PHYSICAL REVIEW E VOLUME 61, NUMBER 4 APRIL 2000

Complexity and Fisher information

P.-M. Binder
Departamento de Física, Universidad de Los Andes, Apartado Aéreo 4976, Bogotá, Colombia
(Received 17 November 1999)

The connection between an observer-based theory of measurement and a hierarchical classification of complex systems in terms of topological exponents is discussed. This leads to generalized definitions of complexity that capture different aspects of the structure of the trajectory space of complex systems.

PACS number(s): 05.45.−a, 05.30.−d, 02.50.Wp, 89.80.+h

In recent years, an observer-based theory of measurement has been developed by Frieden *et al.* ([1,2] and references therein) in which many known physical laws can be derived from an extremum principle for the difference between the so-called Fisher information of a measurement and the information bound in the physical quantity being measured. In an apparently unrelated development (see [3] and references therein), many years of work about the meaning and definition of complexity in physical systems have led to a hierarchical definition by Badii and Politi [4] that combines computational and physical aspects, through the use of irreducible forbidden words and topological entropy. In this Rapid Communication, I establish and discuss a connection between these two theoretical advances. In particular, I show that the lowest-level definition of complexity, $C^{(1)}=K_0$, or topological entropy of a language, is related to a modified form of the Fisher information of the space of *trajectories* generated by the language. This suggests generalizations of the hierarchical definition so that it includes metric aspects and information about local minima. Based on this, I comment on higher-level definitions and the statistical inference process itself, which I conjecture to be related to an extremization of physical information performed by the observer. To begin, I will review Frieden's work on measurement and information theory and the hierarchical definition already mentioned.

In the 1920s the statistician Fisher [5] proposed a metric of an efficient measurement procedure (or channel capacity),

$$I=\int dx\,\frac{[p'(x)]^2}{p(x)}, \tag{1}$$

which is more convenient for our purposes in its discrete form,

$$I=\Delta x^{-1}\sum_n \frac{[p(x_{n+1})-p(x_n)]^2}{p(x_n)}. \tag{2}$$

In the above, $p(x)$ is the probability that a measurement will yield the value x, and the prime indicates a derivative. This metric differs greatly from the usual $\Sigma_n p(x_n)\log p(x_n)$ metric in that it contains local information about the distribution $p(x_n)$, in the form of derivatives. For example, I gives information about the ruggedness of a distribution; see Ref. [1], pp. 29–30. The Fisher metric is not very familiar to physicists, except perhaps for its relation with cross (or Kullback-Leibler) entropy [6,7] or its importance in time-series analysis; see [8], p. 373. Following Brillouin's treatment [9] of the relation between Boltzmann entropy and Shannon information in the measurement of the position of a particle, Frieden proposes that in addition to I there must be a second quantity, J, the bound information which is intrinsic to the physical phenomenon being measured. Then, from the variational principle $\delta(I-J)=0$, many of the Lagrangians used in physics can be derived. This is called by Frieden the principle of extreme physical information (EPI). For example, from an attempt to measure the classical four-position of a boson (fermion), the Klein-Gordon (Dirac) equation can be derived. Frieden's book (Ref. [1]) has further details on this work.

The Badii and Politi definitions apply to finite-alphabet sequences that represent the trajectory of a dynamical system in phase space, with one symbol per time step. Ideally, the symbols are determined by a generating partition related, for example, to homoclinic tangencies [10] or invariant manifolds [11]. The characterization of the particular language that corresponds to a given dynamical system in terms of topological exponents goes as follows. The first exponent $C^{(1)}$ is the large-m limit of $(1/m)\ln N(m)$, where $N(m)$ is the number of allowed words of length m. This limit is the topological entropy of the set of allowed words. $C^{(1)}$ clearly is a measure of the cardinality of the system. For higher exponents, one must find the set of irreducible forbidden words of the language, and the topological entropy of this set yields $C^{(2)}$. An irreducible word is one that cannot be decomposed into subwords strictly shorter than itself. $C^{(2)}$ is a measure of the difficulty of approximating the original language through subshifts of finite type with increasing memory (see [3], pp. 255–260). Next, one finds the topological entropy of the irreducible forbidden words in the set of irreducible forbidden words, which yields $C^{(3)}$, and so on successively. It is expected that $C^{(k+1)}\leqslant C^{(k)}$, and that eventually the topological exponents for a language become zero starting with some given integer k. An attractive feature of this classification is that it assigns $C^{(1)}=0$ to all periodic sequences, and $C^{(2)}=0$ to all random sequences. Further properties and examples are given in [3].

In what follows, consider a system for which we wish to infer information regarding its dynamics. So, instead of a single variable, we want each measurement to consist of a sequence of measurements as a function of time. Clearly, the simplest realization of this is a finite-alphabet string of length m and p possible symbols. This string will be one point in the possible m-dimensional space of trajectories, a hypercube in

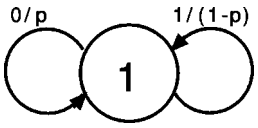

FIG. 1. One-state deterministic finite automaton with unequal transition probabilities. The arrow labels indicate symbol emitted and probability of the transition.

the two-symbol case. This space has $q=p^m$ points. The language of our particular dynamical system will produce different sequences (trajectories) with different probabilities, which we can label respectively x_n and $p(x_n)$ in our trajectory space. We can obtain purely topological information by modifying Eq. (2) in the following way:

$$I_m=\frac{1}{q}\sum_j \frac{H(|p(x_{j+1})-p(x_j)|)}{f[p(x_j)]}, \tag{3}$$

where H is the Heaviside step function, and the sum is evaluated only over the set of singular points for which $p(x_j)=0$. In other words, I_m counts the divergences of $1/f[p(x_n)]$. The denominator not only preserves the form of Fisher information, but also allows some flexibility: depending on whether f is zero or nonzero, we can count isolated or nonisolated forbidden words. Equation (3) then calculates the fraction of trajectories which are not allowed, and it follows that $I_m+K_0=1$.

So, while the motivation of Badii and Politi's topological exponents is derived largely from computer-theoretic ideas, we see that it relates to the structure of the space of trajectories: $p(x_n)$ can be seen as an energy landscape in m-dimensional space, for which both I_m and K_0 are counting the fraction of global minima (forbidden words), or its complement, and higher topological exponents attempt to further compress information about their location. It is natural, then, that the Badii and Politi measure should turn out to be related to the modified Fisher information, which describes local features (minima) of the distribution $p(x_n)$. The exact relation between topological exponents, I_m, and both the Shannon and Fisher information certainly deserves further study.

But, in general, energy landscapes can be quite complicated, and have a hierarchical structure with many local minima as well, as happens with spin glasses [12]. The structure of trajectory space is no exception to this, and ultimately may need to be described with ultrametric concepts [13], especially since our problem does share some features with the spin glass problem [14]: very high dimensionality, and a relatively sparse number of minima, especially after the compression through irreducible forbidden words is performed several times.

To give a simple example of how to add a metric element to K_0, consider a one-state finite automaton, with transitions to itself emitting a 0 with probability p, and a 1 with probability $1-p$; see Fig. 1. This system has no forbidden words, and $K_0=1$. However, we can now relax the conditions of Eq. (3), and calculate the contributions to the original Fisher information Eq. (2) from all points, not just from divergences; this may require a treatment similar to that given to four-vectors in [1], or approximating $p'(x)\sim\|\nabla p(x)\|$.

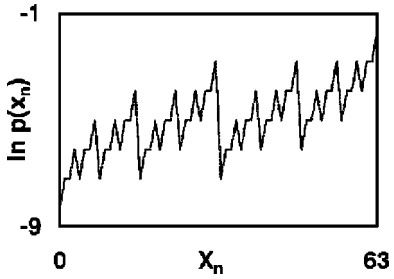

FIG. 2. Landscape $p(x_n)$ vs x_n in trajectory space, with $p=1/4$, $m=6$ for the previous figure (schematic).

When we do this, small differences will appear for different values of p. For example, the landscape is totally flat for $p=1/2$ ($I_m=0$, $K_0=1$), but not otherwise, because different (neighboring) trajectories will occur with different probabilities; see Fig. 2, with $p=1/4$ and $m=6$. Note that the trajectory space represented in this figure is really six-dimensional. Here trajectory labels (the horizontal axis) encode the corresponding binary trajectory, for example, 19=010 011, and the real topology is that of a six-dimensional hypercube: trajectory 7 has as neighbors 3, 5, 6, 15, 23, and 29 (those with Hamming distance equal to one).

For more complex languages the landscape can get extremely complicated. In that case, a characterization not only of global minima but also of the distribution of minima of $p(x_n)$ can be attempted by the methods of Badii and Politi. However, some languages do not have a metric structure and some of the comments above may not apply.

Conversely, topological exponents suggest possible generalizations of I, which include information about higher derivatives than the first. While this will probably not contribute to the derivation of physically meaningful Lagrangians, perhaps it can help to characterize probability distribution functions with a rich structure. For example, a metric of the form

$$I'=\int dx\,\frac{[p''(x)]^2}{p'(x)} \tag{4}$$

will pick out information about the number of extrema, in the same way that the standard Fisher information yields information about absolute minima. Note that an additional term in the numerator of the form $[1-\mathrm{sgn}\,p'(x)]$ is needed to single out the minima.

Finally, I will discuss a possible connection of I with the process of model inference. I will specifically consider systems describable by a deterministic finite automaton, for which there has been much recent work [15–18]. Consider a j-state automaton with transitions that emit only two symbols, 0 and 1 (see Fig. 3). In this case the ''physics'' is given

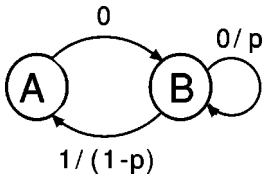

FIG. 3. Two-state deterministic finite automaton. The labels have the same convention as in Fig. 1.

by an allowed sequence of real states (e.g., *ABBBABA* . . .) visited by the system in time, while the measurement, as before, is the corresponding sequence of symbols (for this example, 0001010 . . .). The process of identifying particular segments of sequences with particular states of the system (for example, identifying every emitted one as a return to state *A*) is part of statistical inference theory.

There are several ways (see, for example, [16–18]) to infer from a long sequence which system states will lead to statistically equivalent futures. These are called causal states, and the models that are built on them are known as epsilon-machines; they can be inferred from measured data with help of binary trees, conditional distributions, and other methods. Many interesting properties of these models have been derived [17], and in specific cases states (and hence trajectory segments) can be associated with particular segments of the system's (not the trajectory's) phase space [18]. A goal of statistical inference is to allow prediction in the best possible way. This is achieved when the causal states correspond exactly with the system states, which is not always possible starting from a finite measurement sequence [19]. However, I conjecture that the best possible encoding of histories to system states will minimize $I-J$, this is, the difference between measurement and bound information; see [20], where the quality of an inferred model is measured through the Kullback-Leibler entropy, which is related to *I*. In this case, it is the statistician who is doing the EPI process. Note that the statistical inference process already described checks certain projections of the $p(x_n)$ landscape (the ''past,'' or initial $m/2$ symbols of a trajectory) to see if they have similar shapes in the complementary dimensions (the ''future,'' or final $m/2$ symbols). In this sense, this process is trying to pick out certain patterns in the $p(x_n)$ landscape which are different, and complementary to those of topological exponents. This leads to the following statement: definitions of complexity should provide information about the structure of $p(x_n)$ in trajectory space.

The results of this Rapid Communication can be summarized as follows: (i) topological entropy can be calculated, with a modified form of the Fisher information of the trajectories of a system in phase space, which counts forbidden words directly and is therefore related to the topological entropy of the space of trajectories. In contrast to the single-measurement situations considered by Frieden, here we are concerned with the *dynamics* of the system, and hence sequences of measurements. (ii) We can see the Badii and Politi measures as a characterization of the degree of structure in the *m*-dimensional trajectory space of the system, in particular, about the global minima in this space and the minimal encoding of information about their location. This suggests an extension of the $C^{(k)}$ measures to include a metric component, which has been illustrated with the case of a one-state deterministic finite automaton; this gives additional information about local minima of the distribution $p(x_n)$ in trajectory space. Conversely, this also leads to possible extensions of Fisher information to characterize additional structure in probability distribution functions. (iii) For the simple case of systems describable by a deterministic finite automaton, I conjecture that a statistical inference process that assigns optimal system state labels to particular trajectories or sets of trajectories corresponds to an extremization of physical information (EPI) performed by the scientist analyzing the system. This process requires the identification of patterns in $p(x_n)$ not necessarily related to global (or local) minima. (iv) Finally, combining the last two points suggests that all definitions of complexity should provide information about the structure of $p(x_n)$ in trajectory space.

The author thanks the Santa Fe Institute, where part of this work was done, for their hospitality.

[1] B. R. Frieden, *Physics from Fisher Information: A Unification* (Cambridge University Press, Cambridge, England, 1998).

[2] B. R. Frieden, A. Plastino, A. R. Plastino, and B. H. Soffer, Phys. Rev. E **60**, 48 (1999).

[3] R. Badii and A. Politi, *Complexity: Hierarchical Structure and Scaling in Physics* (Cambridge University Press, Cambridge, England, 1997).

[4] R. Badii and A. Politi, Phys. Rev. Lett. **78**, 444 (1997).

[5] R. A. Fisher, Philos. Trans. R. Soc. London **222**, 309 (1922).

[6] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).

[7] G. V. Vstovsky, Phys. Rev. E **51**, 975 (1995).

[8] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting* (Springer, New York, 1996). Fisher information is related to the properties of maximum likelihood estimators.

[9] L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1956).

[10] L. Jaeger and H. Kantz, Physica D **105**, 79 (1997); J. Phys. A **30**, L567 (1997), and references therein.

[11] F. Christiansen and A. Politi, Phys. Rev. E **51**, 3811 (1995).

[12] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1986).

[13] R. Rammal, G. Toulouse, and M. A. Virasoro, Rev. Mod. Phys. **58**, 765 (1986).

[14] See I. A. Campbell and L. de Arcangelis, Europhys. Lett. **13**, 587 (1990) for a study of the configuration space of an Ising spin glass and its relation to its dynamics and dynamical phase transitions.

[15] P. Grassberger, Z. Naturforsch. Teil A **43a**, 671 (1988).

[16] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).

[17] J. P. Crutchfield and C. R. Shalizi, Phys. Rev. E **59**, 275 (1999); C. R. Shalizi and J. P. Crutchfield, Santa Fe Institute, Working Paper No. WP-99-07-044, 1999 (unpublished). In these papers it is proved that epsilon-machines optimize prediction of the future, that the causal states of a process are a sufficient statistic for predicting the process, that they are simpler than any rival model which is as good at predicting, and that they have higher internal determinism than any rival model.

[18] N. Perry and P.-M. Binder, Phys. Rev. E **60**, 459 (1999).

[19] J. P. Crutchfield, in *Inside vs. Outside*, edited by H. Atmanspacher (Springer, Berlin, 1994), pp. 234–272.

[20] R. Badii, Europhys. Lett. **13**, 599 (1990).