Regularities Unseen, Randomness Observed:
Levels of Entropy Convergence

James P. Crutchfield
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
Electronic Address: chaos@santafe.edu

David P. Feldman
College of the Atlantic, 105 Eden St., Bar Harbor, ME 04609
and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
Electronic Address: dpf@santafe.edu

(February 8, 2001)

We study how the Shannon entropy of sequences produced by an information source converges to the source's entropy rate. We synthesize several phenomenological approaches to applying information theoretic measures of randomness and memory to stochastic and deterministic processes by using successive derivatives of the Shannon entropy growth curve. This leads, in turn, to natural measures of apparent memory stored in a source and the amounts of information that must be extracted from observations of a source in order for it to be optimally predicted and for an observer to synchronize to it. One consequence of ignoring these structural properties is that the missed regularities are converted to apparent randomness. We demonstrate that this problem arises particularly for small data sets; e.g., in settings where one has access only to short measurement sequences.

PACS: 02.50.Ey 05.45.-a 05.45.Tp 89.75.Kd Santa Fe Institute Working Paper 01-02-012

Contents

I. INTRODUCTION

A. Apparent Randomness

Natural processes appear unpredictable to varying degrees and for several reasons. First, and most obviously, one may not know the "rules" or equations that govern a particular system. That is, an observer may have only incomplete knowledge of the forces controlling a process. Laplace was well aware of these sources of apparent randomness; as he commented two centuries ago in motivating his *Philosophical Essay on Probabilities* [1]:

> But ignorance of the different causes involved in the production of events, ... taken together with the imperfection of analysis, prevents our reaching the same certainty about the vast majority of phenomena. Thus there are things that are uncertain for us, things more or less probable, and we seek to compensate for the impossibility of knowing them by determining their different degrees of likelihood.

Second, there may be mechanisms intrinsic to a process that amplify unknown or uncontrolled fluctuations to unpredictable macroscopic behavior. Manifestations of this sort of randomness include *deterministic chaos* and *fractal separatrix* structures bounding different basins of attraction. As Poincarè noted [2]:

> ... it may happen that small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible, and we have the fortuitous phenomenon.

Unpredictability of this kind also arises from sensitive dependence on parameters, such as that seen in non-structurally stable systems with continuous bifurcations [3] or from sensitive dependence on boundary conditions. Knowledge of the governing equations of motion does little to make these kinds of intrinsic randomness go away.

Third, and more subtly, there exists a wide array of observer-induced sources of apparent randomness. For one, the choice of representation used by the observer may render a system unpredictable. For example, representing a square wave in terms of sinusoids requires specifying an infinite number of amplitude coefficients. Truncating the order of approximation leads to errors, even for a source as simple and predictable as a square wave. Similarly, an observer's choice and design of its measuring instruments is an additional source of apparent randomness. As one example, Ref. [4] shows how irreducible unpredictability arises from a measurement instrument's distortion of a spatio-temporal process's internal states.

Fourth, the measurement process engenders apparent randomness in other, perhaps more obvious ways, too. Even if one knows the equations of motion governing a system, accurate prediction may not be possible: the measurements made by an observer may be inaccurate, or, if the measurements are precise, there may be an insufficient volume of measurement data. Or, one may simply not have a sufficiently long measurement stream, for example, to disambiguate several internal states and, therefore, their individual consequences for the process's future behavior cannot be accurately accounted for. Examples of these sorts of measurement-induced randomness are considered in Refs. [5–7]. In all of these cases, the result is that the process appears more random than it actually is.

Fifth, and finally, if the dynamics are sufficiently complicated it may simply be too computationally difficult to perform the calculations required to go from measurements of the system to a prediction of the system's future behavior. The existence of deeply complicated dynamics for which this was a problem was first appreciated by Poincaré more than a century ago as part of his detailed analysis of the three-body problem [8].

Of course, most natural phenomena involve, to one degree or another, almost all of these separate sources of "noise". Moreover, the different mechanisms interact with each other. It is no surprise, therefore, that describing and quantifying the degree of a process's apparent randomness is a difficult yet essential endeavor that cuts across many disciplines.

B. Untangling the Mechanisms

A central goal here is to examine ways to untangle the different mechanisms responsible for apparent randomness by investigating several of their signatures. As one step in addressing these issues, we analyze those aspects of apparent randomness over which an observer may have some control. These include the choice of how to quantify the degree of randomness (e.g., through choices of statistic or in modeling representation) and how much data to collect. We describe the stance taken by the observer toward the process to be analyzed in terms of the *measurement channel* — an adaptation [9] of Shannon's notion of a communication channel. One of the central questions addressed in the following is, how does an observer, apprised of a process's possible states and its dynamics, come to know in what internal state the process is? We will show that this is related to another question, How does an observer come to accurately estimate how random a source is? In particular, we shall investigate how finite-data approximations converge to this asymptotic value. We shall see a variety of different convergence behaviors and will present several different quantities that capture the nature of this convergence. As the title of this work suggests, we shall see that regularities that are unseen are "converted" to apparent randomness.

It is important to emphasize, and this will be clear through our citations, that much of our narrative about levels of entropy convergence touches on and restates results and intuitions known to a number of researchers in information theory, dynamical systems, stochastic processes, and symbolic dynamics. Our attempt here, in light of this, is several-fold. First, we put this knowledge into a single framework, using the language of discrete derivatives and integrals. We believe this approach unifies and clarifies a number of extant quantities. Second, and more importantly, by considering numerous examples, we shall see that examining levels of entropy convergence can give important clues about the computational structure of a process. Finally, our view of entropy convergence will lead naturally to a new quantity, the *transient information* $\mathbf{T}$. We shall prove that the transient information captures the total uncertainty an observer must overcome in synchronizing to a Markov process.

We begin in Sec. II by fixing notation and briefly reviewing the motivation and basic quantities of information theory. In Sections III and IV we use discrete derivatives and integrals to examine entropy convergence. In so doing, we recover a number of familiar measures of randomness, predictability, and "complexity". Then, in Sec. V we introduce, motivate, and interpret a new information theoretic measure of structure, the transient information. In particular, we shall see that the transient information provides a quantitative measure of the manner in which an observer synchronizes to a source. We then illustrate the utility of the quantities discussed in Sec. III–V by considering a series of increasingly rich examples in Sec. VI. In Sec. VII we look at relationships between the quantities discussed previously. In particular, we show several quantitative examples of how regularities that go undetected are converted into apparent randomness. Finally, we conclude in Sec. VIII and offer thoughts on possible future directions for this line of research.

II. INFORMATION THEORY

A. The Measurement Channel

In the late 1940's Claude Shannon founded the field of communication theory [10], motivated in part by his work in cryptography during World War II [11]. His attempt to analyze the basic trade-offs in disguising information from third parties in ways that still allowed recovery by the intended receiver led to a study of how signals could be compressed and transmitted efficiently and error free. His basic conception was that of a *communication channel* consisting of an *information source* which produces *messages* that are encoded and passed through a possibly noisy and error-prone channel. A *receiver* then decodes the channel's output in order to recover the original messages. Shannon's main assumptions were that an information source was described by a distribution over its possible messages and that, in particular, a message was "informative" according to how surprising or unlikely its occurrence was.

We adapt Shannon's conception of a communication channel as follows: We assume that there is a *process* (source) that produces a *data stream* (message) — an infinite string of symbols drawn from some finite alphabet. The task for the *observer* (receiver) is to estimate the probability distribution of sequences and, thereby, estimate how random the process is. Further, we assume that the observer does not know the process's structure; the range of its states and their transition structure — the process's internal dynamics — are hidden from the observer. (We will, however, occasionally relax this assumption below.) Since the observer does not have direct access to the source's internal, hidden states, we picture instead that the observer can estimate to arbitrary accuracy the probability of measurement sequences. Thus, we do not address the eminently practical issue of how much data is required for accurate estimation of these probabilities. For this see, for example, Refs. [7,12,13]. In our scenario, the observer detects sequence blocks directly and stores their probabilities as histograms. Though apparently quite natural in this setting, one should consider the histogram to be a particular class of representation for the source's internal structure — one that may or may not correctly capture that structure.

This *measurement channel* scenario is illustrated in Fig. 1. In this case, the source is a three-state deterministic finite automaton. However, the observer does not see the internal states $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$. Instead, it has access to only the measurement symbols $\{0, 1\}$ generated on state-to-state transitions by the hidden automaton. In this sense, the measurement channel acts like a communication channel; the channel maps from a internal-state sequence $\ldots\mathbf{BCBAACBC}\ldots$ to a measurement sequence $\ldots 0111010\ldots$. The process shown in Fig. 1 belongs to the class of stochastic process known as *hidden Markov models*. The transitions from internal state to internal state are Markovian, in that the probability of a given transition depends only upon which state the process is currently in. However, these internal states are not seen by the observer — hence the name "hidden" Markov model [14,15].

Given this situation, a number of issues arise for the observer. One fundamental question is how many of the system's properties can be inferred from the observed binary data stream. In particular, can the observer build a model of the system that allows for accurate prediction? According to Shannon's coding theorem, success in answering these questions depends on whether the system's entropy rate falls below the measurement channel capacity. If it does, then the observer can build a model of the system. Conversely, if the entropy rate is above the measurement channel's capacity, then the theorem tells us that the observer cannot exactly reconstruct all properties of the system. In this case, source messages — sequences over internal states — cannot be decoded in an error-free manner. In particular, optimal prediction will not be possible. In the following, we assume that the channel capacity is larger than the entropy rate and, hence, that optimal prediction is — in theory, at least — possible.

Similar questions of building models from data produced by various kinds of information sources are found in the fields of machine learning and computational learning theory. See the appendices in Ref. [16] for comments on the similarities and differences with the approach taken here.

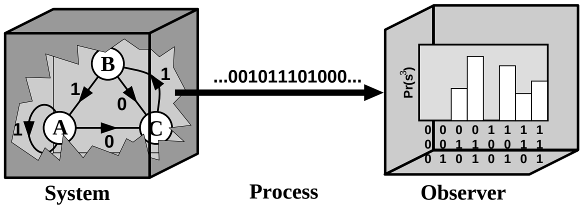


FIG. 1. The measurement channel: The internal states $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ of the system are reflected, only indirectly, in the observed measurement of 1's and 0's. An observer works with this impoverished data to build a model of the underlying system. After Ref. [17].

B. Stationary Stochastic Processes

The measurement streams we shall consider will be stationary stochastic processes. In this section we introduce this idea more formally, fix notation, and define a few classes of stochastic process to which we shall return when considering examples in Sec. VI.

The main object of our attention will be a one-dimensional chain

$$\overleftrightarrow{S} \equiv \ldots S_{-2} S_{-1} S_0 S_1 \ldots \tag{1}$$

of random variables S_t that range over a finite set $\mathcal{A}$. We assume that the underlying system is described by a shift-invariant measure μ on infinite sequences $\cdots s_{-2} s_{-1} s_0 s_1 s_2 \cdots ; s_t \in \mathcal{A}$ [18]. The measure μ induces a family of distributions, $\{\Pr(s_{t+1}, \ldots, s_{t+L}) : s_t \in \mathcal{A}\}$, where $\Pr(s_t)$ denotes the probability that at time t the random variable S_t takes on the particular value $s_t \in \mathcal{A}$ and $\Pr(s_{t+1}, \ldots, s_{t+L})$ denotes the joint probability over blocks of L consecutive symbols. We assume that the distribution is stationary; $\Pr(s_{t+1}, \ldots, s_{t+L}) = \Pr(s_1, \ldots, s_L)$.

We denote a block of L consecutive variables by $S^L \equiv S_1 \ldots S_L$. We shall follow the convention that a capital letter refers to a random variable, while a lowercase letter denotes a particular value of that variable. Thus, $s^L = s_0 s_1 \cdots s_{L-1}$, denotes a particular symbol block of length L. We shall use the term *process* to refer to the joint distribution $\Pr(\overleftrightarrow{S})$ over the infinite chain of variables. A process, defined in this way, is what Shannon referred to as an information source.

For use later on, we define several types of processes. First, and most simply, a process with a *uniform distribution* is one in which all sequences occur with equiprobability. We will denote this distribution by U^L;

$$U(s^L) \;=\; 1/|\mathcal{A}|^L \;. \tag{2}$$

Next, a process is *independently and identically distributed* (IID) if the joint distribution $\Pr(\overleftrightarrow{S}) = \Pr(\ldots, S_i, S_{i+1}, S_{i+2}, S_{i+3}, \ldots)$ factors in the following way:

$$\Pr(\overleftrightarrow{S}) \;=\; \ldots \Pr(S_i) \Pr(S_{i+1}) \Pr(S_{i+2}) \ldots \;, \tag{3}$$

and $\Pr(S_i) = \Pr(S_j)$ for all i, j.

We shall call a process *Markovian* if the probability of the next symbol depends only on the previous symbol seen. In other words, the joint distribution factors in the following way:

$$\Pr(\overleftrightarrow{S}) = \ldots \Pr(S_{i+1}|S_i) \Pr(S_{i+2}|S_{i+1}) \ldots \tag{4}$$

More generally, a process is *order-R Markovian* if the probability of the next symbol depends only on the previous R symbols:

$$\Pr(S_i | \ldots, S_{i-3}, S_{i-1}) = \Pr(S_i | S_{i-R}, \ldots, S_{i-1}) \;. \tag{5}$$

Finally, a *hidden Markov process* consists of an internal order-R Markov process that is observed only by a function of its internal-state sequences. These are sometimes called *functions of a Markov chain* [14,15]. We refer to all of these processes as *finitary*, since there is a well defined sense, discussed below, in which they have a finite amount of memory.

C. Basic Quantities of Information Theory

Here, we briefly state the definitions and interpretations of the basic quantities of information theory. For more details, see Ref. [19]. Let X be a random variable that assumes the values $x \in \mathcal{X}$, where $\mathcal{X}$ is a finite set. We denote the probability that X assumes the particular value x by $\Pr(x)$. Likewise, let Y be a random variable that assumes the values $y \in \mathcal{Y}$.

The *Shannon entropy* of X is defined by:

$$H[X] \equiv - \sum_{x \in \mathcal{X}} \Pr(x) \log_2 \Pr(x) \ . \tag{6}$$

Note that $H[X] \geq 0$. The units of $H[X]$ are *bits*. The entropy $H[X]$ measures the uncertainty associated with the random variable X. Equivalently, it measures the average amount of memory, in bits, needed to store outcomes of the variable X. The *conditional entropy* is defined by

$$H[X|Y] \equiv - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \Pr(x, y) \log_2 \Pr(x|y) \ , \tag{7}$$

and measures the average uncertainty associated with variable X, if we know Y.

The *mutual information* between X and Y is defined as

$$I[X;Y] \equiv H[X] - H[X|Y] \ . \tag{8}$$

In words, the mutual information is the average reduction of uncertainty of one variable due to knowledge of another. If knowing Y on average makes one more certain about X, then it makes sense to say that Y carries information about X. Note that $I[X;Y] \geq 0$ and that $I[X;Y] = 0$ when either X and Y are independent (there is no "communication" between X and Y) or when either $H[X] = 0$ or $H[Y] = 0$ (there is no information to share). Note also that $I[X;Y] = I[Y;X]$.

The *information gain* between two distributions $\Pr(x)$ and $\widehat{\Pr}(x)$ is defined by:

$$\mathcal{D}[\Pr(x) || \widehat{\Pr}(x)] \equiv \sum_{x \in \mathcal{X}} \Pr(x) \log_2 \frac{\Pr(x)}{\widehat{\Pr}(x)} \ , \tag{9}$$

where $\widehat{\Pr}(x) = 0$ only if $\Pr(x) = 0$. Quantitatively, $\mathcal{D}[P||Q]$ is the number of bits by which the two distributions P and Q differ [19]. Informally, $\mathcal{D}[P||Q]$ can be viewed as the distance between P and Q in a space of distributions. However, $\mathcal{D}[P||Q]$ is not a metric, since it does not obey the triangle inequality.

Similarly, the *conditional entropy gain* between two conditional distributions $\Pr(x|y)$ and $\widehat{\Pr}(x|y)$ is defined by:

$$\mathcal{D}[\Pr(x|y) || \widehat{\Pr}(x|y)] \equiv \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \Pr(x, y) \log_2 \frac{\Pr(x|y)}{\widehat{\Pr}(x|y)} \ , \tag{10}$$

D. Block Entropy and Entropy Rate

We now examine the behavior of the Shannon entropy $H(L)$ of $\Pr(s^L)$, the distribution over blocks of L consecutive variables. We shall see that examining how the Shannon entropy of a block of variables grows with L leads to several quantities that capture aspects of a process's randomness and different features of its memory.

The *total Shannon entropy* of length-L sequences is defined

$$H(L) \ \equiv \ - \sum_{s^L \in \mathcal{A}^L} \Pr(s^L) \log_2 \Pr(s^L) \ , \tag{11}$$

where $L > 0$. The sum is understood to run over all possible blocks of L consecutive symbols. If no measurements are made, there is nothing about which to be uncertain and, thus, we define $H(0) \equiv 0$. Below we will show that $H(L)$ is a non-decreasing function of L; $H(L) \geq H(L-1)$. We shall also see that it is concave; $H(L) - 2H(L-1) + H(L-2) \leq 0$.

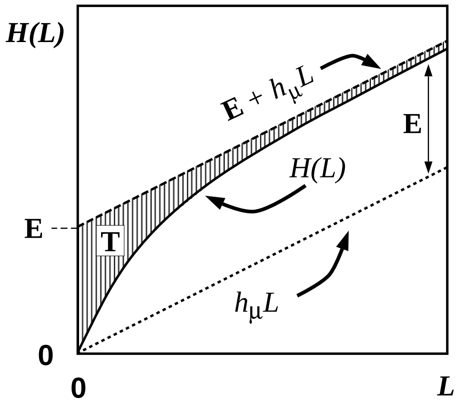


FIG. 2. Total Shannon entropy growth for a finitary information source: a schematic plot of $H(L)$ versus L. $H(L)$ increases monotonically and asymptotes to the line $\mathbf{E} + h_\mu L$, where $\mathbf{E}$ is the excess entropy and h_μ is the source entropy rate. This dashed line is the $\mathbf{E}$-memoryful Markovian source approximation to a source with entropy growth $H(L)$. The entropy growth of the memoryless-source approximation of the source is indicated by the short-dashed line $h_\mu L$ through the origin with slope h_μ. The shaded area is the transient information $\mathbf{T}$. For more discussion, see text.

Note that the maximum average information per observation is $\log_2 |\mathcal{A}|$, $H(1) \leq \log_2 |\mathcal{A}|$, and, more generally,

$$H(L) \leq L \ \log_2 |\mathcal{A}| \ . \tag{12}$$

Equality in Eq. (12) occurs only when the distribution over L-blocks is uniform; i.e., given by U^L. Figure 2 shows $H(L)$ for a typical information source. The various labels and the interpretation of $H(L)$ there will be discussed fully below.

The *source entropy rate* h_μ is the rate of increase with respect to L of the total Shannon entropy in the large L limit:

$$h_\mu \equiv \lim_{L\to\infty} \frac{H(L)}{L} \ , \tag{13}$$

where μ denotes the measure over infinite sequences that induces the L-block joint distribution $\Pr(s^L)$; the units are *bits/symbol*. The limit in Eq. (13) exists for all stationary measures μ [19]. The entropy rate h_μ quantifies the irreducible randomness in sequences produced by a source: the randomness that remains after the correlations and structures in longer and longer sequence blocks are taken into account. The entropy rate is also known as the *thermodynamic entropy density* in statistical mechanics or the *metric entropy* in dynamical systems theory.

As Shannon proved in his original work, h_μ also measures the length, in bits per symbol, of the optimal, uniquely decodable, binary encoding for the measurement sequence. That is, a message of L symbols requires (as $L \to \infty$) only $h_\mu L$ bits of information rather than $\log_2|\mathcal{A}|L$ bits. This is consonant with the idea of h_μ as a measure of randomness. On the one hand, a process that is highly random, and hence has large h_μ, is difficult to compress. On the other hand, a process with low h_μ has many correlations between symbols that can be exploited by an efficient coding scheme.

As noted above, the limit in Eq. (13) is guaranteed to exist for all stationary sources. In other words,

$$H(L) \sim h_\mu L \ \text{ as } L \to \infty \ . \tag{14}$$

However, knowing the value of h_μ indicates nothing about how $H(L)/L$ approaches this limit. Moreover, there may be — and indeed usually are — sublinear terms in $H(L)$. For example, one may have $H(L) \sim c + h_\mu L$ or $H(L) \sim \log L + h_\mu L$. We shall see below that the sublinear terms in $H(L)$ and the manner in which $H(L)$ converges to its asymptotic form reveal important structural information about a process.

E. Redundancy

Before moving on to our main task — considering what can be learned from looking at the entropy growth curve $H(L)$ — we introduce one additional quantity from information theory. Since we are using an alphabet of size $|\mathcal{A}|$, if nothing else is known about the process or the channel we can consider the measurement channel used to observe the process to have a *channel capacity* of $\mathcal{C} = \log_2|\mathcal{A}|$. Said another way, the maximum observable entropy rate for the channel output (the measurement sequence) is $\log_2|\mathcal{A}|$.

Frequently, however, the observed h_μ is less than its maximum value. This difference is measured by the *redundancy* $\mathbf{R}$:

$$\mathbf{R} \equiv \log_2|\mathcal{A}| - h_\mu \ . \tag{15}$$

Note $\mathbf{R} \geq 0$. If $\mathbf{R} > 0$, then the series of random variables $\ldots, S_i, S_{i+1}, \ldots$ has some degree of regularity: either the individual variables are biased in some way or there are correlations between them. Recall that the entropy rate measures the size, in bits per symbol, of the optimal binary compression of the source. The redundancy, then, measures the amount by which a given source can be compressed. If a system is highly redundant, it can be compressed a great deal.

For another interpretation of the redundancy, one can show that $\mathbf{R}$ is the information gain of the source's actual distribution $\Pr(s^L)$ with respect to the uniform distribution $\mathrm{U}(s^L)$ in the $L \to \infty$ limit:

$$\mathbf{R} = \lim_{L\to\infty} \frac{\mathcal{D}[\Pr(s^L)||\mathrm{U}(s^L)]}{L} \ , \tag{16}$$

where $\mathcal{D}$ is defined in Eq. (9). Restated, then, the redundancy $\mathbf{R}$ is a measure of the information gained when an observer, expecting a uniform distribution, learns the actual distribution over the sequence.

III. LEVELS OF ENTROPY CONVERGENCE: DERIVATIVES OF $H(L)$

With these preliminaries out of the way, we are now ready to begin the main task: examining the growth of the entropy curve $H(L)$. In particular, we shall look carefully at the manner in which the block entropy $H(L)$ converges to its asymptotic form — an issue that has occupied the attention of many researchers [5,6,12,13,20–41]. In what follows, we present a systematic method for examining entropy convergence. To do so, we will take discrete derivatives of $H(L)$ and also form various integrals of these derivatives. This method allows one to recover a number of quantities that have been introduced some years ago and that can be interpreted as different aspects of a system's memory or structure. Additionally, our discrete derivative framework will lead us to define a new quantity, the *transient information*, which may be interpreted as a measure of how difficult it is to synchronize to a source, in a sense to be made precise below.

Before continuing, we pause to note that the representation shown in the entropy growth curve of Fig. 2 of a finitary process is *phenomenological*, in the sense that $H(L)$ and the other quantities indicated derive only from the observed distribution $\Pr(s^L)$ over sequences. In particular, they do not require any additional or prior knowledge of the source and its internal structure.

A. Discrete Derivatives and Integrals

We begin by briefly collecting some elementary properties of discrete derivatives. Consider an arbitrary function $F : \mathbb{Z} \rightarrow \mathbb{R}$. In what follows, the function F will be the Shannon block entropy $H(L)$, but for now we consider general functions. The discrete derivative is the linear operator defined by:

$$(\Delta F)(L) \equiv F(L) - F(L-1) . \tag{17}$$

The picture is that the operator Δ acts on F to produce a new function ΔF which, when evaluated at L, yields $F(L) - F(L-1)$. Higher-order derivatives are defined by composition:

$$\Delta^n F \equiv (\Delta \circ \Delta^{n-1}) F , \tag{18}$$

where $\Delta^0 F \equiv F$ and $n \geq 1$. For example, the second discrete derivative is given by:

$$\begin{aligned} \Delta^2 F(L) &\equiv (\Delta \circ \Delta) F(L) & (19) \\ &= F(L) - 2F(L-1) + F(L-2) . & (20) \end{aligned}$$

One "integrates" a discrete function $\Delta F(L)$ by summing:

$$\sum_{L=A}^{B} \Delta F(L) = F(B) - F(A-1) . \tag{21}$$

An integration-by-parts formula also holds:

$$\begin{aligned} \sum_{L=A}^{B} L \Delta F(L) &= B F(B) - \\ & A F(A-1) - \sum_{L=A}^{B-1} F(L) . \end{aligned} \tag{22}$$

Note the shift in the sum's limits on the right-hand side.

B. $\Delta H(L)$: Entropy Gain

We now consider the effects of applying the discrete derivative operator Δ to the entropy growth curve $H(L)$. We begin with the first derivative of $H(L)$:

$$\Delta H(L) \equiv H(L) - H(L-1) , \tag{23}$$

where $L > 0$. The units of $\Delta H(L)$ are *bits/symbol*. A plot of a typical $\Delta H(L)$ vs. L is shown in Fig. 3. We refer to $\Delta H(L)$ as the *entropy gain* for obvious reasons.

If a measurement has not yet been made, the apparent entropy rate is maximal. Thus, we define $\Delta H(0) = \log_2 |\mathcal{A}|$. In a Bayesian modeling setting this is equivalent to being told only that the source has $|\mathcal{A}|$ symbols and then assuming the process is independent identically distributed and uniformly distributed over individual symbols.

Having made a single measurement in each experiment in an ensemble or, equivalently, only looking at single-symbol statistics in one experiment, the entropy gain is the single-symbol Shannon entropy: $\Delta H(1) = H(1) - H(0) = H(1)$, since we defined $H(0) = 0$.

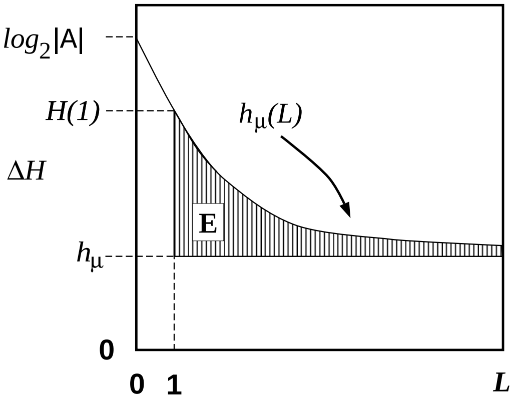


FIG. 3. Entropy-rate convergence: A schematic plot of $h_\mu(L) = \Delta H(L)$ versus L using the finitary process's $H(L)$ shown in Fig. 2. The entropy rate asymptote h_μ is indicated by the lower horizontal dashed line. The shaded area is the excess entropy $\mathbf{E}$.

Let's now look at some properties of $\Delta H(L)$.

Proposition 1 *$\Delta H(L)$ is an information gain:*

$$\Delta H(L) = \mathcal{D}[\Pr(s^L) || \Pr(s^{L-1})] , \tag{24}$$

where $L > 1$.
Proof: Since many of the proofs are straightforward, direct calculations, we have put most of them in Appendix A so as not to interrupt the flow of ideas in the main sections. Proposition 1 is proved in App. A 1.□

Note that Eq. (24) is a slightly different form for the information gain than that defined in Eq. (9). Unlike Eq. (9), in Eq. (24) the two distributions do not have the same support: one $\{s^L\}$ is a refinement of the other $\{s^{L-1}\}$. When this is the case, we extend the length $L-1$ distribution to a distribution over length L sequences by concatenating the symbols s_{L-1} with equal probability onto $s_0, \ldots, s_{L-2}$. We then sum the terms in $\mathcal{D}$ over the set of length L sequences.

Note that since the information gain is a non-negative quantity [19], it follows from Prop. 1 that $\Delta H(L) \equiv H(L) - H(L-1) \geq 0$, as remarked earlier. In a subsequent section, we shall see that $\Delta^2 H(L) \leq 0$; hence, $\Delta H(L)$ is monotone decreasing.

The derivative $\Delta H(L)$ may also be written as a conditional entropy. Since

$$\frac{\Pr(s^L)}{\Pr(s^{L-1})} = \Pr(s_L | s^{L-1}) . \tag{25}$$

it immediately follows from Eq. (23) that

$$\Delta H(L) \;=\; H[S_L|S^{L-1}] \;. \tag{26}$$

This observation helps strengthen our interpretation of h_μ. Recall that the entropy rate h_μ was defined in Eq. (13) as $\lim_{L\to\infty} H(L)/L$. As is well known (see, e.g., Ref. [19]), the entropy rate may also be written as:

$$h_\mu \;=\; \lim_{L\to\infty} H[S_L|S^{L-1}] \;. \tag{27}$$

That is, h_μ is the average uncertainty of the variable S_L, given that an arbitrarily large number of preceding symbols have been seen.

By virtue of Eq. (26), we see that

$$h_\mu = \lim_{L\to\infty} \Delta H(L) \;. \tag{28}$$

Following Refs. [28–30,39,40], we denote $\Delta H(L)$ by $h_\mu(L)$:

$$\begin{aligned} h_\mu(L) &\equiv \Delta H(L) \\ &\equiv H(L) - H(L-1) \;,\; L \geq 1 \;. \end{aligned} \tag{29}$$

The function $h_\mu(L)$ is the estimate of how random the source appears if only blocks of variables up to length L are considered. Thus, $h_\mu(L)$ may be thought of as a finite-L approximation to the entropy rate h_μ — the *apparent entropy rate* at length L. Alternatively, the entropy rate h_μ can be estimated for finite L by appealing to its original definition [42], i.e., Eq. (13). We thus define another finite-L entropy rate estimate:

$$h'_\mu(L) \equiv \frac{H(L)}{L} \;,\; L \geq 1 \;, \tag{30}$$

where we also take $h'_\mu(0) \equiv \log_2 \mathcal{A}$. Note that while we have

$$\lim_{L\to\infty} h'_\mu(L) = \lim_{L\to\infty} h_\mu(L) \;, \tag{31}$$

in general, it is the case that

$$h'_\mu(L) \;\neq\; h_\mu(L) \;,\; L < \infty \;. \tag{32}$$

Moreover, $h'_\mu(L)$ converges more slowly than $h_\mu(L)$. (The examples later illustrate the slow convergence.)

Lemma 1 *:*

$$h'_\mu(L) \geq h_\mu(L) \geq h_\mu \;. \tag{33}$$

Proof: See App. A 8. □

C. Entropy Gain and Redundancy

The entropy gain can also be interpreted as a type of redundancy. To see this, first recall that the redundancy, Eq. (15), is the difference between $\log_2|\mathcal{A}|$ and h_μ, where $\log_2|\mathcal{A}|$ is the entropy given no knowledge of the source apart from the alphabet size, and h_μ is the entropy of the source given knowledge of the distribution of arbitrarily large L-blocks. But what is the redundancy if the observer already knows the actual distribution $\Pr(s^L)$ of words up to length L?

This question is answered by the L-redundancy:

$$\mathbf{R}(L) \;\equiv\; H(L) - h_\mu L \;. \tag{34}$$

Here, $H(L)$ is the entropy given that $\Pr(s^L)$ is known, and the product $h_\mu L$ is the entropy of an L-block if one uses only the asymptotic form of $H(L)$ given in Eq. (14). Note that $\mathbf{R}(L) \leq \mathbf{R}$, where $\mathbf{R}$ is defined in Eq. (15)

We now define the per-symbol L-redundancy:

$$\mathbf{r}(L) \;\equiv\; \Delta\mathbf{R}(L) \;\equiv\; h_\mu(L) - h_\mu \;. \tag{35}$$

The quantity $\mathbf{r}(L)$ gives the difference between the per-symbol entropy conditioned on L measurements and the per-symbol entropy conditioned on an infinite number of measurements. In other words, $\mathbf{r}(L)$ measures the extent to which the length-L entropy rate estimate exceeds the actual per-symbol entropy. Any difference indicates that there is redundant information in the L-blocks in the amount of $\mathbf{r}(L)$ bits. Ebeling [38] refers to $r(L)$ as the local (i.e., L-dependent) predictability.

D. $\Delta^2 H(L)$: Predictability Gain

If we interpret $h_\mu(L)$ as an estimate of the source's unpredictability and recall that it decreases monotonically to h_μ, we can look at $\Delta^2 H(L)$ — the rate of change of $h_\mu(L)$ — as the rate at which unpredictability is lost. Equivalently, we can view $-\Delta^2 H(L)$ as the improvement in our predictions in going from $L-1$ to L blocks. This is the change in the entropy rate estimate $h_\mu(L)$ and is given by the *predictability gain*:

$$\Delta^2 H(L) \equiv \Delta h_\mu(L) \;=\; h_\mu(L) - h_\mu(L-1) \;, \tag{36}$$

where $L > 0$; the units of $\Delta^2 H(L)$ are *bits/symbol*^{2}. (See Fig. 4.) Since we defined $h_\mu(0) \equiv \log_2|\mathcal{A}|$, we have that

$$\Delta^2 H(1) = H(1) - \log_2|\mathcal{A}| \;. \tag{37}$$

The quantity $\Delta^2 H(0)$ is not defined.

A large value of $|\Delta^2 H(L)|$ indicates that going from statistics over $(L-1)$-blocks to L-blocks reduces the uncertainty by a large amount. Speaking loosely, we shall see in Sec. VI that a large value of $|\Delta^2 H(L)|$ suggests that the L^{th} measurement is particularly informative.

Proposition 2 $\Delta^2 H(L)$ *is a conditional information gain:*

$$\Delta^2 H(L) = -\mathcal{D}[\Pr(s_{L-1}|s^{L-2})||\Pr(s_{L-2}|s^{L-3})] \ , \quad (38)$$

for $L > 3$.

Proof: See App. A 2. □

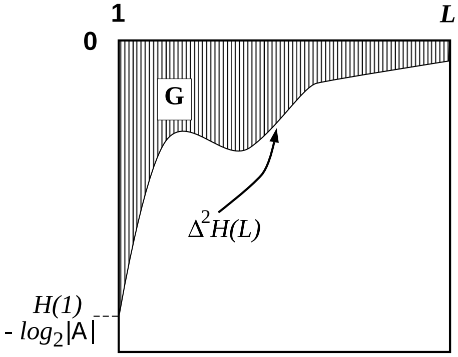


FIG. 4. Predictability gain: A schematic plot of $\Delta^2 H(L)$ versus L using the "typical" $h_\mu(L)$ shown in Fig. 3. The shaded area is the total predictability $\mathbf{G}$.

Since the information gain is non-negative, it follows from Prop. 2 that $\Delta^2 H(L) \leq 0$ and so $H(L)$ is a concave function of L.

The observation contained in Prop. 2 first appeared in Refs. [43] and [34]. There, $-\Delta^2 H(L)$ is referred to as the *correlation information*. However, we feel that the term "predictability gain" is a more accurate name for this quantity. The quantity $-\Delta^2 H(L)$ measures the reduction in per-symbol uncertainty in going from $(L-1)$- to L-block statistics. While $-\Delta^2 H(L)$ is related to the correlation between symbols L time steps apart, it does not directly measure their correlation. The information theoretic analog of the two-variable correlation function is the mutual information between symbols L steps apart: $I[S_t; S_{t+1}]$, averaged over t. For a discussion of two-symbol mutual information and how they compare with correlation functions, see Refs. [44] and [45].

E. Entropy-Derivative Limits

Ultimately, we are interested in how $H(L)$ and its derivatives converge to their asymptotic values. As we will now show, this question is well posed because the derivatives of $H(L)$ have well defined limiting behavior. First, as mentioned above, for stationary sources, $\lim_{L\to\infty} \Delta H(L) = h_\mu$. An immediate consequence of this is is the following.

Lemma 2 *For stationary processes, the higher derivatives of* $H(L)$ *vanish in the* $L \to \infty$ *limit:*

$$\lim_{L\to\infty} \Delta^n H(L) = 0, \ n \geq 2 \ . \quad (39)$$

Proof. To see this, first recall that the limit $h_\mu = \lim_{L\to\infty} \Delta H(L)$ exists for a stationary source [19] and so the sequence $\Delta H(0), \Delta H(1), \Delta H(2), \ldots$ converges. It follows from this that $\lim_{L\to\infty}[\Delta H(L) - \Delta H(L-1)] = \lim_{L\to\infty} \Delta^2 H(L) = 0$. This proves the $n = 2$ case of Eq. (39). The $n \geq 3$ cases of Eq. (39) then follow via identical arguments.□

To recapitulate, for the finitary processes we are considering in the $L \to \infty$ limit we have that

$$H(L) \sim h_\mu L \ , \quad (40)$$

plus possible sublinear terms. We also have that

$$\lim_{L\to\infty} \Delta H(L) \ = \ h_\mu \ , \quad (41)$$

and

$$\lim_{L\to\infty} \Delta^n H(L) \ = 0 \ , \ \text{ for } n \geq 2 \ . \quad (42)$$

IV. ENTROPY CONVERGENCE INTEGRALS

Since limits at each level of the entropy-derivative hierarchy exist, we can ask *how* the derivatives converge to their limits by investigating the following "integrals":

$$\mathcal{I}_n \equiv \sum_{L=L_{\rm n}}^{\infty} \left[\Delta^n H(L) - \lim_{L\to\infty} \Delta^n H(L) \right] \ . \quad (43)$$

The lower limit $L_{\rm n}$ is taken to be the first value of L at which $\Delta^n H(L)$ is defined. The picture here is that at each L, $\Delta^n H(L)$ over- or under-estimates the asymptotic value $\lim_{L\to\infty} \Delta^n H(L)$ by an amount $\Delta^n H(L) - \lim_{L\to\infty} \Delta^n H(L)$. Summing up all of these estimates provides a measure, perhaps somewhat coarse, of the manner in which an entropy derivative converges to its asymptotic value. The larger the sum, the slower the convergence. The latter, in turn, indicates correlations within larger L sequences, thus suggesting that a process possesses a greater degree of structure — internal structure that is responsible for maintaining the correlations.

A. Predictability

We first examine $\mathcal{I}_2$. Recall that $\lim_{L\to\infty} \Delta^2 H(L) = 0$ and that $\Delta^2 H(L)$ is defined for $L \geq 1$. For reasons that will become clear shortly, we refer to $\mathcal{I}_2$ as the *total predictability* $\mathbf{G}$. It is defined as:

$$\mathbf{G} \ \equiv \ \mathcal{I}_2 \ = \sum_{L=1}^{\infty} \Delta^2 H(L) \ , \quad (44)$$

Geometrically, $\mathbf{G}$ is the area under the $\Delta^2 H(L)$ curve, as shown in Fig. 4. The units of $\mathbf{G}$ are *bits/symbol*, as may be inferred geometrically from Fig. 4, where the units of the horizontal axis are bits and those of the vertical axis are bits/symbol^2. Alternatively, this observation follows directly from Eq. (47), when one takes into account the implied ΔL $(= 1)$ in the sum. An interpretation of $\mathbf{G}$ is established by the following result.

Proposition 3 *[43,34] The magnitude of the total predictability is equal to the redundancy, Eq. (15):*

$$\mathbf{G} = -\mathbf{R} \, . \tag{45}$$

Proof: See App. A 3. □

This establishes an accounting of the maximum possible information $\log_2|\mathcal{A}|$ available from the measurement channel in terms of intrinsic randomness h_μ and total predictability $\mathbf{G}$:

$$\log_2|\mathcal{A}| = |\mathbf{G}| + h_\mu . \tag{46}$$

That is, the raw information $\log_2|\mathcal{A}|$ obtained when making a single-symbol measurement can be considered to consist of two kinds of information: that due to randomness h_μ, on the one hand, and that due to order or redundancy in the process $\mathbf{G}$, on the other hand.

Alternatively, we see that $\mathbf{G} = \log_2|\mathcal{A}| - h_\mu$. Thus, viewing h_μ as measuring the unpredictable component of a process, and recalling that $\log_2|\mathcal{A}|$ is the maximum possible entropy per symbol, it follows that $\mathbf{G}$ measures is the source's predictable component. For this reason we refer to $\mathbf{G}$ as the total predictability. Note that this result turns on *defining* the appropriate boundary condition as $h_\mu(0) = \log_2|\mathcal{A}|$.

There is another form for $\mathbf{G}$ that provides an additional interpretation. The total predictability can be expressed as an average number of measurement symbols, or average length, where the average is weighted by the third derivative, $\Delta^3 H(L)$.

Proposition 4 *The total predictability can be expressed as a type of average length, where the average is weighted by the third derivative, $\Delta^3 H(L)$.*

$$\mathbf{G} = -\sum_{L=2}^{\infty} (L-1) \Delta^3 H(L) \, , \tag{47}$$

when the sum is finite.

Proof: See App. A 4. □

Eq. (47) shows that if $\Delta^3 H(L)$ is slow to converge to 0, then $\mathbf{G}$ will be large. Ignoring dimensional considerations, $\mathbf{G}$ can be viewed as an average length, since Eq. (47) expresses $\mathbf{G}$ as L averaged by $\Delta^3 H(L)$. (Note, however, that $\mathbf{G}$ is not a correlation length; a correlation length is typically defined as the L at which a correlation function has decayed to $1/e$ of its maximum.) Alternatively, $\mathbf{G}$ can be viewed as an average of $\Delta^3 H(L)$, weighted by L.

Speaking informally, $\mathbf{G}$ could be viewed as a measure of "disequilibrium", since it measures the difference between the actual entropy rate h_μ and the maximum possible entropy rate $\log_2|\mathcal{A}|$. The extent to which h_μ falls below the maximum measures the deviation from uniform probability, which some authors have interpreted as an equilibrium condition. In this vein, several have proposed complexity measures based on multiplying $\mathbf{G}$ by h_μ [46,47]. However, we and others have shown that this type of complexity measure fails to capture structure or memory, since they are only a function of disorder h_μ [48–50]. For additional critiques of this type of complexity measure, see Refs. [41,51].

Finally, note that for any periodic process, $\mathbf{G} = \log_2|\mathcal{A}|$, since $h_\mu = 0$. The total predictability assumes its maximum value for a completely predictable process. However, $\mathbf{G}$ does not tell us how difficult it is to carry out this prediction, nor how many symbols must be observed before the process can be optimally predicted. To capture these properties of the system, we need to look at other entropy convergence integrals.

B. Excess Entropy

Having looked at how $\Delta^2 H(L)$ converges to 0, we now ask: How does $\Delta H(L) = h_\mu(L)$ converge to h_μ? One answer to this question is provided by $\mathcal{I}_1$. For reasons that will be discussed below, we refer to $\mathcal{I}_1$ as the *excess entropy* $\mathbf{E}$:

$$\mathbf{E} \equiv \mathcal{I}_1 = \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu] \, , \tag{48}$$

The units of $\mathbf{E}$ are *bits*. We may view $\mathbf{E}$ graphically as the area indicated in the entropy-rate convergence plot of Fig. 3. For now, let us assume that the above sum is finite. For many cases of interest, however, this assumption turns out to not be correct; a point to which we shall return at the end of this section.

The excess entropy has a number of different interpretations, which will be discussed below. Excess entropy also goes by a variety of different names. References [21,39,48] use the term "excess entropy". Reference [12] uses "stored information" and Refs. [13,24,34,43,41] use "effective measure complexity". References [27,52] refer to the excess entropy simply as "complexity". References [5,6] refer to the excess entropy as "predictive information". In Refs. [26,35], the excess entropy is called the "reduced Rényi entropy of order 1".

1. **E** *as predictability-gain-weighted length*

Proposition 5 *The excess entropy may also be written as:*

$$\mathbf{E} = -\sum_{L=2}^{\infty} (L-1) \Delta^2 H(L) \; . \qquad (49)$$

Proof: See App. A 5. □

Eq. (49) shows that **E** may also be viewed as an average L, weighted by the predictability gain $\Delta^2 H(L)$, a view emphasized in Ref. [24]. However, this is not a dimensionally consistent interpretation, since **E** has units of bits. Alternatively, Eq. (49) shows that the excess entropy can be seen as an average of $\Delta^2 H(L)$, weighted by the block-length L.

2. **E** *as intrinsic redundancy*

The length-L approximation $h_\mu(L)$ typically overestimates the entropy rate h_μ at finite L. Specifically, $h_\mu(L)$ overestimates the latter by an amount $h_\mu(L) - h_\mu$ that measures how much more random single measurements appear knowing the finite L-block statistics than knowing the statistics of infinite sequences. In other words, this excess randomness tells us how much additional information must be gained about the sequences in order to reveal the actual per-symbol uncertainty h_μ. This merely restates the fact that the difference $h_\mu(L) - h_\mu$ is the per-symbol redundancy $\mathbf{r}(L)$, defined originally in Eq. (35). Though the source appears more random at length L by the amount $\mathbf{r}(L)$, this amount is also the information-carrying capacity in the L-blocks that is not actually random, but is due instead to correlations. We conclude that entropy-rate convergence is controlled by this redundancy in the source. Presumably, this redundancy is related to structures and memory intrinsic to the process. However, specifying how this memory is organized cannot be done within the framework of information theory; a more structural approach based on the theory of computation must be used. We return to the latter in the conclusion.

There are many ways in which the finite-L approximations $h_\mu(L)$ can converge to their asymptotic value h_μ. (Recall Fig. 3.) Fixing the values of $H(1)$ and h_μ, for example, does not determine the form of the $h_\mu(L)$ curve. At each L we obtain additional information about how $h_\mu(L)$ converges, information not contained in the values of $H(L)$ and $h_\mu(L)$ at smaller L. Thus, roughly speaking, each $h_\mu(L)$ is an independent indicator of the manner by which $h_\mu(L)$ converges to h_μ.

Since each increment $h_\mu(L) - h_\mu$ is an independent contribution in the sense just described, one sums up the individual per-symbol L-redundancies to obtain the total amount of apparent memory in a source [12,13,21,23,24,27,33,53]. Calling this sum *intrinsic redundancy*, we have the following result.

Proposition 6 *The excess entropy is the intrinsic redundancy of the source:*

$$\mathbf{E} = \sum_{L=1}^{\infty} \mathbf{r}(L) \; . \qquad (50)$$

Proof: This follows directly from inserting the definition of intrinsic redundancy, Eq. (35), in Eq. (48). □

The next proposition establishes a geometric interpretation of **E** and an asymptotic form for $H(L)$.

Proposition 7 *The excess entropy is the subextensive part of* $H(L)$*:*

$$\mathbf{E} = \lim_{L \to \infty} \left[H(L) - h_\mu L \right] \; . \qquad (51)$$

Proof: See App. A 6. □

This proposition implies the following asymptotic form for $H(L)$:

$$H(L) \; \sim \; \mathbf{E} + h_\mu L \, , \text{as } L \to \infty \; . \qquad (52)$$

Thus, we see that **E** is the $L = 0$ intercept of the linear function Eq. (52) to which $H(L)$ asymptotes. This observation, also made in Refs. [5,6,12,13,27], is shown graphically in Fig. 2. Note that $\mathbf{E} \geq 0$, since $H(L) \geq h_\mu L$. Note also that if $h_\mu = 0$, then $\mathbf{E} = \lim_{L \to \infty} H(L)$.

A useful consequence of Prop. 7 is that it leads one to use Eq. (52) instead of the original (very simple) scaling of Eq. (14). Later sections address how ignoring Eq. (52) leads to erroneous conclusions about a process's unpredictability and structure.

3. **E** *as mutual information*

Yet another way to understand excess entropy is through its expression as a mutual information.

Proposition 8 *The excess entropy is the mutual information between the left and right (past and future) semi-infinite halves of the chain* $\overleftrightarrow{S}$*:*

$$\mathbf{E} \; = \; \lim_{L \to \infty} I[S_0 S_1 \cdots S_{2L-1} ; S_{2L} S_{2L+1} S_{2L-1}] \; . \qquad (53)$$

when the limit exists.

Proof: See App. A 7. □

Note that **E** is not a two-symbol mutual information, but is instead the mutual information between two semi-infinite blocks of variables.

Eq. (53) says that **E** measures the amount of historical information stored in the present that is communicated to the future. For a discussion of some of the subtleties associated with this interpretation, however, see Ref. [16].

Prop. 8 also shows that $\mathbf{E}$ can be interpreted as the *cost of amnesia*: If one suddenly loses track of a source, so that it cannot be predicted at an error level determined by the entropy rate h_μ, then the entire string appears more random by a total of $\mathbf{E}$ bits.

4. Finitary processes

We have argued above that the excess entropy $\mathbf{E}$ provides a measure of one kind of memory. Thus, we refer to those processes with a finite excess entropy as finite-memory sources or, simply, *finitary processes*, and those with infinite memory, *infinitary processes*.

Definition 1 *A process is* finitary *if its excess entropy is finite.*

Definition 2 *A process is* infinitary *if its excess entropy is infinite.*

Proposition 9 *For finitary processes the entropy-rate estimate $h_\mu(L)$ decays faster than $1/L$ to the entropy rate h_μ. That is,*

$$h_\mu(L) - h_\mu < \frac{A}{L} \ , \tag{54}$$

for large L and where A is a constant. For infinitary processes $h_\mu(L)$ decays at or slower than $1/L$.

Proof: By direct inspection of Eq. (48). □

One consequence is that the entropy growth for finitary processes scales as $H(L) \sim \mathbf{E} + h_\mu L$ in the $L \to \infty$ limit, where $\mathbf{E}$ is a constant, independent of L. In contrast, an infinitary process might scale as

$$H(L) \ \sim \ c_1 + c_2 \log L + h_\mu L \ , \tag{55}$$

where c_1 and c_2 are constants. For such a system, the excess entropy $\mathbf{E}$ diverges logarithmically and $h_\mu(L) - h_\mu \sim L^{-1}$.

In Sec. VI we shall determine $\mathbf{E}$, h_μ, and related quantities for several finitary sources and one infinitary source. There are, however, a few particularly simple classes of finitary process for which one can obtain general expressions for $\mathbf{E}$, which we state here before continuing.

Proposition 10 *For a periodic process of period p, the excess entropy is given by*

$$\mathbf{E} = \log_2 p \ . \tag{56}$$

Proof: One observes that $H(L) = \log_2 p$, for $L > p$. □

Proposition 11 *For an order-R Markovian process, the excess entropy is given by*

$$\mathbf{E} = H(R) - R h_\mu \ . \tag{57}$$

Recall that an order-R Markovian process was defined in Eq. (5).

Proof: This result will be proved in Sec. VI C, when we consider an example Markovian process. Also, see Refs. [20], [39], and [40]. □

For finitary processes that are not finite-order Markovian, the entropy-rate estimate $h_\mu(L)$ often decays exponentially to the entropy rate h_μ:

$$h_\mu(L) - h_\mu \ \sim \ A 2^{-\gamma L} \ , \tag{58}$$

for large L and where γ and A are constants.

Exponential decay was first observed for various kinds of one-dimensional map of the interval and a scaling theory was developed based on that ansatz [28]. Later Eq. (58) was proven to hold for one-dimensional, fully chaotic maps with a unique invariant ergodic measure that is absolutely continuous with respect to the Lebesgue measure [36]. To our knowledge, there is not a direct proof of exponential decay for more general finitary processes. There is, however, a large amount of empirical evidence suggesting this form of convergence [13,22,25,28]. Nevertheless, several lines of reasoning suggest that exponential decay is typical and to be expected. For further discussion, see Appendix B.

Corollary 1 *For exponential-decay finitary processes the excess entropy is given by*

$$\mathbf{E} \approx \mathbf{E}_\gamma \equiv \frac{H(1) - h_\mu}{1 - 2^{-\gamma}} \ , \tag{59}$$

where γ is the decay exponent of Eq. 58 and $H(1)$ is the single-symbol entropy.

Proof: One directly calculates the area between two curves in the entropy convergence plot of Fig. 3. The first is the constant line at h_μ. The second is the curve specified by Eq. 58 with the boundary condition $h_\mu(1) = H(1)$. Alternatively, Eq. (58) may be inserted into Eq. (49); Eq. (59) then follows after a few steps of algebra. □

Note that Eq. (59) is an approximate result; it is exact only if Eq. (58) holds for all L. In practice, for small L $h_\mu(L) - h_\mu$ is larger than its asymptotic form $A 2^{-\gamma L}$ and, thus, $\mathbf{E}_\gamma$ gives an upper bound on $\mathbf{E}$.

5. Finite-L expressions for E

There are at least two different ways to estimate the excess entropy $\mathbf{E}$ for finite L. First, we have the partial-sum estimate given by

$$\begin{aligned}\mathbf{E}(L) &\equiv H(L) - Lh_\mu(L) \\ &= \sum_{M=1}^{L} \left[h_\mu(M) - h_\mu(L)\right] \ . \end{aligned} \tag{60}$$

The second equality follows immediately from the integration formula, Eq. (21), and the boundary condition $H(0) = 0$.

Alternatively, a finite-L excess entropy can be defined as the mutual information between $L/2$-blocks:

$$\mathbf{E}'(L) \ \equiv \ I[S_0 S_1 \cdots S_{L/2-1}; S_{L/2} S_{L/2+1} S_{L-1}] \ , \tag{61}$$

for L even. If L is odd, we define $\mathbf{E}'(L) = \mathbf{E}'(L-1)$. The expression in Eq. (61), however, is not as good an estimator of $\mathbf{E}$ as that of equation Eq. (60), as established by the following lemma:

Lemma 3 *:*

$$\mathbf{E}'(L) \leq \mathbf{E}(L) \leq \mathbf{E} \ . \tag{62}$$

Proof: See App. A 9. □

V. TRANSIENT INFORMATION

Thus far, we have discussed derivatives of the entropy growth curve $H(L)$, and we have also defined and interpreted two integrals: the total predictability $\mathbf{G}$ and the excess entropy $\mathbf{E}$. Both $\mathbf{G}$ and $\mathbf{E}$ have been introduced previously by a number of authors.

In this section, however, we introduce a new quantity, by following the same line of reasoning that led us to the total predictability, $\mathbf{G} = \mathcal{I}_2$, and the excess entropy, $\mathbf{E} = \mathcal{I}_1$. That is, we ask: How does $H(L)$ converge to its asymptote $\mathbf{E} + h_\mu L$? The answer to this question is provided by $\mathcal{I}_0$. For reasons that will become clear below, we shall call $-\mathcal{I}_0$ the *transient information* $\mathbf{T}$:

$$\mathbf{T} \equiv -\mathcal{I}_0 = \sum_{L=0}^{\infty} \left[\mathbf{E} + h_\mu L - H(L)\right] \ . \tag{63}$$

Note that the units of $\mathbf{T}$ are *bits* × *symbols.*

The following result establishes an interpretation of $\mathbf{T}$.

Proposition 12 *The transient information may be written as:*

$$\mathbf{T} \ = \ \sum_{L=1}^{\infty} L\left[h_\mu(L) - h_\mu\right] \ . \tag{64}$$

Proof: The proof is a straightforward calculation, however, since it is a new result, we include it here. We begin by writing the right-hand side of Eq. (64) as a partial sum:

$$\begin{aligned}&\sum_{L=1}^{\infty} L\left[h_\mu(L) - h_\mu\right] \ = \\ &\qquad \lim_{M\to\infty} \sum_{L=1}^{M} \left[L\Delta H(L) - h_\mu L\right] \ . \end{aligned} \tag{65}$$

Using Eq. (22), this becomes:

$$\begin{aligned}&\sum_{L=1}^{\infty} L\left[h_\mu(L) - h_\mu\right] \ = \\ &\qquad \lim_{M\to\infty} \left\{ MH(M) - \sum_{L=1}^{M-1} H(L) - \sum_{L=1}^{M} h_\mu L] \right\} \ . \end{aligned} \tag{66}$$

Using $MH(M) = \sum_{L=0}^{M-1} H(M)$ and $\lim_{M\to\infty} H(M) = \mathbf{E} + h_\mu M$, and rearranging slightly, we have:

$$\begin{aligned}&\sum_{L=1}^{\infty} L\left[h_\mu(L) - h_\mu\right] \ = \\ &\qquad \lim_{M\to\infty} \left\{ \sum_{L=0}^{M-1} [\mathbf{E} - H(L)] + \sum_{L=0}^{M-1} h_\mu M - \sum_{L=1}^{M} h_\mu L \right\} \ . \end{aligned} \tag{67}$$

But,

$$\begin{aligned}&\sum_{L=0}^{M-1} h_\mu M - \sum_{L=1}^{M} h_\mu L \\ &\qquad = h_\mu \left[M^2 - \frac{1}{2}M(M+1)\right] &(68)\\ &\qquad = h_\mu \frac{1}{2} M(M-1) &(69)\\ &\qquad = \sum_{L=0}^{M-1} h_\mu L \ . &(70)\end{aligned}$$

Using this last line in Eq. (67), we have

$$\begin{aligned}&\sum_{L=1}^{\infty} L\left[h_\mu(L) - h_\mu\right] \ = \\ &\qquad \lim_{M\to\infty} \left\{ \sum_{L=0}^{M-1} [\mathbf{E} + h_\mu L - H(L)] \right\} \ . \end{aligned} \tag{71}$$

The right-hand side of the above equation is $\mathbf{T}$, completing the proof. □

Recall that $\mathbf{E} + h_\mu L$ is the entropy growth curve for a finitary process, as discussed in Sec. IV B 4. Thus, $\mathbf{T}$ may be viewed as a sum of redundancies, $(\mathbf{E} + h_\mu L) - H(L)$, between the source's actual entropy growth $H(L)$ and the $\mathbf{E} + h_\mu L$ finitary-process approximation.

A. T and Synchronization Information

For finitary processes $H(L)$ scales as $\mathbf{E} + h_\mu L$ for large L. When this scaling form is reached, we say that the observer is *synchronized* to the process. In other words, when

$$\mathbf{T}(L) \equiv \mathbf{E} + h_\mu L - H(L) = 0 , \qquad (72)$$

we say the observer is synchronized at length-L sequences. As we will see below, observer-process synchronization corresponds to the observer being in a condition of knowledge such that it can predict the process outputs at an error rate determined by the process's entropy rate h_μ.

On average, how much information must an observer extract from measurements so that it is synchronized to the process in the sense described above? As argued in the previous section, an answer to this question is given by the transient information $\mathbf{T}$.

We now establish a direct relationship between the transient information $\mathbf{T}$ and the amount of information required for observer synchronization to block-Markovian processes. We begin by stating the question of observer synchronization information theoretically and fixing some notation.

Assume that the observer has a correct model $\mathcal{M} = \{\mathcal{V}, T\}$ of a process, where $\mathcal{V}$ is a set of states and T the rule governing transitions between states. That is, T is a matrix whose components T_{AB} give the probability of making a transition to state B, given that the system is in state A, where $A, B \in \mathcal{V}$. Contrary to the scenario shown in Fig. 1, in this section we assume that the observer *directly* measures the process's states. That is, we have a Markov process, rather than a *hidden* Markov process.

The task for the observer is to make observations and determine in which state $v \in \mathcal{V}$ the process is. Once the observer knows with certainty in which state the process is, the observer is *synchronized* to the source and the average per-symbol uncertainty is exactly h_μ. We are interested in describing how difficult it is to synchronize to a directly observed Markov process.

The observer's knowledge of $\mathcal{V}$ is given by a distribution over the states $v \in \mathcal{V}$. Let $\Pr(v|s^L, \mathcal{M})$ denote the distribution over $\mathcal{V}$ given that the particular sequence of symbols s^L has been observed. The entropy of this distribution over the states measures the observer's average uncertainty in $v \in \mathcal{V}$:

$$H[\Pr(v|s^L, \mathcal{M})] \equiv -\sum_{v \in \mathcal{V}} \Pr(v|s^L, \mathcal{M}) \log_2 \Pr(v|s^L, \mathcal{M}) . \qquad (73)$$

Averaging this uncertainty over the possible length-L observations, we obtain the *average state-uncertainty*:

$$\mathcal{H}(L) \equiv$$
$$-\sum_{s^L} \Pr(s^L) \sum_{v \in \mathcal{V}} \Pr(v|s^L, \mathcal{M}) \log_2 \Pr(v|s^L, \mathcal{M}) . \qquad (74)$$

The quantity $\mathcal{H}(L)$ can be used as a criterion for synchronization. The observer is synchronized to the source when $\mathcal{H}(L) = 0$ — that is, when the observer is completely certain about in which state $v \in \mathcal{V}$ the mechanism generating the sequence is. And thus, when the condition in Eq. (72) is met, we see that $\mathcal{H}(L) = 0$, and the uncertainty associated with the prediction of the model $\mathcal{M}$ is exactly h_μ.

While the observer is still unsynchronized, though, $\mathcal{H}(L) > 0$. We refer to the average total uncertainty experienced by an observer during the synchronization process as the *synchronization information* $\mathbf{S}$:

$$\mathbf{S} \equiv \sum_{L=0}^{\infty} \mathcal{H}(L) . \qquad (75)$$

The synchronization information measures the average total information that must be extracted from measurements so that the observer is synchronized.

In the following, we assume that our model is Markovian of order R. Additionally, we assume that the set of Markovian states $\mathcal{V}$ is associated with the $|\mathcal{A}^R|$ possible values of R consecutive symbols; henceforth the latter are referred to as *R-blocks*. Specifically, there is a one-to-one correspondence between the states v and the R-blocks, and hence there exists a one-to-one, invertible function $\varphi : s^R \to \mathcal{V}$. This function φ enables us to move back and forth between the states v and the R-blocks. For example, we may use φ to rewrite the set of states:

$$\mathcal{V} = \{\varphi(s_1 s_2 \cdots s_R) : s_i \in \mathcal{A}, 1 \le i \le R\} . \qquad (76)$$

The matrix T gives the transition probabilities between symbol blocks. Note that the Markovian states are "sliding" in the sense that a transition from one state to another corresponds to a transition from, say, symbol block $s_0 s_1 \cdots s_{R-1}$ to $s_1 s_2 \cdots s_R$. Thus, it is not hard to see that the transition matrix T is sparse; there are at most $|\mathcal{A}^{R+1}|$ nonzero entries in the $|\mathcal{A}^R \times \mathcal{A}^R|$ matrix T.

Theorem 1 *For a block-Markovian process, the synchronization information* $\mathbf{S}$ *is given by:*

$$\mathbf{S} = \mathbf{T} + \frac{1}{2} R(R+1) h_\mu . \qquad (77)$$

Proof: See App. C.□

Thus, the transient information $\mathbf{T}$, together with the entropy rate h_μ and the order R of the Markov process, measures how difficult it is to synchronize to a process. If a system has a large $\mathbf{T}$, then, on average, an observer will be highly uncertain about the internal state of the process while synchronizing to it. The transient information measures a structural property of the system — a property not captured by the excess entropy $\mathbf{E}$.

Corollary 2 *For periodic processes,* $\mathbf{S} = \mathbf{T}$.

Proof: For periodic processes, $h_\mu = 0$. Plugging this in to Eq. (77), the corollary follows. □

In Sec. VI B 2 we shall see that, while the excess entropy is the same ($\log_2 p$) for all period p processes, the transient informations are different. Thus, the transient information allows one to draw structural distinctions between different periodic sequences.

Corollary 3 *For exponential-decay finitary processes we have*

$$\mathbf{T} \approx \mathbf{T}\gamma \equiv \frac{H(1) - h_\mu}{(1 - 2^{-\gamma})^2} \, . \qquad (78)$$

Proof: Inserting Eq. (58) into the expression for $\mathbf{T}$ given in Eq. (64), the result follows after several steps. □

Combining Eqs. (59) and (78), we arrive at an exact relationship between the approximate expressions for the excess entropy and the transient information:

$$\mathbf{T}_\gamma = \frac{\mathbf{E}_\gamma^2}{H(1) - h_\mu} \, . \qquad (79)$$

B. Summary

This completes our exposition of entropy convergence and our method of differentiating and integrating $H(L)$ to move between levels. Table I summarizes the first levels of the entropy convergence hierarchy as investigated in the preceding sections.

Entropy-Convergence Hierarchy					
Level	Derivatives	L_n	$L \to \infty$ Limit	Integrals	
				At Level n	From Level $n+1$
0	$H(L)$	$L_0 = 0$	∞ or $\log_2 p$	$\mathbf{T} \equiv \sum_{L=0}^{\infty}[\mathbf{E} + h_\mu L - H(L)]$	$\mathbf{T} = \sum_{L=1}^{\infty}(L)[\Delta H(L) - h_\mu]$
1	$\Delta H(L)$	$L_1 = 0$	h_μ	$\mathbf{E} \equiv \sum_{L=1}^{\infty}[\Delta H(L) - h_\mu]$	$\mathbf{E} = -\sum_{L=2}^{\infty}(L-1)\Delta^2 H(L)$
2	$\Delta^2 H(L)$	$L_2 = 1$	0	$\mathbf{G} \equiv \sum_{L=1}^{\infty} \Delta^2 H(L)$	$\mathbf{G} = -\sum_{L=2}^{\infty}(L-1)\Delta^3 H(L)$
...	...	...	...	...	...
n	$\Delta^n H(L)$	$L_n = n - 1$	0	$\mathcal{I}_n \equiv \sum_{L=L_n}^{\infty} [\Delta^n H(L) - \lim_{L\to\infty} \Delta^n H(L)]$	$-\sum_{L=L_n}^{\infty}(L-1)\Delta^{n+1} H(L)$

TABLE I. Moving up and down the first levels of entropy convergence.

VI. EXAMPLES

This section analyzes several variously structured processes to illustrate a range of different entropy convergence behaviors. The results demonstrate what the preceding quantities — such as, the entropy rate, the excess entropy and the transient information — do and do not indicate about a process's organization.

A. Independent, Identically Distributed Processes

We begin with the simplest stochastic process: binary variables independently and identically distributed (IID), as in Eq. (3). Figure 5 shows the entropy growth curve $H(L)$ for two IID processes: a fair coin and a biased coin with a bias of 0.7.

For both coins $H(L)$ grows linearly. Hence, $\Delta H(L)$ is constant for these and all other IID processes. Note, however, that the two systems have different entropy rates h_μ. The fair coin has an h_μ of 1 bit per symbol, while the biased coin, being less unpredictable, has $h_\mu \approx .8813$. As a result, from Eq. (46) the total predictability $\mathbf{G} = \log_2 |\mathcal{A}| - h_\mu = 0$ bits for the fair coin and 0.1187 bits for the biased coin. The predictability of each process is rather low, as expected.

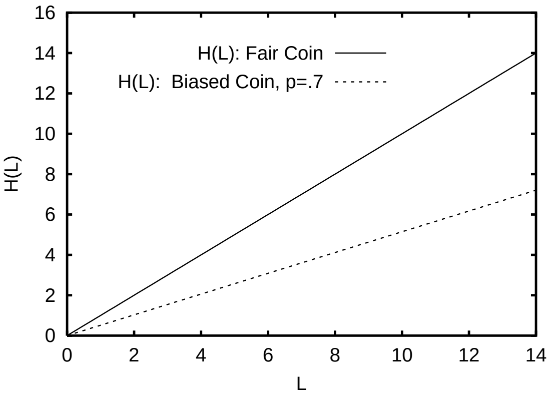
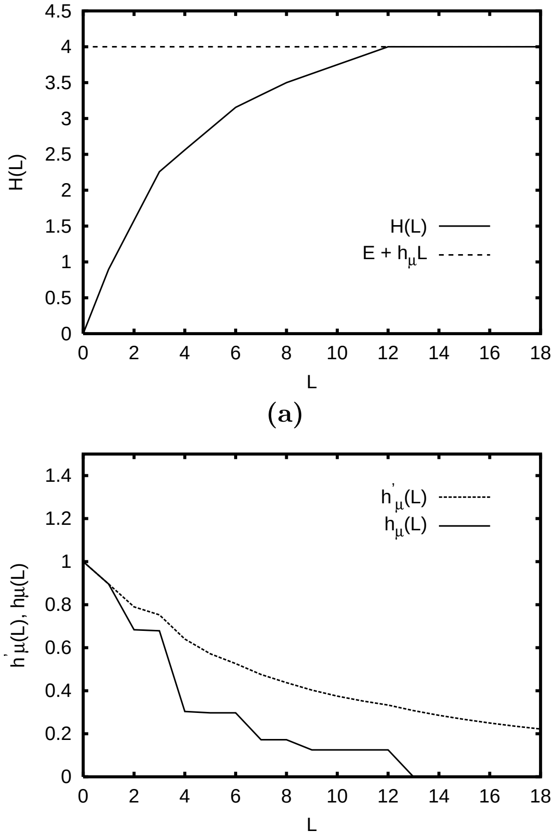


FIG. 5. Entropy growth for IID processes: a fair coin (solid line) and a coin (dashed line) with bias $p = 0.7$.

As is clear from Fig. 5, for both processes the excess entropy $\mathbf{E}$ and the transient information $\mathbf{T}$ are zero. This makes sense in light of the interpretations of $\mathbf{E}$ and $\mathbf{T}$

given in the previous sections. Each coin flip does not depend on past flips, and so there is no mutual information between the past and the future. Thus, $\mathbf{E} = 0$. Similarly, no information is needed to synchronize to the source — $H(L)$ assumes its asymptotic form at $L = 1$ — and so $\mathbf{T} = 0$. That is, the statistics of isolated flips are all that is required to optimally predict both processes. Historical information does not improve predictability.



(a)

(b)

FIG. 6. Entropy curves for the period-16 process: $\cdots(1010111011101110)^\infty\cdots$. (a) Entropy growth (solid line) and $\mathbf{E} + h_\mu L$ (dashed line). (b) Entropy convergence for the two estimators $h_\mu(L)$ (solid line) and $h'_\mu(L)$ (dotted line).

B. Periodic Processes

1. A period-16 process

We now consider periodic processes. We begin with a period-16 process, whose $H(L)$ is shown in Fig. 6(a). The sequence consists of repetitions of the length-16 block $s^{16} = 1010111011101110$. In Fig. 6(b) we show the convergence of entropy density estimates to the asymptotic value, $h_\mu = 0$. As for all period processes, the entropy rate h_μ for the period-16 process is zero; at sufficiently large L the process is perfectly predictable. In addition to $h_\mu(L)$, defined above as $H(L) - H(L-1)$, we show $h'_\mu(L) = H(L)/L$. The total entropy converges at $L = 12$. The value of $H(12) = 4$ bits reflects the fact that there are 16 equally probable sequences at each $L \geq 12$.

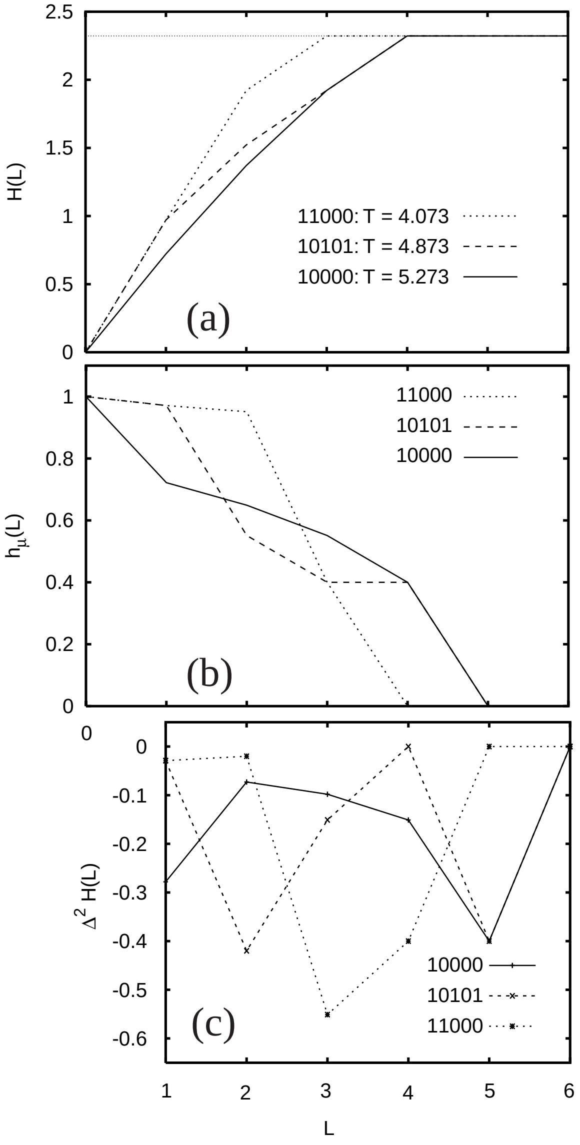


FIG. 7. (a) Entropy growth for all period-5 processes, along with the asymptote $\mathbf{E} + h_\mu L = \log_2 5 \approx 2.321$ (thin dashed line). (b) Entropy convergence, for the same period-5 processes. (c) Predictability gain $\Delta^2 H(L)$.

Template Word	Number of Observations to Synchronize [symbols]	$\mathbf{T}$ [bit-symbols]
11000	2.4	4.073
10000	2.8	5.273
10101	3.2	4.873

TABLE II. Synchronizing to period-5 processes: Comparing the transient information $\mathbf{T}$ to the average number of observations required to synchronize to the three distinct period-5 sequences.

The excess entropy $\mathbf{E}$ for the period-16 process is $\log_2 16 = 4$ bits; the sequence's past carries 4 bits of phase information about the future. Geometrically, $\mathbf{E}$ is the vertical-intercept of the horizontal asymptote on Fig. 6(a) (dashed line) or the area under $h_\mu(L)$ on Fig. 6(b). The predictability is $\mathbf{G} = \log_2 2 = 1$ bit per symbol; the system can be predicted perfectly. Finally, the transient information for this period-16 process is $\mathbf{T} \approx 16.6135$ bit-symbols. Since this process is Markovian, Thm. 1 applies. Thus, we conclude that, on average, an observer would measure a total uncertainty $\mathbf{S}$ of 16.6135 bits during the process of synchronization.

2. **T** *distinguishes period-p processes*

For any periodic process of period p, $h_\mu = 0$ and $\mathbf{E} = \log_2 p$. However, there are important structural differences between different sequences with the same period. To show this, we consider all binary period-5 processes that are distinct up to permutations and ($0 \leftrightarrow 1$)-exchanges in their "template" words. There are only three such processes: $(11000)^\infty$, $(10101)^\infty$, and $(10000)^\infty$. By the symmetries of the Shannon entropy function these processes illustrate the only three types of entropy convergence behavior possible for period-5 sequences.

Figure 7(a) shows the entropy growth curves for each; Fig. 7(b) gives the entropy convergence curves; and Fig. 7(c) gives the predictability gain $\Delta^2 H(L)$. By $L = 4$, $H(L)$ converges to $\mathbf{E} = \log_2 5 \approx 2.321$ bits. We see that $h_\mu(L) = 0$ at this and larger L. For all three processes, $\mathbf{G} = 1 - h_\mu = 1$ bit per symbol: Again, the information in each measurement concerns the periodic component of the process. The predictability gain per measurement vanishes at $L = 6$, since at that point all length-5 templates have been completely parsed and the process appears completely predictable. It is a useful exercise in understanding $\Delta^2 H(L)$ to work through each template symbol-by-symbol to see which symbols are more and less informative about each template's phase. For example, on the one hand, observing the fourth symbol of the $(10101)^\infty$ process does not improve predictability. On the other hand, the third symbol for the $(11000)^\infty$ process is highly informative and predictability increases markedly.

Corollary 2 applies here and, since $h_\mu = 0$, says that the synchronization information $\mathbf{S}$ is equal to $\mathbf{T}$; and so, we can directly interpret $\mathbf{T}$ as the synchronization information. Table VI B 1 gives the values of the transient information $\mathbf{T}$, which are all different, indicating that an observer comes to synchronize to the distinct templates differently. Table VI B 1 also gives the average number of observations required to synchronize. From this table, we see that $\mathbf{T}$ is not directly proportional to the *number* of measurements to synchronize. Rather, it is the total amount of *information* that must be extracted to synchronize.

In summary, this example shows that there are structural differences between different periodic processes of the same period. The transient information is able to capture these differences, while the excess entropy is unable to. Since many chaotic systems, for example, are a combination of periodicity and randomness, one sees that the transient information is useful in detecting synchronization to the ordered component of such processes.

C. Markovian Processes

We now consider a simple Markovian process with a nonzero entropy rate h_μ. (The periodic systems of the previous section are Markovian, but with $h_\mu = 0$.) In particular, we shall consider the golden mean (GM) process, a Markov chain of order one.

In terms of the sequences produced, the underlying golden mean system produces all binary strings with no consecutive 0s. The probabilistic version — the *golden mean process* — generates 0s and 1s with equal probability, except that once a 0 is generated, a 1 is seen. One can write down a simple two-state Markov chain for this process. The GM process is so named because the logarithm of the total number of allowed sequences grows with L at a rate given by the logarithm of the golden mean, $\phi = \frac{1}{2}(1 + \sqrt{5})$.

The various entropy convergence curves for the GM process are shown in Fig. 8. The entropy rate of the GM process is $h_\mu = 2/3$ bits per symbol and the predictability is $\mathbf{G} = 1/3$ bit per symbol. The convergence of $h_\mu(L)$ to h_μ occurs at sequence length $L = 2$. In other words, once the statistics over all possible length-2 sequences are known, one gains no additional predictability by keeping track of the occurrence of blocks of larger length. There is, however, a large predictability gain in going from blocks of length 1 to blocks of length 2. Observing that 00 is missing is the key observation that makes this system predictable. The predictability gain per symbol $\Delta^2 H(L)$ is shown in Fig. 8(c). Note that the second measurement is more informative than the first.

We find that $\mathbf{E} \approx 0.2516$ bits, and $\mathbf{T} = \mathbf{E}$, which can be easily deduced from the $H(L)$ versus L graph in Fig. 8(a). From these small values for $\mathbf{E}$ and $\mathbf{T}$ one concludes that not much historical information is needed to perform optimal prediction, nor is there much uncertainty associated with synchronization.

For this system we find that $H(1) \approx 0.9183$ bits. Plugging this and our result for h_μ into Eq. (57), we see that the expression for the excess entropy of a Markovian process is verified.

The behavior shown in Fig. 8 is typical for Markovian processes. For an order-R Markovian process, the entropy density estimates $h_\mu(L)$ will always converge exactly to h_μ by $L = R$. This follows immediately from inserting Eq. (5) into the expression for $h_\mu(L)$, Eq. (26).

Given this, we know that at $H(R) = \mathbf{E} + h_\mu R$. Solving this for $\mathbf{E}$, we arrive at Eq. (57).

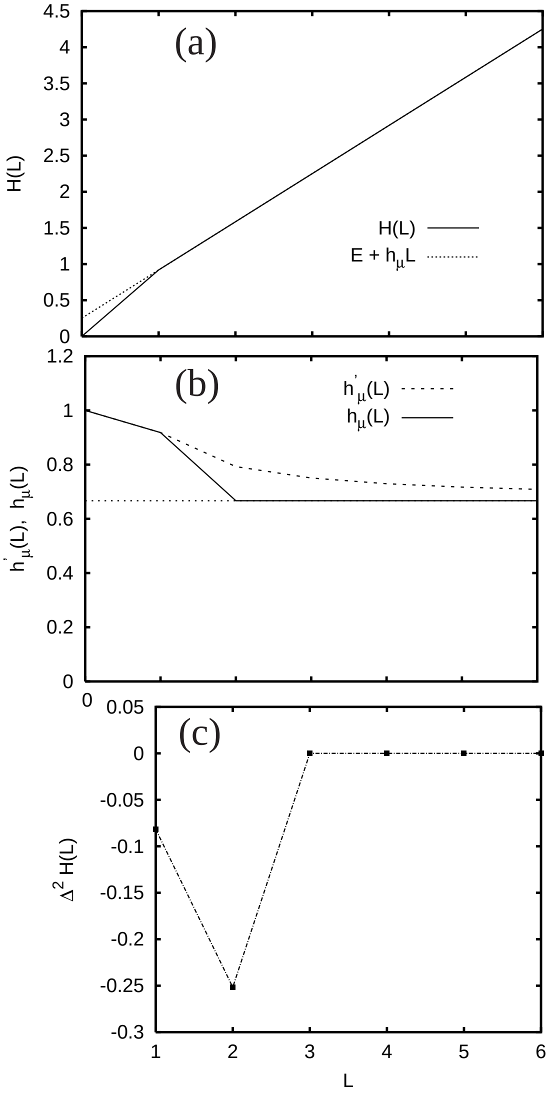


FIG. 8. (a) Entropy growth (solid line) for the golden mean process, along with the asymptote $\mathbf{E} + h_\mu L$ (dashed line). (b) Entropy convergence, both $h_\mu(L)$ (solid line) and $h'_\mu(L)$ (dashed line), for the same. (c) Predictability gain $\Delta^2 H(L)$ versus sequence length.

D. Hidden Markov Processes I: Complex Transient Structure

For our next three examples, we consider three different finitary hidden Markov processes. Each of these examples contains some interesting surprises. We begin by considering a process that consists of two successive random symbols chosen to be 0 or 1 with equal probability and a third symbol that is the logical Exclusive-OR (XOR) of the two previous. We call this the random-random-XOR (RRXOR) process. The entropy growth and convergence plots are given in Figs. 9(a) and 9(b).

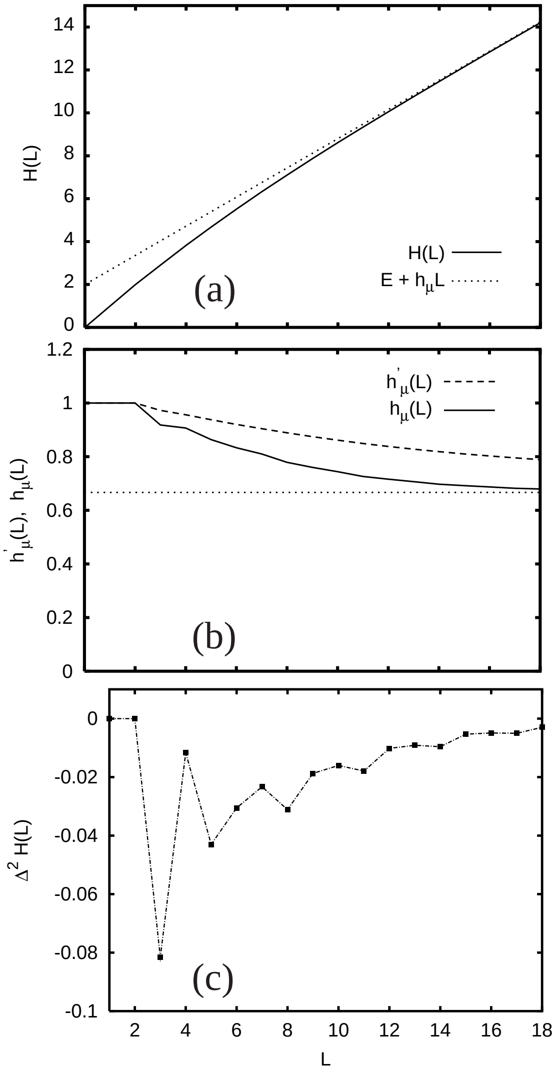


FIG. 9. (a) Entropy growth (solid line) for the random-random-XOR process, along with the asymptote $\mathbf{E}+h_\mu L$ (dashed line). (b) Entropy convergence, both $h_\mu(L)$ (solid line) and $h'_\mu(L)$ (dashed line), for the same. (c) Predictability gain $\Delta^2 H(L)$ versus sequence length.

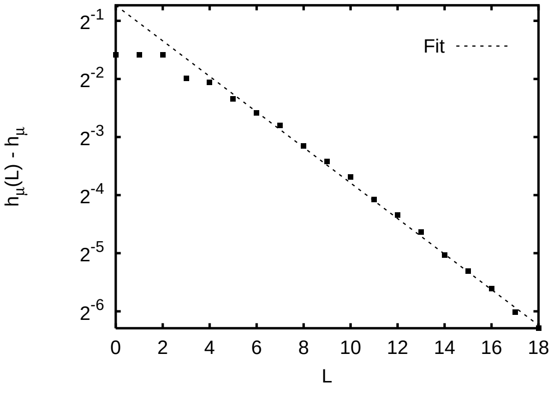


FIG. 10. A least-squares fit (dashed line) to the exponential decay of $h_\mu(L)$ (squares) for the RRXOR process.

The entropy rate h_μ is 2/3 bits per symbol. To see this, note that two out of every three symbols are completely random, while one third of the symbols are determined by the previous two. Note further that the RRXOR process has the same h_μ, and hence the same $\mathbf{G}$, as the GM process of the previous section. This serves as yet another reminder that the entropy rate is not sufficient to distinguish the structural properties of a source.

At first blush, one might expect the entropy growth curve to reach its asymptotic form at $L = 3$, just as $H(L)$ did at $L = 2$ for the golden mean process. However, Fig. 9(a) shows that this is not the case. The reason that it does not converge exactly at $L = 3$ is that the RRXOR process is not Markovian; specifically, the observed sequences of 0's and 1's are not finite-order Markovian. The RRXOR is a hidden Markov process; its internal states are Markovian, but the observed "states" are not.

Instead of converging exactly at finite L, the convergence of $h_\mu(L)$ to h_μ is exponential:

$$h_\mu(L) - h_\mu = A2^{-\gamma L} \ , \tag{80}$$

where we find $A = .60 \pm 0.02$ and $\gamma = .306 \pm .004$. This fit is illustrated in Fig. 10.

The excess entropy is $\mathbf{E} = 2$ bits: one needs to know which of the four possible random symbol-pairs has occurred before one reaches a condition of optimal predictability. Thus, the process has $\log_2 4 = 2$ bits of memory. However, the transient information is quite large; $\mathbf{T} \approx 9.43$ bit-symbols. This indicates that the process is difficult to synchronize to: Even after observing a large number of symbols, there is still some uncertainty about which internal, hidden state the process is in. Nevertheless, the transient information is finite.

For this system, $H(1) = 1$. Using Eqs. (59) and (78), we find $\mathbf{E}_\gamma \approx 1.74$ bits and $\mathbf{T}_\gamma \approx 9.12$ bit-symbols. The differences from the near-exact values above indicate the amount of deviation from a pure exponential decay of $h_\mu(L)$.

Intriguingly, the behavior of the predictability gain $\Delta^2 H(L)$ of Fig. 9(c) shows strong hints of the structure of the hidden Markov model that generates the observed sequences. At lengths $L = 1$ and $L = 2$ symbols are not informative at all: $\Delta^2 H(L) = 0$. This reflects the fact, given by the process's definition, that two of the symbols are produced by fair coin flips. For larger L, note that $\Delta^2 H(L)$ shows oscillations of period three. The RRXOR hidden Markov process also has a period-3 structure: after the two random bits and the XOR bit, the hidden Markov model always resets to the same state. Recall, however, that $\Delta^2 H(L)$ is formed from statistics over the observed symbols, not the hidden states of the process. Given this, it is somewhat surprising that $\Delta^2 H(L)$ picks up the period-3 nature of the transitions between hidden states.

E. Hidden Markov Processes II: Measure Sofic Process

We now consider another hidden Markov process: the *even process* [17], a stochastic process whose support (the set of allowed sequences) is a sofic system called the *even system* [54]. The even system generates all binary strings consisting of blocks of an even number of 1s bounded by 0s. Having observed a process's sequences, we say that a word (finite sequence of symbols) is *forbidden* if it never occurs. A word is an *irreducible forbidden word* if it contains no proper subwords which are themselves forbidden words. A system is *sofic* if its list of irreducible forbidden words is infinite. The even system is one such sofic system, since its set $\{01^{2n+1}0, n = 0, 1, \ldots\}$ of irreducible forbidden words is infinite. Note that no finite-order Markovian source can generate this or, for that matter, any other strictly sofic system. The even process then associates probabilities with each of the even system's sequences by choosing a 0 or 1 with fair probability after generating either a 0 or a pair of 1s. The result is a *measure sofic process* — a distribution over a sofic systems sequences. Like the RRXOR process, the even system is not Markovian, but a hidden Markov process.

The various entropy convergence curves for the even process are shown in Fig. 11. The entropy rate of the even process is $h_\mu = 2/3$ bits per symbol and the predictability $\mathbf{G}$ is 1/3 bits per symbol. Note that these values are the same as those for the RRXOR and GM processes, again emphasizing the poverty of h_μ as a structural measure. The convergence of $h_\mu(L)$ is exponential. A fit to

$$h_\mu(L) - h_\mu = A2^{-\gamma L} \ , \tag{81}$$

shown in Fig. 12, yields $A = .388 \pm 0.019$ and $\gamma = .501 \pm .007$. We find that $\mathbf{E} \approx 0.902$ bits. This is the amount of storage required on average to hold the information that a given observed 1 is the "even" or "odd" symbol in a block of 1s. The transient information is $\mathbf{T} \approx 3.03$ bit-symbols: The even process is moderately difficult to synchronize to, although it is much easier to synchronize to than the RRXOR process in the previous example. Since $H(1) \approx 0.918$, we find that $\mathbf{E}_\gamma \approx 0.86$ and $\mathbf{T}_\gamma \approx 2.92$, both of which agree well with the values measured for $\mathbf{E}$ and $\mathbf{T}$.

Again, the predictability gain per symbol $\Delta^2 H(L)$, shown in Fig. 11(c), oscillates as it converges to zero. The plot indicates that odd-length measurement sequences are more informative than even-length ones. As in the RRXOR example, the oscillation of $\Delta^2 H(L)$ provides a strong hint as to the underlying structure of the hidden Markov process responsible for the observed sequences. This process has two states and, thus, a strong period-2 component. This periodic behavior in the hidden states is picked up in $\Delta^2 H(L)$, despite the fact that $\Delta^2 H(L)$ is based only on the statistics of the observed, nonhidden symbols.

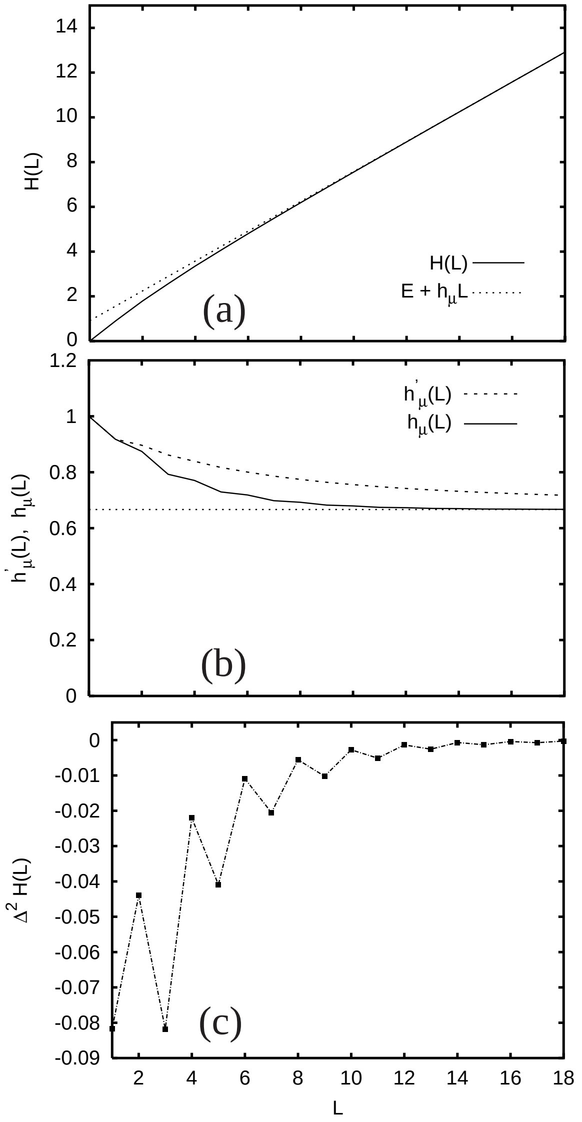


FIG. 11. (a) Entropy growth (solid line) for the even process, along with the asymptote $\mathbf{E}+h_\mu L$ (dashed line). (b) Entropy convergence, both $h_\mu(L)$ (solid line) and $h'_\mu(L)$ (dashed line), for the same. (c) Predictability gain $\Delta^2 H(L)$ versus sequence length.

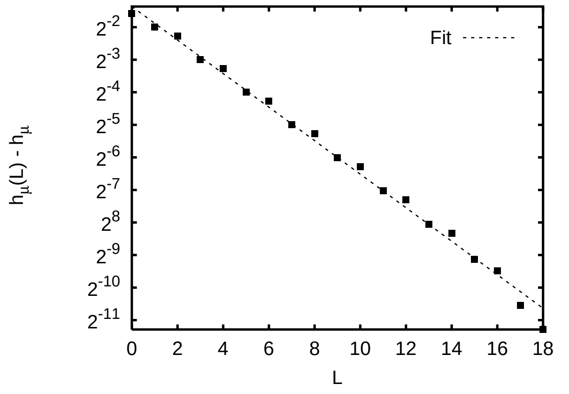


FIG. 12. A least-squares fit (dashed line) to the exponential decay of $h_\mu(L)$ (squares) for the even process.

F. Hidden Markov Processes III: The Simple Nondeterministic Source

We now consider a process known as the *simple nondeterministic source* (SNS). This process was constructed to illustrate how measurement distortion can contribute its own kind of apparent structural complexity to a simple, but hidden, information source. In particular, the SNS describes the process obtained via a non-generating partition of the logistic map [55]. For an introduction to issues of measurement-induced complexity see Ref. [55], and for a full mathematical treatment see Ref. [56]. Spatial versions of this class of hidden process were introduced in Ref. [4] and analyzed from a computation theoretic view in Ref. [57].

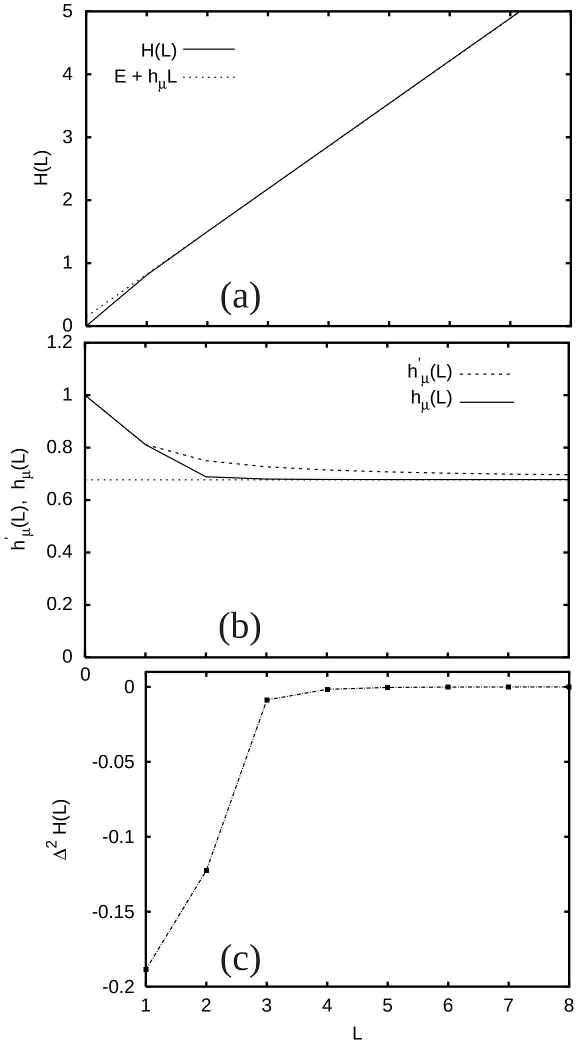


FIG. 13. (a) Entropy growth (solid line) for the simple nondeterministic source, along with the asymptote $\mathbf{E}+h_\mu L$ (dashed line). (b) Entropy convergence, both $h_\mu(L)$ (solid line) and $h'_\mu(L)$ (dashed line), for the same. (c) Predictability gain $\Delta^2 H(L)$ versus sequence length.

The SNS, a hidden Markov process, generates symbol sequences as follows. The system has three internal, hidden states: **A**, **B**, and **C**. The observer, however, only sees the binary outputs 0 and 1. The probabilities of generating the observed symbols, when the process is in each of the internal states, are given by the transition matrices $T^{(0)}$ and $T^{(1)}$, respectively:

$$T^{(1)} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix} \tag{82}$$

and

$$T^{(0)} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix} . \tag{83}$$

The elements of the transition matrices are identified with the set of internal states $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ in the natural way. For example, $T^{(1)}_{23} = 1/2$ indicates that the probability of being in state **B**, producing a 1, and making a transition to state **C** is 1/2.

Assuming that the observer knows the internal structure of the process — i.e., $T^{(0)}$ and $T^{(1)}$ — then whenever a 1 is measured the observer knows that the internal state is **C**. However, for every 0 measured after this, the observer becomes and then remains uncertain as to whether the internal state is **A** or **B**. This also explains the label “nondeterministic” for this process: the measurement of 0 does not *determine* the internal state. In contrast, all the previous examples we have considered have been deterministic, in the sense that specifying the output symbol determines the next internal state.

A central consequence of this nondeterminism is that the number of *effective states* seen by an observer that attempts to reconstruct the hidden process is infinite, even though the internal process is a simple, three-state Markov chain [55,56]. The SNS is arguably one of the simplest such examples for which this infinite-state divergence occurs.

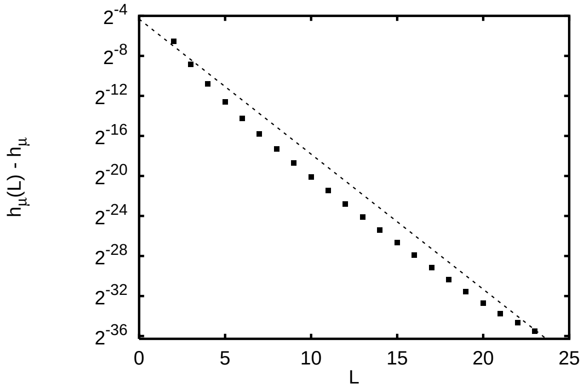


FIG. 14. A semilog plot to test for an exponential decay of $h_\mu(L)$ (squares). The latter are calculated exactly for sequences from $L = 2$ to $L = 25$. The dashed line represents an exponential decay.

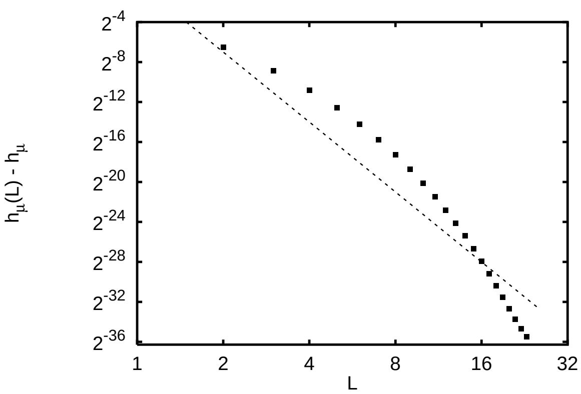


FIG. 15. A log-log plot to test for a power-law decay of $h_\mu(L)$ (squares). The latter are calculated exactly for sequences from $L = 2$ to $L = 25$. The dashed line represents a power-law decay.

The various entropy convergence curves for the SNS process are shown in Fig. 13. The entropy rate, calculable analytically, is $h_\mu \approx 0.6778$ bits per symbol and the predictability is $\mathbf{G} \approx 0.3222$ bits per symbol. We find that $\mathbf{E} \approx 0.147$ bits, there is not much mutual information between the past and future, and $\mathbf{T} \approx 0.175$ bit-symbols.

Interestingly, the functional form of $h_\mu(L) - h_\mu$ is not clear. An exponential decay

$$h_\mu(L) - h_\mu = A 2^{-\gamma L} , \tag{84}$$

is shown as the dashed line, with $A = 0.05$ and $\gamma = 1.35$, in Fig. 14. One can also test a power-law entropy decay of the form

$$h_\mu(L) - h_\mu = c L^{-\alpha} . \tag{85}$$

This is shown as the dashed line, with $c = 1.0$ and $\alpha = 7.0$, in Fig. 15. Neither form is ideal: entropy convergence is slower than exponential and faster a power law. Based on Figs. 14 and 15 one cannot infer a simple functional form for $h_\mu(L) - h_\mu$; perhaps it is some version of a stretched exponential.

In short, the simple nondeterministic source has low predictability and low apparent memory. Moreover, since **T** is small, synchronizing to it entails overcoming very little uncertainty. These would seem to be in accord with the fact that one can write down a compact nondeterministic representation for it that has only a few hidden states. However, to perform optimal prediction, a deterministic representation is needed and for the SNS that representation has an infinite number of states [55]. This degree of complexity is not suggested by the relatively small values for the information theoretic measures of structure considered here. Thus, relying only on information theoretic quantities, one is misled as to the process's actual complexity. Nonetheless, the fact that entropy convergence is not clearly exponential, in contrast to the even and RRXOR processes, provides indirect evidence that the SNS is different from these other finitary sources.

G. Aperiodic Infinitary Process

We now consider an infinite-memory process that is aperiodic and has zero entropy rate. The *Thue-Morse (TM) sequence* is the fixed point of the substitution σ defined by:

$$\sigma(0) = 01 \; , \tag{86}$$

$$\sigma(1) = 10 \; . \tag{87}$$

For example, starting from the initial string $s = 1$, the fifth iterate in the TM sequence is:

$$\sigma^5(s) = 10010110011010010110100110010110 \, . \tag{88}$$

The *Thue-Morse language* L_{TM} is the subset of all words in the TM sequence:

$$L_{TM} = \text{sub}\left(\lim_{t\to\infty} \sigma^t(1)\right) \; , \tag{89}$$

where sub(s) gives all of the subwords in string s. The *Thue-Morse process* is then given by assigning the natural measure — the frequency of occurrence in $\sigma^\infty(1)$ — to the words in L_{TM}. Unlike the previous three examples we have considered, the Thue-Morse process is *not* generated by a finitary hidden Markov process. In fact, there is no finite-state process that can generate the Thue-Morse sequence.

The various entropy convergence curves for the TM process are shown in Figs. 16 and 17. These curves were calculated using the results of Ref. [58], which show that:

$$h_\mu(1) = 1 \; , \tag{90}$$

$$h_\mu(2) = \log_2 3 - \frac{2}{3} \; , \tag{91}$$

$$h_\mu(3) = \frac{2}{3} \; , \tag{92}$$

and, for $k \geq 1$:

$$h_\mu(L) = \begin{cases} 4/(3\cdot 2^k), & \text{if } \; 2^k + 1 \leq L - 1 \leq 3 \cdot 2^{k-1} \\ 2/(3\cdot 2^k), & \text{if } \; 3\cdot 2^{k-1} + 1 \leq L - 1 \leq 2^{k+1} \end{cases} \; . \tag{93}$$

From this, one concludes that $h_\mu = 0$ and that the entropy-rate estimates converge according to a power law: $h_\mu(L) \propto 1/L$. Thus, the total entropy grows logarithmically: $H(L) \propto \log_2 L$; as shown in Fig. 16(a). Despite the slow convergence to $h_\mu = 0$, the predictability is high: $\mathbf{G} = 1$ bit per symbol. Each measurement gives the maximal amount of information about the nonrandom part of the process.

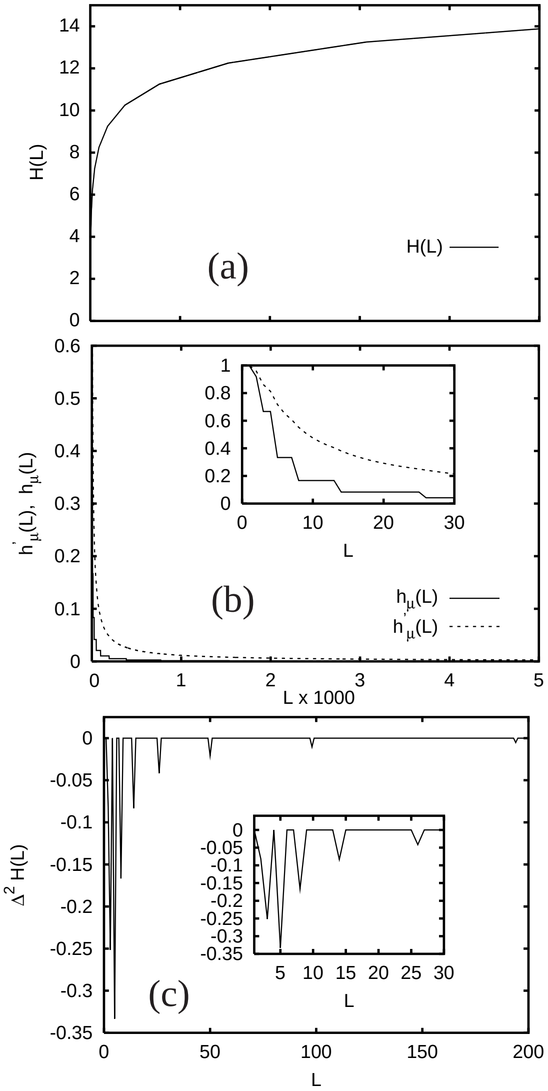


FIG. 16. (a) Entropy growth (solid line) for the Thue-Morse process. (b) Entropy convergence, both $h_\mu(L)$ (solid line) and $h'_\mu(L)$ (dashed line), for the same. In (a) and (b) sequence length goes up to $L = 5000$. (c) Predictability gain $\Delta^2 H(L)$ over small ranges of L.

Nonetheless, the excess entropy diverges; $\mathbf{E}(L) \propto \log_2 L$, indicating an infinite-memory process. (See Fig. 17(a).) This can be also be inferred from Fig. 16(a), where $\mathbf{E}$ is simply the height of the $H(L)$ curve, since $h_\mu = 0$. Finally, the transient information estimate $\mathbf{T}(L)$ also diverges, linearly, as shown in Fig. 17(b). This linear divergence is explained by looking at Eq. (64). If one substitutes $h_\mu(L) \sim L^{-1}$ and $h_\mu = 0$ into the expression there for $\mathbf{T}$, the linear divergence follows immediately.

It is clear from Fig. 16(c) that there are long sequences of measurements that are uninformative. These are punctuated occasionally with isolated symbols that do improve predictability. These occur at sequence lengths $L_i = 3 \times 2^{i-3} + 2, \; i = 3, 4, 5, \ldots$. To determine why $\Delta^2 H(L)$ behaves in this manner requires a computation theoretic approach, such as that given in Ref. [59] for the symbolic dynamics produced at the period-doubling accumulation point of the logistic map. For another similar approach, see Ref. [60].

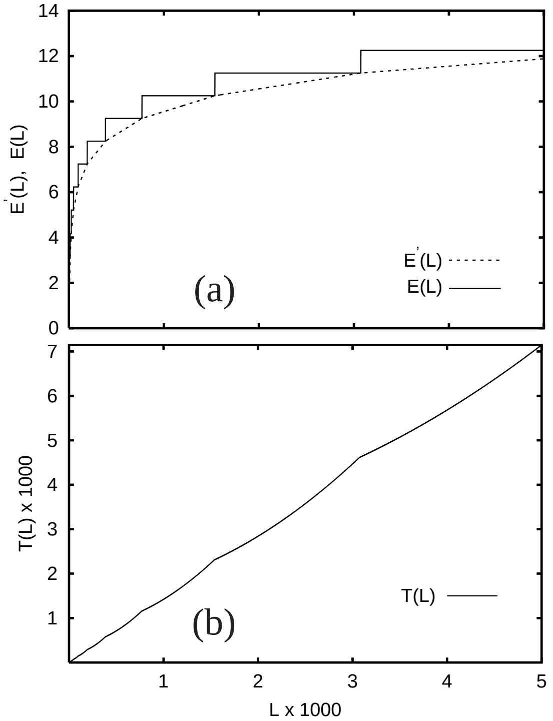


FIG. 17. (a) Excess entropy estimate divergence — $\mathbf{E}(L)$ (solid line) and $\mathbf{E}'(L)$ (dashed line). (b) Transient information estimate divergence — $\mathbf{T}(L)$ (solid line). Note that the sequence length goes up to $L = 5000$ and that both plots have large vertical scales.

We conclude this section by noting that, based on our results and those of several other authors [13,31,58,61], this $1/L$ entropy convergence is typical of aperiodic sequences generated by substitutions rules like those of Eq. (87). Moreover, Freund, Ebeling, and Rateitschat have given an argument for why this entropy convergence form is characteristic of aperiodic sequences [31].

H. Other Infinitary Processes

Before concluding this section, we review the results of several other investigations of entropy convergence. Szépfalusy and Györgyi [23] found that $h_\mu(L) - h_\mu \sim L^{-\alpha}$ with $\alpha \geq 5/2$ for a class of one-dimensional intermittent maps. Thus, this class consists of finitary processes. However, for a different model of an intermittent process, Freund found a similar decay form, but with $\alpha \approx 0.492$ [31]. Examining temporal-block sequences in elementary one-dimensional cellular automata, Grassberger [13] also found a power law decay, with $\alpha = 0.6 \pm 0.1$ for rules 30 and 45 and $\alpha = 1.0 \pm 0.1$ for rule 120. These are examples of infinitary processes.

A number of researchers have examined entropy convergence for written texts—such as The Bible, Grimms' Tales, Moby Dick, the gnuplot manual, and Gleick's popular book "Chaos" [37,38,62–64]. The picture that emerges is that entropy convergence can be fit to a power law $h_\mu(L) - h_\mu \sim L^\alpha$ with α ranging from 0.4 to 0.6. Interestingly, for a Beethoven sonata an exponent of $\alpha \approx .75$ has been found [63]. Again, these results indicate infinitary processes.

Recently, Nemenman [6] and Bialek, Nemenman, and Tishby [5] have found power-law convergences for different one-dimensional Ising models. For long-range coupling, where the coupling constants decay as the inverse lattice separation, they found an α of 0.5. They also examined an Ising model with short-range interactions, but in which the coupling constant changes every 400,000 sites within a lattice of 10^9 spins. The coupling constant was drawn from a Gaussian distribution with zero mean. For this system they found a power-law decay with an exponent of $\alpha = 1$.

I. Summary of Examples

For comparison, Table VI I collects the various analytical and numerical estimates of the information theoretic quantities for the preceding examples we analyzed.

Process	h_μ	$\mathbf{G}$	γ	$\mathbf{E}$	$\mathbf{T}$
Fair Coin	1	0		0	0
Biased Coin	0.881	0.119		0	0
Period-16	0	1		4	16.6135
$(11000)^\infty$	0	1		$\log_2 5$	4.073
$(10000)^\infty$	0	1		$\log_2 5$	5.273
$(10101)^\infty$	0	1		$\log_2 5$	4.873
Golden Mean	2/3	1/3		0.252	0.252
Even	2/3	1/3	0.501	0.902	3.03
Random-Random-XOR	2/3	1/3	0.306	2	9.43
Nondeterministic	0.678	0.322	1.35^*	0.147	0.175
Thue-Morse	0	1		$\propto \log L$	$\propto L$

TABLE III. Summary of Examples. *This can also be fit to a power law, $L^{-\alpha}$ with $\alpha \approx 7$.

VII. APPLICATIONS AND IMPLICATIONS

Being cognizant of various types of entropy convergence, of different classes of process, and of how to quantitatively distinguish between them is useful general knowledge. To this end, we reviewed information-theoretic quantities, introduced a new one, the transient information, and put forth a unified framework for relating them all in terms of discrete derivatives and integrals. Then, in the preceding section, we analyzed a number of examples. We return now to the set of questions posed in the introduction: How can we untangle different sources of apparent randomness? In particular, what happens to our estimates of the entropy rate if we ignore a process's structure?

Addressing these questions is the task of this last section. Here we show that there are direct and empirically important consequences for ignoring structural properties. We consider several different questions:

1. What happens when an observer ignores entropy-rate convergence?

2. What happens when the process's apparent memory is ignored?

3. On the one hand, what happens if the observer ignores synchronization?

4. On the other hand, what happens if the observer assumes it is synchronized to the process?

The answers, given below, show that ignoring a process's structural properties leads to a range of misleading inferences about randomness and organization. In addition to highlighting the negative consequences, we also comment on the fact that the associated problems can be alleviated to some extent, even in cases where data is limited.

A. Disorder as the Price of Ignorance

The first two questions are closely related and rather straightforward to answer. The preceding sections defined several different quantities — h_μ, $\mathbf{G}$, $\mathbf{E}$, and $\mathbf{T}$ — that measure randomness, predictability, memory, synchronization, and other features of a process. For the most part, these are asymptotic quantities in the sense that they involve the behavior of the function $H(L)$ in the $L \to \infty$ limit. Thus, their exact empirical estimation demands that an infinite number of measurements (for accurate estimates of sequence probabilities) of infinitely long sequences be made. Obviously, other than by analytic means, it is not possible to exactly calculate such quantities. Exact, $L \to \infty$ results are known for only a few special systems which are analytically tractable.

This leads one to ask, even when sequence probabilities are accurately known, how well can these various source properties be estimated at finite L? What errors are introduced and are these errors related in any way?

The simplest such question, the first one listed above, arises when one attempts to estimate source randomness h_μ via the approximation $h_\mu(L)$. Stopping the estimate at finite L gives one a rate $h_\mu(L)$ that is larger than the actual rate h_μ. That is, the source appears *more random* if we ignore correlations between variables separated by more than L steps. This observation follows directly from the definitions of h_μ and $h_\mu(L)$. However, it turns out that this form of overestimation of h_μ is related to the excess entropy $\mathbf{E}$. We shall see that there is a quantitative trade-off between randomness and memory.

Assume an observer makes measurements of a process with entropy rate h_μ and excess entropy $\mathbf{E} > 0$. Recall the definition, Eq. (13), of the entropy rate. Using this definition to estimate h_μ is tantamount to assuming that $\mathbf{E} = 0$ — see the dashed line $h_\mu L$ in Fig. 2. But, by assumption, $\mathbf{E} > 0$. Thus, at a given L, we can ask what the entropy estimate $h'_\mu(L) = H(L)/L$ is. Lemma 1 established that $h'_\mu(L) > h_\mu$. But by how much more? This is answered in a straightforward way by the following proposition.

Proposition 13 *When the observer is synchronized to the process,*

$$h'_\mu(L) - h_\mu = \frac{\mathbf{E}}{L} \ . \tag{94}$$

Proof: The claim follows immediately from the graphical construction given in Fig. 18. Saying that the observer is synchronized to the source means using an L such that $H(L) = \mathbf{E} + h_\mu L$. Thus,

$$\begin{aligned} h'_\mu(L) &= \frac{H(L)}{L} \\ &= \frac{\mathbf{E} + h_\mu L}{L} \ . \end{aligned} \tag{95}$$

Eq. (13) follows directly. □

In this way, $\mathbf{E}$ bits of memory are converted into additional, apparent randomness. The process appears more random due to the observer ignoring one of its structural properties.

One can object to this estimate: Typically one does not know the process's properties (e.g., $\mathbf{E}$ and h_μ) and so even these must be estimated. Thus, expressing the estimator h'_μ in terms of the asymptotic quantities $\mathbf{E}$ and h_μ may not be that useful. However, $\mathbf{E}'(L)$ in Eq. (61) is a non-asymptotic, L-dependent estimator of memory. Namely, $\mathbf{E}'(L)$ is a measure of the mutual information between two halves of an L-block. Using this estimator we can restate Proposition 13.

Proposition 14

$$h'_\mu(L) - h_\mu(L) \geq \frac{\mathbf{E}'(L)}{L} \ . \tag{96}$$

Proof: Again, see Fig. 18. Appealing to the monotonicity and convexity of $H(L)$, the monotonicity of $h_\mu(L)$, and Lemma 1, we can rewrite the definition

$$\mathbf{E}(L) = H(L) - h_\mu(L)L \ , \tag{97}$$

as

$$\frac{H(L)}{L} - h_\mu(L) = \frac{\mathbf{E}(L)}{L} \ . \tag{98}$$

Since $\mathbf{E}'(L)$ is bounded above by $\mathbf{E}(L)$ by Lemma 3, we have

$$h'_\mu(L) - h_\mu(L) \geq \frac{\mathbf{E}'(L)}{L} \ . \tag{99}$$

which directly proves the claim. □

This result establishes how $h_\mu(L)$ lower bounds $h'_\mu(L)$, as indicated by Lemma 1. In particular, it emphasizes that their difference is controlled by the excess entropy, a measure of memory.

Although $\mathbf{E}$ is an L-asymptotic quantity, the error $\mathbf{E}/L$ in the entropy-rate estimate dominates at small L. Moreover, being restricted to small L is typical of experimental situations with limited data or in which drift is present. That is, one cannot reliably estimate the L-block probabilities $\Pr(s^L)$ at large L due to the exponential growth in their number or the nonstationarity of block probabilities, respectively.

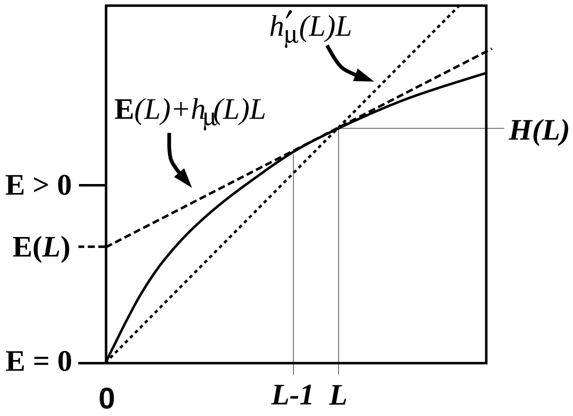


FIG. 18. Ignored memory is converted to randomness: Illustration of how ignoring memory, in this case implicitly assuming $\mathbf{E} = 0$ as Eq. (13) implies, when actually $\mathbf{E} > 0$, leads to an overestimate $h'_\mu(L)$ of the actual entropy rate h_μ.

B. Predictability and Instantaneous Synchronization

Conversely, if one assumes a fixed amount of memory $\mathbf{E}$, we shall see that this leads to an underestimate of the entropy rate h_μ. Assuming a fixed excess entropy is not something that one would be likely to do in the particular setting here, in which an observer empirically measures entropy density and related quantities from observed symbol sequences. In a more general modeling setting, however, one always runs the risk of over-fitting and, in so doing, "projecting" some particular structure—such as, additional memory capacity—onto the system. Assuming a fixed, nonzero value for the excess entropy is, in an abstract sense, an example of over-fitting. Given this, we ask, What is the consequence of assuming a fixed value for $\mathbf{E}$?

Equivalently, what happens if the observer assumes that it is synchronized to the process at some finite L, implying that $H(L) = \mathbf{E} + h_\mu L$? The geometric construction for this scenario is given in Fig. 19. In effect the source is erroneously considered to be a completely observable Markovian process in which, as we have seen, $H(L)$ converges to its asymptotic form exactly at some finite L. If the observer then uses Eq. (57) to estimate h_μ using its assumed value for $\mathbf{E}$, one arrives at the estimator $\widehat{h_\mu}$ where

$$\widehat{h_\mu} \equiv \frac{H(L) - \mathbf{E}}{L} \neq h_\mu \ . \tag{100}$$

At a given L the effect is that the observer considers the source to have a larger $\mathbf{E}$ than it actually has at that L. The line $\mathbf{E} + \widehat{h_\mu}L$ is fixed at $\mathbf{E}$ when that intercept should be lower. The result, easily gleaned from Fig. 19, is that the entropy rate h_μ is underestimated as $\widehat{h_\mu}$. In other words, the source appears more predictable than it actually is.

Proposition 15 *An observer monitors a process with excess entropy $\mathbf{E} > 0$. If the observer assumes it is synchronized when it is not, then*

$$\widehat{h_\mu} \leq h_\mu \ . \tag{101}$$

Proof: From Fig. 19 or Eq. (100), one sees that

$$\widehat{h_\mu} = \frac{H(L) - \mathbf{E}}{L} \ . \tag{102}$$

The observer is assuming that it is seeing $H(L) = \mathbf{E} + \widehat{h_\mu}L$. But since $H(L) \leq \mathbf{E} + h_\mu L$, we have that

$$\mathbf{E} + \widehat{h_\mu}L \leq \mathbf{E} + h_\mu L \ , \tag{103}$$

and so $\widehat{h_\mu} \leq h_\mu$. □

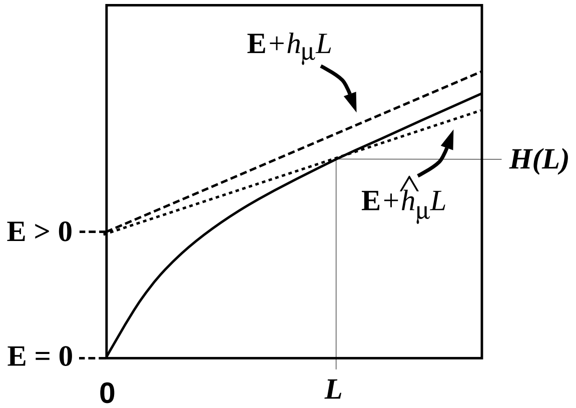


FIG. 19. Assumed synchronization converted to false predictability: Schematic illustration of how assuming one is synchronized to a process, leads to an underestimate $\widehat{h_\mu}$ for a source with excess entropy $\mathbf{E} > 0$ and entropy rate h_μ.

C. Assumed Synchronization Implies Reduced Apparent Memory

In addition to analyzing the effects on the apparent entropy rate due to assuming synchronization, we can ask a complementary question: What are the effects on estimates of the apparent memory $\widehat{\mathbf{E}}$? Figure 20 illustrates this situation.

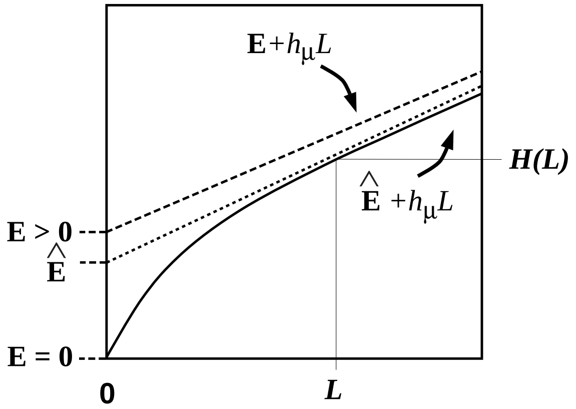


FIG. 20. Assumed synchronization leads to less apparent memory: Schematic illustration of how assuming synchronization to a source, in this case implicitly assuming $H(L) = \mathbf{E} + h_\mu L$, leads to an underestimate $\widehat{\mathbf{E}}$ of the actual memory $\mathbf{E} > 0$.

If, at a given L, we approximate the entropy rate estimate $H(L) - H(L-1)$ by the true entropy h_μ, then the offset between the asymptote and $H(L)$ is simply

$$\Delta\mathbf{E} = \mathbf{E} + h_\mu L - H(L) \ . \tag{104}$$

Translating this back to the original we have a reduced apparent memory $\widehat{\mathbf{E}} \leq \mathbf{E}$ of

$$\widehat{\mathbf{E}} = H(L) - h_\mu L \ . \tag{105}$$

In fact, since the estimated entropy rate is larger than h_μ, the reduction in apparent memory is even larger.

Thus, assuming synchronization, in the sense that $h_\mu(L) = h_\mu$, leads one to underestimate the apparent memory, as measured by the excess entropy $\mathbf{E}$.

VIII. CONCLUSION

Looking back, we have introduced a variety of information theoretic measures of a process's randomness and a variety of structural properties. Along the way, we put forth a new quantity, the transient information $\mathbf{T}$. One of the central results of this work is contained in Theorem 1, where we proved that $\mathbf{T}$ is related to the total state-uncertainty experienced while synchronizing to a Markov process.

We also calculated these information theoretic quantities for a range of differently structured processes. A natural question, then, is: To what extent does this information theoretic approach allow us to distinguish between processes that are structured in fundamentally different ways?

A. Process Classification

To summarize our results from Section VI, we now give a rough classification of several types of information source based on the quantities studied here. Similar, although coarser, classifications have been put forth by Szépfalusy [25], Ebeling [38], and Crutchfield [55].

First, we have the zero entropy rate, asymptotically predictable processes.

1. *Periodic processes*: For period-p processes, $H(L)$ becomes a constant and $h_\mu(L)$ vanishes for $L \geq p$.

2. *Aperiodic processes*: These are infinitary processes, since they need, in a crude sense, an infinite amount of memory to maintain their aperiodicity. Having $h_\mu = 0$, they cannot be aperiodic by virtue of an internal source of randomness. $\mathbf{T}$ diverges, indicating that one is never fully synchronized.

Then we have the positive entropy rate, irreducibly unpredictable processes.

1. *Memoryless processes*: For these, $H(L)$ scales as $h_\mu L$ and $h_\mu(L)$ converges immediately to h_μ. We have $\mathbf{E} = 0$ and $\mathbf{T} = 0$. Independent, identically distributed (IID) processes are examples of this class. They have no temporal memory and no structural complexity.

2. *Finitary processes*: In this class $H(L)$ scales as $\mathbf{E} + h_\mu L$. The entropy density $h_\mu(L)$ typically converges exponentially to h_μ. We have $0 < \mathbf{E} < \infty$ and $\mathbf{T} > 0$. Ref. [53] established a useful connection between information and ergodic theories for this class: finite $\mathbf{E}$ means that a process is weak Bernoulli. Within the finitary class further structural distinctions are possible:

   (a) *Markov processes*: The basic property of Markovian sources is that one synchronizes to them exactly at some finite block length L. For these processes, the effective states can be taken to be single symbols or symbol blocks of some finite length. Once that length of sequence has been parsed, the observer is synchronized and can then optimally predict the process.

   (b) *Deterministic hidden Markov processes*: These processes are characterized by an exponential convergence of $h_\mu(L)$, in contrast to the exact convergence at finite L exhibited by a Markov process. Depending on the transition structure of the hidden states, these processes can have relatively large values for the excess entropy and transient information. Within this broad class of hidden Markov processes lies the interesting case of a measure sofic process — a system whose support set contains an infinite list of irreducible forbidden words. In a limited sense, these systems have an infinite memory that keeps track of the (infinite) list of irreducible forbidden words. Nevertheless, the measure sofic process considered here, the even process, had finite $\mathbf{E}$ and $\mathbf{T}$. As noted above, the behavior of $\Delta^2 H(L)$ for these processes seems to provide strong hints of the structure of the hidden state transitions responsible for the infinite memory. In particular, we find that $\Delta^2 H(L)$ oscillates with a periodicity given by the periodic structure of the transitions between hidden states.

   (c) *Nondeterministic hidden Markov processes*: It would appear that this class of process may not be overtly different from other finitary hidden Markov processes. However, the example we considered, the simple nondeterministic source, showed a markedly different entropy convergence behavior than the other hidden Markov examples.

3. *Infinitary Sources*: At this point in time, this remains a catch-all category of processes — those falling outside the finitary classes. These include, for example, various context-free languages, such as positive-entropy-rate variations on the Thue-Morse process and other stochastic analogues from higher up the Chomsky hierarchy. Presumably, within the infinitary sources there are many interesting structural distinctions waiting to be discovered; some analogous to the automata-architectural distinctions recognized by discrete computation theory [65] and some distinctions related to the nature of the measure over the infinite sequences.

The ultimate goal of this type of classification would be an amalgamation of the structural distinctions made in the Chomsky hierarchy of computation theory [65] and statistical categories found in the ergodic theory hierarchy of stochastic processes [66].

Recent work by Nemenman [6] and Bialek, Nemenman, and Tishby [5] may be a helpful step in this direction. In Refs. [5,6] they show that the excess entropy — the "predictive information" in their parlance — is, in some circumstances, related to the number of parameters in the model producing the process. However, this result holds in a slightly different context than ours. Rather than using histograms of larger and larger variables blocks, they consider a procedure in which an observer is trying to learn a distribution through successive samplings.

B. Inferring Models from Finite Resources

In Section VII we considered various trade-offs between finite-L estimates of the excess entropy $\mathbf{E}$, the transient information $\mathbf{T}$, and the entropy rate h_μ. In particular, we have shown that not taking one or another into account leads one to systematically *over-* or *underestimate* a source's entropy rate h_μ. For example, there can be an inadvertent conversion of ignored memory into apparent randomness. The magnitude of this effect is proportional to the difference between source memory and the upper bound on memory that the observer can estimate. In a complementary way, one can inadvertently convert assumed memory into false predictability. One eventually comes to see that a process's structural features must be accounted for, even if one's focus is only on an apparently simpler question of (say) how random a process is [67].

C. Future Directions

We conclude by mentioning some important open questions and suggesting several directions for future research. First, at a number of points we have referred to "structure", without actually defining it. Is there a better, more systematic, and principled approach for determining the structure of an information source than the pure information-theoretic one just outlined? Refs. [16] and [55], for example, argue that *computational mechanics* is a viable approach to quantifying source structure and the patterns produced by information sources. They show that the ϵ-machine representation used there captures all of a source's structure. Thus, one natural question is how one can determine entropy convergence behavior given a process's ϵ-machine.

Second, it would be helpful to make a direct connection between the source characterization developed here — in terms of average source properties measured by h_μ, $\mathbf{E}$, $\mathbf{T}$, and $\mathbf{G}$ — and the difficulty of estimating these quantities and of inferring models of the sources. Analyzing the computational complexity of these two problems is the domain of computational learning theory [68,69].

Third, establishing that the source entropy rate h_μ is a metric invariant is one of the hallmarks of ergodic and dynamical systems theories [70–72]. What status do $\mathbf{E}$ and $\mathbf{T}$ hold in the same setting?

Finally, there is, of course, the question of how the information theoretic approach to structure outlined here can be extended to more than one dimension. There has been some preliminary work in this direction [13,51,52,73–76]; however, many questions remain. One of the central difficulties is that, unlike in one dimension where the various expressions for the excess entropy are equivalent, they yield different results when extended to two dimensions [77]. Careful definitions and distinct interpretations of the different forms of two-dimensional excess entropy and related quantities will have to be given in order to develop a useful, fully two-dimensional approach to pattern and structure. Our hope is that the preceding development is sufficiently clear and thorough that it can serve as a firm foundation for an information theory of structure in higher-dimensional processes.

ACKNOWLEDGMENTS

The authors thank Kristian Lindgren for helpful discussions and Robert Haslinger and Cosma Shalizi for comments on the manuscript. The cost-of-amnesia view of $\mathbf{E}$ is due to Dan Upper. This work was supported at the Santa Fe Institute under the Computation, Dynamics, and Inference Program via SFI's core grants from the National Science and MacArthur Foundations. Direct support was provided from DARPA contract F30602-00-2-0583 and AFOSR via NSF grant PHY-9970158. DPF thanks the Department of Physics and Astronomy at the University of Maine, Orono, for their hospitality during the summer of 2000, when part of this work was completed.

[1] P. S. de Laplace. *A Philosophical Essay on Probabilities.* Dover Publications, New York, 1951. Translated from the 6th French ed. by F. W. Truscott and F. L. Emory.

[2] H. Poincaré. *Science and Hypothesis.* Dover Publications, New York, 1952.

[3] S. E. Newhouse. The creation of non-trivial recurrence in the dynamics of diffeomorphisms. In G. Iooss, R. H. G. Helleman, and R. Stora, editors, *Chaotic Behaviour of Deterministic Systems.* North-Holland, Amsterdam, Netherlands, 1981.

[4] J. P. Crutchfield. Unreconstructible at any radius. *Phys. Lett. A*, 171:52 – 60, 1992.

[5] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. physics/0007070v2, 2000.

[6] I. Nemenman. *Information Theory and Learning: A Physical Approach.* PhD thesis, Princeton University, 2000.

[7] T. Schürmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6:414–427, 1996.

[8] H. Poincaré. *Les Methodes Nouvelles de la Mecanique Celeste.* Gauthier-Villars, Paris, 1892.

[9] J. P. Crutchfield. *Noisy Chaos.* PhD thesis, University of California, Santa Cruz, 1983. Published by University Microfilms Intl, Ann Arbor, Michigan.

[10] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 1948. as reprinted in "The Mathematical Theory of Communication", C. E. Shannon and W. Weaver, University of Illinois Press, Champaign-Urbana (1963).

[11] C. E. Shannon. Communication theory of secrecy systems. In N. J. A. Sloan and A. D. Wyner, editors, *Claude Elwood Shannon: Collected Papers*, pages 84–143. IEEE Press, 1993.

[12] R. Shaw. *The Dripping Faucet as a Model Chaotic System.* Aerial Press, Santa Cruz, California, 1984.

[13] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25(9):907–938, 1986.

[14] D. Blackwell and L. Koopmans. On the identifiability problem for functions of Markov chains. *Ann. Math. Statist.*, 28:1011–1015, 1957.

[15] R. J. Elliot. *Hidden Markov Models: Estimation and Control.* Springer-Verlag, 1995.

[16] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 2001. in press.

[17] J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317–359, Reading, Massachusetts, 1992. Addison-Wesley.

[18] R. M. Gray. *Entropy and Information Theory.* Springer-Verlag, New York, 1990.

[19] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[20] L. Gatlin. *Information Theory and the Living System.* Columbia University Press, 1972.

[21] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica D*, 7:201–223, 1983.

[22] G. Györgyi and P. Szépfalusy. Calculation of the entropy in chaotic systems. *Phys. Rev. A*, 31:3477–3479, 1985.

[23] P. Szépfalusy and G. Györgyi. Entropy decay as a measure of stochasticity in chaotic systems. *Phys. Rev. A*,

33(4):2852–2855, 1986.

[24] K. Lindgren and M. G. Norhdal. Complexity measures and cellular automata. *Complex Systems*, 2(4):409–440, 1988.

[25] P. Szepfalusy. Characterization of chaos and complexity by properties of dynamical entropies. *Physica Scripta*, T25:226–229, 1989.

[26] A. Csordás and P. Szépfalusy. Singularities in Rényi information as phase transitions in chaotic states. *Phys. Rev. A.*, 39(9):4767–4777, 1989.

[27] W. Li. On the relationship between complexity and entropy for Markov chains and regular languages. *Complex Systems*, 5(4):381–399, 1991.

[28] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433–466, 1982.

[29] J. P. Crutchfield and N. H. Packard. Noise scaling of symbolic dynamics entropies. In H. Haken, editor, *Evolution of Order and Chaos*, pages 215–227, Berlin, 1982. Springer-Verlag.

[30] N. H. Packard. *Measurements of Chaos in the Presence of Noise*. PhD thesis, University of California, Santa Cruz, 1982.

[31] J. Freund, W. Ebeling, and K. Rateitschak. Self-similar sequences and universal scaling of dynamical entropies. *Phys. Rev. E*, 54(5):5561–5566, 1996.

[32] J. Freund. Asymptotic scaling behavior of block entropies for an intermittent process. *Phys. Rev. E*, 53:5793–5799, 1996.

[33] K. Lindgren. Microscopic and macroscopic entropy. *Phys. Rev. A*, 38(9):4794–4798, 1988.

[34] K. Lindgren. Entropy and correlations in dynamical lattice systems. In P. Manneville, N. Boccara, G. Y. Vichniac, and R. Bidaux, editors, *Cellular Automata and Modeling of Complex Physical Systems*, volume 46 of *Springer Proceedings in Physics*, pages 27–40, Berlin, 1989. Springer-Verlag.

[35] Z. Kaufmann. Characteristic quantities of multifractals—application to the Feigenbaum attractor. *Physica D*, 54:75–84, 1991.

[36] A. Csordas, G. Gyorgyi, P. Szepfalusy, and T. Tel. Statistical properties of chaos demonstrated in a class of one dimensional maps. *Chaos*, pages 31–49, 1993.

[37] W. Ebeling and T. Poschel. Entropy and long-range correlations in literary english. *Europhysics Letters*, 26:241–246, 1994.

[38] W. Ebeling. Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO. *Physica D*, 109:42–52, 1997.

[39] J. P. Crutchfield and D. P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E*, 55(2):1239R–1243R, 1997.

[40] D. P. Feldman and J. P. Crutchfield. Measures of statistical complexity: Why? *Physics Letters A*, 238:244–252, 1998.

[41] W. Ebeling, L. Molgedey, J. Kurths, and U. Schwarz. Entropy, complexity, predictability and data analysis of time series and letter sequences. http://summa.physik.hu-berlin.de/tsd/, 1999.

[42] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.

[43] K-E. Eriksson and K. Lindgren. Structural information in self-organizing systems. *Physica Scripta*, 35:388–97, 1987.

[44] W. Li. Mutual information functions versus correlation functions. *J. Stat. Phys.*, 60(5/6):823–837, 1990.

[45] H. Herzel and I. Grosse. Measuring correlations in symbol sequences. *Physica A*, 216:518–542, 1995.

[46] R. Lopez-Ruiz, H. L. Mancini, and X. Calbet. A statistical measure of complexity. *Phys. Lett. A*, 209:321–326, 1995.

[47] J. S. Shiner, M. Davison, and P. T. Landsberg. Simple measure for complexity. *Phys. Rev. E*, 59:1459–1464, 1999.

[48] D. P. Feldman and J. P. Crutchfield. Discovering noncritical organization: Statistical mechanical, information theoretic, and computational views of patterns in simple one-dimensional spin systems. *J. Stat. Phys.*, 1998. (Submitted).

[49] J. P. Crutchfield, D. P. Feldman, and C. R. Shalizi. Comment I on Simple Measure for Complexity. *Phys. Rev. E*, 62:2996–7, 2000.

[50] P. M. Binder. Comment II on Simple Measure for Complexity. *Phys. Rev. E*, 62:2998–9, 2000.

[51] Y. A. Andrienko, N. V. Brilliantov, and J. Kurths. Complexity of two-dimensional patterns. *Europ. Phys. J. B*, 15:539–546, 2000.

[52] D. Arnold. Information-theoretic analysis of phase transitions. *Complex Systems*, 10:143–155, 1996.

[53] A. del Junco and M. Rahe. Finitary codings and weak Bernoulli partitions. *Proc. AMS*, 75:259, 1979.

[54] B. Weiss. Subshifts of finite type and sofic systems. *Monastsh. Math.*, 77:462, 1973.

[55] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[56] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997.

[57] K. Lindgren, C. Moore, and M. Nordahl. Complexity of two-dimensional patterns. *J. Stat. Phys.*, 91:909–951, 1998.

[58] V. Berthé. Conditional entropy of some automatic sequences. *J. Phys. A*, 27:7993–8006, 1994.

[59] J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. H. Zurek, editor, *Complexity, Entropy and the Physics of Information*, volume VIII of *Santa Fe Institute Studies in the Sciences of Compexity*, pages 223–269. Addison-Wesley, 1990.

[60] Y. Wang, L. Yang, and H. Xie. Complexity of unimodal maps with aperiodic kneading sequences. *Nonlinearity*, 12:1151–76, 1999.

[61] T. Gramss. Entropy of the symbolic sequence for critical circle maps. *Phys. Rev. E*, 50(4):2616–2620, 1994.

[62] W. Ebeling and G. Nicolis. Entropy of symbolic sequences: The role of correlations. *Europhys. Lett.*, 14:191–6, 1991.

[63] V. S. Anishchenko, W. Ebeling, and A. B. Neiman. Power law distributions of spectral density and higher order en-

tropies. *Chaos, Solitons, and Fractals*, 4:69–81, 1994.

[64] W. Ebeling, T. Poschel, and K.-F. Albrecht. Entropy, transinformation and word distribution of information-carrying sequences. *Bifurcation and Chaos*, 5:51–61, 1995.

[65] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, 1979.

[66] J. L. Lebowitz and O. Penrose. Modern ergodic theory. *Physics Today*, 26(2):23–29, 1973.

[67] J. Ford. How random is a coin toss? *Physics Today*, pages 40–47, April 1983.

[68] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, Massachusetts, 1994.

[69] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, Berlin, 1995.

[70] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk. SSSR*, 119:861–864, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.

[71] A. N. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR*, 124:754, 1959. (Russian) Math. Rev. vol. 21, no. 2035b.

[72] Ja. G. Sinai. On the concept of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768–771, 1959.

[73] K. Lindgren. Entropy and correlations in discrete dynamical systems. In J. L. Casti and A. Karlqvist, editors, *Beyond Belief: Randomness, Prediction and Explanation in Science*, pages 88–109. CRC Press, Boca Raton, Florida, 1991.

[74] K. Lindgren, C. Moore, and M. G. Nordahl. Complexity of two-dimensional patterns. *J. Stat. Phys.*, 91:909–951, 1998.

[75] C. DeW. Van Siclen. Information entropy of complex structures. *Phys. Rev. E*, 56:5211–15, 1997.

[76] R. Piasecki. Entropic measure of spatial disorder for systems of finite-sized objects. *Physica A*, 277:157–73, 2000.

[77] D. P. Feldman. *Computational Mechanics of Classical Spin Systems.* PhD thesis, University of California, Davis, 1998.

[78] C. Beck and F. Schlögl. *Thermodynamics of Chaotic Systems.* Cambridge University Press, 1993.

[79] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman. *The Theory of Critical Phenomena: An Introduction to the Renormalization Group.* Oxford Science Publications, 1992.

[80] J. L. Lebowitz and O. Penrose. Decay of correlations. *Phys. Rev. Lett.*, 31:749–52, 1973.

APPENDIX A: PROOFS

1. Prop. 1

Proposition 1: $\Delta H(L) = \mathcal{D}[\Pr(s^L)||\Pr(s^{L-1})]$.

Proof: By direct calculation we have the following.

$$\mathcal{D}[\Pr(s^L)||\Pr(s^{L-1})] = \sum_{\{s^L\}} \Pr(s^L)\log_2\frac{\Pr(s^L)}{\Pr(s^{L-1})} \quad \text{(A1)}$$

$$= \sum_{\{s^L\}} \Pr(s^L)\log_2\Pr(s^L) - \sum_{\{s^{L-1}\}}\sum_{\{s_{L-1}\}} \Pr(s^L)\log_2\Pr(s^{L-1}) \quad \text{(A2)}$$

$$= H(L) - \sum_{\{s^{L-1}\}} \log_2\Pr(s^{L-1}) \sum_{\{s_{L-1}\}} \Pr(s^L) \quad \text{(A3)}$$

$$= H(L) - H(L-1)\ , \quad \text{(A4)}$$

since $\Pr(s^{L-1}) = \sum_{\{s_{L-1}\}} \Pr(s^L)$. □

2. Prop. 2

Proposition 2:
$\Delta^2 H(L) = -\mathcal{D}[\Pr(s_{L-1}|s^{L-2})||\Pr(s_{L-2}|s^{L-3})]$.

Proof: By the expressions for the second discrete derivative, Eq. (20) and Eq. (A1), we have:

$$\Delta^2 H(L) = \Delta H(L) - \Delta H(L-1) \quad \text{(A5)}$$

$$= -\sum_{\{s^L\}} \Pr(s^L)\log_2\Pr(s_{L-1}|s^{L-1}) + \sum_{\{s^{L-1}\}} \Pr(s^{L-1})\log_2\Pr(s_{L-2}|s^{L-2}) \quad \text{(A6)}$$

$$= -\sum_{\{s^L\}} \Pr(s^L)\log_2\frac{\Pr(s_{L-1}|s^{L-1})}{\Pr(s_{L-2}|s^{L-2})} \quad \text{(A7)}$$

$$= -\mathcal{D}[\Pr(s_{L-1}|s^{L-1})||\Pr(s_{L-2}|s^{L-2})]\ . \quad \text{(A8)}$$

□

3. Prop. 3

Proposition 3: $-\mathbf{G} = \mathbf{R}$.

Proof: We write the sum of Eq. (44) and use the antidifferentiation formula Eq. (21) to get:

$$\sum_{L=1}^{M} \Delta^2 H(L) = \Delta H(M) - \Delta H(0)\ . \quad \text{(A9)}$$

Since $\lim_{L\to\infty}\Delta H(L) = h_\mu$ and since we have defined $\Delta H(0) = \log_2|\mathcal{A}|$, it follows immediately that

$$-\mathbf{G} = \lim_{M\to\infty}[\Delta H(0) - \Delta H(M)] \quad \text{(A10)}$$

$$= \log_2|\mathcal{A}| - h_\mu\ , \quad \text{(A11)}$$

which is $\mathbf{R}$ by Eq. (15).□

4. Prop. 4

Proposition 4: $\mathbf{G} = -\sum_{L=2}^{\infty}(L-1)\Delta^3 H(L)$.

Proof: We write Eq. (47) as a partial sum as follows:

$$\sum_{L=2}^{\infty}(L-1)\Delta^3 H(L) \; = \\ \lim_{M\to\infty}\left[\sum_{L=2}^{M} L\Delta^3 H(L) - \sum_{L=2}^{M}\Delta^3 H(L)\right] . \qquad \text{(A12)}$$

We use Eqs. (22) and (21) on the first and second terms on the right-hand side and obtain, after simplifying:

$$-\sum_{L=2}^{\infty}(L-1)\Delta^3 H(L) \; = \\ -\lim_{M\to\infty}\left[M\Delta^2 H(M) - \sum_{L=1}^{M}\Delta^2 H(L)\right] . \qquad \text{(A13)}$$

From the definition of $\mathbf{G}$, Eq. (44), and since we assume that $\mathbf{G}$ is finite, $\lim_{M\to\infty} L\Delta^2 H(L) = 0$. From this we see immediately that

$$-\sum_{L=2}^{\infty}(L-1)\Delta^3 H(L) = \sum_{L=1}^{\infty}\Delta^2 H(L) \equiv \mathbf{G} . \qquad \text{(A14)}$$

□

5. Prop. 5

Proposition 5: $\mathbf{E} = -\sum_{L=2}^{\infty}(L-1)\Delta^2 H(L)$.

Proof: Writing the right-hand side of the above equation as a partial sum, and then using the integration-by-parts formula Eq. (22) we obtain, after some algebra:

$$-\sum_{L=2}^{\infty}(L-1)\Delta^2 H(L) \; = \\ \lim_{L\to\infty}\left\{-M\Delta H(M) \; + \sum_{L=1}^{M}\Delta H(L)\right\} . \qquad \text{(A15)}$$

Recalling that $\Delta H(L) = h_\mu(L)$ and that $h_\mu(M) \to h_\mu$ in the $M \to \infty$ limit, we see at once that

$$-\sum_{L=2}^{\infty}(L-1)\Delta^2 H(L) \; = \; \sum_{L=1}^{\infty}[h_\mu(L) - h_\mu] \; \equiv \; \mathbf{E} . \quad \text{(A16)}$$

The last equality follows from the definition of $\mathbf{E}$, Eq. (48). □

6. Prop. 7

Proposition 7: $\mathbf{E} = \lim_{L\to\infty}[H(L) - h_\mu L]$.

Proof: Writing out the partial sum of the infinite sum in Eq. (48) and evaluating it using the integration formula, Eq. (21):

$$\sum_{L=1}^{M}[\Delta H(L) - h_\mu] = H(M) - H(0) - h_\mu M . \qquad \text{(A17)}$$

Since $H(0) \equiv 0$, it then follows immediately that

$$\mathbf{E} = \lim_{M\to\infty}[H(M) - h_\mu M] . \qquad \text{(A18)}$$

Since, by Eq. 34, the left-hand side is $\mathbf{R}(L)$, the proof is complete. □

7. Prop. 8

Proposition 53: $\mathbf{E} = I[\overrightarrow{S}; \overleftarrow{S}]$.

Proof: We rewrite the definition so that we can use the finite-L forms of various entropies:

$$I[\overrightarrow{S}; \overleftarrow{S}] \equiv \lim_{L\to\infty} I[\overrightarrow{S}^L ; \overleftarrow{S}^L] . \qquad \text{(A19)}$$

We begin with the definition of mutual information, Eq. (8), which expresses I as the difference between two entropies:

$$I[\overrightarrow{S}^L ; \overleftarrow{S}^L] = H[\overrightarrow{S}^L] - H[\overrightarrow{S}^L \mid \overleftarrow{S}^L] . \qquad \text{(A20)}$$

Recall that $H[\overrightarrow{S}^L] = H(L)$.

Using the conditional entropy chain rule [19] we have

$$H[\overrightarrow{S}^L \mid \overleftarrow{S}^L] = H[S_0, S_1, \ldots, S_{L-1} | S_{-L}, \ldots, S_{-1}] \quad \text{(A21)}$$

$$= \sum_{i=0}^{L-1} H[S_i | S_{-L} S_{-L+1} \cdots S_{i-1}] . \qquad \text{(A22)}$$

Putting these together we have

$$I[\overrightarrow{S}; \overleftarrow{S}] \; = \\ \lim_{L\to\infty}\left[H(L) - \sum_{i=0}^{L-1} H[S_i | S_{-L} S_{-L+1} \cdots S_{i-1}]\right] . \qquad \text{(A23)}$$

In the $L \to \infty$ limit, each term in the summand is equal to h_μ. Thus, we see that

$$I[\overrightarrow{S}; \overleftarrow{S}] \; = \; \lim_{L\to\infty}[H(L) - L h_\mu] , \qquad \text{(A24)}$$

which is $\mathbf{E}$ by Prop. 7. □

8. Lemma 1

Lemma 1: $h'_\mu(L) \geq h_\mu(L) \geq h_\mu$.

Proof: We prove the right inequality first. Since conditioning reduces entropy,

$$h_\mu(L) \geq h_\mu(L') \ , \ \forall L > L' \ . \tag{A25}$$

Now, recall that

$$\lim_{L\to\infty} h_\mu(L) = h_\mu \ . \tag{A26}$$

Since, by Eq. (A25), the $h_\mu(L)$'s are nonincreasing as L increases, it follows that $h_\mu(L) \geq h_\mu$.

We now prove the left inequality in the proposition.

$$\begin{aligned} h'_\mu(L) &\equiv \frac{H(L)}{L} \\ &= \frac{1}{L}\sum_{i=1}^{L} H[S_i|S_{i-1}S_{i-2}\cdots S_i] \ . \end{aligned} \tag{A27}$$

For all $i < L$,

$$H[S_i|S_{i-1}S_{i-2}\cdots S_i] \geq H[S_L|S_{L-1}S_{L-2}\cdots S_1] \ . \tag{A28}$$

Thus,

$$h'_\mu(L) \geq \frac{1}{L}\sum_{i=1}^{L} H[S_L|S_{L-1}S_{L-2}\cdots S_1] \tag{A29}$$

$$= \frac{1}{L} L H[S_L|S_{L-1}S_{L-2}\cdots S_1] \tag{A30}$$

$$= h_\mu(L) \ . \tag{A31}$$

□

9. Lemma 3

Lemma 3: $\mathbf{E}'(L) \leq \mathbf{E}(L) \leq \mathbf{E}$.

Proof: We first prove the right inequality. Recall that

$$\mathbf{E}'(L) \equiv H(L) - Lh_\mu(L) = \sum_{M=1}^{L} (h_\mu(M) - h_\mu(L)) \ . \tag{A32}$$

Since $M \geq L$ for all terms in the summand, all elements of the sum are positive. Now, the excess entropy is defined as

$$\mathbf{E} \equiv \lim_{L\to\infty} \sum_{M=1}^{L} (h_\mu(M) - h_\mu(L)) \ . \tag{A33}$$

Thus, $\mathbf{E}(L)$ is the partial sum of the above term. Since all terms in the sum are non-negative, it follows immediately that the partial sum $\mathbf{E}(L)$ is less than the infinite sum $\mathbf{E}$.

We now prove the left inequality. Using stationarity,

$$\mathbf{E}'(L) = 2H(L/2) - H(L) \ . \tag{A34}$$

Recall that for odd L, we defined $\mathbf{E}'(L) = \mathbf{E}'(L-1)$. To prove the left inequality, it will suffice to show that:

$$2H(L/2) - H(L) \leq H(L) - Lh_\mu(L) \ . \tag{A35}$$

Rearranging, we have:

$$2H(L/2) \leq 2H(L) - Lh_\mu(L) \ . \tag{A36}$$

By the concavity of $H(L)$, $2H(L/2) \geq H(L)$, and thus the above equation becomes:

$$H(L) \leq 2H(L) - Lh_\mu(L) \ . \tag{A37}$$

Rearranging again, we see that we need to show:

$$H(L) \geq Lh_\mu(L) \ . \tag{A38}$$

That this equation is true can be seen geometrically by inspecting Fig. 18. Note that the inequality is saturated if and only if the process is independent identically distributed.

To verify Eq. (A38) algebraically, we use the chain rule on the left hand side and obtain:

$$\sum_{M=1}^{L} H[S_M|S_{M-1}S_{M-2}\cdots S_1] \geq Lh_\mu(L) \ . \tag{A39}$$

But,

$$\sum_{M=1}^{L} H[S_M|S_{M-1}S_{M-2}\cdots S_1]$$

$$\geq \sum_{M=1}^{L} H[S_L|S_{L-1}S_{L-2}\cdots S_1] \tag{A40}$$

$$= Lh_\mu(L) \ . \tag{A41}$$

Thus, Eq. (A38) is true, and the proof is complete. □

APPENDIX B: EXPONENTIAL CONVERGENCE TO THE ENTROPY RATE

It was claimed in the main text that $h_\mu(L) - h_\mu$ often vanishes exponentially fast for finitary sources. Why is this behavior so common? There are several ways to argue for the ubiquity of exponential entropy convergence.

First, note that if $\Delta^2 H(L)$ converges to 0 exponentially fast, then $h_\mu(L) = \Delta H(L)$ must also converge exponentially fast. Then, a direct calculation shows that $\Delta^2 H(L) \leq I(L)$, where $I(L)$ is the mutual information between two variables separated by L symbols. Now, the two-variable mutual information is related to the two-variable correlation function $C(L)$. In particular, $I(L) \propto C^2(L)$. This result was first shown for binary sequences by Li [44] and later generalized to larger alphabets by Herzel and Grosse [45]. As a result, if the correlations decay exponentially, then the two-symbol mutual information decays exponentially. This, in turn, allows one to conclude that the entropy-rate estimate converges exponentially and so $\mathbf{E}$ is finite.

The conclusion from these observations is that exponential convergence of correlation functions implies the exponential convergence of the entropy rate. However, this only transfers the convergence question from entropy rates to correlation functions. So *why* is it that correlation functions typically decay exponentially? There are several answers to this question.

Mathematically, many stochastic processes can be re-expressed as one-dimensional spin models; see, e.g., Ref. [78]. Thus, we expect that what is typical for spin systems will also be typical for the more general stochastic processes of interest to us here. In a one-dimensional statistical mechanical model with finite interaction strengths, one can always express the partition function as an infinite product of transfer matrices. The correlation function between two spins L lattice sites apart is proportional to $(\lambda_0/\lambda_1)^L$, where λ_0 is the largest eigenvalue of the transfer matrix and λ_1 the second largest eigenvalue. The Perron-Frobenius theorem guarantees that the largest eigenvalue is nondegenerate, thus establishing the exponential decay of the correlation function. This result is standard; see, for example, Ref. [79].

Physically, in a spin system the sum of the correlation functions yields the magnetic susceptibility χ. The exponential decay of the correlation function thus ensures that χ is finite. Hence, away from a critical point, where we expect finite response functions such as χ, we also expect exponentially decaying correlation functions — or at least correlations that decay faster than $1/L$.

Mathematically, it has been shown that, under a fairly wide range of circumstances, a statistical mechanical system with an analytic partition function necessarily has correlation functions that decay exponentially [80]. Unlike the Perron-Frobenius transfer matrix argument, the results in Ref. [80] hold for systems in more than one dimension.

APPENDIX C: PROOF OF THE SYNCHRONIZATION INFORMATION THEOREM

We begin by restating the theorem:

Theorem 1: If the source is order-R Markovian, then

$$\mathbf{S} = \mathbf{T} + \frac{1}{2}R(R+1)h_\mu \; . \qquad \text{(C1)}$$

Proof: Since the transition probabilities are normalized, T is a stochastic matrix; $\sum_b T_{ab} = 1$. The eigenvector corresponding to the eigenvalue 1 shall be denoted by π and is normalized in probability;

$$\sum_a \pi_a T_{ab} \; = \; \pi_b , \quad \sum_a \pi_a = 1 \; . \qquad \text{(C2)}$$

As is well-known, π_a gives the asymptotic probability of the state $A \in \mathcal{V}$. Equivalently, in terms of the R-blocks,

$$\pi_A \; = \; \Pr(\varphi^{-1}(A)) \; . \qquad \text{(C3)}$$

Or, simply

$$\pi_A \; = \; \Pr(s^R) \; , \qquad \text{(C4)}$$

where s^R is understood to correspond to the Ath state.

Initially, before any measurements are made, we assume our distribution over $\mathcal{V}$ is given by π;

$$\Pr(\mathcal{V}|\lambda, \mathrm{M}) = \pi \; , \qquad \text{(C5)}$$

where λ is the empty string. Hence, $\mathcal{H}(0) = H\{\pi\}$. Equivalently, it follows from Eqs. (76) and (C3) that

$$\mathcal{H}(0) = H(R) \; . \qquad \text{(C6)}$$

If we observe a particular symbol s'_1, we now know that the process must be in one of the states that correspond to symbol blocks whose first symbol is s'_1. We denote this set of states by:

$$\mathcal{V}_{s'_1} \equiv \{\varphi(s'_1 s_2 \cdots s_R) : s_i \in \mathcal{A}, 2 < i < R\} \; . \qquad \text{(C7)}$$

Likewise, after we've observed the particular length L sequence ${s'}^L$, $L < R$, we know that the process must be in one of the states that corresponds to an R-symbol block whose first L symbols are ${s'}^L$;

$$\begin{aligned}\mathcal{V}_{s'^L} \; \equiv \; \{\varphi({s'}^L s_{L+1} s_{L+2} \cdots s_R) :& \\ s_i \in \mathcal{A}, L+1 < i < R\} \; , \; L \leq R \; .& \qquad \text{(C8)}\end{aligned}$$

The following properties of $\mathcal{V}_{s'^L}$ follow immediately from the definition, Eq. (C8):

$$\mathcal{V}_{s^L} \subset \mathcal{V} \; , \qquad \text{(C9)}$$

$$\mathcal{V}_{s^L} \bigcap \mathcal{V}_{s'^L} \; = \; \emptyset \; \text{ if and only if } s^L \neq {s'}^L \; , \qquad \text{(C10)}$$

and,

$$\bigcup_{s^L} \mathcal{V}_{s^L} \; = \; \mathcal{V}, \; . \qquad \text{(C11)}$$

Thus, the set of L-blocks $\{s^L\}$ induces a partition of the set of states $\{\mathcal{V}\}$. For a given L there are at most $\mathcal{A}^L$ sets $\mathcal{V}_{s^L}$, each of which is a proper subset of $\mathcal{V}$. (There are exactly $\mathcal{A}^L$ subsets of $\mathcal{V}$ if and only if there are no forbidden sequences.) The set $\mathcal{V}_{s^L}$ has at most $\mathcal{A}^{R-L}$ elements. So, as more and more symbols are observed —

i.e., as L grows — the subsets $\mathcal{V}_{s^L}$ of $\mathcal{V}$ become more and more refined. For the Markovian case considered here, eventually enough symbols will be observed so that we know with probability 1 the state of the process. Since the Markovian states are in a one-to-one relation with the R-blocks, we are guaranteed to know the state with certainty after R symbols have been observed. Hence, $\mathcal{H}(R) = 0$. Observing subsequent symbols will not add to the state uncertainty since each observation uniquely determines the subsequent state. Thus, $\mathcal{H}(L) = 0$ for $L \geq R$.

For $L < R$, the distribution over the Markovian states $v \in \mathcal{V}$ is given by:

$$\Pr(v|s^L, \mathcal{M}) = \frac{\pi^{s^L}}{\Pr(s^L)} , \tag{C12}$$

where π^{s^L} is a vector whose $|\mathcal{V}|$ components are given by:

$$(\pi^{s^L})_v = \begin{cases} \pi_v, & \text{if } v \in \mathcal{V}_{s^L} \\ 0, & \text{otherwise} \end{cases} . \tag{C13}$$

We are interested in calculating $\mathcal{H}(L)$, the average state-uncertainty after observing L symbols. In order to perform this calculation, the following two properties of π^{s^L} will be necessary.

First, for fixed s^L, observe that summing $(\pi^{s^L})_v$ over its components v results in $\Pr(s^L)$, the probability of that particular s^L. This follows from the definition of (π^{s^L}), Eq. (C13):

$$\begin{aligned}\sum_{\{v\}} (\pi^{s^L})_{v\in\mathcal{V}} &= \sum_{v \in \mathcal{V}_{s^L}} \pi_v && \text{(C14)}\\ &= \sum_{\{s^R : \varphi(s^R) \in \mathcal{V}_{s^L}\}} \Pr(s^R) && \text{(C15)}\\ &= \sum_{\{s_{L+1}s_{L+2}\cdots s_R\}} \Pr(s^R) && \text{(C16)}\\ &= \Pr(s^L) . && \text{(C17)}\end{aligned}$$

Hence, $\Pr(\mathcal{V}|s^L, \mathcal{M})$ as given in Eq. (C12) is normalized over s^L.

Second, notice that $(\pi^{s^L})_v$ has only one nonzero entry for fixed L and fixed state A. This follows from noting that the particular state $A \in \mathcal{V}$ is associated with a particular R-block $\varphi^{-1}(A)$. More formally, suppose that $(\pi^{s^L})_v$ has a nonzero entry for two different L-blocks, say s^L and ${s'}^L$;

$$(\pi^{s^L})_v = (\pi^{{s'}^L})_v > 0, \;\; s^L \neq {s'}^L . \tag{C18}$$

Then, by Eq. (C13), it follows that:

$$A \in \mathcal{V}_{s^L}, \;\text{ and } v \in \mathcal{V}_{{s'}^L} , \tag{C19}$$

which, in turn, implies that:

$$\mathcal{V}_{s^L} \cap \mathcal{V}_{{s'}^L} \neq \emptyset \text{ and } s^L \neq {s'}^L . \tag{C20}$$

This last equation contradicts Eq. (C10). Thus, the original proposition must be true: $(\pi^{s^L})_v$ has only one nonzero entry — namely π_v — for all possible s^L's.

We are now ready to complete our calculation of $\mathcal{H}(L)$. Plugging Eq. (C12) into Eq. (74) and simplifying slightly, we have:

$$\begin{aligned}\mathcal{H}(L) = &- \sum_{\{s^L\}} \sum_{v\in\mathcal{V}} (\pi^{s^L})_v \log_2 (\pi^{s^L})_v \; + \\ &\sum_{\{s^L\}} \sum_{v\in\mathcal{V}} (\pi^{s^L})_v \log_2 \Pr(s^L) . \end{aligned} \tag{C21}$$

Parenthetically, we note that $\mathcal{H}(L)$ is the information gain: $\mathcal{H}(L) = \mathcal{D}[\pi^{s^L} || \Pr(s^L)]$. By Eq. (C17), we can perform the sum over v in the second term on the right-hand side of the above equation, and we obtain the entropy of an L-block, $H(L)$.

To evaluate the first term on the right-hand side, recall that $(\pi^{s^L})_v$ has only one nonzero entry for fixed L and fixed v. Using this, we see that

$$\begin{aligned}\sum_{s^L} & \left(- \sum_{v\in\mathcal{V}} (\pi^{s^L})_v \log_2 (\pi^{s^L})_v \right) \\ &= - \sum_{v\in\mathcal{V}} \pi_v \log_2 \pi_v && \text{(C22)}\\ &= H[\pi] && \text{(C23)}\\ &= H(R) . && \text{(C24)}\end{aligned}$$

Thus, it follows that:

$$\mathcal{H}(L) = \begin{cases} H(R) - H(L) & \text{if } 0 \leq L \leq R \\ 0 & \text{if } L > R \end{cases} . \tag{C25}$$

We now have an expression for $\mathcal{H}(L)$ in terms of $H(L)$, and we finish the proof with a direct calculation. Looking at Eq. (C1), one sees that it will suffice to show that:

$$\sum_{L=0}^{\infty} \mathcal{H}(L) - \mathbf{T} = \frac{h_\mu}{2} R(R{+}1) . \tag{C26}$$

By assumption, the process is order-R Markovian. This implies that $\mathcal{H}(L) = 0$ and $H(L) = \mathbf{E} + h_\mu L$ for all $L \geq R$. As a result of this latter equation, the summand of the infinite sum that defines $\mathbf{T}$, Eq. (63), is zero for all $L \geq R$. That is, the last nonzero contribution to the sum comes at $L = R - 1$. As a result, the left-hand side of Eq. (C26) can be written as:

$$\begin{aligned}\sum_{L=0}^{\infty} \mathcal{H}(L) - \mathbf{T} = & \\ \sum_{L=0}^{R-1} & \left[H(R) - H(L) - \mathbf{E} - h_\mu L + H(L) \right] && \text{(C27)}\\ &= \sum_{L=0}^{R-1} \left[H(R) - \mathbf{E} - h_\mu L \right] . && \text{(C28)}\end{aligned}$$

But $H(R) = \mathbf{E} + h_\mu R$, since we assume that synchronization occurs at $L = R$. Plugging this into the above equation, we have:

$$\sum_{L=0}^{\infty} \mathcal{H}(L) - \mathbf{T} = \sum_{L=0}^{R-1} [\mathbf{E} + h_\mu R - \mathbf{E} - h_\mu L] \tag{C29}$$

$$= h_\mu \sum_{L=0}^{R-1} (R - L) \tag{C30}$$

$$= h_\mu (R^2 - \frac{1}{2} R(R-1) \tag{C31}$$

$$= \frac{h_\mu}{2} R(R+1) \ . \tag{C32}$$

This last equation is Eq. (C26), thus completing the proof. □.