An Introduction to Information Theoretic Learning, Part I: Foundations

Daniel G. Silva, Denis G. Fantinato, Jânio C. Canuto, Leonardo T. Duarte, Aline O. Neves, Ricardo Suyama, Jugurta Montalvão and Romis Attux

***Abstract*—With the increasing number of machine learning problems that are out of the linear and Gaussian paradigm, information theoretic learning (ITL) rises as a research field that proposes a modeling method with a wealthier statistical treatment of the adaptation criterion. In the first part of this tutorial, we introduce the main concepts of ITL and a key set of estimators that enable the implementation of algorithms, in the context of a wider view independent of the differentiability property.**

***Index Terms*—ITL, information theory, entropy, Rényi.**

I. Introduction

THANKS to the pioneering work of researchers like Wiener [1] and Kolmogorov [2], adaptive signal processing emerged in the last century as an engineering discipline based on the notion of machine learning and on extensive statistical modeling. The development of this new discipline took place according to two central premises: (i) linearity of the processing structure and (ii) parameter adaptation based on second-order statistics, such as covariance, correlation or the second moment of an error signal.

The choice for linear models is related to factors like parsimony and mathematical tractability. Moreover, if the signals of interest obey a Gaussian model, which can be the case due to intrisic features or to the "Gaussianizing effect" of the central limit theorem [3], a linear structure can be statistically optimal (e.g. according to a maximum likelihood formulation) [4]. However, the extraordinary development of computer technology and an increasing demand for high-performance information processing has been responsible for popularizing the use of machine learning methods, including adaptive nonlinear models — like neural networks and fuzzy systems. In addition to that, the consolidation of unsupervised filtering theory has broadened the statistical range of adaptation criteria with the adoption, for instance, of higher-order statistics.

Daniel G. Silva is with the Dep. of Electrical Engineering, University of Brasília - UnB, Brasília, DF, Brazil. e-mail: danielgs@ene.unb.br.

Denis G. Fantinato and Romis Attux are with the Lab. of Signal Processing for Communications - DSPCom, University of Campinas - Unicamp, Campinas, SP, Brazil. e-mail: {denisgf, attux}@dca.fee.unicamp.br

Leonardo T. Duarte is with the School of Applied Sciences (FCA), University of Campinas - Unicamp, Limeira, SP, Brazil. e-mail: leonardo.duarte@fca.unicamp.br

Jânio C. Canuto and Jugurta Montalvão are with the Dep. of Electrical Engineering, Federal University of Sergipe, São Cristóvão, SE, Brazil. e-mail: janio.canuto@gmail.com, jmontalvao@ufs.br.

Ricardo Suyama and Aline O. Neves are with the Engineering, Modeling and Applied Social Sciences Center, Federal University of ABC, Santo André, SP, Brazil. e-mail: {ricardo.suyama, aline.neves}@ufabc.edu.br.

Digital Object Identifier: 10.14209/jcis.2016.6

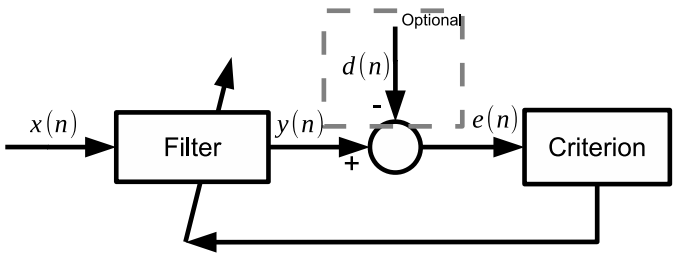


Fig. 1. The general information filtering problem.

These two trends — the use of more powerful signal processing structures and of more extensive statistical information regarding the underlying data — are, in a certain sense, combined and even harmonized under the aegis of the notion of information theoretic learning (ITL). ITL criteria and algorithms share an essential feature: they are based on statistical entities derived from information theory (IT) [5], like entropy and mutual information. The relevance of these entities can be justified in terms of their probabilistic structure, which, in principle, allows a more thorough statistical characterization than that provided by second- or even specific higher-order moments. This broader statistical perspective is what explains the close association between ITL and nonlinear / nongaussian scenarios.

Although the origins of the discipline of ITL can be traced to several key branches of the fields of adaptive filtering and machine learning, it is beyond any doubt that the efforts of the group led by Prof. José Principe were instrumental in providing it with a unified conceptual / theoretical basis and with an important and versatile framework based on Rényi's entropy [6]. Presently, it can be safely stated that ITL is well established in the context of information processing theory, being the knowledge of its main formulations and algorithms essential to all researchers working in this field.

In the big picture, ITL algorithms tackle problems which one can often see as a generic filtering task, as depicted in Figure 1. Three aspects, in this context, should be considered: (i) the criterion to use, (ii) the definition of a filtering model and its free parameters, and (iii) the strategy to adapt the parameters according to the criterion.

This work focuses on aspects (i) and (ii), by discussing important information-theoretic estimators that enable the subsequent derivation of ITL criteria and by illustrating problems where a formulation consisting of a linear / nonlinear model adapted by an ITL criterion is a possible (and

mostly interesting) solution. As the reader shall see, the estimators and, consequently, the criteria set do not necessarily engender differentiable cost functions (which, according to [7], is recommended). Hence, the decision concerning the aforementioned aspect (iii) can be made within an extended scope, beyond gradient-based search strategies and including, for example, use of bio-inspired meta-heuristics (e.g. genetic algorithms [8], artificial immune systems [9]). Furthermore, the non-obligation of differentiability makes the choice of a proper IT-based criterion for the problem more flexible.

This tutorial was conceived with exactly this motivation in mind. It is structured in two parts: the first presents the foundations of ITL — information theory, Rényi's entropy and statistical estimation; the second part focuses on ITL methods and applications, giving the reader an overview of representative formulations and practical scenarios. This first part is organized as follows: Section II presents elements of information theory; Section III discusses Rényi's proposal, which extends the mathematical treatment of the concepts brought forward by Shannon; Section IV discusses some of the essential estimators employed in ITL criteria and Section V brings the conclusions and final remarks.

II. Information Theory

The last century saw a revolutionary development in the technologies of data transmission, storage and processing. This process has given rise to a number of disciplines that have been unified under the aegis of the concept of information [10].

In spite of important contributions like [11], [12] and [13], it is generally accepted that the research field referred to as information theory took a definite shape in a work authored by one of the most remarkable scientific figures of the 20th century: Claude Elwood Shannon. The work's title indicates the vastness of its scope — "A Mathematical Theory of Communication" [14] — and, nonetheless, after the final page, the reader cannot but have the impression that the text has been up to the highest expectations.

A. Discrete sources

Shannon's work deals with information sources of discrete and continuous natures, and also establishes divisions between transmission processes with and without noise. In its first part, which is devoted to the analysis of the discrete and noiseless case, Shannon presents a fundamental quantity H:

$$H(X) = -K \sum_{x \in \mathcal{X}} p_X(x) \log p_X(x), \tag{1}$$

where $p_X(x)$ is the probability mass function (PMF) of the random variable (RV) X.

After indicating that K can be set to unity without loss of generality, Shannon mentions that quantities of the form shown in (1) are relevant as "measures of information, choice and uncertainty" [14]. A connection with statistical mechanics is duly established and $H(\cdot)$ is termed *entropy*. There is a story, mentioned in [15], that Shannon's choice was the result of John von Neumann's advice, who supposedly said that the use of the term "entropy" would give him "...a great edge in debates, because no one knows what entropy really is."

The proposed expression for H is justified in terms of some properties that a definition of entropy should have, like continuity over probability values and monotonic increase with respect to the number of possible events in the uniform case [14]. Afterwards, some important properties of H are given, some of which are useful to our future discussions:

1) $H = 0$ if and only if there is a single event with non zero (i.e. unit) probability.
2) Given a number of possible outcomes n, the entropy is maximal if all probabilities are set to $1/n$. This means that the most entropic case arises from a uniform distribution — uncertainty is maximal.
3) The joint entropy of two RVs, X and Y, is defined as[1]:

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \tag{2}$$

This definition gives rise to an important inequality:

$$H(X,Y) \leq H(X) + H(Y) \tag{3}$$

being equality possible only when X and Y are statistically independent, i.e., when:

$$p(x,y) = p(x)p(y). \tag{4}$$

The inequality in (3) is also intuitive: whenever there is a certain degree of dependence between variables, the uncertainty associated with their joint knowledge will be smaller than the sum of the uncertainties associated with them in separate. When the variables are independent, considering one of them is useless to reduce the amount of uncertainty associated with the other, equality holds.
4) If one defines the entropy of a conditional distribution as a conditional entropy of the form:

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \tag{5}$$

it can be shown that:

$$H(X,Y) = H(X) + H(Y|X) \tag{6}$$

This means that the joint uncertainty of X and Y can be understood as the amount of uncertainty associated with X plus the amount of uncertainty associated with Y when X is known. Notice that (6) can also be written in terms of $H(Y)$ and $H(X|Y)$.
5) From (3) and (6), it is possible to obtain the following expression:

$$H(Y) \geq H(Y|X) \tag{7}$$

This expression reveals that the uncertainty associated with a random variable is never increased by knowledge of another random variable. The limit case — that of independence — accounts for potential equality.

[1] In order to simplify the notation, we shall henceforth omit the subscript of the RV associated with the PMF.

After defining the powerful concept widely known as that of *typical sequence*[2], Shannon proves a theorem that establishes the entropy of the source as a limit to the achievable efficiency of any coding process (notice that, in this noiseless case, the underlying idea is that of lossless compression).

In the following, the case of discrete channels in the presence of noise is dealt with. The idea is to analyze the possibility of reliable data transmission even when it is known that there is always a chance of equivocation in the reconstruction process performed at the receiver. Naturally, "raw data" transmission is ruled out — some sort of coding is the only hope, but to determine the extent of what can be achieved is a far from trivial task.

Shannon, with his lucid and light style, argues that, for a noisy channel, the "rate of actual transmission" [14] is given by $H(X)-H(X|Y)$ — being X related to the transmitter and Y to what is received. This quantity is what we call mutual information (MI), and we will use in its definition the notation that will be adopted throughout this work:

$$I(X;Y) = H(X) - H(X|Y). \tag{8}$$

His line of reasoning is direct: $H(X)$ represents the entropy of the source and $H(X|Y)$ is a measure of equivocation i.e. of the average ambiguity of the received signal. Notice that (8) can be rewritten in two forms:

$$I(X;Y) = H(Y) - H(Y|X), \tag{9}$$

and

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{10}$$

Shannon provides the reader with nice interpretations of both expressions in the context at hand. Expression (9) indicates the amount of received information minus that which is due to noise, and (10) is "the sum of the two amounts less the joint entropy and therefore in a sense is the number of bits per second common to the two." [14]. Also, if one writes $I(X;Y)$ in terms of probabilities, it is possible to obtain the relation between the Kullback-Leibler divergence

$$D_{KL}(p;q) = \sum_u p(u) \log \frac{p(u)}{q(u)}, \tag{11}$$

where p and q are two PMFs, and the mutual information:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D_{KL}(p(x,y); p(x)p(y)). \end{aligned} \tag{12}$$

He then proceeds to prove an astounding result: if a value C called channel capacity is not exceeded by the rate of information production at the source, there is a coding system capable of giving rise to an arbitrarily small error rate. Thus, a noisy channel can be used to send information with an arbitrarily small reconstruction error with a rate that does not have to tend towards zero, but can actually be set to a finite bounded value. We will not discuss in detail the elegant method used to prove this result, but it is important to remark that the capacity of a given channel is defined in terms of the maximization of the mutual information between the variables associated with the transmitter and the receiver:

$$C = \max_{p(x)} I(X;Y). \tag{13}$$

This maximization is performed with respect to the source probability structure.

B. Continuous sources

In the sequence, the case of continuous sources becomes the focus of the work. A first aspect that will be of paramount importance here is the extension of the definition of entropy to the case of a continuous random variable with probability density function (PDF)[3] $f(x)$:

$$h(X) = -\int f(x) \log f(x) dx. \tag{14}$$

This is an intuitive choice that preserves many of the properties valid for the discrete case, but there are also important differences. A relevant aspect pointed out by Shannon is that, in contrast with the discrete case, in which entropy corresponds to an absolute uncertainty measure, in the continuous case, the definition leads, in general, to different results for different coordinate systems. However, the difference between entropies is not affected by this potential modification, which means that quantities like channel capacity will be immune to it [14]. The entropy of a continuous variable, as defined in (14), is also called *differential entropy* [5].

Properties (6) and (7) are directly applicable to this case, as well as the definitions given in (9), (10) and (12), considering that the differential entropy definition, the integral operator and probability density functions substitute the discrete-valued counterparts. Two important properties are also given in Shannon's paper:

1) Under the constraint that a RV is bounded to a finite volume of the space, the probability density with maximum entropy is the uniform density.
2) If the second-order moment of a RV is fixed *a priori*, the probability density with maximum entropy is the Gaussian density.

The remainder of Shannon's paper establishes a number of key results for continuous sources and channels. Due to this work, information theory became a research field *per se* and a great amount of studies were developed to understand and consolidate Shannon's initial contributions. As these results are beyond the scope of the next sections, we will no longer follow the thread of his work. Instead, we shall focus on a generalized measure of information that extended the original definition in (1) and allowed the development of many ITL algorithms.

[2]It is important to mention the idea of asymptotic equipartition, the basis for defining "typical". If $X_1, X_2, ..., X_n$ are independent and identically distributed, with PMF $p(\cdot)$, it is possible to show that the following relationship is, in probability, valid: $-1/n \log p(x_1, ..., x_n) \rightarrow H(X)$. In [5], this is illustrated by the sentence "almost all events are almost equally surprising".

[3]With the same purpose of simplification, we shall omit the subscript in $f(x)$.

III. RÉNYI'S ENTROPY

Alfred Rényi, in the mid 1950s, developed a mathematical generalization of Shannon's entropy, usually called Rényi's α-entropy. Its definition for the one-dimensional and continuous case is [16]:

$$h_\alpha(X) = \frac{1}{1-\alpha}\log\int f^\alpha(x)dx. \tag{15}$$

Rényi's intention was to develop a generalized measure of information having the additivity property of statistically independent systems and respecting Kolmogorov's axioms of probability. This definition remained practically ignored in the field of communications, where Shannon's proposal has been successfully adopted, possibly due to its direct connections with information flow in data transmission systems. However, the application of Rényi's entropy took place in other areas, such as coding theory, quantum mechanics, chaotic dynamic systems and as a measure of diversity in economy [7].

The α parameter of Rényi's entropy allows several measures of uncertainty associated with the same distribution. Two scenarios are worth mentioning: (*i*) if $\alpha \to 1$, Renyi's measure converges to Shannon's entropy; and (*ii*) if $\alpha = 2$, we obtain the *quadratic entropy*

$$h_2(X) = -\log\int f^2(x)dx = -\log E[f(x)]. \tag{16}$$

The quadratic entropy plays an important role in ITL because it engenders a family of estimators that have interesting characteristics (from the machine learning perspective) such as being non-parametric, continuous and computationally straightforward. Although it is possible to obtain estimators for any value of α, the quadratic case, allied to a kernel-based method for probability density estimation, allows a convenient evaluation of the integral in (16), which will be shown in Section IV.

The argument of the log function, $E[f(x)]$, is called *information potential* (IP) — $V_2(x)$ — and, in terms of adaptation, we may consider just the optimization of $V_2(x)$, since $h_2(X)$ is a monotonic function and we are only interested in the extrema of the cost functions. Furthermore, the expression $V_2(X) = E[f(x)]$ carries meaning in itself as the expected value of the probability distribution, in a context where $f(x)$ is a transformation of the original random variable.

Rényi also proposed in his studies [6] a generalized divergence measure in probability spaces, the Rényi's α-divergence:

$$D_\alpha(f;g) = \frac{1}{\alpha-1}\log\int f(x)\left(\frac{f(x)}{g(x)}\right)^{\alpha-1}dx. \tag{17}$$

Analogously to the entropy, the α-divergence converges to the Kullback-Leibler divergence when $\alpha \to 1$. Moreover, it is straightforward to obtain the order α mutual information:

$$I_\alpha(X;Y) = \frac{1}{\alpha-1}\log\int\int\frac{f^\alpha(x,y)}{(f(x)f(y))^{\alpha-1}}dxdy, \tag{18}$$

where $f(x)$ and $f(y)$ are the marginal densities of X and Y, respectively. The general case for more than two random variables is simple to obtain in a manner analogous to that of Shannon's mutual information [5].

The α-divergence and the Kullback-Leibler divergence both present a drawback, which is their asymmetric character. Thus, they are not strictly distances, but there are other proposals that fulfill the simmetry requirement, such as the Cauchy-Schwarz Divergence [17]:

$$D_{CS}(f;g) = -\frac{1}{2}\log\frac{\left(\int f(x)g(x)dx\right)^2}{\int f^2(x)dx\int g^2(x)dx}. \tag{19}$$

With this distance metric it is possible to obtain an alternative independence measure, the Cauchy-Schwarz Quadratic Mutual Information (QMI) [7]. When comparing to classical Shannon's definition or Rényi's α-order mutual information, the QMI measure is a different approach that may provide a more appealing mathematical treatment, in terms of estimation, and also presents interesting results as a criterion for a series of ITL algorithms [7].

IV. ESTIMATORS

As presented in the previous sections, all definitions of information measures proposed by either Shannon or Rényi require knowledge of the probability mass function, in the discrete case, or of the probability density function in the continuous case.

Since, in many practical cases, only data samples whose probability structure is not known in advance are available, probability and information-theoretic estimators are required to implement ITL adaptive algorithms. Initially, the estimators proposed by Principe et al. [7] are introduced, which raise the possibility of deriving gradient-based algorithms; afterwards, we present estimators that are not necessarily differentiable, as well as two basic estimators for discrete data.

The nonparametric estimators presented in this text are just a few members of a much larger set, which include relevant contributions as the entropy estimator based on maximum entropy distributions [18] and other relevant work that the reader should refer to [19], [20] for a more profound review on this particular subject.

A. The Parzen window density estimator as an extreme case of Gaussian mixture model

The Parzen window method for probability density function estimation is a kernel based strategy that can be used to approximate the PDF $f(\mathbf{x})$ of a vector of continuous random variables X. The problem might be stated as follows: let $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ be a set of N d-dimensional samples drawn according to the unknown PDF $f(\mathbf{x})$. We assume that exists an adequate approximation $\hat{f}(\mathbf{x})$ given by

$$\hat{f}(\mathbf{x}) = \sum_{k=1}^{N}\alpha_k\phi(\mathbf{x}-\mathbf{x}_k, h_k) \tag{20}$$

where $\phi(\cdot)$ is the window function and h is the window width parameter. Parzen showed that $\hat{f}(\mathbf{x})$ converges to the true density if $\phi(\cdot)$ and h are properly selected [21]. The window function is required to be a finite-valued non-negative density function where

$$\int\phi(\mathbf{y}, h)d\mathbf{y} = 1, \tag{21}$$

and the width parameter must be a function of N such that

$$\lim_{N\to\infty} h(N) = 0, \tag{22}$$

and

$$\lim_{N\to\infty} Nh^d(N) = \infty. \tag{23}$$

Rectangular and Gaussian window functions are commonly used. For the latter case, we might rewrite (20) as an extreme case of a Gaussian mixture model (GMM) [22], whose general formulation is:

$$\hat{f}(\mathbf{x}|\mathbf{\Theta}) = \sum_{k=1}^{M} \alpha_k G(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) \tag{24}$$

where M is typically much smaller than N, $\sum \alpha_k = 1$ and $\alpha_k \geq 0$. We further denote $\mathbf{\Theta} = \{A, U, C\}$ as the mixture parameter, where $A = [\alpha_1, ..., \alpha_M]$, $U = [\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_M]$, $C = [\mathbf{\Lambda}_1, ..., \mathbf{\Lambda}_M]$) and $G(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k)$ is the k-th Gaussian kernel, with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{\Lambda}_k$:

$$G(\boldsymbol{x}|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Lambda}_k|^{\frac{1}{2}}} e^{\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \mathbf{\Lambda}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}{2}\right)}. \tag{25}$$

Denoting the likelihood of $\mathbf{\Theta}$ as $L(\mathbf{\Theta}) = \hat{f}(\boldsymbol{x}|\mathbf{\Theta})$, the problem consists of finding the optimal parameter vector, $\mathbf{\Theta}_o$, that maximizes the log-likelihood:

$$\mathbf{\Theta}_o = \arg\max_{\mathbf{\Theta}} (\log L(\mathbf{\Theta})). \tag{26}$$

If the N samples are independent, we may write

$$\log L(\mathbf{\Theta}) = \log\left(\prod_{k=1}^{N} \hat{f}(\boldsymbol{x}_k|\mathbf{\Theta})\right) = \sum_{k=1}^{N} \log \hat{f}(\boldsymbol{x}_k|\mathbf{\Theta}) \tag{27}$$

A well-known and widely used method for solving this problem is the EM (Expectation-Maximization) algorithm [23]. However, aside from not being the fastest method, this algorithm presents some other issues, such as [24]: (i) potentially bad convergence, depending on data distribution and initial parameter choices; (ii) the likelihood-based criterion presents local maxima that might result in bad models, especially for small datasets.

As a matter of fact, it is known that the likelihood is not particularly suited to high-dimension problems, and even to some low dimensional cases [25]. A solution can be to make use of regularization methods, which constrain the optimization problem to enhance the generalization performance [26].

Although Parzen model is nonparametric, whereas GMM can be considered a semi-parametric model, as mentioned before in (24), an interesting point of view for the Parzen model is that of an extreme case of an intrinsically regularized Gaussian mixture model [27] where $M = N$, thus yielding (*i*) $\alpha_k = 1/N$ (i.e. all Gaussian kernels have the same weight); (*ii*) $\boldsymbol{\mu}_k = \boldsymbol{x}_k$, the samples are the Gaussian kernels centers; and (*iii*) $\mathbf{\Lambda}_k = \sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is a $d \times d$ identity matrix, in such way that all kernels are identical and isotropic.

Given such constraints, the single free parameter to be optimized is σ, the Parzen window width – which can be optimized through cross-validation, instead of the EM (see Section 4.3 of [19]) –, and (24) becomes:

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^{N} G(\mathbf{x}|\mathbf{x}_k, \sigma^2\mathbf{I}). \tag{28}$$

Alternatively, the value of σ can be directly obtained via an adaptive kernel density estimator which relies on concepts from linear diffusion processes [28]. The estimator obtains (28) as the solution of a partial differential equation, the well-known Fourier heat equation and, additionally, calculates the Gaussian kernel bandwidth automatically, without the common assumption of data normality.

B. Rényi's entropy estimators

The Parzen window method provides an estimate of the underlying probability function and, consequently, we can use it in either Shannon's or Rényi's definition of differential entropy, resulting in entropy estimators commonly known as *plug-in* estimators [20].

Shannon's entropy, nonetheless, still requires a possibly demanding evaluation of an integral, while, for Rényi's entropy, [7] shows that, in the quadratic case and with Gaussian kernels, this operator will result in a straightforward expression[4].

Consider the unidimensional case: if we rewrite (16) replacing the theoretical PDF with Parzen's estimator (28), then

$$\begin{aligned}
\hat{h}_2(X) &= -\log \int \hat{f}^2(x)dx \\
&= -\log \int \left[\frac{1}{N}\sum_{k=1}^{N} G(x|x_k, \sigma^2)\right]^2 dx \\
&= -\log \frac{1}{N^2} \int \left[\sum_{j=1}^{N}\sum_{k=1}^{N} G(x|x_k, \sigma^2)G(x|x_j, \sigma^2)\right] dx \\
&= -\log \frac{1}{N^2} \sum_{j=1}^{N}\sum_{k=1}^{N} \int G(x|x_k, \sigma^2)G(x|x_j, \sigma^2)dx \\
&= -\log \frac{1}{N^2} \sum_{j=1}^{N}\sum_{k=1}^{N} G(x_k|x_j, 2\sigma^2).
\end{aligned} \tag{29}$$

Note that the integral of the product of two Gaussians is exactly a Gaussian evaluated at the difference between arguments, with a variance equal to the sum of the original two variance values [7]:

$$\int G(x|\mu_1, \sigma_1^2)G(x|\mu_2, \sigma_2^2)dx = G(\mu_1|\mu_2, \sigma_1^2 + \sigma_2^2). \tag{30}$$

As already mentioned in Section III, the quadratic entropy estimator is quite appealing from the perspective of machine learning because it is non-parametric, continuous and differentiable. Such properties give support to the design of adaptive algorithms that obtain their solutions based on gradient search techniques.

[4]The Gaussian kernel is able to maintain its functional form under convolution. Notwithstanding, any other kernel function with peak at the origin can be equally employed: the resulting kernel function, in this case, is the convolution of the original kernel with itself [7].

Analogously to the theoretical expression of quadratic entropy, only the argument of the log function in (29) can be considered for adaptation purposes, which yields the information potential estimator:

$$\hat{V}_2(X) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} G(x_k|x_j, 2\sigma^2). \tag{31}$$

When concerning ITL algorithms that employ differentiable objective functions, the IP estimator is one of the most important techniques. Expression (31) shows that it depends on σ — the Parzen window width —, which must be carefully selected by the user due to direct (and hard to predict) effects that may provoke on the shape of optimization surface. Besides the aforementioned method based on linear diffusion processes [28], the use of cross-validation routines [4] is a pragmatic and well studied strategy (see Section 5.2 of [19]), in this case, to select a kernel bandwidth in accordance to the data range and problem characteristics.

The aforementioned plug-in approach can be repeated for any value of α, obtaining a generalized version of Rényi's entropy estimator, with the caveat that the integral is no longer exactly evaluated and the corresponding expectation is replaced by a sample mean estimation [7].

To develop estimators for the mutual information and divergence measures presented in Section III, the same procedure of adopting Parzen window with Gaussian kernels leads to expressions that are defined in terms of the IP quantity. For example, the Cauchy-Schwarz divergence estimator is

$$\hat{D}_{CS}(f;g) = \log\frac{\hat{V}_f\hat{V}_g}{\hat{V}_c^2}, \tag{32}$$

where $\hat{V}_f$ and $\hat{V}_g$ are the IPs with respect to the underlying PDFs $f(x)$ and $g(x)$, respectively, and $\hat{V}_c$ is the cross-information potential, given by

$$\hat{V}_c = \frac{1}{N_f N_g}\sum_{j=1}^{N_g}\sum_{k=1}^{N_f} G(x_{f_k}|x_{g_j}, 2\sigma^2), \tag{33}$$

where N_f and N_g are the total number of samples associated with distributions $f(x)$ and $g(x)$, respectively.

C. Entropy estimators based on order statistics

In this section we return our focus to Shannon's differential entropy, presented in (14), which can be alternatively rewritten in terms of the quantile function of X, denoted as $Q(u)$, as shown, for instance, by Pham [29]. In fact, since the quantile function is defined as the inverse function of the cumulative distribution function of X, $F(x)$, it follows that

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{Q'(F(x))}. \tag{34}$$

Therefore, after replacing (34) into the logarithm term in (14), we obtain

$$h(X) = \int f(x)\log Q'(F(x))\,dx. \tag{35}$$

Finally, if one considers the change of variables $du = f(x)dx$, then (35) can be written as follows

$$h(X) = \int_0^1 \log Q'(u)\,du. \tag{36}$$

According to expression (36), a possible way to estimate entropy can be based on a two-step procedure consisting of (i) estimating the quantile function of X, and (ii) replacing the obtained estimate into (36). The key point here is that good approximations of the quantile function can be obtained by a straightforward approach based on order statistics [29], [30]. Indeed, given a set of N samples, it can be shown that [30]

$$Q\left(\frac{i}{N+1}\right) \approx x_{(i:N)},\ i = 1, \ldots, N, \tag{37}$$

where $x_{(i:N)}$ corresponds to a sample of the i-th order statistics, so that:

$$x_{(1:N)} \leq x_{(2:N)} \leq \ldots x_{(N:N)}.$$

In other words, the quantile function can be estimated by simply sorting these realizations. Then, after some calculation considering (36) and (37), the following entropy estimator can be obtained

$$\hat{h}(X) = \sum_{l=2}^{L}\log\left(\frac{x_{(c_l:N)} - x_{(c_{l-1}:N)}}{u_l - u_{l-1}}\right)\frac{u_l - u_{l-1}}{u_L - u_1}, \tag{38}$$

where

$$c_l = \left\lceil \frac{N u_l}{u_L} \right\rceil. \tag{39}$$

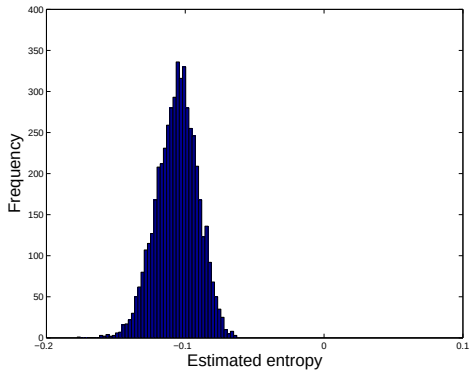
This approximation is necessary because a given u_l within the integration grid will not necessarily satisfy the condition $i/(N+1)$. In this case, $Q(u_l)$ is approximated by $Q(c_l/(N+1))$, which, in turn, can be approximated by $x_{(c_l:N)}$ according to (37).

If, on the one hand, the estimator in (38) provides a simpler solution in terms of computational complexity, on the other hand, it requires a significant number of samples to achieve a good performance. To illustrate this point, let us consider the problem of estimating the differential entropy of a random variable uniformly distributed in $[0, 1]$. In Figure 2, we show the histograms of the estimated entropies obtained by (38) after 5000 executions for the cases of 500 and 5000 samples. Bearing in mind that the theoretical value of the entropy in this is case is 0, one can note a great amount of bias in the case where 500 samples are considered. Moreover, the variance of the estimator in that case is high. Finally, another limitation of order-statistics-based estimators is that they can only be applied in the case of 1-D random variables.

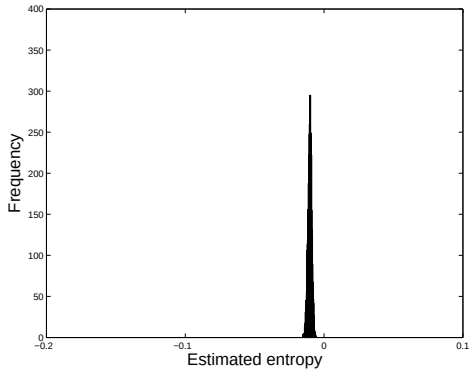
D. K-nearest neighbor and graph-based entropy estimation

In consonance with (14), differential entropy can also be defined in terms of the following statistic:

$$h(X) = -E[\log f(x)]. \tag{40}$$



(a) $N = 500$ samples



(b) $N = 5000$ samples

Fig. 2. Entropy estimation through (38).

Therefore, if a estimator for $\log f(x)$, denoted as $\overline{\log f(x_i)}$ is available, then the entropy can be estimated through the following average

$$\hat{h}(X) = -\frac{1}{N}\sum_{i=1}^{N}\overline{\log f(x_i)}. \quad (41)$$

This approximation is the basis of different approaches for estimating entropy, such as histogram-based methods, which are discussed in Section IV-E, and k-nearest neighbor (k-NN) based methods, which are briefly described in this section.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ denote N outcomes of a random variable X of dimension d. It is possible to define a RV ϵ associated with the distance, e.g. the Euclidean distance, between $\mathbf{x}_i$ and its k-nearest neighbor. As shown, for instance, in [31], ϵ can be used to approximate $\log f(\mathbf{x}_i)$ in the following manner

$$\log f(\mathbf{x}_i) \approx \psi(k) - \psi(N) - \log(c_d) - dE[\log(\epsilon)], \quad (42)$$

where $\psi(\cdot)$ is the digamma function and c_d is the volume of a unit sphere in dimension d. By using this expression in (41), the following estimator can be derived:

$$\hat{h}(X) = -\psi(k) + \psi(N) + \log(c_d) + \frac{d}{N}\sum_{i=1}^{N}\log\epsilon_i. \quad (43)$$

The first three terms in this estimator have constant values, while the last one requires the calculation of the distance of $\mathbf{x}_i$ to its k-nearest neighbor.

An interesting feature of k-nearest neighbor-based entropy estimators is that they can be applied regardless of the dimension of data. In [31], for instance, the authors proposed a method to estimate mutual information between random variables based on this conceptual framework.

In [32] the k-NN strategy is revisited in the context of a graph-based approach. Basically, a directed graph is built connecting all samples of a certain distribution to a given set of nearest neighbors. The sum of the p-th powers of the Euclidean lengths of its edges is then calculated, and, using this distance, it is possible to estimate Rényi's α-entropy using a straightforward formula that requires also the estimation of a constant value derived from the structure of a graph engendered by a uniform distribution. Using the notion of invariance of mutual information to rescaling and the so-called copula transformation, the obtained results can be adapted to the estimation of the related mutual information. Theoretical analyses regarding the consistency of the estimator and simulation results reveal the usefulness of the method.

Finally, it is worth mentioning that methods based on minimum spanning graphs [33] can be employed to estimate both Rényi's α-entropy and divergence. These methods are also based on the sum of a power of the Euclidean distances of the edges of a minimal graph, and are, in contrast with plug-in methods, faster in terms of convergence and economical in terms of *a priori* choice and fine tuning of parameters.

E. Histogram-based estimators for mutual information

A histogram is a straightforward but coarse PDF estimator, whose main attractiveness lies in its simplicity of implementation and use. As an illustration for the use of histograms in MI estimation, we adapt the very illustrative example presented in [34], where a collection of N simultaneous measurements of two continuous variables, X and Y, are considered. First, a 2-D histogram is used, with $M_x \times M_y$ bins, each bin corresponding to a rectangular area, $a_{i,j}$, with center at (c_i, c_j), $(i = 1, 2, ..., M_x$ and $j = 1, 2, ..., M_y)$. Being $k_{i,j}$ the number of measurements that lie within $a_{i,j}$, the probability of a random measurement (X, Y) falling onto $a_{i,j}$ is estimated as being:

$$P[(X,Y) \in a_{i,j}] \approx f_{i,j} = \frac{k_{i,j}}{N}$$

and the *naive* mutual information estimator is:

$$\hat{I}_N(X;Y) = -\sum_{i=1}^{M_x} f_i \log f_i - \sum_{j=1}^{M_y} f_j \log f_j + \sum_{i=1}^{M_x}\sum_{j=1}^{M_y} f_{i,j}\log f_{i,j} \quad (44)$$

where $f_i = \sum_{j=1}^{M_y} f_{i,j}$ and $f_j = \sum_{i=1}^{M_x} f_{i,j}$.

To illustrate this naive estimator with a simple experiment, Figure 3 presents a plot of $N = 300$ independent and uniformly distributed points $(x_n, y_n) : x_n, y_n \in [0, 1]$, along with a regular grid of bins, with $M_x = M_y = 10$. The figure also shows a histogram of 500 independent instances

of $\hat{I}_N(X;Y)$, where a clear estimator bias is noticed, since the true value of the mutual information $I(X;Y)$ is zero.

Indeed, through this naive method based on histograms and relative frequencies, mutual information is systematically overestimated. For the illustration corresponding to Figure 3, $\hat{I}_N(X;Y) \approx 0.15 \pm 0.02$, instead of the true value $I(X;Y) = 0$. This bias is due to all terms $\log(f_i)$, in (44), because, though relative frequency is an unbiased maximum likelihood probability estimator, the logarithmic (nonlinear) transformation of it produces a bias.

To reduce this effect, Miller [35] proposed a popular $O(1/N)$ compensation rule

$$E[\hat{H}(X)] \approx H(X) - \frac{M_x - 1}{2N}. \quad (45)$$

Moreover, in order to extend the estimation up to $O(1/N^2)$ corrections, Harris [36] proposed

$$E[\hat{H}(X)] \approx H(X) - \frac{M_x - 1}{2N} + \frac{\left(1 - \sum_{i=1}^{M_x} \frac{1}{f_i}\right)}{12N^2}. \quad (46)$$

More recently, Paninski [37] applied Bernstein approximating polynomials to obtain corrections of order greater than or equal to two. Further details can be found in [38].

One interesting alternative for MI estimation with histograms is due to Darbellay and Vajda [39], who proposed an adaptive partitioning of the observation space to keep bins symmetrically balanced, thus reducing the bias. Darberllay's method is closely related to a rarely cited method proposed by M. P. Gessaman, in 1970, for nonparametric density estimation based on statistically equivalent blocks, cited and briefly explained in Section 5.2 of [19]. The authors of [39] empirically concluded that "the nonparametric estimator appears to be asymptotically unbiased and efficient." The method explanation presented in [39] is too long to be reproduced here, so we just illustrate it, step-by-step, through one of the experiments presented in Section III - Table I of the original paper, with an alleviated notation.

We start by considering $N = 500$ pairs (x_i, y_i), $i = 1, 2, ..., N$, independently drawn from a Gaussian source, with null mean vector and covariance matrix given by $R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$. The mutual information between the corresponding variables X and Y, in this case, is analytically given by

$$I_{\text{Gauss}}(X;Y) = -\frac{1}{2}\log(1 - r^2).$$

Figure 4 illustrates 500 pairwise instances of these variables with $r = 0.3$.

The method starts by independently splitting X-data and Y-data by their respective middles, thus yielding 2 subgroups of 250 points in each subspace. The cartesian product of these subsets are represented in Figure 4 as A, B, C and D, which corresponds to a first level partitioning of the whole set, or a first coarse quantization grid. If X and Y were independent, we would expect 125 points in each subset. Instead, in one random experiment with $r = 0.3$, we found 151, 99, 99 and 151 points in subsets A, B, C and D, respectively.

This deviation between expected and observed number of points per subset is a clear evidence of non-independence, and at least one subset must be split again into smaller sub-subsets. This subdivision is applied only to subsets which yield, in turn, unbalanced sub-subsets (i.e. if it is measured a minimal amount of conditional dependence inside the sub-subsets). In the illustration, it happens to subsets B and C, thus yielding sub-subsets B1, B2, B3, B4 and C1,C2, C3, C4.

Now, to simplify explanation, we arbitrarily define two measures associated with a given subset S, as illustrated in Figure 4, namely: the actual probability of randomly selecting a point from S, $P(S) = \frac{n(S)}{N}$, and the idealized probability $P_0(S) = \frac{1}{4}^L$, where $n(S)$ stands for the number of points in

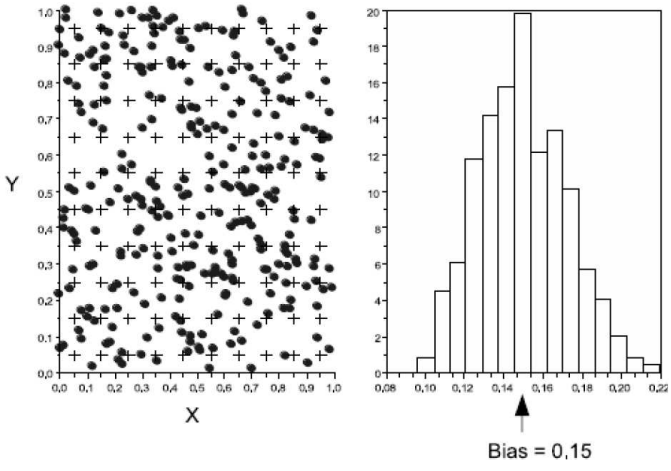


Fig. 3. Biased MI estimated from 300 uniform 2D data points and regular histogram — crosses represent bin centers whereas dots represent data points.

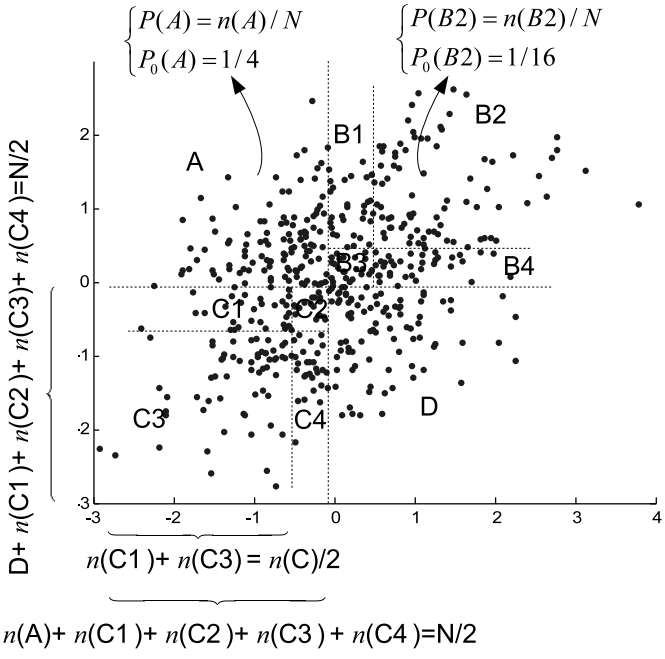


Fig. 4. $N = 500$ pairwise instances of two dependent gaussian random variables, X versus Y, with mutual information given by $I_{\text{gauss}}(X;Y) = 0.0472$ (i.e. $r = 0.3$).

S and L is the partitioning level where subset S is found (see Figure 4).

For instance, if we have

$$n(A) = n(D) = 93, n(B) = n(C) = 157$$

at the first partition level, and

$$n(B1) = 32, n(B2) = 46, n(B3) = 47, n(B4) = 32$$
$$n(C1) = 46, n(C2) = 33, n(C3) = 32, n(C4) = 46$$

at the second partition level, then

$$I_A = \frac{93}{500}\log\left(\frac{93/500}{1/4}\right), ..., I_{C4} = \frac{46}{500}\log\left(\frac{46/500}{1/16}\right)$$

and

$$\hat{I} = I_A + I_D + I_{B1} + ... + I_{C4} = 0.0432.$$

In comparison with the true mutual information value, which is 0.0472, this estimate can be considered pretty good.

Unfortunately, this method is very sensitive to the partitioning (grid-refining) process. Being aware of it, the authors of [39] propose the use of a stopping criterion more accurate than just testing how close to zero the estimated MI in new subpartitions are. This alternative criterion is based on the Chi-square test of independence, in which they use significance levels (which depend on the number of points inside the tested subpartition) instead of static thresholds.

Moreover, being the chain rule of probability the main theoretical support to this approach, it can be easily adapted to other MI estimation methods based on space discretization. For instance, one may use the conventional k-means algorithm, with small values of k, and to keep refining (re-quantizing) clusters upon a test for measuring independence inside it. The same idea may also be adapted to space quantization through kernel methods, such as the Parzen method. Roughly speaking, the idea presented in [39] paves the way for the use of irregular self-adapted grids (space partition) in any histogram-like method.

F. Discrete estimators

The previous sections introduced some important strategies, both from the standpoint of Rényi's and Shannon's definitions, to estimate information measures for continuous signals. Since this work has a broader scope in terms of data characteristics for developing ITL criteria, in the following, the subject is changed to the case of discrete signals.

1) Histogram with plug-in estimator: Similarly to the continuous case, the most direct and simple approach to estimate the PMF of discrete data is based on the histogram of the samples, with the benefit that it is not necessary to employ a discretization process. Consider a set of N independent and identically distributed (iid) observations; the mathematical formulation of the estimator is

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_x(x_i), \tag{47}$$

where $\mathbf{1}_x(\cdot)$ is the indicator function

$$\mathbf{1}_x(x_i) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{otherwise} \end{cases}, \tag{48}$$

and x_i is an observation. Thus, one can plug this estimator directly into the definition (1) of entropy and the definition (10) of mutual information for discrete random variables.

However, this procedure results in a biased estimator; hence there are other proposals to estimate entropy and mutual information of discrete data that attempts do circumvent this issue. As mentioned in Section IV-E, Miller [35] proposed a first-order correction to the estimator and Paninski [37] developed a thorough theoretical analysis of three common entropy estimators (including the previous definition) in the sense of the bias and variance, which also resulted in the presentation of a new estimator that possesses a rigorous bound on the maximum error over all possible distributions.

2) Lempel-Ziv complexity-based estimators: Most algorithms for entropy and mutual information estimation focus on random variables instead of stochastic processes. This is partially justified whenever a process is formed by iid variables through time. More precisely, let $\{X(n)\}$ and $\{Y(n)\}$ be two discrete-time random processes, where $n \in Z$ stands for the discrete-time counter. By defining $H(\{X(n)\})$ and $H(\{Y(n)\})$ as the entropy rate (or information rate, in bits per symbol) of $\{X(n)\}$ and $\{Y(n)\}$, respectively, and by defining $H(\{X(n)\},\{Y(n)\})$ as their joint entropy rate, we also define

$$I(\{X(n)\};\{Y(n)\}) = H(\{X(n)\} + H(\{Y(n)\}) - H(\{X(n)\},\{Y(n)\}) \tag{49}$$

as the MI between the processes $\{X(n)\}$ and $\{Y(n)\}$, i.e., a generalization of MI between random variables (see [5] for more information).

Clearly, if $\{X(n)\}$ is iid, then every random variable in it, $X(n), \forall n \in Z$ is associated to the very same amount of entropy $H_X = H(X(n))$, in bits. Consequently, the entropy rate of the iid process, $H(\{X(n)\})$, in bits per symbol, is numerically equal to H_X, and all usual methods for entropy and MI estimation for random variables are sufficient as tools.

By contrast, if a process is stationary [5] but not independent, then $H(\{X(n)\}) < H_X$, and its entropy rate per symbol is given by:

$$H(\{X(n)\}) = \lim_{N\to\infty}\frac{-1}{N}\sum_{i=1}^{K^N} p_i \log_2(p_i) \tag{50}$$

where K stands for the number of possible states / symbols X may assume, and p_i stands for the joint probability of $(X(n), X(n+1), ..., X(n+N-1))$ being equal to the $i-th$ sequence of states (out of K^N possible sequences).

The main drawback of the *so called* plug-in estimator suggested by (50) (with p_i replaced with relative frequencies, $\hat{p}_i$) is the huge amount of stationary data it demands, because the number of possible sequences, K^N, exponentially grows with the sequence length N.

Some alternatives to cope with it have appeared in literature. In the following, we briefly explain one of the most powerful of them, based on a complexity analysis of finite sequences (instead of the entropy of sources), proposed by Lempel and Ziv [40]. An important aspect of their approach is the lack of *a priori* information regarding the symbol source, which clearly contrasts with the measurement of (source)

Shannon entropy. In spite of these differences, it was shown that [5], under ergodicity conditions, Lempel-Ziv's complexity of increasingly long symbol sequences converges almost surely to the Shannon entropy of the source from which the symbols are drawn.

Lempel-Ziv's (LZ) approach, later simplified for practical reasons, became widely known as the compression algorithm behind many computer programs for file compression — the "zip-like" programs. We should probably credit its success to its universality, in other words, to its lack of demand for *a priori* information. Nevertheless, it should be also highlighted that zip-like programs are just the "tip of the iceberg", for compression is just a single offspring of the elegant theory presented in [40].

Definition 1: Complexity Measure. Let x_1^N represent a sequence (e.g. a single instance of a random process), and let a *Minimal Length Block* (MLB) be a subsequence x_i^j of x_1^N $(1 \leq i, j \leq N)$ such that it does not occur in x_1^{j-1}. Then, there is a unique decomposition of x_1^N into MLB, and the total number of these blocks, p, is the complexity measure of the sequence, denoted as:

$$C(x_1^N) = p$$

Illustration (from [40]): In this illustration, a sequence of $N = 16$ binary $(K = 2)$ symbols is parsed as:

$$x_1^{16} = 0001101001000101$$
$$\downarrow \text{parsing}$$
$$0 \cdot 001 \cdot 10 \cdot 100 \cdot 1000 \cdot 101$$

Note that the last block may produce an exception to the parsing rule, since it may be not unique (i.e. not an MLB), as in this illustration. As a result, we have that the complexity of this specific sequence is

$$C(x_1^{16}) = 6.$$

By comparing the complexity $C(x_1^N)$ to the maximum expected complexity of a hypothetical sequence of same length, which is given by $\frac{N}{\log_K N}$, we obtain the normalized complexity of the sequence, denoted by:

$$c(x_1^N) = \frac{C(x_1^N)}{N/log_K N}, \tag{51}$$

which almost surely [5] converges to the entropy rate given by (50), i.e. the entropy rate of the stochastic process of which x_1^N is likely to be an instance.

In order to illustrate the use of the Lempel-Ziv approach for entropy rate estimation, we reproduce here the experiment presented in [41], where a Markov Chain (whose true entropy rate can be analytically calculated) is used to generate random sequences of 0s and 1s.

The two-state Markov process (in discrete time) used in this experiment has a stationary transition matrix given by:

$$P = \begin{bmatrix} 1 - p_{10} & p_{01} \\ p_{10} & 1 - p_{01} \end{bmatrix}$$

where $p_{01} = P[X(n+1) = 0 | X(n) = 1]$ and $p_{10} = P[X(n+1) = 1 | X(n) = 0]$ are transition probabilities. It can be demonstrated [5] that these two parameters, p_{01} and p_{10} , completely determine the entropy rate of the finite-length process $\{X(n)\}$, $n = 1, 2, ..., N$:

$$H = \frac{-p_{01}(p_{10}\log_2 p_{10} + (1-p_{10})\log_2(1-p_{10}))}{p_{01}+p_{10}} \\ \frac{-p_{10}(p_{01}\log_2 p_{01} + (1-p_{01})\log_2(1-p_{01}))}{p_{01}+p_{10}}. \tag{52}$$

For instance, if $p_{01} = 0.8$ and $p_{10} = 0.1$, we obtain that the resulting binary source asymptotically "produces" 0.497 bits of information per emitted binary symbol.

For all very specific cases where $p_{01}+p_{10} = 1$, $H(\{X(n)\})$ does not depend on N, which makes even the plug-in method less inaccurate. However, these are rather rare cases of Markov processes and, in general, the plug-in method is to be avoided, unless a huge amount of data is available. This is evident if we keep in mind that this method relies upon relative frequency of occurrences of symbol sequences. Clearly, the number of possible sequences exponentially grows with its length, and so does the amount of necessary data to avoid statistical undersampling problems.

On the other hand, the astonishingly 'simple to obtain' measure presented in (51) provides us with accurate estimates of $H(\{X(n)\})$ (although it is aimed at measuring complexity of specific sequences of symbols). As an illustration, in Figure 5, we can observe how fast the normalized complexity measure converges to the true asymptotic information rate of corresponding processes, by using symbolic sequences of length up to 4000 symbols.

Finally, if two random processes share some amount of information, it is also possible to measure it through (49) — i.e. mutual information estimation through the Lempel-Ziv approach —, where $H(\{X(n)\}, \{Y(n)\})$ can be easily

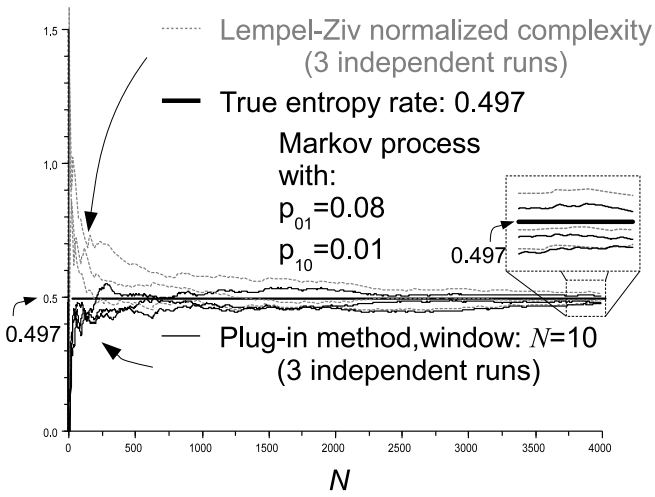


Fig. 5. Three independent runs of the normalized complexity measure, with $p_{01} = 0.8$ and $p_{10} = 0.1$. The entropy rate per symbol estimated through the plug-in method is presented as well, for an arbitrarily chosen observation window of $N = 10$ symbols. In this case, to avoid singularity, we assume that $0 \log(0) = 0$.

computed as the entropy rate of a new "concatenated" process

$$Z(n) = \begin{bmatrix} \{X(n)\} \\ \{Y(n)\} \end{bmatrix}.$$

V. Conclusion

In this work, the first of a two-part tutorial, we presented elements of the three theoretical pillars of ITL: information theory, Rényi's formulations and statistical estimators. The discussion starts from a historical overview of the development of information theory, in which essential concepts like entropy, joint entropy, conditional entropy and mutual information are defined. In the sequence, these concepts are revisited in the context of Rényi's definitions, with emphasis on the quadratic case. Finally, the associated estimation problems are addressed in detail, as well as important paradigms like Parzen windowing and order statistics. In the second part of the tutorial, this theoretical framework will be used in the exposition and analysis of ITL methods and of their application to a number of representative information retrieval tasks.

Acknowledgment

The authors thank FAPESP (Grant 2013/14185-2), CAPES and CNPq for the financial support.

References

[1] N. Wiener, *Nonlinear Problems in Nandom Theory*. MIT, 1958, vol. 1.
[2] A. Kolmogorov, "Interpolation and extrapolation of stationary random processes," *Izv. Akad. Nauk SSSR Ser. Mat.*, vol. 5, pp. 3–14, 1941.
[3] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*. Addison Wesley, 1994.
[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2001.
[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.
[6] A. Rényi, *Probability Theory*. North-Holland, 1970.
[7] J. Príncipe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Verlag, 2010.
[8] T. Back, D. B. Fogel, and Z. Michalewicz, Eds., *Evolutionary Computation 1: Basic Algorithms and Operators*. Bristol, UK: Taylor & Francis, 2000.
[9] L. N. De Castro and J. Timmis, *Artificial Immune Systems: a New Computational Intelligence Approach*. Springer Verlag, 2002.
[10] J. Gleick, *The Information: a History, a Theory, a Flood*. Vintage, 2012.
[11] H. Nyquist, "Certain factors affecting telegraph speed," *Journal of the A.I.E.E.*, vol. 43, no. 12, pp. 1197–1198, 1924. doi: 10.1109/JAIEE.1924.6534511
[12] ——, "Certain topics in telegraph transmission theory," *American Institute of Electrical Engineers, Transactions of the*, vol. 47, no. 2, pp. 617–644, 1928. doi: 10.1109/T-AIEE.1928.5055024
[13] R. V. L. Hartley, "Transmission of Information," *Bell System Technical Journal*, vol. 7, no. 3, pp. 535–563, 1928. doi: 10.1002/j.1538-7305.1928.tb01236.x
[14] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
[15] R. P. Feynman, *Feynman Lectures on Computation*, R. W. A. T. Hey, Ed. Westview Press, 2000.
[16] A. Rényi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.
[17] J. Príncipe, D. Xu, and J. Fischer, *Unsupervised Adaptive Filtering*. Wiley, 2000, vol. 1, ch. Information Theoretic Learning, pp. 265–319.
[18] B. Behmardi, R. Raich, and A. O. Hero III, "Entropy estimation using the principle of maximum entropy," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2008–2011.
[19] A. J. Izenman, "Review papers: Recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 205–224, 1991. doi: 10.1080/01621459.1991.10475021
[20] J. Beirlant, E. J. Dudewicz, L. Gyorfi, and E. C. van der Meulen, "Nonparametric entropy estimation: an overview," *International Journal of Mathematical and Statistics Sciences*, vol. 6, pp. 17–39, 2001.
[21] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.
[22] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, 2006.
[23] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data using the em algorithm," *Journal of the Royal Statistics Society*, vol. 39, pp. 1–38, 1977.
[24] A. Webb, *Statistical Pattern Recognition*. Wiley, 2002.
[25] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
[26] J. Larsen, "Design of neural network filters," Ph.D. dissertation, Institute of Electronics, Technical University of Denmark, 1996.
[27] J. Montalvão and J. Canuto, "Gaussian mixture regularization through parzen method with prunning – an acceptance region based approach," in *Annals of the Congresso Brasileiro de Automática, SBA 2008*, 2008.
[28] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010. doi: 10.1214/10-AOS799
[29] D.-T. Pham, "Blind separation of instantaneous mixture of sources based on order statistics," *IEEE Transactions on Signal Processing*, vol. 48, pp. 363–375, 2000. doi: 10.1109/78.823964
[30] J. Even, "Contributions à la séparation de sources à l'aide de statistiques d'ordre." Ph.D. dissertation, Universitè Joseph Fourier, 2003.
[31] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 29, pp. 1–16, 2004. doi: 10.1103/PhysRevE.69.066138
[32] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1849—-1857.
[33] A. O. Hero, B. Ma, O. J. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 85–95, 2002. doi: 10.1109/MSP.2002.1028355
[34] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, no. Suppl. 2, pp. S231–S240, 2002. doi: 10.1093/bioinformatics/18.suppl_2.S231
[35] G. A. Miller, "Note on the bias of information estimates," *Information Theory in Psychology: Problems and Methods*, vol. 2, pp. 95–100, 1955.
[36] B. Harris, "The statistical estimation of entropy in the non-parametric case," MRC, Tech. Rep. 1605, 1975.
[37] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003. doi: 10.1162/089976603321780272
[38] T. Schürmann, "Bias analysis in entropy estimation," *Journal of Physics A: Mathematical and General*, vol. 37, pp. L295–L301, 2004. doi: 10.1088/0305-4470/37/27/L02
[39] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999. doi: 10.1109/18.761290
[40] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 75–81, 1976. doi: 10.1109/TIT.1976.1055501
[41] J. M. Amigo, J. Szczepanski, E. Wajnryb, and M. V. Sanchez-Vives, "Estimating the entropy rate of spike trains via lempel-ziv complexity," *Neural Computation*, vol. 16, pp. 717–736, 2004. doi: 10.1162/089976604322860677

Daniel G. Silva was born in Botucatu, Brazil, in 1983. He received the B.S. degree in computer engineering and the M.S. and Ph.D. degrees in electrical engineering, all from the University of Campinas (UNICAMP), São Paulo, Brazil, in 2006, 2009, and 2013, respectively. Currently, he is a Professor at the Department of Electrical Engineering (ENE) of the University of Brasília (UnB). His main research interests are information theoretic learning, adaptive signal processing and computational intelligence.

Denis G. Fantinato was born in Americana, Brazil, in 1985. He received the B.S. and M.Sc. degrees in Electrical Engineering from the University of Campinas (UNICAMP) in 2011 and 2013, respectively. Currently, he is a Ph.D. student at the same institution. His main research interests are blind signal processing, adaptive filtering and information theoretic learning.

Jânio C. Canuto was born in Maceió, Brazil, in 1984. He received the B.S. degree in Electrical Engineering (2007) from the Federal University of Sergipe (UFS), M.S. degree in Electrical Engineering (2010) from the University of Campinas (UNICAMP) and Ph.D. degree in Computer Science (2014) from Télécom SudParis. He joined the Department of Computer Science of the Federal University of Sergipe (UFS) in 2016. His main research interests are pattern recognition and machine learning.

Leonardo T. Duarte received the B.S. and the M.Sc. degrees in electrical engineering from UNICAMP (Brazil) in 2004 and 2006, respectively, and the Ph.D. degree from the Grenoble Institute of Technology (Grenoble INP), France, in 2009. He is currently an assistant professor at the School of Applied Sciences at UNICAMP. His research interests are mainly associated with the theory of unsupervised signal processing and include signal separation, independent component analysis, Bayesian methods, and applications in chemical sensors and seismic signal processing. He is also working on unsupervised schemes for adjusting multiple-criteria decision analysis (MCDA) techniques. He is a Senior Member of the IEEE.

Aline O. Neves received the B.S. and M.S. degree in Electrical Engineering from the University of Campinas (UNICAMP), Brazil, in 1999 and 2001 respectively. She received her Ph.D. degree in 2005, also in Electrical Engineering, from the University René Descartes (Paris V), Paris, France. Recently, she is an associate professor at the Engineering, Modeling and Applied Social Science Center of the Federal University of ABC, Santo André, Brasil. Her research interests consist of equalization, channel estimation, source separation and information theoretic learning.

Ricardo Suyama was born in São Paulo, Brazil, in 1978. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the State University of Campinas (Unicamp), Campinas, Brazil, in 2001, 2003 and 2007, respectively. Currently he is an Assistant Professor at the Universidade Federal do ABC (UFABC), Sao Paulo, Brazil. His research interests include blind source separation, adaptive equalization, adaptive nonlinear filtering, and evolutionary algorithms.

Jugurta Montalvão was born in Aracaju, Brazil, in 1968. He received the title of Electrical Engineer (1992) from the University of Campina Grande (UFPB II), Master in Electrical Engineering (1995) from the University of Campinas (UNICAMP) and Doctor in "Automatique et traitement du signal" (2000) from the University Paris-Sud XI. He joined the Department of Electrical Engineering of the Federal University of Sergipe (UFS) in 2005. His main research interests are: pattern recognition and signal processing.

Romis Attux was born in Goiânia, Brazil, in 1978. He received the titles of Electrical Engineer (1999), Master in Electrical Engineering (2001) and Doctor in Electrical Engineering (2005) from the University of Campinas (UNICAMP), Brazil. Currently, he is an associate professor at the same institution. His main research interests are adaptive filtering, computational intelligence, dynamical systems / chaos and brain-computer interfaces.