Title
Prediction and generation of binary Markov processes: Can a finite-state fox catch a Markov mouse?

Journal
Chaos (Woodbury, N.Y.), 28(1)

ISSN
1054-1500

Authors
Ruebeck, Joshua B
James, Ryan G
Mahoney, John R
et al.

Publication Date
2018

DOI
10.1063/1.5003041

Peer reviewed

Prediction and generation of binary Markov processes: Can a finite-state fox catch a Markov mouse?

Cite as: Chaos **28**, 013109 (2018); https://doi.org/10.1063/1.5003041
Submitted: 01 September 2017 . Accepted: 28 November 2017 . Published Online: 10 January 2018

Joshua B. Ruebeck, Ryan G. James, John R. Mahoney, and James P. Crutchfield

Prediction and generation of binary Markov processes: Can a finite-state fox catch a Markov mouse?

Joshua B. Ruebeck,[1,a)] Ryan G. James,[2,b)] John R. Mahoney,[2,c)] and James P. Crutchfield[2,d)]
[1]*Department of Physics and Astronomy, Carleton College, One North College Street, Northfield, Minnesota 55057, USA*
[2]*Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, California 95616, USA*

(Received 1 September 2017; accepted 28 November 2017; published online 10 January 2018)

Understanding the generative mechanism of a natural system is a vital component of the scientific method. Here, we investigate one of the fundamental steps toward this goal by presenting the minimal generator of an arbitrary binary Markov process. This is a class of processes whose predictive model is well known. Surprisingly, the generative model requires three distinct topologies for different regions of parameter space. We show that a previously proposed generator for a particular set of binary Markov processes is, in fact, not minimal. Our results shed the first quantitative light on the relative (minimal) costs of prediction and generation. We find, for instance, that the difference between prediction and generation is maximized when the process is approximately independently, identically distributed. *Published by AIP Publishing.*
https://doi.org/10.1063/1.5003041

Imagine a mouse being chased by a fox. Survival suggests that the mouse should *generate* a path that is difficult for the fox to predict. We might imagine that the mouse brain is designed or trained to maximize the fox's difficulty and, similarly, that the fox somehow has optimized the task of *predicting* the mouse's path. Are these two tasks actually distinct? If so, do there exist escape paths that are easier to generate than predict? Every animal has limited computational resources, and we might reasonably suppose that the mouse has fewer than the fox. Given that mice clearly continue to survive, we can ask whether this disparity in resources exists in tension with the disparity in task-complexity—path-generation versus path-prediction.

I. INTRODUCTION

In lieu of mouse paths, we consider the space of discrete stationary stochastic processes—objects consisting of temporal sequences that span the range from perfectly ordered to completely random. We then frame resource questions quantitatively via hidden Markov model (HMM) representations of these processes. We focus on two particular HMM representations of any given process: the minimal predictive HMM—its computational mechanics' ϵ-machine[1]—and its minimal generative HMM. We then find two primary measures of memory resource: C_μ—defined as the ϵ-machine's state-entropy—quantifies the cost of prediction, while C_g—the state entropy of the generative machine—quantifies the cost of generation. Introduced over two and a half decades ago, the ϵ-machine predictive representation is well studied and can be constructed for arbitrary processes.[2] The generative machine offers more challenges, as it involves a nonconvex constrained minimization over high-dimensional spaces. While there are several known bounds on C_g and restrictions on the construction of generative HMMs,[3–6] they have received significantly less attention than the predictive case and, as a consequence, are markedly less well understood.

The following presents the first construction of the minimal generators for an arbitrary stationary binary Markov process. This allows for the analytic calculation of C_g and other properties of generative models. These models elucidate the differences between the tasks of generation and prediction. The techniques introduced here should also lead to minimal generators for other process classes.

II. MODELS

We represent stochastic processes using edge-emitting (Mealy) *hidden Markov models* (HMMs). Such a representation is specified by a set of states, a set of output symbols, a set of labeled transition matrices, and a stationary distribution over states. We consider stationary processes so that the invariant state distribution is unique and is therefore redundantly determined from the labeled transition matrices, assuming that the state transition structure is mixing.

Clearly, not every HMM corresponds to any given process. If a model is to correspond to a particular process, its states must yield *conditional independence* between the process' past and future. That is, the past $X_{-\infty:0}$ and future $X_{0:\infty}$ random variable chains yielded by a model must be rendered independent by the model's current state $\mathcal{R}_0$. Information theoretically, the past-future mutual information, conditioned on the state vanishes: $\mathrm{I}[X_{-\infty:0} : X_{0:\infty}|\mathcal{R}_0] = 0$. The (unconditioned) mutual information $\mathbf{E} = \mathrm{I}[X_{-\infty:0} : X_{0:\infty}]$ between past and future is called the *excess entropy*. Among other uses, it is the amount of uncertainty about the future one may

[a)]jbruebeck@uwaterloo.ca
[b)]rgjames@ucdavis.edu
[c)]jrmahoney@ucdavis.edu
[d)]chaos@ucdavis.edu

reduce through knowledge of the past. Intuitively then, the state of a correct model must "capture" **E** bits of information; see Fig. 1. (For brevity, the following suppresses infinite variable indices.)

There are an infinity of such models for a given stochastic process. Depending on context, certain models will have merits above those of others. The ability to predict is one such context.

III. PREDICTIVE MODELS

What is *prediction*? Loosely speaking, prediction has to do with a relation between two variables, one which we think of as input and the other as output. In our context of stochastic processes, the input is the past $X_{:0}$ and the output is the future $X_{0:}$. By prediction, we mean that given some instance of the past $x_{:0}$, the task is to yield the exact conditional probability distribution $\Pr(X_{0:\ell}|x_{:0})$ for any length ℓ.

A. ϵ-Machine construction

The minimal predictive model of a process $\mathcal{P}$ is known as its ϵ-machine, and its construction is straightforward. The theory of computational mechanics provides a framework for the detailed characterization of ϵ-machines in topological and information-theoretic terms.[1]

The kernel underlying this construction is the *causal equivalence relation* $\sim_\epsilon$. This is a relation over the set $\{x_{:0}\}$ of semi-infinite pasts such that two pasts, $x_{:0}$ and $x'_{:0}$, belong to the same equivalence class if their conditional futures agree

$$x_{:0} \sim_\epsilon x'_{:0} \iff \Pr(X_{0:}|x_{:0}) = \Pr(X_{0:}|x'_{:0}) \,.$$

Each equivalence class is a state of the system, encapsulating in minimal form the degree to which the past influences the future. Thus, we refer to the classes as *causal states* and denote by $\mathcal{S}_t$ the causal state at time t. The memory required by the ϵ-machine to *implement* the act of prediction is $C_\mu = \mathrm{H}[\mathcal{S}]$—the *statistical complexity*. (This notion of memory applies in the ensemble setting. Single-shot or single-instance memory is also of interest and is studied in Ref. 10).

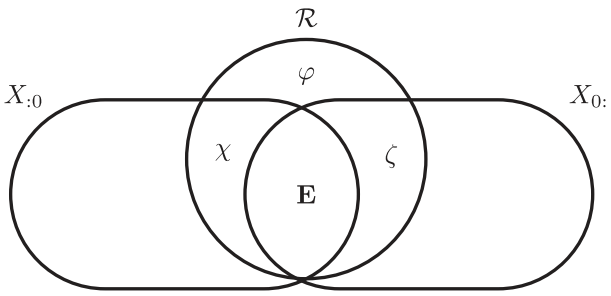

FIG. 1. Information diagram[7] of the $X_{:0} - \mathcal{R}_0 - X_{0:}$ (past-state-future) Markov chain.[8] The state $\mathcal{R}$ of a generating model shields the past $X_{:0}$ and future $X_{0:}$, rendering them conditionally independent. This is reflected by the overlap **E** between the past and future being entirely captured by (contained within) the system state entropy $\mathrm{H}[\mathcal{R}]$—the circle labeled $\mathcal{R}$. The past and future further segment $\mathrm{H}[\mathcal{R}]$ into the *crypticity* χ, *gauge information* φ, and *oracular information* ζ, quantities whose interpretation is explored further in Ref. 9.

Then, transitions $T^k_{i,j}$ between these states follow directly from the equivalence relation

$$T^k_{i,j} = \Pr(X_0 = k, \mathcal{S}_1 = j|\mathcal{S}_0 = i) \,.$$

As previously stated, the excess entropy **E** is the amount of information shared between past and future. The causal equivalence relation induces a particular random variable $\mathcal{S}$ that "captures" **E**. Importantly, **E** is not itself the entropy of a random variable.[11] Thus, the causal-state random variable cannot generally be of size **E** bits. We might then think of the difference $\chi = C_\mu - \mathbf{E}$, also known as the *crypticity*, as the *predictive overhead*.[12] It is an interesting fact that a nonzero predictive overhead χ is generic in the space of all processes.

B. Binary Markov processes

Let us now narrow our focus and construct the predictive models for the particular class of binary Markov processes. More specifically, we consider all stationary stochastic processes over the symbol set $\{0, 1\}$ with the Markov property

$$\Pr(X_0|X_{-\infty:0}) = \Pr(X_0|X_{-1}) \,.$$

Applying the causal equivalence relation, we find that the causal state is completely determined by the previous single symbol, a simple consequence of the process' Markovity. This leads directly to the ϵ-machine in Fig. 2.

Its stationary state distribution is

$$\pi = \left[\frac{1-q}{2-p-q}, \frac{1-p}{2-p-q}\right] .$$

The informational properties of this class of processes—entropy rate, excess entropy, and statistical complexity—can be stated in closed form

$$\begin{aligned} h_\mu &= \pi_A \mathrm{H}(p) + \pi_B \mathrm{H}(q), \\ \mathbf{E} &= \pi - h_\mu, \\ C_\mu &= \mathbf{E} + h_\mu \,, \end{aligned}$$

where $\mathrm{H}(p) = -(p \log p) + -((1-p)\log(1-p))$ denotes Shannon's binary entropy function.[8] The simple relation among these measures follows from the fact that any (nontrivial) binary Markov process is also equivalent to a spin chain—a restricted class of Markov chains.[12]

This class of binary Markov processes spans a variety of structured processes, summarized in Fig. 3. At the extremes of either $p = 0$ or $q = 0$, we have a period-1 (constant) process. If either $p = 1$ or $q = 1$, we have Golden Mean Processes, where 0 s or 1 s occur in isolation, respectively. If

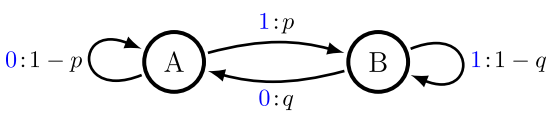

FIG. 2. ϵ-Machine for all binary Markov processes. Cases with $p = 1 - q$ or $p = 0$ or $q = 0$ are single-state ϵ-machines that are minimal in all respects: predictive or generative, entropic, or dimensional.

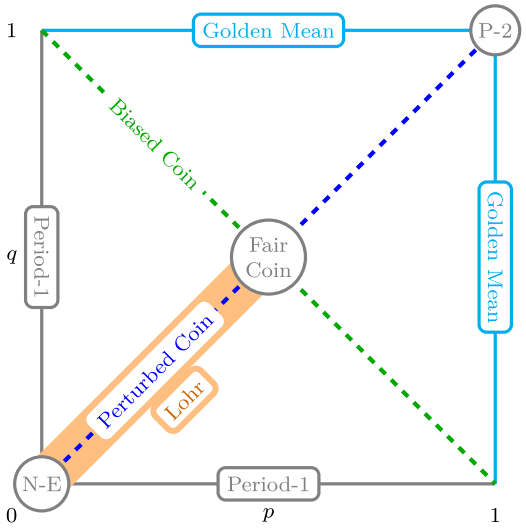

FIG. 3. Process space spanned by binary Markov processes. When either $p=0$ or $q=0$, the process is constant, repeating 0 s or 1 s, respectively. In the limit $p=q=0$, the process is nonergodic (labeled N-E above), realizing only one or the other of the two constant processes. When either $p=1$ or $q=1$, the expressed processes are known as *Golden Mean Processes*, characterized by isolated 0 s or 1 s, respectively. When $p=1$ and $q=1$, the process is period 2 (labeled P-2 above). Along the line $p=1-q$, the process is a biased coin. Along the line $p=q$, the process is known as a Perturbed Coin, where states A and B each represent an oppositely biased coin and the process switches between the two biases based on the symbol just emitted.

$p=1-q$, the process loses its dependence on the prior symbol, and it becomes a biased coin described by an ϵ-machine with a single causal state.

IV. GENERATIVE MODELS

Let us now return to our original topic and describe the second type of process representation—generative models. The only requirement of a generative model is that it be able to correctly *sample* from the distribution $\Pr(X_{0:})$ over futures. More specifically, we require that, given any instance $x_{:0}$ of the past, the generative model yields a next symbol X_0 with the same probability distribution $\Pr(X_0|X_{:0}=x_{:0})$ as specified by the process.

Note that, on the one hand, it may seem obvious that prediction subsumes generation. On the other, it is not so obvious how these two tasks might require distinct mechanisms.

Like the ϵ-machine causal state, a generative state $\mathcal{R}$ must also render past and future conditionally independent. Importantly, as a consequence of the causal equivalence relation ϵ-machines are *unifilar* which, when paired with their minimality, implies that the causal states are functions of the prior observables. Generative models, however, need not have this restriction. Consequently, a given sequence of past symbols (finite or semi-infinite) may induce more than one generative state.

Generative models are much less well understood than their predictive cousins. This is due in large part to the lack of constructive methods for working with and otherwise constructing them. This is why our results here, although addressing only a relatively simple class of processes, mark a substantial step forward.

V. LÖHR EXAMPLE

Let us now focus on a subclass of binary Markov processes—the *Perturbed Coin Processes* for which $0<p=q<1/2$; refer to the orange line in Fig. 3. Reference 4 offers up a three-state HMM generator for this class, which we refer to as the *Löhr model*; see Fig. 4. We see from the HMM that when probability p is near 1/2, the process is nearly *independent, identically distributed* (IID). An IID process has only a single causal state and therefore zero statistical complexity, $C_\mu=0$. However, for any deviation from $p=1/2$, the statistical complexity is a full bit, $C_\mu=1$. Why is it that a generator of a nearly IID process—that is, a nearly memoryless process—still needs a full bit of memory?

The motivation for constructing this three-state model is that it might concentrate the IID behavior into a single state and use the other states only for those infrequent deviations that "make up the difference". So, the state-entropy may be reduced even though there are three states instead of two. A priori it is not obvious that it is possible to yield the correct process in this construction. It is, however, straightforward to check that the Löhr model produces the correct conditional statistics. It is a generator of the process. Note that in general it is sufficient to check these probabilities for all words of length $2N-1$, where $N=\max(|\mathcal{S}|,|\mathcal{R}|)$ (Ref. 13, Corollary 4.3.9).

We find that the Löhr model has the stationary state distribution

$$\pi=[1/2-p,2p,1/2-p]\,.$$

As noted, the statistical complexity $C_\mu=1$ for p's entire range. The state entropy $\mathrm{H}[\mathcal{R}]=\mathrm{H}[\pi]$ of the Löhr model is strictly smaller than $C_\mu=1$ for the parameter range $p\in(0.38645\ldots,0.5)$. Importantly, this is sufficient to show that prediction and generation are generally different tasks—they (must) have different optimal solutions. This was previously shown in Ref. 4. However, the question remained whether or not the Löhr model is *minimal*. Surprisingly, although subsequent works on generative complexity have appeared, to the best of our knowledge, this example is the only HMM published that is entropically smaller than the (finite-state) ϵ-machine.

We will now construct the provably *minimal* generator for these processes. Furthermore, we extend our analysis not only to the range $p>1/2$ but also to the entire (p,q) domain of Fig. 3.

A. Bounds

Recall that, for some p, the Löhr model is entropically smaller than the ϵ-machine and it achieves this while having

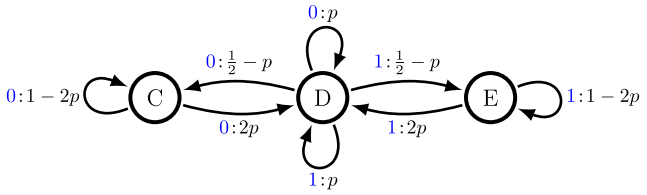

FIG. 4. Löhr model: A three-state HMM that generates the same process as that in Fig. 2 when $0\le p=q\le 1/2$. Its principle interest arises since it has a smaller state entropy than the ϵ-machine for a range of p values: $\mathrm{H}[\mathcal{R}]\le C_\mu$. Supplementary Material Sec. I gives the relationship between the Lohr states and causal states.

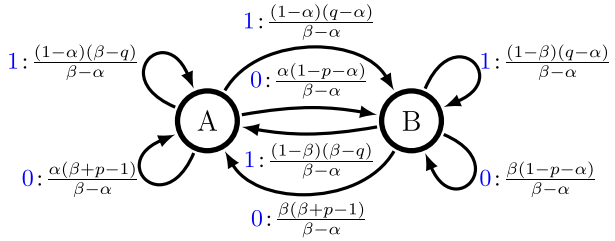

FIG. 5. Parametrized HMM for the complete set of 2-state machines that generate the space of binary Markov chain processes when $0 \leq \alpha \leq \min\{q, 1-p\}$ and $1 \geq \beta \geq \max\{q, 1-p\}$, and we assume $\alpha < \beta$. A second isomorphic class follows from the assumption $\beta < \alpha$.

three states instead of two. The important point is that minimization of entropy in the generative context does not limit the number of states in the same way as in the predictive one. (Recall that among predictive models, the ϵ-machine is minimal in both entropy and state number.[2])

A recent result shows that the maximum number of states in an entropically minimal *channel* Z is $|Z| \leq \min\left\{|X||Y|, 2^{\min\{|X|,|Y|\}} - 1\right\}$, where X and Y are the channel input and output processes and $|\cdot|$ is the size of the random variable's event set.[14] Since a generative model is a form of communication channel from the past to the future, we find that the number of states of the minimal generative model is bounded by $|\mathcal{R}| \leq \min\left\{|X_{:0}||X_{0:}|, 2^{\min\{|X_{:0}|,|X_{0:}|\}} - 1\right\}$. Of course, this result is useless on its own: $|X_{:0}|$ and $|X_{0:}|$ are generically infinite.

This bound can be made practical by combining the data processing inequality for exact common information $\mathrm{G}[X:Y]$[14] with the existence of the following two Markov chains:[15]

$$X_{:0} - \mathcal{S}^+ - \mathcal{S}^- - X_{0:}, \text{ and}$$
$$\mathcal{S}^+ - X_{:0} - X_{0:} - \mathcal{S}^-.$$

We denote forward- and reverse-time causal states $\mathcal{S}^+$ and $\mathcal{S}^-$, respectively. Combined, these tell us that $\mathrm{G}[X_{:0} : X_{0:}] = \mathrm{G}[\mathcal{S}^+ : \mathcal{S}^-]$. Therefore, the bound can be tightened to $|\mathcal{R}| \leq \min\left\{|\mathcal{S}^+||\mathcal{S}^-|, 2^{\min\{|\mathcal{S}^+|,|\mathcal{S}^-|\}} - 1\right\}$. This is a particularly helpful application of causal states.

B. Binary Markov chains

In the particular case of processes represented by binary Markov chains, the reverse process is also represented by a binary Markov chain. So, both $|\mathcal{S}^+| = 2$ and $|\mathcal{S}^-| = 2$. From the above bounds, we find that $|\mathcal{R}| \leq 3$. Closely following the proof in Ref. 14, one can then show that no three-state representation is minimal. Since a single state model can only represent IID processes, this leaves only models with

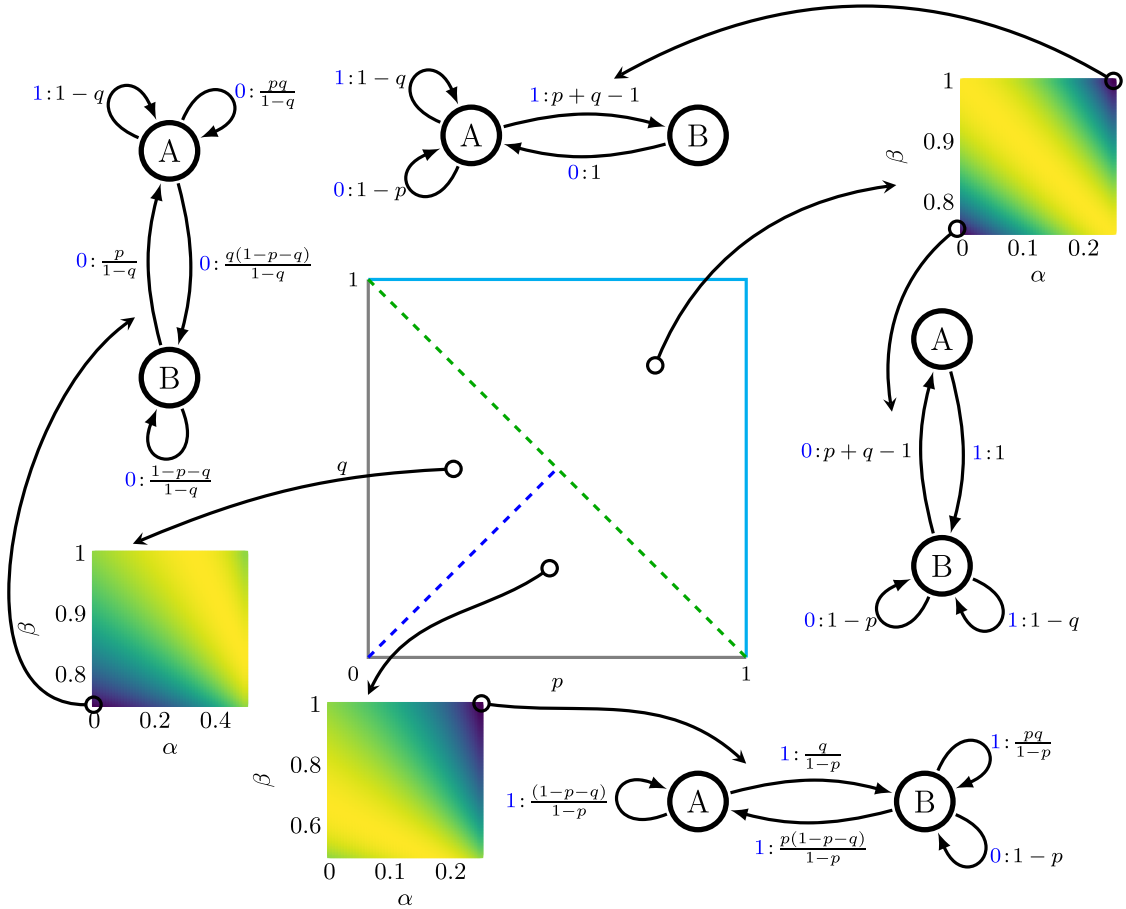

FIG. 6. Two-parameter process space of binary Markov processes and their generators: Consider three points within this space. For each, there is a two-parameter model space. Within each model space, we examine the model's state entropy and identify the global minima. We exhibit the corresponding HMMs. Topological changes in these minimal HMMs induce a three-region partition on process space.

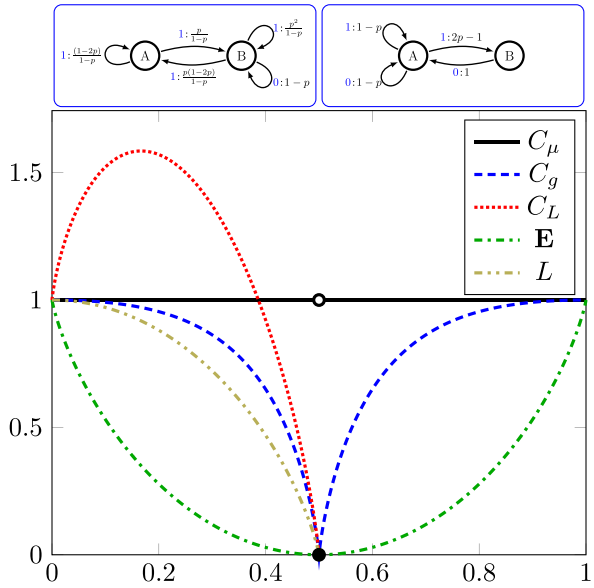

FIG. 7. State entropy of various models of the Perturbed Coin Process: The excess entropy **E** is the amount of information any model of a process must possess. A stronger lower bound L claimed by Löhr is also plotted. Entropies of the three models: C_μ for the ϵ-machine, C_L for Löhr's model, and C_g for the generative model. (The HMMs used to calculate the latter are displayed at the top.) While C_L is less than C_μ for some values of p, C_g is less than both C_μ and C_L everywhere.

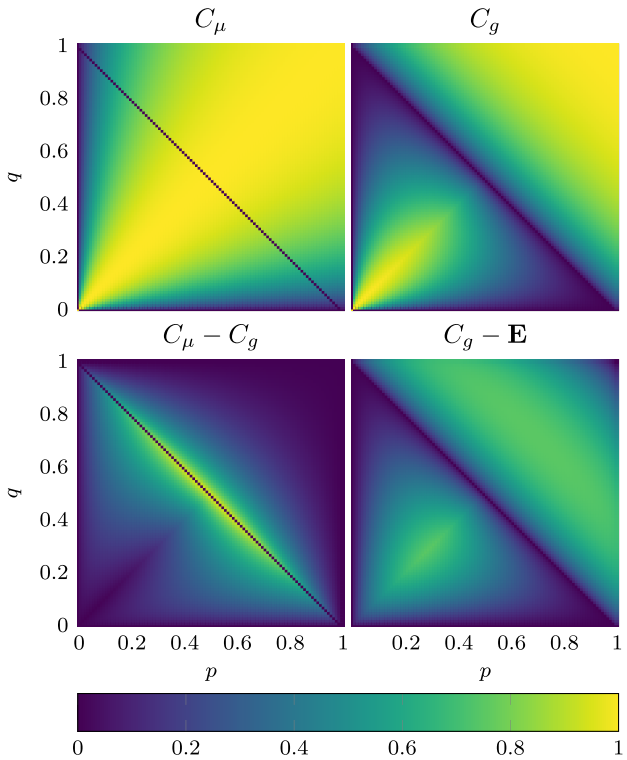

FIG. 8. State complexity of the two canonical models: ϵ-machine and generative machine. The predictive overhead, $C_\mu - C_g$, quantifies the information required to enable prediction above and beyond generation. The generative overhead, $C_g - \mathbf{E}$, quantifies the amount of information a model of a process requires beyond that minimally required by the observable correlations.

two states as the possible minimal representations. Thus, Löhr's model cannot be entropically minimal.

To find the entropically minimal representation, we begin with the assumption that an observation X_0 maps stochastically to a state $\mathcal{R}_0$, which then stochastically maps to a symbol X_1. Constraining this pair of channels to produce observations X_0 and X_1 consistent with the binary Markov chain yields the parametrized hidden Markov model found in Fig. 5. (Supplementary Material Sec. III gives the background calculations.)

For each point (p, q) in the binary Markov *process*-space (Fig. 3), we now have a two-parameter *model*-space of HMMs, specified by (α, β). The constraint that conditional probabilities be between zero and one restricts our model-space parameters to a rectangle $0 \leq \alpha \leq \min\{q, 1-p\}$ and $1 \geq \beta \geq \max\{q, 1-p\}$. One can now compute the state entropy within this constrained model-space and identify the minima.

Since the entropy is concave in α and β and the allowable regions in (α, β)-space are convex (rectangles), it is sufficient to search for local minima along the boundary.

Figure 6 illustrates this for three different points in process space. We find that at each of the points $(p = 1/4, q = 1/2)$ and $(p = 1/2, q = 1/4)$, there is a single global minimum. For the point $(p = 3/4, q = 3/4)$, we find that there are two minima equivalent in value but corresponding to nonisomorphic HMMs. Both representations are biased toward producing a periodic sequence with fluctuations interjected at different phases of the period.

In this way, one can discover the minimal generator for any binary Markov chain. Examining these minimal topologies at each point, we find that process-space is divided into three triangular regions with topologically distinct generators. This is in somewhat surprising contrast with the fact that this model class requires only one predictive topology (Fig. 2).

Let us briefly return to the restricted process previously considered—the Perturbed Coin (Fig. 3). We may now quantitatively compare the three state-entropies of interest. In Fig. 7, we see that the statistical complexity $C_\mu = 1$ everywhere, except at $p = 1/2$, where it vanishes, $C_\mu = 0$. The Löhr model state-entropy C_L falls below C_μ but only for a subset of p values. However, the generative complexity C_g (a continuous function) is everywhere less than both C_μ and C_L. (The generative models for $p < 1/2$ and $p > 1/2$ from which C_g is calculated are shown at the top of Fig. 7.) This demonstrates that the proposed Löhr model is not the generative model for any value of p.

As implied by the conditional independence requirement, the excess entropy **E** remains a lower bound on each of these state-entropies. Löhr[4] constructed a tighter lower bound (denoted L in Fig. 7) on any model of the Perturbed Coin. We see that C_g is slightly larger than this bound. It may be useful to generalize this lower bound for other processes.

The minimal generators are defined over all of (p, q)-space. We can compare the cost C_μ of prediction with the cost C_g of generation and the information necessarily captured by a model—the excess entropy **E**. This comparison is seen in Fig. 8.

Focusing on the upper two panels of Fig. 8, we see that both C_μ and C_g display $p \leftrightarrow q$ symmetry. Furthermore, C_g has a discontinuous derivative along this line of symmetry but only in the southwest (SW).

For C_μ, the line $p + q = 1$ is special in that it marks a causal-state collapse—two causal states merge into one

under the equivalence relation. For C_g, however, this line marks a qualitative change in behavior (SW versus NE). Since the generative complexity is lower semi-continuous,[3] we know that a predictive gap $C_\mu - C_g$ must exist around this line.

The lower two panels of Fig. 8 suggest that the costs of generation and of prediction may have different causes. The parameters for which $C_g - \mathbf{E}$ is high are disjoint from those where $C_\mu - C_g$ is high. C_g is high when p and q are correlated (near the p–q symmetry line) but only for $p, q < 1/2$. In the other half of parameter space, C_g is high when p and q are anti-correlated and away from the causal collapse. In contrast, C_μ is high exclusively near the line of causal collapse. (A fuller information diagram analysis is given in supplementary material Sec. II.)

VI. CONCLUSION

We presented the minimal generators of binary Markov stochastic processes. Curiously, the literature appears to contain no other examples of generative models for processes with finite-state ϵ-machines. So, our contribution here is a substantial step forward. It allows us to begin to understand the difference between prediction and generation through direct calculation. It also opens these new models to analysis by a host of previously developed techniques including the information diagrams presented here.

To put the results in a larger setting, we note that HMMs have found application in many diverse settings, ranging from speech recognition to bioinformatics. So, there are many reasons to care about the states and information-theoretic properties of these models, some obvious and some not. It is common to imbue a state with greater explanatory power than, say, a random variable that merely exhibits the correct correlations for the observables at hand. For instance, we may seek *independent* means of determining the state. Whether or not this is appropriate, the fact remains that the different tasks of prediction and generation are associated with different kinds of state, each with different kinds of explanatory usefulness. This distinction seems to us to be rarely if ever made in HMM applications.

The concept of model state is central, for example, in model selection. A simple and common method for selecting one model over another is through application of a penalty related to the number of states (or entropy thereof).[16] Since the predictive model will never have a lower entropy than the corresponding generative one, an entropic penalty should never yield the predictive model; however, a state-number penalty might. Similarly, in model parameter inference, if one distinguishes between the predictive and generative classes, the maximum likelihood estimated parameters will differ between the two classes.

Finally, we close by drawing out the consequences for fundamental physics. Understanding states bears directly on thermodynamics. Landauer's Principle states that erasing memory comes at a minimum, unavoidable cost—a heat dissipation proportional to the size of the memory erased.[17] One can consider HMMs as abstract representations of processes with memory (the states) that must be modified or erased as time progresses. Applying Landauer's Principle assigns thermodynamic consequences to the HMM time evolution. Which HMM (and corresponding set of states) is appropriate, though? We now see that prediction and generation, two very natural tasks for a thermodynamic system to perform, actually deliver two different answers. It is important to understand how physical circumstances relate to this choice of task—it will be expressed in terms of heat.

SUPPLEMENTARY MATERIAL

See supplementary material for details of the relationship between the generative model and the Löhr model, the informational structure of generative states, and their derivation.

ACKNOWLEDGMENTS

We thank the Santa Fe Institute for its hospitality during visits, where J.P.C. is an External Faculty member. This material was based upon work supported by, or in part by, John Templeton Foundation Grant No. 52095, Foundational Questions Institute Grant No. FQXi-RFP-1609, the U.S. Army Research Laboratory and the U.S. Army Research Office under Contract Nos. W911NF-13-1-0390 and W911NF-13-1-0340, and the UC Davis Intel Parallel Computing Center. J.R. was funded by the 2016 NSF Research Experience for Undergraduates program.

[1]J. P. Crutchfield, "Between order and chaos," Nat. Phys. **8**(1), 17–24 (2012).
[2]C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," J. Stat. Phys. **104**, 817–879 (2001).
[3]W. Löhr, "Predictive models and generative complexity," J. Syst. Sci. Complexity **25**(1), 30–45 (2012).
[4]W. Löhr and N. Ay, "Non-sufficient memories that are sufficient for prediction," in *International Conference on Complex Sciences* (Springer, 2009), pp. 265–276.
[5]W. Löhr and N. Ay, "On the generative nature of prediction," Adv. Complex Sys. **12**(02), 169–194 (2009).
[6]A. Heller, "On stochastic processes derived from Markov chains," Ann. Math. Stat. **36**, 1286–1291 (1965).
[7]R. W. Yeung, "A new outlook on Shannon's information measures," IEEE Trans. Inf. Theory **37**(3), 466–474 (1991).
[8]T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, New York, 2006).
[9]J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, and R. G. James, "Synchronization and control in intrinsic and designed computation: An information-theoretic analysis of competing models of stochastic computation," Chaos **20**(3), 037105 (2010).
[10]C. Aghamohammadi and J. P. Crutchfield, "Minimum memory for generating rare events," Phys. Rev. E **95**, 032101 (2017).
[11]P. Gács and J. Körner, "Common information is much less than mutual information," Probl. Control Inf. Theory **2**, 149–162 (1973).
[12]J. R. Mahoney, C. J. Ellison, R. G. James, and J. P. Crutchfield, "How hidden are hidden processes? A primer on crypticity and entropy convergence," Chaos **21**(3), 037112 (2011).
[13]D. R. Upper, "Theory and algorithms for hidden Markov models and generalized hidden Markov models," Ph.D. thesis (University of California, Berkeley, 1997).
[14]G. R. Kumar, C. T. Li, and A. El Gamal, "Exact common information," in *2014 IEEE International Symposium on Information Theory (ISIT)* (IEEE, 2014), pp. 161–165.
[15]R. G. James, J. R. Mahoney, and J. P. Crutchfield, "Information trimming: Sufficient statistics, mutual information, and predictability from effective channel states," Phys. Rev. E **95**(6), 060102 (2017).
[16]H. Akaike, "An objective use of Bayesian models," Ann. Inst. Stat. Math. **29A**(9), 9–20 (1977).
[17]R. Landauer, "Irreversibility and heat generation in the computing process," IBM J. Res. Develop. **5**(3), 183–191 (1961).