On the Generative Nature of Prediction

Wolfgang Löhr
Nihat Ay

SFI WORKING PAPER: 2008-02-004

SANTA FE INSTITUTE

On the Generative Nature of Prediction

Wolfgang Löhr[1] & Nihat Ay[1,2]

January 30, 2008

Contents

1 Introduction **1**
1.1 Predictive models theories 1
1.2 The main idea of the paper 2

2 Predictive models of stochastic processes **5**
2.1 Generating a process 5
2.2 Our prediction setting 5

3 What does "predictive" really mean? **9**
3.1 Predictive versus strongly predictive memories 9
3.2 Implications on minimality of predictive models 13

4 Conclusions **15**

5 Acknowledgements **16**

Abstract

We propose a notion of predictive models that extends the corresponding notion generally used in computational mechanics. We show that this extension allows for hidden Markov models that are more concise than the corresponding ε-machines. Thereby it provides more consistency with related theories such as Jaeger's theory of observable operator models. Furthermore, ε-machines turn out to be minimal within the usual context of predictive models.

Keywords: hidden Markov models, computational mechanics, ε-machines, observable operator models, prediction

1 Introduction

1.1 Predictive models theories

This paper is mainly about computational mechanics, a theory introduced and further developed by Crutchfield and coworkers [Crutchfield and Young, 1989, Shalizi and Crutchfield, 2001, Ay and Crutchfield, 2005, Still and Crutchfield, 2007]. It deals with the following problems:

[1]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany
[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

Having an observed stochastic process, what is the best stochastic automaton, sometimes also termed hidden Markov model, in the sense of minimal size and maximal predictive power? If such automata exist, is there a way of explicitly constructing them? In the context of these two problems, the so-called ε-machine and its construction have been proposed as the optimal solution.

Similar approaches have been considered by Heller [Heller, 1965] and later also by Jaeger [Jaeger, 2000] within a more general algebraic setting. It turns out that Jaeger's theory of so-called observable operator models (OOM) provides a constructive method for models that are in general more concise than the corresponding ε-machine. On the other hand, although their mathematical structure represents given stochastic processes in the most natural way, OOMs do not always allow for a probabilistic interpretation as hidden Markov models. Therefore, sometimes their operational nature remains unclear. If an OOM satisfies a particular geometric condition, which is quite intricate and specified in [Heller, 1965], then it *does* allow for an interpretation as hidden Markov model. This dependence on a geometric condition complicates the explicit construction of stochastic automata that have a clear operational meaning. In this regard, the ε-machine construction within computational mechanics is easier to handle. But the assumptions for the minimality of ε-machines clearly require further specification. There are examples in which the OOM construction provides very concise hidden Markov models, whereas the ε-machine has infinitely many hidden states, so-called causal states. An example for such a hidden Markov model is given in Section 3.2. Therefore, minimality of ε-machines can only be claimed within classes of models that satisfy appropriately chosen operational constraints.

This paper makes a step towards revealing these constraints by exploring the generative nature of prediction. We propose a natural notion of predictive models that is less restrictive than the one currently applied in computational mechanics. Extending the class of models to this larger class already allows us to have predictive hidden Markov models that are smaller than the corresponding ε-machine. Recently, one step towards such an extension has been made in [Still and Crutchfield, 2007]. The focus of that paper lies more on the trade-off between predictive power and model size based on the bottleneck method [Tishby et al., 1999] and therefore allows for some prediction error. We show in the present contribution that, however, this extension also is necessary for having concise models with maximal predictive power. On the other hand, it improves the theory only in combination with our notion of predictive models.

Before going into the slightly technical details of the paper, in the following section we discuss the main idea in a somewhat simplified setting. Thereby we provide a more intuitive sketch of the structure and the main results of the paper.

1.2 The main idea of the paper

We consider a pair X_{p} and X_{f} of discrete random variables, which we interpret as past and future observations. Not all information of X_{p} is necessary for predicting X_{f}, so that one tries to compress the relevant information in a memory variable M via a memory map mem. This is shown in Figure 1. Now, computational mechanics assumes so-called *predictive*

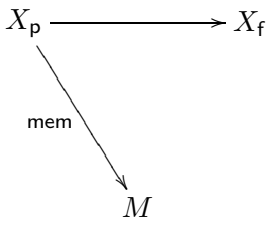

Figure 1: Memory map mem that compresses the information contained in X_p about X_f

memories (termed *prescient* in [Shalizi and Crutchfield, 2001]), which are intended to capture the information of the past that is necessary for predicting the future. This is formalized by the requirement

$$I(M : X_\mathsf{f}) \;=\; I(X_\mathsf{p} : X_\mathsf{f}), \tag{1}$$

where I denotes the mutual information between two variables. It ensures that given the memory state, one knows everything about the future that can possibly be known based on the past. The mutual information $I(X_\mathsf{p} : X_\mathsf{f})$ between past and future is related to a complexity measure which is known as *effective measure complexity*, *excess entropy*, and *predictive information* [Grassberger, 1986, Shalizi and Crutchfield, 2001, Bialek et al., 2001]. We discuss the requirement (1) in Section 3.1 within the setting of stochastic processes. It turns out that (1) is a strong requirement which is not necessary for optimal prediction and can be relaxed in a very natural way. Therefore, we call memories that satisfy (1) *strongly predictive*.

We feel that an appropriate notion of a predictive memory has to take into account the operational nature of prediction performed by a process that generates a future $\widetilde{X}_\mathsf{f}$ with the same statistical properties as the real future X_f, given the past X_p. In Section 2.1 we specify the generating process gen in more detail. According to our view, the diagram of Figure 1 should be extended as shown in Figure 2. We call a memory predictive if it contains sufficient

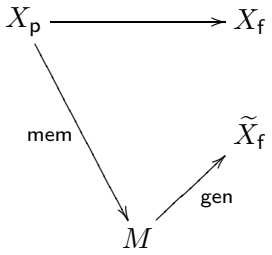

Figure 2: Memory map mem together with a generator gen, which generates $\widetilde{X}_\mathsf{f}$ as a version of X_f that is indistinguishable from X_f based on X_p

information for generating a future trajectory that is indistinguishable from the real future trajectory based on the observed past. More precisely, we assume for all past trajectories ("histories") x_p with positive probability and all future trajectories x_f the following equality:

$$\mathbb{P}(\widetilde{X}_\mathsf{f} = x_\mathsf{f} \mid X_\mathsf{p} = x_\mathsf{p}) \;=\; \mathbb{P}(X_\mathsf{f} = x_\mathsf{f} \mid X_\mathsf{p} = x_\mathsf{p}). \tag{2}$$

The corresponding definitions in the setting of stochastic processes are given in Section 2.2.

Restricted to the situation where the map mem is a deterministic function of X_{p}, both notions of predictive memory turn out to be equivalent (Proposition 3.4). We see the difference only in the case where the memory mem is allowed to be a stochastic map, which corresponds to the extension made in [Still and Crutchfield, 2007]. Within this extension the ε-machine turns out to be minimal (Proposition 3.5). On the other hand, in order to be effective in the sense of allowing for more concise models than the ε-machine, this extension has to be combined with our notion of predictive memory. We make this statement more precise in Proposition 3.5. In general, every strongly predictive memory mem, which is assumed to satisfy (1), is in particular predictive but not vice versa (see the example at the end of this introduction, and Proposition 3.4 within the general setting). Our relaxed notion of predictive memories extends the set of models that can be used for prediction and thereby provides the possibility of more concise models for stochastic processes. Example 3.6 provides a minimal hidden Markov model that is contained in the extended class but not in the class of strongly predictive models, which is usually considered in computational mechanics. This model with two hidden states can be obtained by the OOM construction method, whereas the ε-machine requires infinitely many hidden (causal) states.

We close this introduction by giving the following instructive example of a memory that is predictive but not strongly predictive, and which is extended to the setting of stochastic processes in Example 3.3: We assume that all variables have values 0 or 1, that is $x_{\mathsf{p}}, x_{\mathsf{f}}, m \in \{0,1\}$,

$$\mathbb{P}(X_{\mathsf{p}} = x_{\mathsf{p}}) \;=\; \tfrac{1}{2} \qquad \text{and} \qquad \mathbb{P}(X_{\mathsf{f}} = x_{\mathsf{f}} \mid X_{\mathsf{p}} = x_{\mathsf{p}}) \;=\; \begin{cases} \frac{3}{4}, & \text{if } x_{\mathsf{f}} = x_{\mathsf{p}} \\ \frac{1}{4}, & \text{if } x_{\mathsf{f}} \neq x_{\mathsf{p}} \end{cases}.$$

We define the memory map to be the kernel from X_{p} to X_{f}, i.e.

$$\mathsf{mem}(x_{\mathsf{p}};\, m) \;:=\; \mathbb{P}(X_{\mathsf{f}} = m \mid X_{\mathsf{p}} = x_{\mathsf{p}}).$$

Obviously, this memory is predictive in the sense of our definition (2). As corresponding generator gen we simply choose the identity map which copies the state of the memory into the future variable. On the other hand, this memory mem does not satisfy (1) and is therefore not (strongly) predictive in the usual sense of computational mechanics. This can be seen as follows: We have the joint probabilities

$$\mathbb{P}(M = m,\; X_{\mathsf{f}} = x_{\mathsf{f}}) \;=\; \begin{cases} \frac{5}{16}, & \text{if } x_{\mathsf{f}} = m \\ \frac{3}{16}, & \text{if } x_{\mathsf{f}} \neq m \end{cases} \qquad \text{and} \qquad \mathbb{P}(X_{\mathsf{p}} = x_{\mathsf{p}},\; X_{\mathsf{f}} = x_{\mathsf{f}}) \;=\; \begin{cases} \frac{3}{8}, & \text{if } x_{\mathsf{p}} = x_{\mathsf{f}} \\ \frac{1}{8}, & \text{if } x_{\mathsf{p}} \neq x_{\mathsf{f}} \end{cases}$$

and the corresponding mutual informations

$$I(M : X_{\mathsf{f}}) \;=\; \tfrac{5}{8}\ln(\tfrac{5}{4}) + \tfrac{3}{8}\ln(\tfrac{3}{4}) \;<\; \tfrac{3}{4}\ln(\tfrac{3}{2}) + \tfrac{1}{4}\ln(\tfrac{1}{2}) \;=\; I(X_{\mathsf{p}} : X_{\mathsf{f}})$$

which violate (1).

2 Predictive models of stochastic processes

2.1 Generating a process

Before suggesting our notion of prediction, we first consider the task of generating a process. Generating a predicted future based on memory states is a crucial part of our understanding of prediction. We assume a finite set D (*state space* of the generated process), a countable set M of *memory states* (also called *internal states*) and a Markov kernel, which we call *generator*:

$$\mathsf{gen}\colon \mathsf{M} \to \mathcal{P}(\mathsf{D} \times \mathsf{M}),$$

where $\mathcal{P}(\mathsf{D}\times\mathsf{M})$ is the set of probability measures on $\mathsf{D}\times\mathsf{M}$. We use the notation $\mathsf{gen}(m;\, x, \hat{m})$ to denote the probability of the pair $(x, \hat{m})$ with respect to $\mathsf{gen}(m)$. Together with an initial probability distribution μ on the memory states M, this kernel generates a stochastic process $\widetilde{X}_k$, $k \in \mathbb{N}$, on D and a process M_k, $k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, on M in the following way: Being in a memory state at time k, it (stochastically) produces a new memory state at time $k+1$ and, at the same time, emits a symbol from D. This is shown in Figure 3.

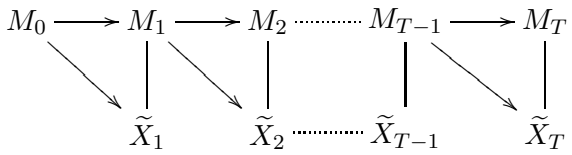

Figure 3: The process of generating memory states M_k and emitting observable states $\widetilde{X}_k$

The joint distribution is computed according to

$$\mathbb{P}(M_{[0,T]} = m_{[0,T]},\ \widetilde{X}_{[1,T]} = x_{[1,T]},) \ =\ \mu(m_0) \prod_{k=1}^{T} \mathsf{gen}(m_{k-1};\, x_k, m_k),$$

where we use the notation $[0,T]$ for the discrete interval $\{\,0,\ldots,T\,\}$ and $M_{[0,T]} = m_{[0,T]}$ for $M_0 = m_0, \ldots, M_T = m_T$. Similarly throughout the paper we also use the notation $X_{\mathbb{T}}$ to denote a stochastic process X_k, $k \in \mathbb{T}$, where $\mathbb{T}$ is the time set of the process.

Definition 2.1 (generating a process)**.** Let $X_{\mathbb{N}}$ be a stochastic process on D. We say that gen *generates* $X_{\mathbb{N}}$ if there exists an initial distribution for gen such that $\widetilde{X}_{\mathbb{N}}$ has the same distribution (the same statistics) as $X_{\mathbb{N}}$.

2.2 Our prediction setting

We use generators as models for the process of prediction. The initial distribution is computed by a memory map from past observations and contains the information of the history. To avoid measure-theoretic technicalities due to an uncountable set of infinite history trajectories, we allow only finite but varying length observations. Unfortunately, the variation leads

to some notational technicalities, especially in Section 3.1.

Throughout this article, we consider a *stationary* stochastic process $X_{\mathbb{Z}}$, the *observable process*, with *finite* state set D. Note that, since $X_{\mathbb{Z}}$ is stationary, it is uniquely determined by its restriction to positive times. For the task of prediction we assume that the outcome of $X_{\mathbb{Z}}$ is known for some finite but arbitrary past time interval $[-t+1, 0]$. Based on these observations, a generator is used as a mechanism for generating an outcome of $\widetilde{X}^t_{[1,T]}$ as prediction of the real future outcome $X_{[1,T]}$. The situation is illustrated in Figure 4 and made more precise by the following definitions.

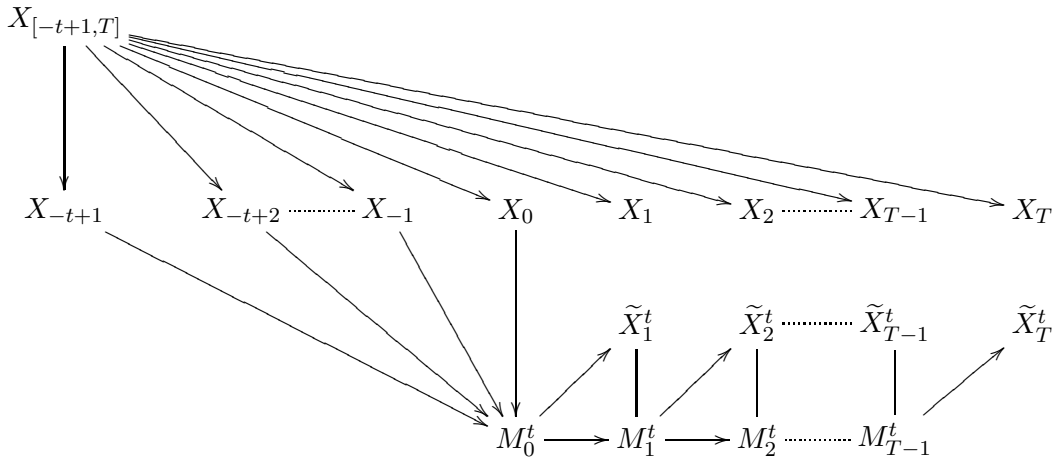

Figure 4: The process of generating $\widetilde{X}^t_{[1,T]}$ as prediction of $X_{[1,T]}$ based on the past trajectory which is fed into the memory M^t_0

Definition 2.2 (memory)**.** A *memory* (*map*) mem assigns to every history $x_{[-t+1,0]} \in \mathsf{D}^{[-t+1,0]}$ of arbitrary but finite length t a probability distribution on a countable set M of *memory states*:

$$\mathsf{mem}\colon \quad \mathsf{D}^* \;:=\; \bigcup_{t\in\mathbb{N}_0} \mathsf{D}^{[-t+1,0]} \;\to\; \mathcal{P}(\mathsf{M}).$$

Note that D^* also contains the "empty history," which corresponds to not having observed anything.

We use the memory map mem and a generator $\mathsf{gen}\colon \mathsf{M} \to \mathcal{P}(\mathsf{D}\times\mathsf{M})$ to define random variables M^t_k and $\widetilde{X}^t_k$ as shown in Figure 4. For every history length t, $X_{[-t+1,0]}$ and mem induce a random variable $M^t = M^t_0$ with distribution

$$\mathbb{P}(M^t = m \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \;=\; \mathsf{mem}(x_{[-t+1,0]};\, m).$$

Now we can start the generator gen in the memory state M^t_0 and obtain the predicted process $\widetilde{X}^t_{\mathbb{N}}$ on D as well as a process of internal states $M^t_{\mathbb{N}_0}$ on M with the joint (conditional)

distribution

$$\mathbb{P}(M^t_{[0,T]} = m_{[0,T]},\ \widetilde{X}^t_{[1,T]} = x_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \\ = \ \mathsf{mem}(x_{[-t+1,0]};\, m_0) \prod_{k=1}^{T} \mathsf{gen}(m_{k-1};\, x_k, m_k).$$

Definition 2.3 (predictive model)**.** We call the memory map mem *predictive* (w.r.t. $X_{\mathbb{Z}}$) if there exists a generator $\mathsf{gen}\colon \mathsf{M} \to \mathcal{P}(\mathsf{D} \times \mathsf{M})$, such that for all t and all $x_{[-t+1,0]}$ satisfying $\mathbb{P}(X_{[-t+1,0]} = x_{[-t+1,0]}) > 0$ the following equality holds:

$$\mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \ = \ \mathbb{P}(\widetilde{X}^t_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \qquad \text{for all } T.$$

We then call the pair $(\mathsf{mem}, \mathsf{gen})$ (*predictive*) *model* of the process $X_{\mathbb{Z}}$.

This definition of a predictive memory corresponds to the requirement (2) which we already discussed in the introduction. To summarize, if we have a predictive model and a finite interval of observations with arbitrary length t, we use the memory map to (stochastically) produce an initial value M^t for the generator. Then we apply the generator to produce a predicted future $\widetilde{X}^t_{[1,T]}$ that follows the same statistics as the "real" future $X_{[1,T]}$, conditioned on the observations $X_{[-t+1,0]}$. It is important that the generator must not depend on the length t of the history.

In the next section, we relate our notion of a predictive model to the definition used in computational mechanics. Before doing so, we give an example showing that one can always find a predictive model of a stochastic process, namely the ε-machine of computational mechanics. This important example is also used in Proposition 3.5 and Example 3.6.

Example 2.4 (ε-machine)**.** In computational mechanics, the *ε-machine* is defined on the so-called causal states. These are defined as equivalence classes of observed histories. Usually these histories are assumed to have infinite length, but finite length histories have also been considered (e.g. [Feldman and Crutchfield, 1998]). In our setting with finite observed histories, the identified histories may have different lengths. The equivalence relation identifies histories with the same conditional expectation on the future, i.e. $x_{[-t+1,0]} \sim x'_{[-s+1,0]}$ if and only if[1]

$$\mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \ = \ \mathbb{P}(X_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]}) \qquad \text{for all } T > 0.$$

The *causal state* of $x_{[-t+1,0]}$ is given by its equivalence class

$$\mathfrak{C}(x_{[-t+1,0]}) \ := \ \left\{ x'_{[-s+1,0]} \;\middle|\; s \in \mathbb{N}_0,\ x'_{[-s+1,0]} \sim x_{[-t+1,0]} \right\}.$$

[1] We assign histories with probability zero, i.e. $\mathbb{P}(X_{[-t+1,0]} = x_{[-t+1,0]}) = 0$, to arbitrary equivalence classes.

As memory set, we take the set $\mathsf{M}_{\mathfrak{C}} := \mathrm{Im}(\mathfrak{C}) := \{\, \mathfrak{C}(x_{[-t+1,0]}) \,\}$ of causal states. The memory function $\mathsf{mem}_{\mathfrak{C}}$ assigns to a history $x_{[-t+1,0]}$ the Dirac measure in the corresponding causal state $\mathfrak{C}(x_{[-t+1,0]})$, i.e. $\mathsf{mem}_{\mathfrak{C}}\big(x_{[-t+1,0]};\, \mathfrak{C}(x_{[-t+1,0]})\big) = 1$. To get a predictive model, we also need a generator. By $x_{[-t+1,0]}x$, we denote the history $y_{[-t,0]}$ of length $t+1$ obtained by appending the symbol x to $x_{[-t+1,0]}$, i.e. $y_0 = x$ and $y_{-k} = x_{-k+1}$ for $k = 1, \ldots, t$. Note that if $\mathfrak{C}(x_{[-t+1,0]}) = \mathfrak{C}(x'_{[-s+1,0]})$, we also have $\mathfrak{C}(x_{[-t+1,0]}x) = \mathfrak{C}(x'_{[-s+1,0]}x)$, provided that $x_{[-t+1,0]}$ and $x'_{[-s+1,0]}$ have positive probability. This is true because

$$\mathbb{P}(X_{[1,T]} \mid X_{[-t,0]} = x_{[-t+1,0]}x) \;=\; \frac{\mathbb{P}(X_0 = x,\; X_{[1,T]} \mid X_{[-t,-1]} = x_{[-t+1,0]})}{\mathbb{P}(X_0 = x \mid X_{[-t,-1]} = x_{[-t+1,0]})},$$

and $X_{\mathbb{Z}}$ is stationary. Therefore, the following generator (the ε-machine transition) is well defined:

$$\mathsf{gen}_{\mathfrak{C}}\big(m;\, x, m'\big) \;:=\; \begin{cases} \mathbb{P}(X_1 = x \mid X_{[-t+1,0]} = x_{[-t+1,0]}), & \text{if } \mathfrak{C}(x_{[-t+1,0]}x) = m' \\ 0, & \text{otherwise} \end{cases},$$

where $x_{[-t+1,0]}$ is any history with positive probability and $\mathfrak{C}(x_{[-t+1,0]}) = m$. We obtain

$$\begin{aligned} &\mathbb{P}(\widetilde{X}^t_{[1,T]} = x_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \\ &\quad= \prod_{k=1}^{T} \mathsf{gen}_{\mathfrak{C}}\big(\mathfrak{C}(x_{[-t+1,0]}x_1 \cdots x_{k-1});\, x_k, \mathfrak{C}(x_{[-t+1,0]}x_1 \cdots x_k)\big) \\ &\quad= \prod_{k=1}^{T} \mathbb{P}(X_1 = x_k \mid X_{[-t-k+2,0]} = x_{[-t+1,0]}x_1 \cdots x_{k-1}) \\ &\quad\overset{\text{(stationary)}}{=} \mathbb{P}(X_{[1,T]} = x_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}). \end{aligned}$$

Thus $(\mathsf{mem}_{\mathfrak{C}}, \mathsf{gen}_{\mathfrak{C}})$ is a predictive model.

As a pair $(\mathsf{mem}, \mathsf{gen})$, a predictive model (of $X_{\mathbb{Z}}$) in particular provides the generator gen, which generates the restriction $X_{\mathbb{N}}$ of the process $X_{\mathbb{Z}}$ to positive times in the sense of Definition 2.1. The appropriate initial distribution is given by the memory map for $t = 0$. In the following proposition we show the converse of this statement: Every generator which generates the positive time restriction $X_{\mathbb{N}}$ can be used in a predictive model with an appropriate memory map. In particular, if the number of memory states in M is large enough to allow for generating the positive time restriction of the process, it is also large enough to admit a predictive model of $X_{\mathbb{Z}}$.

Proposition 2.5 (generator as predictive model)**.** *Let* $\mathsf{gen}\colon \mathsf{M} \to \mathcal{P}(\mathsf{D} \times \mathsf{M})$ *be a generator that generates the positive time restriction of the process* $X_{\mathbb{Z}}$*. Then there is a memory map* $\mathsf{mem}\colon \mathsf{D}^* \to \mathsf{M}$*, such that* $(\mathsf{mem}, \mathsf{gen})$ *is a predictive model of* $X_{\mathbb{Z}}$*.*

Proof. Let the initial distribution for gen be such that $\widetilde{X}_{\mathbb{N}}$ has the same distribution as $X_{\mathbb{N}}$. Define for all $x_{[-t+1,0]}$ with positive probability

$$\mathsf{mem}(x_{[-t+1,0]};\, m) \;:=\; \mathbb{P}(M_t = m \mid \widetilde{X}_{[1,t]} = x_{[-t+1,0]}).$$

The case $t = 0$ is clear because of the fact that $\widetilde{X}^0_{\mathbb{N}}$ and $\widetilde{X}_{\mathbb{N}}$ have the same distribution. Therefore, let $t > 0$:

$$\begin{aligned}
&\mathbb{P}(\widetilde{X}^t_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \\
&\quad = \sum_m \mathbb{P}(M^t_0 = m \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \mathbb{P}(\widetilde{X}^t_{[1,T]} \mid M^t_0 = m,\ X_{[-t+1,0]} = x_{[-t+1,0]}) \\
&\quad = \sum_m \mathsf{mem}(x_{[-t+1,0]};\, m) \mathbb{P}(\widetilde{X}^t_{[1,T]} \mid M^t_0 = m) \\
&\quad = \sum_m \mathbb{P}(M_t = m \mid \widetilde{X}_{[1,t]} = x_{[-t+1,0]}) \mathbb{P}(\widetilde{X}_{[t+1,t+T]} \mid M_t = m) \\
&\quad = \mathbb{P}(\widetilde{X}_{[t+1,t+T]} \mid \widetilde{X}_{[1,t]} = x_{[-t+1,0]}) \overset{\text{(assumption)}}{=} \mathbb{P}(X_{[t+1,t+T]} \mid X_{[1,t]} = x_{[-t+1,0]}) \\
&\quad \overset{\text{(stationary)}}{=} \mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}),
\end{aligned}$$

where we used that $\widetilde{X}^t_{[1,T]}$ is independent of $X_{[-t+1,0]}$ given M^t_0 and $\widetilde{X}_{[t+1,t+T]}$ is independent of $\widetilde{X}_{[1,t]}$ given M_t. □

3 What does "predictive" really mean?

3.1 Predictive versus strongly predictive memories

As we already mentioned in the introduction, our concept of a predictive memory map differs from the concept usually discussed within computational mechanics. There, one tries to compress the observed sequence $x_{[-t+1,0]}$ by the memory map and requires that, at the same time, no information about the future $x_{[1,T]}$ (for all T) that is contained in the history $x_{[-t+1,0]}$ is lost. In the situation where all observed histories have the same length, which in computational mechanics is usually assumed to be infinite, this means requiring that the mutual information between history and future is equal to the mutual information between the memory M and the future, similar to the requirement (1) of the introduction. In our present setting of finite varying history lengths, however, we do not have a single memory state at time zero, but for any history length t a different memory state M^t. Simply assuming the information equality for every history length t separately, i.e.

$$I(M^t : X_{[1,T]}) \;=\; I(X_{[-t+1,0]} : X_{[1,T]}) \qquad \text{for all } T \text{ and all } t, \tag{3}$$

is a weak requirement which does not provide the correct correspondence to (1) in the context of computational mechanics for finite but varying observation lengths. For memory maps satisfying (3), the information about the future need not be contained in the memory state alone, but also in the particular observation length t. The same memory state m can have a completely different implication on the future if it results from different history lengths (see Example 3.1). Therefore, we have to assume that the memory keeps all information about the future without the additional knowledge of t. We give two equivalent versions of the right correspondence to (1).

First, we simply assume, in addition to (3), that conditional probabilities of the future given a memory state do not depend on the observation length t. More precisely, given $m \in \mathsf{M}$, we assume that $\mathbb{P}(X_{[1,T]} \mid M^t = m)$ is independent of t, whenever $\mathbb{P}(M^t = m) > 0$. Since (3) is equivalent to

$$\mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \;=\; \mathbb{P}(X_{[1,T]} \mid M^t = m)$$

whenever $\mathbb{P}(X_{[-t+1,0]} = x_{[-t+1,0]},\; M^t = m) > 0$, we finally get the following condition as correspondence to (1):

$$\begin{aligned} &\mathbb{P}(X_{[-t+1,0]} = x_{[-t+1,0]},\; M^t = m) \;>\; 0, \quad \mathbb{P}(M^s = m) \;>\; 0 \\ &\qquad\Longrightarrow \qquad \mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \;=\; \mathbb{P}(X_{[1,T]} \mid M^s = m) \quad \text{for all } T. \end{aligned} \tag{4}$$

As a second definition, which is equivalent to (4), we provide a t-independent version of (3). To this end, we imagine that t is determined randomly by a $\mathbb{N}_0$-valued random variable τ which is assumed to be independent of all other variables. We call such a variable τ *random time variable*. Combining the family of memory variables M^t, $t \in \mathbb{N}_0$, with a random time variable τ we get a new variable M^τ which is equal to M^t precisely when $\tau = t$. We require that, for all random time variables τ, the corresponding M^τ contains maximal information about the future, even if we don't know the value of τ. More precisely,

$$I(M^\tau : X_{[1,T]}) \;=\; I(X_{[-\tau+1,0]} : X_{[1,T]}) \quad \text{for all } T \in \mathbb{N} \text{ and all random time variables } \tau \tag{5}$$

Note that (5) contains (3) as the special case of constant random times. It is straightforward to show (but omitted here) that (5) is equivalent to (4). We illustrate the difference between (3) and (5) by the following example.

Example 3.1 (the difference via random times)**.** Let $X_{\mathbb{Z}}$ be a non-i.i.d. Markov process on $\mathsf{D} := \{\, 0, 1 \,\}$. Define

$$M^t \;:=\; X_0 \qquad \text{and} \qquad \widehat{M}^t \;:=\; \begin{cases} X_0, & \text{if } t \text{ odd} \\ 1 - X_0, & \text{if } t \text{ even} \end{cases}.$$

Then both M and $\widehat{M}$ satisfy (3), while M also satisfies (5) and $\widehat{M}$ does not. This is because the information $\widehat{M}^\tau = m$ is useless if we don't know whether τ is odd or even.

Definition 3.2 (strongly predictive)**.** We call a memory *strongly predictive*, if it satisfies the equivalent conditions (5) and (4).

The property "strongly predictive" only refers to the map mem and not to the operational aspects of prediction given by the generating "mechanism" of a generator gen. Therefore, it does not refer to the predicted process $\widetilde{X}^t_{[1,T]}$. We feel that an appropriate notion of a predictive model must not neglect the operational view. The following simple example shows that

predictive memory maps in the sense of Definition 2.3 are not necessarily strongly predictive and even do not, in general, satisfy the weaker condition (3). The idea of this example was already given in the introduction (Section 1.2) and is now extended to the more complicated setting of stochastic processes.

Example 3.3 (predictive but not strongly predictive)**.** Let $X_{\mathbb{Z}}$ be the Markov process on $\mathsf{D} := \{0, 1\}$ satisfying

$$\mathbb{P}(X_k = x_k) \;=\; \tfrac{1}{2} \qquad \text{and} \qquad \mathbb{P}(X_{k+1} = x_{k+1} \mid X_k = x_k) \;=\; \begin{cases} \frac{3}{4}, & \text{if } x_{k+1} = x_k \\ \frac{1}{4}, & \text{if } x_{k+1} \neq x_k \end{cases}.$$

We define a generator on $\mathsf{M} := \mathsf{D}$ which emits its internal state as output symbol and chooses its new internal state according to the same transition rule as the Markov process $X_{\mathbb{Z}}$, i.e.

$$\mathsf{gen}(m;\, \tilde{x}, m') \;:=\; \begin{cases} \mathbb{P}(X_1 = m' \mid X_0 = m), & \text{if } \tilde{x} = m \\ 0, & \text{if } \tilde{x} \neq m \end{cases}.$$

By choosing as initial distribution the uniform distribution on the memory states we see that gen generates $X_{\mathbb{N}}$. According to Proposition 2.5, the memory map, which maps the empty history to the uniform distribution and

$$\mathsf{mem}(x_{[-t+1,0]};\, m) \;:=\; \mathbb{P}(X_1 = m \mid X_0 = x_0)$$

for $t > 0$, is predictive. On the other hand, we can use the Markov property to reduce the calculation of informations to the situation of Section 1.2:

$$I(M^t : X_{[1,T]}) \;=\; I(M^t : X_1) \overset{\text{(Section 1.2)}}{<} I(X_0 : X_1) \;=\; I(X_{[-t+1,0]} : X_{[1,T]}).$$

Thus the memory is not strongly predictive.

In the following proposition, we show that every strongly predictive memory is predictive. Furthermore, the Example 3.3 of a predictive but not strongly predictive model requires a stochastic memory map: In the deterministic case, predictive and strongly predictive are equivalent.

Proposition 3.4 (predictive versus strongly predictive)**.**
1. *Every strongly predictive memory map is predictive.*
2. *If a memory map is deterministic and predictive, then it is also strongly predictive.*

Proof.
1. Assume w.l.o.g. that for all $m \in \mathsf{M}$ there is some t_m with $\mathbb{P}(M^{t_m} = m) > 0$ (otherwise, m may be removed from M). Let $\widehat{M}^t$ be constructed from $X_{[-t+2,1]}$ with mem, just like M^t is constructed from $X_{[-t+1,0]}$. We define the generator

$$\mathsf{gen}(m;\, x, \hat{m}) \;:=\; \mathbb{P}(\widehat{M}^{t_m+1} = \hat{m},\; X_1 = x \mid M^{t_m} = m).$$

In view of Proposition 2.5, it suffices to show

$$\mathbb{P}(\widetilde{X}^t_{[1,T]} = x_{[1,T]}) \;=\; \mathbb{P}(X_{[1,T]} = x_{[1,T]}).$$

We show the more general equation (for m with $\mathbb{P}(M^t_0 = m) > 0$)

$$\mathbb{P}(\widetilde{X}^t_{[1,T]} = x_{[1,T]} \mid M^t_0 = m) \;=\; \mathbb{P}(X_{[1,T]} = x_{[1,T]} \mid M^t_0 = m)$$

by induction over T. The case $T = 0$ is trivial. For $T > 0$:

$$\begin{aligned}
&\mathbb{P}(\widetilde{X}^t_{[1,T]} = x_{[1,T]} \mid M^t_0 = m) \\
&\quad= \sum_{\hat{m}} \mathbb{P}(M^t_1 = \hat{m},\ \widetilde{X}^t_1 = x_1 \mid M^t_0 = m)\mathbb{P}(\widetilde{X}^t_{[2,T]} = x_{[2,T]} \mid M^t_1 = \hat{m}) \\
&\quad= \sum_{\hat{m}} \mathsf{gen}(m;\, x_1, \hat{m})\mathbb{P}(\widetilde{X}^{t_m+1}_{[1,T-1]} = x_{[2,T]} \mid M^{t_m+1}_0 = \hat{m}) \\
&\quad\overset{\text{(ind. as.)}}{=} \sum_{\hat{m}} \mathbb{P}(\widehat{M}^{t_m+1} = \hat{m},\ X_1 = x_1 \mid M^{t_m}_0 = m)\mathbb{P}(X_{[1,T-1]} = x_{[2,T]} \mid M^{t_m+1}_0 = \hat{m}). \quad (6)
\end{aligned}$$

Now using stationarity of $X_{\mathbb{Z}}$ and (4), which holds also for $\widehat{M}$ instead of M due to stationarity, we obtain for those $\hat{m}$ with $\mathbb{P}(\widehat{M}^{t_m+1} = \hat{m},\ X_1 = x_1 \mid M^{t_m}_0 = m) > 0$ that

$$\begin{aligned}
\mathbb{P}(X_{[1,T-1]} = x_{[2,T]} \mid M^{t_m+1} = \hat{m}) \;&=\; \mathbb{P}(X_{[2,T]} = x_{[2,T]} \mid \widehat{M}^{t_m+1} = \hat{m}) \qquad (7)\\
&=\; \mathbb{P}(X_{[2,T]} = x_{[2,T]} \mid \widehat{M}^{t_m+1} = \hat{m},\ M^{t_m}_0 = m,\ X_1 = x_1).
\end{aligned}$$

In total we obtain the required equality:

$$\mathbb{P}(\widetilde{X}^t_{[1,T]} \mid M^t_0 = m) \overset{(6)+(7)}{=} \mathbb{P}(X_{[1,T]} \mid M^{t_m}_0 = m) \overset{(4)}{=} \mathbb{P}(X_{[1,T]} \mid M^t_0 = m).$$

2. Let $M^t = f_t(X_{[-t+1,0]})$. For $m = f(x_{[-t+1,0]}) = f(x'_{[-s+1,0]})$, we get from predictiveness

$$\begin{aligned}
\mathbb{P}(X_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]}) \;&=\; \mathbb{P}(\widetilde{X}^s_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]}) \\
&=\; \mathbb{P}(\widetilde{X}^t_{[1,T]} \mid M^t = m) \\
&=\; \mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]})\ .
\end{aligned}$$

Consequently, if $A := f_s^{-1}(m)$ is the set of length-s histories mapped to m,

$$\begin{aligned}
\mathbb{P}(X_{[1,T]} \mid M^s = m) \;&=\; \frac{\sum_{x'_{[-s+1,0]} \in A} \mathbb{P}(X_{[-s+1,0]} = x'_{[-s+1,0]})\mathbb{P}(X_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]})}{\sum_{x'_{[-s+1,0]} \in A} \mathbb{P}(X_{[-s+1,0]} = x'_{[-s+1,0]})} \\
&=\; \frac{\mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \sum \mathbb{P}(X_{[-s+1,0]} = x'_{[-s+1,0]})}{\sum \mathbb{P}(X_{[-s+1,0]} = x'_{[-s+1,0]})} \\
&=\; \mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}).
\end{aligned}$$

This is nothing but equation (4). □

3.2 Implications on minimality of predictive models

Figure 5 illustrates the situation in view of Proposition 3.4 with the abbreviations "DM = deterministic memory," "SPM = strongly predictive memory," and "PM = predictive memory." In computational mechanics, strongly predictive deterministic memories have been studied,

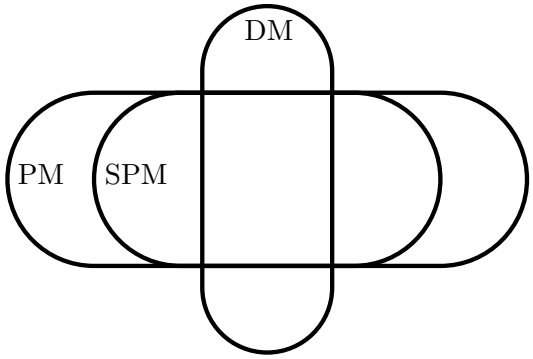

Figure 5: The extension of the memory class suggested by predictive models

that is the intersection of DM and SPM. An extension of this intersection to larger classes of memory maps is natural. According to Proposition 3.4, we have the following hierarchy of possible extensions:

$$\mathrm{DM} \cap \mathrm{SPM} \;=\; \mathrm{DM} \cap \mathrm{PM} \;\subsetneq\; \mathrm{SPM} \;\subsetneq\; \mathrm{PM}. \tag{8}$$

According to the equality in (8), considering predictive memory maps without dropping the determinism requirement does not enlarge the class. Only recently, an extension to the class SPM including also non-deterministic memories has been considered by Still and Crutchfield [Still and Crutchfield, 2007]. It turns out that this extension does not allow for "smaller" models than already captured by deterministic memory maps, as we show in the following proposition. Therefore, we suggest to further extend the class from SPM to PM and show in Example 3.6 that this extension is indeed effective.

Proposition 3.5 (ε-machine minimality in SPM)**.** *The causal state projection* $\mathfrak{C}$ *of Example 2.4 defines a strongly predictive deterministic memory map to* $\mathsf{M}_{\mathfrak{C}}$*. Further, it has minimal number of memory states in the class* SPM*. More precisely:*

$$\mathsf{mem}\colon \mathsf{D}^* \to \mathsf{M} \text{ } \textit{strongly predictive} \qquad \Rightarrow \qquad |\mathsf{M}| \;\geq\; |\mathsf{M}_{\mathfrak{C}}|.$$

Proof. We show (4). Let $m_{\mathfrak{C}} := \mathfrak{C}(x_{[-t+1,0]})$ and s be such that $\mathbb{P}\big(\mathfrak{C}(X_{[-s+1,0]}) = m_{\mathfrak{C}}\big) > 0$. The conditional probability $\mathbb{P}\big(X_{[1,T]} \;\big|\; \mathfrak{C}(X_{[-s+1,0]}) = m_{\mathfrak{C}}\big)$ is a convex combination of $\mathbb{P}(X_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]})$ with $x'_{[-s+1,0]} \in m_{\mathfrak{C}}$. Since all elements of $m_{\mathfrak{C}}$ induce the same conditional probability of $X_{[1,T]}$, we obtain

$$\mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \;=\; \mathbb{P}\big(X_{[1,T]} \;\big|\; \mathfrak{C}(X_{[-s+1,0]}) = m_{\mathfrak{C}}\big)$$

and $\mathsf{mem}_{\mathfrak{C}}$ is strongly predictive. Now assume that mem is another strongly predictive memory. We show that if the supports of $\mathsf{mem}(x_{[-t+1,0]})$ and $\mathsf{mem}(x'_{[-s+1,0]})$ are not disjoint, then

$\mathfrak{C}(x_{[-t+1,0]}) = \mathfrak{C}(x'_{[-s+1,0]})$. In particular, $|\mathsf{M}_{\mathfrak{C}}| \leq |\mathsf{M}|$. Thus, assume some $m \in \mathsf{M}$ with

$$\mathbb{P}(M^t = m \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \ > \ 0 \qquad \text{and} \qquad \mathbb{P}(M^s = m \mid X_{[-s+1,0]} = x'_{[-s+1,0]}) \ > \ 0.$$

From (4), we obtain

$$\mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \ = \ \mathbb{P}(X_{[1,T]} \mid M^s = m) \ = \ \mathbb{P}(X_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]}),$$

hence $\mathfrak{C}(x_{[-t+1,0]}) = \mathfrak{C}(x'_{[-s+1,0]})$, which finishes the proof. □

Example 3.6 below illustrates that our extension allows for minimal memory in PM not captured within SPM. It is an example of a process that admits a memory from the class PM with two memory states. The corresponding minimal number of memory states within SPM, which, according to Proposition 3.5, is realized by the ε-machine, turns out to be infinite. Moreover, the causal states are singletons, so that the causal state projection achieves no compression. On the other hand, the OOM construction method [Jaeger, 2000] yields the two internal states of the hidden Markov model.

Example 3.6 (smaller than ε-machine in PM)**.** In order to specify this example, we consider a generator gen together with an initial distribution on the set M of memory states. This defines stochastic processes $\widetilde{X}_{\mathbb{N}}$ and $M_{\mathbb{N}_0}$. If the joint process $(\widetilde{X}_{\mathbb{N}}, M_{\mathbb{N}})$ is stationary, we can extend this joint process to a stationary process with time set $\mathbb{Z}$ in a unique way. We denote the resulting processes on D and M with $X_{\mathbb{Z}}$ and $S_{\mathbb{Z}}$, respectively, and interpret them as the observable process and a process of internal states. In this concrete example, we take $\mathsf{D} := \mathsf{M} := \{\, 0, 1 \,\}$ as state space for both processes and the uniform distribution on M as initial distribution. With a parameter p, $0 < p < \frac{1}{4}$, we define the generator by

$$\mathsf{gen}(m;\, x, \hat{m}) \ := \ \begin{cases} 1-2p, & \text{if } \hat{m} = x = m \\ p, & \text{if } x \neq m \\ 0, & \text{otherwise} \end{cases} . \tag{9}$$

See Figure 6 for an illustration of the transition graph. It is easy to check that the stationarity condition is indeed satisfied. Because of Proposition 2.5, it is clear that there is a predictive model of the process $X_{\mathbb{Z}}$ with two memory states. We now show that, nevertheless, the causal states are singletons. For this purpose, we define for any output symbol $x \in \mathsf{D}$ a function $f_x \colon [0,1] \to [0,1]$, which keeps track of the probability that the internal state is 0. Concretely,

$$f_x(y) \ := \ \frac{y\,\mathsf{gen}(0;\, x, 0) + (1-y)\,\mathsf{gen}(1;\, x, 0)}{y \sum_{m=0}^{1} \mathsf{gen}(0;\, x, m) + (1-y) \sum_{m=0}^{1} \mathsf{gen}(1;\, x, m)}.$$

We compute the conditional probability that the internal state is 0 as follows:

$$\begin{aligned} &\mathbb{P}(S_0 = 0 \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \\ &\quad = \quad \sum_{m=0}^{1} \frac{\mathbb{P}(S_{-1} = m \mid X_{[-t+1,-1]} = x_{[-t+1,-1]}) \mathbb{P}(S_0 = 0,\ X_0 = x_0 \mid S_{-1} = m)}{\mathbb{P}(X_0 = x_0 \mid X_{[-t+1,-1]} = x_{[-t+1,-1]})} \\ &\quad = \quad f_{x_0}\big(\mathbb{P}(S_{-1} = 0 \mid X_{[-t+1,-1]} = x_{[-t+1,-1]})\big) \\ &\quad \overset{\text{(induction)}}{=} \quad f_{x_0} \circ \cdots \circ f_{x_{-t+1}}\big(\mathbb{P}(S_{-t} = 0)\big) \ = \ f_{x_0} \circ \cdots \circ f_{x_{-t+1}}(\tfrac{1}{2}) \end{aligned}$$

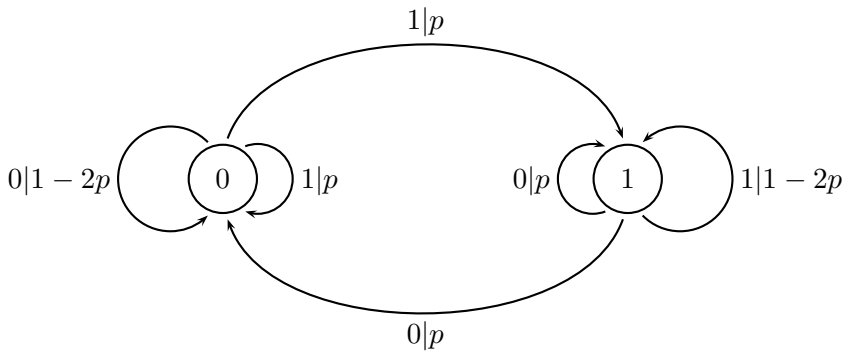

Figure 6: Transition graph of the generator defined by (9). The circled nodes are internal states. The edges are transitions, labeled with output symbol x and transition probability $\hat{p}$ in the form "$x|\hat{p}$".

Obviously, $\mathbb{P}(X_{[1,T]} \mid S_0 = 0) \neq \mathbb{P}(X_{[1,T]} \mid S_0 = 1)$ (as $p \neq \frac{1}{4}$), and therefore

$$\begin{aligned} \mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \;&=\; \mathbb{P}(X_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]}) \\ &\Leftrightarrow \qquad f_{x_0} \circ \cdots \circ f_{x_{-t+1}}(\tfrac{1}{2}) \;=\; f_{x'_0} \circ \cdots \circ f_{x'_{-s+1}}(\tfrac{1}{2}). \end{aligned} \tag{10}$$

Now plugging the definition of gen into the definition of f_x we obtain

$$f_0(y) \;=\; \frac{y(1-3p)+p}{y(1-4p)+2p} \qquad \text{and} \qquad f_1(y) \;=\; \frac{yp}{1-2p-y(1-4p)}.$$

We observe that both f_0 and f_1 are strictly increasing,

$$f_0(]0,1[) \;=]\tfrac{1}{2},1[\qquad \text{and} \qquad f_1(]0,1[) \;=]0,\tfrac{1}{2}[.$$

This implies that $f_{x_0} \circ \cdots \circ f_{x_{-t+1}}(\frac{1}{2})$ and $f_{x'_0} \circ \cdots \circ f_{x'_{-s+1}}(\frac{1}{2})$ are different for distinct $x_{[-t+1,0]}$ and $x'_{[-s+1,0]}$. Because of (10), the causal states are singletons.

4 Conclusions

We do believe that the ε-machine construction is distinguished by operational constraints that can naturally be imposed on models of prediction. Currently, computational mechanics does not specify these constraints with sufficient precision and justification. Revealing them requires a better understanding of the assumptions that are usually made. To this end, in this paper we question the notion of predictive models by exploring the generative nature of prediction. We propose a natural notion less restrictive than the one currently applied in computational mechanics and demonstrate how this provides more consistency with related approaches such as Jaeger's theory of observable operator models. We regard our contribution as a step towards specifying the natural assumptions for models of prediction. Further steps are subject of our current research and will be published elsewhere.

5 Acknowledgements

The authors are grateful for discussions with Jim Crutchfield, Cosma Shalizi, and Susanne Still. Nihat Ay thanks the Santa Fe Institute for hosting him during the initial work on this paper.

References

[Ay and Crutchfield, 2005] Ay, N. and Crutchfield, J. P. (2005). Reductions of hidden information sources. *Journal of Statistical Physics*, 120:659–684.

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

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

[Feldman and Crutchfield, 1998] Feldman, D. P. and Crutchfield, J. P. (1998). Discovering noncritical organization: Statistical mechanical, information theoretic, and computational views of patterns in one-dimensional spin systems. Santa Fe Institute Working Paper 98-04-026.

[Grassberger, 1986] Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.*, 25:907–938.

[Heller, 1965] Heller, A. (1965). On stochastic processes derived from markov chains. *Annals of Mathematical Statistics*, 36:1286–1291.

[Jaeger, 2000] Jaeger, H. (2000). Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398.

[Shalizi and Crutchfield, 2001] Shalizi, C. R. and Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879.

[Still and Crutchfield, 2007] Still, S. and Crutchfield, J. P. (2007). Optimal causal inference. informal publication, `http://arxiv.org/abs/0708.1580`.

[Tishby et al., 1999] Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, 358-377.