Measures of Statistical Complexity: Why?

David P. Feldman
Department of Physics,
University of California, Davis, CA 95616
Electronic Address: dpf@santafe.edu

James P. Crutchfield
Physics Department, University of California,
Berkeley, CA 94720-7300
and Santa Fe Institute,
1399 Hyde Park Road, Santa Fe, NM 87501
Electronic Address: chaos@santafe.edu

(Dated: 11 November 1997)

We review several statistical complexity measures proposed over the last decade and a half as general indicators of structure or correlation. Recently, Lòpez-Ruiz, Mancini, and Calbet [Phys. Lett. A 209 (1995) 321] introduced another measure of statistical complexity C_{LMC} that, like others, satisfies the "boundary conditions" of vanishing in the extreme ordered and disordered limits. We examine some properties of C_{LMC} and find that it is neither an intensive nor an extensive thermodynamic variable and that it vanishes exponentially in the thermodynamic limit for all one-dimensional finite-range spin systems. We propose a simple alteration of C_{LMC} that renders it extensive. However, this remedy results in a quantity that is a trivial function of the entropy density and hence of no use as a measure of structure or memory. We conclude by suggesting that a useful "statistical complexity" must not only obey the ordered-random boundary conditions of vanishing, it must also be defined in a setting that gives a clear interpretation to what structures are quantified.

PACS numbers: 05.20.-y,05.50.+q,64.60.Cn,05.45.+b
Keywords: statistical complexity, excess entropy, mutual information, Shannon entropy, Kolmogorov complexity

I. STATISTICAL COMPLEXITY MEASURES

Theoretical physics has long possessed a general measure of the uncertainty associated with the behavior of a probabilistic process: the Shannon entropy of the underlying distribution [1, 2]—a quantity originally introduced by Boltzmann over 100 years ago. In the '50's, Kolmogorov and Sinai [3, 4] adapted Shannon's information theory to the study of dynamical systems. This work formed the foundation for the statistical characterization of deterministic sources of apparent randomness in the late '60's through the early '80's. These efforts to describe the unpredictability of dynamical systems were largely successful. The metric entropy, Lyapunov exponents, and fractal dimensions now provide widely applicable quantities that can be used to detect the presence of and to quantify the degree of deterministic chaotic behavior.

Since that time, though, it has become better appreciated that measuring the randomness and unpredictability of a system fails to adequately capture the correlational structure in its behavior. *Structure* here is taken to be a statement about the relationship between a system's components. Roughly speaking, the larger and more intricate the "correlations" between the system's constituents, the more structured its underlying distribution. Structure and correlation are not completely independent of randomness, however. It is generally agreed that both maximally random and perfectly ordered systems possess no structure [5–7]. Nevertheless, at a given level of randomness away from these extremes, there can be an enormously wide range of differently structured processes.

These realizations led to a considerable effort to develop a general measure that quantifies the degree of structure or pattern present in a process (c.f. [5–17]). There are many *ad hoc* methods for detecting structure, but none are as widely applicable as entropy is for indicating randomness. The quantities that have been proposed as general structural measures are often referred to as "complexity measures." To reduce confusion, it has become convenient to refer to them instead as *statistical* complexity measures. In so doing they are immediately distinguished from deterministic complexities, such as the Kolmogorov-Chaitin complexity [18–20] which requires the deterministic accounting of every bit—random or not—in an object. In contrast, statistical complexity measures discount for randomness and so provide a measure of the regularities present in an object above and beyond pure randomness. Deterministic complexities are dominated by the random components in an object; the result is that their average-case growth rate is given by the Shannon entropy rate [2].

A number of approaches to measuring statistical complexity have been taken. One line of attack operates within information theory and examines how the Shannon entropy of successively larger subsystems converges to the entropy density of the entire system [6, 8, 9, 11, 13, 14]. These quantities can be interpreted as the average memory stored in configurations.

Another set of approaches appeals to computation theory's classification of devices that recognize different classes of formal languages (sets of strings). Examples include finite memory devices (e.g. the finite-state machines) and infinite memory devices (e.g. push-down automata and Turing machines) [21]. One such computation-based measure of statistical complexity is the *logical depth* [12]. The logical depth of (say) a system's configuration is the time required for a universal Turing machine to run the minimal program that reproduces it. Another example of a computation theoretic approach is found in the statistical complexity of refs. [7, 15], a quantity that measures the amount of memory needed, on average, to statistically reproduce a given configuration. Unlike logical depth, which assumes the use of a universal Turing machine—the most powerful discrete computational model class—this statistical complexity assumes that the simplest possible computational class is used to describe the configuration. A higher level class is used only if lower ones fail to admit a finite description. This "bottom-up" definition of hierarchical information processing has been successfully applied to the symbolic dynamics of chaotic maps [7, 15], cellular automata [22], spin systems [23], and hidden Markov models [24]. For other approaches to statistical complexity see, for example, refs. [5, 10, 16, 17, 25].

II. PROPERTIES OF $C_{\rm LMC}$

Recently, Lòpez-Ruiz, Mancini, and Calbet proposed another measure of statistical complexity $C_{\rm LMC}$ [26]. Consider a discrete random variable Y that can take on N values y. We denote by $\Pr(y)$ the probability that the variable Y assumes the value y. Ref. [26] then defines a complexity measure:

$$C_{\rm LMC}[Y] \equiv H[Y]\, D[Y] \, , \tag{1}$$

where H is the Shannon entropy,

$$H[Y] = -\sum_{\{y\}} \Pr(y) \log_2 \Pr(y) \, , \tag{2}$$

in which the sum runs over all allowed values of y. The quantity D is the "disequilibrium," defined by

$$D[Y] = \sum_{\{y\}} (\Pr(y) - \frac{1}{N})^2 \, , \tag{3}$$

which measures the departure of $\Pr(y)$ from uniformity.

The motivation posited in ref. [26] for the form of $C_{\rm LMC}$ is that it vanishes for distributions that correspond to perfect order and maximal randomness. Ref. [26] argues that perfect order corresponds to a vanishing Shannon entropy and notes that for $H = 0$, $C_{\rm LMC} = 0$. Maximal randomness occurs for $H = \log_2 N$, corresponding to $\Pr(y) = 1/N$. And so, by eq. (3) D and hence $C_{\rm LMC}$ equal zero. Thus, by construction, $C_{\rm LMC}$ vanishes in the extreme ordered and disordered limits.

We now proceed to discuss the behavior of $C_{\rm LMC}$ in the thermodynamic limit. Anteneodo and Plastino have already reported some of $C_{\rm LMC}$'s properties in this limit [27]. However, their line of investigation is rather different from that undertaken here. In ref. [27] the distribution that maximizes $C_{\rm LMC}$ is determined. Numerical and analytic work there indicate that the maximizing distribution for $N \to \infty$ is one in which a single event has probability $2/3$ while all others are equally likely.

As an alternative to looking at the maximizing distribution, we suggest examining how a system's complexity changes as parameters—e.g. temperature, coupling strength, nonlinearity, etc.—are varied. This approach is in keeping with statistical mechanics, where one typically looks at changes in the behavior of quantities in the thermodynamic limit as system parameters are modified. For example, rather than determine the distribution that maximizes the expectation value of the specific heat for a finite-sized system, one usually determines how the specific heat per site behaves in the limit of an infinite system as, say, the temperature is varied.

Consider the class of probability distributions over discrete, finite random variables generated by finite-memory Markov chains. Let $\ldots X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ be a bi-infinite chain of random variables where each value x_i is chosen from a discrete finite alphabet of size k. We denote L consecutive variables by $X_i^L \equiv X_i, X_{i+1}, X_{i+2}, \ldots, X_{i+L-1}$. X_i^L is a system of L variables with $N = k^L$ possible configurations x_i^L. Let $\Pr(x_i)$ be the probability that the $i^{\rm th}$ random variable takes on the particular value x_i. We denote by $\Pr(x_i^L)$ the joint probability distribution over L consecutive random variables. We assume a shift symmetry so that $\Pr(x_i^L) = \Pr(x_0^L)$ and subsequently drop the subscript $i = 0$. The chain of discrete random variables may be viewed as a translationally invariant spin system or, equivalently, as a stationary stochastic process.

As is well known, the Shannon entropy of a block of L such variables typically grows linearly for sufficiently large L. In other words, the limit

$$h_\mu \equiv \lim_{L\to\infty} \frac{1}{L} H[X_0, X_1, \ldots, X_{L-1}] \tag{4}$$

exists and, in the thermodynamic ($L \to \infty$) limit,

$$H[X^L] \propto h_\mu L \, . \tag{5}$$

The quantity h_μ is well-defined as the system size goes to infinity and is known as the entropy rate, the metric entropy, or the thermodynamic entropy density depending on the context. In statistical mechanics parlance, eq. (4) tells us that the Shannon entropy H is an extensive quantity, so that it is possible to define a meaningful entropy density h_μ that characterizes the randomness per variable in the system.

The "disequilibrium" term, eq. (3), is not so well behaved. In fact, under no circumstances does it grow linearly with system size L. One can show, quite generally, that for a system of length L described by a probability distribution over $N = k^L$ events, D is bounded above by $1-1/N$. To see this, we expand the square in eq. (3) and, since the probability distribution is normalized, obtain

$$D[X^L] \, = \, \sum_{\{x^L\}} \Pr(x^L)^2 - \frac{1}{k^L} \; . \tag{6}$$

The sum is understood to run over all k^L possible values of X^L. Since $\Pr(x^L) \leq 1$, it follows that

$$\sum_{\{x^L\}} \Pr(x^L)^2 \, \leq \, \sum_{\{x^L\}} \Pr(x^L) \, = \, 1 \; . \tag{7}$$

Thus,

$$D[X^L] \, \leq \, 1 - k^{-L} \; , \tag{8}$$

and we see that D cannot grow linearly with the system size L. Therefore, the "disequilibrium" is not a thermodynamically extensive quantity.

In fact, it can be shown that D vanishes exponentially with increasing system size for a large class of systems. Let our chain of variables $\ldots X_{-1}, X_0, X_1, X_2, \ldots$ be chosen by a one-step Markov process with transition probabilities given by $T_{ab} = \Pr(b|a), \; a, b \in \{0, 1, \ldots, k-1\}$. That is, T_{ab} gives the probability that the variable X_{i+1} takes on the b^{th} value given that X_i takes on the a^{th} value. (What follows applies to any finite-step Markov chain, as blocks of adjacent variables can be grouped to render the process one-step.) Then, if we assume that the Markovian process is regular—*i.e.*, there exists some K such that $(T_{ab})^K > 0$ for all a and b—then it follows that the disequilibrium of a system of L variables goes to zero exponentially fast in L. This is proved in appendix A.

As a result, C_{LMC} vanishes in the thermodynamic limit for all regular Markov chains, a class of systems that includes all finite-range, one-dimensional spin systems with finite-strength interactions. It seems to us counterintuitive that a (useful) measure of complexity vanishes for all of these systems. While these models exhibit no critical phenomena, there are considerable changes in the structure of the distributions as system parameters are varied [23]. A measure of complexity, as we envision it, should be sensitive to these changes.

As an illustration of the nonextensivity of D, consider the special case where the chain consists of variables that are independent and identically distributed (iid); *i.e.*, $\Pr(x_i, x_j) = \Pr(x_i)\Pr(x_j)$ for all X_i and X_j with $i, j = \ldots, -2, -1, 0, 1, 2, \ldots$. For a spin system, this corresponds to the case where there is no coupling between spins—a paramagnet. For convenience we assume that X_i can take on two values, (say) 0 and 1, and we denote the probability that X_i takes on the value 1 by p. Then, using the binomial theorem, we find that C_{LMC} for a system of L such variables is given by:

$$C_{\text{LMC}}[X^L] \, = \, LH[X] \, \left((1 - 2p + 2p^2)^L \, - \, 2^{-L} \right) \; , \tag{9}$$

where $H[X]$ is the binary entropy function:

$$H[X] = -p \log_2 p \, - \, (1-p) \log_2 (1-p) \; . \tag{10}$$

Eq. (9) is rather curious. In our view, the complexity of a collection of iid binary variables should vanish regardless of their number. A set of independent variables is statistically very simple—there is manifestly no correlational structure whatsoever. Furthermore, it seems to us that if the complexity doesn't vanish for all such systems, it ought to grow linearly as a function of the number of variables. That is, six biased coins should be twice as complex as three biased coins. Eq. (9) shows that the size dependence of C_{LMC} is much more complicated.

In ref. [27] it is found that the maximal value of C_{LMC} goes to 4/27 as the system size goes to infinity. This is not at odds with the exponentially fast vanishing of D (and hence C_{LMC}) for regular Markov processes noted here, since the system that maximizes C_{LMC} is not Markovian. To see this, recall that the maximal distribution reported in ref. [27] has one configuration with probability 2/3 while all others have equal probability. In the thermodynamic limit, this one configuration is infinitely long. Thus, the generating process must keep track of arbitrarily long configurations in order to assign a distinct probability to one and another probability uniformly to all others. As a result, the distribution that maximizes C_{LMC} cannot be generated by a finite-memory Markov process in the thermodynamic limit.

III. REPAIRING NONEXTENSIVITY

Up to this point we have seen that C_{LMC} is not suitable for use in a statistical mechanics context. In particular, it suffers from two related deficiencies: it is not an extensive quantity and it vanishes for a large variety of structured processes. The trouble causing both of these shortcomings resides in the "disequilibrium" factor. Perhaps if one altered the definition of C_{LMC} so that D was extensive, as the Shannon entropy H is, one would obtain a more useful measure of statistical complexity.

To this end, we seek an extensive measure of a distribution's departure from uniformity. As we will be multiplying this measure by the Shannon entropy H which carries units of bits, it also seems natural, although not necessary, to choose a "disequilibrium-like" quantity that also carries units of bits. Information theory is armed with just such a function: the relative information [2].

The relative information, also known as the information gain or the Kullback-Leibler information distance, between two distributions $\Pr(y)$ and $\widehat{\Pr}(y)$ is defined by

$$D(\Pr(y) \,\|\, \widehat{\Pr}(y)) \;\equiv\; \sum_{\{y\}} \Pr(y) \log_2 \frac{\Pr(y)}{\widehat{\Pr}(y)} \; . \qquad (11)$$

The relative information is not a true distance function, neither satisfying the triangle inequality nor being symmetric. Nevertheless, it does provide a measure of how much two distributions differ and it does carry the same units (bits) as the Shannon entropy. It is also an extensive quantity, since it grows linearly with the number of variables in the distribution's support.

So, $D(\Pr(y) \,\|\, \widehat{\Pr}(y))$ where $\widehat{\Pr}(y) = 1/N$ provides an extensive measure of $\Pr(y)$'s departure from uniformity in units of bits. Using this in eq. (3), we define a modified statistical complexity measure:

$$C'[Y] \;\equiv\; H[Y] \, D(\Pr(y) \,\|\, 1/N) \; , \qquad (12)$$

which has units of [bits^2]. From here on we will focus on C', the modified $C_{\rm LMC}$. Note that much of C''s character is shared by $C_{\rm LMC}$.

To see how this new quantity behaves, let's look more closely at $D(\Pr(y) \,\|\, 1/N)$. Consider again X^L, a Markov chain of length L. And for convenience let the X_i be binary variables. The total number N of configurations for such a system is 2^L. First, note that:

$$D(\Pr(x^L) \,\|\, 1/N) \;=\; L - H[X^L] \; . \qquad (13)$$

Since $H[X^L]$ is extensive (eq. (5)), we have:

$$D(\Pr(x^L) \,\|\, 1/N) \;\propto\; L(1 - h_\mu) \; . \qquad (14)$$

Thus, C' consists of the product of two extensive quantities, H and $D(\Pr(x^L) \,\|\, 1/N)$. As a consequence, dividing C' by L^2 yields a quantity that is finite in the $L \to \infty$ limit; specifically,

$$\lim_{L \to \infty} \frac{1}{L^2} C' \;=\; h_\mu (1 - h_\mu) \; . \qquad (15)$$

(See Fig. 1.) This is indeed a function that vanishes in the ordered ($h_\mu = 0$) and disordered ($h_\mu = 1$) extremes. However, note that it is a function *only* of the system's entropy density. That is, the modified statistical complexity measure eq. (12) is a function only of the system's randomness. The relation $C' \propto L^2 h_\mu (1 - h_\mu)$ strikes us as being "over-universal." For example, it's possible for a paramagnet and a system at its critical point, where the correlation length diverges, to have the same value of C'. It does not seem particularly revealing that all systems with the same entropy density have the same statistical complexity.

As a contrast to the C' versus h_μ behavior shown in Fig. (1), consider Fig. (2), where we show the behavior of a different measure of statistical complexity, the *excess entropy* E [6, 8, 9, 11, 13, 14]. The excess entropy of an infinite configuration may be expressed as the mutual information between two semi-infinite halves of the configuration. That is, E is the amount of spatial memory embedded in configurations. In ref. [23] we show that E captures significant structural changes in the configurations of one-dimensional spin systems as external parameters are varied. In Fig. (2) we plot the excess entropy for 10^4 sets of parameter values for a 1D Ising system with nearest-neighbor coupling in the presence of an external field. Note that for *all* these Ising systems $C_{\rm LMC} = 0$, since $C_{\rm LMC}$ vanishes in the thermodynamic limit. Comparing the two figures, it is clear that the excess entropy depends on the entropy density h_μ in a much more subtle way than C' does.

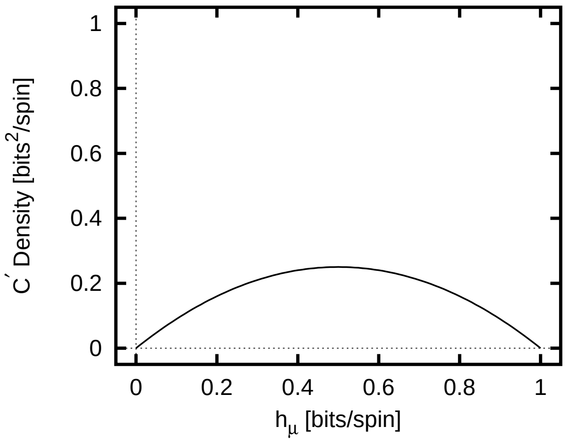

FIG. 1: C' versus entropy density h_μ. Note that C' is a function of h_μ.

Comparing these two plots raises another important issue. Note that the excess entropy does not always equal zero for $h_\mu = 0$, an apparent violation of the "boundary condition" requiring that a complexity measure vanish in the perfectly ordered limit. However, $h_\mu = 0$ corresponds to perfect *asymptotic predictability*, not perfect *order*. A process with a vanishing entropy rate indicates that it can be predicted without error—it says nothing, however, about how much effort or memory is required to perform this prediction. Thus, a zero value of the entropy density is too crude a measure of order. To see this, note that *any* periodic system has $h_\mu = 0$. Yet all periodic systems aren't equally ordered: a configuration with period 1 is certainly more ordered than a configuration of period 1729—which, for example, requires more memory to produce. In fact, at the period-doubling accumulation point of the logistic map, the symbolic dynamics produce periodic configurations of diverging periodicity. Hence, the excess entropy is infinite here, while the entropy rate remains zero [6, 7, 15].

In contrast to C', which vanishes for any periodic system, the excess entropy E for a configuration of period P is $\log_2 P$. Only if the period is 1, indicating trivial ordering and predictability, does the excess entropy vanish for a periodic process. Thus, the $(h_\mu, \mathrm{E}) = (0, 0)$ points in

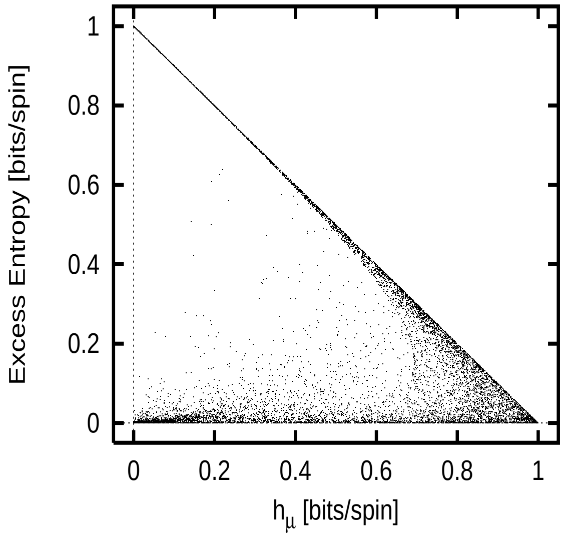

FIG. 2: Excess entropy E, a statistical complexity measure, versus entropy density h_μ for a spin-1/2 one-dimensional Ising spin system: 10^4 (h_μ, E) points. The system parameters were randomly chosen from the following intervals: J (coupling constant) $\in [-3, 3]$; T (temperature) $\in [0.05, 4.05]$; and B (external field) $\in [0, 3]$.

Fig. (2) correspond to the system's ferromagnetic ground states of period 1 and the $(h_\mu, \mathrm{E}) = (0, 1)$ points correspond to the antiferromagnetic ground states of period 2 [23]. Statistical complexity measures such as C' or, for that matter C_{LMC}, that are zero for all $h_\mu = 0$ configurations are very blunt implements with which to detect structure. A measure of complexity should be able to distinguish between structures of different periodicities.

For maximal randomness ($h_\mu = 1$), the excess entropy E vanishes, as expected. At $h_\mu = 1$, corresponding to infinite temperature, the spins decouple and there is no information shared between them. But the excess entropy does more than satisfy the "boundary conditions" of vanishing for $h_\mu = 0$ and 1. The interpretation of E as the memory stored in spatial configurations holds for intermediate values of h_μ as well. As a result, Fig. (2) lets us place an upper bound on the memory stored in spatial configurations for a spin-1/2 nearest neighbor Ising model: $\mathrm{E} \leq 1 - h_\mu$. This result, derived analytically in ref. [23], applies to all one-step Markov chains over a binary alphabet.

We close this section by noting that there is a growing body of evidence indicating that, aside from the requirement of vanishing at the ordered and disordered extremes, entropy and "complexity" (defined in a number of different ways to reflect "structure") are more or less independent. That is, there is a vastly wider range of complexity versus entropy relationships than indicated by eq. (15) [14, 15, 23]. Fig. (2) is just one example of many possible statistical complexity-entropy density relationships.

IV. CONCLUSION

To summarize, we have shown that C_{LMC} vanishes in the thermodynamic limit for finite-memory regular Markov chains. This class of systems includes, at a minimum, all finite-range one-dimensional spin systems. We have also shown that C_{LMC} is not an extensive variable. We have proposed modifying C_{LMC}, replacing the "disequilibrium" of ref. [26] with the relative entropy with respect to the uniform distribution. This results in a quantity C' that grows appropriately in the thermodynamic limit, making it possible to define a meaningful statistical complexity density that, nonetheless, retains the spirit of C_{LMC}. However, the product of this modification is a quantity that is a trivial, "over-universal" function of the entropy density h_μ. In short, based on the above observations, it seems to us that C_{LMC} and C' may be of little use in measuring the complexity of a statistical mechanical system.

We conclude by pointing out that the "boundary conditions" of vanishing in the extreme ordered and disordered limits do not uniquely specify a measure of complexity, an observation also made by Anteneodo and Plastino [27]. In fact, if this is the only feature one demands of a complexity measure, it's not clear to us why one would be motivated to devise a new statistic at all.

Statistical mechanics, for example, is replete with functions that vanish in the high and low temperature limits. Since thermodynamic entropy, a measure of randomness, is a monotonic function of temperature, high (low) temperature corresponds to high (low) randomness. Examples of quantities that vanish in these extremes (assuming there is not a critical point at $T = 0$) include the connected correlation functions, the correlation length, and magnetic susceptibility. These functions can be easily applied to any probability distribution describing a spatially or spatio-temporally extended collection of random variables.

Information theory also comes equipped with a function that vanishes for perfectly ordered and disordered systems: the mutual information I [1, 2]. If two random variables Z and Y are independently distributed, then the mutual information between them, $I[Z;Y]$ vanishes. At the other extreme, if Z and Y are both known with certainty—that is, $H[Z] = H[Y] = 0$—then $I[Z;Y]$ also vanishes. For statistical dependencies between these extremes, $I[Z;Y]$ is positive and measures the amount of information shared between Y and Z.

Given that there are many functions that vanish in the extreme ordered and disordered limits, it is clear that requiring this property does not sufficiently restrict a statistical complexity measure. What other criteria can we use, then, to guide us as we attempt to detect structures and patterns in nature? To this question we offer two suggestions.

First, it is helpful if the candidate statistical complexity has a clear interpretation: *What exactly is the statistical complexity measuring?* The two English words "statistical complexity" do not sufficiently answer the question. Many of the statistical complexity measures proposed over the last decade or so do have clear interpretations. For example, the excess entropy may be interpreted as the mutual information between two halves of an infinite configuration [6, 14, 23]. Logical depth is the run time required by a universal Turing machine executing the minimal program to reproduce a given pattern. These unambiguous interpretations help put these statistical complexity measures on a solid footing.

Second, it is essential to consider the motivations behind a measure of statistical complexity: *How is the measure to be used? What questions might it help answer?* It is possible to meaningfully assess its utility only if the motivations and goals for defining a complexity measure are stated clearly. One set of issues is the detection and quantification of patterns produced by a process. It has been proposed that particular notions of structure adapted from computation theory capture the intrinsic "patterns" and information processing architecture embedded in a system. In this setting, one finds well-defined and easily interpreted measures of statistical complexity [15]. For other views on questions that a measure of statistical complexity might help answer, see ref. [28].

Finally, ref. [27] mentions several different notions of complexity and notes that "there is not yet a consensus on a precise definition." "Complexity" has accepted meanings in other fields—meanings established prior to the recent attempts to use it as a label for structure in natural systems. For example, *Kolmogorov-Chaitin complexity* in algorithmic information theory means something quite different from *computational complexity* in the analysis of algorithms. These, in turn, are each different from the *stochastic complexity* used in model-order estimation in statistics [29]. Though at a future date relationships may be found, at present all of these are different from the notions of statistical complexity discussed here.

Unfortunately, "complexity" has been used without qualification by so many authors, both scientific and nonscientific, that the term has been almost stripped of its meaning. Given this state of affairs, it is even more imperative to state clearly why one is defining a measure of complexity and what it is intended to capture.

We thank Melanie Mitchell and Karl Young for helpful comments on the manuscript. This work was partially supported at UC Berkeley by ONR grant N00014-95-1-0524 and at the Santa Fe Institute by ONR grant N00014-95-1-0975.

[1] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1963.

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

[3] A. N. Kolmogorov. *Dokl. Akad. Nauk. SSSR*, 119:861, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.

[4] Ja. G. Sinai. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[5] B. A. Huberman and T. Hogg. *Physica D*, 22:376–384, 1986.

[6] P. Grassberger. *Intl. J. Theo. Phys.*, 25(9):907–938, 1986.

[7] J. P. Crutchfield and K. Young. *Phys. Rev. Lett.*, 63:105–108., 1989.

[8] J. P. Crutchfield and N. H. Packard. *Physica D*, 7:201–223, 1983.

[9] P. Szépfalusy and G. Györgyi. *Phys. Rev. A*, 33(4):2852, 1986.

[10] S. Wolfram. *Physica D*, 10:1–35, 1984.

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

[12] C. H. Bennett. *Found. Phys.*, 16:585, 1986.

[13] K. Lindgren and M. G. Norhdal. *Complex Systems*, 2(4):409–440, 1988.

[14] W. Li. *Complex Systems*, 5(4):381–99, 1991.

[15] J. P. Crutchfield. *Physica D*, 75:11–54, 1994.

[16] B. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, and H. Scheingraber. *Chaos, Solitons & Fractals*, 4(1):133–173, 1994.

[17] M. Gell-Mann and S. Lloyd. *Complexity*, 2(1):44–52, 1996.

[18] A. N. Kolmogorov. *Prob. Info. Trans.*, 1:1, 1965.

[19] G. Chaitin. *J. ACM*, 13:145, 1966.

[20] M. Li and P. M. B. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer-Verlag, New York, 1993.

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

[22] J. P. Crutchfield and J. E. Hanson. *Physica D*, 69:279–301, 1993.

[23] J. P. Crutchfield and D. P. Feldman. *Phys. Rev. E*, 55(2):1239R–1243R, 1997.

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

[25] H. Pagels and S. Lloyd. *Ann. Physics*, 188:186–213, 1988.

[26] R. Lòpez-Ruiz, H. L. Mancini, and X. Calbet. *Phys. Lett. A*, 209:321–326, 1995.

[27] C. Anteneodo and A. R. Plastino. *Phys. Lett. A*, 223:348–354, 1996.

[28] C. H. Bennett. In W. H. Zurek, editor, *Complexity, Entropy, and the Physics of Information*, pages 137–148. Addison-Wesley, 1990.

[29] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publisher, Singapore, 1989.

[30] R. Bronson. *Matrix Operations.* McGraw-Hill, 1989.

APPENDIX A: D VANISHES EXPONENTIALLY FAST FOR REGULAR MARKOV CHAINS

We will show that D goes to zero exponentially fast with increasing system size for a Markov chain governed by a regular transition matrix $T_{ab} = \Pr(b|a)$, where $0 \leq T_{ab} \leq 1$ and there exists a $K < \infty$ such that $(T^K)_{ab} > 0$

for all a and b. Since the conditional probabilities are normalized, the matrix T is stochastic: $\sum_{b=1}^{k} T_{ab} = 1$.

The probability of a block of L consecutive variables taking on the values $x_1, x_2, \ldots, x_L$ is given by

$$\Pr(x_1, x_2, \ldots, x_L) = p_{x_1} T_{x_1 x_2} T_{x_2 x_3} \cdots T_{x_{L-1} x_L} \ , \quad \text{(A1)}$$

where p is the stationary distribution of a single variable, as given by the left eigenvector of T with eigenvalue 1. The eigenvector p is chosen so as to be normalized in probability, $\sum_{a=1}^{k} p_a = 1$.

Let $\tilde{T}$ be a matrix whose components $\tilde{T}_{ab}$ are given by $(T_{ab})^2$. Note that $\tilde{T} \neq T^2$. Similarly, let $\tilde{p}$ be a vector whose components $\tilde{p}_a$ are given by $(p_a)^2$. Eq. (6) indicates that we are interested in

$$\sum_{\{x^L\}} \Pr(x^L)^2 = \sum_{\{x^L\}} \tilde{p}_{x_1} \tilde{T}_{x_1 x_2} \tilde{T}_{x_2 x_3} \cdots \tilde{T}_{x_{L-1} x_L} . \quad \text{(A2)}$$

The sum runs over all configurations of length L. The effect of the sum is to multiply the matrices together;

$$\sum_{\{x^L\}} \Pr(x^L)^2 = \sum_{x_1} \sum_{x_L} \tilde{p}_{x_1} (\tilde{T}^{L-1})_{x_1 x_L} \ . \quad \text{(A3)}$$

We shall show that in the $L \to \infty$ limit the above expression goes to zero exponentially fast.

We begin by considering the vector $V \equiv \tilde{p}\tilde{T}^{L-1}$ and its L_∞ norm, $\|V\| \equiv \max\{|V_1|, |V_2|, \ldots\}$. Eq. (A3) may be rewritten in terms of V,

$$\sum_{\{x^L\}} \Pr(x^L)^2 = \sum_{i=1}^{k} V_i . \quad \text{(A4)}$$

Since V is finite dimensional and all elements are nonnegative, if $\|V\|$ goes to zero exponentially fast, $\sum_{\{x^L\}} \Pr(x^L)^2$ must also go to zero exponentially fast. To show the former we use some well-known properties of vector and matrix norms [30].

Consider the matrix norm induced by the L_∞ vector norm:

$$\|\tilde{T}\| \equiv \max_a \{ \sum_b \tilde{T}_{ab} \} \ . \quad \text{(A5)}$$

Any matrix norm is compatible with its associated vector norm:

$$\|V\| = \|\tilde{p}\,\tilde{T}^{L-1}\| \leq \|\tilde{p}\|\,\|\tilde{T}^{L-1}\| \ . \quad \text{(A6)}$$

Recall that the components of $\tilde{p}$ are the squares of the components of the stationary probability p of the Markov chain. Except for the trivial case in which there is only one symbol in our chain and T is a one-by-one matrix, the maximum component of p is less than one. Thus, $0 < \|\tilde{p}\| < 1$ and we have

$$\|V\| \leq \|\tilde{p}\|\,\|\tilde{T}^{L-1}\| \leq \|\tilde{T}^{L-1}\| \ . \quad \text{(A7)}$$

Since T is regular and stochastic, there exists a K such that $0 < (T^K)_{ab} < 1$ for all a and b. Each element of T^K is a sum of terms that are products of T's elements. Likewise, each element of $\tilde{T}^K$ is a sum of terms that are products of $\tilde{T}$'s elements. However, since $\tilde{T}_{ab} = (T^2)_{ab}$ and $0 \leq T_{ab} \leq 1$, it follows that each component of $\tilde{T}^K$ is strictly less than the corresponding component of T^K. The product of stochastic matrices is itself a stochastic matrix, so $\sum_{b=1}^{k} (T^K)_{ab} = 1$. Thus, since each component of $\tilde{T}$ is less than the corresponding component of T, $\sum_{b=1}^{k} (\tilde{T}^K)_{ab} < 1$. As a result, $\|\tilde{T}\| < 1$.

Rewriting eq. (A7), we have: $\|V\| \leq \|\tilde{T}^{K(L-1)/K}\|$. Using the consistency condition $\|\tilde{T}^2\| \leq \|\tilde{T}\|\,\|\tilde{T}\|$, obeyed by all matrix norms, we see that in the $L/K \to \infty$ limit:

$$\|V\| \leq \|\tilde{T}^K\|^{L-1} \ . \quad \text{(A8)}$$

(The $L/K \to \infty$ limit is equivalent to the $L \to \infty$ limit since K is finite.) Since $\|\tilde{T}^K\| < 1$ we see that $\|V\|$ is bounded above by a function that decreases exponentially in L. Hence $\|V\|$ itself also decreases exponentially in L.