




## Estimating the variance of Shannon entropy

Leonardo Ricci <sup>1,2,\*</sup>, Alessio Perinelli <sup>2</sup> and Michele Castelluzzo <sup>1</sup>

<sup>1</sup>*Department of Physics, University of Trento, 38123 Trento, Italy*

<sup>2</sup>*CIMeC, Center for Mind/Brain Sciences, University of Trento, 38068 Rovereto, Italy*



(Received 24 April 2021; accepted 10 August 2021; published 25 August 2021)

The statistical analysis of data stemming from dynamical systems, including, but not limited to, time series, routinely relies on the estimation of information theoretical quantities, most notably Shannon entropy. To this purpose, possibly the most widespread tool is provided by the so-called plug-in estimator, whose statistical properties in terms of bias and variance were investigated since the first decade after the publication of Shannon's seminal works. In the case of an underlying multinomial distribution, while the bias can be evaluated by knowing support and data set size, variance is far more elusive. The aim of the present work is to investigate, in the multinomial case, the statistical properties of an estimator of a parameter that describes the variance of the plug-in estimator of Shannon entropy. We then exactly determine the probability distributions that maximize that parameter. The results presented here allow one to set upper limits to the uncertainty of entropy assessments under the hypothesis of memoryless underlying stochastic processes.

DOI: [10.1103/PhysRevE.104.024220](https://doi.org/10.1103/PhysRevE.104.024220)

### I. INTRODUCTION

The estimation of Shannon entropy [1] associated to a probability distribution lies at the very core of information theory and is a fundamental tool in statistical analysis. Its appealing interpretation in terms of *information content* makes Shannon entropy, along with metrics directly stemming from it, a crucial ingredient of many analytical techniques that are used to characterize complex systems. For example, mutual information, which is linked to the entropy difference between a joint distribution and the two related marginal distributions, is used to estimate the degree of correlation between two random variables [2]. In the analysis of time series generated by complex systems, entropy measures such as permutation entropy [3] and approximate and sample entropy [4] are widely used. Tools from information theory and based on entropy are common in neuroscience [5–7], genomics [8–10], physiology [11], economics [12,13], climatology [14–16], and geoscience [17,18].

A key question is how reliably an estimator evaluates the actual entropy. The question becomes relevant especially in experimental situations when a single acquisition of limited size is available and thus only a single estimate can be assessed, for example, in research fields that rely on historical time series or measurements of systems that cannot be controlled by the experimenter. While no general solution is known, the ways this issue is tackled depend on the—possibly unknown—physical mechanisms underlying the evolution of a system of interest. The issue of assessing an estimator's reliability is relevant also in wider contexts of information theory, such as Renyi entropies [19], mutual information [20], and entropy rate [21].

The most basic case occurs when states are randomly and independently accessed, so that their number of visits follow a multinomial distribution. Despite its apparent simplicity, the issue of reliably estimating entropy in the multinomial case is far from being straightforward [22]. Typically, Shannon entropy is estimated by relying on a recording of a system's discrete-time evolution: Upon identifying the states of interest, the related occurrences are counted over an  $N$ -step evolution to yield a histogram of sample rates. Entropy is then evaluated on this empirical distribution by using the so-called plug-in estimator [23]  $\hat{H}$ , also referred to as maximum likelihood [22] or naive [5] estimator. To assess the reliability of the plug-in estimator, crucial parameters are the support size  $M$  and the sample size  $N$ . Early works by, among others, Basharin [24], Miller [25], Harris [26] showed the consistency and the asymptotic normality of  $\hat{H}$ .

In general, both the bias and the variance affecting  $\hat{H}$  depend on  $N^{-1}$ . However, while the bias is well described by the so-called Miller-Madow correction [24–26] equal to  $-(M-1)/(2N)$  (see also Vinck *et al.* [27] for more recent developments), there is no general rule to estimate the variance  $\sigma_{\hat{H}}^2$ . More in detail, it is known that [24–26], upon defining the variance parameter  $\Lambda_0 = \sum_{i=1}^M s_i (H + \ln s_i)^2$ , where the  $s_i$ 's are the prior probabilities that define the multinomial distribution, and provided that  $\Lambda_0 \neq 0$ , one has  $\sigma_{\hat{H}}^2 \sim \Lambda_0/N$ . Unfortunately, whenever the multinomial distribution is unknown, so are  $\Lambda_0$  and  $\sigma_{\hat{H}}^2$ . Nevertheless, as shown by Antos and Kontoyiannis [23],  $\sigma_{\hat{H}}^2$  has an upper bound given by  $(\ln N)^2/N$ . In addition, Roulston [28] derived, via standard error analysis, an estimate of  $\sigma_{\hat{H}}^2$  that has a widespread use (see, for example, Ref. [29]).

In the present work, we address the problem of estimating the variance parameter  $\Lambda_0$  that describes  $\sigma_{\hat{H}}^2$  in the multinomial case. In the first part, we investigate the statistical

\*leonardo.ricci@unitn.it

properties of a *straightforward* plug-in estimator  $\widehat{\Lambda}_0$  for  $\Lambda_0$ . The estimator is shown to be itself consistent and asymptotically normal. Both its bias and variance are shown to be inversely proportional to the finite sample size  $N$  making up the available data set.

In the second part, we derive an upper bound for the variance parameter  $\Lambda_0$  that depends on the support size  $M$ . Provided that this support size is known, the upper bound can be used to set a limit to the degree of confidence on an estimate of  $H$  via  $\widehat{H}$  out of a sample histogram. The maximum variance parameter  $\Lambda_{0, \max}$  is shown to correspond to a distribution having a single, *outlier* state, whereas the access probability to the remaining ones is uniformly distributed.

The paper is organized as follows. In Sec. II, upon summarizing the statistical properties of  $\widehat{H}$ , we introduce the plug-in estimator of the variance parameter  $\Lambda_0$  and we derive its asymptotic distribution. Section III concerns the assessment of the maximum variance parameter  $\Lambda_{0, \max}$  in the multinomial case and the related underlying distributions. Numerical simulations to test the reliability of the estimator of the variance parameter  $\Lambda_0$  and to compare it with the one proposed by Roulston are discussed in Sec. IV. Final remarks are presented in Sec. V.

## II. PLUG-IN ESTIMATOR OF THE VARIANCE PARAMETER $\Lambda_0$

### A. A summary on the plug-in entropy estimator $\widehat{H}$ and its statistical properties

The system of interest consists of  $M$  states and has a purely stochastic evolution: the transition probability from state  $j$  to state  $i$  is independent of  $j$ , where, here and henceforth, both  $i$  and  $j$  run between 1 and  $M$ . Let  $s_i$  be the probability with which the  $i$ th state is visited. The Shannon entropy  $H$  of the system is given by

$$H = - \sum_{i=1}^M s_i \ln s_i. \quad (1)$$

Throughout the paper it is assumed—via continuous extension—that  $x \ln x = 0$  if  $x = 0$ , for each  $n \in \mathbb{N}$ . However, unless otherwise specified, we suppose here that  $M$  coincides with the size of the support of the distribution  $\{s_i\}$ , so that  $s_i > 0$ ,  $\forall i$ .

A way to estimate  $H$  relies on the plug-in estimator  $\widehat{H}$ , as follows. (Henceforth, given a typically unknown quantity  $a$ , the symbol  $\widehat{a}$  refers to an estimator of  $a$  that, as a function of observed data, is a statistic affected by bias and fluctuations). We consider  $N$  consecutive steps of the system's evolution and record the  $M$ -dimensional set  $\{\widehat{j}_i\}$  of the number of visits of each one of the  $M$  states. The set  $\{\widehat{j}_i\}$  is distributed according to a multinomial distribution, and it holds that  $\sum_{i=1}^M \widehat{j}_i = N$ . Upon computing, for each  $i$ , the observed rate  $\widehat{p}_i$  as  $\widehat{p}_i = \widehat{j}_i/N$ , the plug-in estimator  $\widehat{H}$  is defined as

$$\widehat{H} \equiv - \sum_{i=1}^M \widehat{p}_i \ln \widehat{p}_i.$$

The plug-in estimator is affected by a bias term that is known as the Miller-Madow correction [24–26] and is equal

to  $-(M-1)/(2N)$ . With regard to the variance of the plug-in estimator, let first the variance parameter  $\Lambda_0$  be defined as

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

Then, as shown by Basharin [24] and Harris [26], in the multinomial case and provided that  $\Lambda_0 > 0$ , the variance  $\sigma_{\widehat{H}}^2$  is given by

$$\sigma_{\widehat{H}}^2 = \frac{\Lambda_0}{N} + O\left(\frac{1}{N^{3/2}}\right).$$

Note that by writing  $\Lambda_0$  in the form mentioned in Sec. I, namely  $\Lambda_0 = \sum_{i=1}^M s_i (H + \ln s_i)^2$ , it can be promptly shown that  $\Lambda_0 \geq 0$ , where the equality holds if and only if the probabilities  $s_i$  are uniformly distributed.

As pointed out in the introduction, estimating the variance of the plug-in estimator requires the knowledge of the parameter  $\Lambda_0$ , for which an estimator is proposed and analyzed in the following sections.

### B. Plug-in estimator $\widehat{\Lambda}_0$ of the variance parameter $\Lambda_0$

The estimator of the variance parameter  $\Lambda_0$  proposed in this paper is the plug-in estimator that is evaluated out of the rate histogram  $\{\widehat{p}_i\}$  as

$$\widehat{\Lambda}_0(\widehat{p}_1, \dots, \widehat{p}_M) = \sum_{i=1}^M (\widehat{p}_i \ln^2 \widehat{p}_i) - \widehat{H}^2. \quad (3)$$

Because the quantity  $\widehat{\Lambda}_0$  is a sample statistic, it is necessary to assess its reliability in terms of bias and variance. This issue is tackled by determining the asymptotic behavior of  $\widehat{\Lambda}_0$  as  $N \rightarrow \infty$ . To this goal, it is suitable to define, for each  $i$ ,  $\zeta_i \equiv (\widehat{p}_i - s_i)\sqrt{N}$ , so that the rates can be expressed as

$$\widehat{p}_i = s_i + \frac{\zeta_i}{\sqrt{N}}.$$

It follows that

$$\begin{aligned} \widehat{\Lambda}_0 &= \sum_{i=1}^M \left[ \left( s_i + \frac{\zeta_i}{\sqrt{N}} \right) \ln^2 \left( s_i + \frac{\zeta_i}{\sqrt{N}} \right) \right] \\ &\quad - \left[ \sum_{i=1}^M \left( s_i + \frac{\zeta_i}{\sqrt{N}} \right) \ln \left( s_i + \frac{\zeta_i}{\sqrt{N}} \right) \right]^2. \end{aligned}$$

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

$$\begin{aligned} \widehat{\Lambda}_0 &= \Lambda_0 + \frac{1}{\sqrt{N}} \sum_{i=1}^M \zeta_i [\ell_i^2 + 2(1+H)\ell_i] \\ &\quad + \frac{1}{N} \left[ \sum_{i=1}^M \left( \frac{1+H+\ell_i}{s_i} \zeta_i^2 \right) - \sum_{i=1}^M \sum_{j=1}^M (\zeta_i \zeta_j \ell_i \ell_j) \right] \\ &\quad + O\left(\frac{1}{N^{3/2}}\right), \end{aligned}$$

where  $\ell_i \equiv \ln s_i$  and the identity  $\sum_{i=1}^M \zeta_i = 0$ , which follows from the constraint  $\sum_{i=1}^M \widehat{p}_i = 1$ , was used.

Let  $G_{\delta\widehat{\Lambda}_0}(t)$  be the moment-generating function of the residual  $\delta\widehat{\Lambda}_0 \equiv \widehat{\Lambda}_0 - \Lambda_0$ , where  $t$  is a real variable defined in a neighborhood of the origin. It holds that

$$G_{\delta\widehat{\Lambda}_0}(t) = E(e^{t\delta\widehat{\Lambda}_0}).$$

The exponential form  $e^{t\delta\widehat{\Lambda}_0}$  can be expressed as

$$\begin{aligned} e^{t\delta\widehat{\Lambda}_0} &= 1 + \frac{t}{\sqrt{N}} \sum_{i=1}^M \zeta_i [\ell_i^2 + 2(1+H)\ell_i] \\ &\quad + \frac{t}{N} \left[ \sum_{i=1}^M \left( \frac{1+H+\ell_i}{s_i} \zeta_i^2 \right) - \sum_{i=1}^M \sum_{j=1}^M (\zeta_i \zeta_j \ell_i \ell_j) \right] \\ &\quad + \frac{t^2}{2N} \sum_{i=1}^M \sum_{j=1}^M \zeta_i \zeta_j [\ell_i^2 + 2(1+H)\ell_i] \\ &\quad \times [\ell_j^2 + 2(1+H)\ell_j] \\ &\quad + O\left(\frac{1}{N^{3/2}}\right). \end{aligned}$$

Because the starting vector of the system's evolution is randomly chosen according to the distribution  $\{s_i\}$ , the expected value of each  $\zeta_i$  is given by [30]  $E(\zeta_i) = O(1/N)$ , so that the term proportional to  $t/\sqrt{N}$  is absorbed within the term  $O(N^{-3/2})$ . In addition [30], one has  $E(\zeta_i \zeta_j) = \chi_{i,j} + O(N^{-1/2})$ , where  $\chi_{i,j} = \delta_{i,j} s_i - s_i s_j$  is the covariance matrix of rates being distributed according to a multinomial distribution ( $\delta_{i,j}$  is the Kronecker delta). It follows that

$$\begin{aligned} G_{\delta\widehat{\Lambda}_0}(t) &= \frac{t}{N} \left[ \sum_{i=1}^M \left( \frac{1+H+\ell_i}{s_i} \chi_{i,i} \right) - \sum_{i=1}^M \sum_{j=1}^M (\chi_{i,j} \ell_i \ell_j) \right] \\ &\quad + \frac{t^2}{2N} \sum_{i=1}^M \sum_{j=1}^M \chi_{i,j} [\ell_i^2 + 2(1+H)\ell_i] \\ &\quad \times [\ell_j^2 + 2(1+H)\ell_j] \\ &\quad + O\left(\frac{1}{N^{3/2}}\right). \end{aligned} \tag{4}$$

Let the coefficients of  $t/N$  and  $t^2/(2N)$  in Eq. (4) be defined as  $\gamma$  and  $\Gamma$ , respectively:

$$\begin{aligned} \gamma &\equiv \sum_{i=1}^M \left( \frac{1+H+\ell_i}{s_i} \chi_{i,i} \right) - \sum_{i=1}^M \sum_{j=1}^M (\chi_{i,j} \ell_i \ell_j), \\ \Gamma &\equiv \sum_{i=1}^M \sum_{j=1}^M \chi_{i,j} [\ell_i^2 + 2(1+H)\ell_i] [\ell_j^2 + 2(1+H)\ell_j]. \end{aligned}$$

Equation (4) can be then rewritten as

$$G_{\delta\widehat{\Lambda}_0}(t) = \frac{t}{N} \gamma + \frac{t^2}{2N} \Gamma + O\left(\frac{1}{N^{3/2}}\right). \tag{5}$$

It is worth noting that, because  $\chi$  is positive semidefinite, from the definition of  $\Gamma$  one has  $\Gamma \geq 0$ .

As a consequence of the properties of the moment-generating function, the coefficient of  $t$  in this last equation

corresponds to the expected value of  $\delta\widehat{\Lambda}_0$ , i.e., of  $\widehat{\Lambda}_0 - \Lambda_0$ :

$$\mu_{\delta\widehat{\Lambda}_0} = \mu_{\widehat{\Lambda}_0} - \Lambda_0 = \frac{\gamma}{N} + O\left(\frac{1}{N^{3/2}}\right). \tag{6}$$

Also, because the square of this last term is of order  $N^{-2}$ , the coefficient of  $t^2/2$  in Eq. (5) directly provides the variance of  $\delta\widehat{\Lambda}_0$  and thus of  $\widehat{\Lambda}_0$ :

$$\sigma_{\delta\widehat{\Lambda}_0}^2 = \sigma_{\widehat{\Lambda}_0}^2 = \frac{\Gamma}{N} + O\left(\frac{1}{N^{3/2}}\right). \tag{7}$$

Equations (6) and (7) provide the bias and the variance of the plug-in estimator  $\widehat{\Lambda}_0$  of the variance parameter  $\Lambda_0$ , respectively.

Besides these results, Eq. (5) can be further exploited: Starting from it, the moment-generating function of the random variable  $w \equiv (\delta\widehat{\Lambda}_0 - \frac{\gamma}{N})\sqrt{N}$  is given by

$$G_w(t) = \exp\left(\frac{t^2}{2}\Gamma\right) + O\left(\frac{1}{N^{1/2}}\right),$$

where  $t$  is a real variable defined in a neighborhood of the origin. This expression shows that, provided that  $\Gamma$  is nonvanishing,  $w$  is asymptotically normally distributed with zero mean and variance  $\Gamma$ . It is therefore possible to formulate a central limit theorem for the plug-in estimator  $\widehat{\Lambda}_0$  of the variance parameter  $\Lambda_0$ :

$$\widehat{\Lambda}_0 \sim \mathcal{N}\left(\Lambda_0 + \frac{\gamma}{N}, \frac{\Gamma}{N}\right) \quad \text{as } N \rightarrow \infty.$$

The main result of this section can be summarized as the fact that the plug-in estimator  $\widehat{\Lambda}_0$  defined in Eq. (3) is asymptotically normal and is a consistent estimator (both its population mean and variance depend on  $N^{-1}$ ) of the variance parameter  $\Lambda_0$ .

### C. Practical considerations about $\widehat{\Lambda}_0$

From a practical point of view, using the asymptotic expressions above requires the knowledge of the parameters  $\gamma$  and  $\Gamma$ . The two parameters can be expressed upon straightforward algebra as

$$\gamma = MH + M - 1 - \Lambda_0 + \sum_{i=1}^M \ell_i, \tag{8}$$

$$\begin{aligned} \Gamma &= \mu'_4 - 4\mu'_3(H+1) + H^3(3H+4) \\ &\quad + 2\Lambda_0(3H^2 + 6H + 2) - \Lambda_0^2, \end{aligned} \tag{9}$$

where  $\mu_3, \mu_4$  are the third and fourth moments of the distribution of the single-bin entropy  $-\ell_i = -\ln s_i$ , which are generally defined as

$$\mu'_n \equiv \sum_{i=1}^M s_i (-\ell_i)^n,$$

so that, for example,  $\mu'_1 = H$ ,  $\mu'_2 = \Lambda_0 + H^2$ .

The presence, in the expressions above, of  $\mu'_3, \mu'_4, \sum_i \ell_i$ , makes the estimates of  $\gamma, \Gamma$ , and thus of  $\widehat{\Lambda}_0$ , a circular problem. In principle, provided that sufficiently large data samples are available, one could carry out a statistical analysis on subsamples of different size  $N$ , in order to “experimentally

assess” the parameters  $\gamma$ ,  $\Gamma$ . In practice, however, and especially in the case of limited data samples, it is difficult to estimate the reliability of  $\widehat{\Lambda}_0$  and thus of the variance of the estimated Shannon entropy  $\widehat{H}$ . In the next section, this issue is circumvented by evaluating the maximum value of  $\Lambda_0$  as a function of the number  $M$  of the system’s states.

### III. MAXIMA OF THE VARIANCE PARAMETER $\Lambda_0$

The goal of this section is to find the maximum value of  $\Lambda_0$  and the related probability distribution(s)  $\{s_i\}$  for a given support size  $M$ .

Let  $\mathbb{D}$  be the probability  $M$ -simplex defined as

$$\mathbb{D} \equiv \left\{ \mathbf{r} \in \mathbb{R}^M \mid \sum_{i=1}^M r_i = 1, r_i \geq 0 \forall i \right\}.$$

Due to the settings  $x \ln^n x = 0$  if  $x = 0$  and  $\forall n \in \mathbb{N}$ , which rely on continuous extensions of the respective functions, both  $H$  and  $\Lambda_0$  are continuous on the simplex  $\mathbb{D}$ . According to the Weierstrass extreme value theorem [31] and the theorem of necessary condition for extreme values [32], the extreme values of  $\Lambda_0$  can only occur either on the boundary  $\partial\mathbb{D}$  of  $\mathbb{D}$  or in stationary points within  $\mathbb{D}$ . We first discuss the latter case.

#### A. Stationary points in the interior of the probability simplex

Because the probabilities  $s_i$  have to add up to one, only  $M - 1$  values are independent. We then consider  $s_M$  as a function of the other  $M - 1$  probabilities:

$$s_M = 1 - \sum_{i=1}^{M-1} s_i. \quad (10)$$

To find the stationary points of  $\Lambda_0$ , its partial derivative with respect to the independent  $s_i$ ’s, with  $1 \leq i \leq M - 1$ , is considered:

$$\frac{\partial \Lambda_0}{\partial s_i} = (\ln s_i - \ln s_M)(\ln s_i + \ln s_M + 2 + 2H).$$

A vanishing partial derivative requires fulfilling one of two possibilities, namely

$$s_i = p_0 \equiv s_M, \quad (11)$$

$$s_i = q_0 \equiv \frac{1}{s_M} e^{-2-2H}. \quad (12)$$

The two possibilities are mutually exclusive: If it were  $p_0 = q_0$ , then it would follow, first,  $s_i = 1/M$ ,  $\forall i \in [1, M]$ , so that  $H = \ln M$ , and, second,  $s_M = \frac{1}{s_M} e^{-2-2H}$ , so that  $H = 1 + \ln M$ . Therefore, it must hold that  $p_0 \neq q_0$ .

Let  $k$  be the number of  $s_i$ ’s that satisfy the first possibility, Eq. (11). Because  $s_M$  trivially satisfies this condition, one has  $1 \leq k \leq M$ . However, the case  $k = M$  corresponds to a uniform distribution where  $s_i = 1/M$ ,  $\forall i \in [1, M]$ . In this case,  $\Lambda_0 = 0$  and, as shown by Harris [26], the dominant variance term is of order  $N^{-2}$ . The uniform distribution thus yields the vanishing, minimum value of  $\Lambda_0$ , which also occurs in the  $M$  vertices of the simplex, namely whenever a single  $s_i$  is unitary while all the others are vanishing. Henceforth we

therefore restrict the discussion to the range  $1 \leq k \leq M - 1$ . The number of  $s_i$ ’s that satisfy the second possibility, Eq. (12), is then  $M - k$ , which also varies between 1 and  $M - 1$ .

Equations (1), (10), and (12) can be respectively rewritten as follows:

$$-kp_0 \ln p_0 - (M - k)q_0 \ln q_0 = H, \quad (13)$$

$$kp_0 + (M - k)q_0 = 1, \quad (14)$$

$$\ln p_0 + \ln q_0 + 2 + 2H = 0. \quad (15)$$

Equations (13), (14), and (15) make up a system of three equations in the three variables  $H$ ,  $p_0$ ,  $q_0$ , while the integer number  $k$  is a fixed parameter. Inserting the expression for  $H$  given by Eq. (13) into Eq. (15) and replacing  $q_0$  with  $\frac{1-kp_0}{M-k}$ , as given by Eq. (14), yields

$$(1 - 2kp_0) \ln \frac{1 - kp_0}{kp_0} = 2 + (1 - 2kp_0) \ln \frac{M - k}{k}. \quad (16)$$

Crucially, because of Eq. (14) and the fact that we are considering the interior of  $\mathbb{D}$ , it holds that  $0 < kp_0 < 1$  and  $0 < (M - k)q_0 < 1$ .

It is useful to introduce the auxiliary variable  $v$  defined as

$$v \equiv 2kp_0 - 1.$$

It follows that

$$|v| < 1,$$

$$p_0 = \frac{1 + v}{2k}, \quad (17)$$

$$q_0 = \frac{1 - v}{2(M - k)}. \quad (18)$$

By also defining the function  $f(v)$  as

$$f(v) \equiv v \ln \frac{1 + v}{1 - v}, \quad (19)$$

Eq. (16) can be rewritten as

$$f(v) = 2 - v \ln \frac{M - k}{k}. \quad (20)$$

The graph of  $f(v)$  corresponds to the continuous, red line shown in Fig. 1.

It is straightforward to show that  $f(v)$  satisfies the following properties:  $f(-v) = f(v)$ ;  $f(v) \geq 0$ ;  $f(v) \rightarrow +\infty$  as  $v \rightarrow \pm 1$ . In addition, because the derivative  $f'(v)$  is given by

$$f'(v) = \frac{1}{v} \left[ f(v) + \frac{2v^2}{1 - v^2} \right],$$

the function  $f(v)$  is monotonically increasing (decreasing) if  $v > 0$  ( $v < 0$ ) and thus has a single minimum equal to 0 in  $v = 0$ . As shown in Fig. 1, the right-hand term of Eq. (20) is instead a straight line going through the point  $(0, 2)$  and having a slope  $-\ln \frac{M-k}{k}$ .

For any  $k$ , the roots of the transcendental Eq. (20) provide, via Eqs. (17) and (18), stationary points for  $\Lambda_0 = \Lambda_0(p_0, p_1, \dots, p_{M-1})$ . Let  $\tilde{v}_k$  be a root of Eq. (20) for a given  $k$ . According to Eq. (2) and by using Eq. (17) and (18), the variance parameter  $\Lambda_0$  can be promptly rewritten as

$$\Lambda_0 = \frac{1}{\tilde{v}_k^2} - 1. \quad (21)$$

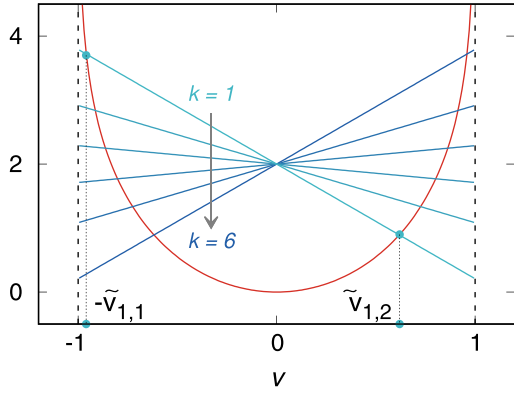


FIG. 1. Graphical representation of Eq. (20). The red, solid line corresponds to the graph of the function  $f(v)$  defined in Eq. (19). The straight lines correspond to the right-hand term of Eq. (20) in the case  $M = 7$  and for different values of the parameter  $k$ . The blue dots mark the intersection of  $f(v)$  with the straight line in the case  $k = 1$ , as well as the two abscissae  $-\tilde{v}_{k,1}$ ,  $\tilde{v}_{k,2}$ .

It then follows that the maximum value of  $\Lambda_0$  occurs for a stationary point characterized by a value  $k$  that yields the minimum  $|\tilde{v}_k|$ .

To assess this  $k$  value and the corresponding minimum  $|\tilde{v}_k|$  we start observing that for each  $k$  there are two roots of Eq. (20), one negative and one positive, which are henceforth referred to as  $-\tilde{v}_{k,1}$ ,  $\tilde{v}_{k,2}$ , respectively, with  $\tilde{v}_{k,1} > 0$ ,  $\tilde{v}_{k,2} > 0$ . From Eq. (20) it holds that  $\tilde{v}_{M-k,2} = \tilde{v}_{k,1}$ . It is therefore sufficient to study the cases  $k \leq M/2$  for which the factor  $\ln \frac{M-k}{k}$  is non-negative.

Inserting  $-\tilde{v}_{k,1}$ ,  $\tilde{v}_{k,2}$  in Eq. (20) and subtracting the results yields

$$f(-\tilde{v}_{k,1}) - f(\tilde{v}_{k,2}) = (\tilde{v}_{k,1} + \tilde{v}_{k,2}) \ln \frac{M-k}{k} \geq 0.$$

Consequently,  $f(\tilde{v}_{k,1}) = f(-\tilde{v}_{k,1}) \geq f(\tilde{v}_{k,2})$  and, because  $f(v)$  is monotonically increasing if  $v > 0$ , it follows  $\tilde{v}_{k,1} \geq \tilde{v}_{k,2}$ . The root  $\tilde{v}_{k,2}$  is therefore the closest one to the origin.

We now show that  $\tilde{v}_{k,2}$  decreases if  $k$  increases by one. Because  $\frac{M-k}{k} > \frac{M-k-1}{k+1}$ , one has

$$\frac{2 - f(\tilde{v}_{k,2})}{\tilde{v}_{k,2}} > \frac{2 - f(\tilde{v}_{k+1,2})}{\tilde{v}_{k+1,2}},$$

i.e.,

$$\frac{2}{\tilde{v}_{k,2}} - \frac{2}{\tilde{v}_{k+1,2}} > \frac{f(\tilde{v}_{k,2})}{\tilde{v}_{k,2}} - \frac{f(\tilde{v}_{k+1,2})}{\tilde{v}_{k+1,2}}.$$

If it were  $\tilde{v}_{k,2} \geq \tilde{v}_{k+1,2}$ , the left-hand term of the previous inequality would be nonpositive, whereas the right-hand term would be non-negative (the function  $f(v)/v = \ln \frac{1+v}{1-v}$  is monotonically increasing if  $0 < v < 1$ ), thus leading to a contradiction. It then must hold that  $\tilde{v}_{k,2} < \tilde{v}_{k+1,2}$ . Consequently,  $\tilde{v}_{1,2}$  provides the minimum absolute value of a root of Eq. (20) and thus, via Eq. (21), the maximum value of  $\Lambda_0$  for a given  $M$  with regard to the interior of the simplex  $\mathbb{D}$ .

As a corollary, the very same argument of the last paragraph can be promptly applied to show that increasing  $M$  leads to a decrease of  $\tilde{v}_{1,2}$  and thus to an increase of  $\Lambda_{0, \max}$ .

This corollary is important in order to discuss the behavior of the variance parameter  $\Lambda_0$  on the boundary of the probability simplex.

It is finally important to note that, when  $k = 1$ , the single *outlier* among the set of probabilities  $\{s_i\}$  is  $s_M$ , namely the one that was chosen as dependent on all the others [see Eq. (10)]. Because  $M$  similar choices are possible, the number of equivalent maxima of  $\Lambda_0$  is  $M$ .

### B. Boundary of the probability simplex

The boundary  $\partial\mathbb{D}$  of the  $M$ -simplex  $\mathbb{D}$  is the union of  $M$  boundary facets. Each facet is defined by a single  $s_j$ , with  $1 \leq j \leq M$ , set to zero, and it thus corresponds to a  $(M-1)$ -simplex.

This last assumption contradicts the fact that  $M$  is the support size of the probability distribution. However, we discuss these cases for the sake of completeness and because, in principle, one could have arbitrarily small  $s_i$  values.

Because each facet corresponds to a simplex, the arguments of the previous section can be iteratively applied as follows. First, in the interior of the  $j$ th facet, the maximum of  $\Lambda_0$  will be  $1/\tilde{v}'_{1,2} - 1$ , where  $\tilde{v}'_{1,2}$  is the negative root of Eq. (20) when  $k = 1$  and  $M$  is replaced by  $M-1$ . Second, the boundary of the  $j$ th facet has to be treated as a  $(M-2)$ -dimensional set of  $(M-2)$  simplexes, and so on. The iteration stops at the two-dimensional simplexes in which only a pair  $(s_i, s_j)$  of probabilities is nonzero: In this case, the boundaries are the pointlike vertices of a segment, in which  $\Lambda_0 = 0$ .

Because of the corollary expressed at the end of the previous section, Sec. III A, the maximum of  $\Lambda_0$  on the  $j$ th facet is strictly less than the value produced by a stationary point lying within  $\mathbb{D}$ , namely  $1/\tilde{v}_{1,2}^2 - 1$ .

### C. Maximum value of $\Lambda_0$ and related distributions

By taking into account the results of Secs. III A and III B, we can conclude that the maximum of  $\Lambda_0$  in the multinomial case occurs when the distribution of the probabilities  $\{s_i | i \in [1, M]\}$  has a single entry equal to the value  $p_0$  and all the other entries uniformly distributed and equal to  $q_0$ .

Upon finding the single positive root  $\tilde{v} \equiv \tilde{v}_{1,2}$  of the transcendental equation

$$f(v) = 2 - v \ln(M-1), \quad (22)$$

which corresponds to Eq. (20) with  $k = 1$ , one has

$$p_0 = \frac{1 + \tilde{v}}{2}, \quad (23)$$

$$\Lambda_{0, \max} = \frac{1}{\tilde{v}^2} - 1. \quad (24)$$

Equation (22) can be approximately solved by replacing  $f(v)$  with the first term of its Taylor expansion, namely  $2v^2$ . Equation (22) then becomes a second-degree equation whose positive solution is given by

$$\tilde{v} \approx \sqrt{\frac{\ln^2(M-1)}{16} + 1} - \frac{\ln(M-1)}{4}, \quad (25)$$



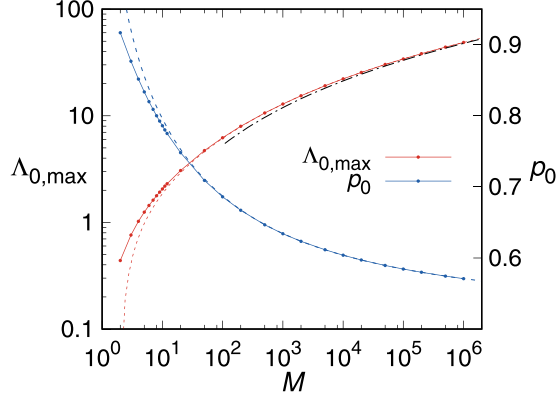


FIG. 2. Maximum variance parameter  $\Lambda_{0,\max}$  as a function of the support size  $M$ : (red, solid line) exact computation; (red, dashed line) evaluation via Eq. (24) by using the approximated value for  $\tilde{v}$  given by Eq. (25); (black, dash-dotted line) approximation described by Eq. (26) and valid for  $M \gtrsim 100$ . Outlier probability  $p_0$  as a function of the support size  $M$ : (blue, solid line) exact computation; (blue, dashed line) evaluation via Eq. (23) by using the approximated value for  $\tilde{v}$  given by Eq. (25).

which, for large  $M$  values, can be further approximated as  $\tilde{v} \sim 2/\ln(M)$ . (These approximations improve with  $M$  because the larger  $M$ , the smaller  $\tilde{v}$ ). Figure 2 shows the plots of  $p_0$ ,  $\Lambda_{0,\max}$  as a function of the dimension  $M$ . The approximation based on Eq. (25) works well for  $M \gtrsim 10$ . It is worth observing that, as  $M \rightarrow \infty$ , one has

$$\Lambda_{0,\max} \cong \frac{\ln^2(M)}{4}, \quad (26)$$

while  $p_0$  tends to  $1/2$ .

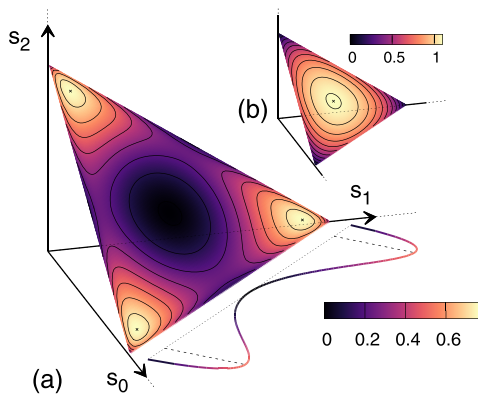


FIG. 3. (a) Color map of the variance parameter  $\Lambda_0$  on the probability simplex in the case  $M = 3$ . Three points, located at one coordinate being equal to  $p_0 \cong 0.88$  and the other two equal to  $q_0 \cong 0.06$ , provide the maximum value of  $\Lambda_0$ , namely  $\Lambda_{0,\max} \cong 0.762$ . The curve placed on the plane  $\pi_{s_0,s_1}$  represents the plot of  $\Lambda_0$  on the simplex facet corresponding to  $s_2 = 0$ . (b) Color map of Shannon entropy  $H$  on the probability simplex in the case  $M = 3$ . The maximum value occurs in the center, where  $s_0 = s_1 = s_2 = 1/3$  and  $\Lambda_0$  vanishes. No additional stationary points are present (which is true at any dimension).

The number of different distributions that maximize  $\Lambda_0$  is equal to  $M$ , one for each state taking on the value  $p_0$ . Figure 3 shows the case  $M = 3$ .

It is important to note that, in the case of a distribution that maximizes  $\Lambda_0$ , the parameter  $\Gamma$ , which is given by Eq. (9) and describes the  $N^{-1}$  dependence of the plug-in estimator  $\widehat{\Lambda}_0$  of  $\Lambda_0$ , vanishes. A proof of this statement is discussed in the Appendix. The situation is similar to what happens to the plug-in entropy estimator  $\widehat{H}$  when the distribution is uniform: In this case,  $H$  takes on its maximum value equal to  $\ln(M)$ , while the quantity that describes the expected value of the variance of the plug-in estimator, namely  $\Lambda_0/N$ , vanishes and it is replaced by higher-order terms (typically  $\sim N^{-2}$  [26]). The same is expected to occur in the case of  $\Gamma$ .

Finally, it is worth considering how “peaked” the maxima of  $\Lambda_0$  are. The partial derivatives with respect to the independent  $s_i$ ’s, evaluated in the locations corresponding to  $\Lambda_0 = \Lambda_{0,\max}$  are given by

$$\begin{aligned} \frac{\partial^2 \Lambda_0}{\partial s_i \partial s_j} &= -\left(\ln \frac{p_0}{q_0}\right) \left[ \frac{\delta_{i,j}}{q_0} - \frac{1}{p_0} + 2 \left(\ln \frac{p_0}{q_0}\right) \right] \\ &= -\Lambda_{0,\max} \frac{4}{1-\tilde{v}^2} \left[ \delta_{i,j}(M-1) \frac{\tilde{v}}{1-\tilde{v}} + \frac{2+\tilde{v}}{1+\tilde{v}} \right] \\ &\approx -8\Lambda_{0,\max} \left[ 1 + \delta_{i,j} \frac{M}{\ln M} \right], \end{aligned}$$

where the last approximation holds provided that  $M \gg 1$  so that  $\tilde{v} \sim 2/\ln(M)$ . In this limit, upon defining  $\delta s_M$  and  $\delta s_{\parallel}$  as

$$\begin{aligned} \delta s_M &= -\sum_{i=1}^{M-1} \delta s_i, \\ \delta s_{\parallel} &= \left[ \sum_{i=1}^{M-1} (\delta s_i)^2 \right]^{1/2}, \end{aligned}$$

where  $\delta s_i$  is the infinitesimal displacement of the respective  $s_i$  from the value  $q_0$ , we have the following Taylor expansion for  $\Lambda_0$ :

$$\frac{\Lambda_0}{\Lambda_{0,\max}} \approx 1 - 4 \left[ (\delta s_M)^2 + \frac{M}{\ln M} (\delta s_{\parallel})^2 \right].$$

#### IV. NUMERICAL EXPERIMENTS

To test the theory discussed in the previous two sections, two different distributions  $\{s_i\}$  were considered, both with  $M = 5$ : an “arithmetic progression” distribution (red inset in Fig. 5), where  $s_i$  is proportional to  $i$ ; and a “maximum variance” distribution (blue inset in Fig. 5), namely one of the distributions that yield a maximum  $\Lambda_0$  in the case  $M = 5$ . Here,  $s_1 = p_0 \cong 0.834$ , whereas the other  $s_i$ ’s are equal to  $q_0 \cong 0.042$ .

For each of the two distributions and for values of  $N$  ranging from  $10^2$  to  $10^6$ , a sample  $\mathbb{S}$  of  $10^4$  evolutions were simulated, and for each evolution the plug-in estimator  $\widehat{\Lambda}_0$  was evaluated. Thereupon, the results were averaged on the sample  $\mathbb{S}$  in order to estimate the expected values of the mean and the variance of  $\widehat{\Lambda}_0$ , according to Eqs. (6) and (7), respectively. By neglecting higher-order terms in those equations,

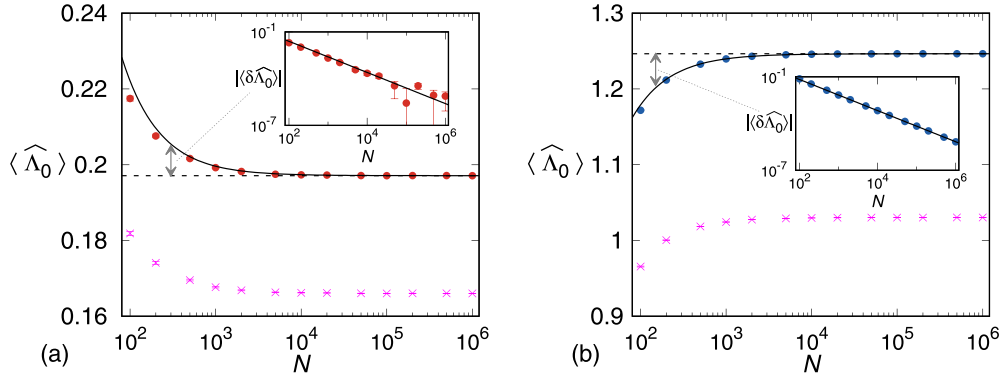


FIG. 4. Plug-in estimator  $\widehat{\Lambda}_0$  of the variance parameter  $\Lambda_0$  as a function of the number of steps  $N$  in the case of (a) the “arithmetic progression” distribution and (b) the “maximum variance” distribution. Each red (a) or blue (b) dot and the related error bar corresponds to the sample mean and the sample standard deviation, respectively, of  $\widehat{\Lambda}_0$  evaluated on a sample  $\mathbb{S}$  of  $10^4$  simulated evolutions of  $N$  steps. The black solid lines correspond to the expected value of  $\langle \widehat{\Lambda}_0 \rangle$  given by Eq. (27). For both panels (a) and (b), the insets show in log-log scale both the same numerical values and theoretical curves referred to the respective asymptotic theoretical values of  $\Lambda_0$  [dashed lines: (a),  $\Lambda_0 \cong 0.197$ ; (b),  $\Lambda_0 \cong 1.246$ ]. Each magenta “x” symbol and the related error bar correspond to the sample mean and the sample standard deviation, respectively, of the coefficient of  $N^{-1}$  in Eq. (40) of Ref. [28] evaluated on the same sample  $\mathbb{S}$ .

the expected value of the sample mean  $\langle \widehat{\Lambda}_0 \rangle$  and the sample variance  $s_{\widehat{\Lambda}_0}^2$  evaluated on a sample  $\mathbb{S}$  are given by

$$E(\langle \widehat{\Lambda}_0 \rangle) \cong \Lambda_0 + \frac{\gamma}{N}, \tag{27}$$

$$E(s_{\widehat{\Lambda}_0}^2) \cong \frac{\Gamma}{N}. \tag{28}$$

Figures 4 and 5 show the results of the numerical experiments for the mean and the variance of the plug-in estimator  $\widehat{\Lambda}_0$ , respectively. For both distributions in the case of the mean and for the “arithmetic progression” distribution in the case of the variance, the agreement with theoretical predictions is evident. In the case of the “maximum variance” distribution,

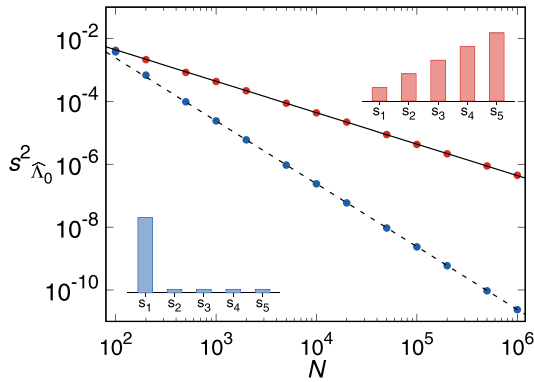


FIG. 5. Variance of the plug-in estimator  $\widehat{\Lambda}_0$  of the variance parameter  $\Lambda_0$  as a function of the number of steps  $N$  in the case of (red) the “arithmetic progression” distribution and (blue) the “maximum variance” distribution. Each dot corresponds to the sample variance of  $\widehat{\Lambda}_0$  evaluated on a sample  $\mathbb{S}$  of  $10^4$  simulated evolutions of  $N$  steps. The black solid line correspond to the expected value of  $s_{\widehat{\Lambda}_0}^2$  given by Eq. (28). In the case of the “maximum variance” distribution, the expected value of  $\sigma_{\widehat{\Lambda}_0}^2$  vanishes (see main text). The black dashed line corresponds to a  $N^{-2}$  power law. Finally, the red (blue) inset shows a plot of the “arithmetic progression” (“maximum variance”) distribution.

we have  $\Gamma = 0$ , so that the corresponding plot of  $s_{\widehat{\Lambda}_0}^2$  in Fig. 5 is expected to depend on higher-order terms in  $N^{-1/2}$  than  $N^{-1}$ . Indeed we observe a  $N^{-2}$  dependence.

In Fig. 4, the average values of  $\widehat{\Lambda}_0$  are compared with the average values of the estimator proposed by Roulston and corresponding to the coefficient of  $N^{-1}$  in Eq. (40) of Ref. [28]:  $\widehat{\Lambda}_{\text{Roulston}} = \sum_i [\ln(\widehat{p}_i) + \widehat{H}]^2 \widehat{p}_i (1 - \widehat{p}_i)$ . This quantity was evaluated on the same data as  $\widehat{\Lambda}_0$ . The numerical results show that the estimator described in the present work is more reliable than Roulston’s one in assessing the uncertainty of the plug-in estimator  $\widehat{H}$  of Shannon entropy. The estimator proposed by Roulston indeed underestimates the variance of  $\Lambda_0$ : A reason is the fact that Roulston’s expression of the variance is derived via propagation of error by assuming the observed number of visits to be mutually independent. However, as mentioned in Sec. II A, the number of visits are constrained by the expression  $\sum_{i=1}^M \widehat{j}_i = N$ , which leads to the set  $\{\widehat{j}_i\}$  being distributed according to a multinomial distribution. The approach presented above takes into account this crucial property.

V. CONCLUSIONS

The results obtained show that, under the null hypothesis of an underlying multinomial distribution generating the data and provided that the size  $M$  of the support is known, the expected uncertainty of an entropy estimation carried out via the plug-in estimator  $\widehat{H}$  on a set of  $N$  data, and neglecting higher-order terms in  $N^{-1/2}$ , is upper bounded as

$$\sigma_{\widehat{\Lambda}_0} \lesssim \left[ \frac{\Lambda_{0, \max}(M)}{N} \right]^{1/2} \cong \frac{\ln(M)}{2\sqrt{N}}.$$

The upper bound can be then improved by using the plug-in estimator  $\widehat{\Lambda}_0$  defined in Eq. (3). The evaluation of this estimator for different values of  $N$  can deliver reliable estimates of the uncertainty affecting the assessments of Shannon entropy obtained by means of the plug-in estimator  $\widehat{H}$ .

Finally, it is worth noting that, in order to “experimentally visit” all the  $M$  states, the number  $N$  of trials has to exceed  $M$ , so that  $N > M > \sqrt{M}$ . Consequently, the upper bound described above turns out to be smaller, and thus more accurate, than the one derived by Antos and Kontoyiannis [23], namely  $(\ln N)/\sqrt{N}$ .

**APPENDIX: PROOF OF VANISHING  $\Gamma$  IN THE CASE OF  $\Lambda_0 = \Lambda_{0, \max}$**

We show here that, in the case of  $\Lambda_0 = \Lambda_{0, \max}$ , the parameter  $\Gamma$  given by Eq. (9) vanishes. To this purpose, we assume that, for a given  $M$ ,

$$s_i = q_0 \quad \text{if } 1 \leq i \leq M - 1, \\ s_i = p_0 \quad \text{if } i = M,$$

where, according to Eqs. (14) and (15), it is

$$p_0 + (M - 1)q_0 = 1, \tag{A1}$$

$$\ln p_0 + \ln q_0 + 2 + 2H = 0. \tag{A2}$$

Instead of relying on Eq. (9), it is convenient to express  $\Gamma$  as the coefficient of  $t^2/2$  in Eq. (4):

$$\Gamma = \sum_{i=1}^M \sum_{j=1}^M \chi_{i,j} [\ell_i^2 + 2(1 + H)\ell_i][\ell_j^2 + 2(1 + H)\ell_j].$$

By relying on Eq. (A2), it is straightforward to show that each of the two terms  $[\ell_i^2 + 2(1 + H)\ell_i]$ ,  $[\ell_j^2 + 2(1 + H)\ell_j]$  appearing within the previous sum is equal to  $-(\ln p_0)(\ln q_0)$ . One then has

$$\Gamma = (\ln p_0)^2 (\ln q_0)^2 \sum_{i=1}^M \sum_{j=1}^M \chi_{i,j}.$$

The sum, and thus  $\Gamma$ , is equal to zero due to the covariance matrix  $\chi$  having vanishing column and row sums [33]. Alternatively, one can compute the sums by taking into account Eq. (A1).

---

[1] C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).  
 [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (John Wiley & Sons, Hoboken, NJ, 2006).  
 [3] C. Bandt and B. Pompe, *Phys. Rev. Lett.* **88**, 174102 (2002).  
 [4] A. Delgado-Bonal and A. Marshak, *Entropy* **21**, 541 (2019).  
 [5] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, *Phys. Rev. Lett.* **80**, 197 (1998).  
 [6] A. G. Dimitrov, A. A. Lazar, and J. D. Victor, *J. Comput. Neurosci.* **30**, 1 (2011).  
 [7] K. Sayood, *Entropy* **20**, 706 (2018).  
 [8] S. Vinga, *Brief. Bioinform.* **15**, 376 (2014).  
 [9] E. V. Koonin, *Philos. Trans. R. Soc. A* **374**, 20150065 (2016).  
 [10] P. Chanda, E. Costa, J. Hu, S. Sukumar, J. Van Hemert, and R. Walia, *Entropy* **22**, 627 (2020).  
 [11] W. W. Burggren and M. G. Monticino, *J. Exp. Biol.* **208**, 3221 (2005).  
 [12] G. A. Darbellay and D. Wuertz, *Phys. A* **287**, 429 (2000).  
 [13] C. A. Zapart, *Phys. A* **388**, 1157 (2009).  
 [14] L. Leung and G. R. North, *J. Climate* **3**, 5 (1990).  
 [15] A. J. Majda and B. Gershgorin, *Proc. Natl. Acad. Sci. USA* **107**, 14958 (2010).  
 [16] G. Balasis, R. V. Donner, S. M. Potirakis, J. Runge, C. Papadimitriou, I. A. Dagleis, K. Eftaxias, and J. Kurths, *Entropy* **15**, 4844 (2013).  
 [17] J. F. Wellmann, *Entropy* **15**, 1464 (2013).  
 [18] R. A. P. Perdigão, U. Ehret, K. H. Knuth, and J. Wang, *Water Resour. Res.* **56**, e2019WR025270 (2020).  
 [19] P. Grassberger, *Phys. Lett. A* **128**, 369 (1988).  
 [20] J. A. Pardo, *Appl. Math. Comput.* **72**, 33 (1995).  
 [21] A. Lesne, J.-L. Blanc, and L. Pezard, *Phys. Rev. E* **79**, 046208 (2009).  
 [22] L. Paninski, *Neural Comput.* **15**, 1191 (2003).  
 [23] A. Antos and I. Kontoyiannis, *Random Struct. Alg.* **19**, 163 (2001).  
 [24] G. P. Basharin, *Theory Probab. Appl.* **4**, 333 (1959).  
 [25] G. Miller, Note on the bias of information estimates, in *Information theory in Psychology II-B* (Free Press, Glencoe, IL, 1955), pp. 95–100.  
 [26] B. Harris, The statistical estimation of entropy in the non-parametric case, in *Topics in Information Theory* (North-Holland, Amsterdam, 1975), pp. 323–355.  
 [27] M. Vinck, F. P. Battaglia, V. B. Balakirsky, A. J. H. Vinck, and C. M. A. Pennartz, *Phys. Rev. E* **85**, 051139 (2012).  
 [28] M. S. Roulston, *Phys. D* **125**, 285 (1999).  
 [29] D. Kugiumtzis, *Eur. Phys. J.: Spec. Top.* **222**, 401 (2013).  
 [30] L. Ricci, *Phys. Rev. E* **103**, 022215 (2021).  
 [31] R. Courant and D. Hilbert, *Methods of Mathematical Physics* (Interscience, New York, 1953), Vol. 1.  
 [32] R. A. Adams and C. Essex, *Calculus: A Complete Course*, 7th ed. (Pearson Canada, Toronto, 1953).  
 [33] L. Ricci, *Chaos Soliton. Fract.* **142**, 110450 (2021).