Journal of Statistical Planning and Inference 142 (2012) 1716–1732

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On the convergence of Shannon differential entropy, and its connections with density and entropy estimation

Jorge F. Silva *, Patricio Parada

University of Chile, Department of Electrical Engineering, Av. Tupper 2007, Santiago 412-3, Chile

ARTICLE INFO

Article history:
Received 9 June 2011
Accepted 13 February 2012
Available online 21 February 2012

Keywords:
Convergence of probability measures
Shannon information measures
Strong consistency
Density estimation
Differential entropy estimation
Consistency in information divergence
Histogram-based estimators

ABSTRACT

This work extends the study of convergence properties of the *Shannon* differential entropy, and its connections with the convergence of probability measures in the sense of total variation and direct and reverse information divergence. The results relate the topics of distribution (density) estimation, and Shannon information measures estimation, with special focus on the case of differential entropy. On the application side, this work presents an explicit analysis of the density estimation, and differential entropy estimation, for distributions defined on a finite-dimension Euclidean space $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$. New consistency results are derived for several histogram-based estimators: the classical product scheme, the *Barron*'s estimator, one of the approaches proposed by *Györfi and Van der Meulen*, and the data-driven partition scheme of *Lugosi and Nobel*.

1. Introduction

The estimation of *Shannon* information measures, such as the differential entropy and the mutual information (Shannon, 1948; Cover and Thomas, 1991; Gray, 1990; Csiszár and Shields, 2004), is fundamentally related to the problem of distribution (density) estimation (Devroye and Györfi, 1985; Devroye and Lugosi, 2001), as these information measures are functionals of a probability distribution. These two important learning scenarios are well understood and have been systematically studied by the statistical learning community.

Density estimation, when posed as a histogram-based problem, has been characterized extensively in the literature, where strong consistency in the L_1 sense is well understood (Devroye and Györfi, 1985). Necessary and sufficient conditions are known in particular for product non-adaptive histogram-based estimates (Abou-Jaoude, 1976) (see also, Devroye and Györfi, 1985). In recent years, some extensions have been derived using data-dependent partitions (Lugosi and Nobel, 1996), and the family of histogram-based estimators proposed by Barron et al. (1992). In the particular case of the *Barron-type* histogram-based estimator, research has addressed consistency under topologically stronger notions, such as consistency in direct information divergence (I-divergence) (Barron et al., 1992; Györfi and Van der Meulen, 1994), in χ^2-divergence and expected χ^2-divergence (Györfi et al., 1998; Vajda and Van der Meulen, 2001) and in the general family of Csiszár's ϕ-divergence (Berlinet et al., 1998).

For the estimation of information measures, there is a large body of literature dealing with mutual information (MI) and Shannon differential entropy estimation for distributions defined on a finite dimensional Euclidean space $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$, (see Beirlant et al., 1997 and references therein for an excellent review). In particular, consistency is well known for histogram-based and

* Corresponding author.
E-mail addresses: josilva@ing.uchile.cl, jorgesil.edu@gmail.com (J.F. Silva), pparada@ing.uchile.cl (P. Parada).
URL: http://www.ids.uchile.cl/~josilva/ (J.F. Silva).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2012.02.023

kernel plug-in estimates (Beirlant et al., 1997; Györfi and Van der Meulen, 1987). In the case of histogram-based estimators, the standard approach considers non-adaptive product partition (Beirlant et al., 1997), and some extensions have been proposed for data-driven partitions (Darbellay and Vajda, 1999; Silva and Narayanan, 2010b).

A natural question to ask is if there is a connection between the many flavors of consistency for density estimation (in total variation—or L_1 in the case of absolutely continuous distribution with respect to the Lebesgue measure, in (direct-reverse) I-divergence, in Csiszár's ϕ-divergence) and the problem of estimating *Shannon* information measures. We are interested in knowing what flavor of consistent for the density estimation, if any, is sufficient, or needed, to achieve a strongly consistent estimate of the differential entropy. A version of this question was originally stated by Györfi and Van der Meulen (1987). Based on their results (on two histogram-based constructions), they conjectured that extra conditions are always needed to make L_1-consistent histogram-based density estimates consistent for the differential entropy. Silva and Narayanan (2010a, 2010b) have found congruent results when working with data-dependent partitions in the context of MI and Kullback–Leibler divergence (KLD) estimation. In particular, they found stronger conditions for estimating MI and KLD than the one obtained for a consistent estimation of the underlying density in the L_1 sense (Lugosi and Nobel, 1996). These findings, although interesting, are partial in the sense that they are valid only for specific constructions (estimators) and, consequently, general conclusions cannot be derived from them. To the best of our knowledge the stipulation of concrete results connecting the topics of information-measure estimation and density estimation remains an open problem. Such a result (or results) would provide cross-fertilization between these two important lines of research, which to our knowledge have been mostly developed as independent tracts.

Moving in this direction, this work addresses lines of studying the Shannon differential entropy as a functional of the space of probability distribution, in particular, in terms of its convergence properties with respect to deterministic sequences of measures. This is the basic ingredient for understanding consistency, since in the learning scenario we also have sequences of measures, although they are random objects driven by an empirical process (Devroye and Györfi, 1985). Along these lines, Piera and Parada (2009) recently studied this problem and derived a number of conditions on a sequence of probability measures $\{P_n, n \in \mathbb{N}\}$ and the limiting distribution P to guarantee that $\lim_{n\to\infty} H(P_n) = H(P)$.

In the first part of this work, we revisit, refine, and extend these convergence results. From them, we derive concrete relationships between convergence in (reverse and direct) I-divergence and the convergence of Shannon differential entropy. These relationships are obtained under different settings, varying from stronger to weaker conditions on the limiting distribution, and from weaker to stronger conditions in the way the sequence converges to P, respectively. Interestingly, in many of these settings, the convergence on I-divergence suffices to guarantee the convergence of the differential entropy. The results ratify the conjecture raised by Györfi and Van der Meulen (1987), in the sense that convergence on total variation is not sufficient to obtain a convergence of the Shannon differential entropy for the continuous alphabet case. These findings also agree with recent results that demonstrate the discontinuity of the Shannon measures in the countable alphabet scenario (Ho and Yeung, 2009, 2010).

In the second part of this article, we report applying those convergence results to the problem of histogram-based estimation. Specifically we studied four particular estimators: the classical product-type partition estimator (Abou-Jaoude, 1976), the data-driven partition estimator (Lugosi and Nobel, 1996), the Barron histogram-based estimator (Barron et al., 1992), and the histogram-based estimator by Györfi and Van der Meulen (1987). We derived new density-free strong consistency results for each estimator, either in the case of density (in the sense of I-divergence), or in the Shannon differential entropy estimation problem.

The rest of the paper is organized as follows. Section 2 introduces notations and the background needed for the rest of the exposition. Section 3 addresses the convergence of Shannon differential entropy. Section 4 presents the applications of the results in the two previously mentioned statistical learning scenarios. Finally, some of the proofs are presented in the Appendix section.

2. Preliminaries

We start with some basic notations and definitions needed for the rest of the exposition. Let $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ denote the standard k-dimensional Euclidean measurable space equipped with the Borel sigma field (Halmos, 1950; Breiman, 1968). Let $\mathbb{X} \in \mathcal{B}(\mathbb{R}^d)$ be a separable and complete subset of $\mathbb{R}^d$ (i.e., $\mathbb{X}$ is a Polish subspace of $\mathbb{R}^d$). For this space, let $\mathcal{P}(\mathbb{X})$ be the collection of probability measures in $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and let $\mathcal{AC}(\mathbb{X}) \subset \mathcal{P}(\mathbb{X})$ denote the set of probability measures absolutely continuous with respect to λ, the *Lebesgue* measure[1] (Halmos, 1950). For any $\mu \in \mathcal{AC}(\mathbb{X})$, $(d\mu/d\lambda)(x)$ denotes the *Radon–Nikodym* (RN) derivative of μ with respect to λ. In addition, let $\mathcal{AC}_+(\mathbb{X})$ denote the collection of probability measures $\mu \in \mathcal{AC}(\mathbb{X})$ where $(d\mu/d\lambda)(x)$ is strictly positive, Lebesgue almost everywhere in $\mathbb{X}$, i.e., the *support*$(d\mu/d\lambda)$ differs from $\mathbb{X}$ in a set of Lebesgue measure zero.[2] Note that when $\mu \in \mathcal{AC}_+(\mathbb{X})$, then μ and λ are mutually absolutely continuous in $\mathbb{X}$, and consequently, $(d\lambda/d\mu)(x)$ is well-defined and, furthermore, is equal to $((d\mu/d\lambda)(x))^{-1}$ for Lebesgue almost every (Lebesgue-a.e.) point $x \in \mathbb{X}$.

[1] A measure σ is absolutely continuous with respect to a measure μ, denoted by $\sigma \ll \mu$, if for any event A such that $\mu(A) = 0$, then $\sigma(A) = 0$. Consequently $d\sigma/d\mu$ is well-defined, which is the *Radon–Nikodym* derivative or density, and furthermore, $\forall A \in \mathcal{B}(\mathbb{X})$, $\sigma(A) = \int_A (d\sigma/d\mu)\, d\mu$.

[2] Let $f : \mathbb{X} \to \mathbb{R}$ be a real function: then its support is the closure of the set $\{x : f(x) > 0\}$.