R}^d)$ such that $D(P\|Q) < \infty$. Let $\Pi = \{T_1, T_2, \cdots\}$ be a TSP scheme driven by the empirical process $Y_1, Y_2, \cdots$ (with $Y_i \sim Q$, $\forall i > 0$) and the local stopping rule governed by a sequence of non-negative numbers $(k_n)_{n\in\mathbb{N}}$.*

If $(k_n) \approx (n^{0.5+l/2})$, $(a_n) \approx (n^{-p})$, $(l, p) \in \mathcal{D} = \left\{(l', p') : l' \in (0, 1), p' \in (0, \frac{l'}{2}), 1 + 4p' < 3l'\right\}$ (illustrated in Fig. 1) and Π satisfies the shrinking cell condition in (14), then $\lim_{n\to\infty} D_{\pi_n(Y_1^n)}(P_n^*\|Q_n) = D(P\|Q)$, $\mathbb{P}$-almost surely.

Proof: We need to verify the conditions **b)** and **c)** of *Theorem* 4, because **a)** and **d)** are obtained from the construction of the estimate and **e)** is assumed. By the stopping criterion $\left|T_n(y_1^n)\right| \leq n/k_n$, $\forall y_1^n \in \mathbb{R}^{d\cdot n}$. Then $\mathcal{M}(\mathcal{A}_n) \leq n/k_n$ and consequently,

$$(n^{-(l-2p)}\mathcal{M}(\mathcal{A}_n)) \leq (\frac{n^{1-(l-2p)}}{k_n}) \approx (m^{0.5-\frac{3}{2}l+2p}),$$

upper bound that tends to zero if $1 + 4p < 3l$. Concerning condition **c)**, we use the arguments proposed by Lugosi and Nobel (1996), specifying that every polytope of $\pi(T_n(y_1^n))$ is induced by at most $\mathcal{M}(\mathcal{A}_n)$ hyperplane splits. Each binary split can dichotomize $n \geq 2$ points in $\mathbb{R}^d$ in at most n^d ways (Cover, 1965). Consequently, $\Delta_n^*(\mathcal{A}_n) \leq (n^d)^{n/k_n}$, then,

$$(n^{-(l-2p)} \log \Delta_n^*(\mathcal{A}_n)) \leq (\frac{n^{1-(l-2p)}}{k_n} \cdot d \log n),$$

upper bound that again tends to zero as long as $1 + 4p < 3l$. □

8.3. Statistically Equivalent Splitting Rule

Going one step further, in this section we consider a version of a *balanced search tree* (see Devroye et al., 1996, *Chapter 20.3)*. More precisely, given $Y_1, Y_2, .., Y_n$ i.i.d. realizations of the reference measure Q we consider a splitting rule $\Psi(U_t, Q_n) \in \mathcal{H}$ that choses a dimension of the space sequentially, function of the depth of t — for instance $i = mod_d(depth(t))$ — and the i axis-parallel halfspace by

$$\Psi(U_t, Q_n) = \left\{x \in \mathbb{R}^d : x(i) \leq \bar{Y}^{(\lceil \bar{n}/2 \rceil)}(i)\right\}, \tag{26}$$

where $\bar{Y}^{(1)}(i) < \bar{Y}^{(2)}(i) <, .., < \bar{Y}^{(\bar{n})}(i)$ denotes the order statistics of the sampling points of interest $\left\{\bar{Y}_1, .., \bar{Y}_{\bar{n}}\right\} = \{Y_1, .., Y_n\} \cap U_t$ projected in the target dimension i. At the end $T_n(Y_1^n)$ can be seen as a statistically equivalent partition of the space. However considering the stopping criterion in (25), it does not guarantee equal empirical mass on their bins, neither to be a balanced tree (expect for the case when n is *dyadic*, i.e., $n = 2^k$ for some $k \in \mathbb{N}$).

To prove that this TSP scheme Π induces a strongly consistent KL divergence estimator, we just need to verify that Π satisfies the shrinking cell condition, under the specific assumptions stated in Theorem 7. For that, some definitions and a result will be needed.

Definition 4. *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 5. *A TSP scheme $\Pi = \{T_1, T_2, \cdots\}$ is a uniform balanced tree-structure scheme, if each partition rule $T_n(\cdot)$ forms a balanced tree of height d_n (only function of n).*

LEMMA 4. *Let $\Pi = \{T_1, T_2, \cdots\}$ be a uniform balanced tree-structure scheme induced by the statistically equivalent splitting rule (26) and with height sequence $(d_n)_{n\in\mathbb{N}}$. Π satisfies the shrinking cell condition of Theorem 2, if there exists a non-negative real sequence $(q_n) \approx (n^\theta)$, for some $\theta > 0$, such that*

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

This result was derived from the ideas presented by Devroye et al. (1996, *Theorem* 20.2) where a weak version of our shrinking cell condition was proved for a similar balanced tree-structured partition scheme. The proof of this stronger result is presented in Appendix F.

Finally we have all the machinery to state our final result.

THEOREM 8. *Let $\Pi = \{T_1, T_2, \cdots\}$ be a TSP scheme with the stopping and splitting rule presented in (25) and (26), respectively. Under the problem statement and the parameter constraints imposed on the sequences (k_n) and (a_n) in Theorem 7, for any pair P and Q in $\mathcal{P}_\lambda(\mathbb{R}^d)$ for which $D(P\|Q) < \infty$, $\lim_{n\to\infty} D_{\pi_n(Y_1^n)}(P_n^*\|Q_n) = D(P\|Q)$ $\mathbb{P}$-almost surely.*

Proof: The proof reduces to verify the shrinking cell condition for Π. By the binary tree structure of Π and the stopping rule, it is simple to show that, $\forall y_1^n \in \mathbb{R}^{d\cdot n}$,

$$r(n) \equiv \lfloor \log_2(n) \rfloor - \lceil \log_2(k_n) \rceil \leq \min_{t\in\mathcal{L}(T_n(y_1^n))} depth(t), \tag{27}$$

and consequently $T_n^{r(n)}(Y_1^n)$ is a balanced tree. Defining $\bar{\Pi} = \left\{T_1^{r(1)}, T_2^{r(2)}, \cdots\right\}$, it suffices to check the shrinking cell condition on $\bar{\Pi}$ [6]. Given that $\bar{\Pi}$ is a uniform balanced tree-structure scheme, we can check the sufficient condition stated in Lemma 4. Let $\bar{d}_n (= r(n))$ denote the height of $T_n^{r(n)}$. By construction $\bar{d}_n \geq \log_2(n/k_n) - 2$ and consequently tends to infinity ($(k_n) \approx (n^{0.5+l/2})$ with $l < 1$). On the other hand, if we consider an arbitrary non-negative sequence $(q_n) \approx (n^\theta)$ with $\theta \in \left(0, \frac{2}{3}\right]$, then

$$\frac{n}{\bar{d}_n 2^{\bar{d}_n}} - \frac{q_n}{\bar{d}_n} \geq \frac{n}{d_n \cdot 2^{\log_2(n/k_n)}} - \frac{q_n}{d_n} = \frac{k_n - q_n}{d_n} \to \infty \tag{28}$$

as $n \to \infty$, because $(d_n) \preceq (\log_2(n))$, $(k_n) \approx (n^{0.5+l/2})$ and in $\mathcal{D}$ we have that $l > 1/3$, which proves the result. □

[6] Π is a refinement of $\bar{\Pi}$ in the sense that $\forall n \in \mathbb{N}$, $\forall y_1^n \in \mathbb{R}^{d\cdot n}$, $T_n^{r(n)}(y_1^n) \ll T_n(y_1^m)$, then by definition the shrinking cell condition of $\bar{\Pi}$ implies the property for Π.

9. Final Remarks

The main result in Theorem 4 and its applications (Theorems 6 and 8) suggest that the information divergence estimation problem put more restrictions in terms of data-driven design conditions when compared with the problem of density estimation, in particular for the reference measure Q, consistent in the L_1 sense (Lugosi and Nobel, 1996). This conjecture agrees with findings on density-free estimation of information theoretic quantities (Györfi and van der Meulen, 1987) and the convergence analysis of the Shannon differential entropy in (Piera and Parada, 2009).

Concerning the Barron's density estimate adopted for estimating $\frac{\partial P}{\partial Q}(x)$, it is interesting to contrast its use here from its original adoption in (Barron et al., 1992). The main difference is that in our problem both distributions need to be estimated from data, while in the work of Barron et al. (1992) a sigma finite measure μ is assumed and used to estimate a measure P (assuming that $P \ll \mu$) from a smooth version of $\frac{\partial P_n}{\partial \mu}$, restricted to the sigma field induced by sequence of μ-statistically equivalent partitions. This last point rises the other important difference from our problem, which is the use of data-driven partitions. Consequently, in this work we demonstrated the utility of Barron's density estimate in a different learning context, as well as its adequate interaction with data-driven partition schemes.

Finally, the presented formulation offers the possibility of extending the role of data-driven histogram-based construction to the estimation of other information theoretic quantities — like the Shannon mutual information (Shannon, 1948) and the general family of ϕ-divergence introduced by Csiszár (1967) (see also, Liese and Vajda (1987); Csiszár and Shields (2004)), as well as using the rich machinery of statistical learning theory (see, eg., Vapnik (1998); Devroye et al. (1996)) to explore for instance, distribution-free rate of convergence results.

10. Acknowledgment

The work of J. Silva was supported by funding from FONDECYT Grant 1090138, CONICYT-Chile. The work of S. S. Narayanan was supported by funding from the National Science Foundation (NSF).

A. Proof of Lemma 2

Proof: Let us restrict to the case when $\mathcal{A}$ is a collection of measurable events. It is simple to show that $\forall a \in [0,1]$,

$$\begin{aligned}
\mathbb{P}\left(\sup_{A\in\mathcal{A}}\left|\mu_n^a(A)-\mu^a(A)\right|>\epsilon\right) &\leq \mathbb{P}\left(\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|+\sup_{A\in\mathcal{A}}|Q_n(A)-Q(A)|>\epsilon\right)\\
&\leq \mathbb{P}\left(\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|>\frac{\epsilon}{2}\right)+\mathbb{P}\left(\sup_{A\in\mathcal{A}}|Q_n(A)-Q(A)|>\frac{\epsilon}{2}\right)\\
&\leq 8\cdot\mathcal{S}_{2n}(\mathcal{A})\exp^{\frac{-n\epsilon^2}{4\cdot 8}}.
\end{aligned}$$

The last inequality is the classical *VC inequality* (Vapnik and Chervonenkis, 1971; Vapnik, 1999), where $S_{2n}(\mathcal{A})$ denotes the *scatter coefficient*[7] of $\mathcal{A}$. Finally, the proof of Lemma 2 follows from the arguments presented in Lugosi and Nobel (1996, *Lemma* 1). □

B. Proof of Theorem 1

Proof: Let us consider an arbitrary $\epsilon > 0$. Then, there exists a finite partition $\pi(\epsilon/3)$, such that,

$$D_{\pi(\epsilon/3)}(P\|Q) > D(P\|Q) - \epsilon/3. \tag{29}$$

Considering that $|\pi(\epsilon/3)| < \infty$ and that $x\log x$ is a continuous real function, $D_{\pi(\epsilon/3)}(P\|Q)$ is a continuous function with respect to the total variational distance in the product space of probability measures on $(\mathbb{R}^d, \sigma(\pi(\epsilon/3)))$ under

[7] The *scatter coefficient* of $\mathcal{A}$ is given by $S_n(\mathcal{A}) = \sup_{x_1^n\in\mathcal{X}^n}|\{\{x_1,x_2,..,x_n\}\cap A : A\in\mathcal{A}\}| \leq 2^n$. It is an indicator of the richness of $\mathcal{A}$ to dichotomize a sequence of points in $\mathcal{X}$ (see, eg., Devroye et al. (1996); Vapnik and Chervonenkis (1971); Vapnik (1999)).

some additional conditions. More precisely for $\epsilon/3$, $\exists\delta_1 > 0$ and $\delta_2 > 0$, such that if, $\sup_{i=1,..,r} \left|P^1(A_i) - P^2(A_i)\right| < \delta_1$, $\sup_{i=1,..,r} \left|Q^1(A_i) - Q^2(A_i)\right| < \delta_2$, and $P^1 \ll Q^1$, $P^2 \ll Q^2$ then,

$$\left|D_{\pi(\epsilon/3)}(P^1\|Q^1) - D_{\pi(\epsilon/3)}(P^2\|Q^2)\right| < \epsilon/3.$$

Then a direct consequence of the hypotheses of the theorem is that for any typical sequence $y_1, y_2, y_3, y_4 \cdots$ (sequence for which Π is simultaneously (P, Q)-approximating) there exists a sequence of measurable approximations of $\pi(\epsilon/3)$, denoted by $\{\pi_n^*, n \in \mathbb{N}\} \subset \mathcal{Q}$ with $\pi_n^* \subset \sigma(\pi_n(y_1^n))$, such that for $\epsilon/3$, $\exists N < \infty$ and $\forall n > N$,

$$D_{\pi_n^*}(P\|Q) > D_{\pi(\epsilon/3)}(P\|Q) - \epsilon/3. \tag{30}$$

Furthermore by construction,

$$\sup_{A\in\pi_n^*} \left|\tilde{P}_n(A) - P(A)\right| \leq \sup_{A\in\sigma(\pi_n(y_1^n))} \left|\tilde{P}_n(A) - P(A)\right| \leq \sup_{A\in\mathcal{S}} \left|\tilde{P}_n(A) - P(A)\right| \leq a_n,$$

with $\lim_n a_n = 0$. Then from the continuity of $x \cdot \log(x)$, for $\epsilon/3$ there exists $\bar{N} > 0$ such that $\forall n > \bar{N}$,

$$D_{\pi_n^*}(\tilde{P}_n\|Q) > D_{\pi_n^*}(P\|Q) - \epsilon/3. \tag{31}$$

Finally, using the three previous inequalities and noting that $\pi_n(y_1^n)$ is a refinement of π_n^*, we have that for every typical sequence $D_{\pi_n(y_1^n)}(\tilde{P}_n\|Q) > D(P\|Q) - \epsilon$, eventually $\forall\epsilon > 0$. Then, $\forall\epsilon > 0$

$$\liminf_{n\to\infty} D_{\pi_n(Y_1^n)}(\tilde{P}_n\|Q) > D(P\|Q) - \epsilon, \ \mathbb{Q}\text{-almost surely.} \tag{32}$$

On the other hand, for an arbitrary sequence $y_1, y_2, \cdots$ and $\forall n > 0$,

$$\begin{aligned} D_{\pi_n(y_1^n)}(\tilde{P}_n\|Q) &\leq \sum_{A\in\pi_n(y_1^n)} \left[(1-a_n)log(P(A))P(A) + a_n log(Q(A))Q(A)\right] - \sum_{A\in\pi_n(y_1^n)} \log(Q(A)) \left[(1-a_n)P(A) + a_n Q(A)\right] \\ &= (1-a_n)D_{\pi_n(y_1^n)}(P\|Q) \leq D(P\|Q), \end{aligned} \tag{33}$$

this by the convexity of $x \cdot \log(x)$, *Jensen's* inequality (Cover and Thomas, 1991) and the fact that $a_n \leq 1$. Finally, from (32) and (33), $\lim_{n\to\infty} D_{\pi_n(Y_1^n)}(\tilde{P}_n\|Q) = D(P\|Q)$ $\mathbb{Q}$-almost surely. □

C. Proof of Theorem 2

Let us first note that the shrinking cell condition of Q in (14) implies that the same property holds for the measure P [8]. Using the short-hand notation $Y_1^m = Y_1, ..., Y_m$, (14) is equivalent to

$$\lim_{m\to\infty} Q\left(\bigcup_{\substack{A\in\pi_m(Y_1^m) \\ diam(A)>\gamma}} A\right) = 0, \ \mathbb{Q}\text{-almost surely } \forall\gamma > 0. \tag{34}$$

Proof: We will concentrate on showing the result for the measure Q. Let us consider an arbitrary partition $\pi = \{A_1, .., A_r\} \in \mathcal{Q}$. Let $\left\{B_1^m, .., B_r^m\right\}$ be the covering of π induced by $\pi_m(Y_1^m)$, i.e.,

$$B_j^m = \bigcup_{\substack{A\in\pi_m(Y_1^m) \\ A\cap A_j\neq\varnothing}} A, \ \ \forall j \in \{1, .., r\}.$$

[8] This can be derived from the fact that $P \ll Q$ and the *dominated convergence theorem* (Varadhan, 2001; Halmos, 1950).

Based on $\left\{B_1^m, .., B_r^m\right\}$, we can induce a partition $\pi_m^* = \left\{A_1^m, .., A_r^m\right\} \subset \sigma(\pi_m(Y_1^m))$ that approximates π by the following construction: $A_1^m = B_1^m$, $A_2^m = B_2^m \setminus B_1^m$, $\cdots$, $A_r^m = B_r^m \setminus \left(\cup_{j=1}^{r-1} B_j^m\right)$.

Let us consider an arbitrary $\delta > 0$. Given that Q is absolutely continuous with respect to the Lebesgue measure λ, there is a bounded measurable set B such that $Q(B) > 1 - \delta/2$. Let us define $\bar{\pi} = \left\{\bar{A}_1, .., \bar{A}_r\right\}$ with $\bar{A}_j = B \cap A_j$, $\forall j = 1, .., r$ and $\bar{\pi}_m^* = \left\{\bar{A}_1^m, .., \bar{A}_r^m\right\}$ as the partition of B induced by π and π_m^*, respectively. Then for any $A_i \in \pi$,

$$\begin{aligned}\left|Q(A_i) - Q(A_i^m)\right| &< \left|Q(A_i) - Q(\bar{A}_i)\right| + \left|Q(\bar{A}_i) - Q(\bar{A}_i^m)\right| + \left|Q(\bar{A}_i^m) - Q(A_i^m)\right| \\ &< \delta + \left|Q(\bar{A}_i) - Q(\bar{A}_i^m)\right|, \ \forall m \in \mathbb{N}, \end{aligned} \tag{35}$$

the last inequality from the construction of B.

In addition, for any measurable set $A \in \mathcal{B}(\mathbb{R}^d)$ let us define its γ-*open covering* by $A^{\gamma+} \equiv \bigcup_{x \in A} B(x, \gamma)$, and its γ-*residue* by $\delta_\gamma(A) \equiv A^{\gamma+} \setminus A \in \mathcal{B}(\mathbb{R}^d)$, with $B(x, \gamma)$ denoting the open ball of radius γ centered at x. Note that by the continuity of λ under monotone set sequences (Halmos, 1950)[9], $\forall A \in \mathcal{B}(\mathbb{R}^d)$ and $\forall \epsilon > 0$, $\exists \gamma > 0$, such that $\lambda(\delta_\gamma(A)) < \epsilon$, where given that $Q \ll \lambda$, the same is true considering the measure Q. Hence, let us fix γ such that $Q(\delta_\gamma(\bar{A}_i) \cup \delta_\gamma((\bar{A}_i)^c)) < \epsilon$ uniformly $\forall i \in \{1, .., r\}$, and let us define the event S_γ^m in $\mathcal{B}(\mathbb{R}^{d \cdot m})$ by

$$S_\gamma^m = \left\{y_1^m \in \mathbb{R}^{d \cdot m} : diam(\pi_m(y_1^m)) < \gamma\right\},$$

with $diam(\pi_m(y_1^m)) = \max_{A \in \pi_m(y_1^m)} diam(A)$. Then,

$$\left|Q(\bar{A}_i) - Q(\bar{A}_i^m)\right| \leq Q(\bar{A}_i \triangle \bar{A}_i^m) \tag{36}$$

$$\leq Q(\delta_\gamma(\bar{A}_i) \cup \delta_\gamma((\bar{A}_i)^c) \cdot \mathbb{I}_{S_\gamma^m}(Y_1^m) + [Q(\delta_\gamma(\bar{A}_i) \cup \delta_\gamma((\bar{A}_i)^c) + Q\left(\bigcup_{\substack{A \in \pi_m(Y_1^m) \\ diam(A) > \gamma}} A\right)] \cdot \mathbb{I}_{(S_\gamma^m)^c}(Y_1^m) \tag{37}$$

$$\leq \epsilon + Q\left(\bigcup_{\substack{A \in \pi_m(Y_1^m) \\ diam(A) > \gamma}} A\right), \ \forall m \in \mathbb{N}, \tag{38}$$

where (37) derives from the construction of $\bar{A}_i^m$ and the fact that conditioning to the event S_γ^m, $\bar{A}_i \triangle \bar{A}_i^m = (\bar{A}_i^m \setminus \bar{A}_i) \cup (\bar{A}_i \setminus \bar{A}_i^m) \subset \delta_\gamma(\bar{A}_i) \cup \delta_\gamma((\bar{A}_i)^c)$, where more generally $\forall \gamma > 0$,

$$\bar{A}_i \triangle \bar{A}_i^m \subset \delta_\gamma(\bar{A}_i) \cup \delta_\gamma((\bar{A}_i)^c) \cup_{\substack{A \in \pi_m(Y_1^m) \\ diam(A) > \gamma}} A. \tag{39}$$

Then, from the hypothesis in (34), (35) and (38),

$$\limsup_{m \to \infty} \left|Q(A_i) - Q(A_i^m)\right| < \delta + \epsilon, \ \mathbb{Q}\text{-almost surely.} \tag{40}$$

Finally noting that this result is valid for any measurable event $A_i \in \pi$ and that ϵ can be chosen arbitrarily small, it follows that,

$$\limsup_{m \to \infty} \sup_{i \in \{1, .., r\}} \left|Q(A_i) - Q(A_i^m)\right| < \delta, \ \mathbb{Q}\text{-almost surely.} \tag{41}$$

The same partition sequence $\left\{\pi_1^*, \pi_2^*, \cdots\right\}$ and arguments can be adopted to show the result for the measure P, which proves the theorem. □

[9] Note that for all $A \in \mathcal{B}(\mathbb{R}^d)$, $\lim_{n \to \infty} \delta_{1/n}(A) = \emptyset$.

D. Proof of Lemma 3

Proof: Let us focus on proving Eq. (17) and consequently in analyzing,

$$\begin{aligned}\mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)}\left|\frac{\tilde{P}_n(A)}{P_n^*(A)}-1\right|>\epsilon\right) &\leq \mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)}\left|\tilde{P}_n(A)-P_n^*(A)\right|>\epsilon\cdot a_n b_n\right)\\ &\leq \mathbb{P}\left(\sup_{\pi\in\mathcal{A}_n}\sup_{A\in\pi}\left|\tilde{P}_n(A)-P_n^*(A)\right|>\epsilon\cdot a_n\cdot b_n\right)\\ &\leq 8\Delta_{2n}^*(\mathcal{A}_n)2^{\mathcal{M}(\mathcal{A}_n)}\exp^{-\frac{n(\epsilon\cdot a_n\cdot b_n)^2}{128}},\end{aligned} \tag{42}$$

where the first inequality is by the hypothesis, i.e., $P_n^*(A)\geq a_n\cdot b_n\ \forall A\in\pi_n(Y_1^n)$ with $b_n\equiv\frac{k_n}{n}\ \forall n>0$, and the last from the VC inequality for mixture distributions in Lemma 2. From (42) and the fact that $(a_n)\geq(n^{-p})$ and $(b_n)\geq(n^{l/2-0.5})$ (condition **a)**),

$$(n^{-\tau}\log\mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)}\left|\frac{\tilde{P}_n(A)}{P_n^*(A)}-1\right|>\epsilon\right))\leq(-n^{1-\tau}\frac{(\epsilon\cdot a_n\cdot b_n)^2}{128})\leq(-\epsilon^2\cdot n^{(l-2p)-\tau}), \tag{43}$$

then it is clear that

$$\lim_{n\to\infty}n^{-\tau}\log\mathbb{P}\left(\sup_{A\in\pi_n(Y_1^n)}\left|\frac{\tilde{P}_n(A)}{P_n^*(A)}-1\right|>\epsilon\right)<0\text{ or diverges to }-\infty,$$

where from the hypothesis $\tau>0$. Then *Borel-Cantelli* lemma proves the result.

The same arguments can be adopted to show Eq.(16) but in this case using the classical VC inequality in Lemma 1. In fact weaker conditions can be stated to prove that this term converges to zero almost surely. In that sense the critical part was to bound the deviation of P_n^* with respect to $\tilde{P}_n$ in $(\mathcal{X},\sigma(\pi_n(Y_1^n)))$. □

E. Proof of Theorem 6

Proof: The conditions **a)** and **d)** are satisfied by construction of the partition scheme. The argument for **b)** extends from the proof of Theorem 5. Concerning **c)**, using the same combinatorial argument, we have that $\Delta_n^*(\mathcal{A}_n)\leq\binom{T_n+n}{n}^d$. Defining $\bar{T}_n=\lfloor n/ln\rfloor\geq T_n$ and $h(\cdot)$ the binary entropy function, we can use the derivations in (23) to show that,

$$n^{-(l-2p)}\log\left(\Delta_n^*(\mathcal{A}_n)\right)\leq n^{-(l-2p)}d\cdot\log(\tbinom{\bar{T}_n+n}{n})\leq 2d\cdot n^{1-(l-2p)}\cdot h\left(\frac{1}{l_n}\right).$$

This last upper bound tends to zero as n goes to infinity because $(l_n)\approx(n^{0.5+l/2})$ and $1+4p<3l$ as shown in *Theorem* 5. To verify the shrinking cell condition, we follow the structure of the proof presented by Devroye et al. (1996, *Theorem* 20.2). In particular, subsections E.1, E.2 and E.3 provide some preliminaries and subsection E.4 provides the final argument.

E.1. Reducing the Problem to a Bounded Measurable Space

Note that the partition scheme Π is *monotone transformation invariant* (Devroye et al., 1996), in the sense that for all $\pi_n\in\Pi$, $\forall x\in\mathbb{R}^d$, $\forall y_1^n\in\mathbb{R}^{d\cdot n}$,

$$\pi_n(x|y_1,..y_n)=\pi_n(F(x)|F(y_1),..F(y_n)),$$

where $F:\mathbb{R}^d\to\mathbb{R}^d$ is an arbitrary function that can be expressed by $F(x)=(f_1(x(1)),\cdots,f_d(x(d)))$, for some collection of strictly increasing real functions $\{f_i(\cdot):i=1,..,d\}$. In particular, we can consider $f_i(\cdot)$ to be the distribution function of the marginal probability Q, restricted to events on the i-coordinate $\forall i\in\{1,..,d\}$. Without loss of generality we can restrict to the case when $\{f_i(\cdot):i=1,..,d\}$ are strictly increasing. Consequently, the induced distributions of the transform space, denoted by $\bar{Q}$ and $\bar{P}$ respectively, have support on $[0,1]^d$ and satisfies that (Gray, 1990)

$$D(P\|Q)=D(\bar{P}\|\bar{Q}), \tag{44}$$

because $F(\cdot)$ is one-to-one continuous mapping from $\mathbb{R}^d$ to $[0,1]^d$ (more precisely $\left\{F^{-1}(A) : A \in \mathcal{B}([0,1]^d)\right\} = \mathcal{B}(\mathbb{R}^d)$). Moreover, if we apply Π in the transform domain, i.e., we estimate the empirical distributions using the transform i.i.d. realizations $F(X_1), .., F(X_n)$ and $F(Y_1), .., F(Y_n)$ — denoted by $\bar{P}_n^*$ and $\bar{Q}_n$ on $\sigma(\pi(F(Y_1), .., F(Y_n)))$ — and estimate the divergence by (8), it is simple to check that

$$D_{\pi(F(Y_1),..,F(Y_n)))}(\bar{P}_n^*, \bar{Q}_n) = D_{\pi(Y_1^n)}(P_n^*, Q_n). \tag{45}$$

Then from (44) and (45), we can reduce the problem to checking the the shrinking cell condition for the case when Q and P are defined on $([0,1]^d, \mathcal{B}([0,1]^d)$.

E.2. Formulation of a Sufficient Condition

Given that $\pi_n(Y_1^n)$ is induced by axis-parallel hyperplanes, every cell $U \in \pi_n(Y_1^n)$ is a finite dimensional rectangle of the form $\otimes_{i=1}^d [l_i, u_i)$ (with the possible open and closed interval variations). In this scenario, $\forall U \in \pi_n(Y_1^n)$,

$$diam(U) \leq \sum_{i=1}^{d} length_i(U), \tag{46}$$

with $length_i(U)$ denoting the Lebesgue measure of the projection of U on the i-coordinate. Then from *Markov's inequality*, for proving the shrinking cell condition it suffices to show that (Devroye et al., 1996),

$$\lim_{n\to\infty} \mathbb{E}_Q\left(\sum_{i=1}^{d} length_i(\pi_n(X|Y_1^n))\right) = \lim_{n\to\infty} \int_{[0,1]^d} \sum_{i=1}^{d} length_i(\pi_n(x|Y_1^n))\partial Q(x) = 0, \tag{47}$$

almost surely with respect to the process distribution of $Y_1, Y_2 \cdots$.

E.3. ϵ-Statistically Equivalent Partitions

Definition 6. *Let $A \subset [0,1]^d$ be a finite dimensional rectangle of the form $\bigotimes_{i=1}^d [l_i, u_i)$ with $l_i < u_i$. Let $\pi_j(A)$ be a partition of A induced by axis parallel hyperplanes on the j coordinate. We say that $\pi_j(A)$ is ϵ-*statistically equivalent *with respect to a measure Q if,*

$$\max_{B\in\pi_j(A)} Q(B) \leq \frac{Q(A)}{\left|\pi_j(A)\right|} \cdot \sqrt{1+\epsilon}, \tag{48}$$

where in this case by construction,

$$\sum_{B\in\pi_j(A)} \sum_{i=1}^{d} length_i(B) \cdot Q(B) \leq length_j(A) \cdot \frac{Q(A)\sqrt{1+\epsilon}}{\left|\pi_j(A)\right|} + \sum_{i\neq j} length_i(A) \cdot Q(A). \tag{49}$$

Note that our data-dependent construction can be seen as a concatenation of the type of axis parallel partition presented in *Definition* 6. Then, the following result holds.

PROPOSITION 1. *Let $\pi_n(Y_1^n)$ be a data-dependent Gessaman's partition of $[0,1]^d$ with T_n splits per coordinate. If during the construction of $\pi_n(Y_1^n)$, all its axis-paralled partitions are ϵ–statistically equivalent with respect to the reference measure Q, then $\forall n > 0$,*

$$\mathbb{E}_Q\left(\sum_{i=1}^{d} length_i(\pi_n(X|Y_1^n))\right) \leq \frac{d \cdot \sqrt{1+\epsilon}.}{T_n}. \tag{50}$$

Proof: By construction, $\pi_n(Y_1^n)$ can be seen as the concatenation of $1 + T_n + T_n^2 + \cdots T_n^{d-1}$ family of axis-paralled partitions. Then the proof can be derived from a recursive application of (49). □

E.4. Final Argument

Let $B_n(\epsilon) \subset \mathcal{B}(\mathbb{R}^{d\cdot n})$ be the set of realizations of the empirical process where $\pi_n(y_1^n)$ is concatenation of ϵ–statistically equivalent partitions with respect to Q, then from (50)

$$\mathbb{E}_Q\left(\sum_{i=1}^{d} length_i(\pi_n(X|Y_1^n))\right) \leq \frac{d\cdot\sqrt{1+\epsilon}.}{T_n}\cdot \mathbb{I}_{B_n(\epsilon)}(Y_1^n) + d\cdot T_n \cdot \mathbb{I}_{B_n(\epsilon)^c}(Y_1^n). \tag{51}$$

For proving (47), we are interested in the event

$$A_n(\epsilon) = \left\{y_1^n : \left|\mathbb{E}_Q\left(\sum_{i=1}^{d} length_i(\pi_n(X|y_1^n))\right)\right| > \epsilon\right\} \in \mathcal{B}(\mathbb{R}^{d\cdot n}).$$

Note that fixing $\epsilon_0 > 0$, $\forall \epsilon > 0$ we have that eventually $A_n(\epsilon) \subset B_n(\epsilon_0)^c$ and consequently $(\mathbb{Q}^n(A_n(\epsilon)) \preceq (\mathbb{Q}^n(B_n(\epsilon_0)^c))$, where $\mathbb{Q}^n$ denotes the probability measure on $(\mathbb{R}^{d\cdot n}, \mathcal{B}(\mathbb{R}^{d\cdot n}))$ induced by restricting the empirical process to the finite block Y_1^n. In addition, by the sub-additivitive of $\mathbb{Q}^n$,

$$\mathbb{Q}^n(B_n(\epsilon_0)^c) \leq \frac{T_n^d - 1}{T_n - 1}\mathbb{Q}^n(B_n^o(\epsilon_0)), \tag{52}$$

where $B_n^o(\epsilon_0) \in \mathcal{B}(\mathbb{R}^{d\cdot n})$ denotes the event that one of the $1 + T_n + T_n^2 + \cdots T_n^{d-1}$ axis-paralled partitions of $\pi_n(y_1^n)$ is not ϵ_0–statistically equivalent with respect to Q.

To find an expression for $\mathbb{Q}^n(B_n^o(\epsilon_0))$, without loss of generality, let us consider $A = [0,1]^d$, a coordinate $j \in \{1,..,d\}$ and $\pi_j(A) = \{A_1,..,A_{T_n}\}$ a partition of A based on $\bar{n}$ i.i.d. samples points projected on the j-coordinate, say $\bar{Y}_1(j) < \bar{Y}_2(j), \cdots < \bar{Y}_{\bar{n}}(j)$. If $F(x)$ and $\hat{F}_{\bar{n}}(x)$ denote the j-marginal distribution function and its empirical counterpart, respectively (associated with the reference measure Q), it is simple to show that if $\pi_j(A)$ is not ϵ_0–statistically equivalent, then $\sup_{x\in[0,1]}\left|\hat{F}_{\bar{n}}(x) - F(x)\right| > \frac{\sqrt{1+\epsilon_0}-1}{T_n}$ (see Devroye et al., 1996, Chapter 20.3). Consequently,

$$\begin{aligned}
\mathbb{Q}^n(B_n^o(\epsilon_0)) &\leq \mathbb{Q}^n\left(\left\{\sup_{x\in[0,1]}\left|\hat{F}_{\bar{n}}(x) - F(x)\right| > \frac{\sqrt{1+\epsilon_0}-1}{T_n}\right\}\right)\\
&\leq 2\cdot\exp\left(-2\cdot\bar{n}\cdot\left(\frac{\sqrt{1+\epsilon_0}-1}{T_n}\right)^2\right)\\
&\leq 2\cdot\exp\left(-2\cdot l_n\cdot\left(\frac{\sqrt{1+\epsilon_0}-1}{T_n}\right)^2\right),
\end{aligned} \tag{53}$$

the second inequality is obtained from the large deviation result in (Devroye et al., 1996, *Theorem* 12.9), where the last inequality is because $\forall A \in \pi_n(Y_1^n)$, $Q_n(A) \geq \frac{l_n}{n}$ and $\bar{n} \geq l_n$. Then, from (52) and (53),

$$\begin{aligned}
\mathbb{Q}^n(B_n(\epsilon_0)^c) &\leq 2\cdot\frac{T_n^d - 1}{T_n - 1}\cdot\exp\left(-2\cdot l_n\cdot\left(\frac{\sqrt{1+\epsilon_0}-1}{T_n}\right)^2\right)\\
&\leq 2\cdot T_n^d\cdot\exp\left(-2\cdot\frac{l_n}{\lfloor(n/l_n)^{1/d}\rfloor^2}\cdot\left(\sqrt{1+\epsilon_0}-1\right)^2\right)\\
&\leq 2\cdot\frac{n}{l_n}\cdot\exp\left(-2\cdot\frac{l_n^2}{n}\cdot\left(\sqrt{1+\epsilon_0}-1\right)^2\right),
\end{aligned} \tag{54}$$

where the third inequality uses that $\lfloor(n/l_n)^{1/d}\rfloor^2 \leq \lfloor(n/l_n)^{1/2}\rfloor^2 \leq n/l_n$ (considering $d \geq 2$). Finally, noting that $(l_n) \approx (n^{0.5+l/2})$ with $l \in (0,1)$,

$$(\mathbb{Q}^n(B_n(\epsilon_0)^c))) \preceq (n^{0.5-l/2}\cdot\exp\left(-2\cdot n^l\cdot(\sqrt{1+\epsilon_0}-1)^2\right)), \tag{55}$$

then the *Borel Cantelli lemma* proves the result. □

F. Proof of Lemma 4

Proof: For the rest, let $\pi_n(Y_1^n) = \pi(T_n(Y_1^n))$ denote the n-sample partition rule of the TSP scheme Π. Note that Π is monotone transformation invariant, then we can restrict to the case where P and Q are defined on $([0,1]^d, \mathcal{B}([0,1]^d))$ (see Appendix E.1). Also for proving the shrinking cell condition, this reduces to checking the condition presented in *Appendix* E.2: i.e., $\lim_{n\to\infty} \mathbb{E}_Q\left(\sum_{i=1}^d length_i(\pi_n(X|Y_1^n))\right) = 0$, $\mathbb{Q}$-almost surely.

F.1. Preliminaries: ϵ-good median cuts

Let $U = \otimes_{i=1}^d [l_i, u_i]$ be a rectangle in $\mathcal{B}([0,1]^d)$ and let $\left\{H_0^0, H_0^1, H_1^1, \cdots, H_0^{d-1}, .., H_{2^{d-1}-1}^{d-1}\right\}$ be a sequence of axis-parallel hyperplanes used to recursively split U in every coordinate. This partitions U in 2^d cells. More precisely, H_0^0 parallel to the 1-coordinate splits $U_0^0 = U$ into two rectangles U_0^1, U_1^1, then H_0^1 and H_1^1 parallel to the 2-coordinate split U_0^1 and U_1^1 into U_0^2, U_1^2, and U_2^2, U_3^2 respectively, and inductively at the end of the process a TSP for U is created $\left\{U_j^d : j = 0, .., 2^d - 1\right\}$.

Definition 7. *In the aforementioned construction, let $p_j^l = Q(U_j^l)$ be the probability of every induced rectangle, then we say that $\left\{H_0^0, H_0^1, H_1^1, \cdots, H_0^{d-1}, .., H_{2^{d-1}-1}^{d-1}\right\}$ is a* sequence of ϵ-good median cuts *for U if: $\forall l \in \{0, .., d-1\}$ and $j \in \left\{0, .., 2^l - 1\right\}$,*

$$\max(p_{2j}^{l+1}, p_{2j+1}^{l+1}) \leq \frac{1}{2}(1+\epsilon)^{1/d} \cdot p_j^l. \tag{56}$$

PROPOSITION 2. *Let U be a finite dimensional rectangle in $\mathcal{B}([0,1]^d)$ with probability $Q(U) = p > 0$, and $\left\{U_j^d : j = 0, .., 2^d - 1\right\}$ a partition of U induced by sequence of ϵ-good median cuts. Then,*

$$\sum_{j=0}^{2^d-1} p_j^d \cdot \sum_{i=1}^d length_i(U_j^d) \leq \frac{1+\epsilon}{2} \cdot p \cdot \sum_{i=1}^d length_i(U). \tag{57}$$

The proof is a simple consequence of (56).

F.2. Shrinking cell condition for balanced TSP

Let us focus on our balanced TSP $\Pi = \{T_1, T_2, \cdots\}$ of height (d_n), i.e. $\left|\pi_n(Y_1^n)\right| = 2^{d_n}$, $\forall n > 0$. In addition, let us consider $\bar{\Pi} = \left\{\bar{T}_1, \bar{T}_2, \cdots\right\}$, with partition rule $\bar{\pi}_n(y_1^n) \equiv \pi(\bar{T}_n(y_1^n))$, where $\bar{T}_n(y_1^n) \equiv T_n^{\bar{d}_n}(y_1^n)$ and $\bar{d}_n = d \cdot \lfloor d_n/d \rfloor$. It is sufficient to prove the shrinking cell condition for the pruned balanced TSP $\bar{\Pi}$. The reason for this reduction is that by construction the height of $\bar{T}_n(Y_1^n)$ is power of d, and then we can recursively use *Proposition* 2 to bound $\mathbb{E}_Q\left(\sum_{i=1}^d length_i(\bar{\pi}_n(X|Y_1^n))\right)$. More precisely, if we condition to the event $B_n(\epsilon) \in \mathcal{B}(\mathbb{R}^{d\cdot n})$, where all the axis-parallel hyperplanes that induce $\bar{T}_n(Y_1^n)$ are ϵ-good median cuts, from (57) we have the following bound,

$$\mathbb{E}_Q\left(\sum_{i=1}^d length_i(\bar{\pi}_n(X|Y_1^n))\right) \leq \left[\frac{1+\epsilon}{2}\right]^{r_n} \cdot d, \tag{58}$$

with $r_n = \lfloor d_n/d \rfloor$. Let us choose $\epsilon_0 > 0$ sufficiently small in order that $1 + \epsilon_0 < 2$. Then from (58) as $r_n \to \infty$ (when $n \to \infty$), the event $A_n(\epsilon) = \left\{y_1^n \in \mathbb{R}^{d\cdot n} : \mathbb{E}_Q\left(\sum_{i=1}^d length_i(\bar{\pi}_n(X|y_1^n))\right) > \epsilon\right\} \in \mathcal{B}(\mathbb{R}^{d\cdot n})$ is eventually contained in $B_n(\epsilon_0)^c$, $\forall \epsilon > 0$. Consequently, let us focus on the analysis of $\mathbb{Q}^n(B_n(\epsilon_0)^c)$. By definition $B_n(\epsilon_0)^c$ is the event that one of the cuts of $\bar{T}_n(Y_1^n)$ is not ϵ_0-median good. By construction the number of hyperplanes splitting $\bar{T}_n(Y_1^n)$ is given by $(1 + 2 + \cdots + 2^{\bar{d}_n - 1})$, then

$$\mathbb{Q}^n(B_n(\epsilon_0)^c) \leq 2^{\bar{d}_n} \cdot \mathbb{Q}^n(B_n^o(\epsilon_0)) \tag{59}$$

with $B_n^o(\epsilon_0)$ denoting the event that one cut is not ϵ_0-median good. Devroye et al. (1996, *Theorem 20.2)* showed for this case of balanced trees that,

$$\mathbb{Q}^n(B_n^o(\epsilon_0)) \leq 2 \cdot \exp\left(-\frac{n}{2^{\bar{d}_n+2}} \cdot ((1+\epsilon_0)^{1/d} - 1)^2\right), \tag{60}$$

for n sufficiently large. Consequently, from (59) and (60), there exists $K > 0$ such,

$$\mathbb{Q}^n(B_n(\epsilon_0)^c) \leq K \cdot \exp\left(\log(2) \cdot \bar{d}_n - \frac{n}{2^{\bar{d}_n+2}} \cdot ((1+\epsilon_0)^{1/d} - 1)^2\right), \tag{61}$$

$\forall n \in \mathbb{N}$. From the definition of $\bar{d}_n$, we have that $d_n - d < \bar{d}_n \leq d_n$, and consequently from the hypothesis, there exists $(a_n) \approx n^p$ for some $p > 0$, such that

$$\frac{n}{\bar{d}_n 2^{\bar{d}_n}} - \frac{a_n}{\bar{d}_n} \to \infty, \tag{62}$$

which from (61) is sufficient to show that,

$$\frac{\mathbb{Q}^n(B_n(\epsilon_0)^c)}{\exp(-n^p)} \to 0 \tag{63}$$

as n tends to infinity. Finally, $\limsup_n A_n(\epsilon) \subset \limsup_n B_n(\epsilon_0)^c$, $\forall \epsilon > 0$, then given that $\sum_n \mathbb{Q}^n(B_n(\epsilon_0)^c) < \infty$ from (63), and the *Borel-Cantelli lemma*, $\mathbb{E}_Q\left(\sum_{i=1}^d length_i(\bar{\pi}_n(X|Y_1^n))\right)$ tends to zero with probability one. □

References

Barron, A., Györfi, L., van der Meulen, E. C., 1992. Distribution estimation consistent in total variation and in two types of information divergence. IEEE Transactions on Information Theory 38 (5), 1437–1454.

Beirlant, J., Dudewics, E., Györfi, L., van der Meulen, E. C., 1997. Nonparametric entropy estimation: An overview. Int. Math. Statist. Sci. 6, 17–39.

Berlinet, A., Vajda, I., 2005. On asymtotic sufficiency and optimality of quantizations. Journal of Statistical Planing and Inference 136, 4217–4238.

Berlinet, A., Vajda, I., van der Meulen, E. C., 1998. About the asymptotic accuracy of Barron density estimate. IEEE Transactions on Information Theory 44 (3), 999–1009.

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

Cover, T. M., 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. on Electronic Computers 14, 326–334.

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. Wiley Interscience, New York.

Csiszár, I., 1967. Information-type measures of difference of probability distributions and indirect observations. Studia Scient. Mathe. Hung. 2, 299–318.

Csiszár, I., 1973. Generalized entropy and quantization problems. In: Academia (Ed.), Trans. 6th Prague Confe. Information Theory, Statistical Decision Functions, and Random Processes. pp. 159–174.

Csiszár, I., Shields, P. C., 2004. Information theory and Statistics: A tutorial. Now Inc.

Darbellay, G. A., Vajda, I., 1999. Estimation of the information by an adaptive partition of the observation space. IEEE Transactions on Information Theory 45 (4), 1315–1321.

den Hollander, F., 2000. Large Deviations. American Mathematical Society.

Devroye, L., Györfi, L., 1985. Nonparametric density estimation: The L_1 view. Wiley Interscience, New York.

Devroye, L., Györfi, L., Lugosi, G., 1996. A Probabilistic Theory of Pattern Recognition. New York: Springer-Verlag.

Devroye, L., Lugosi, G., 2001. Combinatorial Methods in Density Estimation. Springer - Verlag, New York.

Do, M. N., Vetterli, M., 2002. Wavelet-based texture retrieval using generalized gaussian densities and Kullback-Leibler distance. IEEE Transactions on Image Processing 11 (2), 146–158.

Gessaman, M. P., 1970. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. Ann. Math. Statist. 41, 1344–1346.

Gray, R. M., 1990. Entropy and Information Theory. Springer - Verlag, New York.

Györfi, L., Liese, F., Vajda, I., van der Meulen, E. C., 1998. Distribution estimates consistent in χ^2- divergence. Statistics 32 (1), 31–57.

Györfi, L., van der Meulen, E. C., 1987. Density-free convergence properties of various estimators of entropy. Computational Statistics and Data analysis 5, 425–426.

Györfi, L., van der Meulen, E. C., 1994. Density estimation consistent in information divergence. In: IEEE International Symposium on Information Theory. pp. 35–35.

Halmos, P. R., 1950. Measure Theory. Van Nostrand, New York.

Jain, A., Moulin, P., Miller, M., Ramchandran, K., 2002. Information-theoretic bounds on target recognition performance based on degraded image data. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (9), 1153–1166.

Juang, B. H., Rabiner, L. R., 1985. A probabilistic distance measure for hidden Markov models. AT&T Technical Journal 64 (2), 391–408.

Kullback, S., 1958. Information theory and Statistics. New York: Wiley.

Liese, F., Morales, D., Vajda, I., 2006. Asymptotically sufficient partition and quantization. IEEE Transactions on Information Theory 52 (12), 5599–5606.

Liese, F., Vajda, I., 1987. Convex Statistical Distances. Teubner-Verlag, Germany.

Lugosi, G., Nobel, A. B., 1996. Consistency of data-driven histogram methods for density estimation and classification. The Annals of Statistics 24 (2), 687–706.

Nguyen, X., Wainwright, M., Jordan, M., 2007. Nonparametric estimation of the likelihood ratio and divergence functionals. In: IEEE International Symposium on Information Theory. IEEE.

Nobel, A. B., 1996. Histogram regression estimation using data-dependent partitions. The Annals of Statistics 24 (3), 1084–1105.
Nobel, A. B., July 1997. Recursive partitioning to reduce distortion. IEEE Transactions on Information Theory 43 (4), 1122–1133.
Nobel, A. B., 2002. Analysis of a complexity-based pruning scheme for classification tree. IEEE Transactions on Information Theory 48 (8), 2362–2368.
Novovicova, J., Pudil, P., Kittler, J., 1996. Divergence based feature selection for multimodal class densities. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (2), 218–223.
Paninski, L., 2003. Estimation of entropy and mutual information. Neural Comput. 15, 1191–1253.
Paninski, L., 2008. Undersmoothed kernel entropy estimation. IEEE Transactions on Information Theory 54 (9), 4384–4388.
Piera, F., Parada, P., 2009. On convergence properties of Shannon entropy. Problems of Information Transmission 45 (2), 75–94.
Poor, H. V., Tomas, J. B., 1977. Applications of Ali-Silvey distance measures in the design generalized quantizers for binary decision systems. IEEE Trans. on Comm. 25 (9), 893–900.
Saito, N., Coifman, R. R., 1994. Local discriminant basis. in Proc. SPIE 2303, Mathematical Imaging: Wavelet Applications in Signal and Image Processing, 2–14.
Scott, C., 2005. Tree pruning with subadditive penalties. IEEE Transactions on Signal Processing 53 (12), 4518–4525.
Shannon, C. E., 1948. A mathematical theory of communication. Bell System Tech, J. 27, 379–423; 623–656.
Silva, J., Narayanan, S., 2007. Universal consistency of data-driven partitions for divergence estimation. In: IEEE International Symposium on Information Theory. IEEE.
Silva, J., Narayanan, S., 2009. Discriminative wavelet packet filter bank selection for pattern recognition. IEEE Transactions on Signal Processing 57 (5), 1796–1810.
Singer, Y., Warmuth, M., 1996. Training algorithm for hidden Markov models using entropy based distance functions. In: Advances in Neural Information Processing System 9. Morgan Kaufmann Publishers.
Vajda, I., 2002. On convergence of information contained in quantized observations. IEEE Transactions on Information Theory 48 (8), 2163–2172.
Vapnik, V., 1998. Statistical Learning Theory. John Wiley.
Vapnik, V., 1999. The Nature of Statistical Learning Theory. Springer - Verlag, New York.
Vapnik, V., Chervonenkis, A. J., 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability Apl. 16, 264–280.
Varadhan, S., 2001. Probability Theory. American Mathematical Society.
Vasconcelos, N., 2000. Bayesian model for visual information retrieval. Ph.D. thesis, Mass. Inst. of Technol.
Vasconcelos, N., 2004a. Minimum probability of error image retrieval. IEEE Transactions on Signal Processing 52 (8), 2322–2336.
Vasconcelos, N., 2004b. On the efficient evaluation of probabilistic similarity functions for image retrieval. IEEE Transactions on Information Theory 50 (7), 1482–1496.
Wang, Q., Kulkarni, S. R., Verdú, S., 2005. Divergence estimation of continuous distributions based on data-dependent partitions. IEEE Transactions on Information Theory 51 (9), 3064–3074.
Wang, Q., Kulkarni, S. R., Verdú, S., 2009. Divergence estimation for multidimensional densities via k-nearest-neighbor distance. IEEE Transactions on Information Theory 55 (5), 2392–2405.