# Computation and Estimation of Generalized Entropy Rates for Denumerable Markov Chains

Gabriela Ciuperca, Valérie Girardin, Loïck Lhote

# Computation and Estimation of Generalized Entropy Rates for Denumerable Markov Chains

Gabriela Ciuperca, Valerie Girardin, Loïck Lhote

***Abstract*—We study entropy rates of random sequences for general entropy functionals including the classical Shannon and Rényi entropies and the more recent Tsallis and Sharma-Mittal ones.**

**In the first part, we obtain an explicit formula for the entropy rate for a large class of entropy functionals, as soon as the process satisfies a regularity property known in dynamical systems theory as the quasi-power property. Independent and identically distributed sequence of random variables naturally satisfy this property. Markov chains are proven to satisfy it too, under simple explicit conditions on their transition probabilities. All the entropy rates under study are thus shown to be either infinite or zero except at a threshold where they are equal to Shannon or Rényi entropy rates up to a multiplicative constant.**

**In the second part, we focus on the estimation of the marginal generalized entropy and entropy rate for parametric Markov chains. Estimators with good asymptotic properties are built through a plug-in procedure using a maximum likelihood estimation of the parameter.**

***Index Terms*—entropy rate, entropy functional, parametric Markov chain, plug-in estimation, Rényi entropy, Tsallis entropy.**

## I. Introduction

In [21], Shannon adapted to the field of information theory the concept of entropy introduced by Boltzmann and Gibbs in the XIX-th century. Entropy measures the randomness or uncertainty of a random phenomenon. It now applies to various areas such as information theory, finance, statistics, cryptography, physics, artificial intelligence, etc.; see [8] for details. Rényi proposed in [19] a one parameter family of entropies extending Shannon entropy to new applications. Since then, many different generalized entropies have been defined to adapt to many different fields. Among them, Tsallis [23] or Sharma-Mittal [24] entropies are instances of what we will call $(h, \phi)$-entropies, thus following [20] – where only parametric probability density functions are considered. Precisely, we set

$$\mathbb{S}_{h(y),\phi(x)}(\nu) = h\left[\sum_{i\in E}\phi(\nu(i))\right] \tag{1}$$

for any measure $\nu$ on a countable space $E$ such that the quantity is finite.

G. Ciuperca is with the Laboratoire de Probabilité, Combinatoire et Statistique, Université LYON I, Domaine de Gerland, Bât. Recherche B, 50 Av. Tony-Garnier, 69366 Lyon cedex 07, France, gabriela.ciuperca@pop.univ-lyon1.fr

V. Girardin is with the Laboratoire de Mathématiques N. Oresme, UMR6139, Campus II, Université de Caen, BP5186, 14032 Caen, France, girardin@math.unicaen.fr

L. Lhote is with ENSICAEN, GREYC, Campus II, Université de Caen, BP5186, 14032 Caen, France, loick.lhote@info.unicaen.fr

In this paper, we address the problems of computing and then estimating the $(h, \phi)$-entropy rates of random sequences – especially Markov chains, taking values in countable spaces, either finite or denumerable. This entropy rate is defined as the limit of the time average of the entropy of the considered random sequence, that is as the entropy per unit time of the sequence.

For an independent identically distributed (i.i.d.) sequence, Shannon and Rényi entropy rates are well-known to be the entropy of the common distribution. The Shannon entropy rate of an ergodic and homogeneous Markov chain with a countable state space is an explicit function of its transition distributions and stationary distribution; it is also known to be related to the dominant eigenvalue of some perturbation of the transition matrix, a result proven for Rényi entropy in [18]. [12] deals with the denumerable case but the proofs contain flaws (see end of Section IV-B). Up to our knowledge, no other result exists in the literature concerning explicit determination of $(h, \phi)$-entropy rates. The aim of the first part of the present paper is to fill this gap, with a particular interest to Markov chains.

The entropy of the stationary distribution of a Markov chain is the (asymptotic) entropy of the chain at equilibrium; if this distribution is taken as initial distribution of the chain, its entropy is also the marginal entropy of the chain. In both cases, the entropy rate is more representative of the whole trajectory of the sequence. Having the marginal entropy and entropy rate of Markov chains under an explicit form allows one to use them efficiently in all applications involving Markov modeling. When only observations of the chain are available, the need for estimation obviously appears. We consider the case of countable parametric chains, for which transition probabilities are functions of a finite set of parameters. Since the entropy is an explicit function of the transition probabilities, and hence of the parameters, plug-in estimators of the marginal entropy and entropy rate are obtained by replacing the parameters by their maximum likelihood estimators (MLE).

Up to our knowledge, no result exists in the literature on the estimation neither of the generalized entropy of the stationary distribution nor of the generalized entropy rate of a countable Markov chain. Concerning Shannon entropy, see [5] for results on the estimation of the marginal entropy through a Monte Carlo method, and [7] for the estimation of the marginal entropy and entropy rate of finite chains. The special case of two-state Markov chains is studied in [11].

The paper is organized as follows. Basics on generalized entropies and entropy rates are given in Section II. In Section III, the regularity property called quasi-power property is

| $h(y)$ | $\phi(x)$ | $(h,\phi)-$ entropies |
|---|---|---|
| $y$ | $-x\log x$ | Shannon (1948) |
| $(1-s)^{-1}\log y$ | $x^s$ | Renyi (1961) |
| $[t(t-r)]^{-1}\log y$ | $x^{r/t}$ | Varma (1966) |
| $y$ | $(1-2^{1-r})^{-1}(x-x^r)$ | Havrda and Charvat (1967) |
| $(t-1)^{-1}(y^t-1)$ | $x^{1/t}$ | Arimoto (1971) |
| $(r-1)^{-1}[y^{(r-1)/(s-1)}-1]$ | $x^s$ | Sharma and Mittal 1 (1975) |
| $(r-1)^{-1}[\exp(r-1)y-1]$ | $-x\log x$ | Sharma and Mittal 2 (1975) |
| $y$ | $-x^r\log x$ | Taneja (1975) |
| $y$ | $(t-r)^{-1}(x^r-x^t)$ | Sharma and Taneja (1975) |
| $(r-1)^{-1}(1-y)$ | $x^r$ | Tsallis (1988) |

TABLE I
SOME $(h,\phi)$-ENTROPIES.

stated and shown to induce convergence of the time average entropy to an explicit limit. In Section IV, mild conditions are shown to be sufficient for countable Markov chains to satisfy the quasi-power property. Estimation of the generalized marginal entropy and entropy rate for parametric countable Markov chains is studied in Section V.

## II. GENERALIZED ENTROPIES AND ENTROPY RATES

### A. Generalized $(h,\phi)$-entropies

All throughout the paper, $E$ will be a countable set, either finite or denumerable. Both functions $h:\mathbb{R}_+\to\mathbb{R}$ and $\phi:[0,1]\to\mathbb{R}_+$ are twice continuously differentiable functions, with $h$ monotonous and either $\phi$ concave or convex. Both $\phi$ and $h\circ\phi$ will be supposed to be positive for simplification, but all the results below can be adapted to negative functions.

We define the $(h,\phi)$-entropy $\mathbb{S}_{h(x),\phi(y)}(\nu)$ of any measure $\nu$ on $E$ as in (1) if $\sum_{i\in E}\phi(\nu(i))$ is finite, and as $+\infty$ either. For the sake of simplicity, we will suppose that $\mathbb{S}_{h(x),\phi(y)}(\nu)$ is nonnegative.

The conditions on both $\phi$ and $h$ are the usual conditions for entropy that are for instance satisfied by all the entropies of Table I. The function $h$ may not be positive (see for instance Rényi or Varma entropies) but the parameters in $h$ and $\phi$ behave such that $h\circ\phi$ is indeed positive. Note also that $h(x)$ is finite for all $x\in\mathbb{R}_+$, but that $h$ may not be bounded on $\mathbb{R}_+$.

For a a random variable $X$ with distribution $\nu$, we set $\mathbb{S}_{h(x),\phi(y)}(X)=\mathbb{S}_{h(x),\phi(y)}(\nu)$. For a stationary random sequence $(X_n)_{n\in\mathbb{N}}$ with common distribution $\nu$, we will call marginal entropy of the sequence the quantity $\mathbb{S}_{h(x),\phi(y)}(\nu)=\mathbb{S}_{h(x),\phi(y)}(X_n)$.

Definition (1) includes all classical entropies. First, we get Shannon entropy for $\phi(x)=-x\log x$ and $h$ the identity function, so that

$$\mathbb{S}(\nu)=\mathbb{S}_{y,-x\log x}(\nu)=-\sum_{i\in E}\nu(i)\log\nu(i).$$

Shannon entropy is concave and is additive (and fits well to extensive systems), meaning that the Shannon entropy of the product of marginal measures is the sum of the entropies of the marginal measures.

Rényi entropy is obtained for $h_s(y)=(1-s)^{-1}\log y$ and $\phi_s(x)=x^s$ with $s>0$, that is

$$\mathbb{R}_s(\nu)=\mathbb{S}_{h_s(y),\phi_s(x)}(\nu)=\frac{1}{1-s}\log\sum_{i\in E}\nu(i)^s;$$

Shannon entropy corresponds to $s\to 1$. Rényi entropy is additive, but is concave only for $s\le 1$. Note that [2] proves that additive entropies are necessarily non linear transforms of Rényi entropies. We refer to [13] for detailed applications of Rényi entropies.

Standard thermodynamical extensivity is lost in strong mixing, long range interacting or non-Markovian physical systems. This led Tsallis to postulate in [24] a nonadditive generalization of Shannon entropy which now bears his name, thus allowing for superextensivity (when $r<1$) or subextensivity (when $r>1$). Note that Tsallis entropy equals Havrda-Charvat entropy up to a multiplicative term depending only on the parameter. Tsallis entropy involves the functions $\phi_r(x)=x^r$ for some positive $r\neq 1$, and $h_r(y)=(r-1)^{-1}(1-y)$, so that

$$\mathbb{T}_r(X)=\mathbb{S}_{h_r(y),\phi_r(x)}(\nu)=\frac{1}{r-1}\left[1-\sum_{i\in E}\nu(i)^r\right].$$

Tsallis entropy is concave and appears as the unique solution of a generalized Khinchin's set of conditions. Shannon entropy corresponds to the value $r=1$. Rényi entropy is a monotonically decreasing function of Tsallis entropy, precisely $\mathbb{R}_s(X)=(1-s)^{-1}\log[1-(s-1)\mathbb{T}_s(X)]$, but concavity is not preserved through monotonicity. See Tsallis [24] for details in statistical mechanics, and for hints for determining $r$ from fitting physical constraints. See [27] and the references therein for other applications in statistical mechanics, in thermodynamics, in the study of DNA binding sites, etc.

Both Rényi and Tsallis entropies appear as particular cases of Sharma-Mittal entropy introduced in [23] with $h_{s,r}(y)=(r-1)^{-1}[1-y^{(1-r)/(1-s)}]$ and $\phi_s(x)=x^s$, that is

$$\mathbb{S}_{s,r}(\nu)=\mathbb{S}_{h_{s,r}(y),\phi_s(x)}(\nu)=\frac{1}{r-1}\left(1-\left[\sum_{i\in E}\nu(i)^s\right]^{\frac{1-r}{1-s}}\right).$$

Rényi entropy corresponds to $r\to 1$ and Tsallis entropy to $s\to r$. The case $s\to 1$ is sometimes called Gaussian entropy;

precisely,

$$\lim_{s\to 1}\mathbb{S}_{s,r}(X)=\frac{1}{1-r}\left(1-\exp\left[(r-1)\mathbb{S}(X)\right]\right).$$

In general, Sharma-Mittal entropy is neither extensive nor concave.

A list of $(h,\phi)$-entropies is given in Table I; we refer to [15] and to the references therein for details.

### B. Entropy rates

For a discrete-time process $\mathbf{X}=(X_n)_{n\in\mathbb{N}}$, under suitable conditions (see [10]), the entropy of $(X_0,\dots,X_{n-1})$ divided by $n$ converges to a limit called the entropy rate of the process, say $\mathbb{H}_{h(y),\phi(x)}(\mathbf{X})$. Precisely,

$$\mathbb{H}_{h(y),\phi(x)}(\mathbf{X})=\lim_{n\to\infty}\frac{1}{n}\mathbb{S}_{h(y),\phi(x)}(X_0,\dots,X_{n-1}).$$

See [9] for an interesting study of Tsallis and other non-extensive entropies and entropy rates.

For all additive entropy functionals, the entropy rate of an i.i.d. sequence $\mathbf{X}=(X_n)_{n\in\mathbb{N}}$ with common distribution $\nu$ is the entropy of $\nu$, so that in particular $\mathbb{H}_s(\mathbf{X})=\mathbb{S}_s(\nu)$ for Rényi entropy and $\mathbb{H}(\mathbf{X})=\mathbb{S}(\nu)$ for Shannon entropy. For an ergodic homogeneous Markov chain $\mathbf{X}$ with countable state space, the Shannon entropy rate depends on the transition probabilities $P=(p(i,j))_{i,j\in E}$ and stationary distribution $\pi=(\pi(i))_{i\in E}$ (such that $\pi P=\pi$) through the well-known expression stated in [21],

$$\mathbb{H}(\mathbf{X})=-\sum_{i,j\in E}\pi(i)p(i,j)\log p(i,j). \tag{2}$$

Rached *et al* proved in [18] that the Rényi entropy rate of a finite Markov chain is

$$\mathbb{H}_s(\mathbf{X})=\frac{1}{1-s}\log\lambda(s), \tag{3}$$

using that the perturbated matrix $P_s=(p(i,j)^s)_{i,j\in E}$ has a unique dominant eigenvalue $\lambda(s)$ for any $s>0$. We will show in Section IV-B that (3) holds true for a denumerable Markov chain too under mild regularity conditions. Note that by letting $s$ go to 1 in (3), since $\lambda(1)=1$, Shannon entropy rate is also related to the dominant eigenvalue through $\mathbb{H}(\mathbf{X})=-\lambda'(1)$.

A random sequence can also be described in terms of symbolic dynamical systems theory. We refer to [25] for the definition in terms of dynamical systems of i.i.d. sequences (also called Bernoulli processes) and finite Markov chains, and to [6] for Markovian dynamical systems with countable state spaces. Both deal, among other topics, with the Shannon entropy rate of processes by means of functional operators also called transfer operators. These operators play the same role as perturbations of transition matrices do in [18]; they also have a unique dominant eigenvalue $\lambda(s)$, and the Shanon entropy rate is thus proven to be $-\lambda'(1)$. In the next section, we will use similar operators techniques for determining $(h,\phi)$-entropy rates.

## III. Quasi-power property and $(h,\phi)$-entropy rate

We will first introduce the quasi-power property and then prove that the $(h,\phi)$-entropy rate of a random sequence satisfying that property can be computed explicitly for a large class of $(h,\phi)$ functions. The proof will involve the series

$$\Lambda_n(s)=\sum_{i_0^{n-1}\in E^n}\nu_n(i_0^{n-1})^s \tag{4}$$

for $s>0$, and its formal derivatives for $k\geq 1$,

$$\Lambda_n^{(k)}(s)=\sum_{i_0^{n-1}\in E^n}[\log\nu_n(i_0^{n-1})]^k\nu_n(i_0^{n-1})^s, \tag{5}$$

where $\nu_n$ denotes the marginal distribution of order $n$ of the random sequence $\mathbf{X}=(X_n)_{n\in\mathbb{N}}$; in other words, $\nu_n(i_0^{n-1})=\mathbb{P}(X_0=i_0,\dots,X_{n-1}=i_{n-1})$. In dynamical systems theory, $\Lambda_n(s)$ is called the Dirichlet series of fundamental measures of depth $n$. It is a central tool for studying general sources, in pattern matching or in the analysis of data structures; see for instance [6]. This series is also introduced in [17] (see $V(n,s)$ page 36).

The simplest case of a random sequence is an i.i.d. sequence with a non-degenerated distribution $\nu$ over a finite set $E$. Since its marginal distribution of order $n$ is $\nu_n(i_0^{n-1})=\nu(i_0)\nu(i_1)\dots\nu(i_{n-1})$, the Dirichlet series $\Lambda_n(s)$ defined in (4) can be simply written as the $n$-th power of an analytic function, precisely

$$\Lambda_n(s)=\left[\sum_{i\in E}\nu(i)^s\right]^n.$$

The quasi-power property next stated says that $\Lambda_n(s)$ behaves for more general random sequences satisfying it like the $n$-th power of some analytic function up to some error term.

**Property 1** *Let $\mathbf{X}=(X_n)_{n\in\mathbb{N}}$ be a random sequence taking values in a countable set $E$ and let $\nu_n$ denote the marginal distribution of $(X_0,\dots,X_{n-1})$. Then $\mathbf{X}$ is said to satisfy the quasi-power property with parameters $[\sigma_0,\lambda,c,\rho]$ if both following conditions are fulfilled:*

*1. $\sup_{i_0^{n-1}\in E^n}\nu_n(i_0^{n-1})$ converges to zero when $n$ tends to infinity.*

*2. there exists a real number $\sigma_0<1$, such that for all real number $s>\sigma_0$ and all integer $n\geq 0$, the series $\Lambda_n(s)$ defined in (4) is convergent and satisfies*

$$\Lambda_n(s)=c(s)\cdot\lambda(s)^{n-1}+R_n(s), \tag{6}$$

*with $|R_n(s)|=O\left(\rho(s)^{n-1}\lambda(s)^{n-1}\right)$, where $c$ and $\lambda$ are strictly positive analytic functions for $s>\sigma_0$, and $\lambda$ is strictly decreasing with $\lambda(1)=c(1)=1$, and $R_n$ is also analytic, and finally $\rho(s)<1$.*

Obviously, any i.i.d. sequence taking values in a finite set $E$ satisfies the quasi-power property for $\sigma_0=0$ and functions $\lambda$, $c$ and $\rho$ defined by

$$\lambda(s)=\sum_{i\in E}\nu(i)^s,\quad c(s)=\lambda(s)\quad\text{and}\quad\rho(s)=0.$$

The result extends to the case of a denumerable set $E$ as soon as some $\sigma_0<1$ exists such that, for all $s>\sigma_0$, the series

$\lambda(s)$ converges. Note that Shannon entropy rate is then equal to $-\lambda'(1)$ and Rényi entropy rate to $(1-s)^{-1}\log\lambda(s)$.

We can now state the main result for determinating the generalized entropy rates of random sequences satisfying the quasi-power property.

**Theorem 1** *Let $\mathbf{X}=(X_n)_{n\in\mathbb{N}}$ be a random sequence satisfying the quasi-power property with parameters $[\sigma_0,\lambda,c,\rho]$. Suppose that*

$$\phi(x)\underset{x\to 0}{\sim} c_1\cdot x^s\cdot(-\log x)^k \tag{7}$$

*with $s>\sigma_0$, $c_1\in\mathbb{R}_+^*$ and $k\in\mathbb{N}^*$. Then Table II gives the entropy rate $\mathbb{H}_{h,\phi}(\mathbf{X})$ according to the behavior of $h$ around $0$ for $s>1$ and around $+\infty$ for $s\leq 1$.*

Note that condition (7) is satisfied for all the usual entropies listed in Table I.

*Proof:* Point 1 of the quasi-power property and (7) together induce that for any $\varepsilon>0$, there exists $n_0\in\mathbb{N}$ such that for all $n\geq n_0$ and $i_0^{n-1}\in E^n$,

$$(1-\varepsilon)(-1)^k c_1\nu_n(i_0^{n-1})^s\log^k\nu_n(i_0^{n-1})\leq\phi(\nu_n(i_0^{n-1}))$$

and

$$\phi(\nu_n(i_0^{n-1}))\leq(1+\varepsilon)(-1)^k c_1\nu_n(i_0^{n-1})^s\log^k\nu_n(i_0^{n-1}).$$

Therefore,

$$(1-\varepsilon)(-1)^k c_1\Lambda_n^{(k)}(s)\leq\Sigma_n\leq(1+\varepsilon)(-1)^k c_1\Lambda_n^{(k)}(s), \tag{8}$$

where we have set $\Sigma_n=\sum_{i_0^{n-1}\in E^n}\phi(\nu_n(i_0^{n-1}))$ for simplification. Due to the analyticity of all the functions involved in (6), for $n$ large enough,

$$\Lambda_n^{(k)}(s)=c(s)\cdot\lambda'(s)^k\cdot n^k\cdot\lambda(s)^{n-k-1}\cdot\left[1+O\left(\frac{1}{n}\right)\right].$$

Putting this into (8) yields

$$\Sigma_n\sim c_1\cdot c(s)\cdot(-\lambda'(s))^k\cdot n^k\cdot\lambda(s)^{n-k-1}, \tag{9}$$

with $-\lambda'(s)=|\lambda'(s)|$ since $\lambda$ is a decreasing function. Let us now study the three different cases concerning $s$.

First, suppose that $s=1$. Since $\lambda(1)=c(1)=1$, (9) simplifies into

$$\Sigma_n\sim c_1\cdot|\lambda'(1)|^k\cdot n^k.$$

Since $\phi$ is a positive function, $\Sigma_n$ converges polynomially to $+\infty$. Depending on the conditions on $h$, this leads to the next equivalences: $h(\Sigma_n)\sim c_2\cdot c_1^{1/k}\cdot|\lambda'(1)|\cdot n$ in Case (I), $h(\Sigma_n)\sim o(n)$ in Case (II), and $h(\Sigma_n)\sim s_n\cdot n$ with $s_n\to+\infty$ in Case (III). By definition, the $(h,\phi)$-entropy rate is the limit of $h(\Sigma_n)/n$ when $n$ tends to infinity, so the results given in Table II for $s=1$ follow immediately.

For either $s<1$ or $s>1$, the function $\lambda$ is strictly decreasing with $\lambda(1)=1$. Hence $\lambda(s)<1$ for $s>1$ and $\lambda(s)>1$ for $s<1$, from which we deduce that $\Sigma_n$ tends exponentially to $+\infty$ for $s<1$ and to $0^+$ for $s>1$. Depending on conditions on $h$, this leads to the next equivalences: $h(\Sigma_n)\sim c_2\cdot\log\lambda(s)\cdot n$ in cases (IV) and (VII), $h(\Sigma_n)\sim o(n)$ in cases (V) and (VIII), and $h(\Sigma_n)\sim s_n\cdot n$ with $s_n\to+\infty$ in cases (VI) and (IX). Then, we get exactly in the same way as for $s=1$ the results listed in Table II for $s<1$ and $s>1$. □

Table III shows applications of Theorem 1 to various entropy rates for random sequences satisfying the quasi-power property with parameters $[\lambda,c,\rho,\sigma_0]$. Remark that almost all the entropy rates are finite and non-zero only at a threshold where they are equal to Shannon or Rényi entropy rates up to a multiplicative factor. Elsewhere, they are either null or infinite, which limits their practical interest in applications.

## IV. Markov chains and the quasi-power property

Let us begin this section by connecting Markov chains to dynamical sources. A dynamical source is defined by five objects: a countable alphabet $\overline{E}$, a topological partition $(I_i)_{i\in\overline{E}}$ of the interval $I=[0,1]$, a coding function $\sigma:I\to\overline{E}$ such that $\sigma(I_i)=i$ for all symbols $i$ of $\overline{E}$, a density function $f_0$ on $I$ and finally a shift function $T$ which is twice continuously differentiable and strictly monotonous on each interval of the partition. The random sequence $\mathbf{X}=(X_n)$ associated to the dynamical source corresponds to the trace of the iterates $T^n(x)$ for some $x$ chosen according to the distribution $f_0$. Precisely,

$$X_n=\sigma(T^n(X_0)),$$

where $X_0$ is a random variable with density function $f_0$ on $I$.

The dynamical source $(T,I,\overline{E},(I_i)_{i\in\overline{E}},f_0)$ is Bernoulli if $T$ is surjective (i.e., $T(I_i)=I$) and affine on each $I_i$ and if $f_0$ is constant. Then, it is easy to check that the associated random sequence $\mathbf{X}$ is i.i.d.. The dynamical source is said to be Markovian if the image of each interval is the union of images of intervals of the partition. If, furthermore, $T$ is piecewise affine and $f_0$ is constant on each interval of the partition, the associated random sequence $\mathbf{X}$ is a Markov chain.

Conversely, any Markov chain $\mathbf{X}$ over a countable state space $E$ can be represented by a (non unique) Markovian dynamical source. Let $P=(p(i,j))_{i,j\in E}$ denote the transition matrix of $\mathbf{X}$ (that is $p(i,j)=\mathbb{P}(X_n=j|X_{n-1}=i)$) and $\mu=(\mu(i))_{i\in E}$ its initial distribution (that is $\mu(i)=\mathbb{P}(X_0=i)$). For instance, we can consider a topological partition $(I_i)_{i\in E}$ of $I=[0,1]$ and then a second one $(I_{j|i})_{i,j\in E}$ such that for all $i,j\in E$,

$$|I_{j|i}|=p(i,j)\cdot|I_i|\quad\text{and}\quad\bigcup_{j\in E}\overline{I_{j|i}}=\overline{I_i},$$

where $\overline{I}$ denotes the closure of the interval $I$. A dynamical source simulating the Markov chain $\mathbf{X}$ is then given by the following five elements: the alphabet, say $\overline{E}=\{j|i,\ (i,j)\in E^2\}$, the topological partition $(I_{j|i})_{i,j\in E}$, the coding function $\sigma$ defined by $\sigma(I_{j|i})=j|i$, the piecewise constant density function $f$ defined by $f(I_i)=\mu(i)/|I_i|$, and finally the piecewise linear function $T$ defined by $T(I_{j|i})=I_j$. Note that even if the state spaces of the Markov chain and of the associated dynamical source seem different, a clear bijection exists between both processes.

Finally, note that [6] give sufficient conditions on countable Markovian dynamical systems ensuring the quasi-power property. The associated transfer operator (see (12) below) has

| Value of $s$ | Condition on $h$ | Entropy rate | Case |
|---|---|---|---|
| $s=1$ | $h(x) \underset{x\to+\infty}{\sim} c_2 \cdot x^{1/k}$ | $c_2 \cdot c_1^{1/k} \cdot \lambda'(1)$ | $(I)$ |
| | $h(x) \underset{x\to+\infty}{=} o(x^{1/k})$ | $0$ | $(II)$ |
| | $x^{1/k} \underset{x\to+\infty}{=} o(h(x))$ | $+\infty$ | $(III)$ |
| $s>1$ | $h(x) \underset{x\to0^+}{\sim} c_2 \cdot \log x$ | $c_2 \cdot \log \lambda(s)$ | $(IV)$ |
| | $h(x) \underset{x\to0^+}{=} o(\log x)$ | $0$ | $(V)$ |
| | $\log x \underset{x\to0^+}{=} o(h(x))$ | $+\infty$ | $(VI)$ |
| $\sigma_0 < s < 1$ | $h(x) \underset{x\to+\infty}{\sim} c_2 \cdot \log x$ | $c_2 \cdot \log \lambda(s)$ | $(VII)$ |
| | $h(x) \underset{x\to+\infty}{=} o(\log x)$ | $0$ | $(VIII)$ |
| | $\log x \underset{x\to+\infty}{=} o(h(x))$ | $+\infty$ | $(IX)$ |

TABLE II
VALUE OF THE ENTROPY RATE $\mathbb{H}_{h,\phi}$, ACCORDING TO THE BEHAVIOR OF $h$

| Entropy | Parameters | Entropy rate |
|---|---|---|
| Shannon | | $-\lambda'(1)$ |
| Rényi | $s=1$ | $-\lambda'(1)$ |
| | $s\neq 1$ | $\frac{1}{1-s}\log\lambda(s)$ |
| Varma | $r=t$ | $-\frac{1}{t^2}\lambda'(1)$ |
| | $r\neq t$ | $\frac{1}{t(t-r)}\log\lambda(r/t)$ |
| Havrda and Charvat | $r>1$ | $0$ |
| | $r=1$ | $\frac{-1}{\log 2}\lambda'(1)$ |
| | $r<1$ | $+\infty$ |
| Arimoto | $t>1$ | $+\infty$ |
| | $t=1$ | $-\lambda'(1)$ |
| | $t<1$ | $0$ |
| Sharma-Mittal 1 | $r<1$ | $+\infty$ |
| | $r>1$ | $0$ |
| | $s=r=1$ | $-\lambda'(1)$ |
| | $r=1\neq s$ | $\frac{1}{1-s}\log\lambda(s)$ |
| Sharma-Mittal 2 | | $\frac{1}{1-r}(\exp[-(r-1)\lambda'(1)]-1)$ |
| Taneja | $r<1$ | $+\infty$ |
| | $r=1$ | $-\lambda'(1)$ |
| | $r>1$ | $0$ |
| Sharma and Taneja | $r<1$ or $t<1$ | $+\infty$ |
| | $r>1$ and $t>1$ | $0$ |
| | $r=1$ and $t>1$ | $0$ |
| | $r=1$ and $t=1$ | $-\lambda'(1)$ |
| | $r>1$ and $t=1$ | $0$ |
| Tsallis | $r<1$ | $+\infty$ |
| | $r=1$ | $-\lambda'(1)$ |
| | $r>1$ | $0$ |

TABLE III
VALUES OF CLASSICAL ENTROPY RATES OF A RANDOM SEQUENCE SATISFYING THE QUASI-POWER PROPERTY WITH PARAMETERS $[\lambda, c, \rho, \sigma_0]$.

only one eigenvalue with maximum modulus, isolated from the remainder of the spectrum by a spectral gap. The dominant eigenvector is strictly positive and a spectral decomposition of the operators exists. Unfortunately, to exhibit the right topological partition $(I_j)_{j\in E}$ for which these conditions hold is quite difficult in terms of transition matrices. Therefore, in the following sections, we prefer to state conditions especially fitted to transition matrices under which we prove that the quasi-power property holds true for countable Markov chains.

### A. Finite Markov chains

Let $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ be an ergodic Markov chain with finite state space $E$, transition matrix $P = (p(i,j))_{i,j\in E}$ and initial distribution $\mu = (\mu(i))_{i\in E}$. The marginal distribution $\nu_n$ of $(X_0,\dots,X_{n-1})$ satisfies

$$\nu_n(i_0^{n-1}) = \mu(i_0)p(i_0,i_1)p(i_1,i_2)\dots p(i_{n-2},i_{n-1}).$$

The series $\Lambda_n(s)$ defined in (4) can be written in matrix form

$$\Lambda_n(s) = \mu_s \cdot P_s^{n-1} \cdot \mathbf{1},$$

where $P_s = (p(i,j)^s)_{i,j\in E}$ and $\mu_s$ is the column vector $(\mu(i)^s)_{i\in E}$. Since $P$ is irreducible and aperiodic, the same is true for $P_s$ for any $s$. In particular, $P_s$ has a unique dominant eigenvalue with maximum modulus. This eigenvalue $\lambda(s)$ is positive and its associated left and right eigenvectors, say $\mathbf{l}_s$ and $\mathbf{r}_s$, are also positive in the sense that all their components are positive. We deduce from these spectral properties that

$$\mathbf{v}\cdot P_s^{n-1} = \lambda(s)^{n-1}\cdot < \mathbf{v}, \mathbf{r}_s > \mathbf{l}_s + \mathbf{v}\cdot R^{n-1}(s),$$

where the spectral radius of $R(s)$ is strictly less than $\lambda(s)$. This defines the functions $\lambda$, $c$ and $\rho$ of the quasi-power property. They are analytic due to perturbation arguments that are detailed in [14].

Note that this result was indirectly proven in [18], thus inducing the explicit determination of the Rényi entropy rate of finite Markov chains.

### B. Denumerable Markov chains

Let $\mathbf{X} = (X_n)$ be a Markov chain with denumerable state space $E$, transition matrix $P = (p(i,j))_{i,j\in E}$ and initial distribution $\mu = (\mu(i))_{i\in E}$. The following assumptions will be proven to be sufficient for $\mathbf{X}$ to satisfy the quasi-power property.

**Assumptions 1**

*A.* $\sup_{(i,j)\in E^2} p(i,j) < 1$;

*B. there exists* $\sigma_0 < 1$ *such that for all* $s > \sigma_0$,

$$\sup_{i\in E}\sum_{j\in E} p(i,j)^s < \infty$$

*and*

$$\sum_{i\in E}\mu(i)^s < \infty;$$

*C. for all* $\varepsilon > 0$ *and all* $s > \sigma_0$, *there exists some* $A \subset E$ *with a finite number of elements, such that*

$$\sup_{i\in E}\sum_{j\in E\setminus A} p(i,j)^s < \varepsilon.$$

Before stating the main result of this section, let us prove a technical lemma which essentially transforms Point C of Assumption 1 into a more convenient form.

**Lemma 1** *If Assumptions 1 holds true, then for all* $s > \sigma_0$, *the operator* $P_s : (\mathcal{L}^1, ||\ .\ ||_1) \to (\mathcal{L}^1, ||\ .\ ||_1)$ *defined by* $P_s[\mathbf{v}] = \mathbf{v}\cdot P_s$ *is a compact operator on*

$$\mathcal{L}^1 = \{u = (u_i)_{i\in\mathbb{N}} : \|u\|_1 = \sum_{i\in\mathbb{N}} |u_i| < +\infty\}.$$

*Proof:* For the sake of simplicity, since any denumerable state space $E$ can be enumerated as a sequence $E = (i_k)_{k\in K}$ with $K = \mathbb{N}$, we will here set $E = \mathbb{N}$, so that Point C of Assumption 1 takes the form

$$\forall\varepsilon > 0,\ \forall s > \sigma_0,\ \exists N\in\mathbb{N},\quad \sup_{i\in\mathbb{N}}\sum_{j>N} p(i,j)^s < \varepsilon.$$

First, let us prove that for all $s > \sigma_0$, there exists a sequence of integers $N_k$ increasing to infinity such that

$$\sup_{i\in\mathbb{N}}\sum_{j>N_k} p(i,j)^s < \frac{1}{k}. \tag{10}$$

Point C of Assumption 1 says that for all $k\in\mathbb{N}^*$, some $N_k\in\mathbb{N}$ exists such that (10) holds true. If $N_k$ is replaced by $\sup_{l\le k} N_l$, the inequality remains true and the sequence is clearly increasing. If $N_k$ did not converge to infinity, some $j_0\in\mathbb{N}$ would exist such that $j_0 > N_k$ for all $k$, and hence,

$$\frac{1}{k} \ge \sup_{i\in\mathbb{N}}\sum_{j>N_k} p(i,j)^s \ge \sup_{i\in\mathbb{N}} p(i,j_0)^s.$$

Letting $k$ tend to infinity, we would obtain that $p(i,j_0)$ is zero for all $i$, which is untrue since the chain is irreducible.

Now, let us prove that $P_s$ is indeed a compact operator. Let $(u^n)_n$ denote a sequence of elements $u^n = (u_i^n)_{i\in\mathbb{N}}$ of $\mathcal{L}^1$ such that $||u^n||_1 \le 1$ and define $v^n = u^n\cdot P_s$. By induction on $k$, we can build a sequence of functions $s_k : \mathbb{N}\to\mathbb{N}$ such that $(s_{k+1}(n))_n$ is a subsequence of $(s_k(n))_n$ and such that for all $i \le N_k$, with $N_k$ such as in (10), $v_i^{s_k(n)}$ converges to some $v_i$. Then $v = (v_i)_{i\in\mathbb{N}}$ belongs to $\mathcal{L}^1$; indeed, for all $\varepsilon > 0$ and all $M\in\mathbb{N}$, there exists $N_k\in\mathbb{N}$, such that $M < N_k$ and for $n$ large enough,

$$\begin{aligned}\sum_{i<M}|v_i| &\le \sum_{i<N_k}|v_i - v_i^{s_k(n)}| + \sum_{i<N_k}|v_i^{s_k(n)}|\\ &\le \varepsilon + ||P_s||_1,\end{aligned}$$

and $||P_s||_1$ is finite by Point B of Assumption 1.

Let us now extract a subsequence of $(v^n)_n$ converging to $v$ for the $\mathcal{L}^1$-norm. Since $v$ belongs to $\mathcal{L}^1$, for all $k\in\mathbb{N}^*$, there exists $M_k\in\mathbb{N}$ such that

$$\sum_{i>M_k}|v_i| < \frac{1}{k}.$$

Let us set $k^* = \max(k,k')$, where $k'$ is such that $M_k < N_{k'}$. Then

$$\sum_{i\in\mathbb{N}}|v_i^{s_{k^*}(n)} - v_i| \le$$

$$\textstyle\sum_{i<N_{k^*}}|v_i^{s_{k^*}(n)} - v_i| + \sum_{i>N_{k^*}}|v_i^{s_{k^*}(n)}| + \sum_{i>N_{k^*}}|v_i|.$$

For $\tilde{n} = \tilde{n}(k)$ such that the first term in the sum is less than $1/k$, we get

$$\sum_{i\in\mathbb{N}} |v_i^{s_{k^*}(\tilde{n})} - v_i| \leq \frac{2}{k} + \sum_{i>N_{k^*}} |v_i^{s_{k^*}(n)}|.$$

Finally, since by Point C of Assumption 1 again,

$$\begin{aligned}\sum_{i>N_{k^*}} |v_i^{s_{k^*}(n)}| &\leq \sum_{j\in\mathbb{N}} |u_j^{s_{k^*}(n)}| \sum_{i>N_{k^*}} p(i,j)^s \\ &\leq ||u_j^{s_{k^*}(n)}||_1 \frac{1}{k^*} \leq \frac{1}{k},\end{aligned}$$

we obtain

$$\sum_{i\in\mathbb{N}} |v_i^{s_{k^*}(\tilde{n})} - v_i| \leq \frac{3}{k},$$

which concludes the proof that $P_s$ is compact. □

Lemma 1 allows us to prove that denumerable Markov chains satisfy the quasi-power property under mild conditions. Note that all results remain true if Point A of Assumption 1 is replaced by: there exist $n \in \mathbb{N}$ and $\eta < 1$ such that all the coefficients of $P^n$ are less than $\eta$.

**Theorem 2** *Let $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ be an irreducible and aperiodic Markov chain with transition matrix $P = (p(i,j))_{(i,j)\in E^2}$ and initial distribution $\mu = (\mu(i))_{i\in E}$. If Assumption 1 holds true, then $\mathbf{X}$ satisfies the quasi-power property.*

*Proof:* It follows from Lemma 1 that $P_s$ is compact for all $s > \sigma_0$. Therefore, the spectrum of $P_s$ over $\mathcal{L}^1$ is a sequence converging to zero. Hence, $P_s$ has a finite number of eigenvalues with maximum modulus and there exists a spectral gap separating these dominant eigenvalues from the remainder of the spectrum; details can be found in [14].

Further, since $\mathbf{X}$ is irreducible and aperiodic, $P_s$ is a non-negative irreducible and aperiodic infinite matrix, so has a unique dominant eigenvalue $\lambda(s)$ which, moreover, is positive. We deduce from these spectral properties the following spectral decomposition of the iterates of $P_s$,

$$u \cdot P_s^n = \lambda(s)^n \cdot u \cdot Q_s + u \cdot R_s^n, \qquad u \in \mathcal{L}^1, \tag{11}$$

where $Q_s$ is the projector over the dominant eigenspace and $R_s$ is the projector over the remainder of the spectrum. In particular, the spectral radius of $R_s$ can be written $\rho(s)\cdot\lambda(s)$ with $\rho(s) < 1$. Finally, $\Lambda_n(s)$ is given by the $\mathcal{L}^1$-norm of $\mu_s \cdot P_s^{n-1}$, where $\mu_s = (\mu(i)^s)_{i\in E}$, so that

$$\Lambda_n(s) = \lambda(s)^{n-1} ||\mu_s \cdot Q_s||_1 [1 + O(\rho(s)^{n-1}\lambda(s)^{n-1})],$$

which means that $\mathbf{X}$ satisfies the quasi-power property. The analyticity with respect to $s$ of all the functions involved in (11) is due jointly to the analyticity of $s \to P_s$ and to perturbation arguments detailed in [14]. Moreover, for $s < t$, due to Point A of Assumption 1,

$$\begin{aligned}\Lambda_n(t) &= \sum_{i_0^{n-1}\in E^n} \Pr(i_0^{n-1})^t \\ &\leq \eta^{n(t-s)} \sum_{i_0^{n-1}\in E^n} \Pr(i_0^{n-1})^s = \eta^{(n-1)(t-s)}\Lambda_n(s),\end{aligned}$$

where $\eta = \sup_{(i,j)\in E^2} p(i,j)$, so that $s \to \lambda(s)$ is strictly decreasing. □

Then, applying Theorems 1 and 2 to $\phi$ satisfying (7), the $(h,\phi)$-entropy rate of the chain is given by

$$\mathbb{H}_{h(x),\phi(y)}(\mathbf{X}) = \lim_{n\to+\infty} \frac{1}{n} h\left(c_1 c(s) \lambda'(s)^k n^k \lambda(s)^{n-k-1}\right).$$

**Remarks**

1. In dynamical sources theory, the perturbated matrices $P_s$ are replaced for Bernoulli sources by the transfer operators $\mathbf{G}_s$ defined by

$$\mathbf{G}_s[f](y) = \sum_{x:T(x)=y} \frac{f(x)}{|T'(x)|^s},$$

and for Markovian sources by the secant operators $\mathbb{G}_s$ defined by

$$\mathbb{G}_s[F](y_1,y_2) = \sum_{i\in E} \sum_{(x_1,x_2)\in I_i(y_1,y_2)} \frac{|x_1-x_2|^s}{|T(x_1)-T(x_2)|^s} F(x_1,x_2), \tag{12}$$

where $I_i(y_1,y_2) = \{(x_1,x_2) \in I_i | T(x_1) = y_1, T(x_2) = y_2\}$.

2. As noticed in the introduction, Golshani *et al* deal in [12] with the denumerable case for the Rényi entropy rate but their proofs contain some errors. Indeed, they use results issued from the $R$-theory of non-negative matrices developed in [26]. In particular, they invoke the following asymptotic argument: if $T$ is a positive irreducible and aperiodic matrix with radius of convergence $R$, then for all states $i,j \in E$, the $(i,j)$ coefficient of $R^nT^n$ converges to some finite value $\mu_{i,j}$. Actually, in [12], the expression of the Rényi entropy rate involves the double sum $S_n = \sum_{i,j\in E}(R^nT^n)_{i,j} f_j$ for large $n$, where $(f_j)_{j\in E}$ is related to the initial distribution of the Markov chain, and $T$ is a perturbation of the transition matrix of the chain. The authors exchange the limit with respect to $n$ with the double infinite sum whereas the uniform convergence is not proven to hold true. On the contrary, [26, Theorem 6.2] states necessary and sufficient conditions to allow the change when $T$ is $R$-positive recurrent; unfortunately, these conditions involve the generally unknown $R$-invariant vectors, which makes them difficult to check in practice. Furthermore, even if the transition matrix of the chain was supposed to be positive recurrent, it would remain to prove that its perturbation $T$ shares the same property.

In the above first part of the paper, we have explicitly obtained the generalized $(h,\phi)$-entropy rate of random sequences satisfying the quasi-power property. We have also given simple assumptions on countable Markov chains for the quasi-power property to hold. In the second part below, we will focus on the estimation of the entropy rates for parametric Markov chains using the expressions given in Table III and plug-in estimators built from the MLE of the parameter.

## V. Estimation of entropy for denumerable Markov chains

We suppose that the transition probabilities of the chain depend on an unknown parameter $\theta \in \Theta^d$, where $\Theta$ is an

open subset of some Euclidean space and $d \geq 1$.

The partial derivatives will be denoted with a subscript, as for example $f_u = \partial f/\partial\theta_u$. The expectation under the value $\theta$ of the parameter will be denoted $\mathbb{E}_\theta$. The true value of the parameter will be denoted by $\theta^0$.

### A. Estimation of the parameters

Let $\mathbf{X} = (X_0, \ldots, X_n)$ be a sample of the chain. Let $(x_0, \ldots, x_n) \in E^{n+1}$ denote an observation; the associated log-likelihood is

$$\log \mu(x_0) + \sum_{m=0}^{n-1} \log p(x_m, x_{m+1}; \theta),$$

where $\mu$ denotes the initial distribution of the chain. The information contained in the observation of this initial distribution does not increase with $n$. Hence, for a large sample theory, it is convenient to consider the value of $\theta$ maximizing the pseudo log-likelihood

$$\sum_{m=0}^{n-1} \log p(x_m, x_{m+1}; \theta).$$

Asymptotic results on the MLE $\widehat{\theta}_n$ of the parameter $\theta$ thus obtained are proven in [3] under the following regularity assumptions.

**Assumptions 2**

*A. For any $x$, the set of $y$ for which $p(x, y; \theta) > 0$ does not depend on $\theta$.*

*B. For any $(x, y)$, the partial derivatives $p_u(x, y; \theta)$, $p_{uv}(x, y; \theta)$ and $p_{uvw}(x, y; \theta)$ exist and are continuous with respect to $\Theta$.*

*C. For all $\theta \in \Theta$, there exists a neighborhood $N$ such that for any $u, v, x, y$, the functions $p_u(x, y; \theta)$ and $p_{uv}(x, y; \theta)$ are uniformly bounded in $L^1(\mu(dy))$ on $N$ and*

$$\mathbb{E}_\theta[\sup_{\theta' \in N} \mid p_u(x, y; \theta') \mid^2] < +\infty.$$

*D. There exists $\alpha > 0$ (possibly depending on $\theta$) such that $\mathbb{E}_\theta[\mid p_u(x, y; \theta) \mid^{2+\alpha}]$ is finite, for $u = 1, \ldots, d$.*
*E. The $d \times d$ Fisher information matrix $\sigma(\theta) = (\sigma_{uv}(\theta))$ is non singular, where $\sigma_{uv}(\theta) = \mathbb{E}_\theta[p_u(x, y; \theta) p_v(x, y; \theta)]$.*

**Proposition 1** *If Assumptions 2 are satisfied, then a strongly consistent MLE $\widehat{\theta}_n$ of $\theta$ exists. Moreover, $\sqrt{n}(\widehat{\theta}_n - \theta_n)$ is asymptotically centered and normal, with covariance matrix $\sigma^{-1}(\theta^0)$.*

Note that if $n$ is large, there is exactly one MLE in $N$. We refer to [16] for weaker differentiability assumptions on the transition functions.

Any finite Markov chain can be considered as a parametric chain, with parameters $\theta_{i,j} = p(i, j)$, for $i \neq j$. The MLE of the transition probabilities are the empirical ones, defined by

$$\widehat{p}_n(i, j) = \frac{\mathbf{N}_n(i, j)}{\mathbf{N}_n(i)} \mathbf{1}_{N_n(i) > 0},$$

where

$$\mathbf{N}_n(i, j) = \sum_{m=1}^{n} \mathbf{1}_{\{X_{m-1} = i, X_m = j\}},$$

and

$$\mathbf{N}_n(i) = \sum_{j \in E} \mathbf{N}_n(i, j) = \sum_{m=0}^{n-1} \mathbf{1}_{X_m = i}, \quad i, j \in E,$$

The estimators $\widehat{p}_n(i, j)$ are strongly consistent and asymptotically normal. Precisely, when $n$ tends to infinity, the vector $(\sqrt{n}[\widehat{p}_n(i, j) - p(i, j)])$ converges in distribution to a centered Gaussian vector with covariances $\delta_{ik}[\delta_{jl} p(i, j) - p(i, j)p(i, l)]/\pi(i)$ for $1 \leq i, j, k, l \leq |E|$.

A natural estimator of the stationary distribution $\pi$ is the empirical estimator

$$\widehat{\pi}_n(i) = \frac{\mathbf{N}_n(i)}{n}, \quad i \in E.$$

It is strongly consistent and asymptotically normal. Precisely, when $n$ tends to infinity,

$$\sqrt{n}\left[\widehat{\pi}_n(i) - \pi(i)\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(0,\ \pi(i)^2[2\mathbb{E}_\pi \tau(i) - 1] - \pi(i)\right),$$

where $E_\pi \tau(i)$ is the expectation of the return time $\tau(i)$ of the chain to state $i$ when the initial distribution is $\pi$.

These asymptotic properties derive from the law of large numbers and central limit theorem for Markov chains; see [7] for details.

For a finite chain with state space $E$, the transition probabilities may also be functions of a number $d$ of parameters strictly smaller than $|E|(|E|-1)$. In this case, Points C to E of Assumption 2 reduce to: for any $\theta$, the $d \times d$ matrix $\left(\frac{p_u}{p}(i, j)\right)$ has rank $d$; see [3].

### B. Estimation of marginal entropy

Since the transition probabilities of the chain depend on $\theta$, the stationary distribution also depends on $\theta$. It is natural to consider the plug-in estimator $\mathbb{S}_{h(y),\phi(x)}(\pi(\hat{\theta}_n))$ of $\mathbb{S}_{h(y),\phi(x)}(\pi(\theta))$.

**Theorem 3** *Let $\mathbf{X}$ be an ergodic homogeneous finite Markov chain satisfying the quasi-power property. If Assumption 2 is satisfied, then the estimator $\mathbb{S}_{h(y),\phi(x)}(\pi(\hat{\theta}_n))$ is strongly consistent. If, moreover, the differential function $D_\theta \mathbb{S}_{h(y),\phi(x)}(\pi)$ is not null at $\theta^0$, then the plug-in estimator is asymptotically normal. Precisely*

$$\sqrt{n}[\mathbb{S}_{h(y),\phi(x)}(\pi(\hat{\theta}_n)) - \mathbb{S}_{h(y),\phi(x)}(\pi(\theta))] \to \mathcal{N}(0, \Sigma_\pi),$$

*where*

$$\Sigma_\pi = \left[D_{\theta^0} \mathbb{S}_{h(y),\phi(x)}(\pi)\right]^t \sigma^{-1}(\theta^0) \left[D_{\theta^0} \mathbb{S}_{h(y),\phi(x)}(\pi)\right].$$

*Proof:* We know from Proposition 1 that $\widehat{\theta}_n$ converges almost surely to $\theta^0$. Due to operators properties detailed in [1] (see particularly p94), the eigenvector $\pi$ is known to be a continuously differentiable function of the operator; using Point B of Assumption 2 shows that $\pi(x, y; \theta)$ is absolutely continuous with respect to $\theta$. The continuous mapping theorem

implies that $\mathbb{S}_{h(y),\phi(x)}(\pi(\hat{\theta}_n))$ converges almost surely to $\mathbb{S}_{h(y),\phi(x)}(\pi(\theta^0))$.

Then, the normality of $\mathbb{S}_{h(y),\phi(x)}(\pi(\hat{\theta}_n))$ follows from Proposition 1 by the delta method. □

### C. Estimation of entropy rates

Table III shows that when the $(h,\phi)$-entropy rate is neither null nor infinite, only two cases happen. Either, the entropy rate is equal to $-\lambda'(1)$, that is to Shannon entropy rate, or it is a simple function of Rényi entropy rate, that is of $(1-s)^{-1}\log\lambda(s)$. Therefore, we will only detail the estimation of Shannon and Rényi entropy rates.

The estimation of Shannon entropy rate has already been considered by two of the authors in [7], mainly for finite chains for which estimation is detailed under different schemes of observation, with a plug-in method based on (2). It allowed them to prove the asymptotic normality of plug-in estimators for finite chains but does not apply to the denumerable case. We will solve the problem here, for any countable parametric chains, by applying results from operators theory.

Let us define the plug-in estimators $\mathbb{H}(\hat{\theta}_n) = -\lambda'(1;\hat{\theta}_n)$ of Shannon entropy rate $\mathbb{H}(\theta)$, and $\mathbb{H}_s(\hat{\theta}_n) = (1-s)^{-1}\log\lambda(s;\hat{\theta}_n)$ of Rényi entropy rate $\mathbb{H}_s(\theta)$.

**Theorem 4** *Let* $\mathbf{X}$ *be an ergodic homogeneous countable Markov chain satisfying the quasi-power property. If Assumption 2 is satisfied, then the estimators* $\mathbb{H}(\hat{\theta}_n)$ *and* $\mathbb{H}(s;\hat{\theta}_n)$ *for* $s \neq 1$*, are strongly consistent and asymptotically normal. Precisely*

$$\sqrt{n}[\mathbb{H}(\hat{\theta}_n) - \mathbb{H}(\theta)] \to \mathcal{N}(0,\Sigma_1),$$

*where*

$$\Sigma_1 = \left\{\frac{\partial}{\partial\theta}[-\lambda'(1;\theta)]\right\}^t \sigma^{-1}(\theta)\frac{\partial}{\partial\theta}[-\lambda'(1;\theta)]$$

*and* $\sqrt{n}[\mathbb{H}_s(\hat{\theta}_n) - \mathbb{H}_s(\theta^0)] \to \mathcal{N}(0,\Sigma_s)$*, where*

$$\Sigma_s = \frac{1}{(1-s)^2}\left\{\frac{\partial}{\partial\theta}\lambda(s;\theta)\right\}^t \sigma^{-1}(\theta)\frac{\partial}{\partial\theta}\lambda(s;\theta).$$

*Proof:* For a parametric chain depending on $\theta$, let us set $p_s(x,y,\theta) = p(x,y,\theta)^s$. Due to operators properties (see again [1, p94]), the eigenvalue $\lambda(s,\theta)$ of the perturbated operator defined by $P_s = (p_s(x,y,\theta))$ and its derivative $\lambda'(s,\theta)$ are known to be continuous with respect to $P_s$. Point B of Assumption 2 induces that $P_s$ too is a continuously differentiable function of $\theta$. Therefore both $\lambda(s;\theta)$ and $\lambda'(s;\theta)$ are continuous with respect to $\theta$. The results follow from the continuous mapping theorem and the delta method.□

## REFERENCES


[1] Ahues, A., Largillier A. and Limaye B. V. *Spectral Computations for Bounded Operators* Chapman & Hall/CRC, 2001.

[2] Amblard, P.-O., and Vignat, C., A note on bounded entropies, *Physica A*, vol. 365 (1) pp50–56, 2006.

[3] Billingsley, P., *Statistical Inference for Markov Processes* The university of Chicago Press, 1961.

[4] Bourdon, J., Nebel M. E.and Vallée B., On the Stack-Size of General Tries, *ITA,* vol. 35 (2), pp163-185, 2001.

[5] Chauveau, D., and Vandekerkhove, P., Monte Carlo estimation of the entropy for Markov chains. *Meth. Comp. Appl. Probab.*, vol. 9 (1), pp133–149, 2007.

[6] Chazal, F., and Maume-Deschamps, V., Statistical properties of General Markov dynamical sources: applications to information theory *Discrete Math. Theor. Comp. Sc.* vol. 6 (2), pp283–314, 2004.

[7] Ciuperca, G., and Girardin, V., Estimation of the Entropy Rate of a Countable Markov Chain *Comm. Stat. Th. Methods*, vol. 36, pp2543–2557, 2007.

[8] Cover, L., and Thomas, J., *Elements of information theory.* Wiley series in telecommunications, New-York, 1991.

[9] Furuichi, S., Information theoretical properties of Tsaliis entropies *J. Math. Physics*, vol. 47, 2006.

[10] Girardin, V., On the Different Extensions of the Ergodic Theorem of Information Theory, in *Recent Advances in Applied Probability*, R. Baeza-Yates, J. Glaz, H. Gzyl, J. Hüsler and J. L. Palacios (Eds), Springer-Verlag, San Francisco, pp163–179, 2005.

[11] Girardin, V., and Sesboüé, A., Comparative Construction of Plug-in Estimators of the Entropy Rate of Two-State Markov Chains, *Method. Comput. Appl. Probab.*, V11, pp181–200, 2009.

[12] Golshani, L., Pasha, E., and Yari, G., Some properties of Rényi entropy and Rényi entropy rate, *Inf. Sci.*, vol. 179 (14), pp2426–2433, 2009.

[13] Harremoës, P., Interpretations of Rényi Entropies and Divergences *Physica A* vol. 365 (1), pp57–62, 2006.

[14] Kato, T., *Perturbation Theory for Linear Operators*, 2d edition, Springer-Verlag, Berlin, 1976.

[15] Menéndez, M.L., Morales, D., Pardo, L., and Salicrú, M., $(h,\Phi)$-entropy differential metric *Appl. Math.*, vol. 42 (2), pp81-98, 1997.

[16] Prakasa Rao, B.L.S., Maximum Likelihood Estimation for Markov Process. *Ann. Inst. Stat. Math.* vol. 24, pp333–345, 1972.

[17] Rached, Z., *Renyi's Entropy for Discrete Markov Sources.* Master of Science Project, September, 1998.

[18] Rached, Z., Alajaji, F., and Campbell, L. L., Rényi's Entropy Rate for Discrete Markov Sources, *Proc. CISS*, pp613-618, 1999.

[19] Rényi, A., On measures of information and entropy, *Proc. 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pp547-561, 1960.

[20] Salicrú, M., Menéndez, M. L., Morales, D., and Pardo, L. Asymptotic distribution of $(h,\phi)$-entropies, *Comm. Stat. (Theory and Methods)* vol. 22 (7), pp2015–2031, 1993.

[21] Shannon, C., A mathematical theory of communication. *Bell Syst.Techn. J.* vol. 27, pp379–423 and pp623-656, 1948.

[22] Shao, J., *Mathematical Statistics* Springer-Verlag, New York, 2003.

[23] Sharma, B.D., and Mittal, P., New non-additive measures of relative information J. *Comb. Inform. and Syst. Sci.* vol.2, pp122–133, 1975.

[24] Tsallis, C., Possible generalization of Boltzmann-Gibbs statistics, *J. Stat. Physics* vol. 52 pp479–487, 1988.

[25] Vallée, V., Dynamical Sources in Information Theory: Fundamental Intervals and Word Prefixes, *Algorithmica* vol. 29, pp262–306, 2001.

[26] Vere-Jones, D., Ergodic properties of nonnegative matrices. I and II *Pacific J. Math.* vol. 22 (2) pp361–386, 1967 and vol.26, issue 3, pp601-620, 1968.

[27] Wachowiak, M.P., Smolikova, R., Tourassi, G.D., Elmaghray A.S., Estimation of generalized entropies with sample spacing, *Pattern Anal. Applic.* vol. 8, pp95–101, 2005.