Signatures of infinity: Nonergodicity and resource scaling in prediction, complexity, and learning

James P. Crutchfield and Sarah Marzen

Phys. Rev. E **91**, 050106 — Published 27 May 2015

DOI: 10.1103/PhysRevE.91.050106

Santa Fe Institute Working Paper 15-04-010
arxiv.org:1504.00386 [cond-mat.stat-mech]

Signatures of Infinity: Nonergodicity and Resource Scaling in Prediction, Complexity, and Learning

James P. Crutchfield[1, *] and Sarah Marzen[2, †]

[1] *Complexity Sciences Center and Department of Physics,*
University of California at Davis, One Shields Avenue, Davis, CA 95616
[2] *Department of Physics, University of California at Berkeley, Berkeley, CA 94720-5800*

We introduce a simple analysis of the structural complexity of infinite-memory processes built from random samples of stationary, ergodic finite-memory component processes. Such processes are familiar from the well known multi-arm Bandit problem. We contrast our analysis with computation-theoretic and statistical inference approaches to understanding their complexity. The result is an alternative view of the relationship between predictability, complexity, and learning that highlights the distinct ways in which informational and correlational divergences arise in complex ergodic and nonergodic processes. We draw out consequences for the resource divergences that delineate the structural hierarchy of ergodic processes and for processes that are themselves hierarchical.

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey
Keywords: mutual information, excess entropy, statistical learning theory, computational complexity, Kolmogorov-Chaitin complexity, asymptotic normality

I. INTRODUCTION

Truly complex stochastic processes—the *infinitary processes* [1] whose mutual information between past and future diverges—arise in many physical and biological systems [2–5], such as those in critical states. They are implicated in many natural phenomena, from the geophysics of earthquakes [6] and physiological measurements of neural avalanches [7] to semantics in natural language [8] and cascading failure in power transmission grids [9]. Their apparent infinite memory makes empirical estimation and modeling particularly challenging. The difficulty is reflected in the computational complexity of inference [10]: the resources required to predict and model them diverge in sample size, in memory for storing model parameters, and in memory required for prediction. Resource scaling suggests that for infinitary processes we look for statistical signatures that track divergences. Since resource divergences are sensitive to a process's inherent randomness and organization, one hopes that their scaling forms are uniquely revealing indicators of process complexity and can guide the selection of appropriate models.

To date, though, there are few tractable constructions with which to explore possible general relationships between prediction, complexity, and learning for infinitary processes. One of the few tractable and general constructions is the class of *Bandit processes* consisting of repeated trials of an experiment whose properties are, themselves, varying stochastically from trial to trial [11, 12]. Even if each individual trial is a realization generated by a stationary process with finite memory and exponentially decaying correlations, the resulting process over many trials can be infinitary [3–5].

Why can the past-future mutual information of Bandit processes diverge? The answer is remarkably simple: Bandit processes are nonergodic. More to the point, the divergence is driven by memory in the nonergodic part of their construction—the mechanism in each trial that selects and then remembers the operant ergodic component. Here, we use that insight to provide a simple, alternative derivation of information divergence for this class of infinitary process: a structural complexity scaling that directly accounts for nonergodicity.

Information divergence in Bandit processes has been interpreted as reflecting a universal property of learning: a unique indicator of the number of process parameters [3]. The derivation presented here recovers the connection between the complexity of parameter estimation and divergence in past-future information. However, it also identifies other structural features, such as infinitary ergodic components, that can drive divergences. Thus, information divergences in Bandit processes reflect particular structural properties of this class, rather than overarching principles of prediction, complexity, and learning for infinitary processes. Nonetheless, the issues raised highlight the need for a more balanced view of truly complex processes and their challenges. We hope our simplified analysis introduces tools appropriate to further, detailed scaling analysis of both ergodic and nonergodic infinitary processes.

Analyzing structural complexity is often conflated with statistical and computation-theoretic approaches to com-

[*] chaos@ucdavis.edu
[†] smarzen@berkeley.edu

plex processes. To ameliorate this, the next section reviews these alternatives. Then we move on to construct Bandit processes and analyze their structural complexity. We then discuss the results, draw out contrasts with computation-theoretic and statistical approaches, highlight the structural hierarchy of ergodic processes, and close with a brief discussion of hierarchical processes with nested organization.

II. PREDICTION, COMPLEXITY, AND LEARNING

There is a relationship between, on the one hand, the inherent unpredictability and memory in a process and, on the other, the difficulty of learning a model from time series samples and predicting the time series. Alternative framings lead to different views of this relationship. There are those that attempt to exactly describe a time series, those that try to express persistent regularities, and those that consider the consequences for inference. Their methods are closely related.

The *Kolmogorov-Chaitin complexity* monitors the computational resources—specifically, length of the minimal program for a given Universal Turing Machine (UTM)—required to reconstruct an individual time series [13–18]. It is a measure of randomness: A random time series has no smaller description than itself. Elaborating on this, *logical depth* [19] and *sophistication* [20] track complementary computational resources. Logical depth is the number of compute steps the minimal UTM program requires to generate the time series. Sophistication is the length of that part of the UTM program which captures regularities and organization, effectively discounting the time series' irreducible randomness. All these are uncomputable, though, even if one is given a generative model.

Fortunately, for a process' typical realizations the Kolmogorov-Chaitin complexity grows linearly with time series length, with coefficient equal to *Shannon source entropy rate* h_μ (a measure of a process' unpredictability) and offset equal to the *statistical complexity* C_μ (a measure of a process' memory) [21, and references therein]. Given a generative model called the *ϵ-machine*—a process' minimal maximally predictive model—both the entropy rate and statistical complexity are computable; if the ϵ-machine is finite, they are calculable in closed form [22].

We say that h_μ, C_μ, and the finite-time excess entropy discussed later are intrinsic measures of a process' structure, randomness, and organization. By *intrinsic*, we mean that these measures exist independently of the amount of data that we have observed. The aforementioned algorithmic complexities explicitly depend on the amount of data seen so far, but if the process is ergodic, then algorithmic complexities are also (almost always) intrinsic to a process in the limit of an arbitrarily large amount of data.

Such analyses of intrinsic properties should be contrasted with how statistical inference approaches complex processes. Statistical learning theory [23, 24] analyses and machine learning complexity controls [25–28] are not intrinsic in the sense that they show how to choose the best in-class model, but the choice of that class remains subjective. The problem of out-of-class modeling always exists as a practical necessity, but it is rarely, if ever, tackled directly. Of course, in the happy circumstance a correct generative model is in-class, then one has identified something intrinsic about a process. This, however, begs the question of discovering the class in the first place. And, practically, such luck is rarely the case. Worse, when they do not work well, complexity controls give no prescription for choosing an alternative class.

Intrinsic complexity characterizations have been most constructively and thoroughly developed for finite-memory, finite-randomness processes, despite the fact that many important natural processes are infinitary. The latter include the critical phenomena [29] of statistical physics and the routes to chaos in nonlinear dynamics [2], to mention only two. They exhibit arbitrarily long-range spatiotemporal correlations, infinite memory, and infinite parameter space dimension. The relationship between prediction, complexity, and learning is especially interesting when confronted with infinitary processes, and, paralleling Ref. [3], we re-investigate that relationship for nonergodic Bandit processes.

III. BANDIT PROCESS CONSTRUCTION

The simplest construction of a Bandit process is the following. Consider the stochastic process generated by a biased coin whose bias $\mathbf{P}$ is itself a random variable. First, a coin bias p is chosen from a user-specified distribution $\Pr(\mathbf{P})$; next, a bi-infinite sequence $\mathbf{x}^1 = \ldots x_{-1}x_0x_1x_2\ldots$ is generated from a coin with this particular bias; then, this is repeated for an arbitrarily large number of such trials; generating an ensemble $\{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \ldots\}$ of sequences at different biases. The process of interest is this sequence ensemble. We denote the random variable block between times a and b, but not that at time b, as $X_{a:b} = X_a X_{a+1} \ldots X_{b-1}$. We suppress denoting indices that are infinite. And so, the process of interest is denoted $X_:$. To denote the random variable block conditioned on a random variable Z taking realization z we use $X_{a:b}|Z = z$. So here, the subprocess $X_:|\mathbf{P} = p$ is that produced by a coin with bias p.

A single one of these bi-infinite sequences comes from an ergodic process that is memoryless in every sense of the word. In particular, since in each trial past and future are independent, the conditional past-future mutual information $I[X_{-M:0}; X_{0:N}|\mathbf{P} = p]$ vanishes for any M, N, and p. However, each of these bi-infinite chains is statistically distinct. The mean number of heads, say, in one is very different than the mean number of heads in another. For sufficiently long chains, such differences are almost surely not the consequence of finite-sample fluctuations.

The overall process $X_{:}$ does not distinguish between sequences generated by different biased coins. So, by making the coin bias a random variable, the past and future are no longer independent. Both share information about the underlying coin bias p. As we will now show, the shared information or *excess entropy* $\mathbf{E}(M, N) = I[X_{-M:0}; X_{0:N}]$ diverges with M and N when $\mathbf{P}$ is a continuous random variable.

IV. INFORMATION ANALYSIS

To see why, we abstract to a more general case. What follows is an alternative, direct derivation of results in Ref. [3, Sec. 4] that, due to its simplicity, lends additional transparency to the mechanisms driving the divergence.

Let Θ be a random variable with realizations θ in a (parameter) space of dimension K. Θ has some as-yet unspecified relationship with observations $X_{:} = \ldots X_{-2}, X_{-1}, X_0, X_1, \ldots$. We can always perform the following information-theoretic decomposition of the composite process's excess entropy:

$$\begin{aligned} I[X_{-M:0}; X_{0:N}] = I&[X_{-M:0}; X_{0:N}|\Theta] \\ &+ I[X_{-M:0}; X_{0:N}; \Theta] \ . \end{aligned} \tag{1}$$

The first term quantifies the range of temporal correlations of the observed process *given* Θ, and the second term quantifies the dependencies between past and future purely due to Θ. When the fixed-parameter process $X_{:}|\Theta = \theta$ is ergodic and the composite process $X_{:}$ is not, then Eq. (1) can be viewed as a decomposition of $I[X_{-M:0}; X_{0:N}]$ into ergodic and nonergodic contributions, respectively.

The second term $I[X_{-M:0}; X_{0:N}; \Theta]$ is a multivariate mutual information [30] or *co-information* [31]. It is closely related to parameter estimation, as expected [3], since it provides information about the dimension K of Θ. Standard information-theoretic identities yield:

$$\begin{aligned} I[X_{-M:0}; X_{0:N}; \Theta] = H[\Theta] &+ H[\Theta|X_{-M:N}] \\ &- H[\Theta|X_{-M:0}] - H[\Theta|X_{0:N}] \ . \end{aligned} \tag{2}$$

The first term $H[\Theta]$ quantifies our intrinsic uncertainty in the bias. When Θ is a continuous random variable, $H[\Theta]$ is a differential entropy. The subsequent terms describe how our uncertainty in Θ decreases after seeing blocks of lengths $M + N$, M, or N.

Altogether, Eqs. (1) and (2) give:

$$\begin{aligned} I[X_{-M:0}; X_{0:N}] = I[X_{-M:0}; X_{0:N}|\Theta] &+ H[\Theta] \\ + H[\Theta|X_{-M:N}] &- H[\Theta|X_{-M:0}] \\ - H[\Theta|X_{0:N}] \ . \end{aligned} \tag{3}$$

Thus, assuming one chose a prior with finite entropy $H[\Theta]$, divergences in $I[X_{-M:0}; X_{0:N}]$ can come from divergences in $I[X_{-M:0}; X_{0:N}|\Theta]$ or from divergences in $H[\Theta|X_{-M:N}] - H[\Theta|X_{-M:0}] - H[\Theta|X_{0:N}]$.

Let's take the cases covered in Ref. [3, Secs. 4.1-4.4]. There, Θ consists of the model parameters, θ are realizations of Θ, and $X_{:}|\Theta = \theta$ consists of (noisy, potentially temporally correlated) sequences generated by the model with parameters θ. For instance, Θ could be the firing rate of a Poisson neuron and $X_{:}|\Theta = \theta$ could be the time-binned spike trains at firing rate θ. Or, Θ could be transition probabilities in a finite Hidden Markov Model (HMM) and $X_{:}|\Theta = \theta$ could be the generated process given transition probabilities θ. The result, in any case, is a nonergodic process $X_{:}$ constructed from a mixture of ergodic component processes $X_{:}|\Theta = \theta$.

In these examples, the component-process excess entropy $I[X_{-M:0}; X_{0:N}|\Theta] = \langle I[X_{-M:0}; X_{0:N}|\Theta = \theta]\rangle_\theta$ does not diverge with M or N, since finite HMMs have finite excess entropy, which is bounded by the internal state entropy [4, 32]. In fact, the excess entropy for many ergodic stochastic processes is finite, even if generated by infinite-state HMMs. Any divergence in the composite process $I[X_{-M:0}; X_{0:N}]$ therefore comes from divergences in $H[\Theta|X_{-M:N}] - H[\Theta|X_{-M:0}] - H[\Theta|X_{0:N}]$.

Since the composite process includes sequences $\mathbf{x}^i$ from trials with different θ, one's intuition might suggest that $\Pr(\Theta = \theta|X_{-M:0} = x_{-M:0})$ is multimodal for most $x_{-M:0}$. However, existing results [33–36] on the asymptotic normality of posteriors carry over to this setting, since they essentially rely on the log-likelihood function $\log \Pr(X_{-M:0} = x_{-M:0}|\Theta = \theta)$ being sufficiently well behaved.

For instance, consider the Bandit process construction of Sec. III. A crude derivation of the asymptotic normality of $\Pr(\Theta = \theta|X_{-M:0} = x_{-M:0})$ [37] starts with Bayes Rule:

$$\begin{aligned} &\Pr(\Theta = \theta|X_{-M:0} = x_{-M:0}) \\ &\qquad = \frac{\Pr(X_{-M:0} = x_{-M:0}|\Theta = \theta)\Pr(\Theta = \theta)}{\Pr(X_{-M:0} = x_{-M:0})} \ . \end{aligned}$$

The denominator $\Pr(X_{-M:0} = x_{-M:0})$ is quite complicated to calculate, but this normalization factor does not affect the θ-dependence of $\Pr(\Theta = \theta|X_{-M:0} = x_{-M:0})$. More to the point, the prior's contribution $\Pr(\Theta = \theta)$ is dwarfed by the likelihood:

$$\Pr(|X_{-M:0} = x_{-M:0}|\Theta = \theta) \\ = \theta^{\sum_{i=0}^{M-1} x_i} (1-\theta)^{M - \sum_{i=0}^{M-1} x_i} ~,$$

in the large-M limit. Let θ^* be the unique maximum of $\Pr(\Theta = \theta | X_{-M:0} = x_{-M:0})$: $\theta^* = \frac{1}{M} \sum_{i=0}^{M-1} x_i + O(1/M)$. Taylor-expanding $\log \Pr(\Theta = \theta | X_{-M:0} = x_{-M:0})$ about θ^* suggests that $\Pr(\Theta = \theta | X_{-M:0} = x_{-M:0})$ is approximately normal in the large-M limit, with variance decaying as $\sim 1/M$. (Any one of the many sources [33–36] on asymptotic normality of posteriors provides rigorous and generalized statements.)

Armed with such asymptotic normality, we now turn our attention to find the asymptotic form of $H[\Theta | X_{-M:0} = x_{-M:0}]$, $H[\Theta | X_{0:N} = x_{0:N}]$, and $H[\Theta | X_{-M:N} = x_{-M:N}]$ in the large-M and -N limits. The differential entropy of a normal distribution is $\frac{1}{2} \log |\det \Sigma|$, where Σ is the covariance matrix; here, $\det \Sigma \sim 1/M$. This captures the error distribution for each of the K parameters. So, this and asymptotic normality of the posterior imply that: $H[\Theta | X_{-M:0} = x_{-M:0}] \sim -\frac{1}{2} K \log M$, plus corrections of $O(1)$ in M, and thus: $H[\Theta | X_{-M:0}] \sim -\frac{1}{2} K \log M$, where K is the parameter space dimension.

At first blush, the result is counterintuitive. In the limit that M and N tend to infinity, and we see longer and longer sequences $x_{-M:0}$, we become more certain as to Θ's value. This increasing certainty should mean that the conditional entropy $H[\Theta | X_{-M:0} = x_{-M:0}]$ vanishes. However, if Θ is a continuous random variable (such as a Poisson rate), then $H[\Theta | X_{-M:0} = x_{-M:0}]$ is a differential entropy. As our variance in $\Theta | X_{-M:0} = x_{-M:0}$ decreases to 0, the differential entropy $H[\Theta | X_{-M:0} = x_{-M:0}]$ diverges to negative infinity. It is exactly this well known divergence that causes a divergence in $I[X_{-M:0}; X_{0:N}]$ for the nonergodic processes we are considering.

From these results and Eq. (3), one has:

$$I[X_{-M:0}; X_{0:N}; \Theta] \sim \frac{K}{2} \log \frac{MN}{M+N} ~.$$

And, recalling that the ergodic-component information does not diverge, we immediately recover:

$$I[X_{-M:0}; X_{0:N}] \sim \frac{K}{2} \log \frac{MN}{M+N} ~. \tag{4}$$

Lower-order terms in M and N include the expected log-determinant of the Fisher information matrix for maximum likelihood estimates of Θ [38]. The joint divergence in past (M) and future (N) lengths is new here; cf. Ref. [3] which examined the case of $\mathbf{E}(-\infty, N)$.

A similar information-theoretic decomposition can be used to upper-bound the excess entropy of ergodic processes as well. For instance, App. A, uses a similar decomposition to show that the temporal excess entropy of an Ising spin on a two-dimensional Ising lattice at criticality is finite.

Logarithmic divergences in excess entropy also occur in stationary ergodic processes, such as exhibited at the onset of chaos through period-doubling [2]. And, alternative scalings are known, such as power-law divergences [3, Sec. 4.5]. For natural language texts there is empirical evidence that the excess entropy diverges. One form is referred to as Hilberg's Law [8, 39, 40]: $I[X_{-N:0}; X_{0:N}] \propto \sqrt{N}$.

In contrast with Sec. IV's rather direct calculation, it is far less straightforward to analyze these power-law divergences: $I[X_{:0}; X_{0:N}] \sim N^{\gamma}$, with $\gamma \in [0, 1)$. While there are results on asymptotics of posteriors for nonparametric Bayesian inference, many aim to establish asymptotic normality of the posterior; e.g., as in Refs. [41, 42]. As far as we know, no result yet recovers the aforementioned power-law divergence; likely, since existing asymptotic analyses avoid the essential singularity for the prior utilized in Ref. [3, Sec. 4.5] to obtain power-law divergence.

V. DISCUSSION

We investigated one large, but particular class of infinitary processes in terms of how information measures diverge; recovering, in short order, a previously reported logarithmic divergence in Bandit-like process past-future mutual information. Practically, this suggests that one could use the scaling of empirical estimates of past-future information as a function of sequence length to estimate a process's parameter space dimension.

Section IV's scaling analysis left open the possibility that information divergences can be driven by the ergodic components themselves. So, what is known about information divergences in ergodic processes? An information divergence hints at a structural level in the space of ergodic processes; a space that is itself highly organized. This is seen in the hierarchy of divergences separating processes into classes of distinct architecture, depicted in Fig. 1. (See also Table 1, Fig. 18, and Sec. 5 in Ref. [43].) Processes at each level are distinguished by different scalings for their complexity and in how difficult they are to learn and predict.

At the lowest level (*Markov*) are processes described by finite ϵ-machines with finite history dependence (finite Markov order R); e.g., those described by existing Maximum Caliber models [44] or by measure subshifts of finite type [45]. Though very commonly posited as models, they inhabit a vanishingly small measure in the space

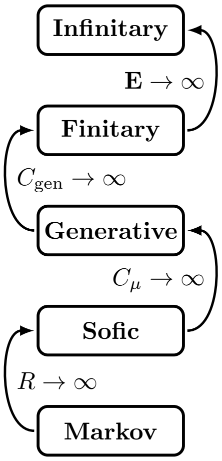

FIG. 1. Prediction hierarchy for stationary ergodic processes: Each level describes a process class with finite informational quantities. A class above finitely models the processes in the class below. Classes are separated by divergence in the corresponding informational quantity. Moving up the hierarchy corresponds to it diverging. Example processes that are finitely presented at each level, but infinitely presented at the preceding lower level. *Sofic*: typical unifilar HMMs, e.g., Even Process [1]; *Generative*: typical nonunifilar HMMs [32]; *Finitary*: typical infinite nonunifilar HMMs; *Infinitary*: highly atypical infinite HMMs with long-range memory, e.g., the ergodic construction in Ref. [4].

of processes [46]. At the next level (*Sofic*) of structure are processes described by ϵ-machines with finite C_μ. These typically have infinite Markov order; e.g., the measure-sofic processes. Above this level are processes generated by general (that is, nonunifilar) HMMs with uncountable recurrent causal states and divergent statistical complexity that, nonetheless, have finite generative complexity, $C_{\text{gen}} < \infty$ [32]. Processes at the *generative* level not only have infinite Markov order and storage, but also require a growing amount of memory for accurate prediction. One consequence is that they are inherently unpredictable by any observer with finite resources. Note, however, that predictability is complicated at all levels by *cryptic processes* [47]—those with arbitrarily small excess entropy, but large statistical complexity. When the smallest generative model is infinite but the process still has short-term memory, we arrive at the class of *finitary* processes ($\mathbf{E} < \infty$).

Processes with divergent excess entropy—infinitary processes—inhabit the upper reaches of this hierarchy. Predicting such processes necessarily requires infinite resources, but accurate prediction can also return infinite dividends. We agree, here, with Ref. [3]: the asymptotic rate of information divergence is a useful proxy for process complexity. Historically, this view appears to have been anticipated in Shannon's introduction of the *dimension rate* [48, App. 7] of an ensemble of functions: $\lambda = \lim_{\delta\to 0, \epsilon\to 0, T\to\infty} N(\epsilon,\delta,T)/T\log\epsilon$, where $N(\epsilon,\delta,T)$ is the smallest number of elements that can be chosen such that all elements of the ensemble, apart from a set of measure δ, are within the distance ϵ of at least one of those chosen.

However, it is as important to know which process mechanism drives the divergence as it is to know the divergence rate. Infinitary Bandit processes store memory entirely in their nonergodic component. Our analysis identified the divergence in this memory with the well known divergence in the differential entropy of highly peaked distributions of vanishing width. Generalizing Bandit processes to have *structured* ergodic components, we now see that even finite ϵ-machines trivially generate infinitary processes when their transition probabilities are continuous random variables.

Thus, in this case, we also agree that information divergence is a "necessary but not sufficient" criteria for process complexity [5]. (Appendix A, however, calls out a caveat.) This leaves open a broad challenge to understand the sufficient mechanisms for information divergences. For example, we have yet to develop similar informational and computation-theoretic analyses for the infinitary ergodic processes in Refs. [4, 5].

Looking forward, the simplicity of our structural complexity analysis opens up the possibility to better frame information in hierarchical processes [43, Sec. 5], such as the structural hierarchy in biology [49, Fig. 6], epochal evolution [50], and knowledge hierarchies in social systems such as semantics in human language [51]. These are processes in which multiple levels of mechanism are manifest and operate simultaneously and in which each level is separated from those below via phase transitions that lead to distinct signatures of informational and structural divergence.

ACKNOWLEDGMENTS

We thank N. Ay, W. Bialek, R. D'Souza, C. Ellison, C. Hillar, and I. Nemenman for useful conversations. This work was supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-12-1-0288, W911NF-13-1-0390, and W911NF-13-1-0340. SM was funded by NSF Graduate Student Research and UCB Chancellor's Fellowships.

Appendix A: Truly Complex Spin Systems?

Reference [5] pointed out that many infinitary processes do not satisfy intuitive definitions for complexity: divergence in $\mathbf{E}$ is a "necessary but not sufficient condition" for a process being truly complex. While intuitively compelling, perhaps divergent $\mathbf{E}$ is not even a necessary condition. Let's explain.

Spin systems at criticality are one of the most familiar examples of truly complex processes: global correlations emerge from purely local interactions [29]. Evidence of this complexity appears even if we are only allowed to observe a single spin's interaction with another on the lattice. At the critical temperature, the interaction has a power-law autocorrelation; at all other temperatures, the spin's autocorrelation is asymptotically exponential. The configurations' *spatial* excess entropy appears to diverge at criticality [52], too. However, does the *temporal* excess entropy $\mathbf{E}(M, N)$—roughly, the interaction a single spin with itself at later times—also diverge at criticality?

Surprisingly, it is finite, even at the critical temperature, *unless* there are nonlocal spatial interactions between lattice spins. Consider evolving the lattice configurations via Glauber dynamics for concreteness [29]. That is, spin j's next state σ^j_{t+1} is determined stochastically by its previous state σ^j_t and its effective magnetic field $h^j_t = \sum_i J_{ij}\sigma^i_t$. In other words, h^j_t and σ^j_t causally shield the past $\overleftarrow{\sigma}^j_t$ from the future $\overrightarrow{\sigma}^j_t$, implying that: $I[\sigma^j_{t-M:t};\sigma^j_{t+1:t+N}|h^j_t] = I[\sigma^j_t;\sigma^j_{t+1}|h^j_t] \leq H[\sigma^j_t]$. Given a finite set of spin values and local interactions, h^j_t can only take a finite number of values. Thus, $H[h^j_t] < \infty$, and so: $\left|I[\sigma^j_{t-M:t};\sigma^j_{t+1:t+N};h^j_t]\right| \leq H[h^j_t] < \infty$, as well.

A more familiar example makes this concrete. For the standard two-dimensional Ising lattice $J_{ij} = J$, if i and j are nearest neighbors, and $J_{ij} = 0$, otherwise. There, h^j_t can only take 5 possible values—$h^j \in \{0,\ J,\ 2J,\ 3J,$ and $4J\}$—giving: $\left|I[\sigma^j_{t-M:t};\sigma^j_{t+1:t+N};h^j_t]\right| \leq H[h^j_t] \leq \log_2 5$ bits.

The information-theoretic decomposition in Eq. (1) applies in this particular situation. Here, observed variables X_t are spins σ_t, and the parameters Θ are replaced by h^j. The bounds above then directly imply that $\mathbf{E}(M, N) < \infty$ for all M and N. In fact, for the standard two-dimensional Ising lattice, we find that $\mathbf{E}(-\infty, \infty) \leq 1 + \log_2 5 = 3.4$ bits. We expect excess entropy to diverge only when h^j is a continuous random variable. This can happen when J_{ij} is nonzero for an infinite number of i's. However, this necessitates global, not local, spin-spin couplings.

This does not negate $\mathbf{E}$'s utility as a generalized order parameter [53]. It is still likely maximized at the critical point, even if its temporal version does not diverge. Rather, our analysis shows that phenomena—here, spin lattices with purely local couplings—do not necessarily have divergent $\mathbf{E}$ even when many would consider their dynamics to be truly complex at criticality.

At first glance, this contradicts the experiments in Fig. 1 of Ref. [3] for the Ising lattice with local interactions. A careful look reveals that there is none: coupling strengths were randomly changed every $400,000$ iterations. So, the resultant time series is a concatenation of samples from a Bandit process. Section IV then predicts the observed logarithmic scaling in Fig. 1 there for $N \lesssim 25$. However, it also implies that $\mathbf{E}(-\infty, N)$ will stop increasing logarithmically at or before $N = 400,000$.

[1] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[2] J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. Zurek, editor, *Entropy, Complexity, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, pages 223 – 269, Reading, Massachusetts, 1990. Addison-Wesley.

[3] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.

[4] N. Travers and J. P. Crutchfield. Infinite excess entropy processes with countable-state generators. *Entropy*, 16:1396–1413, 2014.

[5] L. Debowski. On hidden Markov processes with infinite excess entropy. *J. Theo. Prob.*, 27(2):539–551, 2012.

[6] D. L. Turcotte. *Fractals and Chaos in Geology and Geophysics.* Cambridge University Press, Cambridge, United Kingdom, second edition, 1997.

[7] J. M. Beggs and D. Plenz. Neuronal avalanches in neocortixal circuits. *J. Neurosci.*, 23(35):11167–11177, 2003.

[8] L. Debowski. Excess entropy in natural language: Present state and perspectives. *CHAOS*, 21(3):037105, 2011.

[9] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *CHAOS*, 17(2):26103, 2007.

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

[11] T. M. Cover and M. E. Hellman. The two-armed-bandit problem with time-invariant finite memory. *IEEE Trans. Info. Th.*, IT-16(2):185–195, 1970.

[12] D. A. Berry. A Bernoulli two-armed bandit. *Ann. Math. Stat.*, 43(3):871–897, 1972.

[13] A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.

[14] G. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 13:145, 1966.

[15] P. Martin-Lof. The definition of random sequences. *Info. Control*, 9:602–619, 1966.

[16] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russ. Math. Surveys*, 38:29–40, 1983.

[17] A. A. Brudno. Entropy and the complexity of the trajectories of a dynamical system. *Trans. Moscow Math.*

Soc., 44:127, 1983.

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

[19] C. H. Bennett. Dissipation, information, computational complexity, and the definition of organization. In D. Pines, editor, *Emerging Syntheses in the Sciences.* Addison-Wesley, Redwood City, 1988.

[20] M. Koppel and H. Atlan. An almost machine-independent theory of program-length complexity, sophistication, and induction. *Information Sciences*, 56(1-3):23–33, 1991.

[21] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012.

[22] J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. 2013. Santa Fe Institute Working Paper 13-09-028; arXiv:1309.3792 [cond-mat.stat-mech].

[23] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge, Cambridge, United Kingdom, 2003.

[24] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Spinger, New York, second edition, 2011.

[25] C. S. Wallace and F. P. Freeman. Estimation and inference by compact coding. *J. R. Statist. Soc. B*, 49:240, 1987.

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

[27] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Auto. Control*, 19(6):716–723, 1974.

[28] H. Akaike. An objective use of Bayesian models. *Ann. Inst. Statist. Math.*, 29A:9, 1977.

[29] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman. *The Theory of Critical Phenomena.* Oxford University Press, Oxford, 1992.

[30] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *CHAOS*, 21(3):037109, 2011.

[31] A. J. Bell. The co-information lattice. In S. Makino S. Amari, A. Cichocki and N. Murata, editors, *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*, volume ICA 2003, New York, 2003. Springer.

[32] W. Lohr. Properties of the statistical complexity functional and partially deterministic HMMs. *Entropy*, 11(3):385–401, 2009.

[33] A. M. Walker. On the asymptotic behaviour of posterior distributions. *J. Roy. Stat. Soc. Series B (Methodological)*, 31:80–88, 1969.

[34] C. C. Heyde and I. M. Johnstone. On asymptotic posterior normality for stochastic processes. *J. Roy. Stat. Soc. Series B (Methodological)*, 41(2):184–189, 1979.

[35] T. J. Sweeting. On asymptotic posterior normality in the multiparameter case. *Bayesian Statistics*, 4:825–835, 1992.

[36] R. C. Weng and W-C. Tsai. Asymptotic posterior normality for multiparameter problems. *J. Stat. Plan. Infer.*, 138(12):4068–4080, 2008.

[37] J. A. Hartigan. Asymptotic normality of posterior distributions. In *Bayes Theory*, pages 107–118. Springer, 1983.

[38] E. L. Lehman and G. Casella. *Theory of Point Estimation.* Springer, second edition, 1998.

[39] W. Ebeling and G. Nicolis. Entropy of symbolic sequences: The role of correlations. *Europhys. Lett.*, 14:191–196, 1991.

[40] W. Ebeling and T. Poschel. Entropy and long-range correlations in literary English. *Europhys. Lett.*, 26:241–246, 1994.

[41] L. Wasserman. Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 293–304. Springer, 1998.

[42] S. Ghosal. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68, 2000.

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

[44] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.*, 85:1115–1141, Jul 2013.

[45] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding.* Cambridge University Press, New York, 1995.

[46] R. G. James, J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Many roads to synchrony: Natural time scales and their algorithms. *Phys. Rev. E*, 89:042135, 2014.

[47] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield. Prediction, retrodiction, and the amount of information stored in the present. *J. Stat. Phys.*, 136(6):1005–1034, 2009.

[48] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

[49] Q. Shen, Q. Hao, and S. M. Gruner. Macromolecular phasing. *Physics Today*, 59(3):46–52, 2006.

[50] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell. Finite populations induce metastability in evolutionary search. *Phys. Lett. A*, 229:144–150, 1997.

[51] N. Chomsky. Three models for the description of language. *IRE Trans. Info. Th.*, 2:113–124, 1956.

[52] H. W. Lau and P. Grassberger. Information theoretic aspects of the two-dimensional Ising model. 2012. arxiv.org:1210.5707 [cond-mat.stat-mech].

[53] D. P. Feldman. *Computational Mechanics of Classical Spin Systems.* PhD thesis, University of California, Davis, 1998. Published by University Microfilms Intl, Ann Arbor, Michigan.