Prediction and Generation of Binary Markov Processes:
Can a Finite-State Fox Catch a Markov Mouse?

Joshua Ruebeck[*]
Department of Physics and Astronomy, Carleton College,
One North College Street, Northfield, MN 55057

Ryan G. James,[†] John R. Mahoney,[‡] and James P. Crutchfield[§]
Complexity Sciences Center and Physics Department,
University of California at Davis, One Shields Avenue, Davis, CA 95616
(Dated: August 1, 2017)

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.

I. INTRODUCTION

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.

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 computa-

tional 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, as well as 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

---

[*] ruebeckj@carleton.edu
[†] rgjames@ucdavis.edu
[‡] jrmahoney@ucdavis.edu
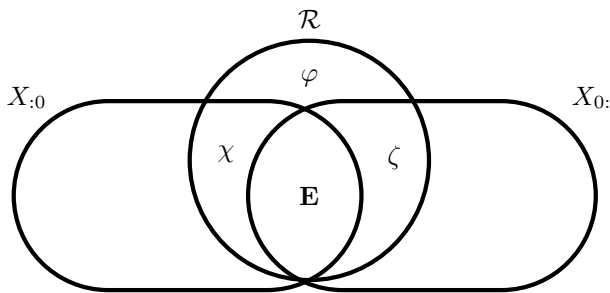[§] chaos@ucdavis.edu

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

symbols, a set of labeled transition matrices, and a stationary distribution over states. We consider stationary processes so, assuming the state transition structure is mixing, the invariant state distribution is unique and is therefore redundantly determined from the labeled transition matrices.

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 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, we have: $\mathrm{I}[X_{-\infty:0} : X_{0:\infty}|\mathcal{R}_0] = 0$. 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 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 indexes.)

There are an infinity of 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: Given

some instance of the past $x_{:0}$, the task is to yield the exact conditional probability distribution $\mathrm{Pr}(X_{0:\ell}|x_{:0})$ 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:

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

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.* [1]

Then, transitions between these states follow directly from the equivalence relation:

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

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. 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* [10]. 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

---

[1] This notion of memory applies in the ensemble setting. Single-shot or single-instance memory is also of interest and is studied in **??**.
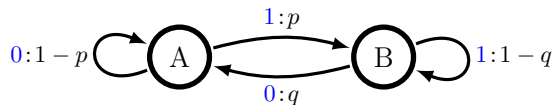
FIG. 2. ϵ-Machine for all binary Markov processes. Cases with $p = 1 - q$ or $p = 0$ or $q = 0$ are single-state ϵ-machines that are minimal in all respects: predictive or generative, entropic or dimensional.

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:

$$h_\mu = \pi_A \, \mathrm{H}\,(p) + \pi_B \, \mathrm{H}\,(q) \, ,$$
$$\mathbf{E} = \mathrm{H}\,[\pi] - h_\mu,$$
$$C_\mu = \mathbf{E} + h_\mu \ ,$$

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 [10].

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 0s or 1s, respectively, occur in isolation. If $p = 1 - q$, the process loses its dependence on the prior symbol and it becomes a biased coin.

IV. GENERATIVE MODELS

Let's 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 $\Pr(X_{0:})$
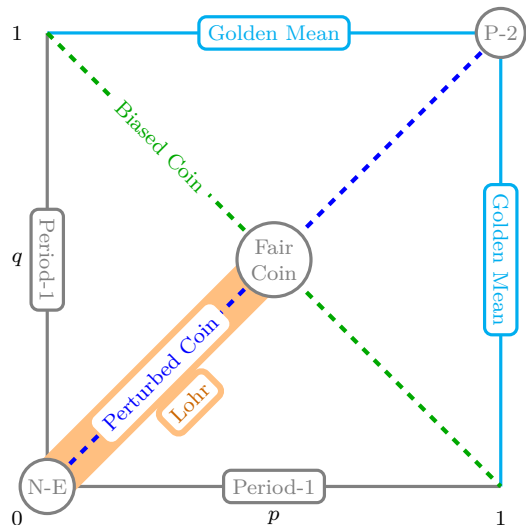

FIG. 3. Process space spanned by binary Markov processes. When either $p = 0$ or $q = 0$, the process is constant, repeating 0s or 1s, respectively. In the limit $p = q = 0$ the process is nonergodic, realizing only one or the other of the two constant processes. When either $p = 1$ or $q = 1$, the expressed processes are known as *Golden Mean Processes*, characterized by isolated 0s or 1s, respectively. 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 symbols just emitted.

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 prefer different 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, though addressing only on a relatively simple class of processes, mark a substantial step forward.
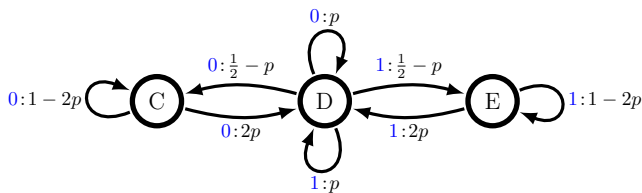
FIG. 4. Löhr model: A three-state HMM that generates the same process as that in Fig. 2 when $0 \leq p = q \leq 1/2$. Its principle interest arises since it has a smaller state entropy than the ϵ-machine for a range of p values: $H[\mathcal{R}] \leq C_\mu$.
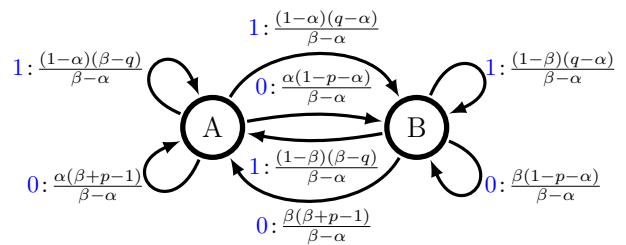

FIG. 5. Parametrized HMM for the complete set of 2-state machines that generate the space of binary Markov chain processes when $\alpha \leq \min\{q, 1-p\}$ and $\beta \geq \max\{q, 1-p\}$ and we assume $\alpha < \beta$. A second isomorphic class follows from the assumption $\beta < \alpha$.

V. LÖHR EXAMPLE

Let us now focus on a subclass of binary Markov processes, those 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". And 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 processes. Note that in general it is sufficient to check these probabilities for all words of length $2N - 1$ where $N = \max(|\mathcal{S}|, |\mathcal{R}|)$ [11, 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 $H[\mathcal{R}] = H[\pi]$ of the Löhr model is smaller than $C_\mu = 1$ for the range of values $p \in (0.38645\ldots, 0.5)$. Importantly, this is sufficient to show that prediction and generation are generally different tasks—they have different optimal solutions. This was previously shown in Ref. [4]. However, the question

remained whether or not the Löhr model is *minimal*. Surprisingly, though 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. Further, we extend our analysis, not only to the range $p > 1/2$, but 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 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 $|Z| \leq \min\{|X||Y|, 2^{\min\{|X|, |Y|\}} - 1\}$, where X and Y are the channel input and output processes [12]. 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 $|\mathcal{R}| \leq \min\{|X_{:0}||X_{0:}|, 2^{\min\{|X_{:0}|, |X_{0:}|\}} - 1\}$. 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 $G[X : Y]$ [12] with the existence of the following two Markov chains [13]:

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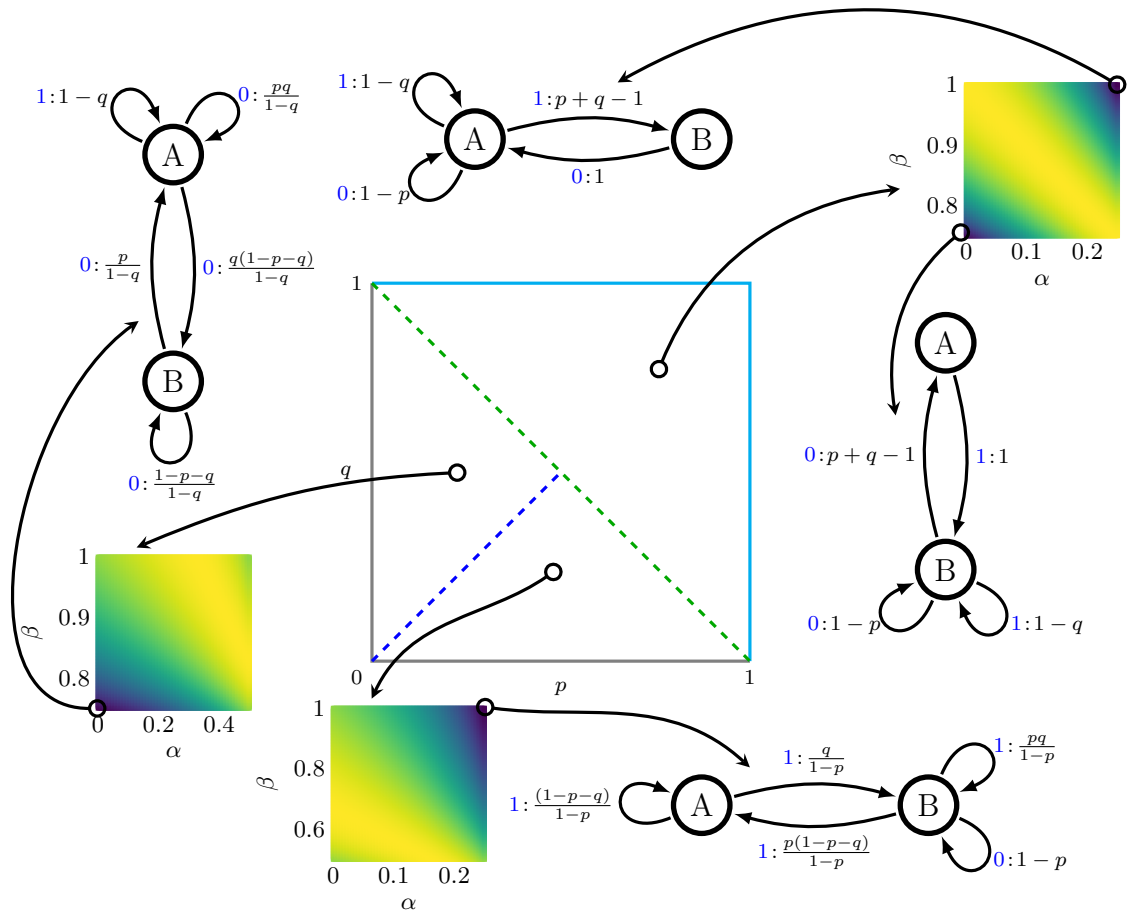$$\mathcal{S}^+ - X_{:0} - X_{0:} - \mathcal{S}^- \ .$$

FIG. 6. Two-parameter process space of binary Markov processes: 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.

We denote forward- and reverse-time causal states $\mathcal{S}^+$ and $\mathcal{S}^-$, respectively. Combined, these tell us that $G[X_{:0} : X_{0:}] = G[\mathcal{S}^+ : \mathcal{S}^-]$. 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. And 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. [12], one can then show that no three-state representation is minimal. And, since a single state model can only represent IID processes, this leaves only models with two-states as the possible minimal representations.

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. (Appendix III gives the background calculations.)

For each point (p, q) in the binary Markov *process*-space, 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 $\alpha \leq \min\{q, 1-p\}$ and $\beta \geq \max\{q, 1-p\}$. 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 a somewhat surprising contrast with the fact that this model class requires only one predictive topology.

Let us briefly return to the restricted process previously considered—the perturbed coin. We may now quantitatively compare the three state-entropies of interest. In Fig. 7 we see that the statistical complexity $C_\mu = 1$ everywhere, but at $p = {}^1/_2$, where it vanishes, $C_\mu = 0$. The Löhr model's state-entropy C_L falls below C_μ, but only for a subset of p values. However, the generative complexity C_g (a smooth function) is everywhere less than both C_μ and C_L. (The generative models for $p < {}^1/_2$ and $p > {}^1/_2$ are shown above.) This shows 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 $\mathbf{E}$ remains a lower bound on each of these state-entropies. Löhr [4] constructed a tighter lower bound, denoted L, 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 $\mathbf{E}$. 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 this lines 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
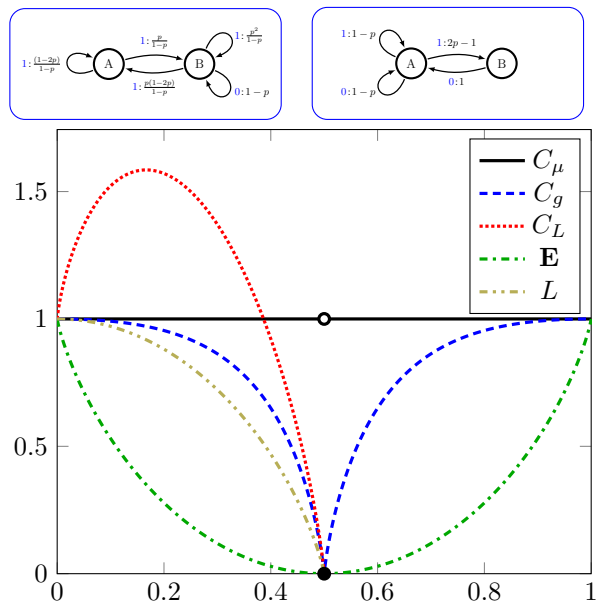

FIG. 7. State entropy of various models of the perturbed coin: The excess entropy $\mathbf{E}$ is the amount of information any model of a process must possess. A stronger lower bound L claimed by Löhr is also plotted. Entropies of the three models: C_μ for the ϵ-machine, C_L for Löhr's model, and C_g for the generative model. 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.

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.

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. And 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. And so, there are many reasons to care about the states and information-theoretic properties of these models, some
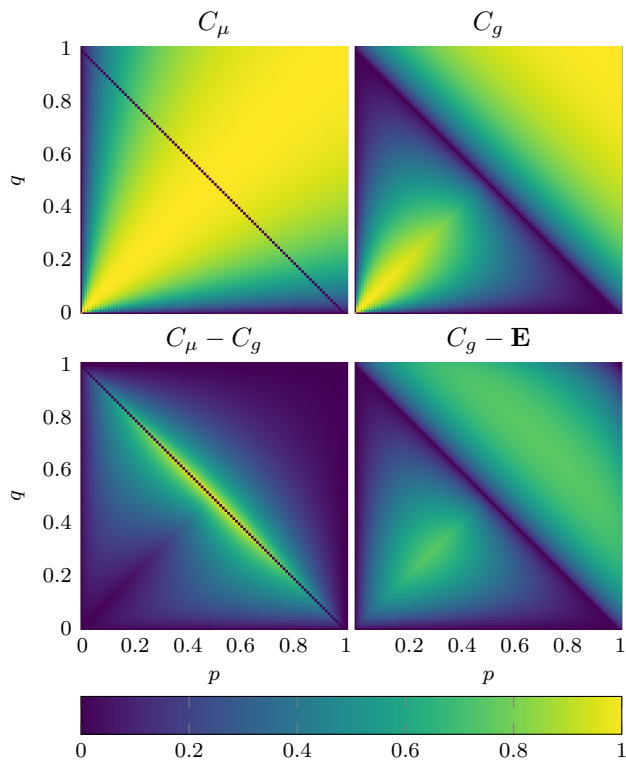
FIG. 8. State complexity of the two canonical models: ϵ-machine and generative machine. The predictive overhead, $C_\mu - C_g$, quantifies the information required to enable prediction above and beyond generation. The generative overhead, $C_g - \mathbf{E}$, quantifies the amount of information a model of a process requires beyond that minimally required by the observable correlations.

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) [14]. 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 [15]. One can consider HMMs as abstract representations of processes with memory (the state) 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 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.

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.

[1] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(1):17–24, 2012. 1, 2

[2] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001. 1, 4

[3] W. Löhr. Predictive models and generative complexity. *J. Systems Sci. Complexity*, 25(1):30–45, 2012. 1, 6

[4] W. Löhr and N. Ay. Non-sufficient memories that are sufficient for prediction. In *International Conference on Complex Sciences*, pages 265–276. Springer, 2009. 4, 6

[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. Mat. Stat.*, 36:1286–1291, 1965. 1

[7] R. W. Yeung. A new outlook on Shannon's information measures. *IEEE Trans. Info. Th.*, 37(3):466–474, 1991. 2, 1

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006. 2, 3

[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. 2, 1

[10] 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. 2, 3

[11] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan. 4

[12] G. R. Kumar, C. T. Li, and A. El Gamal. Exact common information. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 161–165. IEEE, 2014. 4, 5

[13] Ryan G James, John R Mahoney, and James P Crutchfield. Information trimming: Sufficient statistics, mutual information, and predictability from effective channel states. *Physical Review E*, 95(6):060102, 2017. 4, 1

[14] H. Akaike. *Ann. Inst. Statist. Math.*, 29A:9, 1977. 7

[15] R. Landauer. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.*, 5(3):183–191, 1961. 7

[16] S. Kamath and V. Anantharam. A new dual to the Gács-Körner common information defined via the Gray-Wyner system. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1340–1346. IEEE, 2010. 2

Supplementary Materials
Prediction and Generation of Binary Markov Processes:
Can a Finite-State Fox Catch a Markov Mouse?

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

These supplementary materials give additional information on the processes and models analyzed in the main body. They cover the relationship between causal states and Löhr states, process information diagram analysis, including crypticity and oracular information, and the derivation of the general, parametrized binary Markov process model itself.

I. CAUSAL STATES AND LÖHR STATES

Since the process considered in the main article is Markov order $R = 1$ and its causal states are in one-to-one correspondence with the symbol last seen, we can compactly represent the relation between pasts, causal states, and Löhr states. The mapping from causal states A and B (or observation symbols 0 and 1) to states C, D, and E of the Löhr presentation is given by:

$$\Pr(\mathcal{R}|\mathcal{S}) = \begin{array}{c} \\ A/0 \\ B/1 \end{array} \begin{array}{c} C \quad\quad D \quad\quad E \\ \left(\begin{array}{ccc} 1-2p & 2p & 0 \\ 0 & 2p & 1-2p \end{array} \right) \end{array} \quad \text{(S1)}$$

For example, if the last symbol was $x = 0$, this induces causal state A. The corresponding Löhr states occur with respective probabilities $\Pr(C, D, E) = (1 - 2p, 2p, 0)$.

II. INFORMATION DIAGRAM ANALYSIS

The information diagram [7] is a tool that has become increasingly useful for analyzing and characterizing the information content of process presentations [13]. It gives a visual representation of multivariate information-theoretic dependencies. It arises from a duality between information measures and set theory: set-intersection corresponds to mutual information and set-union corresponds to joining distributions. Here, we analyze several information measures introduced in Ref. [9] for the processes considered in the main article; the associated I-diagram is depicted in Fig. 9.

A generic model state $\mathcal{R}$ has nonzero intersection (mutual information) with both the process' past $X_{:0}$ and
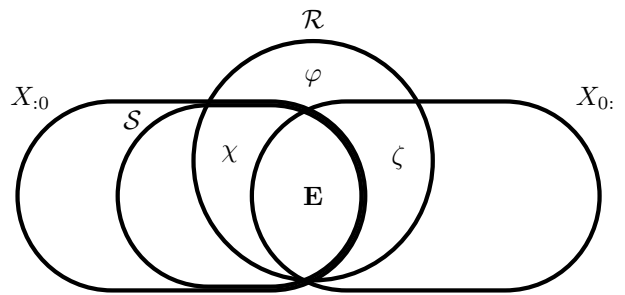

FIG. 9. I-diagram depicting how different kinds of information are shared among a process' past $X_{:0}$, future $X_{0:}$, (forward) causal states $\mathcal{S}$, and generative states $\mathcal{R}$. Several information atoms are labeled: the process' past-future mutual information or excess entropy $\mathbf{E}$, the crypticity χ, the oracular information ζ, and the gauge information φ.

future $X_{0:}$: $\mathrm{I}\,[\mathcal{R} : X_{:0}] > 0$ and $\mathrm{I}\,[\mathcal{R} : X_{0:}] > 0$. Importantly, the state information of a model that correctly generates a process—a process presentation—must intersect the past-future union, which completely contains the past-future intersection. We might say that the state of the model "captures" this atom—the excess entropy $\mathbf{E} = \mathrm{I}\,[X_{:0} : X_{0:}]$. Any model that does not capture at least $\mathbf{E}$ of the past and future cannot correctly generate the process. (Recall that if a model is predictive, then it is also generative.)

The ϵ-machine is simple not only in the sense that it is constructible, but it is informationally simple as it appears in the information diagram as well. This simplicity follows from the fact that causal states are functions of the past. As a result, ϵ-machine state information has no intersection with the future beyond that of that past's—it contains no, what we call, *oracular* information $\zeta = \mathrm{I}\,[\mathcal{R} : X_{0:}|X_{:0}]$. It also contains no information outside the past-future union—it contains not *gauge* information $\varphi = \mathrm{H}\,[\mathcal{R}|X_{:0}, X_{0:}]$. Beyond the excess entropy atom, it has only one potential region—what we call the process' *crypticity* $\chi_{\epsilon\text{-machine}} = C_\mu - \mathbf{E}$, where $C_\mu = \mathrm{H}\,[\mathcal{S}]$ is the process' *statistical complexity*.

The situation is richer for general representations, including generative ones; denote them g. The more general definition of crypticity is $\chi_g = \mathrm{I}\,[\mathcal{R} : X_{:0}|X_{0:}]$. A general representation may have oracular and gauge information. And, it may have a crypticity greater than, equal to, or less than the ϵ-machine crypticity.

Generative models are restricted in two ways. First, their crypticity is never greater than the ϵ-machine's crypticity: $\chi_g \leq \chi_{eM}$. This follows straightforwardly since if it were greater, then the ϵ-machine would necessarily have smaller state entropy, leading to a contradiction. Further, the sum of the generative model's crypticity, gauge information, and oracular information must be less than
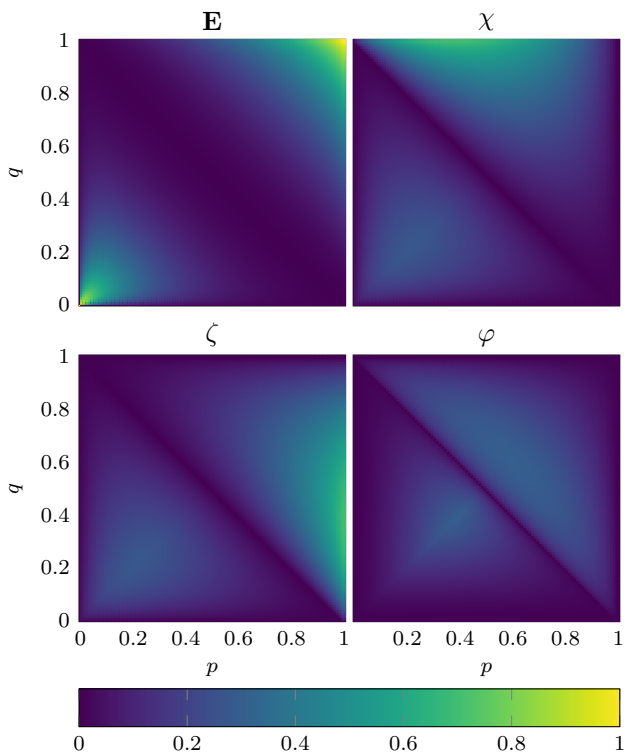
FIG. 10. Decomposition of the generative state information C_g into excess entropy $\mathbf{E}$, crypticity χ, oracular information ζ, and gauge information φ over process space (p, q).

the ϵ-machine's crypticity:

$$\chi_g + \varphi_g + \zeta_g \leq \chi_{eM} \qquad (S2)$$

Second, appealing to our newly introduced generative models for binary Markov chains, we can compare the generative-model crypticity to that of the ϵ-machine. A generative model distinct from the ϵ-machine must have nonzero oracular information [16]. Also, a generative model may only have nonzero gauge information if it has both crypticity and oracular information [16]:

$$\varphi_g > 0 \implies \begin{cases} \chi_g > 0 \\ \zeta_g > 0 \end{cases}, \qquad (S3)$$

while still satisfying Eq. (S2). Effectively, this means that gauge information can be "minimized" away, unless it supports the existence of both cryptic and oracular information.

For our parametrized generative models of binary Markov chains we find that C_g decomposes as shown in the two mosaics in Figs. 10 and 11. The gauge information φ_g is, generally, a rather small portion of C_g, though it is largest in the same regions as $C_g - \mathbf{E}$. Both the crypticity χ_g and the oracular information ζ_g are large along the
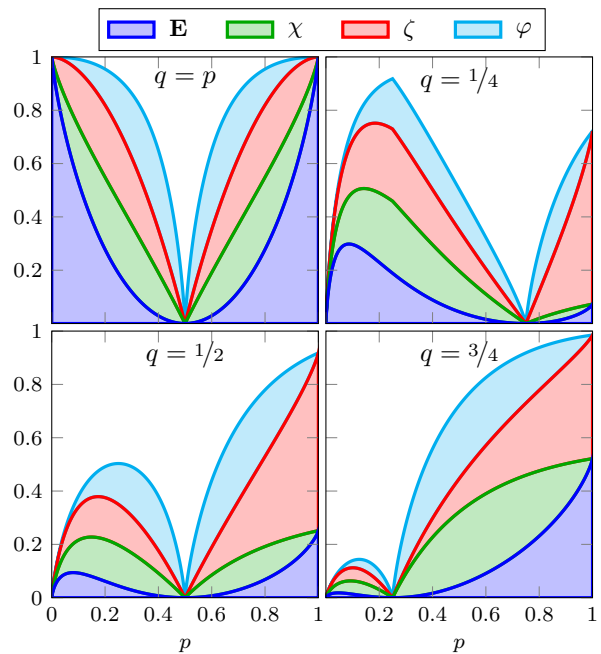

FIG. 11. Decomposition of C_g for several slices of process space, revealing more directly their functional dependencies.

$q = 1$ and $p = 1$ edges, respectively. This implies that $\mathcal{R}$ "leans left" when $q \approx 1$ and "leans right" when $p \approx 1$. This relationship flips if we use the alternative, equivalent model in the $p > 1 - q$ half of the process space.

Note that while gauge information is not required of generative models, it is present in all of the binary Markov chain generators, except along boundary and causal collapse lines. Surveys of other processes suggest that the presence of gauge information may be somewhat rare.

III. PARAMETRIZED BINARY MARKOV PROCESS HMM: DERIVATION

We derive the most generic parametrized, two-state HMM of binary Markov processes. The target binary Markov process is represented by the conditional probability matrix:

$$\Pr(X_t | X_{t-1}) = \begin{matrix} 0 \\ 1 \end{matrix} \begin{pmatrix} 1 - p & p \\ q & 1 - q \end{pmatrix} \qquad (S4)$$

As an intermediate step to deriving the full HMM, we find the conditional probabilities involving the machine states relative to the symbols emitted on the transitions to and from a state. We write down two template matrices, denoting the states A and B and the random variable
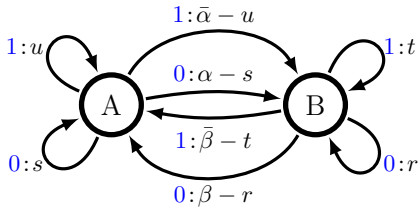
FIG. 12. As an intermediate step, we construct a skeleton for the fully general HMM. r, s, t, u are temporary variables to be solved for eventually. The other transitions are labeled using the fact that, for example, the probabilities of the paths coming from state A and emitting a 0 must sum to α in order to agree with Eq. S5.

over them $\mathcal{R}$:

$$\Pr(X_t|\mathcal{R}) = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} 0 \qquad 1 \\ \begin{pmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{pmatrix} \end{array} \qquad (S5)$$

and:

$$\Pr(\mathcal{R}|X_{t-1}) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} A \qquad B \\ \begin{pmatrix} \gamma & 1-\gamma \\ \delta & 1-\delta \end{pmatrix} \end{array} . \qquad (S6)$$

there is ambiguity here in labeling the states. If we switch the rows of Eq. (S5) with each other and the columns of Eq. (S6) with each other, then we obtain different matrices that describe the same model. Only the the state labels have been swapped. To avoid double-counting such machines, we restrict $\alpha \leq \beta$.

Multiplying out this factorization and requiring it to be consistent with the target provides two constraints:

$$\gamma = \frac{\beta + p - 1}{\beta - \alpha} \quad , \quad \delta = \frac{\beta - q}{\beta - \alpha}. \qquad (S7)$$

(We keep the denominators as $\beta - \alpha$ so that both the numerator and denominator are positive.) We ask that all these be valid probabilities. Requiring $0 \leq \gamma, \delta \leq 1$ and substituting Eq. (S7) yields the constraints:

$$\alpha \leq \min\{q, 1-p\},$$
$$\beta \geq \max\{q, 1-p\} .$$

Now, consider the incomplete HMM that defines the helper variables r, s, t, u in Fig. 12. To determine the latter's values, we evaluate the probabilities of a string

(a) being generated by this incomplete HMM:

$$\begin{aligned} \Pr(01) &= \Pr(01|A)\Pr(A) + \Pr(01|B)\Pr(B) \\ &= [s(1-\alpha) + (\alpha - s)(1-\beta)]\Pr(A) \\ &\quad + [(\beta - r)(1-\alpha) + r(1-\beta)]\Pr(B) \quad (S8) \end{aligned}$$

and (b) being generated by the Markov process described in Eq. (S4):

$$\begin{aligned} \Pr(01) &= \Pr(1|0)\Pr(0) \\ &= p\Pr(0) . \qquad (S9) \end{aligned}$$

We used the string 01 as our first case. And, $\Pr(0)$, $\Pr(1)$, $\Pr(A)$, and $\Pr(B)$ are given by the stationary distributions over the symbols and states, respectively:

$$\begin{aligned} \Pr(0) &= \frac{q}{p+q} \\ \Pr(1) &= \frac{p}{p+q} \\ \Pr(A) &= \frac{q\gamma + p\delta}{p+q} \\ \Pr(B) &= \frac{q(1-\gamma) + p(1-\delta)}{p+q} . \end{aligned}$$

Setting Eqs. (S8) and Eq. (S9) equal to one another constrains r, s, t, u. To fully specify all four, though, we need three more independent equations. We obtain these by evaluating the probabilities of the strings 11, 110, and 000 in a similar fashion. Respectively, these yield:

$$\begin{aligned} (1-q)\Pr(1) &= [u(1-\alpha) + (1-\alpha-u)(1-\beta)]\Pr(A) \\ &\quad + [t(1-\beta) + (1-\beta-t)(1-\alpha)]\Pr(B) , \\ q(1-q)\Pr(1) &= [u((1-\alpha-u)\beta + u\alpha) \\ &\quad + (1-\alpha-u)(t\beta + (1-\beta-t)\alpha)]\Pr(A) \\ &\quad + [t(t\beta + (1-\beta-t)\alpha) \\ &\quad + (1-\beta-t)((1-\alpha-u)\beta + u\alpha)]\Pr(B), \text{ and} \\ (1-p)^2\Pr(0) &= [s((\alpha - s)\beta + s\alpha) \\ &\quad + (\alpha - s)(r\beta + (\beta - r)\alpha)]\Pr(A) \\ &\quad + [r((\beta - r)\alpha + r\beta) \\ &\quad + (\beta - r)((\alpha - s)\beta + s\alpha)]\Pr(B) . \end{aligned}$$

Solving this system yields:

$$r = \frac{\beta(1 - p - \alpha)}{\beta - \alpha} \ ,$$

$$s = \frac{\alpha(\beta + p - 1)}{\beta - \alpha} \ ,$$

$$t = \frac{(1 - \beta)(q - \alpha)}{\beta - \alpha} \ , \text{ and}$$

$$u = \frac{(1 - \alpha)(\beta - q)}{\beta - \alpha} \ .$$

Finally, we substitute these into Fig. 12 to recover the parametrized HMM the main article introduced in Fig. 5.