Entropy estimation of symbolic sequences: How short is a short sequence?

Annick LESNE, Jean-luc BLANC and Laurent PEZARD

Institut des Hautes Études Scientifiques
35, route de Chartres
91440 – Bures-sur-Yvette (France)

Décembre 2008

IHES/P/08/63

Entropy estimation of symbolic sequences: How short is a short sequence?

Annick Lesne,[1,*] Jean-Luc Blanc,[2] and Laurent Pezard[2]

[1] *Institut des Hautes Études Scientifiques*

Le Bois-Marie, 35 route de Chartres, F-91440, Bures-sur-Yvette, France.

[2] *Laboratoire de Neurosciences Intégratives et Adaptatives*

UMR 6149 CNRS Aix-Marseille Université,

3 Place Victor Hugo, F-13331 Marseille Cedex 03, France.

(Dated: December 19, 2008)

Abstract

While entropy per unit time is a meaningful index to quantify the dynamic features of experimental time series, its estimation is often hampered by the finite length of the data. We here investigate the performance of entropy estimation procedures, relying either on block entropies or Lempel-Ziv complexity, when only *very short symbolic sequences* are available. Heuristic analytical arguments point at the influence of temporal correlations on the bias and statistical fluctuations, and put forward a reduced effective sequence length suitable for error estimation. Numerical studies are conducted using, as benchmarks, the wealth of different dynamic regimes generated by the family of logistic maps and stochastic evolutions generated by a Markov chain of tunable correlation time. Practical guidelines and validity criteria are proposed, based on the result that the quality of entropy estimation is sensitive to the sequence temporal correlation hence self-consistently depends on the entropy value itself.

PACS numbers: 05.45.Tp, 05.45.-a

Keywords: Time-series analysis, symbolic sequences, Shannon entropy, block-entropy, metric entropy, Lempel-Ziv complexity, finite-size effects, correlations.

*Permanent address: LPTMC, Université Pierre et Marie Curie-Paris 6, 4 place Jussieu, F-75252 Paris Cedex 05, France.

I. INTRODUCTION

Investigating non-linear dynamic features of a system from a single experimental time series is an ubiquitous issue that gave rise to a very rich methodology [1]. An important characteristic is entropy (per unit time) which has been defined in a coherent manner in dynamical system theory, stochastic dynamics and information theory [2, 3]. It quantifies global temporal organization of time series and provides a meaningful statistics in surrogate data tests for discriminating linear and various non-linear dynamic models [4]. We focus here on entropy estimation for a symbolic sequence originating either from the intrinsic discreteness of the system states (e.g. linguistic data [5], DNA sequences [6, 7]), from an adequate partition of the phase space [8, 9] or from an adapted encoding (e.g. behavioral sequences [10, 11], speech analysis [12, 13], spike emission in neurons [14, 15]). We address the specific sub-question of controlling entropy estimation for *very short* time series (for which undersampling is likely to be critical) and, for a given dynamics, what is the minimal sequence size for which entropy analysis gives significant results. We shall henceforth consider only very short symbolic sequences of length $N \leq 1000$. Actually, such a limitation on the data length is often encountered in biological, medical or social data, mainly due to the necessary restriction to a time window in which the system evolution can be considered as stationary.

The paper is organized as follows. In the next section, we recall the definition of entropy (per unit time, denoted h throughout the paper) in terms of the limiting behavior of block entropies, and we present the associated estimation procedures. We then investigate the bias and statistical fluctuations in entropy estimation, focusing on *non* asymptotic results (i.e. for very short symbolic sequences) and *without* the usual assumption of an independent and identically distributed (i.i.d.) data sample. We thus propose to consider an effective sequence length $N_{\rm eff}$ (instead of N) to account for sample correlation in error computation. Next we briefly present the Lempel-Ziv compression algorithms and their relation to entropy, yielding an alternative estimation procedure for h. Then numerical investigation of finite-size effects for the different methods of entropy estimation is conducted using as benchmarks short symbolic sequences of known dynamics, either deterministic (logistic maps) or stochastic (Markov chains). We evidence precisely how correlation time of the sequence controls statistical errors. Our aim being to give practical guidelines on the relevant procedures that can be implemented to extract discriminating information on the temporal organization of

very short symbolic sequences, we focus our conclusions on the self-consistent dependence of the estimation quality on the entropy value itself.

II. BLOCK-ENTROPIES AND ASSOCIATED ENTROPY ESTIMATORS

Considering a stationary source emitting at each time step a symbol from a finite alphabet of size k, its *block-entropy of order* n is defined as the Shannon entropy of the probability distribution $p_n(w)$ of the n-words [41]:

$$H_n \equiv -\sum_w p_n(w) \ln p_n(w) \tag{1}$$

where the sum runs over all the possible n-words w, hence depends on the dynamics of the source over time intervals of n steps. $n \mapsto H_n$ is monotonous non decreasing and concave. Accordingly, $h_n = H_{n+1} - H_n$ decreases towards a limit h as the word length n tends to ∞, that defines the *entropy rate* of the source [42]:

$$h = \lim_{n\to\infty} H_{n+1} - H_n = \lim_{n\to\infty} \frac{H_n}{n} \tag{2}$$

The ratio $h_{n,av} = H_n/n$ also decreases, from which follows that $H_n \geq nh$ and $h_{n,av} \geq h_n \geq h$.

For an i.i.d. sequence, h coincides with H_1. Other simple situations where the entropy rate can be determined analytically are periodic regimes and Markov sources; explicitly, $h_n = h$ as soon as $n \geq q$ for a Markov chain of order q, and $h_n = 0$ hence $h = 0$ and $h_{n,av} \sim 1/n$ for a periodic sequence of period $T \leq n$.

The influence of the symbol distribution on the entropy is fully encapsulated in H_1, with $H_1 \leq \ln k$ (where the equality holds for uniform distribution only). For an i.i.d. sequence, the knowledge of the symbol distribution p_1 thus gives a full account of the source properties and $h = H_1$. When correlations are present and p_n is not a product of n replicas of the symbol distribution p_1, we have $h_{n,av} < h_{1,av}$ and $h_n < h_1$. Consequently, correlations are encapsulated in the way h differs from H_1. But the respective contributions of symbol distribution and temporal structures are not additive and $(H_1 - h)/h$ cannot be used as a simple measure of the correlation strength. In fact, the entropy h accounts for the *whole statistical dependencies* in an integrated way [43] (i.e. more thoroughly than correlation functions that are restricted to linear correlations). Although, a general qualitative principle states that h decreases if correlations increase and $1/h$ has the meaning of a correlation time, the relation between entropy and sequence correlations is not straightforward.

In practice, h should most often be estimated from a *single* observed sequence $[s] \equiv (s_i)_{1\leq i\leq N}$ of length N. We shall henceforth denote $\hat{X}$ the estimator of a quantity X, without mentioning explicitly that it depends on the sequence $[s]$ and its length N. Any 'hatted' quantity is thus a random variable, relative to a finite-length sequence and reflecting its length-dependent distribution.

Several variants have been proposed for estimating Shannon entropy (block entropies in the present context), like bias-corrected or jackknifed estimators [16] but they all rely on the i.i.d. nature of the data. This assumption typically fails when investigating the entropy rate of a correlated times series. So, we limit ourselves to the maximum-likelihood estimator $\hat{H}_n$ of H_n directly following from the definition (1) using the maximum-likelihood estimator $\hat{p}_n(w)$ of the n-word distribution, i.e. the frequency of w. Then an estimator $\hat{h}$ of h is provided by one of the following possibilities:
(i) the difference $\hat{h}_n = \hat{H}_{n+1} - \hat{H}_n$, for n large enough.
(ii) the average $\hat{h}_{av,n} = \hat{H}_n/n$ for n large enough;
(iii) the slope $\hat{h}_{slope}$ of the linear region on the graph $n \mapsto \hat{H}_n$, before it saturates at a value $\hat{H}_{max} \sim \ln N$ due to finite sampling. Since slope determination is known to be computationally unstable, we will not consider this third method to get an estimation of h but we shall exploit it to appreciate the strength of the finite-size effects (see figure 1).

Note that all three ways provide a *non parametric* estimation of h, insofar as no parameterized expression of $p_n(w)$ is required, nor any assumption about the source other than stationarity and ergodicity.

The maximum-likelihood estimation of $p_n(w)$ needs the extraction of n-words from the sequence $[s]$ from which one can obtain $N - n + 1$ overlapping n-words. It should be underlined that this word sequence $[w] = (w_i)_{1\leq i\leq N-n+1}$ forms a *correlated sample* of n-words, not only because of the overlap between the successive words, but also because of inherent statistical dependencies within the symbol sequences $[s]$ (precisely those entropy h accounts for). Heuristic criteria based on the statistical meaning of entropy h can be derived to account for the influence of these correlations.

Given the sequence length N, there is an upper bound on the word length n that it is possible to investigate with a sufficient statistical quality. It is currently derived by considering that there are only N/n independent n-words in the sequence which leads to the constraint $N \geq nk^n$ [17]. But this estimation is valid for i.i.d. sequences only, where

the correlations between n-words follow only from their overlapping. In the general case, non-trivial correlations between n-words originating from correlations within the sequence should be taken into account. Using the interpretation of h as the average independent information (in $\ln k$ units) brought by the observation of an additional symbol, the original sequence length should be replaced by an effective length

$$N_{\text{eff}} \approx \frac{Nh}{\ln k} \tag{3}$$

to account for intrinsic correlations. The criterion bounding the word length thus writes more stringently

$$Nh \geq n\, k^n\, \ln k \tag{4}$$

In fact, the dependence of this constraint, and more generally of the entropy estimation quality, with respect to the number k of symbols might be not so strong. Indeed, for n large enough, Shannon-McMillan-Breiman theorem [18] states that the number of n-words of non-negligible probability that actually contribute to the entropy is not k^n but e^{nh} and the criterion reduces to

$$Nh \gg n\, e^{nh}\, \ln k \tag{5}$$

Even though the theorem cannot be strictly applied to short sequences due to their bounded acceptable word length, its contents hints at a dependence of the quality of the estimation with respect to the entropy value h in a non trivial way since e^{nh}/h is minimal in $h = 1/n$. We thus expect better estimation for moderate h than for high h, and strong difficulties to assess very small or vanishing values of h.

A similar criterion follows from a recurrence argument. It has been established that the minimal recurrence time T_n at the level of n-words (that is, the smallest time t such that $(x_0, x_1, \ldots, x_{n-1} = x_{-t}, x_{-t+1}, \ldots x_{-t+n-1})$) behaves as e^{nh}, namely $\lim_{n\to\infty}(1/n)\ln T_n \to h$ in probability [19]. On qualitative grounds, it yields a similar criterion $Nh \gg ne^{nh}\ln k$ to ensure that it is not improbable to observe a typical n-word a sufficient number of times to get a good statistics for estimating $\hat{p}_n$ (at least for typical words). Here again, the effective size of the sample of n-words is N_{eff}/n in order to account for time correlations.

These clues will be the guideline of our quality assessment of entropy estimation and systematic numerical study.

III. BIAS AND FLUCTUATIONS IN FINITE-SIZE BLOCK-ENTROPY ESTIMATORS

Previous studies focused either on the convergence ($N \to \infty$) of the estimators towards the exact entropy and the asymptotic error estimation [20, 21, 22, 23], or on the scaling behavior of entropy and error estimators [17, 24, 25, 26]. None of these two classes of results are relevant in experimental studies where only *very short* sequences are available. In this case, the asymptotic regime is presumably out of reach and the range of accessible lengths (word length n or sequence length N) too narrow to validate any scaling behavior. Asymptotic error estimators might be not only meaningless but possibly misleading, e.g. subtracting an asymptotic estimation of the bias might actually not lead to an improvement in entropy estimation. Moreover, error bars and finite-size effects on the estimation have been established within restricted dynamic models, most often i.i.d. random variables sequences invalid for correlated time series.

Our aim is rather to provide a *model-free* quality assessment of entropy estimation. We shall thus analyze the discrepancy between H_n and the estimated value $\hat{H}_n$ as a function of the sequence length N and the block-size n using its decomposition into a *bias* (deterministic contribution in the error) and *fluctuations* (of vanishing average). Two complementary sampling situations should be considered:
(i) the case of *good statistics*, where typical n-words are adequately sampled and their probability $p_n(w)$ properly estimated by their frequency of occurrence $\hat{p}_n(w)$. It corresponds to the condition $\sup_w Np_n(w) \gg 1$ or in practice $\sup_w N\hat{p}_n(w) \gg 1$.
(ii) the case of *bad statistics*, where a word occurs at most a few times and frequencies of occurrence are meaningless. It corresponds to $Np_n(w) \leq \mathcal{O}(1)$ for all n-words.

Since we are interested in the influence of time correlation on the entropy estimation, we shall use the integrated correlation time as an indicator of the statistical dependencies. The integrated correlation time $\tau_{int}(w)$ of the process $i \to \delta_{ww_i}$ is defined as follows [3, 27, 28]:

$$\tau_{int}(w) = \frac{1}{2} \sum_{-\infty}^{+\infty} \frac{C_w(t)}{C_w(0)} \tag{6}$$

where $C_w(t) = \langle \delta_{ww_i} \delta_{ww_{i+t}} \rangle - \langle \delta_{ww_i} \rangle \langle \delta_{ww_{i+t}} \rangle$ is the correlation function of the n-word w within the sequences generated by the source [29] with δ_{ww_i} being the Kronecker symbol, such that $\delta_{ww_i} = 1$ if $w = w_i$ else 0.

A. Good statistics

In the case of good statistics, entropy estimation simply parallels the definition of H_n, setting the estimator $\hat{H}_n$ equal to the block-entropy of the maximum-likelihood estimator of the n-word probability distribution:

$$\hat{p}_n(w) = \frac{1}{N-n+1} \sum_{i=1}^{N-n+1} \delta_{ww_i} \tag{7}$$

Under the assumption of statistical stationarity of the process $i \to s_i$, hence of the process $i \to w_i$ for any word-length n, the random variables δ_{ww_i} (for w given and w_i random) are identically distributed, of mean $\langle \delta_{ww_i} \rangle = p_n(w)$, so that the estimator $\hat{p}_n(w)$ is unbiased. As mentioned above, these variables δ_{ww_i} are *correlated*, all the more since the source is itself correlated. The law of large numbers:

$$\lim_{N\to\infty} \hat{p}_n(w) = p_n(w) \qquad \text{almost surely} \tag{8}$$

nevertheless applies provided the correlations decrease rapidly enough at infinity i.e. $\tau_{int}(w) < \infty$ [3, 27, 28].

We introduce the error $\delta\hat{p}_n(w) = \hat{p}_n(w) - p_n(w)$ which is centered since $\hat{p}_n(w)$ is unbiased. Under the assumption $\tau_{int}(w) < \infty$, a generalized central limit theorem [44] applies and the error can be characterized by its variance

$$\langle \delta\hat{p}_n(w)^2 \rangle = \frac{2p_n(w)[1-p_n(w)]\tau_{int}(w)}{N} \tag{9}$$

In computing this error $\delta\hat{p}_n(w)$, we identified $p_n(w)$ and $\tau_{int}(w)$ with their estimators, since the difference yields an higher-order contribution.

The discrepancy $\Delta\hat{H}_n$ between estimated and real values of the block entropy is currently decomposed into a statistical error corresponding to the fluctuation $\delta\hat{H}_n \equiv \hat{H}_n - \langle \hat{H}_n \rangle$, and the systematic error corresponding to the bias $b_{\hat{H}_n} \equiv \langle \hat{H}_n \rangle - H_n$:

$$\Delta\hat{H}_n \equiv \hat{H}_n - H_n = \delta\hat{H}_n + b_{\hat{H}_n} \tag{10}$$

Although $\hat{p}_n(w)$ is unbiased, $\hat{H}_n$ depends in a nonlinear way on $\hat{p}_n(w)$ and $\langle \hat{H}_n \rangle \neq H_n$. Explicit expression of these finite-size corrections are obtained by expanding $\hat{H}_n$ with respect to $\delta\hat{p}_n(w)$:

$$\Delta\hat{H}_n = -\left(\sum_w [1+\ln p_n(w)]\delta\hat{p}_n(w) + \frac{1}{2p_n(w)}\,\delta\hat{p}_n(w)^2 + \mathcal{O}(\delta\hat{p}_n(w)^3)\right) \tag{11}$$

where the sum runs over the set of all n-words present in the sequence, i.e. having a non-vanishing probability. The validity of the expansion only requires that $|\delta\hat{p}_n(w)| \ll p_n(w)$, namely that $2\tau_{int}(w) \ll Np_n(w)$ for all typical words if we consider statements in the sense of L_2 convergence, i.e. statements about the moments of $\Delta\hat{H}_n$.

Taking the average of the above expansion yields the expression for the bias $b_{\hat{H}_n} \equiv \langle\Delta\hat{H}_n\rangle$ [45]:

$$b_{\hat{H}_n} = -\sum_w \frac{1}{2p_n(w)}\,\langle\delta\hat{p}_n(w)^2\rangle + h.o. \tag{12}$$

$$\approx -\sum_w \frac{2\tau_{int}(w)[1-p_n(w)]}{N} \tag{13}$$

Due to the above-mentioned restriction on the range of the sum, it comes at lower order:

$$b_{\hat{H}_n} \approx -\frac{2\tau_n M_n}{N} \tag{14}$$

where M_n is the number of n-words of non-vanishing probability and $\tau_n = \langle\tau_{int}\rangle_n$ the average of the correlation time. It amounts to replace the term N/n in the bias estimators given in the literature [16, 20, 21, 22, 23, 24, 25] by an effective number

$$\frac{N_{\text{eff}}}{n} = \frac{N}{2\tau_n} \tag{15}$$

so as to account for the contribution of the correlations between the n-words. The definition (3) of N_{eff} is thus supported by the extension of central limit theorem to the case of correlated sequences. We recover $N_{\text{eff}} = N$ for an i.i.d. sequence and N_{eff} is all the smaller that the range of correlations is larger. For a strongly correlated sequence, the very definition of h as a compression rate yields: $2\tau_n \sim n\ln k/h$, and $N_{\text{eff}} \sim Nh/\ln k$, recovering equation (3), becoming dramatically small in case of long-range correlations, for $h \ll 1$. The statistical error on the word probability distribution finally writes

$$\frac{\delta\hat{p}_n(w)}{\hat{p}_n(w)} \approx \sqrt{\frac{n}{N_{\text{eff}}p_n(w)}} \tag{16}$$

Accordingly, the criterion for good statistics is more stringent than that derived on the basis of the current error estimators for independent samples i.e. $\sup_w Np_n(w) \gg 1$. It writes

rather $\sup_w N_{\rm eff} p_n(w) \gg n$, especially for long-range correlated sequences where $h \ll \ln k$ hence $N_{\rm eff} \ll N$. This yields an upper bound $n^*(N,h)$ on the word-length n, such that $n \ll n^*(N,h)$ corresponds to a situation of good statistics, where

$$n^*(N,h) \sim \begin{cases} \frac{\ln N}{h} & \text{if } h = \mathcal{O}(1) \\ \\ \frac{Nh}{\ln k} & \text{if } h \to 0 \end{cases} \tag{17}$$

The statistical error on H_n is directly related to the variance of $\Delta\hat{H}_n$, which can be estimated as follows (noticing that $\sum_w \delta\hat{p}_n(w) = 0$ since both p_n and $\hat{p}_n$ are normalized to 1):

$$\begin{aligned} \langle (\delta\hat{H}_n)^2 \rangle &= \langle (\Delta\hat{H}_n)^2 \rangle - b^2_{\hat{H}_n} = \mathrm{Var}(\Delta\hat{H}_n) \\ &= \left\langle \left[\sum_w \ln p_n(w) \delta\hat{p}_n(w) \right]^2 \right\rangle + \text{h.o.} \end{aligned} \tag{18}$$

where again correlations between the sampled n-words control the amplitude of $\delta\hat{p}_n(w)$. More explicitly, let us compute for n large enough the leading-order terms using the Shannon-McMillan-Breiman theorem: either the n-word w is one of the $M_n \sim e^{nh}$ typical n-words, and $p_n(w) \approx e^{-nh}$, else $p_n(w) \approx 0$. Hence, from (14) it comes:

$$b_{\hat{H}_n} \approx -\frac{n e^{hn}}{N_{\rm eff}} \tag{19}$$

and plugging (9) in (18) yields

$$\langle (\delta\hat{H}_n)^2 \rangle^{1/2} \approx n\ h\ e^{nh/2}\ \sqrt{n/N_{\rm eff}} \tag{20}$$

Although the bias scales as $1/N$ and the fluctuation as $1/\sqrt{N}$, in agreement with current wisdom, the prefactors coming from correlations cannot be ignored for practical purposes: considering N instead of $N_{\rm eff}$ would drastically underestimate the errors.

B. Bad statistics

In the case of very short sequences, the regime of good statistics allowing to estimate n-word probabilities will not be valid for n larger than a few units, and entropy estimation has to be done in an undersampling situation [30]. It is no longer justified to perform

an expansion of $\hat{H}_n$ around $p_n(w)$ in powers of $\delta\hat{p}_n(w)$, and another procedure has to be implemented.

Denoting $\hat{K}_N(w) = \sum_{i=1}^{N-n+1} \delta_{ww_i}$ the number of occurrences of the word w in the considered sequence (it also implicitly depends on the length n of the word w), the dominant contributions to $\hat{H}_n$ comes from $\hat{K}_N(w) = 1$ or $\hat{K}_N(w) = 2$, a larger number of occurrences being highly improbable. Nevertheless, $\hat{K}_N(w) = 2$ or $\hat{K}_N(w) = 1$ is simply a matter of chance, reflecting a finite-size fluctuation and not the value of $p_n(w)$. This means that the word count does not approximate the probability distribution $p_n(w)$. The relevant expansion should now be performed around $\hat{K}_N(w) = 0$ [21, 24]:

$$\langle \hat{H}_n \rangle = -\sum_K \sum_w \frac{K}{N} \left(\ln \frac{K}{N} \right) \text{Prob}(\hat{K}_N(w) = K) \tag{21}$$

At the leading order, neglecting the multiple occurrences of some words, it simply remains

$$\hat{H}_n = \ln N \tag{22}$$

which accounts for the saturation of the curve $n \to \hat{H}_n$ predicted above heuristically. The correlation between n-words in the sample are usually not taken into account, based on the argument that the probability of joint occurrence is very weak [25]. Indeed, sticking to a first-order expansion, involving only $\text{Prob}(\hat{K}_N(w) = 1)$, allows to neglect this issue. But as soon as one tries to estimate more precisely $\hat{H}_n$ and takes into account the next terms with $K \geq 2$, then it is necessary to consider these correlations, that might notably affect the probabilities $\text{Prob}(\hat{K}_N(w) = K)$.

A straightforward estimation of the crossover location is currently obtained by matching the linear part $\hat{H}_n \approx hn$ and the saturation value[46] $\hat{H}_n \approx \ln N$ due to undersampling at large n; or a given sequence length N, this yields $n^*(N, h) \sim \ln N/h$. This result becomes paradoxical when $h \to 0$ since it would indicate that large n-values could be faithfully considered as $h \to 0$, the larger the closer h is to 0, whereas we expect an opposite behavior for the estimation quality, as discussed above in Section II for correlated sequences. The paradox is solved if we use the refined criterion $N_{\text{eff}} \gg ne^{nh}$ for good statistics stated in equation (5): at very low value of h, it yields a crossover value $n^*(N, h)$
$simNh/\ln k$ that consistently decreases with h. Although the criterion (5) is a rough approximation for small n since the Shannon-McMillan-Breiman theorem is an asymptotic result, it yet shows how the value of h influences the very procedure of its estimation, here the upper bound $n^*(N, h)$ above which drastic finite-size effects (bad statistics) arise.

This analytical study demonstrates that estimating entropy is a self-consistent problem, since the convergence rate and error bars depend on the estimated value of h insofar as it reflects the time correlations of the source.

IV. LEMPEL-ZIV COMPLEXITY

The viewpoint adopted in computing Lempel-Ziv complexity is a priori far different from that associated with Shannon entropy rate h. Indeed, the definition of Shannon entropy rate h involves a global feature of the dynamics, namely its invariant measure. It can be computed from the knowledge of a single trajectory insofar as the measure is ergodic and allows the reconstruction of the probability distribution of the source from the observation of a single typical sequence. But it is not in its own right meaningful as a feature of a single sequence. By contrast, Lempel-Ziv complexity provides a measure of the compressibility of the considered single symbolic sequence, in other words the information contents per symbol. Under the assumption that the source is stationary and ergodic, Lempel-Ziv theorems [31] ensure that Lempel-Ziv complexity coincides with h up to a factor $\ln k$ involving the number k of symbols in the alphabet. This assumption indeed implies that almost all symbolic sequences have the same compressibility features, hence the computation can be equivalently performed with any typical sequence [47] and its result coincides with the average.

According to the Lempel-Ziv scheme, the sequence of length N is parsed into $\mathcal{N}_w$ words. Two different parsings have been proposed, either "LZ77" [32]:

1 • 0 • 01 • 10 • 11 • 100 • 101 • 00 • 010 • 11...

where the parsing considered as a new word the shortest one that has not yet been encountered, or "LZ76" [33]:

1 • 0 • 01 • 101 • 1100 • 1010 • 001011 • ...

where the parsing considers as a new word any subsequence that has not yet been encountered (the fourth word in the above example is thus 101 and not the 2-sequence 10 since the latter has already been seen). One then computes

$$\hat{L} = \frac{\mathcal{N}_w[1 + \log_k \mathcal{N}_w]}{N} \quad \text{with} \quad \lim_{N\to\infty} \hat{L} = \frac{h}{\ln k} \tag{23}$$

An alternative and simpler computation involves

$$\hat{L}_0 = \frac{\mathcal{N}_w \ln N}{N} \quad \text{with} \quad \lim_{N\to\infty} \hat{L}_0 = h \tag{24}$$

Replacing $\log_k N$ by $\ln N$ makes the limit directly comparable to h, whereas the original definition is normalized with a common upper bound equal to 1. We respectively specify $\mathcal{N}_w^{(76)}$, $\hat{L}^{(76)}$ or $\mathcal{N}_w^{(77)}$, $\hat{L}^{(77)}$ according to the chosen parsing. [48]

Computation of error bars on Lempel-Ziv complexity does not follow from a standard limit theorem. There is indeed no analytical way to check the internal consistency of the estimation and its accuracy. The only internal test of validity is to check the convergence of $\hat{L}_0$ as a function of N, ensuring that the limiting behavior is reached. For a stationary ergodic source, both finite-size estimators decrease to their limit [18]. A numerical fit $\hat{L}_0 \sim h + (a \log_k N)/N^\gamma$ for large N has been proposed in [17]. Other asymptotic estimators are given and shown in [18, 19]. A simple expression of the standard deviation has been proposed in [34]

$$\hat{\sigma} = (\hat{L}_0)^{3/2} \frac{s}{\sqrt{N \log_k N}} \tag{25}$$

where s is the standard deviation of the word length in the parsing (according to [33]). Nevertheless, this computation relies on the questionable assumption that the words in the parsing are i.i.d. according to a Gaussian distribution $N(\Lambda, s^2)$ where $\Lambda = N/\mathcal{N}_w$ is the average length of these words; its relevance has been yet supported by numerical simulations only in simple cases and for N large enough ($N > 10^5$).

We rather investigate short ($N \leq 10^3$) and correlated sequences where such asymptotic error estimation and assumptions on word distribution are irrelevant; we precisely focus on the influence of time correlations on estimation quality, in particular the relative performance of the two parsings for very short and correlated sequences.

V. NUMERICAL INVESTIGATIONS

In order to investigate the relative performance of the different entropy estimators on short symbolic sequences as a function of their size N and correlation time τ_n, we performed a panel of numerical tests. We used the family of logistic maps and a family of Markov chains with tunable correlation time as benchmarks. Logistic maps provide a paradigmatic example of a deterministic evolution in a continuous phase space where the trajectories are turned into symbolic sequences using a generating partition, while Markov chains exemplify a stochastic evolution between discrete states. The entropy h is equally well-defined and known (either numerically or analytically) in both cases. The flexibility of such numerical

models with controlled entropy allows to investigate the relation between the value of h and the quality of its estimation in very different dynamic regimes.

A. Dynamical models

1. Logistic maps

We first used logistic maps $x_{n+1} = ax_n(1 - x_n)$, where $x_n \in [0, 1]$ and $a \in [3.5, 4]$, taking benefit of the almost exhaustive knowledge available about this one-parameter dynamics. Several different dynamic regimes are encountered as the control parameter a varies [35], ranging from periodic to fully random through critical and chaotic. The entropy $h(a)$ can be computed exactly as the Lyapunov exponent (on a sufficiently long run) according to the Pesin equality [36]. We coded the sequences using the available generating partition $[0, 1/2[\cup[1/2, 1]$. By varying a, we are able to tune the amount of correlations in the source and the entropy of the generated sequences. We shall therefore use the intrinsic entropy rate to characterize the dynamics as regards its time organization and correlations.

2. Markov chain with tunable correlation time

We also considered binary sequences of length N generated by a Markov chain, with transition matrix

$$R(a, b) = \begin{pmatrix} 1 - a & b \\ a & 1 - b \end{pmatrix} \tag{26}$$

(with notation $R_{i \leftarrow j}$) having eigenvalues 1 and $1 - a - b$, hence a characteristic time

$$\tau(a, b) = \frac{1}{-\ln|1 - a - b|} \tag{27}$$

that can be shown to coincide with the correlation time of the evolution. The stationary distribution writes

$$p_{eq} = \begin{pmatrix} \frac{b}{a+b} \\ \frac{a}{a+b} \end{pmatrix} \tag{28}$$

and the entropy:

$$h(a, b) = -\frac{b^2}{a + b} \ln b - \frac{b(1 - a)}{a + b} \ln(1 - a) - \frac{a^2}{a + b} \ln a - \frac{a(1 - b)}{a + b} \ln(1 - b) \tag{29}$$

For simplicity, we shall present the results obtained in the case where $b = a \leq 1/2$, which corresponds to a one-parameter family of Markov chain, with transition matrix

$$R(a) = \begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix} \tag{30}$$

with entropy:

$$h_1(a) = -a \ln a - (1-a) \ln(1-a) \tag{31}$$

and correlation time:

$$\tau(a) = \frac{1}{-\ln(1-2a)} \tag{32}$$

B. Simulations and results

1. *Convergence and saturation of block-entropy estimator*

We first considered the behavior of the block-entropy maximum-likelihood estimator $\hat{H}_n$ as a function of the word length n in three typical dynamical regimes: chaotic, critical and periodic, see figure 1, and investigated the crossover between good and bad statistics. For the chaotic regime, where $h = \mathcal{O}(1)$, the vertical dashed line indicates the location of the theoretical crossover $n^*(N, h) = \ln N/h$. At the onset of chaos, it has been proved for logistic map ($a = a_c$) [20, 24] that $H_n(a_c) = \ln(3n/2)$ for n equal to a power of 2. Even with short sequences, the sublinear increases of $n \to H_n$ can nevertheless be detected, i.e. $n \to H_n$ markedly departs from a straight line $n \to hn$. It also markedly departs from the behavior observed for a periodic sequences (although $h = 0$ in both cases) since as soon as n is larger than the period, one has $H_n = const.$

2. *Quality assessment of entropy estimation*

Quality assessment of entropy estimation was quantified using the difference $\hat{h} - h$ between the estimated value of entropy and the true one considering different entropy estimators: blocks-entropy estimators ($\hat{h}_n = \hat{H}_{n+1} - \hat{H}_n$ and $\hat{h}_{av} = \hat{H}_n/n$ for $n = 5$) and Lempel-Ziv complexity estimators ($\hat{L}^{(76)}$ and $\hat{L}^{(77)}$). In every numerical simulation, a normalized version of Lempel-Ziv complexity estimator $\hat{L}^{(7.)}$ was used namely $\hat{L}^{(7.)} = \hat{L} \ln(2)/\max[\hat{L}]$ where $\max[\hat{L}]$ is the maximum value of complexity obtained for random i.i.d. binary sequences

[37, 38]; the factor ln(2) appears to facilitate the comparison between Lempel-Ziv complexity and entropy h. These estimators and conventions were used for all the results presented here.

Self-consistency of entropy estimation was investigated using the dependence of the entropy estimation quality as a function of entropy value h (figure 2) or correlation time τ_n (figure 3). As expected, a visible decreasing trend appears on figure 2 where the quality of the entropy estimation is better for high entropy dynamics than for low ones. It agrees with the heuristic argument based on the effective length $N_{\text{eff}} = Nh/\ln k$ which predicts larger errors for low h. As a general observation, all the entropy estimators overestimate h with the noticeable exception of $\hat{L}^{(76)}$ which tends to underestimate high values of entropy. Nevertheless, $\hat{L}^{(76)}$ outperforms all other estimators at small h. Thus, contrary to the current claim that $\hat{L}^{(77)}$ is more efficient than $\hat{L}^{(76)}$ for correlated sequences [26], we observe that $\hat{L}^{(76)}$ gives far better estimation than $\hat{L}^{(77)}$ at small h. Our results show that $\hat{h}_{n,av}$ and $\hat{L}^{(77)}$ always overestimate h although they are efficient for weakly correlated sequences, of high entropy. Nevertheless, their performance are far too low compared to that of the two other estimators ($\hat{h}_n$ and $\hat{L}^{(76)}$) and $\hat{h}_{n,av}$ and $\hat{L}^{(77)}$ should thus be rejected in entropy estimation. $\hat{h}_n$ gives equally (but moderately) good results, so it could be specially fruitful in case of an automated computation a priori covering low-h and high-h situations; one of its virtues is to always provide an overestimation of h. Actually, for high values of h, $\hat{h}_n$ appears to be the best estimator, while $\hat{L}^{(76)}$ outperforms the other estimators for correlated sequences i.e. for low values of h.

These results show that our claim on the influence of temporal correlations on the convergence of entropy estimators, hence the necessity of a self-consistent quality assessment, are valid for both deterministic and stochastic evolutions. It is thus essential to have a first guess about h or τ_n before choosing the estimation procedure.

Convergence with sequence length was investigated considering the convergence of the different entropy estimators: $\hat{h}_n$, $\hat{h}_{av}$,$\hat{L}^{(76)}$, $\hat{L}^{(77)}$ described above (see figure 4). In the case of random dynamics, the performances of all the estimators are comparable for short sequence length ($N > 500$). In the critical deterministic case, the constancy of $\hat{h}_n$ and $\hat{h}_{n,av}$ (at a non vanishing value) reflects the dramatically slow convergence of these estimators in case of strong correlations and $\hat{L}^{(76)}$ appears to be the best estimator. It provides valuable estimation of h even for moderate sequence length ($N > 1200$). In the stochastic critical

case, the high variability depicted in the estimation with $\hat{L}^{(77)}$ and $\hat{h}_{av}$ reflects dramatically the finite-size effects for long-range correlations. These results thus underline the dual aspect of finite-size effects in block-entropy approach for estimating h: (i) *a convergence issue* of H_n and $h_{n,av}$ towards their limit h (constraining the word-length n) and (ii) *a statistical issue* in the reconstruction of the word probability distribution from a single sequence (constraining the sequence length N). In Lempel-Ziv approach, there is a priori no statistical issue since Lempel-Ziv complexity is relative to a single sequence; nevertheless, statistics somehow reappears in the sequence dependence (non uniformity) of the convergence to h as $N \to \infty$.

VI. CONCLUSION

Our analysis has been based on two assumptions: ergodicity and stationarity of the dynamics. Ergodicity assumption underlying estimator definitions means that the entropy estimation deals with the invariant measure sampled in the observation, i.e. the measure of interest. In this respect, it is not a limitation and does not affect the consistency of the results. By contrast, stationarity assumption is a strict requirement, that precisely leads to consider restricted time windows and to characterize the dynamics from the analysis of (very) short sequences.

We have seen that the validity and the optimality of the different (model-free) ways of estimating the source entropy h, as well as their quality assessment, depend crucially on the very value of h. Since time correlation of the source, as quantified by h itself, influences dramatically the bias and statistical fluctuations, error estimation should not be done assuming i.i.d. sequence. As regards the error bars, all happens as if the sequence were of effective length $N_{\text{eff}} = Nh/\ln k$. This self-consistency of entropy estimation and its quality hints at using a two-step method, where a preliminary rough and quick estimation of h indicates what method should be implemented to get the most accurate and most faithful estimation of h, with a possible trade-off between accuracy and assessment of definite bounds. In most of the cases, algorithmic estimation using Lempel-Ziv complexity performs this valuable compromise for very short sequences.

[1] H. Kantz and T. Schreiber, *Nonlinear time series analysis* (Cambridge University Press, Cambridge, 1997).

[2] R. Badii and A. Politi, *Complexity. Hierarchical structures and scaling in physics* (Cambridge University Press, Cambridge, 1999).

[3] P. Castiglione, M. Falcioni, A. Lesne, and A. Vulpiani, *Chaos and coarse graining in statistical mechanics* (Cambridge University Press, 2008).

[4] J. Theiler and D. Prichard, Physica D **94**, 221 (1996).

[5] W. Ebeling and T. Pöschel, Europhysics Letters **26**, 241 (1994).

[6] H. Herzel, W. Ebeling, and A. Schmitt, Physical Review E **50**, 5061 (1994).

[7] C. Peng, S. Buldyrev, A. Goldberger, S. Havlin, M. Simons, and S. H.E., Physical Review E **47**, 3730 (1993).

[8] E. Bollt, T. Stanford, Y. Lai, and K. Zyczkowski, Physical Review Letters **85**, 3524 (2000).

[9] E. Bollt, T. Stanford, Y. Lai, and K. Zyczkowski, Physica D **154**, 259 (2001).

[10] M. Paulus, M. Geyer, L. Gold, and A. Mandell, Proceedings of the National Academy of Sciences USA **87**, 723 (1990).

[11] P. Faure, H. Neumeister, D. Faber, and H. Korn, Fractals **11**, 233 (2003).

[12] K. Doba, L. Pezard, A. Lesne, V. Christophe, and J. Nandrino, Psychological Reports **101**, 237 (2007).

[13] K. Doba, J. Nandrino, A. Lesne, J. Vignau, and L. Pezard, New Ideas in Psychology **26**, 295 (2008).

[14] S. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek, Physical Review Letters **80**, 197 (1998).

[15] J. Amigo, J. Szczepanski, E. Wajnryb, and M. Sanchez-Vives, Neural Computation **16**, 717 (2004).

[16] L. Paninski, Neural Computation **15**, 1191 (2003).

[17] T. Schürmann and P. Grassberger, Chaos **6**, 414 (1996).

[18] T. Cover and J. Thomas, *Elements of information theory* (Wiley, New York, 1991).

[19] A. Wyner and J. Ziv, IEEE Transactions on Information Theory **35**, 1250 (1989).

[20] W. Ebeling and G. Nicolis, Chaos, Solitons and Fractals **2**, 635 (1992).

[21] H. Herzel, A. Schmitt, and W. Ebeling, Chaos, Solitons and Fractals **4**, 97 (1994).

[22] H. Herzel and I. Grosse, Physica A **216**, 518 (1995).

[23] S. Panzeri and A. Treves, Network: Computation in Neural Systems **7**, 87 (1996).

[24] P. Grassberger, Physics Letters A **128**, 369 (1988).

[25] M. Roulston, Physica D **125**, 285 (1998).

[26] T. Schürmann, J. Phys. A: Math. Gen. **35**, 1589 (2002).

[27] G. Lawler and A. Sokal, Transactions of the American Mathematical Society **309**, 557 (1988).

[28] A. Sokal and L. Thomas, Journal of Statistical Physics **54**, 797 (1989).

[29] W. Li, Journal of Statistical Physics **60**, 823 (1990).

[30] L. Paninski, IEEE Transactions on Information Theory **50**, 2200 (2004).

[31] J. Ziv and A. Lempel, IEEE Transactions in Information Theory **24**, 530 (1978).

[32] J. Ziv and A. Lempel, IEEE Transactions on Information Theory **23**, 337 (1977).

[33] A. Lempel and J. Ziv, IEEE Transactions on Information Theory **22**, 75 (1976).

[34] J. Amigo and M. Kennel, Chaos **16**, 043102 (2006).

[35] R. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, and H. Scheingraber, Chaos, Solitons and Fractals **4**, 133 (1994).

[36] P. Collet and J. Eckmann, *Iterated maps of the interval as dynamical systems* (Birkhäuser, Basel, 1981).

[37] P. Rapp, C. Cellucci, K. Korslund, T. Watanabe, and J.-M. no, Physical Review E **64**, 016209 (2001).

[38] M. Aboy, R. Hornero, D. Abásolo, and Álvarez D., IEEE Trans. Biomed. Eng. **53**, 2282 (2006).

[39] G. Miller and W. Madow, Tech. Rep. AFCRC-TR-54-75, Air Force Cambridge Research Center (1954).

[40] A. Wyner and J. Ziv, IEEE Transactions on Information Theory **37**, 878 (1991).

[41] We here consider natural logarithm, at odds with Shannon definition where a binary logarithm $\log_2$ is used, but in agreement with the definition currently used in dynamical systems theory.

[42] If $\lim_{n\to\infty} H_n - H_{n-1} = h$ exists, then $\lim_{n\to\infty} H_n/n$ exists and takes the same value h. The converse is not true. We shall here consider situations where the two limits exist, hence coincide.

[43] The expression $h = \lim_{n\to\infty} H(s_{n+1}|s_1, \ldots, s_n)$ shows that all possible temporal correlations are taken into account.

[44] The central limit theorem extends to the case of correlated sequences of random variables (here n-words) provided the variance of the sum is renormalized by a quantity proportional to the integrated correlation time [3].

[45] In contrast to [23, 25, 39], we shall not assume that the errors $\delta\hat{p}_n(w)$ are statistically independent.

[46] Note that since the saturation value $\hat{H}_n = \ln N$ originates from undersampling at large n and is not related to the temporal structure of the sequence, surrogate sequences obtained by randomly shuffling the data would exhibit the same saturation value $\ln N$ but with a different crossover value $n^* = \ln N/H_1$

[47] Note that the set of typical sequences have a full measure, hence sequences drawn at random or observed experimentally are typical; only sequences generated from a specially chosen non generic initial condition might happen to be non typical.

[48] Several variants and improvements of the original Lempel-Ziv algorithms have been developed, see for instance [40].

Figures

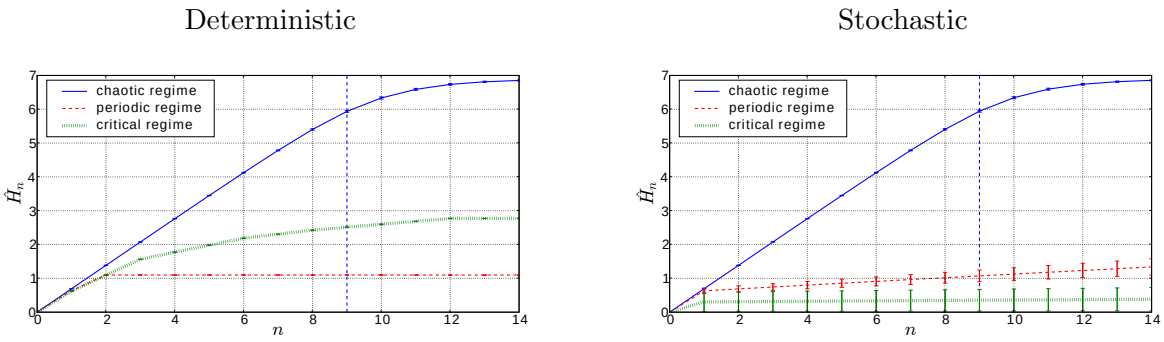

FIG. 1: Behavior of the block-entropy estimator $\hat{H}_n$ when increasing the word length n in the deterministic case (left) and the stochastic case (right) for binary sequences of length $N = 1000$. In the deterministic case, sequences were obtained using the logistic map with different parameter values: $a = a_c \approx 3.569$ ($h = 0$) for the critical behavior with long-range correlations, $a = 3.83$ for the periodic regime ($h = 0$) and $a = 4$ for the fully chaotic regime (where $H_n = n \ln 2$ and $h = \ln 2$). In the stochastic case, fully chaotic behavior was obtained using $a = 1/2$, periodic dynamics using $a = 0.1$ and critical behavior using $a = 10^{-3} \ll 1$. The vertical dashed line indicates the theoretical location $n^* = \ln N/h$ of the crossover between good and bad statistics when $h > 0$ (chaotic regime) above which $\hat{H}_n$ saturates to a value $\ln N$.

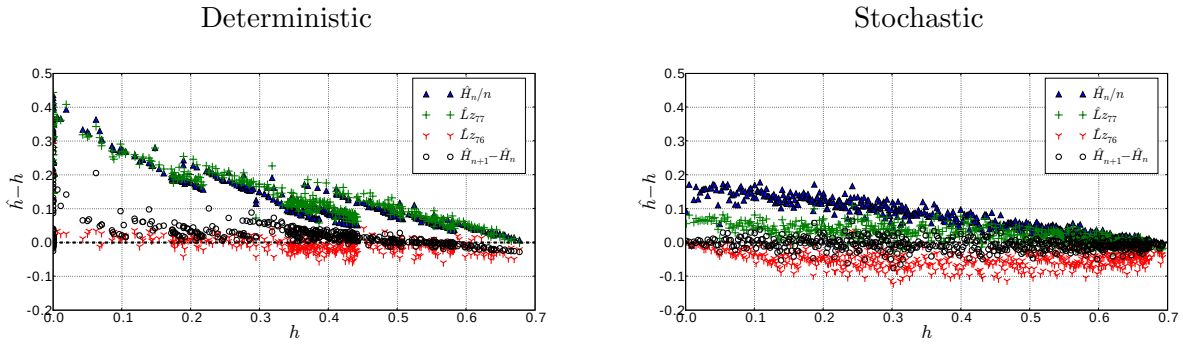

FIG. 2: Entropy estimation quality $\hat{h} - h$ as a function of the true entropy value h in the case of a deterministic evolution (left) and in the stochastic case (right). In the deterministic case, sequences are generated from a logistic map with control parameter $a \in [3.5, 4]$. In the stochastic case, sequences are generated from a Markov chain with a transition matrix parameter $a \in [0, 1/2]$. The exact value h of the entropy is computed as the Lyapounov exponent from a very long typical trajectory in the logistic case; it is known analytically in the Markov case. The estimators are calculated for 100 symbolic sequences of length $N = 1000$ generated with random initial conditions for 500 values of the control parameter a. Values represented on the figure correspond to the mean value over these 100 sequences.

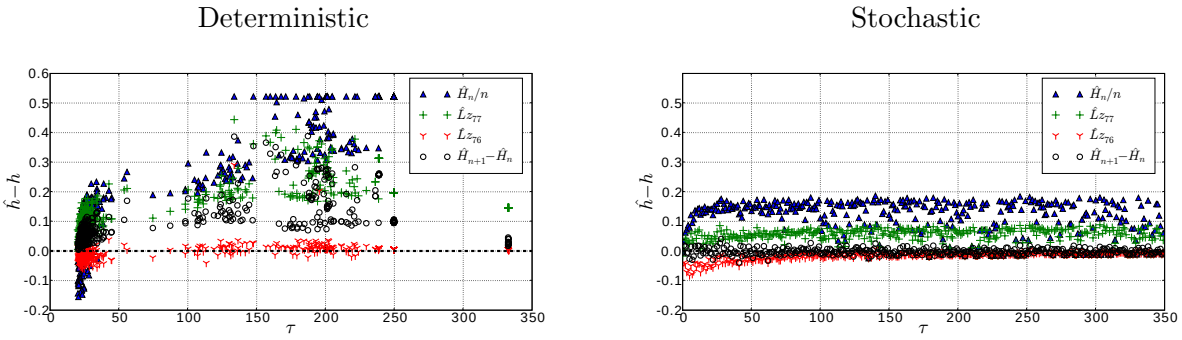

FIG. 3: Entropy estimation quality $\hat{h} - h$ as a function of the average integrated correlation time τ_n, in the deterministic case (left) and stochastic case (right). The numerical procedure is similar to that described in figure 2.

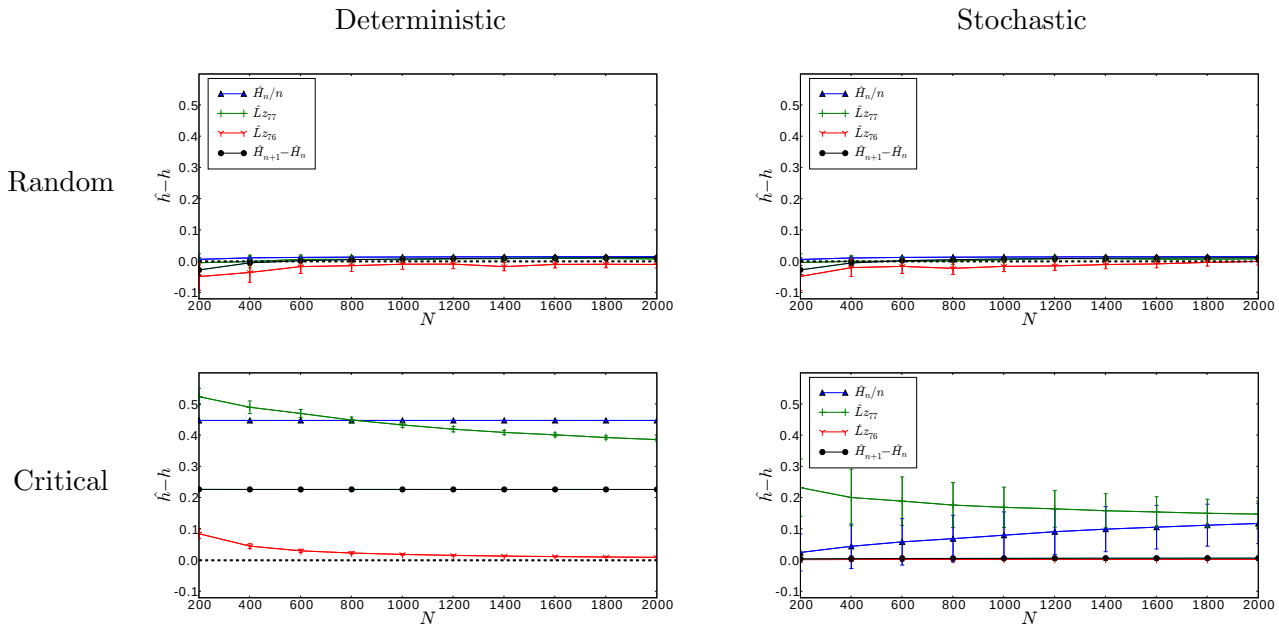

FIG. 4: Entropy estimation quality $\hat{h}-h$ as a function of the sequence length N in the deterministic case (left column) and stochastic case (right column) for full randomness (top row) and infinite-range correlations (bottom row). Fully random sequences are generated using the chaotic logistic map with parameter $a = 4$ (deterministic case) and an uncorrelated Markov chain with parameter $a = 1/2$ (stochastic case). Sequences with infinite-range correlations are generated using the critical logistic map with parameter $a = a_c \approx 3.569$ (deterministic case) and Markov chain with parameter $a = 10^{-3} \ll 1$ i.e. with correlation time $\tau(a) \approx 1/a \gg 1$ (stochastic case). The entropy of 200 symbolic sequences with random initial conditions and varying length $N \in [200, 2000]$ was computed using each estimator. Values represented on the figures correspond to the mean value over these 200 sequences. Standard deviation is depicted using vertical error bars. For the deterministic case and critical behavior, the caption inset is the same as for the other plots but for clarity, it is not depicted on the figure.