Published at: 10.1103/PhysRevE.103.022215

Asymptotic distribution of sample Shannon entropy in the case of an underlying finite, regular Markov chain

Leonardo Ricci*
Dipartimento di Fisica, Università di Trento, 38123 Trento, Italy†
(Dated: February 23, 2021)

The inference of Shannon entropy out of sample histograms is known to be affected by systematic and random errors that depend on the finite size of the available data set. This dependence was mostly investigated in the multinomial case, in which states are visited in an independent fashion. In this paper the asymptotic behavior of the distribution of the sample Shannon entropy, also referred to as plug-in estimator, is investigated in the case of an underlying finite Markov process characterized by a regular stochastic matrix. As the size of the data set tends to infinity, the plug-in estimator is shown to become asymptotically normal, though in a way that substantially deviates from the known multinomial case. The asymptotic behavior of bias and variance of the plug-in estimator are expressed in terms of the spectrum of the stochastic matrix and of the related covariance matrix. Effects of initial conditions are also considered. By virtue of the formal similarity with Shannon entropy, the results are directly applicable to the evaluation of permutation entropy.

I. INTRODUCTION

Given a stationary source that generates a sequence of symbols $\{A_0, \ldots, A_{N-1}\}$ belonging to a finite alphabet, or alternatively a system whose evolution leads to visiting states that belong to a finite set, the Shannon entropy of the source or system is experimentally determined out of sample histograms $\{\hat{p}_i\}$ reporting the rate with which each symbol has been observed, or the state visited. First Basharin [1] and then Harris [2] described the bias and the fluctuation properties of the so called plug-in estimator $\hat{H} = -\sum_i \hat{p}_i \ln \hat{p}_i$ of the system's entropy H_0, where the index i runs on the M accessible—i.e. with a non-vanishing visit probability—states of the system. Those descriptions were limited to the multinomial case, namely when an independent random access to the states occurs. Main results of these works are leading terms in N^{-1}, where N is the sequence length, for both the bias affecting the plug-in estimator and its variance.

Besides the dependency on N, in the multinomial case the bias only depends on the size M and not on how the *a priori*, non-vanishing probabilities of visiting the different states are distributed. This fact can be exploited to experimentally assess an unknown size M [3]. If the probability to visit each state is uniform, the dominant term in N^{-1} of the variance vanishes and is replaced by a term proportional to N^{-2} [2]. Finite sample corrections to the plug-in estimator were first studied for Shannon entropy [4] and then generalized to Renyi entropies [5]. Most research concerns entropy bias [6–9]. Several works addressed general conditions for a central limit theorem for the plug-in entropy estimator [10–12].

A crucial step-up in sophistication consists of relaxing the assumption of independent random access to the states by considering an underlying Markov chain. In the field of information theory the quantity of interest is then typically *entropy rate* [13]. Shannon entropy H_0 is known to set an upper limit to the source's entropy rate. Entropy rate estimation out of finite sets of samples is a fundamental issue in many research fields (see [14] for a recent review) like, for example, machine learning [15]. Different kinds of estimators have been investigated: from those relying on compression algorithms [16] to estimators based on the maximum likelihood estimation of the stochastic matrix [17], nearest-neighbor entropy rate estimators [18], and straight plug-in estimators [19]. An estimator's reliability is mostly assessed in probabilistic terms, i.e. by determining, for example, how large N needs to be in order to estimate entropy rate with probability $1 - \varepsilon$, where ε is an arbitrarily small positive number [20].

The goal of the present paper is to study the asymptotic behavior of the plug-in estimator $\hat{H}$ of the Shannon entropy H_0 when a stationary, regular Markov chain, described by a stochastic matrix P, underlies the access to the states. Rather than the probabilistic one mentioned above, the approach followed here is asymptotic: the sequence length N is assumed to be long enough so that higher-order terms in $N^{-1/2}$ can be neglected. The results show how the plug-in estimator bias and variance depend on the parameters that define the Markov chain, and in particular on the spectrum of the stochastic matrix P and on the related covariance matrix χ. While the multinomial case is recovered whenever the stochastic matrix P satisfies $P^\infty = P$, in the general case the bias and the variance of the plug-in estimator are shown to deviate from the multinomial behavior, though preserving the same dependence on the sequence length N. In the asymptotic limit $N \to \infty$, the distribution of $\hat{H}$ is shown to be normal, corresponding to a central limit theorem formulation for this statistic. Initial conditions, i.e. the choice of the starting state, are shown to affect the evolution of the system by contributing on the estimator bias in terms that can be comparable with, or even larger than, the initial-condition-independent contribution.

The results of the following analysis have a wide range of possible applications, from the enduring issue of distinguishing between deterministic and stochastic sources to, for example, a recently proposed Markov modeling via ordinal partitions [21]. In addition, due to the formally identical definition, statistical properties of Shannon entropy estimators can be directly translated into those of permutation entropy [22].

* leonardo.ricci@unitn.it
† also at: CIMeC, Center for Mind/Brain Sciences, University of Trento, 38068, Rovereto, Italy.

Within this context, the results discussed in the present paper can be applied to investigate clustering [23], dispersion among trajectories characterized by the same ordinal pattern [24], time series stationarity detected via permutation entropy [25], and its use to distinguish between different underlying time series sources [26].

The paper is organized as follows. The derivation of the asymptotic distribution of the plug-in estimator of the Shannon entropy starting from the multivariate central limit theorem for Markov chains is described in Sec. II. Two subsections, Sec. II A and Sec. II B, are devoted to a detailed description of the properties of the parameters that define the asymptotic behavior of the plug-in estimator and to the effect of initial conditions. The case of a two-state system, which highlights some crucial properties of those parameters, is the topic of Sec. III.

II. ASYMPTOTIC DISTRIBUTION OF THE PLUG-IN ESTIMATOR OF THE SHANNON ENTROPY

A finite system is considered, which consists of M states and whose evolution is governed by a left, regular stochastic (or transition) matrix P, where $P_{i,j}$ corresponds to the transition probability from state j to state i. Here and henceforth both i and j run from 1 to M. Let $\mathbf{f}(0)$ be the starting stochastic vector, which represents the starting distribution of the states' occupation, and $\mathbf{f}(n)$ its evolution at step n. It holds:

$$\mathbf{f}(n) = P^n \mathbf{f}(0)\,.$$

In real-world cases and numerical simulations thereof, a starting vector $\mathbf{f}(0)$ usually represents a "pure" seed state θ, with $1 \leqslant \theta \leqslant M$, rather than a statistical mixture, so that $f_i(0) = \delta_{i,\theta}$.

Let then $\hat{k}_i$ be the number of visits of state i during an evolution that covers N steps (from 0 to $N-1$), and $\hat{p}_i$ the related rate: $\hat{p}_i = \hat{k}_i/N$. The symbol $\hat{\mathbf{p}}$ represents the vector of rates. The multivariate central limit theorem for Markov chains [27–30] states that the random vector $\hat{\mathbf{p}}$ asymptotically follows a multivariate normal distribution:

$$\hat{\mathbf{p}} \sim \mathcal{N}\left(\mathbf{s}, \frac{\chi}{N}\right) \qquad \text{as } N \to \infty\,,$$

where $\mathbf{s}$ is the stationary vector, namely the only stochastic eigenvector that pertains to the eigenvalue $\lambda_1 = 1$ of P, and χ is the covariance matrix whose elements are given by

$$\chi_{i,j} = \sum_{l=2}^{M} \frac{1+\lambda_l}{1-\lambda_l} \frac{s_j(\Delta_l)_{i,j} + s_i(\Delta_l)_{j,i}}{2}\,. \tag{1}$$

In this last expression, the $M-1$ parameters λ_l, with $2 \leqslant l \leqslant M$, are the remaining eigenvalues of P. Let $\lambda_{\max} \equiv \sup_{2 \leqslant l \leqslant M} \{|\lambda_l|\}$; then $\lambda_{\max} < 1$. Also, the $M-1$ matrices Δ_l, with $2 \leqslant l \leqslant M$, are defined through their elements as

$$(\Delta_l)_{i,j} \equiv D_{i,l}\left(D^{-1}\right)_{l,j}\,, \tag{2}$$

where the matrix D diagonalizes P:

$$\left(D^{-1} P D\right)_{i,j} = \delta_{i,j}\lambda_i\,.$$

Here $\delta_{i,j}$ is the Kronecker delta. The covariance matrix is real, symmetric, and positive semi-definite.

An alternative form of the central limit theorem for Markov chains states that the probability generating function of the *quasi-standard* random variable $\zeta \equiv (\hat{\mathbf{p}} - \mathbf{s})\sqrt{N}$ is given by

$$G_\zeta(\mathbf{t}) = \exp\left(\mathbf{t}^\dagger \chi\, \mathbf{t}\right) + O\left(\frac{1}{N^{1/2}}\right).$$

Consequently, by virtue of the Edgeworth's theorem, the distribution of $\hat{\mathbf{p}}$ differs from a purely normal one by a factor $\left[1 + O\left(N^{-1/2}\right)\right]$. From the last expression it follows

$$\begin{aligned} \mathrm{E}(\zeta_i) &= O\left(\frac{1}{N^{1/2}}\right), \\ \mathrm{E}(\zeta_i \zeta_j) &= \chi_{i,j} + O\left(\frac{1}{N^{1/2}}\right). \end{aligned} \tag{3}$$

However, for reasons that will become clear below, it is necessary to more carefully express $\mathrm{E}(\zeta_i)$ at step N.

Indeed, if N is chosen large enough so that

$$\frac{N}{\ln N} \geqslant \frac{1}{2\,|\ln \lambda_{\max}|}\,,$$

then

$$\mathrm{E}(\zeta_i) = \frac{1}{\sqrt{N}} \mathbf{r}_i^\dagger B\, \mathbf{f}(0) + O\left(\frac{1}{N}\right), \tag{4}$$

where $\mathbf{r}_i$ represents the unitary vector whose j-th component $(\mathbf{r}_i)_j$ is given by $(\mathbf{r}_i)_j = \delta_{i,j}$, and B is the matrix defined as

$$B \equiv \sum_{l=2}^{M} \frac{1}{1-\lambda_l} \Delta_l\,. \tag{5}$$

The derivation of Eq. (4) is discussed in the Appendix.

The plug-in estimator of the Shannon entropy is evaluated out of the rate histogram $\{\hat{p}_1, \dots\, \hat{p}_M\}$:

$$\hat{H}_N(\hat{p}_1, \dots\, \hat{p}_M) = -\sum_{i=1}^{M} \hat{p}_i \ln \hat{p}_i\,.$$

The quantity $\hat{H}_N$ is a sample statistic. To tackle the problem of determining its asymptotic distribution as $N \to \infty$, it is suitable to express the rates as

$$\hat{p}_i = s_i + \frac{\zeta_i}{\sqrt{N}}\,,$$

so that

$$\hat{H}_N = -\sum_{i=1}^{M} \left(s_i + \frac{\zeta_i}{\sqrt{N}}\right) \ln\left(s_i + \frac{\zeta_i}{\sqrt{N}}\right).$$

Expanding the previous expression in terms of order $N^{-n/2}$, with $n \in \mathbb{N}$, leads to

$$\hat{H}_N = H_0 - \frac{1}{\sqrt{N}} \sum_{i=1}^{M} \zeta_i \ln s_i - \frac{1}{2N} \sum_{i=1}^{M} \frac{\zeta_i^2}{s_i} + O\left(\frac{1}{N^{3/2}}\right), \tag{6}$$

where

$$H_0 = -\sum_{i=1}^{M} s_i \ln s_i$$

is the Shannon entropy of the system. In the derivation of Eq. (6) the identity $\sum_{i=1}^{M} \zeta_i = 0$, which follows from the constraint $\sum_{i=1}^{M} \hat{p}_i = 1$ and the definition of the ζ_i's, was used.

Let $G_{\delta\hat{H}}(t)$ be the probability generating function of the entropy residual $\delta\hat{H} \equiv \hat{H} - H_0$ (the dependence of $\hat{H}$ on N is henceforth omitted for the sake of clarity). Then,

$$G_{\delta\hat{H}}(t) = \mathrm{E}\left(e^{t\,\delta\hat{H}}\right).$$

The exponent $e^{t\,\delta\hat{H}}$ can be expressed as

$$e^{t\,\delta\hat{H}} = 1 - \frac{t}{\sqrt{N}} \sum_{i=1}^{M} \zeta_i \ln s_i - \frac{t}{N} \sum_{i=1}^{M} \frac{\zeta_i^2}{2\,s_i} + \frac{t^2}{2N} \sum_{i=1}^{M} \sum_{j=1}^{M} \zeta_i \zeta_j \left(\ln s_i\right)\left(\ln s_j\right) + O\left(\frac{1}{N^{3/2}}\right).$$

The expected value of each ζ_i is given by Eq. (4): the dominant contribution, namely $N^{-1/2}\mathbf{r}_i^\dagger B\,\mathbf{f}(0)$, provides in the previous expression a term of order N^{-1}, whereas higher-order corrections are absorbed within the term $O\left(N^{-3/2}\right)$. The expected value of the second and third sums are evaluated by using Eq. (3) and observing that higher-order corrections are again absorbed within the term $O\left(N^{-3/2}\right)$. It follows:

$$G_{\delta\hat{H}}(t) = 1 - \frac{t}{N} \sum_{i=1}^{M} \left[\left(\ln s_i\right)\mathbf{r}_i^\dagger B\,\mathbf{f}(0) + \frac{\chi_{i,i}}{2\,s_i}\right] + \frac{t^2}{2N} \sum_{i=1}^{M} \sum_{j=1}^{M} \chi_{i,j} \left(\ln s_i\right)\left(\ln s_j\right) + O\left(\frac{1}{N^{3/2}}\right). \quad (7)$$

It is worth defining $\mathbf{h}$ as the vector whose i-th component is equal to the single-bin entropy $-\ln s_i$. In addition let

$$\beta \equiv \sum_{i=1}^{M} \frac{\chi_{i,i}}{2\,s_i}, \quad (8a)$$

$$\Lambda \equiv \mathbf{h}^\dagger \chi \mathbf{h}. \quad (8b)$$

Because χ is positive semi-definite, one has $\Lambda \geqslant 0$. The coefficient of t in Eq. (7) corresponds to the expected value of $\delta\hat{H}$, which can be expressed by using the definitions of $\mathbf{h}, \beta$ as follows:

$$\mu_{\delta\hat{H}} = -\frac{1}{N}\left[\beta - \mathbf{h}^\dagger B\,\mathbf{f}(0)\right] + O\left(\frac{1}{N^{3/2}}\right). \quad (9)$$

Because the square of this last term is of order N^{-2}, the coefficient of $t^2/2$ in Eq. (7) directly provides the dominant term of the variance of $\delta\hat{H}$, and thus of $\hat{H}$:

$$\sigma^2_{\delta\hat{H}} = \frac{\Lambda}{N} + O\left(\frac{1}{N^{3/2}}\right), \quad (10)$$

where the variable Λ defined in Eq. (8b) was used.

Let the *quasi-standard* random variable $\delta\hat{h}$ be defined as

$$\delta\hat{h} \equiv \left(\delta\hat{H} + \frac{\beta - \mathbf{h}^\dagger B\,\mathbf{f}(0)}{N}\right)\sqrt{N}.$$

From Eq. (7) it is straightforward to show that the related probability generating function is

$$G_{\delta\hat{h}}(t) = \exp\left(\frac{t^2}{2}\Lambda\right) + O\left(\frac{1}{N^{1/2}}\right).$$

This last expression shows that, provided that Λ is non-vanishing, $\delta\hat{h}$ is asymptotically normally distributed with zero mean and variance Λ. Consequently, and provided $\Lambda > 0$, it is possible to formulate a central limit theorem for the Shannon entropy of a finite system whose evolution is governed by a regular stochastic matrix as follows:

$$\hat{H} \sim \mathcal{N}\left(H_0 - \frac{\beta - \mathbf{h}^\dagger B\,\mathbf{f}(0)}{N}, \frac{\Lambda}{N}\right) \qquad \text{as } N \to \infty.$$

Remarkably, while the variance is asymptotically unaffected by initial conditions, the mean is biased by a contribution that depends on the starting stochastic vector $\mathbf{f}(0)$.

A. Properties of the parameters β, Λ

From Eq. (1) and noting that $\mathrm{Tr}\,\Delta_l = 1$, $\forall l$, it is straightforward to show that the coefficient β, defined in Eq. (8a), is given by

$$\beta = \frac{1}{2}\sum_{l=2}^{M} \frac{1+\lambda_l}{1-\lambda_l} = \frac{M-1}{2} + \sum_{l=2}^{M} \frac{\lambda_l}{1-\lambda_l}. \quad (11)$$

In the multinomial case, the probability of visiting state i does not depend on the previous state j, so that all columns of P are equal. It follows that, first, P is idempotent, i.e. $P^2 = P$ and thus $P^n = P$, $\forall n \in \mathbb{N} \setminus \{0\}$, and, second, $P\mathbf{v} = \mathbf{s}$ for any stochastic vector $\mathbf{v}$. It is then straightforward to show that each eigenvalue λ_l with $2 \leqslant l \leqslant M$ turns out to be equal to zero [27, 30]. Consequently, according to Eqs. (8a), (9), (11), the bias term that is independent of initial conditions reduces to the well-known Miller-Madow bias correction [1–3, 31] equal to $-(M-1)/(2N)$. Interestingly, this bias term only depends on the size M of the set of nonzero probabilities $\{s_i\}$, whereas it is independent of their values. Furthermore, according to Eqs. (5), (A4), the term $\mathbf{h}^\dagger B\,\mathbf{f}(0)/N$, which provides the bias contribution that depends on the initial conditions, is equal to $\left[\mathbf{h}^\dagger\mathbf{f}(0) - H_0\right]/N$. This term is better explained by considering the expected value of the entropy plug-in estimator $\hat{H}$ as a whole: in the multinomial case and omitting higher-order corrections $O\left(N^{-3/2}\right)$, one gets

$$\mu_{\hat{H}} \cong \frac{(N-1)H_0 + \mathbf{h}^\dagger\mathbf{f}(0)}{N} - \frac{M-1}{2N}.$$

Besides the Miller-Madow bias correction, there is an entropy contribution due to the starting stochastic vector of the evolution (typically a pure state θ yielding $\mathbf{h}^\dagger \mathbf{f}(0) = -\ln s_\theta$) and having a weight $1/N$, and a contribution by the random part of the sequence, which consists of $N-1$ elements and thus has a weight $(N-1)/N$.

In the multinomial case it is known that $\chi_{i,j} = s_i\delta_{i,j} - s_i s_j$. On the other hand, setting $\lambda_l = 0$ with $2 \leqslant l \leqslant M$ into Eq. (1) yields $\chi_{i,j} = \sum_{l=2}^{M} \left[s_j (\Delta_l)_{i,j} + s_i (\Delta_l)_{j,i} \right]/2$. It is then straightforward to rewrite Eq. (1) in the general case as

$$\chi_{i,j} = s_i\delta_{i,j} - s_i s_j + \sum_{l=2}^{M} \frac{\lambda_l}{1-\lambda_l} \left[s_j (\Delta_l)_{i,j} + s_i (\Delta_l)_{j,i} \right] . \quad (12)$$

Consequently, with regard to Λ one has

$$\Lambda = \Lambda_0 + \sum_{l=2}^{M} \frac{2\,\lambda_l}{1-\lambda_l} \left[\sum_{i=1}^{M} \sum_{j=1}^{M} s_i \left(\ln s_i\right) \left(\ln s_j\right) (\Delta_l)_{j,i} \right] , \quad (13)$$

where the parameter Λ_0 is a sort of *population variance* of the single-bin entropy $h_i = -\ln(s_i)$:

$$\Lambda_0 \equiv \sum_{i=1}^{M} s_i \ln^2 s_i - \left(\sum_{i=1}^{M} s_i \ln s_i \right)^2 .$$

For each $l \geqslant 2$, it is worth defining ξ_l as

$$\xi_l \equiv \sum_{i=1}^{M} \sum_{j=1}^{M} s_i \left(\ln s_i\right) \left(\ln s_j\right) (\Delta_l)_{j,i} . \quad (14)$$

From the expression $\chi_{i,j} = \sum_{l=2}^{M} \left[s_j (\Delta_l)_{i,j} + s_i (\Delta_l)_{j,i} \right]/2$, valid in the multinomial case, it follows that

$$\Lambda_0 = \sum_{l=2}^{M} \xi_l .$$

The expression for Λ given by Eq. (13) can be therefore rewritten as

$$\Lambda = \Lambda_0 + \sum_{l=2}^{M} \frac{2\,\lambda_l}{1-\lambda_l} \xi_l = \sum_{l=2}^{M} \frac{1+\lambda_l}{1-\lambda_l} \xi_l . \quad (15)$$

In the multinomial case, according to Eqs. (8b), (10), (15) the variance of the plug-in entropy estimator is given by Λ_0/N, a result first obtained by Basharin [1] and Harris [2].

A final interesting case is provided when the stationary vector $\mathbf{s}$ is uniform, i.e. $s_i = 1/M$, $\forall i$, a situation that can occur independently of the multinomial case. Because the covariance matrix χ has vanishing column and row sums [30], from Eq. (8b) it follows $\Lambda = 0$. In this case the dominant term of the variance turns out to be $(M-1)/(2N^2)$ (see Harris [2]).

Incidentally, while the term Λ_0 can be straightforwardly proven to vanish if the stationary vector $\mathbf{s}$ is uniform, this last condition also implies $\xi_l = 0$, $\forall l \,|\, 2 \leqslant l \leqslant M$. This fact follows from a uniform $\mathbf{s}$ being a left eigenvector of P with unitary eigenvalue: $\mathbf{s}^\dagger P = \mathbf{s}^\dagger$. Because $P = D\Lambda D^{-1}$, it follows $\mathbf{s}^\dagger D\Lambda = \mathbf{s}^\dagger D$. Therefore, $\mathbf{s}^\dagger D$ is a left eigenvector of the diagonal matrix Λ with unitary eigenvalue. Because the only unitary diagonal element of Λ is $\Lambda_{1,1}$, it holds $\mathbf{s}^\dagger D\mathbf{r}_l = 0$, $\forall l \,|\, 2 \leqslant l \leqslant M$. By using this last expression along with the definitions of Δ_l, ξ_l, respectively given by Eqs. (2), (14), it follows $\xi_l = 0$, $\forall l \,|\, 2 \leqslant l \leqslant M$ in the case of a uniform $\mathbf{s}$.

B. Effect of initial conditions on bias

As shown above, the bias contains a term $\mathbf{h}^\dagger B\,\mathbf{f}(0)/N$ that is dependent on the initial conditions, i.e. on the starting stochastic vector $\mathbf{f}(0)$. If $\mathbf{f}(0)$ would be given by the stationary vector $\mathbf{s}$, the term would identically vanish, as explained in the last paragraph of the Appendix. On the other hand, as mentioned at the beginning of Sec. II, in numerical simulations of real-world cases a starting vector usually represents a pure state θ, with $1 \leqslant \theta \leqslant M$: $f_i(0) = \delta_{i,\theta}$.

It is worth considering a set of evaluations of the plug-in estimator $\hat{H}$ carried out on an "experimental" sample $\mathbb{S}$ of ω numerically-simulated evolutions of equal length N and whose pure seed states are distributed according to a multinomial distribution $\{\phi_\theta\}$, with $1 \leqslant \theta \leqslant M$: ϕ_θ corresponds to the occurrence probability of the pure seed state θ in the set $\mathbb{S}$. Let $\langle \hat{H} \rangle$, $s^2_{\hat{H}}$ be the sample mean and the sample variance, respectively, of the entropy plug-in estimator $\hat{H}$ evaluated on the set $\mathbb{S}$. By neglecting higher-order terms, the expected value of the sample mean $\langle \hat{H} \rangle$ is given by

$$\mathrm{E}\left(\langle \hat{H} \rangle \right) \cong H_0 - \frac{\beta - \mathbf{h}^\dagger B\,\phi}{N} , \quad (16)$$

where ϕ is the vector whose i-th component is given by ϕ_i. Clearly, if $\phi = \mathbf{s}$, the average contribution proportional to $\mathbf{h}^\dagger B\,\phi$ will vanish.

With regard to the variance, besides Λ/N, which is unaffected by initial conditions, the starting state variability yields a contribution of order N^{-2} and that can be therefore neglected for sufficiently large values of N. By also neglecting the higher-order terms of Eq. (10), the expected value of the sample variance $s^2_{\hat{H}}$ is then given by

$$\mathrm{E}\left(s^2_{\hat{H}} \right) \cong \frac{\Lambda}{N} . \quad (17)$$

III. TWO-STATE SYSTEM

In this section, a two-state system is discussed as a special case of the results of the previous section. Let

$$P = \begin{pmatrix} 1-a & b \\ a & 1-b \end{pmatrix}$$

be a generic regular stochastic matrix of dimension $M = 2$. Regularity requires $0 < a < 1$, $0 < b < 1$. It is straightforward to show that the eigenvalue λ_2, the stationary vector $\mathbf{s}$, and the

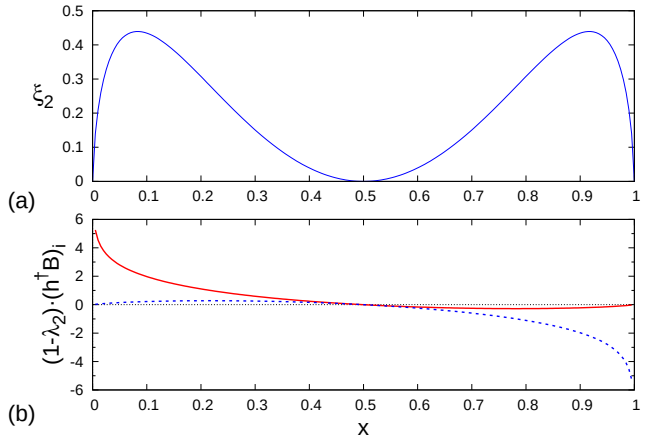

FIG. 1. (a) Parameter ξ_2, defined in Eq. (14), as a function of the first component of the stationary vector $\mathbf{s}$, $x = b/(a+b)$. (b) Components of the vector $\left(\mathbf{h}^\dagger B\right)_i$, multiplied by the factor $1-\lambda_2$, again as a function of the first component of the stationary vector $\mathbf{s}$, $x = b/(a+b)$: red, solid line, $i = 1$; blue, dotted line, $i = 2$. The former plot tends to $+\infty$ as $x \to 0$, the latter to $-\infty$ as $x \to 1$.

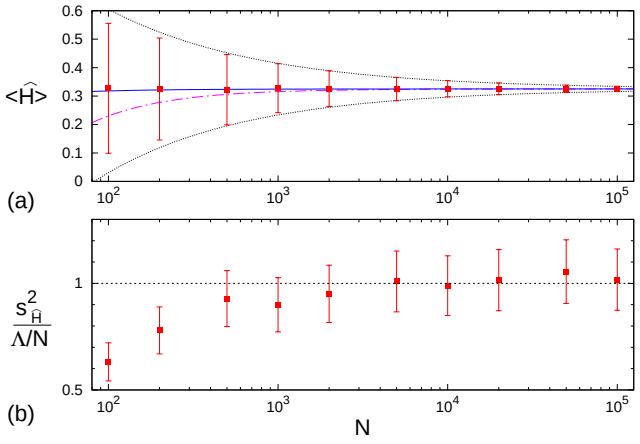

FIG. 2. Entropy plug-in estimator $\hat{H}$ as a function of the number of steps N for the Markov chain described by the stochastic matrix of Eq. (18). (a) Each red dot and the related errorbar correspond to the sample mean and the sample standard deviation, respectively, of $\hat{H}$ evaluated on a set $\mathbb{S}$ of $\omega = 1000$ simulated evolutions of N steps. The starting state of each simulation is randomly chosen with equal probability between state 1 and 2, i.e. $\phi = \left(\frac{1}{2}, \frac{1}{2}\right)$. The magenta, dashed line and the blue, solid line correspond to Eq. (16) where the term proportional to $\mathbf{h}^\dagger B\,\phi$ is neglected and considered, respectively. The two black, dotted lines correspond to the expected sample mean value plus/minus the standard deviation given by Eq. (17). (b) Each red dot correspond to the sample variance, normalized by the expected value given by Eq. (17), of the evaluations described for the plot (a). The uncertainty of each sample variance value s^2 is evaluated as $s^2\,(2/\omega)^{1/2}$.

matrices P^∞, D, Δ_2 are given by

$$\lambda_2 = 1 - a - b \quad \Rightarrow \quad |\lambda_2| < 1\,,$$
$$\mathbf{s} = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad P^\infty = \begin{pmatrix} x & x \\ y & y \end{pmatrix},$$
$$D = \begin{pmatrix} x & -1 \\ y & 1 \end{pmatrix}, \qquad \Delta_2 = \begin{pmatrix} y & -x \\ -y & x \end{pmatrix},$$

where $x = b/(a+b)$, $y = a/(a+b) = 1 - x$.

It follows

$$\chi = \frac{1+\lambda_2}{1-\lambda_2}\, x\,(1-x) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

so that

$$\beta = \frac{1}{2}\,\frac{1+\lambda_2}{1-\lambda_2}\,,$$
$$\Lambda = \frac{1+\lambda_2}{1-\lambda_2}\left[x\,(1-x)\,\ln^2\left(\frac{1-x}{x}\right)\right].$$

In addition,

$$\mathbf{h}^\dagger B = \frac{1}{1-\lambda_2}\,\ln\left(\frac{1-x}{x}\right)\left(1-x\,,\quad x\right).$$

Because $\beta = \frac{1}{(1-\lambda_2)} - \frac{1}{2}$ and λ_2 can vary between (and excluding) -1 and 1, it follows that β is an increasing function of λ_2 that can take on any positive value.

The parameter Λ can be expressed as $\Lambda = 2\beta\xi_2$, where ξ_2 defined in Eq. (14) corresponds to the x-dependent term in square brackets in the expression of Λ. As shown in Fig. 1(a), the function ξ_2 reaches the maximum value $\xi_{2,\max} \cong 0.439$ when either x or $y = 1 - x$ are equal to $x_0 \cong 0.0832$, whereas it vanishes—as expected from the discussion at the end of Sec. II A—in the uniform case $x = y = 1/2$.

The multinomial (binomial) case occurs when $\lambda_2 = 0$, i.e. when $a + b = 1$. Consequently, $\beta = 1/2$, which corresponds to the term providing the Miller-Madow bias correction, while $\Lambda = \Lambda_0$ can be trimmed between 0 and $\xi_{2,\max}$ by acting on x, i.e. on b.

Finally, the two components of the vector $\mathbf{h}^\dagger B$, which provide, via Eq. (9), the respective initial condition biases, are shown in Fig. 1(b) upon multiplication by the factor $1 - \lambda_2$. The first component, $(\mathbf{h}^\dagger B)_1$, decreases from $+\infty$ at $x = 0$ down to $\cong -0.278/(1-\lambda_2)$ at $x \cong 0.782$, to thereupon monotonically increase up to zero at $x = 1$. The behavior of $(\mathbf{h}^\dagger B)_2$ follows from that of $(\mathbf{h}^\dagger B)_1$ by noting that $(\mathbf{h}^\dagger B)_2$ in x is equal to $-(\mathbf{h}^\dagger B)_1$ in $1 - x$. The two components are simultaneously zero at $x = \frac{1}{2}$.

As an example, the Markov chain corresponding to the stochastic matrix

$$P = \begin{pmatrix} 0.91 & 0.01 \\ 0.09 & 0.99 \end{pmatrix}, \tag{18}$$

is analyzed. In this case, $\lambda_2 = 0.9$, $x = 0.1$. Figure 2 shows the sample mean of the plug-in estimator and the related sample standard deviation evaluated on sets of $\omega = 1000$ simulations of the Markov chain evolution, where the pure seed states are randomly chosen with equal probability. Starting from $N \simeq 500$, the results of the numerical simulations are perfectly described by the theory developed above.

It is worth noting that the parameters that describe the asymptotic behavior of the plug-in estimator have—in the case of β—or can have—in the case of Λ and $\mathbf{h}B$—a divergent behavior as $\lambda_2 \to 1$. This result can be generalized to regular Markov chains of any size M: if $\lambda_{\max}$ approaches 1, namely if the *absolute spectral gap* [19] of the Markov chain, defined as $\gamma^*(P) \equiv 1 - \lambda_{\max}$, tends to zero, the bias parameter β becomes arbitrarily large. A similar situation typically occurs when there are at least two states, or two subsets of states, in which the system tends to remain during its evolution, i.e with a relatively small probability to escape. This behavior is consistent with the relaxation time of a Markov chain being defined as the reciprocal of $\gamma^*(P)$ [19]. Besides the mentioned condition on the eigenvalues, relatively large values of Λ and of the components of $\mathbf{h}B$ require non-uniform distributions of the components of the stationary vector.

Appendix A: Expected values of the number of visits

By using the matrix D that diagonalizes P and the vectors $\mathbf{r}_i$'s (all these quantities are defined in Sec. II), and noting that $\lambda_1 = 1$, one can write:

$$P^n = \sum_{l=1}^{M} \lambda_l^n D\mathbf{r}_l\mathbf{r}_l^\dagger D^{-1} = D\mathbf{r}_1\mathbf{r}_1^\dagger D^{-1} + \sum_{l=2}^{M} \lambda_l^n \Delta_l \,, \tag{A1}$$

where the definition of the matrices Δ_l given by Eq. (2) was used. Because $|\lambda_l| < 1\,,\ \forall l\,|\,2 \leqslant l \leqslant M$, it follows:

$$P^\infty = D\mathbf{r}_1\mathbf{r}_1^\dagger D^{-1}\,,$$

so that Eq. (A1) can be rewritten as

$$P^n = P^\infty + \sum_{l=2}^{M} \lambda_l^n \Delta_l \,. \tag{A2}$$

From this last expression it follows that, starting from the stochastic vector $\mathbf{f}(0)$ at step 0, the resulting stochastic vector $\mathbf{f}(n)$ at step n is given by:

$$\mathbf{f}(n) = P^n\mathbf{f}(0) = P^\infty\mathbf{f}(0) + \sum_{l=2}^{M} \lambda_l^n \Delta_l\mathbf{f}(0) = \mathbf{s} + \sum_{l=2}^{M} \lambda_l^n \Delta_l\mathbf{f}(0)\,. \tag{A3}$$

As a corollary, setting $n = 0$ in Eq. (A2) yields

$$\sum_{l=2}^{M} \Delta_l = \mathbb{1} - P^\infty \,. \tag{A4}$$

Let $X_i(n)$ be the dichotomic random variable that describes the occupation of the i-th state at the n-th step: $X_i(n)$ only takes on the values 0 or 1. The expected value of $X_i(n)$ is given by

$$\mathrm{E}\left[X_i(n)\right] = \mathrm{Prob}\left[X_i(n) = 1\right] = f_i(n)\,.$$

By using Eq. (A3), one has:

$$\mathrm{E}\left[X_i(n)\right] = s_i + \sum_{l=2}^{M} \lambda_l^n\ \mathbf{r}_i^\dagger \Delta_l \mathbf{f}(0)\,.$$

The expected values of the number of times $\hat{k}_i$ that, prior to the N-th step, the i-th state was visited during the system's evolution is therefore:

$$\mathrm{E}\left(\hat{k}_i\right) = \sum_{n=0}^{N-1} \mathrm{E}\left[X_i(n)\right] = Ns_i + \sum_{l=2}^{M} \frac{1-\lambda_l^N}{1-\lambda_l}\ \mathbf{r}_i^\dagger \Delta_l \mathbf{f}(0)\,.$$

By using the definition of ζ, namely $\zeta_i \equiv \left(\hat{k}_i/N - s_i\right)\sqrt{N}$, one gets:

$$\begin{aligned}\mathrm{E}(\zeta_i) &= \frac{1}{\sqrt{N}} \sum_{l=2}^{M} \frac{1-\lambda_l^N}{1-\lambda_l}\ \mathbf{r}_i^\dagger \Delta_l \mathbf{f}(0) \\ &= \frac{1}{\sqrt{N}} \left[\mathbf{r}_i^\dagger B\,\mathbf{f}(0) - \sum_{l=2}^{M} \frac{\lambda_l^N}{1-\lambda_l}\ \mathbf{r}_i^\dagger \Delta_l \mathbf{f}(0)\right], \end{aligned} \tag{A5}$$

where the matrix B is defined in Eq. (5). By using the triangle inequality, the absolute value of the sum within the square brackets in the last expression satisfies

$$\left|\sum_{l=2}^{M} \frac{\lambda_l^N}{1-\lambda_l}\ \mathbf{r}_i^\dagger \Delta_l \mathbf{f}(0)\right| \leqslant \lambda_{\max}^N \sum_{l=2}^{M} \left|\frac{1}{1-\lambda_l}\mathbf{r}_i^\dagger \Delta_l \mathbf{f}(0)\right| .$$

Because $\lambda_{\max} < 1$, for any $\alpha > 0$ it is possible to set a sufficiently large value N_0 so that $N \geqslant N_0$ implies $\lambda_{\max}^N \leqslant N^{-\alpha/2}$, or equivalently

$$\frac{N}{\ln N} \geqslant \frac{\alpha}{2\,|\ln \lambda_{\max}|}\,.$$

It follows that, if $N \geqslant N_0$, Eq. (A5) can be expressed as

$$\mathrm{E}(\zeta_i) = \frac{1}{\sqrt{N}}\mathbf{r}_i^\dagger B\,\mathbf{f}(0) + O\left(\frac{1}{N^{(\alpha+1)/2}}\right),$$

which, in the case $\alpha = 1$, corresponds to Eq. (4).

As a final remark, setting $\mathbf{f}(0) = \mathbf{s}$ in Eq. (A3) leads to

$$\sum_{l=2}^{M} \lambda_l^n \Delta_l \mathbf{s} = 0 \qquad \forall n \geqslant 0\,.$$

Consequently, if $\mathbf{f}(0) = \mathbf{s}$, each expected value $\mathrm{E}(\zeta_i)$ is equal to zero at any N. In addition, $B\,\mathbf{s} = 0$.

ACKNOWLEDGMENTS

The author wishes to acknowledge A. Politi for fruitful discussions, A. Perinelli for critical reading of the manuscript, and the anonymous referees for valuable suggestions.

[1] G. P. Basharin, Theor. Probability Appl. **4**, 333 (1959).
[2] B. Harris, "The statistical estimation of entropy in the non-parametric case," in *Topics in Information Theory* (North-Holland, Amsterdam, 1975) pp. 323–355.
[3] L. Paninski, Neural Computation **15**, 1191 (2003).
[4] H. Herzel, Sys. Anal. Mod. Sim. **5**, 435 (1988).
[5] P. Grassberger, Physics Letters A **128**, 369 (1988).
[6] M. S. Roulston, Physica D: Nonlinear Phenomena **125**, 285 (1999).
[7] T. Schürmann, J. Phys. A: Math. Gen. **37**, L295 (2004).
[8] J. A. Bonachela, H. Hinrichsen, and M. A. Muñoz, Journal of Physics A: Mathematical and Theoretical **41**, 202001 (2008).
[9] A. M. T. Ramos, H. L. Casagrande, and E. E. N. Macau, Physica A **549**, 124301 (2020).
[10] A. Antos and I. Kontoyiannis, Random Structures and Algorithm **19**, 163 (2001).
[11] Z. Zhang and X. Zhang, IEEE Transactions on Information Theory **58**, 2745 (2012).
[12] C. Chen, M. Grabchak, A. Stewart, J. Zhang, and Z. Zhang, Entropy **20**, 371 (2018).
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition* (John Wiley & Sons, 2006).
[14] S. Verdú, Entropy **21**, 720 (2019).
[15] Y. Han, J. Jiao, C.-Z. Lee, T. Weissman, Y. Wu, and T. Yu, in *Proc. Conf. Neural Inf. Process. Syst.* (2018) pp. 9803–9814.
[16] P. Grassberger, IEEE Transactions on Information Theory **35**, 669 (1989).
[17] G. Ciuperca and V. Girardin, "On the estimation of the entropy rate of finite markov chains," in *Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis* (ENST Bretagne, 2005) pp. 1109–1117.
[18] A. Kaltchenko and N. Timofeeva, Advances in Mathematics of Communications **2**, 1 (2008).
[19] S. Kamath and S. Verdú, in *2016 IEEE International Symposium on Information Theory (ISIT)* (2016) pp. 685–689.
[20] G. Valiant and P. Valiant, J. ACM **64**, 37 (2017).
[21] K. Sakellariou, T. Stemler, and M. Small, Phys. Rev. E **100**, 062307 (2019).
[22] C. Bandt and B. Pompe, Phys. Rev. Lett. **88**, 174102 (2002).
[23] D. J. Little and D. M. Kane, Phys. Rev. E **95**, 052126 (2017).
[24] A. Politi, Phys. Rev. Lett. **118**, 144101 (2017).
[25] Y. Cao, W. w. Tung, J. B. Gao, V. A. Protopopescu, and L. M. Hively, Phys. Rev. E **70**, 046217 (2004).
[26] O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, Phys. Rev. Lett. **99**, 154102 (2007).
[27] C. M. Grinstead and J. L. Snell, *Introduction to Probability* (AMS, 2003).
[28] C. J. Geyer, "Introduction to Markov Chain Monte Carlo," in *Handbook of Markov Chain Monte Carlo* (CRC Press, 2011).
[29] E. Omey, J. Santos, and S. Van Gulck, Applicable Analysis and Discrete Mathematics **2**, 38 (2008).
[30] L. Ricci, Chaos Sol. Fractals **142**, 110450 (2021).
[31] G. Miller, "Note on the bias of information estimates," in *Information theory in Psychology II-B* (Free Press, Glencoe, IL, 1955) pp. 95–100.

Published at: 10.1103/PhysRevE.103.022215