Reductions of Hidden Information Sources

Nihat Ay[1, 2, *] and James P. Crutchfield[3, 2, †]

[1]*Mathematical Institute, Friedrich-Alexander University Erlangen-Nuremberg,*
Bismarckstr. 1, 91054 Erlangen, Germany
[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*
[3]*Computational Science & Engineering Center and Physics Department,*
University of California, Davis, California 95616

(Dated: May 14, 2005)

In all but special circumstances, measurements of time-dependent processes reflect internal structures and correlations only indirectly. Building predictive models of such hidden information sources requires discovering, in some way, the internal states and mechanisms. Unfortunately, there are often many possible models that are observationally equivalent. Here we show that the situation is not as arbitrary as one would think. We show that generators of hidden stochastic processes can be reduced to a minimal form and compare this reduced representation to that provided by computational mechanics—the ϵ-machine. On the way to developing deeper, measure-theoretic foundations for the latter, we introduce a new two-step reduction process. The first step (internal-event reduction) produces the smallest observationally equivalent σ-algebra and the second (internal-state reduction) removes σ-algebra components that are redundant for optimal prediction. For several classes of stochastic dynamical systems these reductions produce representations that are equivalent to ϵ-machines.

PACS numbers: 02.50.Ey, 02.50.Ga, 05.45.Tp, 89.70.+c

Santa Fe Institute Working Paper 04-05-011; arxiv.org/abs/math.PR/0405425

Contents

I. INTRODUCTION

Experiment and simulation often produce voluminous amounts of data—data that the scientist or analyst attempts to understand by building predictive models. The best models, however, do more than simply predict the data. In the best of circumstances, models also capture the internal structures, active degrees of freedom, correlations, and so on that underlie the observations. In this way modeling enhances understanding and leads to new insights about the forces that shape our world.

Unfortunately, measurements generally are only indirect indicators of internal structure. This makes the process of model building difficult and often highly nonunique. One would hope that there is some principled approach to model building and inference that would guide us in inferring structural properties from data. The possibilities for such an approach are bounded by two extremes: (i) Are there formal constraints that guide the discovery of good representations? (ii) Can the observations themselves tell us which representation to use or, perhaps, how to refine an existing model or to correct an initially faulty hypothesis?

These days, however, the problem of building useful predictors of hidden information sources is compounded by the fact that the systems studied are quite complicated, in the sense of consisting of many components, for example. Genomic, geophysical, neurobiological, Internet traffic, and World Wide Web systems easily come to mind as complex in this sense and as particularly desirable to model. This very practical observation, in turn, argues even more forcefully for a principled approach to discovering and describing hidden structure. That is, we now need to understand the process of model building for such complicated systems well enough to teach machines how to do it.

[*] Electronic address: ay@mi.uni-erlangen.de
[†] Electronic address: chaos@cse.ucdavis.edu

Here, building on previous work [1–3], we address one piece in this puzzle—what we call the *Forward Modeling Problem*: Given a generator of an observed stochastic process, Is there a minimal, optimal predictor of it? In answering this question positively, we have two goals. The first, naturally enough, is to articulate the notion of minimal generators of observed stochastic processes and show that they exist. The second, though, is to lay more rigorous and broader foundations than currently available for the *Reverse Modeling Problem*—Given observations, can one reconstruct the hidden mechanisms?

A. Background

Reviewing a little background on previous work will help put the current formal results in perspective and motivate our development. We then comment on closely related work in which similar questions arise, but which take different approaches to structural inference. Then, after outlining our approach, the mathematical development begins.

If we are to build a predictive model of an information source that produces a time series, the most basic assumption to make is that the source, at each moment of time, is in some "state". Over time, the source transitions from state to state. As we noted already, though, in the general setting we do not have access to these states, we only have indirect information about them—information that we call *measurements*. So the modeling question reduces to the following. Given that all we have are sequences of observations, what kind of "state" should be formed from them and used for modeling? The answer is rather straightforward, and seemingly tautological: the "states" that we should use are those that are effective for prediction.

This is the starting point for how *computational mechanics* [1–3] builds optimal models. One of the notable results in computational mechanics, though, is that the representation, which emerges from focusing on states that are effective for prediction, captures all of a process's internal causal structure. In fact, computational mechanics shows that there is a preferred representation for modeling, which is called an *ϵ-machine*.

To start, we consider a time series of observations $\overleftrightarrow{s} = \ldots, s_{-2}, s_{-1}, s_0, s_1, \ldots$, in which the individual measurements are symbols in a finite alphabet: $s_i \in \Delta$. An ϵ-machine consists of states—called *causal states* and denoted $\boldsymbol{\mathcal{S}}$—and transitions between them. The causal states are defined as those *sets of histories* $\overleftarrow{s}_t = \ldots, s_{t-3}, s_{t-2}, s_{t-1}$ that are equivalent for predicting the *future* $\overrightarrow{s}_t = s_t, s_{t+1}, s_{t+2}, \ldots$. That is, two histories—$\overleftarrow{s}$ and $\overleftarrow{s}'$—are associated with a given causal state, when the *sets* of possible futures "look" the same having seen them. More precisely, this modeling principle defines an equivalence relation $\sim$ over the set $\overleftarrow{\mathbf{S}}$ of histories:

$$\overleftarrow{s} \sim \overleftarrow{s}' \text{ if and only if } \mathrm{P}(\overrightarrow{s} \,|\overleftarrow{s}) = \mathrm{P}(\overrightarrow{s} \,|\overleftarrow{s}') \ , \qquad (1)$$

where in the conditional distribution equality we mean that each individual future is given the same probability. The resulting equivalence classes are the causal states.

From this, one can show that the ϵ-machine for an information source is the optimal, minimal, and unique predictor of an information source. In the language of mathematical statistics, the ϵ-machine is a minimal sufficient statistic for the observed stochastic process produced by an information source. More than being a good predictor that is small, the semigroup determined by the causal states and transitions captures all of the information source's internal structure—regularities, symmetries, and so on. And, due to minimality, one can show that the *statistical complexity* C_μ—the "size" of an ϵ-machine measured as the Shannon entropy of the set of causal states—measures the amount of historical information that the source stores. That is, ϵ-machine minimality is not only helpful in terms of compact representations, but it is essential, since C_μ gives one a quantitative way to say how structured a hidden information source is.

Although the emphasis here is on the mathematical foundations of computational mechanics, we should note that it has been used to analyze structural complexity in a wide range of information sources. These include cellular automata [4], one-dimensional maps [1, 5], and the one-dimensional Ising model [6, 7], as well as several experimental systems, such as the dripping faucet [8], atmospheric turbulence [9], geomagnetic data [10], complex materials [11, 12], and molecular dynamics [13, 14].

In the present work we begin to address the problems posed in Appendix H.3 of Ref. 3 on founding computational mechanics more fully on stochastic process and measure theories by considering one part of the Forward Modeling Problem noted above. The results here differ from previous work on computational mechanics in two ways. First, the development is mathematically rigorous, in the sense that we use measure theory to explore the notion of minimal representations, which underlies ϵ-machines. What is novel compared to stochastic process theory is that we ask for minimal representations of a stochastic process and express them in terms of the minimal σ-algebra. We also introduce two new components of the minimization procedure—internal-event and internal-state reduction—which complement the existing concept of causal-state reduction for ϵ-machines. Analyzing the Forward Modeling Problem in this way allows us to draw parallels with the computational mechanics development of ϵ-machines, comparing and contrasting the various kinds of reduction method. We show that in a number of cases these reductions are equivalent and so provide an extension of the original concept of an ϵ-machine to a broader class of processes than previously possible.

B. Related Work

The modeling questions that we address here, and that are also addressed by computational mechanics, do not arise in a vacuum. Here we briefly mention related work that is motivated by similar concerns of the equivalence of observed processes and of structural inference, but that adopts different approaches. In a later section, when we turn to discuss our results, we broaden the discussion of related work to mention additional areas in which one might find useful applications.

One of the first attempts to address the difficulties of analyzing (known) hidden information sources is that of Ref. 15. The problem, which comes under the heading of the *identifiability of functions of Markov chains*, was to calculate the source entropy rate, given an internal finite-state Markov chain, the states of which are observed with a probabilistic measurement function. (Note that today one refers to this class of information sources as *hidden Markov models* [16, 17].) There it was shown that in the majority of cases there are no closed-form expressions for the entropy rate. A corollary of this result is that one needs to determine the effective states (and these might be infinite in number) in order to calculate a property as basic as the entropy rate—that is, simply attempting to determine how random a finite-state information source is. This contrasts, of course, with Shannon's closed-form expression for finite Markovian sources [18]. We take Ref. 15's result as one of the first indications of the nontrivial nature of inferring the structure of hidden information sources. Another testimony to this difficulty is that the problem of identifiability itself, though posed by Blackwell and Koopmans in the late 1950s, was not solved for almost 40 years [19]. Moreover, the existence of minimal representations of these same hidden sources was not established until a few years later still [20, 21].

Similar concerns about inference, representation, and causality are found in the fields of causal inference [22], graphical models [23], and nonlinear time series analysis and state-space reconstruction [24]. Most of the work in these areas proceeds by *assuming* a given set of observed and hidden variables (and their connectivity) and then asks for efficient algorithms to estimate various kinds of marginal, conditional, and joint distributions. The goals are to infer from the latter the relationship between these variables and so, on that basis, to draw structural conclusions. That is, in these cases one begins with strong structural priors about the internal architecture of a hidden information source in order to initiate analysis. (Indeed, such priors are often a useful and necessary starting point for modeling.) Notably, only the last of these fields concentrates on temporal dynamics and sources with memory. Here we are interested in both architectural and temporal properties of memoryful hidden sources and wish to understand these employing a minimum of structural priors.

C. Outline

The principle focus of the following is to develop the notion of a minimal reduction of a given (hidden) Markov process. To do this, the development is organized as follows. In the next section we characterize the (rather general) class of stochastic processes—hidden information sources—in a way that respects the distinction between a process's internal structure and the measurements, which indirectly reflect the internal state, available to an observer. This then allows us to define generators of stochastic processes as *Markov transition kernels* and so state the problem of observationally equivalent generators. The succeeding section establishes how different generators can be mapped onto each other while maintaining observational equivalence. Then, in the next section, we address the central problem and show that one can maximally reduce the representation of a process's internal structure—it's generator—while still producing the same observed stochastic process. The reduction is achieved in two steps—the first, called *internal-event reduction*, produces the smallest σ-algebra and the second, *internal-state reduction*, reduces the internal structure further, removing components that are not necessary for optimal prediction. During the development we illustrate the ideas with several examples that show how the new formulation extends the range of applicability of computational mechanics.

II. GENERATORS OF STOCHASTIC PROCESSES

An information source is a process that at each time step emits an output or measurement symbol. Only the probabilistic nature of the output process is specified in order to describe the observed information processing. Indeed, often in information theory a source is mathematically described as a stochastic process without concrete specification of internal mechanisms. In many theories of complexity, however, one often uses explicitly structural notions (e.g., *automata*) from the theory of discrete computation [25] to describe the resources required to reproduce or model an observed process. So that we will have a mathematical model that both captures the observed stochastic process and allows for a range of internal structures, we adapt the concept of finite-state automata to the setting of stochastic processes as follows; cf. Refs. 26 and 17.

We consider a finite set Q of *internal states* of the system and also a finite set Δ of *output states*, which are the observed symbols. The internal structure is modeled in various ways. First, it can be specified by a deterministic (det) transition map:

$$T_{\mathrm{det}}: \; Q \;\to\; Q \times \Delta, \qquad x \;\mapsto\; T_{\mathrm{det}}(x) = (y, s) \; . \quad (2)$$

This map assigns to each internal state $x \in Q$ the next internal state y and, at the same time, also the next

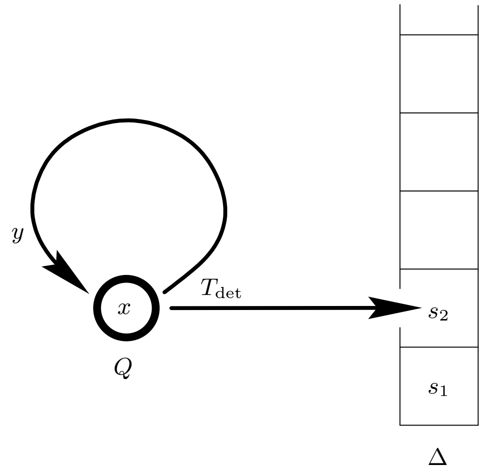

FIG. 1: The transition structure of a deterministic machine that generates a stochastic output process.

output symbol $s \in \Delta$. Figure 1 illustrates the transition structure.

A nondeterministic (non) version of such a *machine (without input)* can be introduced as a map

$$T_{\text{non}}: \; Q \; \to \; 2^{Q\times\Delta} \; , \qquad x \mapsto C \; . \tag{3}$$

This machine assigns to each internal state x a set C of possible next-state pairs (y, s). This extends the deterministic machine T_{det} of Eq. (2), which can be interpreted within the nondeterministic framework as follows:

$$T_{\text{non}}(x) \; := \; \{T_{\text{det}}(x)\} \; \in \; 2^{Q\times\Delta} \; .$$

Finally, a further extension is provided by the following probabilistic (pr) interpretation of a nondeterministic machine T_{non}:

$$T_{\text{pr}}: \; Q \times 2^{Q\times\Delta} \; \to \; [0,1] \; ,$$

where

$$(x, C) \; \mapsto \; T_{\text{pr}}(x, C) := \frac{|C \cap T_{\text{non}}(x)|}{|T_{\text{non}}(x)|} \; .$$

The function T_{pr} satisfies

$$T_{\text{pr}}(x, C_1 \uplus C_2) \; = \; T_{\text{pr}}(x, C_1) + T_{\text{pr}}(x, C_2)$$

and

$$T_{\text{pr}}(x, Q \times \Delta) \; = \; 1 \; .$$

Therefore T_{pr} is a *Markov transition kernel* on finite symbols.

This interpretation allows for an extension of finite-state machines to machines given by general Markov transition kernels that are not restricted to finite symbols, for example. Here, we allow the internal states to be described by an arbitrary measurable space $(Q, \mathcal{Q})$.

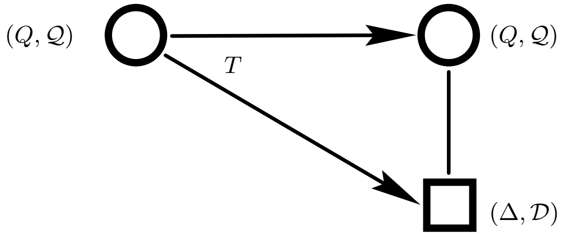

FIG. 2: The Markov transition kernel: The internal structure is specified by $(Q, \mathcal{Q})$ and the observed process by $(\Delta, \mathcal{D})$.

Again, Q is the set of internal states or, in terms of probability theory, the set of (internal) elementary events. The σ-algebra $\mathcal{Q}$ represents all internal events of interest. The output is modeled by a measurable space $(\Delta, \mathcal{D})$, too. A machine is now considered to be a Markov transition kernel:

$$T: \; Q \times \big(\mathcal{Q} \otimes \mathcal{D}\big) \; , \qquad (x, C) \; \mapsto \; T(x, C) \; .$$

More precisely, T is assumed to satisfy the following conditions:

1. For all $x \in Q$, the function $T(x, \cdot)$ is a probability distribution on $\mathcal{Q} \otimes \mathcal{D}$.

2. For all $C \in \mathcal{Q} \otimes \mathcal{D}$, the function $T(\cdot, C)$ is $\mathcal{Q}$-measurable.

Finally, we assume that all measurable spaces $(Q, \mathcal{Q})$—the measurable space of internal states or hidden events—and $(\Delta, \mathcal{D})$—the measurable space of visible events—are Polish spaces.

We should point out that the well established notion of machines that manipulate finitely many (or a countable number of) symbols may seem more appropriate for implementations in physical systems than our broad approach to computation using general Markov transition kernels. Putting the natural ideas of computation into the probabilistic setting, however, allows us to employ measure-theoretic concepts and techniques. This approach turns out to be very useful in understanding the relations between the probabilistic nature of the observed processes and the underlying internal computational structures. In particular, problems on minimality properties of machines can be handled in an efficient way and for a broader class of processes than those over discrete symbols.

Given a Markov transition kernel T from $(Q, \mathcal{Q})$ to $(Q \times \Delta, \mathcal{Q} \otimes \mathcal{D})$, we consider it as a temporal "map", as illustrated in Fig. 2. In order to specify observable stochastic processes in $(\Delta, \mathcal{D})$, we consider an initial distribution μ on $(Q, \mathcal{Q})$ and measurable sets $B_1, \ldots, B_n \in$

$\mathcal{D}$. The finite-dimensional marginals $\mathrm{P}_n^{\mu,T}$ on $(\Delta^n, \mathcal{D}^n)$ are obtained by iteration of T, as shown in Fig. 3.

This suggests the following expression for the finite-dimensional marginals of the observed stochastic process:

$$\mathrm{P}_n^{\mu,T}(B_1 \times \cdots \times B_n) := \int_Q \int_{Q\times B_1} \cdots \int_{Q\times B_n} T(x_{n-1}, d(x_n, y_n)) \cdots T(x_0, d(x_1, y_1))\, \mu(dx_0)\ .$$

(Throughout the following $d(x,y)$ denotes the differential of two variables. This notation should not be confused with a distance measure between x and y.)

Proposition 2.1. *Up to equivalence, there is exactly one stochastic process Y_n, $n = 1, 2, \ldots,$ in $(\Delta, \mathcal{D})$, such that for all $n \in \mathbb{N}$ and all $B_i \in \mathcal{D}$, $i = 1, \ldots, n$,*

$$\Pr\{Y_1 \in B_1, \ldots, Y_n \in B_n\} \ = \ \mathrm{P}_n^{\mu,T}(B_1 \times \cdots \times B_n)\ . \quad (4)$$

We can identify this process or, more precisely, the class of corresponding equivalent processes, with a probability distribution $\mathrm{P}^{\mu,T}$ on $(\Delta^{\mathbb{N}}, \mathcal{D}^{\mathbb{N}})$.
Proof. This follows from Kolmogorov's extension theorem [27]. □

Definition 2.2. We call a Markov transition kernel T from $(Q, \mathcal{Q})$ to $(Q \times \Delta, \mathcal{Q} \otimes \mathcal{D})$ a *generator* and denote it by $[(Q, \mathcal{Q}), T, (\Delta, \mathcal{D})]$ or simply by T. We say that a stochastic process $(Y_n)_{n\in\mathbb{N}}$ in $(\Delta, \mathcal{D})$ is *generated* by T if there exists a probability distribution μ on $(Q, \mathcal{Q})$ such that Eq. (4) is satisfied for all $n \in \mathbb{N}$ and all $B_i \in \mathcal{D}$, $i = 1, \ldots, n$.

Given a stochastic process $Y = (Y_n)_{n\in\mathbb{N}}$, a natural question is whether there always exists a generator that generates Y. The following trivial shift ansatz shows that this is indeed the case.

Example 2.3 (Shift Generator): We set $Q := \Delta^{\mathbb{N}}$ and $\mathcal{Q} := \mathcal{D}^{\mathbb{N}}$. Consider the shift map $s : Q \to Q$, where

$$x = (y_n)_{n\in\mathbb{N}} \ \mapsto \ s(x) = (y_{n+1})_{n\in\mathbb{N}}\ ,$$

and the projection onto the first coordinate $\pi : Q \to \Delta$, where

$$x = (y_n)_{n\in\mathbb{N}} \ \mapsto \ \pi(x) = y_1\ .$$

Furthermore, we define

$$T(x, A \times B) \ := \ \begin{cases} 1, & \text{if } s(x) \in A \text{ and } \pi\big(s(x)\big) \in B \\ 0, & \text{otherwise}\ . \end{cases}$$

Now, a stochastic process $Y = (Y_n)_{n\in\mathbb{N}}$ in $(\Delta, \mathcal{D})$ can be identified with a probability distribution μ on $(Q, \mathcal{Q}) = (\Delta^{\mathbb{N}}, \mathcal{D}^{\mathbb{N}})$. It is easy to prove that T generates Y by verifying Eq. (4) with initial distribution μ. □

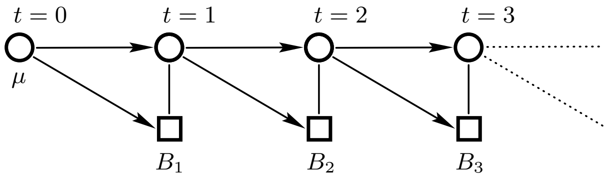

FIG. 3: The finite-dimensional marginals $\mathrm{P}_n^{\mu,T}$ on $(\Delta^n, \mathcal{D}^n)$.

The shift generator (Example 2.3) is maximal in the sense that it generates all processes in $(\Delta, \mathcal{D})$. (It is the mathematical analog of the experimentalist who reports each and every measurement, without summary, compression, or explanation.) For an arbitrary generator T, we consider the map

$$G_T : \ \boldsymbol{P}(Q, \mathcal{Q}) \ \to \ \boldsymbol{P}(\Delta^{\mathbb{N}}, \mathcal{D}^{\mathbb{N}}), \qquad \mu \ \mapsto \ G_T(\mu)\ .$$

(Throughout, for a general measurable space $(X, \mathcal{X})$, $\boldsymbol{P}(X, \mathcal{X})$ denotes the set of probability measures on $(X, \mathcal{X})$.) The image $\mathrm{im}(G_T)$ of G_T is the set of processes that are generated by T. Here, we mainly focus on the following problem.

Problem Statement 2.4: *Given a generator T, can we find a substitute T' for T, which, on the one hand, generates the same set of processes, that is $\mathrm{im}(G_T) = \mathrm{im}(G_{T'})$, and, on the other, is minimal in some sense?*

From Eq. (4) it follows directly that G_T is affine in the sense that for all $\mu_1, \mu_2 \in \boldsymbol{P}(Q, \mathcal{Q})$ and all $0 \le t \le 1$,

$$G_T\big((1-t)\,\mu_1 + t\,\mu_2\big) \ = \ (1-t)\,G_T(\mu_1) + t\,G_T(\mu_2)\ . \quad (5)$$

This implies that $\mathrm{im}(G_T)$ is a convex set, and we have the following constraint on the solution of Problem 2.4: The set $\mathrm{ext}\big(\mathrm{im}(G_T)\big)$ of the extreme points of $\mathrm{im}(G_T)$ represents a "lower bound" for the set Q of internal states. More precisely, we have the following onto mapping $Q \to \mathrm{ext}\big(\mathrm{im}(G_T)\big)$:

$$x \ \mapsto \ \delta_x \ \mapsto \ G_T(\delta_x)\ .$$

Thus, we cannot expect to have a notion of minimality that reduces the internal states more than given by the extreme points of $\mathrm{im}(G_T)$.

However, identifying internal states x_1 and x_2 if $G_T(\delta_{x_1}) = G_T(\delta_{x_2})$ leads to a partition of Q into equivalence classes—classes that are the analogs of the *causal states* in computational mechanics. The corresponding canonical projection of internal states to their equivalence classes is called *causal-state reduction*, which is intended to reduce the internal structure in such a way that a given observed stochastic process is still generated by the reduced generator.

This is different from the intention stated in Problem 2.4, which is to reduce a given generator without affecting the *whole* set of observable stochastic processes. We solve this problem by applying reductions within a natural category of generators. The morphisms of this category will be introduced in Section III. Based on the results there, we present our reduction procedures in Section IV. We leave to the future discussing causal-state reduction in terms of morphisms in a larger category than the one studied here.

III. TRANSFORMATION RULES FOR GENERATORS

We interpret generators as objects of a category and define the morphisms between these objects in the following way: Let $[(Q_i, \mathcal{Q}_i), T_i, (\Delta_i, \mathcal{D}_i)]$, $i = 1, 2$, be two generators. A morphism $T_1 \to T_2$ consists of a pair (f, g) of measurable maps $f : Q_1 \to Q_2$ and $g : \Delta_1 \to \Delta_2$ such that for all $x \in Q_1$, $A \in \mathcal{Q}_2$, and $B \in \mathcal{D}_2$ the following commutativity rule holds:

$$T_2(f(x), A \times B) \ = \ T_1\left(x, f^{-1}(A) \times g^{-1}(B)\right) \ . \qquad (6)$$

The diagram in Fig. 4 illustrates this commutativity.

With the product map

$$(f{\times}g) : \ \ Q_1{\times}\Delta_1 \ \to \ Q_2{\times}\Delta_2, \qquad (x, y) \ \mapsto \ (f(x), g(y)) \ ,$$

we can rewrite Eq. (6) as

$$T_2(f(x), A \times B) \ = \ T_1\big(x, (f \times g)^{-1}(A \times B)\big) \ .$$

Thus, the property of Eq. (6) is equivalent to

$$T_2(f(x), C) \ = \ T_1\big(x, (f \times g)^{-1}(C)\big) \ , \qquad (7)$$

with $C \in \mathcal{Q}_2 \otimes \mathcal{D}_2$. Here, one has to use the fact that two probability measures are equal if they coincide on an intersection closed system of measurable sets that generates the underlying σ-algebra [27]. Rewriting (7) gives us

$$T_2(f(x), \cdot) \ = \ (f \times g)_*\big(T_1(x, \cdot)\big) \ , \qquad (8)$$

where $(f \times g)_*(\mu)$ denotes the $(f \times g)$-image of a probability distribution μ.

In the following, a morphism (f, g) is called a *transition-preserving map.* In order to define the composition of transition-preserving maps, we consider three generators $[(Q_i, \mathcal{Q}_i), T_i, (\Delta_i, \mathcal{D}_i)]$, $i = 1, 2, 3$, and transition-preserving maps $(f_i, g_i) : T_i \to T_{i+1}$, $i = 1, 2$. Now define the composition as

$$(f_2, g_2) \circ (f_1, g_1) \ := \ (f_2 \circ f_1, g_2 \circ g_1) \ .$$

We prove that this composition is a transition-preserving map $T_1 \to T_3$ by verifying Eq. (6):

$$\begin{aligned} T_3\big((f_2 \ \circ \ f_1)(x), A \times B\big) &= T_3\Big(f_2\big(f_1(x)\big), A \times B\Big) \\ &= \ T_2\big(f_1(x), f_2^{-1}(A) \times g_2^{-1}(B)\big) \\ &= \ T_1\Big(x, f_1^{-1}\big(f_2^{-1}(A)\big) \times g_1^{-1}\big(g_2^{-1}(B)\big)\Big) \\ &= \ T_1\big(x, (f_2 \circ f_1)^{-1}(A) \times (g_2 \circ g_1)^{-1}(B)\big) \ . \end{aligned}$$

Proposition 3.1. *Let $[(Q_i, \mathcal{Q}_i), T_i, (\Delta_i, \mathcal{D}_i)]$, $i = 1, 2$, be two generators, and let (f, g) be a transition-preserving map from T_1 to T_2, and let μ be a probability distribution on $(Q_1, \mathcal{Q}_1)$. Then, denoting the f-image of μ by $f_*(\mu)$, for all $B_1, \ldots, B_n \in \mathcal{D}_2$,*

$$\mathrm{P}_n^{f_*(\mu), T_2}(B_1{\times}\cdots{\times}B_n) \ = \ \mathrm{P}_n^{\mu, T_1}(g^{-1}(B_1){\times}\cdots{\times}g^{-1}(B_n)) \ .$$

Proof. With the general transformation rule for integrals we have

$$\begin{aligned} \mathrm{P}_n^{f_*(\mu), T_2}(B_1 \times \cdots \times B_n) \ &= \ \int_{Q_1}\int_{Q_2\times B_1}\cdots\int_{Q_2\times B_n} T_2(x'_{n-1}, d(x'_n, y'_n)) \cdots T_2(x'_0, d(x'_1, y'_1))\, f_*(\mu)(dx'_0) \\ &= \ \int_{Q_1}\int_{Q_2\times B_1}\cdots\int_{Q_2\times B_n} T_2(x'_{n-1}, d(x'_n, y'_n)) \cdots T_2(f(x_0), d(x'_1, y'_1))\, \mu(dx_0) \\ &\quad \text{(transformation rule)} \\ &\overset{(8)}{=} \ \int_{Q_1}\int_{Q_2\times B_1}\cdots\int_{Q_2\times B_n} T_2(x'_{n-1}, d(x'_n, y'_n)) \cdots (f \times g)_*\big(T_1(x_0, \cdot)\big)(d(x'_1, y'_1))\, \mu(dx_0) \end{aligned}$$

$$\begin{aligned}
\mathrm{P}_n^{f_*(\mu),T_2}(B_1\times\cdots\times B_n) &= \int_{Q_1}\int_{Q_1\times g^{-1}(B_1)}\cdots\int_{Q_2\times B_n} T_2(x'_{n-1},d(x'_n,y'_n))\,\cdots\,T_1(x_0,d(x_1,y_1))\,\mu(dx_0)\\
&\quad \text{(transformation rule)}\\
&\quad \vdots \qquad \vdots \qquad \vdots\\
&= \int_{Q_1}\int_{Q_1\times g^{-1}(B_1)}\cdots\int_{Q_2\times g^{-1}(B_n)} T_1(x_{n-1},d(x_n,y_n))\,\cdots\,T_1(x_0,d(x_1,y_1))\,\mu(dx_0)\\
&= \mathrm{P}_n^{\mu,T_1}(g^{-1}(B_1)\times\cdots\times g^{-1}(B_n))\ .
\end{aligned}$$

□

Theorem 3.2. *If T_1 generates $(Y_n)_{n\in\mathbb{N}}$ and (f,g) is a transition-preserving map from T_1 to T_2, then T_2 generates $(g\circ Y_n)_{n\in\mathbb{N}}$.*

Proof. This statement follows directly from Proposition 3.1. □

Theorem 3.2 has important and direct implications for two special cases. In the first case, we fix g as the identity map and, in the second case, we fix f as the identity map. In these cases, without reference to the identity maps, f and g are called transition-preserving. The implications are stated in the following two corollaries.

Corollary 3.3. *Let $[(Q_i,\mathcal{Q}_i),T_i,(\Delta,\mathcal{D})]$, $i=1,2$, be two generators, and let f be a transition-preserving map. Then*

$$G_{T_1} \;=\; G_{T_2}\circ f_*\ .$$

In particular, this implies

$$\mathrm{im}(G_{T_1}) \;\subseteq\; \mathrm{im}(G_{T_2})\ ,$$

where the equality holds if f_* is onto.

Corollary 3.4. *Let $[(Q,\mathcal{Q}),T_1,(\Delta_1,\mathcal{D}_1)]$ be a generator of a stochastic process $(Y_n)_{n\in\mathbb{N}}$ in $(\Delta_1,\mathcal{D}_1)$, and let $g:(\Delta_1,\mathcal{D}_1)\to(\Delta_2,\mathcal{D}_2)$ be a measurable map. Then $T_2: Q\times(\mathcal{Q}\otimes\mathcal{D}_2)\to[0,1]$ with*

$$T_2(x,A\times B) \;:=\; T_1\big(x,A\times g^{-1}(B)\big)$$

is a generator of the stochastic process $(g\circ Y_n)_{n\in\mathbb{N}}$ in $(\Delta_2,\mathcal{D}_2)$.

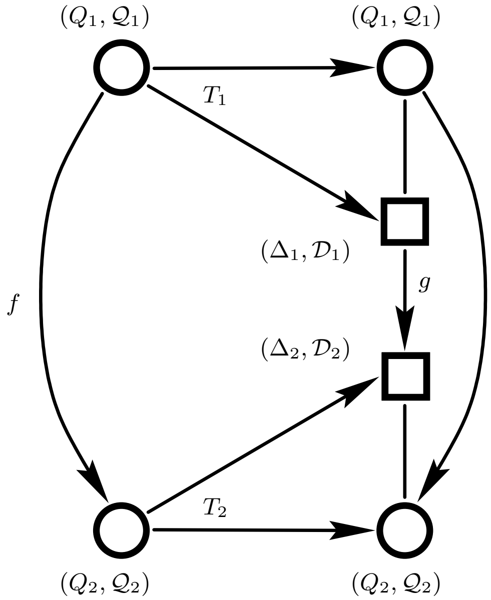

FIG. 4: Commutativity for generators of equivalent observed processes.

IV. REDUCTIONS OF GENERATORS

After having derived some basic transformation rules for generators in Section III, we are now ready to concentrate on the main problem, namely to maximally reduce a given generator T while keeping the set $\mathrm{im}(G_T)$ of generated processes unchanged. The solution of this problem is given by Theorem 4.5 below and is based on a combination of reduction methods, which we present in this section. First, we attempt to reduce the σ-algebra $\mathcal{Q}$ of internal events as much as possible, by considering only those events in $\mathcal{Q}$ that are necessary for maintaining the output process unchanged. The following theorem formalizes this idea.

Theorem (Internal-Event Reduction) 4.1. *Let $[(Q,\mathcal{Q}),T,(\Delta,\mathcal{D})]$ be a generator. Then there exists a smallest σ-subalgebra $\sigma_Q(T)$ of $\mathcal{Q}$ with the property that for all $C\in\sigma_Q(T)\otimes\mathcal{D}$, $T(\cdot,C)$ is $\sigma_Q(T)$-measurable. The generator $[(Q,\sigma_Q(T)),\bar{T},(\Delta,\mathcal{D})]$ with the restriction $\bar{T}:=T|_{Q\times(\sigma_Q(T)\otimes\mathcal{D})}$ then satisfies*

$$\mathrm{im}(G_{\bar{T}}) \;=\; \mathrm{im}(G_T)\ .$$

Proof. Let $\mathcal{A}_i$, $i\in I$, be the family of all σ-subalgebras of $\mathcal{Q}$ that satisfy the following condition: for all $C\in\mathcal{A}_i\otimes\mathcal{D}$,

$T(\cdot,C)$ is $\mathcal{A}_i$-measurable. Now define

$$\sigma_Q(T) \;:=\; \bigcap_{i\in I} \mathcal{A}_i \;.$$

Then for $C \in \sigma_Q(T)\otimes\mathcal{D}$, $T(\cdot,C)$ is $\mathcal{A}_i$ measurable for all $i\in I$, and therefore also $\sigma_Q(T)$-measurable.

For the reason that trivially

$$\bar{T}(\mathrm{id}_Q(x), A\times D) \;=\; T(x, \mathrm{id}_Q^{-1}(A)\times \mathrm{id}_\Delta^{-1}(D)) \;,$$

Corollary 3.3 implies that T and $\bar{T}$ generate the same set of stochastic processes. □

Theorem 4.1 guarantees the existence of a minimal sufficient σ-subalgebra of $\mathcal{Q}$. Now we provide a way to calculate it explicitly in the case where we have a deterministic internal dynamics $f : Q \to Q$ and a visible process given by a measurement $g : Q \to \Delta$. This case generalizes the shift generator of Example 2.3.

Theorem 4.2. *Let $(Q,\mathcal{Q})$ and $(\Delta,\mathcal{D})$ be two measurable spaces, and let $f : Q\to Q$ and $g : Q\to\Delta$ be two measurable maps. Consider the generator $[(Q,\mathcal{Q}),T,(\Delta,\mathcal{D})]$ defined by*

$$\begin{aligned} T(x, A\times B) \;&:=\; 1_{f\in A,\, g\circ f\in B}(x) \\ &=\; \begin{cases} 1, & \textit{if } f(x)\in A \textit{ and } g\big(f(x)\big)\in B \\ 0, & \textit{otherwise} \end{cases} \;. \end{aligned}$$

Then

$$\sigma_Q(T) \;=\; \sigma(g\circ f, g\circ f^2,\dots) \;. \qquad (9)$$

Proof. We prove inclusion in each direction separately.

1. We establish that $\sigma_Q(T) \subseteq \sigma(g\circ f, g\circ f^2,\dots)$ by showing that for all $C\in\sigma(g\circ f, g\circ f^2,\dots)\otimes\mathcal{D}$, $T(\cdot,C)$ is measurable with respect to $\sigma(g\circ f, g\circ f^2,\dots)$. From

$$\begin{aligned} \sigma(g \;\circ\; &f, g\circ f^2,\dots)\otimes\mathcal{D} \\ &=\; \sigma\big((g\circ f, g\circ f^2,\dots)\times \mathrm{id}_\Delta\big) \end{aligned}$$

   we know that there exists a measurable set $C'\in\mathcal{D}^{\mathbb{N}}\otimes\mathcal{D}$ with

$$C \;=\; \big((g\circ f, g\circ f^2,\dots)\times \mathrm{id}_\Delta\big)^{-1}(C') \;.$$

   This implies

$$\begin{aligned} T(\cdot,C) \;&=\; 1_{(f,g\circ f)\,\in\, C} \\ &=\; 1_{(f,g\circ f)\,\in\, \big((g\circ f,g\circ f^2,\dots)\times\mathrm{id}_\Delta\big)^{-1}(C')} \\ &=\; 1_{\big((g\circ f,g\circ f^2,\dots)\times\mathrm{id}_\Delta\big)\circ(f,g\circ f)\,\in\, C'} \\ &=\; 1_{((g\circ f^2,g\circ f^3,\dots),g\circ f)\,\in\, C'} \;\cdot \end{aligned}$$

   Thus, $T(\cdot,C)$ is measurable with respect to $(g\circ f, g\circ f^2,\dots)$.

2. Now we prove that $\sigma_Q(T) \;\supseteq\; \sigma(g\circ f, g\circ f^2,\dots)$ by applying an induction argument to show that $\sigma(g\circ f^k)\subseteq\sigma_Q(T)$ for all $k=1,2,\dots$:

   (a) "$k=1$": Let A be a $(g\circ f)$-measurable set. Then there exists a measurable set $B\in\mathcal{D}$ with $A=(g\circ f)^{-1}(B)$. From $Q\times B\in\sigma_Q(T)$, and

$$\begin{aligned} 1_A \;&=\; 1_{g\circ f\in B} \\ &=\; 1_{(f,g\circ f)\in Q\times B} \\ &=\; T(\cdot, Q\times B) \;, \end{aligned}$$

   it follows that A is $\sigma_Q(T)$-measurable.

   (b) "$k\to k+1$": We assume that $\sigma(g\circ f^k)$ is a σ-subalgebra of $\sigma_Q(T)$, and we have to show that this is also true for $\sigma(g\circ f^{k+1})$. To this end, we choose a measurable set $A\in\sigma(g\circ f^{k+1})$. There exists a measurable set $B\in\mathcal{D}$ with $A=(g\circ f^{k+1})^{-1}(B)$, and we have

$$\begin{aligned} 1_A \;&=\; 1_{g\circ f^{k+1}\in B} \\ &=\; 1_{f\in(g\circ f^k)^{-1}(B)} \\ &=\; T(\cdot, (g\circ f^k)^{-1}(B)\times\Delta) \;. \end{aligned}$$

   This implies $A\in\sigma_Q(T)$, because according to the induction hypothesis $(g\circ f^k)^{-1}(B)\in\sigma_Q(T)$.

□

Examples 4.3.

1. **Complete Randomness.** Consider a probability space $(Q,\mathcal{Q},\mu)$. This defines the following generator $[(Q,\mathcal{Q}),T,(Q,\mathcal{Q})]$ which is completely random in the sense that the next internal state, which coincides with the next output state, is independent of the current internal state:

$$T:\;\; Q\times(\mathcal{Q}\otimes\mathcal{Q})\to[0,1] \;,$$

   and

$$T(x, A\times B) \;:=\; \mu(A\cap B) \;.$$

   In this case

$$\sigma_Q(T) \;=\; \{\emptyset, Q\}.$$

   In other words, as expected, the process has no memory. Only a single internal event is required to generate the process $\mu^{\otimes\mathbb{N}}$ and $\mu^{\otimes\mathbb{N}}$ is the only process in $\mathrm{im}(G_T)$.

2. **Rotation of the Unit Circle.** Consider the unit circle $K=\{x\in\mathbb{C} : |x|=1\}$ and its upper half $A_1=\{e^{i\,\varphi} : \varphi\in[0,\pi)\}$ and its lower half $A_2=\{e^{i\,\varphi} : \varphi\in[\pi,2\,\pi)\}$. With a number $a\in K$, we construct the generator T according to Theorem 4.2 using $f(x)=a\,x$ and $g(x)=k$ for $x\in A_k$. There are two qualitatively different cases:

(a) Assume that a is a root of unity. Then there is a natural number $p \neq 0$ with $a^p = 1$. This implies $f^p = \mathrm{id}_K$ and, therefore,

$$\begin{aligned} \sigma_Q(T) &= \sigma(g \circ f, g \circ f^2, \dots) \\ &= \sigma(g \circ f, g \circ f^2, \dots, g \circ f^{p-1}) \ . \end{aligned}$$

Since g has just two different values, $\sigma_Q(T)$ is finite in this case and we have an effective internal-event reduction.

(b) Assume that a is not a root of unity. Then $\sigma_Q(T)$ is the Borel algebra of the unit circle, and we have no internal-event reduction.

□

In addition to the reduction method given by Theorem 4.1, we now consider another way to reduce the generator's internal structure. Given a generator $[(Q,\mathcal{Q}),T,(\Delta,\mathcal{D})]$, we identify each two elements $x_1, x_2 \in Q$ if $T(x_1,\cdot) = T(x_2,\cdot)$. The equivalence class of x is denoted by $[x]$. Furthermore, we define

$$[Q] \ := \ \big\{[x] \ : \ x \in Q\big\}$$

and

$$[\mathcal{Q}] \ := \ \big\{A' \subseteq [Q] \ : \ [\cdot]^{-1}(A') \in \mathcal{Q}\big\} \ .$$

The σ-algebra $[\mathcal{Q}]$ is just the terminal algebra of the canonical projection $[\cdot] : x \mapsto [x]$. It is easy to see that the following transition kernel $[T] : [Q] \times \big([\mathcal{Q}] \otimes \mathcal{D}\big) \to [0,1]$ is well defined

$$[T]([x], A' \times B) \ := \ T(x, [\cdot]^{-1}(A') \times B) \ .$$

Theorem (Internal-State Reduction) 4.4. *Let $[(Q,\mathcal{Q}),T,(\Delta,\mathcal{D})]$ be a generator. Then $[([Q],[\mathcal{Q}]),[T],(\Delta,\mathcal{D})]$ is a generator, which generates the same set of processes in $(\Delta,\mathcal{D})$ as T, that is,*

$$\mathrm{im}(G_{[T]}) \ = \ \mathrm{im}(G_T) \ .$$

Proof. We show that $[T]$ is a Markov transition kernel in two stages.

1. We fix $[x]$ and prove that $[T]([x],\cdot)$ is a probability measure:

$$\begin{aligned} [T]\left([x], \biguplus_{n=1}^{\infty} C_n\right) &= T\left(x, \big([\cdot] \times \mathrm{id}_\Delta\big)^{-1}\left(\biguplus_{n=1}^{\infty} C_n\right)\right) \\ &= T\left(x, \biguplus_{n=1}^{\infty} \big([\cdot] \times \mathrm{id}_\Delta\big)^{-1}(C_n)\right) \\ &= \sum_{n=1}^{\infty} T\Big(x, \big([\cdot] \times \mathrm{id}_\Delta\big)^{-1}(C_n)\Big) \\ &= \sum_{n=1}^{\infty} [T]\big([x], C_n\big) \ , \end{aligned}$$

and

$$\begin{aligned} [T]\left([x], [Q] \times \Delta\right) &= T\Big(x, \big([\cdot] \times \mathrm{id}_\Delta\big)^{-1}\big([Q] \times \Delta\big)\Big) \\ &= T\left(x, Q \times \Delta\right) \\ &= 1 \ . \end{aligned}$$

2. Now we fix $C \in [\mathcal{Q}] \otimes \mathcal{D}$ and prove that $[T](\cdot, C)$ is $[\mathcal{Q}]$-measurable. To this end, it is sufficient to prove that for all ε with $0 \leq \varepsilon \leq 1$, the set $\{[T](\cdot,C) \leq \varepsilon\}$ is an element of $[\mathcal{Q}]$ or equivalently $[\cdot]^{-1}\big(\{[T](\cdot,C) \leq \varepsilon\}\big) \in \mathcal{Q}$. This is shown as follows.

$$\begin{aligned} &[\cdot]^{-1}\big(\{[T](\ \cdot\ , C) \leq \varepsilon\}\big) \\ &\qquad = \ [\cdot]^{-1}\big(\{[x] \in [Q] \ : \ T'([x], C) \leq \varepsilon\}\big) \\ &\qquad = \ \{x \in Q \ : \ [T]([x], C) \leq \varepsilon\} \\ &\qquad = \ \Big\{x \in Q \ : \ T\Big(x, \big([\cdot] \times \mathrm{id}_\Delta\big)^{-1}(C)\Big) \leq \varepsilon\Big\} \\ &\qquad \in \ \mathcal{Q} \ . \end{aligned}$$

□

Combining the reduction methods provided by Theorem 4.1 and Theorem 4.4, we can reduce every generator to a minimal generator. This statement is specified in the following theorem.

Theorem (Solution of Problem 2.4) 4.5. *Let $[(Q,\mathcal{Q}),T,(\Delta,\mathcal{D})]$ be a generator, and let $[(Q',\mathcal{Q}'),T',(\Delta,\mathcal{D})]$ be the generator obtained from T by applying first the reduction method of Theorem 4.1 and then the method of Theorem 4.4. Then T' satisfies*

$$\mathrm{im}(G_{T'}) \ = \ \mathrm{im}(G_T)$$

and is minimal in the sense that given another generator $[(Q'',\mathcal{Q}''),T'',(\Delta,\mathcal{D})]$ with $\mathrm{im}(G_{T''}) = \mathrm{im}(G_{T'})$, every transition-preserving map f from T' to T'' is injective.

Proof. Again there are two steps.

1. We prove

$$\sigma(f \circ [\cdot]) \ = \ \sigma_Q(T) \ . \qquad (10)$$

(a) "$\subseteq$": This inclusion follows directly from the measurability of

$$\begin{aligned} (Q,\mathcal{Q}) &\xrightarrow{\mathrm{id}_Q} (Q, \sigma_Q(T)) \\ &\xrightarrow{[\cdot]} ([Q], [\sigma_Q(T)]) = (Q',\mathcal{Q}') \\ &\xrightarrow{f} (Q'',\mathcal{Q}'') \ . \end{aligned}$$

(b) "$\supseteq$": Let $C \in \sigma(f \circ [\cdot]) \otimes \mathcal{D}$. We prove that $T(\cdot, C)$ is $(f \circ [\cdot])$-measurable, from which $\sigma_Q(T) \subseteq \sigma(f \circ [\cdot])$ follows, because $\sigma_Q(T)$ is the smallest σ-algebra with that invariance property: From

$$\begin{aligned} \sigma(f \circ [\cdot]) \otimes \mathcal{D} &= \sigma(f \circ [\cdot]) \otimes \sigma(\mathrm{id}_\Delta) \\ &= \sigma\big((f \circ [\cdot]) \times \mathrm{id}_\Delta\big) \ , \end{aligned}$$

it follows that there exists $C'' \in \mathcal{Q}'' \otimes \mathcal{D}$ with

$$\big((f \circ [\cdot]) \times \mathrm{id}_\Delta\big)^{-1}(C'') \;=\; C.$$

This implies the $(f \circ [\cdot])$-measurability of $T(\cdot, C)$:

$$\begin{aligned} T(x, C) \;&=\; T\Big(x, \big((f \circ [\cdot]) \times \mathrm{id}_\Delta\big)^{-1}(C'')\Big) \\ &=\; T''\big((f \circ [\cdot])(x), C''\big) \\ &=\; \big(T''(\cdot, C'') \circ (f \circ [\cdot])\big)(x) \;. \end{aligned}$$

2. Using Eq. (10), we now prove that f is injective. Assume $f([x_1]) = f([x_2])$ where $[x_i]$ are equivalence classes in Q; that is, $[x_1], [x_2] \in [Q]$. In order to prove injectivity of f, we have to show $[x_1] = [x_2]$:

$$\begin{aligned} \bar{T}(x_1, C) \;&=\; \bar{T}\Big(x_1, \big((f \circ [\cdot]) \times \mathrm{id}_\Delta\big)^{-1}(C'')\Big) \\ &=\; T''(f([x_1]), C'') \\ &=\; T''(f([x_2]), C'') \\ &=\; \bar{T}\Big(x_2, \big((f \circ [\cdot]) \times \mathrm{id}_\Delta\big)^{-1}(C'')\Big) \\ &=\; \bar{T}(x_2, C) \;. \end{aligned}$$

□

Examples (Continuation of Examples 4.3) 4.6.

1. **Complete Randomness.** Applying the internal-state reduction leads to an internal state space Q' consisting of one point, namely $Q' = \{Q\}$. The reduced generator is then given by

$$T'(x, \{Q\} \times B) \;=\; \mu(B) \;.$$

2. **Rotation of the Unit Circle.**

   (a) Identifying points according to the internal-state reduction leads to the grouping of all elements in a given atom of the finite σ-algebra $\sigma_Q(T)$. Thus, in this case we have finite transition kernel T' resulting from Theorem 4.5.

   (b) In this case, the internal-state reduction leads to equivalence classes that consist of individual points, so that effectively there is no reduction.

As pointed out at the end of Section II, our goals differ from those underlying causal-state reduction in computational mechanics. Nonetheless, it is not hard to see the following close relationship: In the situation of Theorem 4.2, identifying x_1 and x_2 if and only if $G_T(\delta_{x_1}) = G_T(\delta_{x_2})$ is equivalent to the identification of x_1 and x_2 if and only if $T(x_1, C) = T(x_2, C)$ for all $C \in \sigma_Q(T)$. The first identification leads to the analogs of the causal states in computational mechanics and the second identification is the one used in Theorem 4.5. For completeness, we conclude this section with the proof of this relationship.

Corollary 4.7. *Let $[(Q, \mathcal{Q}), T, (\Delta, \mathcal{D})]$ be a generator as in Theorem 4.2, and let $x_1, x_2 \in Q$. Then*

$$G_T(\delta_{x_1}) \;=\; G_T(\delta_{x_2})$$

is equivalent to

$$T(x_1, C) \;=\; T(x_2, C) \quad \textit{for all } C \in \sigma_Q(T) \;.$$

Proof.

$$\begin{aligned} & G_T(\delta_{x_1}) = G_T(\delta_{x_2}) \\ \Leftrightarrow \quad & g\big(f^k(x_1)\big) = g\big(f^k(x_2)\big) \\ & \text{for all } k = 1, 2, \dots \\ \Leftrightarrow \quad & 1_{(g\circ f, g\circ f^2, \dots) \in C'}(x_1) = 1_{(g\circ f, g\circ f^2, \dots) \in C'}(x_2) \\ & \text{for all } C' \in \mathcal{D}^{\mathbb{N}} \\ \Leftrightarrow \quad & T(x_1, C) = T(x_1, C) \\ & \text{for all } C \in \sigma\{g \circ f, g \circ f^2, \dots\} \\ \Leftrightarrow \quad & T(x_1, C) = T(x_1, C) \\ & \text{for all } C \in \sigma_Q(T) \qquad \text{(Theorem 4.2)} \;. \end{aligned}$$

□

V. DISCUSSION

After this long development, it will be helpful to discuss more informally what was achieved and how to interpret the results. We began by characterizing the class of hidden information sources in a way that respected the distinction between a source's internal structure and its observed process. That allowed us to define generators of stochastic processes as Markov transition kernels and to state the problem of observationally equivalent generators. We then established how different generators can be mapped onto each other while maintaining equivalence of the observed stochastic process. We showed that one can maximally reduce the representation of a source's generator under the same constraint. The reduction was achieved in two steps: first by internal-event reduction which produced the smallest σ-algebra and the second by internal-state reduction which collapsed σ-algebra components redundant for optimal prediction.

"Prediction" here refers to the hidden internal state *and* to the observed state of the machine in the next time step. Within computational mechanics, however, predictions are made for the whole future of the observed process, which seems more natural than trying to make predictions of the hidden states. For the class of generators that have the structure of Theorem 4.2 it turns out that both approaches are equivalent (see Corollary 4.7). We expect this equivalence to be valid for a larger class of generators but leave this to future investigations.

One interpretation of these results is that the seemingly intractable nonuniqueness of inferring models of hidden information sources can be directly addressed. There are more constraints on one's choice of representation than one thinks, at first blush. The new reductions and their sometimes-equivalence to ϵ-machine representations suggest that there might be a preferred minimal representation of general stochastic processes—the ϵ-machine or some generalization of it. Even if these minimal models are unachievable when inferring from finite data, they are, nevertheless, the goal toward which modeling should strive. We hoped to show, and partly illustrated this by the examples, that the new formulations of reductions and their relationship to causal-state reduction greatly extends the class of processes to which computational mechanics can be applied.

VI. APPLICATION AREAS

The developments here properly lie in the domains of measure theory and stochastic processes. However, we believe the results on reductions are relevant to a number of areas outside of those fields. To emphasize this, and also to suggest possible directions for future work, we shall point out the similarities with some areas and possible applications that would follow from the similarities. The areas considered are not, by any means, exhaustive. The observations are intended only to be suggestive.

Very generally, in statistical physics theories assume that a system is Markovian [28]. There is, for example, little concern about minimal representations. One consequence of this is that one sees an only indirect interest in calculating the structural and information-processing properties of physical systems. Historically, as reflected in the invention and use of order parameters, structural aspects are what the theorist introduces at the beginning of analysis. The difficulty that arises is that the systems of genuine interest often produce "order"—behaviors and structures—that is not directly determined by the fundamental equations of motion, but only arises over long times and large spatial scales. In these cases, one must adopt something like the inferential stance to discovering the emergent order, rather than assume it at the outset. All of which is to say that applying the reductions discussed here to problems in statistical mechanics should lead to novel and useful notions of structure and to quantitative methods for measuring degrees of structuredness.

In communication theory hidden information sources are called *channels* [18]. Overwhelmingly, the cases that are considered and analyzed and that, more importantly, are the basis for the central results of information theory assume channels with *no* memory [29]. Here, though, in effect we addressed channels with memory in the sense that the output symbols were not in one-to-one relationship to the channel's internal states. Indeed, to the extent the set of causal states is nontrivial, then one is confronted with memoryful information sources. Looking forward, the results on reductions should help in analyzing memoryful information sources and in quantitatively addressing the size of encoders and decoders under fixed channel fidelity.

VII. CONCLUSION

The process of model building is sometimes characterized as equivalent to data compression. While this might be true from a pragmatic engineering perspective, from the scientific, one must disagree. Model building is much more than data compression, especially to the extent that one attempts to explain and understand hidden structures and mechanisms. (See, for example, the discussion in the last section of Ref. 30.)

Building a good model certainly helps with compressing the original data, since the predictable components of a process that the model captures can be used in encoding and decoding to send only the "random" portions. However, the goal of modeling in the sciences is understanding the (possibly hidden) mechanisms and structures—elements that help explain observed phenomena and lead to new insights about how nature organizes itself. In this, minimal models—the theme of the present work—play a particularly important role. Not only do small models make for more tractable analysis and manipulation, they express how a process is structured and, in this, they allow for improved scientific understanding.

Here we addressed the Forward Modeling Problem of maximally reducing a given generator while keeping the observed process unchanged. Future work will focus on the Reverse Modeling Problem, the goal of which is to construct a minimal generator based on a distribution of measurement sequences alone. We envision a two-step approach. In the first, one constructs a possibly large but sufficient generator that, in the second step, is reduced using the results developed above. Unfortunately, the problem of ambiguity arises at the end of this procedure. From previous work in computational mechanics, however, we expect uniqueness of minimal generators up to isomorphism.

Acknowledgments

The authors thank D. Eric Smith for helpful discussions. This work was supported at the Santa Fe Institute under the Networks Dynamics Program funded by the Intel Corporation and under the Computation, Dynamics, and Inference Program via SFI's core grants from the National Science and MacArthur Foundations. Direct support was provided by DARPA Agreement F30602-00-2-0583. NA was supported by a Santa Fe Institute Post-doctoral Fellowship.

[1] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[2] J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E*, 59(1):275–283, 1999.

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

[4] James E. Hansen. *Computational Mechanics of Cellular Automata*. PhD thesis, University of California, Berkeley, 1993.

[5] James P. Crutchfield and Karl Young. Computation at the onset of chaos. In W. H. Zurek, editor, *Complexity, Entropy, and the Physics of Information*, volume VIII of *Santa Fe Institute Studies in the Sciences of Complexity*. Addison-Wesley, 1990.

[6] David P. Feldman. *Computational Mechanics of Classical Spin Systems*. PhD thesis, University of California, Davis, 1998.

[7] James P. Crutchfield and David P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E*, 55:R1239–R1242, 1997.

[8] W. M. Gonçalves, R. D. Pinto, J. C. Sartorelli, and M. J. de Oliveira. Inferring statistical complexity in the dripping faucet experiment. *Physica A*, 257:385–389, 1998.

[9] A. J. Palmer, C. W. Fairall, and W. A. Brewer. Complexity in the atmosphere. *IEEE Trans. Geosci. Remote Sens.*, 38:2056–2063, 2000.

[10] Richard W. Clarke, Mervyn P. Freeman, and Nicholas W. Watkins. The application of computational mechanics to the analysis of geomagnetic data. *Phys. Rev. E*, 67:016203, 2003.

[11] Dowman Parks Varn. *Language Extraction from ZnS*. PhD thesis, University of Tennessee, Knoxville, 2001.

[12] Dowman P. Varn, Geoffrey S. Canright, and James P. Crutchfield. Discovering planar disorder in close-packed structures from x-ray diffraction: Beyond the fault model. *Phys. Rev. B.*, 66(17):174110, 2002.

[13] D. Nerukh, G. Karvounis, and R. C. Glen. Complexity of classical dynamics of molecular systems. I. Methodology. *J. Chem. Phys.*, 117:9611–9617, 2002.

[14] D. Nerukh, G. Karvounis, and R. C. Glen. Complexity of classical dynamics of molecular systems. II. Finite statistical complexity of water-Na$^+$ system. *J. Chem. Phys.*, 117:9618–9622, 2002.

[15] D. Blackwell and L. Koopmans. On the identifiability problem for functions of Markov chains. *Ann. Math. Statist.*, 28:1011, 1957.

[16] L. R. Rabiner. A tutorial on hidden Markov models and selected applications. *IEEE Proc.*, 77:257, 1989.

[17] R. J. Elliot, L. Aggoun, and J. B. Moore. *Hidden Markov Models: Estimation and Control*, volume 29 of *Applications of Mathematics*. Springer, New York, 1995.

[18] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.

[19] H. Ito, S.-I. Amari, and K. Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Info. Th.*, 38:324, 1992.

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

[21] 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.

[22] C. Glymour and G. F. Cooper, editors. *Computation, Causation, and Discovery*, Menlo Park, California, 1999. AAAI Press.

[23] M. I. Jordan, editor. *Learning in Graphical Models*, Cambridge, Massachusetts, 1999. MIT Press.

[24] M. Casdagli and S. Eubank, editors. *Nonlinear Modeling*, SFI Studies in the Sciences of Complexity, Reading, Massachusetts, 1992. Addison-Wesley.

[25] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, 1969.

[26] A. Paz. *Introduction to Probabilistic Automata*. Academic Press, New York, 1971.

[27] H. Bauer. *Probability Theory and Elements of Measure Theory*. International Series in Decision Processes. Holt, Reinhardt, and Winston, Inc., New York, 1972.

[28] O. Penrose. *Foundations of statistical mechanics; a deductive treatment*. Pergamon Press, Oxford, 1970.

[29] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

[30] J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317 – 359, Reading, Massachusetts, 1992. Addison-Wesley.