On the estimation of the entropy rate of finite Markov chains

Gabriela Ciuperca[1] and Valerie Girardin[2]

[1] Université LYON I, LaPCS, 50 Av. Tony-Garnier, 69366 Lyon cedex 07, France, gabriela.ciuperca@pop.univ-lyon1.fr
[2] LMNO, UMR6139, Campus II, Université de Caen, BP5186, 14032Caen, France, girardin@math.unicaen.fr

Abstract. We consider here ergodic homogeneous Markov chain with finite state spaces. We study an estimator of the entropy rate of the chain based on the maximum likelihood estimation of the transition matrix. We prove its asymptotic properties for estimation from one sample with long length or many independent samples with given length. This result has potential applications in all the real situations modeled by Markov chains, as detailed in the introduction.
Keywords: entropy rate, homogeneous Markov Chain, maximum likelihood estimation.

1 Introduction

Markov chains and entropy are linked since the introduction of entropy in probability theory by Shannon [24]. He defined the entropy of a distribution P taking values in a finite set, say $E = \{1, \ldots, s\}$, as $\mathbb{S}(P) = -\sum_{i=1}^{s} p_i \log p_i$, with the convention $0 \ln 0 = 0$.

For a discrete-time process $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$, the entropy at time n is defined as the Shannon entropy of the n-dimensional marginal distribution of $\mathbf{X}$. Under suitable conditions, the entropy at time n divided by n converges. When the limit $\mathbb{H}(\mathbf{X})$ is finite, it is called the entropy rate of the process.

The entropy rate was first defined in [24] for an ergodic Markov chain with a finite state space E as the sum of the entropies of the transition probabilities $(p_{ij})_{j=1,\ldots,s}$ weighted by the probability of occurrence of each state according to the stationary distribution π of the chain, namely

$$\mathbb{H}(\mathbf{X}) = -\sum_{i=1}^{s} \pi_i \sum_{j=1}^{s} p_{ij} \log p_{ij}. \tag{1}$$

Shannon [24] proved the convergence of $n^{-1} \log \mathbb{P}(X_1 = i_1, \ldots, X_n = i_n)$ to $\mathbb{H}(\mathbf{X})$ in probability. McMillan [16] proved the convergence in mean for any stationary ergodic process with a finite state space. This constitutes the Shannon-McMillan theorem. The almost sure convergence proven by Breiman [4] is known as the Shannon-McMillan-Breiman theorem. Many extensions have been proven since (see [10] and the references therein), but the

entropy rate has an explicit form only for Markov or semi-Markov processes (see [12]).

The entropy $\mathbb{S}(P)$ of a distribution P is widely used in all applications involving random variables; see [6], [8] and the references therein. The entropy rate $\mathbb{H}(\mathbf{Y})$ of an i.i.d. sequence with distribution P is the entropy $\mathbb{S}(P)$ of P. A whole statistical tool-box has been developed in this regard and applied to a wide range of applied domains. Having an explicit form for the entropy rate of a Markov chain allows one to use it similarly in all applications involving Markov modeling. For example, maximum entropy methods can be considered (see [9]).

It is well-known that in information theory, the entropy rate of a source measures its degree of complexity (see [6]), but the entropy rate is used in many other applied fields. In time series theory, the ApEnt coefficient describes the degree of hazard in a time series, and Pincus [20] proved that for a Markovian model, the ApEnt is equal to the entropy rate of the chain. In finance, Kelly [14] introduced entropy for gambling on horse races, and Breiman [5] for investments in general markets; Shannon-McMillan-Breiman theorem appears as an ergodic theorem for the maximum growth of compounded wealth when gambling on a sequence of random variables (see [6]), and the admissible self-financing strategy achieving the maximum entropy is a growth optimal strategy (see [15]).

When observations of the process are available, the need for estimating the entropy rate obviously arises.

Approximations of entropy can be obtained by numerical algorithms. The Ziv-Lempel algorithm allows one to get an approximation of the entropy of a binary process, whichever be its distribution. Plotkin & Wyner [21] derive an algorithmic estimator of the entropy rate for a queueing problem in telecommunication networks, for measuring the scattering and clustering of cells. Abundo et al. [1] compute numerical approximations of the entropy rate via the ApEnt to explain the degree of cooperativity of proteins in a Markov model with binomial transition distributions.

Basharin [3] introduced estimation of the entropy rate in the statistical theory of random processes by considering the maximum likelihood (ML) estimator $\widehat{p}_i = n^{-1} \sum_{k=1}^{n} \mathbb{1}_{(X_k=i)}$ and the plug-in estimator $\widetilde{H} = -\sum_{i=1}^{s} \widehat{p}_i \log \widehat{p}_i$ of $\mathbb{H}(\mathbf{Y})$, for an i.i.d. sequence $\mathbf{Y} = (Y_n)$ with distribution $P = (p_1, \ldots, p_s)$ on a finite state space $E = \{1, \ldots, s\}$. He proved that $\widetilde{H}$ is biased but strongly consistent and asymptotically normal. Misevichyus [18] considers an estimator of the entropy rate of an homogeneous stationary Markov chain with finite state space, based on the ML estimation of the transition probabilities.

For an estimation based on one sample of long length, problems may arise from the non-observation of some states, especially if s is large. Several procedures exist in order to avoid these problems.

Meeden [17] constructs an estimator of the transition matrix by a ML method modified by a Bayes procedure. He proves that this estimator is admissible when the loss function is the sum of individual squared error losses.

Another procedure consists in the series schemes (the number of observed states, their probabilities and the transition probabilities may vary with n). The main issue of these methods is the determination of the asymptotic distribution (possibly normal, but also Poisson, centered or non-centered chi-square, etc.) of the estimators thus obtained. For an i.i.d. sequence, Zubkov [27] gives conditions on the series scheme for the asymptotic normality of $\widetilde{H}$. Mukhamedkhanova [19] studies the class of asymptotic distributions of an estimator based on the ML estimation of the transition probabilities for a two-state stationary Markov chain.

Another approach consists in using several samples of finite length in which all the states are observed infinitely often; see [2], [13, Chapter V] or [23]. Moreover, practically, it may be simpler to observe many independent trajectories of the chain with short length rather than one long trajectory.

We study here ergodic homogeneous but non necessarily stationary Markov chains with finite state spaces. We study the estimator of the entropy rate for non-stationary chains and prove its asymptotic properties for an estimation based one sample in Section 3. We generalize it to an estimation based on several samples in Section 4. Some extension prospects are given in Section 5.

2 Notation and definitions

Let (X_n) be an homogeneous ergodic (that is irreducible and aperiodic) Markov chain with finite state space $E = \{1, \dots, s\}$ and stationary distribution $(\pi_i)_{i=1,\dots,s}$. Set, for $i, j = 1, \dots, s$,

$$\begin{aligned} p_i^{(n)} &= \mathbb{P}(X_n = i), \quad n \geq 0, \\ p_{ij} &= \mathbb{P}(X_n = j | X_{n-1} = i), \quad n \geq 1, \\ p_{(i,j)}^{(n)} &= p_{ij} p_i^{(n)} = \mathbb{P}(X_n = j, X_{n-1} = i), \quad n \geq 1, \end{aligned}$$

in which p_{ij} does not depend on n due to the homogeneity of the chain. We know from the ergodic theorem of Markov chains that $p_i^{(n)}$ converges to π_i when n tends to infinity (see, e.g., [11]).

We will also consider the bidimensional Markov chain (X_n, X_{n-1}), which is homogeneous and ergodic too, with transition probabilities

$$\mathbb{P}(X_{n+1} = l, X_n = k | X_n = j, X_{n-1} = i) = p_{ij} \delta_{jk}, \tag{2}$$

(where δ_{jk} denotes Kronecker's symbol). Its stationary distribution is given by $\pi_{(i,j)} = \pi_i p_{ij}$. Indeed, since π is the stationary distribution of $\mathbf{X}$, we have

$\sum_{i'=1}^{s} \pi_{i'} p_{i'i} = \pi_i$, or

$$\sum_{i'=1}^{s} \pi_{i'} p_{ij} p_{i'i} = \pi_i p_{ij}, \quad i,j = 1,\dots,s,$$

which is equivalent to

$$\sum_{i',j'=1}^{s} \pi_{(i',j')} p_{ij} \delta_{j'i} = \pi_{(i,j)} \quad i,j = 1,\dots,s.$$

Note that $p_{(i,j)}^{(n)}$ converges to $\pi_{(i,j)}$ when n tends to infinity.

The entropy rate of the chain $\mathbf{X}$, given in (1), can be written

$$\mathbb{H}(\mathbf{X}) = \sum_{i=1}^{s} \pi_i \log \pi_i - \sum_{i=1}^{s} \sum_{j=1}^{s} \pi_{(i,j)} \log \pi_{(i,j)}, \tag{3}$$

This decomposition will be the basis of the definition of the estimators of $\mathbb{H}(\mathbf{X})$ considered in the following.

3 Estimation from one sample with long length

Suppose we are given one observation of the chain, say $X = (X_0,\dots,X_n)$. Let us set for $i,j = 1,\dots,s$,

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

It is well-known (see [2, Section 5] and the references therein, and also [23]) that the following estimators of the transition probabilities (p_{ij}),

$$\widehat{p}_{ij} = \frac{\mathbf{N}_n(i,j)}{\mathbf{N}_n(i)},$$

are their ML estimators. Clearly, the stationary distribution (π_i) is estimated by

$$\widehat{\pi}_i = \frac{\mathbf{N}_n(i)}{n}, \quad i,j = 1,\dots,s,$$

Note that when $\mathbf{N}_n(i) = 0$, it is necessary to set $\widehat{p}_{ij} = 0$ for all $j = 1,\dots,s$, and $\widehat{\pi}_i = 0$. When $\mathbf{N}_n(i) \neq 0$ and $\mathbf{N}_n(i,j) = 0$, we also have $\widehat{p}_{ij} = 0$ and suppose that $p_{ij} = 0$. Note that the scheme of estimation considered below in Section 4 constitutes a means of avoiding such problems of non-observation.

The asymptotic properties given in the following proposition derive from the law of large numbers and central limit theorem for Markov chains (see, e.g., [7]).

Proposition 1 *The estimators $\widehat{p}_{ij}$ and $\widehat{\pi}_i$ are strongly consistent and asymptotically normal, in mathematical words, when n tends to infinity,*

$$\widehat{\pi}_i \xrightarrow{a.s.} \pi_i \quad \textit{and} \qquad \sqrt{n}(\widehat{\pi}_i - \pi_i) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \pi_i(1-\pi_i)),$$
$$\widehat{p}_{ij} \xrightarrow{a.s.} p_{ij} \textit{ and } \sqrt{n}\pi_i(\widehat{p}_{ij} - p_{ij}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p_{ij}(1-p_{ij})).$$

Replacing in (3) the probabilities by their estimators, we get the following estimator for the entropy rate,

$$\widehat{\mathbb{H}}_n = \sum_{i=1}^{s} \widehat{\pi}_i \log \widehat{\pi}_i - \sum_{i=1}^{s}\sum_{j=1}^{s} \widehat{\pi}_{(i,j)} \log \widehat{\pi}_{(i,j)},$$

where $\widehat{\pi}_{(i,j)} = \widehat{\pi}_i \widehat{p}_{ij} = n^{-1}\mathbf{N}_n(i,j)$.

Misevichyus [18] proved the following theorem in the particular case of a stationary chain (whose initial distribution is the stationary one). We give here a shorter proof which holds true for any chain.

Theorem 1 *Let $\mathbf{X}$ be an homogeneous ergodic Markov chain with a finite state space. Then the estimator $\widehat{\mathbb{H}}_n(K)$ of $\mathbb{H}(\mathbf{X})$ is*
1. strongly consistent;
2. asymptotically normal and unbiased when n tends to infinity.

Proof of Theorem 1

1. For proving that $\widehat{\mathbb{H}}_n$ converges almost surely to H when n tends to infinity, it is sufficient to apply [26, Theorem 1.10, p59].

2. Set

$$\widehat{\mathbb{H}}_1 = \sum_{i=1}^{s} \widehat{\pi}_i \log \widehat{\pi}_i \quad \text{and} \quad \widehat{\mathbb{H}}_2 = -\sum_{i=1}^{s}\sum_{j=1}^{s} \widehat{\pi}_{(i,j)} \log \widehat{\pi}_{(i,j)}.$$

Since by Proposition 1, $\widehat{\pi}_i$ converges almost surely to π_i when n tends to infinity, the Taylor's formula for $x \log x$ at π_i, for $\pi_i \neq 0$, implies that

$$\widehat{\mathbb{H}}_1 = H_1 + \sum_{i=1}^{s} (\log \pi_i + 1)(\widehat{\pi}_i - \pi_i) - \frac{1}{2}\sum_{i=1}^{s} \frac{(\widehat{\pi}_i - \pi_i)^2}{[\pi_i + \Theta_1(\widehat{\pi}_i - \pi_i)]^3},$$

for some $0 < \Theta_1 < 1$.

Clearly, $\mathbb{E}[\widehat{\pi}_i - \pi_i]$ converges to zero when n tends to infinity. We get from Proposition 1 that $\mathbb{E}[\widehat{\pi}_i - \pi_i]^2 = \mathrm{O}(n^{-1})$. Hence $\widehat{\mathbb{H}}_1$ is asymptotically unbiased.

By Proposition 1, $\sqrt{n}(\widehat{\pi}_i - \pi_i)$ converges in distribution to $\mathcal{N}(0, \pi_i(1-\pi_i))$ when n tends to infinity, hence the delta method (see, e.g., [25]) applies to prove that $\sqrt{n}(\widehat{\mathbb{H}}_1 - H_1)$ is asymptotically centered and normal.

Since $(\pi_{(i,j)})_{i,j=1,\ldots,s}$ is the stationary distribution of the bidimensional chain given in (2), the same arguments hold for H_2, and the conclusion follows. □

4 Estimation based on several independent samples with fixed length

Suppose we are given K independent observations of the chain, say $X^{(k)} = (X_0^{(k)}, \ldots, X_n^{(k)})$, $k = 1, \ldots, K$, for a fixed integer n. Let us set

$$\mathbf{n}_K(i) = \sum_{k=1}^{K} \mathbb{1}_{\{X_0^{(k)}=i\}} = \mathrm{Card}\{X_0^{(k)} = i : \ k = 1, \ldots, K\},$$

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

$$\text{and} \quad \mathbf{N}_{n,K}(i) = \sum_{j=1}^{s} \mathbf{N}_{n,K}(i,j).$$

The following ML estimators of the transition probabilities (p_{ij}),

$$\widehat{p}_{ij}(n,K) = \frac{\mathbf{N}_{n,K}(i,j)}{\mathbf{N}_{n,K}(i)}, \quad i,j = 1, \ldots, s,$$

have been computed and studied in [2].

Suppose that when K tends to infinity, $\mathbf{n}_K(i)/K$ converges to a finite quantity, say η_i, for all $i = 1, \ldots, s$ (with $\eta_i > 0$ and $\sum_{i=1}^{s} \eta_i = 1$). Then, the ML estimators $\widehat{p}_{ij}(K)$ are strongly consistent and Anderson & Goodman [2] proved that

$$\sqrt{\mathbf{N}_{n,K}(i)}\ [\widehat{p}_{ij}(K) - p_{ij}] \xrightarrow{\mathcal{L}} \mathcal{N}(0, p_{ij}(1-p_{ij})).$$

Note that for the above result to hold, the initial distribution of the chain $\mathbf{n}_K(i)$ can be supposed to be either non-random, with multinomial distribution $\mathcal{M}(K, (\eta_i)_{i=1,\ldots,s})$ or equal to the stationary distribution of the chain.

For estimating the stationary distribution from samples with finite length, it is easy to see that it is necessary for the chain to be stationary, with then

$$\widehat{\pi}_i(K) = \frac{\mathbf{n}_K(i)}{K}, \quad i = 1, \ldots, s.$$

Proposition 2 *Suppose that the chain is stationary and that K is such that $\mathbf{n}_K(i)/K$ converges to a finite quantity, say η_i, for all $i = 1, \ldots, s$, when K tends to infinity. Then, the estimators $\widehat{\pi}_i(K)$ and $\widehat{p}_{ij}(K)$ are strongly consistent and asymptotically normal, in mathematical words,*

$$\widehat{\pi}_i \xrightarrow{a.s.} \pi_i \text{ and } \sqrt{K}(\widehat{\pi}_i - \pi_i) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \pi_i(1-\pi_i)) \tag{4}$$

$$\widehat{p}_{ij}(K) \xrightarrow{a.s.} p_{ij}, \text{ and } \sqrt{K}\widehat{\pi}_i(\widehat{p}_{ij}(K) - p_{ij}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p_{ij}(1-p_{ij})). \tag{5}$$

Proof of Proposition 2 Since the K samples are supposed to be independent, (4) is a straightforward consequence of the strong law of large numbers and of the central limit theorem for i.i.d. sequences. Finally, (5) is proven in [2]. □

Setting $\widehat{\pi}_{(i,j)}(K) = \widehat{\pi}_i(K)\widehat{p}_{ij}(K) = n^{-1}\mathbf{N}_{n,K}(i,j)$ and replacing in (3) the probabilities by their estimators, we get the following estimator for the entropy rate,

$$\widehat{\mathbb{H}}(K) = \sum_{i=1}^{s} \widehat{\pi}_i(K) \log \widehat{\pi}_i(K) - \sum_{i=1}^{s}\sum_{j=1}^{s} \widehat{\pi}_{(i,j)}(K) \log \widehat{\pi}_{(i,j)}(K).$$

Theorem 2 *Let* $\mathbf{X}$ *be a stationary homogeneous ergodic Markov chain with a finite state space. Suppose that* $\mathbf{n}_K(i)/K$ *converges to a finite quantity, say* η_i*, for all* $i = 1, \ldots, s$*, when* K *tends to infinity. Then the estimator* $\widehat{\mathbb{H}}_n(K)$ *of* $\mathbb{H}(\mathbf{X})$ *is*

1. *strongly consistent;*
2. *asymptotically normal and unbiased when* K *tends to infinity.*

The proof follows the same lines as the proof of Theorem 1, with n replaced by K.

5 Conclusion

The above results have potential extensions in several directions. Extensions to a countable state space or to a general Borel state space can be considered. The parametric case, that is a Markov chain whose transition matrix depends continuously on a parameter with dimension less than s, would also be of interest for many applications; see for example [21] for a Bernoulli traffic source, [1] for a Markov chain with binomial transition probabilities modeling proteins interactions, or [6] for binary information source models.

References

1.Abundo, M., Accardi, L., Rosato, N. and Stella, L., Analyzing protein energy data by a stochastic model for cooperative interactions: comparison and characterization of cooperativity, *J. Math. Bio.* V44, pp341–359 (2002).

2.Anderson, T. W. and Goodman, L. A., Statistical inference about Markov Chains. *Ann. Math. Stat.* V 28, pp89–110 (1957).

3.Basharin, G.P., On a statistical estimation for the entropy of a sequence of independent random variables.*Theory Probab. Appl.* V4, pp333–36 (1959).

4.Breiman, L., The individual ergodic theorem of information theory. *Ann. Math. Stat.* V28, pp809–11 (1957) and V31, pp809–10 (1960).

5.Breiman, L., Optimal gambling system for favorable games in *Proc. 4th Berkeley Symp. Math. Stat. Prob. Berkeley, Ca: Univ. California Press* V1 pp65–78 (1960).
6.Cover, L., and Thomas, J., *Elements of information theory.* Wiley series in telecommunications, New-York (1991).
7.Dacunha-Castelle, D., and Duflo, M., *Probabilités et Statistiques. 2. Problèmes à temps mobile.* 2e édition, Masson, Paris (1994).
8.Föllmer, H., and Schied, A., *Stochastic Finance: An Introduction in Discrete Time.* Walter de Gruyter, Berlin (2002).
9.Girardin, V., Entropy maximization for Markov and semi-Markov processes. *Method. Comp. Appl. Probab.* V6, pp109–127 (2004).
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 (2005).
11.Girardin, V., and Limnios, N., *Probabilités en vue des applications*, Vuibert, Paris (2001).
12.Girardin, V. and Limnios, N., Entropy rate and maximum entropy methods for countable semi-Markov chains. *Commun. in Stat. : Theory and Methods* V33, pp609–622 (2004).
13.Gouriéroux, C., *Econométrie des variables quantitatives.* Economica, Paris (1984).
14.Kelly, J. L., A new interpretation for the information rate. *Bell Syst. Tech. J.* V35 pp917–26 (1956).
15.Li, P., and Yan, J., The growth optimal portfolio in discrete-time financial markets. *Adv. Math.* V31, pp537–42 (2002).
16.Mcmillan, M., The basic theorems of information theory. *Ann. Math. Stat.* V24, pp196–219 (1953).
17.Meeden, G., The admissibility of the maximum likelihood estimator for estimating a transition matrix. *Sankhya* V51, pp37–44 (1989).
18.Misevichyus, E. V., On the statistical estimation of the entropy of an homogeneous Markov chain. (in Russian) *Liet. Mat. Rink.* V6, pp393–95 (1966).
19.Mukhamedkhanova, R., The class of limit distributions for a statistical estimator of the entropy of Markov Chains with two states. *Soviet Math. Dokl.* V29, pp155–58 (1984).
20.Pincus, S. M., Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* V88, pp2297–301 (1994).
21.Plotkin, N. and Wyner, A., An entropy estimator algorithm and telecommunications applications. in *Maximum Entropy and Bayesian Methods.* Heidbreder, G.R. (Ed.), Kluwer Academic Publishers, pp35–50 (1996).
22.Sadek, A., and Limnios, N., Asymptotic properties for maximum likelihood estimators for reliability and failure rates of Markov chains. *Commun. Stat., Theory Methods* V31, pp1837–61 (2002).
23.Sadek, A., *Estimation des processus markoviens avec application en fiabilité.* Thèse Univ. Techn. Compiègne (2003).
24.Shannon, C., A mathematical theory of communication. *Bell Syst., Techn. J.* V27, pp379–423, 623-656 (1948).
25.van der Vaart, A.W., and Wellner, J.A., *Weak Convergence and Empirical Processes* Springer-Verlag, New-York (1996).
26.Shao, J., *Mathematical Statistics*, Sringer-Verlag, New York (2003).

27.Zubkov, A. M., Limit distribution for a statistical estimator of the entropy. *Theor. Probab. Appl.* V18, pp611–618 (1973).