PHYSICAL REVIEW E **95**, 051301(R) (2017)

Nearly maximally predictive features and their dimensions

Sarah E. Marzen[1,2,*] and James P. Crutchfield[3,†]

[1]*Physics of Living Systems Group, Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[2]*Department of Physics, University of California at Berkeley, Berkeley, California 94720-5800, USA*

[3]*Complexity Sciences Center, Department of Physics, University of California at Davis, One Shields Avenue, Davis, California 95616, USA*

(Received 27 February 2017; published 25 May 2017)

Scientific explanation often requires inferring maximally predictive features from a given data set. Unfortunately, the collection of minimal maximally predictive features for most stochastic processes is uncountably infinite. In such cases, one compromises and instead seeks nearly maximally predictive features. Here, we derive upper bounds on the rates at which the number and the coding cost of nearly maximally predictive features scale with desired predictive power. The rates are determined by the fractal dimensions of a process' mixed-state distribution. These results, in turn, show how widely used finite-order Markov models can fail as predictors and that mixed-state predictive features can offer a substantial improvement.

DOI: 10.1103/PhysRevE.95.051301

Often, we wish to find a minimal maximally predictive model consistent with available data. Perhaps we are designing interactive agents that reap greater rewards by developing a predictive model of their environment [1–6] or, perhaps, we wish to build a predictive model of experimental data because we believe that the resultant model gives insight into the underlying mechanisms of the system [7,8]. Either way, we are almost always faced with constraints that force us to efficiently compress our data [9].

Ideally, we would compress information about the past without sacrificing any predictive power. For stochastic processes generated by finite unifilar hidden Markov models (HMMs), one need only store a finite number of predictive features. The minimal such features are called *causal states*, their coding cost is the *statistical complexity* C_μ [10], and the implied unifilar HMM is the ϵ-*machine* [10,11]. However, most processes require an infinite number of causal states [7] and so cannot be described by finite unifilar HMMs.

In these cases, we can only attain some maximal level of predictive power given constraints on the number of predictive features or their coding cost. Equivalently, from finite data we can only infer a finite predictive model. Thus, we need to know how our predictive power grows with available resources.

Recent work elucidated the tradeoffs between resource constraints and predictive power for stochastic processes generated by countable unifilar HMMs or, equivalently, described by a finite or countably infinite number of causal states [12–14]. Few, though, studied this tradeoff or provided bounds thereof more generally.

Here, we place bounds on resource-prediction tradeoffs in the limit of nearly maximal predictive power for processes with either a countable or an uncountable infinity of causal states by coarse-graining the *mixed-state simplex* [15]. These bounds give an operational interpretation to the fractal dimension of the mixed-state simplex and suggest routes towards quantifying the memory stored in a stochastic process when, as is typical, statistical complexity diverges.

Background. We consider a discrete-time, discrete-state stochastic process $\mathcal{P}$ generated by an HMM $\mathcal{G}$, which comes equipped with underlying states g and labeled transition matrices $T^x_{g,g'} = \Pr(\mathcal{G}_{t+1} = g', X_{t+1} = x | \mathcal{G}_t = g)$ [16]. There is an infinite number of alternate HMMs that generate $\mathcal{P}$ [17–20], so we specify here that $\mathcal{G}$ is the *minimal* generative model—that is, the generative model with the minimal number of hidden states consistent with the observed process [21].

For reasons that become clear shortly, we are interested in the *block entropy* $H(L) = H[X_{0:L}]$, where $X_{a:b} = X_a, X_{a+1}, \ldots, X_{b-1}$ is a contiguous block of random variables generated by $\mathcal{G}$. In particular, its growth—the *entropy rate* $h_\mu = \lim_{L\to\infty} H(L)/L$—quantifies a process's intrinsic "randomness". Finite-length entropy-rate estimates $h_\mu(L) = H[X_0|X_{-L:0}]$ provide increasingly better approximations to the true entropy rate h_μ as L grows large.

The *excess entropy* $\mathbf{E} = \lim_{L\to\infty}[H(L) - h_\mu L]$ quantifies how much is predictable: how much future information can be predicted from the past [22]. Finite-length excess-entropy estimates,

$$\mathbf{E}(L) = H(L) - h_\mu L \tag{1}$$

$$= \sum_{\ell=0}^{L-1} [h_\mu(\ell) - h_\mu], \tag{2}$$

tend to the true excess entropy $\mathbf{E}$ as L grows large [23]. As there, we consider only *finitary* processes, those with finite $\mathbf{E}$ [24].

Predictive features $\mathcal{R}$ from some alphabet $\mathcal{F}$ are formed by compressing the process's past $X_{-\infty:0}$ in ways that implicitly retain information about the future $X_{0:\infty}$. (From here on, our block notation suppresses infinite indices.) A *predictive distortion* quantifies the predictability lost after such a coarse graining:

$$d(\mathcal{R}) = I[X_{:0}; X_{0:}|\mathcal{R}]$$

$$= \mathbf{E} - I[\mathcal{R}; X_{0:}].$$

We choose to use an informational distortion, though in principle, any other predictive distortion would do; for instance, bounds in Appendix A of the Supplemental Material [25] have bearing on total variation. On the one hand, the informational distortion considered here achieves its maximum value $\mathbf{E}$ when

*semarzen@mit.edu

†chaos@ucdavis.edu

$\mathcal{R}$ captures no information from the past that could be used for prediction. On the other, it can be made to vanish trivially by taking the predictive features $\mathcal{R}$ to be all of the possible histories $X_{:0}$.

Results. Ideally, we would identify the minimal number $|\mathcal{F}|$ of predictive features or the coding cost $H[\mathcal{R}]$ [9,26] required to achieve at least a given level d of predictive distortion. This is almost always a difficult optimization problem. However, we can place upper bounds on $|\mathcal{F}|$ and $H[\mathcal{R}]$ by constructing suboptimal predictive feature sets that achieve predictive distortion d.

We start by reminding ourselves of the optimal solution in the limit that $d = 0$—the *causal states* $\mathcal{S}$ [10,11,13]. Causal states can be defined by calling two pasts, $x_{:0}$ and $x'_{:0}$, *equivalent* when by using them our predictions of the future are the same: $\Pr(X_{0:}|X_{:0} = x_{:0}) = \Pr(X_{0:}|X_{:0} = x'_{:0})$. (These conditional distributions are referred to as *future morphs*.) One can then form a model, the ϵ*-machine*, from the set $\mathcal{S}$ of causal-state equivalence classes and their transition operators. The Shannon entropy of the causal state distribution is the *statistical complexity*: $C_\mu = \mathrm{H}[\mathcal{S}]$. A process's ϵ-machine can also be viewed as the minimal unifilar HMM capable of generating the process [27] and C_μ the amount of historical information the process stores in its causal states. Though the mechanics of working with causal states can become rather involved, the essential idea is that causal states are designed to capture everything about the past relevant to predicting the future and *only* that information.

In the lossless limit, when $d = 0$, one cannot find predictive representations that achieve $|\mathcal{F}|$ smaller than $|\mathcal{S}|$ or that achieve $\mathrm{H}[\mathcal{R}]$ smaller than C_μ. Similarly, no optimal lossy predictive representation will ever find $|\mathcal{F}| > |\mathcal{S}|$ or $\mathrm{H}[\mathcal{R}] \geqslant C_\mu$. When the number of causal states is infinite or statistical complexity diverges, as is typical, as we noted, these bounds are quite useless, but otherwise, they provide a useful calibration for the feature sets proposed below.

Markov features. Several familiar predictive models use pasts of length L as predictive features. This feature set can be thought of as constructing an order-L Markov model of a process [28]. The implied predictive distortion is $d(L) = I[X_{:0}; X_{0:}|X_{0:L}] = \mathbf{E} - \mathbf{E}(L)$ [29], while the number of features is the number of length-L words with nonzero probability and their entropy $H[\mathcal{R}] = H(L)$. Generally, $h_\mu(L)$ converges exponentially quickly to the true entropy rate h_μ for stochastic processes generated by finite-state HMMs, unifilar or not; see Ref. [30] and references therein. Then, $h_\mu(L) - h_\mu \sim K e^{-\lambda L}$ in the large L limit. From Eq. (2), we see that the convergence rate λ also implies an exponential rate of decay for $\mathbf{E} - \mathbf{E}(L) \sim K' e^{-\lambda L}$. Additionally, according to the asymptotic equipartition property [26], when L is large, $|\mathcal{F}| \sim e^{h_0 L}$ and $H(L) \approx h_\mu L$, where h_0 is the *topological entropy rate* and h_μ is the entropy rate, both in nats.

In sum, this first set of predictive features—effectively, the construction of order-L Markov models—yields an algebraic tradeoff between the size of the feature set $\mathcal{F}$ and the predictive distortion d:

$$|\mathcal{F}| \sim \left(\frac{1}{d}\right)^{h_0/\lambda}, \tag{3}$$

and a logarithmic tradeoff between the entropy of the features and distortion:

$$H[\mathcal{R}] \sim \frac{h_\mu}{\lambda} \ln\left(\frac{1}{d}\right). \tag{4}$$

In principle, λ can be arbitrarily small, and so h_0/λ and h_μ/λ arbitrarily large, even for processes generated by finite-state HMMs. This can be true even for finite unifilar HMMs (ϵ-machines). To see this, let W be the transition matrix of a process's mixed-state presentation, defined shortly. From Ref. [29], when W's spectral gap γ is small, we have $\mathbf{E} - \mathbf{E}(L) \sim (1-\gamma)^L$, so that $\lambda = \ln \frac{1}{1-\gamma}$ can be quite small.

In short, when the process in question has only a finite number of causal states, $|\mathcal{F}|$ optimally saturates at $|\mathcal{S}|$ and $H[\mathcal{R}]$ optimally saturates at C_μ, but from Eqs. (3) and (4), the size of this Markov feature set can grow without bound when we attempt to achieve zero predictive distortion.

Mixed-state features. A different predictive feature set comes from coarse graining the mixed-state simplex. As described by Blackwell [15], mixed states Y are probability distributions $\Pr(\mathcal{G}_0|X_{-L:0} = x_{-L:0})$ over the internal states $\mathcal{G}$ of a generative model given the generated sequences. Transient mixed states are those at finite L, while recurrent mixed states are those remaining with positive probability in the limit that $L \to \infty$. Recurrent mixed states exactly correspond to causal states $\mathcal{S}$ [31]. When C_μ diverges, recurrent mixed states often lay on a Cantor set in the simplex; see Fig. 1. In this circumstance, one examines the various dimensions that describe the scaling in such sets. Here, for reasons that will become clear, we use the box counting $\dim_0(Y)$ and information dimensions $\dim_1(Y)$ [32].

More concretely, we partition the simplex into cubes of side length ϵ, and each nonempty cube is taken to be a predictive feature in our representation. The number of nonempty cubes is denoted N_ϵ. When there are only a finite number of causal states, then this feature set consists of the causal states $\mathcal{S}$ for some nonzero ϵ. There is no such ϵ if, instead, the process has a countable infinity of causal states. Sometimes, as for the finite $|\mathcal{S}|$ case, the information dimension and perhaps even the box-counting dimension of the mixed states will vanish. When there is an uncountable infinity of causal states and ϵ is sufficiently small, the corresponding number of features scales as:

$$|\mathcal{F}| \sim \left(\frac{1}{\epsilon}\right)^{\dim_0(Y)},$$

where $\dim_0(Y)$ denotes the box-counting dimension [33] and the coding cost scales as:

$$H[\mathcal{R}] \sim \dim_1(Y) \ln \frac{1}{\epsilon},$$

where $\dim_1(Y)$ is the information dimension. These dimensions can be approximated numerically; see Fig. 1 and Appendix A of the Supplemental Material [25].

For processes generated by infinite ϵ-machines, finding the better feature set requires analyzing the scaling of predictive distortion with ϵ. Appendix B of the Supplemental Materials [25] shows that $d(\mathcal{R})$ at coarse graining ϵ scales at most as ϵ in the limit of asymptotically small ϵ. Our upper bound relies on the fact that two nearby mixed states have similar

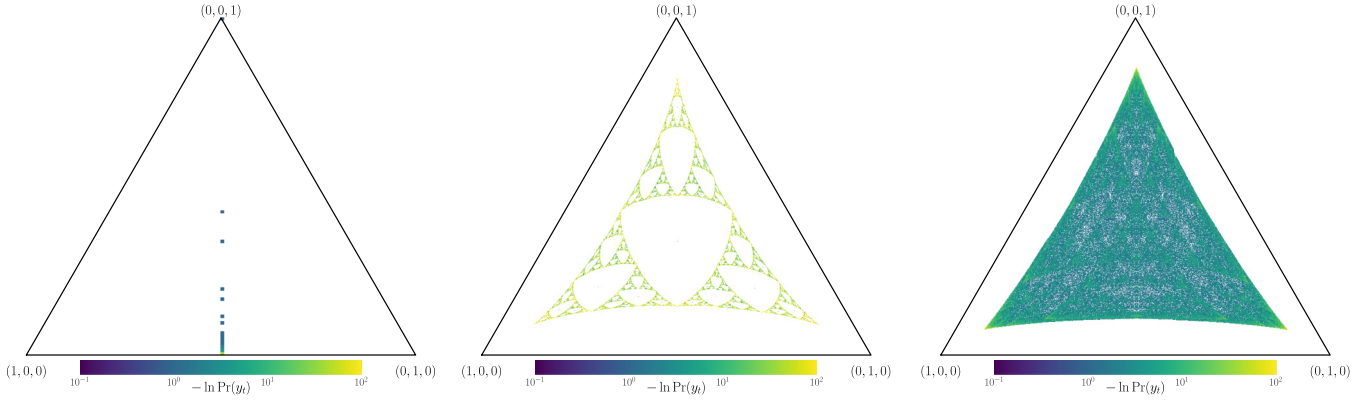

FIG. 1. Mixed-state presentations Y for three different generative models given in Appendix A of the Supplemental Materials [25]. To visualize the mixed-state simplex, the diagrams plot iterates y_t such that the left vertex corresponds to the pure state $\Pr(A,B,C) = (1,0,0)$ or HMM state A, the right to pure state $(0,1,0)$ or HMM state B, and the top to pure state $(0,0,1)$ or HMM state C. Left plot has 55 mixed states plotted in a 1×100 bin histogram; middle plots 2,391,484 mixed states in a 1000×1000 bin histogram; and right plots 21,523,360 mixed states in a 4000×4000 bin histogram. Bin cell coloring is negative logarithm of the normalized bin counts. The box-counting dimensions are respectively $\dim_0(Y) \approx 0.3$ at left, $\dim_0(Y) \approx 1.8$ in the middle, and $\dim_0(Y) \approx 1.9$ at right. Box-counting dimension calculated by estimating the slope of $\ln(1/\epsilon)$ vs $\ln N_\epsilon$ as described in Appendix A of the Supplemental Material [25].

future morphs, i.e., they have similar conditional probability distributions over future trajectories given one's mixed state. From this, we conclude that:

$$|\mathcal{F}| \sim \left(\frac{1}{d}\right)^{\dim_0(Y)} \tag{5}$$

and:

$$H[\mathcal{R}] \sim \dim_1(Y) \ln \frac{1}{d}. \tag{6}$$

In general, both $\dim_0(Y)$ and $\dim_1(Y)$ are bounded above by $|\mathcal{G}|$, the number of states in the minimal generative model.

Resource-prediction tradeoff. Finally, putting Eqs. (3)–(6) together, we find that for a process with an uncountably infinite ϵ-machine, the necessary number of predictive features $|\mathcal{F}^*|$ scales no faster than:

$$|\mathcal{F}^*| \lesssim \left(\frac{1}{d}\right)^{\min(h_0/\lambda, \dim_0(Y))} \tag{7}$$

and the requisite coding cost $H[\mathcal{R}^*]$ scales no faster than:

$$H[\mathcal{R}^*] \lesssim \min(h_\mu/\lambda, \dim_1(Y)) \ln \frac{1}{d}, \tag{8}$$

where d is predictive distortion.

These bounds have a practical interpretation. From finite data, one can only justify inferring finite-state minimal maximally predictive models. Indeed, the criteria of Ref. [34] applied as in Ref. [13] suggests that the maximum number of inferred states should not yield predictive distortions below the noise in our estimate of predictive distortion d from T data points. This noise scales as $\sim 1/\sqrt{T}$, since from T data points, we have approximately T separate measurements of predictive distortion. This, in turn, sets upper bounds on $|\mathcal{F}^*| \lesssim T^{(1/2)\min(h_0/\lambda, \dim_0(Y))}$ and $H[\mathcal{R}^*] \lesssim \min(h_\mu/\lambda, \dim_1(Y)) \ln T$ when the process in question has an uncountable infinity of causal states.

We can test this prediction directly using the Bayesian structural inference (BSI) algorithm [35] applied to the processes described in Fig. 1. The major difficulty in employing BSI is that one must specify a list of ϵ-machine topologies to search over. Since the number of such topologies grows superexponentially with the number of states [36], experimentally probing the scaling behavior of the number of inferred states with the amount of available data is, with nothing else said, impossible.

However, "expert knowledge" can cull the number of ϵ-machine topologies that one should search over. In this spirit, we focus on the simple nonunifilar source (SNS), since ϵ-machines for renewal processes have been characterized in detail [37–40]. The SNS's generative HMM is given in Fig. 2 (left). This process has a mixed-state presentation similar to that of "nond" in Fig. 1 (left). We choose to only search over the ϵ-machine topologies that correspond to the class of eventually Poisson renewal processes.

We must first revisit the upper bound in Eq. (7), as the SNS has a countable (not uncountable) infinity of causal states. When a process's ϵ-machine is countable instead of uncountable, then we can improve upon Eq. (7). However, the magnitude of improvement is process dependent and not easily characterized in general for two main reasons. First, predictive distortion can decrease faster than ϵ for processes generated by countably infinite ϵ-machines since there might only be one mixed state in an ϵ hypercube. (See the left-hand factor in the Supplemental Material, Eq. (B5) [25], $1 - \sum_{r \in \mathcal{R}^{(\epsilon)}: H[Y|\mathcal{R}^{(\epsilon)}=r]=0} \pi(r)$. We are guaranteed that $\lim_{\epsilon \to 0} \sum_{r \in \mathcal{R}^{(\epsilon)}: H[Y|\mathcal{R}^{(\epsilon)}=r]=0} \pi(r) = 0$, but the rate of convergence to zero is highly process dependent.) Second, the scaling relation between N_ϵ and $1/\epsilon$ may be subpower law.

Given a specific process, though, with a countably infinite ϵ-machine—here, the SNS—we can derive expected scaling relations. See Appendix C in the Supplemental Material [25]. We argue that, roughly speaking, we should expect predictive distortion to decay exponentially with N_ϵ, so that $d(\mathcal{R}_\epsilon) \propto \lambda^{N_\epsilon}$ for some $\lambda < 1$. In Appendix C of the Supplemental Material, we empirically argue that N_ϵ scales as $(1/\epsilon)^2$ when

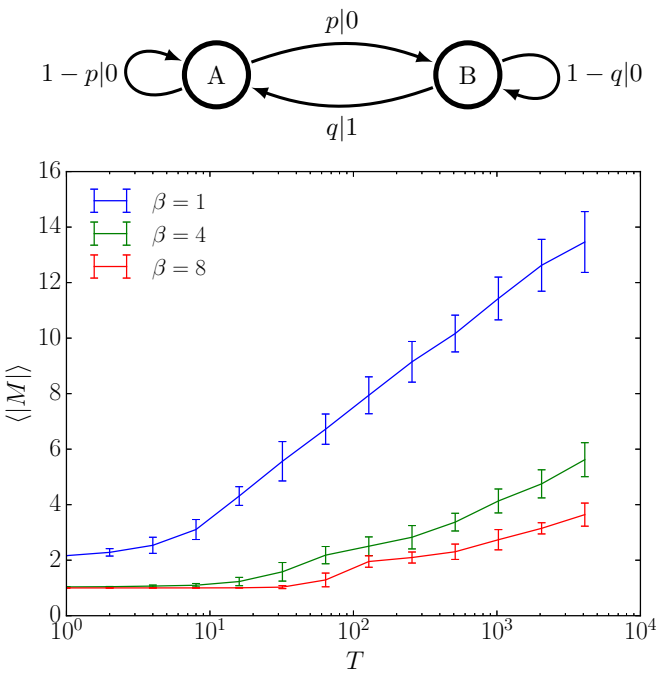

FIG. 2. Model-size scaling for the parametrized simple nonunifilar source (SNS). Top: Generative HMM for the SNS. Bottom: Scaling of $\langle|M|\rangle = \sum_M |M| P(M|x_{0:T})$ with T, with the posterior $P(M|x_{0:T})$ calculated using formulas from BSI [35] with $\alpha = 1$ and the set of model topologies being the eventually Poisson ϵ-machines described in Ref. [37]. Data generated from the SNS with $p = q = \frac{1}{2}$. Linear regression reveals a slope of ≈ 1 for $\ln \ln T$ vs $\ln\langle|M|\rangle$, confirming the expected $N_T \propto \ln T$, where the proportionality constants depend on β. Curves from top to bottom: $\beta = 1,4,8$ and blue, green, and red, respectively.

$p = q = 0.5$. As we expect, since our uncertainty in d scales as $\sim 1/\sqrt{T}$, where T is the amount of data, we expect that the number of inferred states N_T should scale as $\propto \ln T$.

These scaling relationships are confirmed by BSI applied to data generated from the SNS, where the set of ϵ-machine topologies selected from are the eventually Poisson ϵ-machine topologies characterized by Ref. [37]. Given a set of machines $\mathcal{M}$ and data $x_{0:T}$, BSI returns an easily calculable posterior $\Pr(M|x_{0:T})$ for $M \in \mathcal{M}$ with (at least) two hyperparameters: (i) the concentration parameter for our Dirichlet prior on the transition probabilities α, and (ii) our prior on the likelihood of a model M with number of states $|M|$, taken to be proportional to $e^{-\beta|M|}$ for a user-specified β. From this posterior, we calculate an average model size $\langle|M|\rangle(T) = \sum_{M\in\mathcal{M}} |M| \Pr(M|x_{0:T})$ as a function of T for multiple data strings $x_{0:T}$. The result is that we find the scaling of $\langle|M|\rangle(T)$ with T is proportional to $\ln T$ in the large T limit, where the proportionality constant and the initial value depends on hyperparameters α and β.

In theory, this scaling might be affected by our restrictive prior over ϵ-machine topologies, but the scaling is hard to analyze if a naive search over ϵ-machine topologies is used. The number of ϵ-machine topologies grows superexponentially with the number of states, practically limiting unbiased estimates of $\langle|M|\rangle$ to 6 using a single standard computer's memory. We therefore conjecture (without proof) that the scaling of $\langle|M|\rangle$ with T is qualitatively similar if the prior of ϵ-machine topologies allows for consistency.

Conclusion. We now better understand the tradeoff between memory and prediction in processes generated by a finite-state hidden Markov model with infinite statistical complexity. Importantly and perhaps still underappreciated, these processes are more "typical" than those with finite statistical complexity and Gaussian processes, which have received more attention in related literature [12,13].

We proposed a method for obtaining predictive features—coarse-graining mixed states—and used this feature set to bound the number of features and their coding cost in the limit of small predictive distortion. These bounds were compared to those obtained from the more familiar and widely used (Markov) predictive feature set—that of memorizing all pasts of a fixed length.

The general results here suggest that the mixed-state feature set can outperform the standard set—order-L Markov models—in many situations. Examples to the contrary exist, but are difficult to find, given that it is difficult to accurately calculate λ [the exponential rate of decay of $h_\mu(L)$ to h_μ] without at least an approximate mixed-state presentation [41]. As such, rigorously finding such an example in which order-L Markov features outperform a mixed-state presentation coarse graining is an interesting open problem.

Practically speaking, the bounds presented have at least three potential uses. First, they give weight to Ref. [7]'s suggestion to replace the statistical complexity with the information dimension of mixed-state space as a complexity measure when statistical complexity diverges. Second, they suggest a route to an improved, process-dependent tradeoff between complexity and precision for estimating the entropy rate of processes generated by nonunifilar HMMs [42]. Admittedly, space here precludes us from addressing how to estimate the probability distribution over ϵ boxes, and accurate estimation of such is an important open problem. Third, and perhaps most importantly, our upper bounds provide an attempt to calculate the expected scaling of inferred model size with available data. This scaling can be then used to estimate when more memory is needed to store information about the predictive model, even when an online inference algorithm is utilized, and thus how finite memory lower bounds the achievable predictive distortion.

Finally, from a statistics point of view, we characterized the asymptotics of the posterior distribution over model size. It is commonly accepted that inferring time-series models from finite data typically has two components—parameter estimation and model selection—though these two can be done simultaneously. Our focus above was on model selection, as we monitored how model size increased with the amount of available data. One can view the results as an effort to characterize the posterior distribution of ϵ-machine topologies given data or as the growth rate of the optimum model cost [43] in the asymptotic limit. Posterior distributions of estimated parameters (transition probabilities) are almost always asymptotically normal, with standard deviation decreasing as the square root of the amount of available data [44–46]. However, asymptotic normality does not typically hold for this posterior distribution, since using finite ϵ-machine topologies almost always implies out-of-class modeling. Rather, we

showed that the particular way in which the mode of this posterior distribution increases with data typically depends on a gross process statistic—the box-counting dimension of the mixed-state presentation.

Stepping back, we only tackled the lower parts of the *infinitary* process hierarchy identified in Ref. [47]. In particular, "complex" processes in the sense of Ref. [48] have different resource-prediction tradeoffs than analyzed here, since processes generated by finite-state HMMs (as assumed here) cannot produce processes with infinite excess entropy. To do so, at the very least, predictive distortion must be more carefully defined. We conjecture that success will be achieved by instead focusing on a one-step predictive distortion, equivalent to a self-information loss function, as is typically done [43,49]. Luckily, the derivation in Appendix B of the Supplemental Material [25] easily extends to this case, simultaneously suggesting improvements to related entropy-rate-approximating algorithms.

We hope these introductory results inspire future study of resource-prediction tradeoffs for processes.

Acknowledgments. The authors thank the Santa Fe Institute for its hospitality during visits and A. Boyd, C. Hillar, and D. Upper for useful discussions. This material was based upon work supported by, or in part by, the US Army Research Laboratory and the US Army Research Office under Contracts No. W911NF-13-1-0390 and No. W911NF-12-1-0288. S.E.M. was funded by a National Science Foundation Graduate Student Research Fellowship, a U.C. Berkeley Chancellor's Fellowship, and the MIT Physics of Living Systems Fellowship.

[1] S. Still, Information-theoretic approach to interactive learning, Europhys. Lett. **85**, 28005 (2009).

[2] M. L. Littman, R. S. Sutton, and S. P. Singh, Predictive representations of state, Adv. Neural. Inf. Process. Syst. **2**, 1555 (2002).

[3] N. Tishby and D. Polani, in *Perception-Action Cycle* (Springer, New York, 2011), pp. 601–636.

[4] D. Little and F. Sommer, Learning and exploration in action-perception loops, Front. Neural Circuits **7**, 37 (2013).

[5] S. Still and D. Precup, An information-theoretic approach to curiosity-driven reinforcement learning, Theory Biosci. **131**, 139 (2012).

[6] N. Brodu, Reconstruction of ϵ-machine in predictive frameworks and decisional states, Adv. Complex Syst. **14**, 761 (2011).

[7] J. P. Crutchfield, The calculi of emergence: Computation, dynamics, and induction, Physica D (Amsterdam) **75**, 11 (1994).

[8] S. Still and J. P. Crutchfield, Structure or noise? arXiv:0708.0654.

[9] T. Berger, *Rate Distortion Theory* (Prentice-Hall, New York, 1971).

[10] J. P. Crutchfield and K. Young, Inferring Statistical Complexity, Phys. Rev. Lett. **63**, 105 (1989).

[11] C. R. Shalizi and J. P. Crutchfield, Computational mechanics: Pattern and prediction, structure and simplicity, J. Stat. Phys. **104**, 817 (2001).

[12] F. Creutzig, A. Globerson, and N. Tishby, Past-future information bottleneck in dynamical systems, Phys. Rev. E **79**, 041925 (2009).

[13] S. Still, J. P. Crutchfield, and C. J. Ellison, Optimal causal inference: Estimating stored information and approximating causal architecture, Chaos **20**, 037111 (2010).

[14] S. Marzen and J. P. Crutchfield, Circumventing the curse of dimensionality in prediction: Causal rate-distortion for infinite-order markov processes, J. Stat. Phys. **163**, 1312 (2014).

[15] D. Blackwell, *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, held at Liblice near Prague, November 28–30, 1956 (1957), Vol. 28, pp. 13–20.

[16] L. R. Rabiner and B. H. Juang, An introduction to hidden markov models, IEEE ASSP Mag. **3**, 4 (1986).

[17] E. J. Gilbert, On the identifiability problem for functions of finite Markov chains, Ann. Math. Stat. **30**, 688 (1959).

[18] H. Ito, S.-I. Amari, and K. Kobayashi, Identifiability of hidden markov information sources and their minimum degrees of freedom, IEEE Trans. Inf. Theory **38**, 324 (1992).

[19] V. Balasubramanian, Statistical inference, Occam's Razor, and statistical mechanics on the space of probability distributions, Neural Comput. **9**, 349 (1997).

[20] D. R. Upper, *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*, Ph.D. thesis, University of California, Berkeley, 1997 (University Microfilms Intl., Ann Arbor, MI, 1997).

[21] This is not necessarily the same as the generative model with minimal generative complexity [53], since a nonunifilar generator with a large number of sparsely used states might have smaller state entropy than a nonunifilar generator with a small number of equally used states. And, typically, per our main result, the minimal generative model is usually not the same as the minimal prescient (unifilar) model.

[22] Cf. Ref. [23]'s discussion of terminology and Ref. [48].

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

[24] Finite HMMs generate finitary processes, as **E** is bounded from above by the logarithm of the number of HMM states [53].

[25] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.95.051301 for how to estimate fractal dimensions and the derivation of predictive distortion scaling for coarse-grained mixed states [50–53].

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, New York, 2006).

[27] N. F. Travers and J. P. Crutchfield, Equivalence of history and generator ϵ-machines, arXiv:1111.4500.

[28] One can construct a better feature set simply by applying the causal-state equivalence relation to these length-L pasts. We avoid this here since the size of this feature set is more difficult to analyze generically and since the causal-state equivalence relation applied to length L pasts often induces no coarse graining.

[29] P. M. Ara, R. G. James, and J. P. Crutchfield, The elusive present: Hidden past and future correlation and why we build models, Phys. Rev. E **93**, 022143 (2016).

[30] N. F. Travers, Exponential bounds for convergence of entropy rate approximations in hidden Markov models satisfying a path-mergeability condition, Stochastic Proc. Appl. **124**, 4149 (2014).

[31] 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**, 1005 (2009).

[32] Ya. B. Pesin, *Dimension Theory in Dynamical Systems: Contemporary Views and Applications* (University of Chicago Press, 2008).

[33] We assume here that the lower and upper box-counting dimensions are equivalent.

[34] S. Still and W. Bialek, How many clusters? An information-theoretic perspective, Neural Comput. **16**, 2483 (2004).

[35] C. C. Strelioff and J. P. Crutchfield, Bayesian structural inference for hidden processes, Phys. Rev. E **89**, 042119 (2014).

[36] B. D. Johnson, J. P. Crutchfield, C. J. Ellison, and C. S. McTague, Enumerating finitary processes, arXiv:1011.0036.

[37] S. Marzen and J. P. Crutchfield, Informational and causal architecture of discrete-time renewal processes, Entropy **17**, 4891 (2015).

[38] S. Marzen and J. P. Crutchfield, Statistical Signatures of structural organization: The case of long memory in renewal processes, Phys. Lett. A **380**, 1517 (2016).

[39] S. Marzen, M. R. DeWeese, and J. P. Crutchfield, Time resolution dependence of information measures for spiking neurons: Scaling and universality, Front. Comput. Neurosci. **9**, 109 (2015).

[40] S. E. Marzen and J. P. Crutchfield, Informational and causal architecture of continuous-time renewal and hidden semi-Markov processes, arXiv:1611.01099.

[41] J. P. Crutchfield, C. J. Ellison, and P. M. Riechers, Exact complexity: Spectral decomposition of intrinsic computation, Phys. Lett. A **380**, 998 (2016).

[42] E. Ordentlich and T. Weissman, Entropy of hidden markov processes and connections to dynamical systems, London Math. Soc. Lect. Notes **385**, 117 (2011).

[43] J. Rissanen, Universal coding, information, prediction, and estimation, IEEE Trans. Inf. Theory **30**, 629 (1984).

[44] A. M. Walker, On the asymptotic behaviour of posterior distributions, J. R. Stat. Soc. Ser. B **31**, 80 (1969).

[45] C. C. Heyde and I. M. Johnstone, On asymptotic posterior normality for stochastic processes, J. R. Stat. Soc. Ser. B **41**, 184 (1979).

[46] T. J. Sweeting, On asymptotic posterior normality in the multiparameter case, Bayesian Stat. **4**, 825 (1992).

[47] J. P. Crutchfield and S. Marzen, Signatures of infinity: Nonergodicity and resource scaling in prediction, complexity, and learning, Phys. Rev. E **91**, 050106 (2015).

[48] W. Bialek, I. Nemenman, and N. Tishby, Predictability, complexity, and learning, Neural Comput. **13**, 2409 (2001).

[49] N. Merhav and M. Feder, Universal prediction, IEEE Trans. Inf. Theory **44**, 2124 (1998).

[50] S.-W. Ho and R. W. Yeung, The interplay between entropy and variational distance, IEEE Trans. Inf. Theory **56**, 5906 (2010).

[51] T. Tao, *An Introduction to Measure Theory* (American Mathematical Society, Providence, Rhode Island, 2011), Vol. 126.

[52] This upper bound on **E** was noted in Ref. [53].

[53] W. Löhr, Models of discrete-time stochastic processes and associated complexity measures, Ph.D. thesis, University of Leipzig, 2009.