# Max-Planck-Institut
## für Mathematik
## in den Naturwissenschaften
## Leipzig


**On the Generative Nature of Prediction**

(revised version: September 2008)

by


*Wolfgang Löhr, and Nihat Ay*



Preprint no.: 8 2008


# On the Generative Nature of Prediction

Wolfgang Löhr[1] & Nihat Ay[1,2]

August 14, 2008

**Abstract**

Given an observed stochastic process, computational mechanics provides an explicit and efficient method of constructing a minimal hidden Markov model within the class of maximally predictive models. Here, the corresponding so-called $\varepsilon$-machine encodes the mechanisms of prediction. We propose an alternative notion of predictive models in terms of a hidden Markov model capable of generating the underlying stochastic process. A comparison of these two notions of prediction reveals that our approach is less restrictive and thereby allows for predictive models that are more concise than the $\varepsilon$-machine.

*Keywords:* hidden Markov models, computational mechanics, $\varepsilon$-machines, prediction

## Contents



[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

# 1 Introduction

## 1.1 Models of prediction and their constraints

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: Having an observed stochastic process, what is the best model in the sense of minimal size and maximal predictive power? If such a model exists, is there a way of explicitly constructing it? How much memory is needed to generate the process? In the context of these problems, the so-called $\varepsilon$-machine and its construction have been proposed as the optimal solution.

The models considered in computational mechanics are stochastic output automata, also called edge-emitting hidden Markov models (HMMs). We give a short discussion of the different types of HMMs in Section 1.3. The $\varepsilon$-machine is directly derived from the so-called causal states which are the minimal sufficient statistic on the past for predicting the future of the stochastic process. More precisely, they form the coarsest partition of the past that retains all information about the future. Furthermore, they can be obtained in a constructive and efficient way ([Shalizi et al., 2002]).

There may exist, however, HMMs with fewer internal states than the corresponding $\varepsilon$-machine. There are even examples of HMMs with few internal states that lead to an $\varepsilon$-machine with infinite size ([Crutchfield, 1994], see also Example 3.6). Consequently, $\varepsilon$-machines can only be of minimal size within reduced model classes that satisfy appropriately chosen constraints. Unfortunately, in the literature these constraints are mainly implicitly stated, which leads to a misperception and the above-mentioned apparent inconsistency. In particular, the $\varepsilon$-machine is not the minimal model that *generates* a given process, a property occasionally referred to in the literature. Instead, the $\varepsilon$-machine minimality depends on constraints that are related to particular notions of memory and prediction based on the fundamental concept of sufficient statistic. This paper reveals and relaxes these constraints in such a way that, given an HMM, a minimal predictive model of the corresponding output process can not be larger than that HMM. Recently, one step towards such an extension, namely introducing stochastic memories, has been made in [Still and Crutchfield, 2007]. While the focus of that paper lies on the trade-off between predictive power and model size (allowing some prediction error) based on the bottleneck method [Tishby et al., 1999], we show in the present contribution that this extension also is necessary for having concise models with maximal predictive power (in our weakened sense). On the other hand, it allows for smaller models only in combination with our notion of prediction.

In Section 1.2, we provide an intuitive sketch of our main idea in a simplified setting before going into the technical details of precise definitions and results within the main part of the paper. Based on our understanding of prediction as a generative operation we then briefly discuss several generative models for stochastic processes in Section 1.3. The main part, Sections 2 and 3, contains definitions and results for the case of prediction with finite histories and does not require any measure theoretic background. To complete the discussion, in the appendix we apply measure theoretic tools for showing that all results, up to minor modifications, remain true in the cases of infinite history and general state spaces.

### 1.2 The main idea of the paper

In order to illustrate the main idea of the paper, as an initial step we start with a simplified situation of a one-step prediction scenario. More precisely, we consider a pair $X_{\mathsf{p}}$ and $X_{\mathsf{f}}$ of discrete random variables, which we interpret as past and future observations. Not all information of $X_{\mathsf{p}}$ is necessary for predicting $X_{\mathsf{f}}$, so that one tries to compress the relevant information in a (discrete) memory variable $M$ via a memory map $\mathsf{mem}$ as shown in Figure 1.

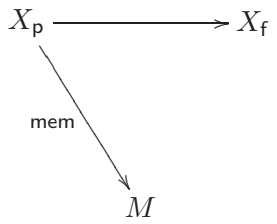


Figure 1: Memory map $\mathsf{mem}$ that compresses the information contained in $X_{\mathsf{p}}$ about $X_{\mathsf{f}}$

Within computational mechanics, mainly deterministic memory maps $\mathsf{mem}$ have been considered, where $M$ is assumed to be a *sufficient statistic*[1] on the past $X_{\mathsf{p}}$ for the future $X_{\mathsf{f}}$. Only recently an extension to stochastic maps has been considered ([Still and Crutchfield, 2007]). We adopt this extension and do not require $\mathsf{mem}$ to be a deterministic function but allow for a stochastic assignment, i.e. the memory map is assumed to be a Markov kernel (transition probability).

In general, the map $\mathsf{mem}$ reduces the information about the future, which is expressed by the following inequality:

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

where $I$ denotes the mutual information between two variables. The mutual information $I(X_{\mathsf{p}} : X_{\mathsf{f}})$ between past and future corresponds to a complexity measure defined within the context of stochastic processes and known as *effective measure complexity*, *excess entropy*, and *predictive information* [Grassberger, 1986, Shalizi and Crutchfield, 2001, Bialek et al., 2001]. We use the term predictive information for $I(X_{\mathsf{p}} : X_{\mathsf{f}})$ also within the simplified setting of this section.

In computational mechanics one assumes that $M$ is *prescient* ([Shalizi and Crutchfield, 2001]) in the sense that the equality holds in (1):

$$I(M : X_{\mathsf{f}}) \;=\; I(X_{\mathsf{p}} : X_{\mathsf{f}}). \tag{2}$$

Note that for a deterministic memory map $\mathsf{mem}$, this property reduces to the notion of sufficient statistic. Furthermore, (2) directly implies that the predictive information is a lower bound of the entropy $H(M)$ which corresponds to a complexity measure known as *statistical complexity* within computational mechanics ([Shalizi and Crutchfield, 2001]):

$$I(X_{\mathsf{p}} : X_{\mathsf{f}}) \;\leq\; H(M) \;\leq\; \ln\bigl(|\mathsf{M}|\bigr), \tag{3}$$

[1] See, e.g., [Kulhavý, 1996] for an introduction to statistics in an abstract context.

where $|\mathsf{M}|$ denotes the cardinality of the set $\mathsf{M}$ of memory states.

The requirement (2) is equivalent to

$$I(X_{\mathsf{f}} : X_{\mathsf{p}} \mid M) \;=\; 0, \tag{4}$$

which means that $M$ captures all the information of the past that is necessary for predicting the future, or, stated differently, the future is independent of the past given the memory:

$$X_{\mathsf{f}} \perp\!\!\!\perp X_{\mathsf{p}} \mid M. \tag{5}$$

This implies that the joint distribution can be represented by the graph of Figure 2.

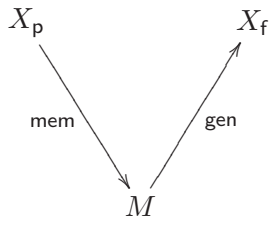


Figure 2: Factorization of the channel $X_{\mathsf{p}} \to X_{\mathsf{f}}$ resulting from a prescient memory

More precisely, there is a Markov kernel $\mathsf{gen}$, illustrated by the arrow $M \to X_{\mathsf{f}}$, that satisfies

$$\mathbb{P}(X_{\mathsf{f}} = x_{\mathsf{f}} \mid X_{\mathsf{p}} = x_{\mathsf{p}}) \;=\; \sum_{m} \mathsf{mem}(x_{\mathsf{p}};\, m)\, \mathsf{gen}(m;\, x_{\mathsf{f}}). \tag{6}$$

Interpreting $\mathsf{gen}$ in a generative way, this means that one can replace the mechanisms that underly the channel $X_{\mathsf{p}} \to X_{\mathsf{f}}$ by the two steps of

1. mapping the past to a memory state using $\mathsf{mem}$ and then, based on that state,
2. generating the future with $\mathsf{gen}$.

It is important to point out that the replacement of the channel $X_{\mathsf{p}} \to X_{\mathsf{f}}$ by the composition of $\mathsf{mem}$ and $\mathsf{gen}$ does not mean that both mechanisms have the same outcome. Due to the intrinsic stochasticity of the kernels, one can not expect that the predicted outcome coincides with the real outcome. Only the distribution is the same. In order to explicitly model this outcome difference, we use the symbol $\widetilde{X}_{\mathsf{f}}$ to denote the prediction variable, given by the mechanisms $\mathsf{mem}$ and $\mathsf{gen}$, and combine the graphs of the Figures 1 and 2 into the graph of Figure 3. In this picture, the condition (6) translates to the following condition:

$$\mathbb{P}(X_{\mathsf{f}} = x_{\mathsf{f}} \mid X_{\mathsf{p}} = x_{\mathsf{p}}) \;=\; \mathbb{P}(\widetilde{X}_{\mathsf{f}} = x_{\mathsf{f}} \mid X_{\mathsf{p}} = x_{\mathsf{p}}). \tag{7}$$

Here, the existence of a kernel $\mathsf{gen}$ that satisfies (7) is a consequence of the assumption that the memory is prescient. On the other hand, one could try to find a kernel $\mathsf{mem}$, not necessarily prescient, for which there exists a kernel $\mathsf{gen}$ such that (7) is satisfied. The requirement (7) simply means that the memory $\mathsf{mem}$ contains sufficient information for generating a future trajectory with $\mathsf{gen}$ that is indistinguishable from the real future trajectory based on the observed past. We call such a memory *predictive*. This is the ansatz of the present paper,

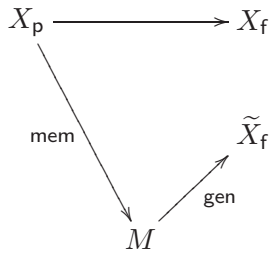


Figure 3: Memory map mem together with a generator gen which generates $\widetilde{X}_{\mathsf{f}}$ as a version of $X_{\mathsf{f}}$ that is indistinguishable from $X_{\mathsf{f}}$ based on $X_{\mathsf{p}}$

which relaxes the notion of a prescient memory. At the end of this section, we give an example of a predictive memory that is not prescient.

With inequality (1) and with the assumption (7) one has the following relations to the predictive information, that is the mutual information between past and future:

$$I(M : X_{\mathsf{f}}) \;\leq\; I(X_{\mathsf{p}} : X_{\mathsf{f}}) \;=\; I(X_{\mathsf{p}} : \widetilde{X}_{\mathsf{f}}) \;\leq\; I(M : \widetilde{X}_{\mathsf{f}}). \tag{8}$$

More precisely, only the equality in (8) requires predictability of the memory. According to the second inequality, a memory can contain more information about the predicted future $\widetilde{X}_{\mathsf{f}}$ than the past $X_{\mathsf{p}}$, whereas, according to the first inequality, no memory can contain more information about the actual future than the past. In view of the usual reference to the data processing inequality within the computational mechanics literature, the second inequality appears unfamiliar and emphasizes a new aspect of prediction. On the other hand, it is simply a direct consequence of the explicit distinction between the actual and predicted future, that is $X_{\mathsf{f}}$ and $\widetilde{X}_{\mathsf{f}}$. Furthermore, as a direct implication of (8) the inequality (3), which holds for prescient memories, remains valid also for predictive memories. Stated explicitly, the entropy of a predictive memory variable $M$ is lower bounded by the predictive information of the observed process. We will see that, as consequence of our extension of the memory class to predictive memories, it is possible to come closer to this lower bound and thereby better exploit the predictive structure of the underlying process.

Example 1.1 shows the difference between prescient and predictive memories within the simplified setting of this introduction. As we already mentioned, all definitions of this section will be extended to the general setting of stochastic processes. The key idea of our main Example 3.6 within that general setting is illustrated by the following construction:

**Example 1.1.** Let $X_{\mathsf{p}}$ and $X_{\mathsf{f}}$ assume values in $\{\,0, 1, 2\,\}$,

$$\mathbb{P}(X_{\mathsf{p}} = x_{\mathsf{p}}) \;=\; \tfrac{1}{3} \qquad \text{and} \qquad \mathbb{P}(X_{\mathsf{f}} = x_{\mathsf{f}} \mid X_{\mathsf{p}} = x_{\mathsf{p}}) \;=\; \begin{cases} \frac{2}{3}, & \text{if } x_{\mathsf{f}} = x_{\mathsf{p}} \neq 2 \\ \frac{1}{3}, & \text{if } x_{\mathsf{p}} = 2 \text{ or } x_{\mathsf{f}} = 2\,. \\ 0, & \text{otherwise} \end{cases}$$

As predictive information we get

$$I(X_{\mathsf{p}} : X_{\mathsf{f}}) \;=\; \tfrac{4}{9}\ln(2),$$

which is a lower bound for the entropy of any prescient or predictive memory. Since all three possibilities for $X_\mathsf{p}$ yield different expectations on the future, every prescient memory variable still has to distinguish between these three states (the mathematical proof of this statement is simple and contained in the proof of Proposition 3.5). Intuitively, if we identify states $x_\mathsf{p}$ with the same conditional distribution on the $x_\mathsf{f}$ we get three so-called causal states as equivalence classes. Obviously, every prescient memory has at least three memory states with an entropy $H(M) \geq \ln(3)$. On the other hand, it is possible to define a predictive memory with only two memory states. To this end, consider a memory with state set $\{0,1\}$ that copies $x_\mathsf{p}$ if $x_\mathsf{p} \in \{0,1\}$ and choses a random state otherwise:

$$\mathbb{P}(M = m \mid X_\mathsf{p} = x_\mathsf{p}) \;=\; \mathsf{mem}(x_\mathsf{p};\, m) \;:=\; \begin{cases} 1, & \text{if } m = x_\mathsf{p} \neq 2 \\ \frac{1}{2}, & \text{if } x_\mathsf{p} = 2 \\ 0, & \text{otherwise} \end{cases} .$$

We see that $\mathsf{mem}$ is predictive by defining $\mathsf{gen}(m) = \mathbb{P}(X_\mathsf{f} \mid X_\mathsf{p} = m)$, $m \in \{0,1\}$ and observing that also for $x_\mathsf{p} = 2$ this yields the correct distribution of the prediction:

$$\begin{aligned} \mathbb{P}(\widetilde{X}_\mathsf{f} \mid X_\mathsf{p} = 2) &= \tfrac{1}{2}\mathbb{P}(\widetilde{X}_\mathsf{f} \mid M = 0) + \tfrac{1}{2}\mathbb{P}(\widetilde{X}_\mathsf{f} \mid M = 1) \\ &= \tfrac{1}{2}\mathbb{P}(X_\mathsf{f} \mid X_\mathsf{p} = 0) + \tfrac{1}{2}\mathbb{P}(X_\mathsf{f} \mid X_\mathsf{p} = 1) \;=\; \mathbb{P}(X_\mathsf{f} \mid X_\mathsf{p} = 2). \end{aligned}$$

Furthermore, $H(M) = \ln(2)$. ◊

Within the general context of stochastic processes, our main results are the following: Every prescient memory map $\mathsf{mem}$, which is assumed to satisfy (2), is also predictive but not vice versa (Proposition 3.4 and Example 3.6). On the other hand, restricted to the situation where the map $\mathsf{mem}$ is deterministic, i.e. $M$ is a statistic on $X_\mathsf{p}$, predictive and prescient turn out to be equivalent (Proposition 3.4). The $\varepsilon$-machine turns out to be minimal in the class of *prescient* memories, even if we allow stochastic memory maps (Proposition 3.5). Every HMM, however, can be considered as a predictive model with the same number of memory states (Proposition 2.5), which provides a way of constructing machines that are more concise than the $\varepsilon$-machine. Example 3.6 is such a (minimal, edge-emitting) HMM with two hidden states whereas the corresponding $\varepsilon$-machine requires infinitely many hidden states.

Although extending the model class from prescient to predictive allows for substantially smaller models, while preserving a notion of predictive power which we consider quite natural, we have to point out two drawbacks. Firstly, constructing a minimal HMM is intrinsically difficult, whereas efficient algorithms are available for the construction of the $\varepsilon$-machine. Secondly and conceptually more important, the property of sufficiency is lost. This means that the memory state is not a complete substitute for the past. Within computational mechanics, the total information that is required for encoding the future distribution is completely contained in the memory state whereas in our model the memory only contains that part of it that is sufficient for generating the future distribution based on the past. When we want to use our model for sampling the future distribution given an $x_\mathsf{p}$, we first choose a memory state $m$ according to $\mathsf{mem}(x_\mathsf{p})$ and then apply the process $\mathsf{gen}(m)$ which generates an $x_\mathsf{f}$. If we repeat this sampling procedure we obtain the correct future distribution. On the other hand, if we "forget" the history state $x_\mathsf{p}$ and, instead of sampling new $m$'s according to $\mathsf{mem}(x_\mathsf{p})$, apply $\mathsf{gen}(m)$ using the same $m$, the resulting distribution of $x_\mathsf{f}$ will be different from $\mathbb{P}(X_\mathsf{f} \mid X_\mathsf{p} = x_\mathsf{p})$. Thus, we have to memorize the *distribution* (the information state) of

the memory states $m$ given $x_{\mathsf{p}}$. One can show that the number of these information states is lower bounded by the number of causal states.

### 1.3 Generative models: Stochastic automata, HMMs, and OOMs

In computational mechanics the $\varepsilon$-machine is defined in terms of a *stochastic (output) automaton*.[2] This name is directly linked to the intuition of a "machine": It has internal states $\mathsf{M}$ and is initialized by one of these states according to some initial probability distribution. At each time step $t$ the internal state $M_t$ is updated, that is $M_{t+1}$ is generated and, at the same time, an output symbol $X_{t+1}$ from a finite alphabet $\mathsf{D}$ is emitted. This is modeled by a joint transition probability from the internal states to output and internal states. Stochastic automata are also widely known as *edge-emitting hidden Markov models*. Here "edge-emitting" means that the output symbol may depend on both the old and the new internal state (output symbol and new internal state are determined by a joint kernel). In transition graph representations of edge-emitting HMMs, the output symbols appear as edge labels.

Probably more common than edge-emitting HMMs are the more restrictive *state-emitting HMMs*. These HMMs have to satisfy the additional condition that the output symbol $X_{t+1}$ depends on either $M_t$ or $M_{t+1}$ but not on both (as in the case of edge-emitting HMMs). More precisely, it is assumed that the transition probability factorizes as follows:

$$\mathbb{P}(M_{t+1},\, X_{t+1} \mid M_t) \;=\; \mathbb{P}(M_{t+1} \mid M_t)\, \mathbb{P}(X_{t+1} \mid M_s), \qquad s \in \{\, t, t+1 \,\}.$$

Even more restrictive are *functions of Markov chains*. Here, the output symbol is a deterministic function of the internal state $M_{t+1}$. Sometimes also a notion less restrictive than edge-emitting HMM is considered, namely that of a *partially observed Markov process*:[3] A (time homogeneous) Markov process $(M_t, X_t)_{t\in\mathbb{Z}}$ on a product space $\mathsf{M} \times \mathsf{D}$, where the component $\mathsf{D}$ is considered to be observable, whereas the other component $\mathsf{M}$ consists of hidden states. Note that here *both* marginal processes $(M_t)_{t\in\mathbb{Z}}$ and $(X_t)_{t\in\mathbb{Z}}$ need not be Markovian.

These four notions, partially observed Markov process, edge-emitting HMM (stochastic automaton), state-emitting HMM, and function of a Markov chain (ordered from more general to more restrictive) are essentially equivalent in the following sense: To every partially observed Markov process, one can naturally associate a function of a Markov chain such that the cardinality of the internal states increases only by the constant factor of the cardinality of the output alphabet. One simply takes as new set $\mathsf{M}'$ of internal states the product $\mathsf{M} \times \mathsf{D}$ of internal and observable states and the projection $m' = (m, x) \mapsto x$ onto the observable component $x$ as function determining the output symbol. In the following, we only consider edge-emitting HMMs and HMM always means edge-emitting.

There are also more algebraic models of stochastic processes, which dismiss the conception of the internal dynamics being described by a stochastic process. Instead, some vector space replaces the internal states and the "dynamics" is described by linear maps (instead of Markov kernels). These models were introduced and termed *stochastic S-modules* by Heller in the very concise and well-written paper [Heller, 1965]. Later, in [Jaeger, 2000], Jaeger made the construction more explicit and transparent for readers not familiar with module theory. He

[2] See, e.g., [Bukharaev, 1995] for an introduction to the theory of stochastic automata.

[3] The term "partially observed Markov process" sometimes refers to edge- or state-emitting HMMs.

introduced the name *observable operator model* (*OOM*), provided ways of interpreting them and extended the theory by learning algorithms. Ergodic theory for OOMs was developed in [Faigle and Schönhuth, 2007, Schönhuth and Jaeger, 2007]. In [Littman et al., 2001], the same model class was also obtained starting from a somewhat different intuition (internal states are constructed as predictions for certain tests) as linear non-controlled *predictive state representations* (*PSRs*).[4] The equivalence of linear non-controlled PSRs and OOMs is shown in [Singh et al., 2004].

OOMs are a generalization of HMMs in the sense that for any HMM there is a naturally associated, equivalent OOM with the internal states of the HMM providing a basis for its internal vector space.[5] Here the OOM state vectors roughly correspond to the so called *information states* (probability distributions over internal states) and the number of internal states of the HMM is equal to the dimension of the associated OOM. However, not every OOM is induced by an HMM (as the transitions need not be positive), and the dimension of the minimal OOM may be substantially smaller than the minimal number of HMM states. In fact, finite dimensional OOMs exist for some processes which do not allow for HMMs with finitely many internal states.

While OOMs can be constructed such that the internal state vector has some probabilistic interpretation (giving the probability of certain future events), it may be impossible to interpret any basis as "internal (pure) states": To simulate a process with an HMM, one can, at each time step, determine the next internal state randomly. Thus one has to store the internal state only (as opposed to the information state). This is not possible for OOMs in general, due to the occurrence of "negative probabilities".

The question whether there exists a (finite) HMM, as well as finding the minimal one, is intrinsically difficult. It depends on an intricate geometrical condition specified in [Heller, 1965]. Thus, although we want to keep the interpretation of a probabilistic machine and work with HMMs as generative models, OOM theory is interesting in that it provides a constructive algorithm for finding the minimal OOM. In some cases, in particular in the situation of our main Example 3.6, this minimal OOM is equivalent to an HMM and thus the minimal HMM can be obtained in a constructive way.

# 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 $\mathsf{D}$ (called *state space* or *alphabet* of the generated process), a countable set $\mathsf{M}$ of *memory states* (also called *internal states*) and a Markov kernel (transition probability) which we call *generator*:

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

[4] PSRs exist also for controlled systems, thus including actions of the observer. In principle, they can be non-linear, but most of the theory considers the linear case.

[5] The vector space consists of the signed measures on the internal states.

where $\mathcal{P}(\mathsf{D} \times \mathsf{M})$ is the set of probability measures on $\mathsf{D} \times \mathsf{M}$. The situation for more general spaces is discussed in the appendix.

**Remark.** A generator together with an initial probability distribution on the internal states is an (edge-emitting) HMM, except that we allow the set of internal states to be countable (instead of finite).

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 $\mu$ on the memory states $\mathsf{M}$, this kernel generates a stochastic process $\widetilde{X}_k$, $k \in \mathbb{N}$, on $\mathsf{D}$ and a process $M_k$, $k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, on $\mathsf{M}$ 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 $\mathsf{D}$. This is shown in Figure 4.

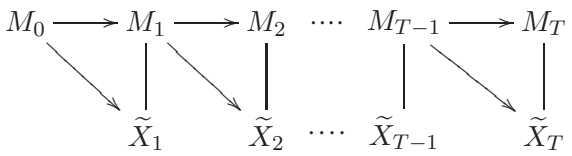


Figure 4: The process of generating memory states $M_k$ and emitting observable states $\widetilde{X}_k$

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 $\mathsf{D}$. We say that $\mathsf{gen}$ *generates* $X_{\mathbb{N}}$ if there exists an initial distribution $\mu$ for $\mathsf{gen}$, such that $\widetilde{X}_{\mathbb{N}}$ has the same distribution as $X_{\mathbb{N}}$.

**Remark.** For *every* stochastic process $X_{\mathbb{N}}$, there exists a generator $\mathsf{gen}$ which generates it. This is true because the set of internal states is allowed to be countable and the time set is only semi-infinite. Thus the generator can store the complete history of output symbols in the internal state. Another generator which can be constructed for any process is the finite-history version of the $\varepsilon$-machine (see Example 2.4).

## 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. Although computational mechanics usually works with infinite length observations (histories), we allow here only finite but varying length observations. This way we avoid measure-theoretic

technicalities due to an uncountable set of infinite history trajectories. In the appendix, we treat the case of infinite histories and show that virtually everything remains valid. Unfortunately, the variation of history length 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 $\mathsf{D}$. 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 5 and made more precise by the following definitions.

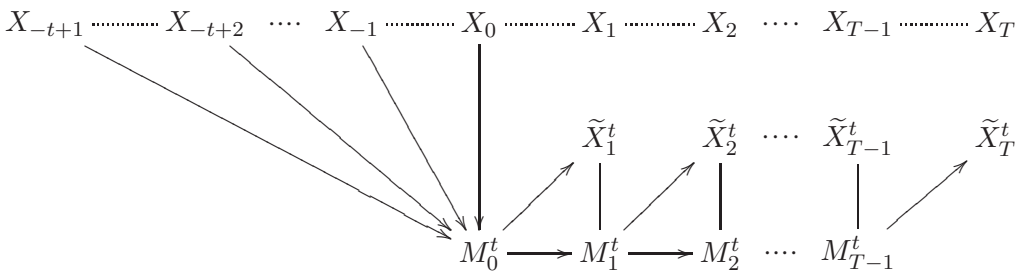


Figure 5: The process of generating $\widetilde{X}^t_{[1,T]}$ as prediction of $X_{[1,T]}$ based on a length-$t$ history which is fed into the memory variable $M^t_0$. The dotted lines symbolize that $X_{\mathbb{Z}}$ may have arbitrary dependencies and need not be Markov.

**Definition 2.2 (memory).** A *memory* (*map*) $\mathsf{mem}$ 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 $\mathsf{M}$ 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 $\mathsf{D}^*$ contains the "empty history," which corresponds to not having observed anything.

We use the memory map $\mathsf{mem}$ 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 5. For every history length $t$, $X_{[-t+1,0]}$ and $\mathsf{mem}$ 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 $\mathsf{gen}$ in the memory state $M^t_0$ and obtain the predicted process $\widetilde{X}^t_{\mathbb{N}}$ on $\mathsf{D}$ as well as a process of internal states $M^t_{\mathbb{N}_0}$ on $\mathsf{M}$ 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 $\mathsf{mem}$ *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 (7) 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 has the same statistic properties 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 Section 3, 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 construct a predictive model of *any* stationary stochastic process, namely (the finite-history version of) the $\varepsilon$-machine of computational mechanics. This important example is also used in Proposition 3.5 and Example 3.6.

**Example 2.4 ($\varepsilon$-machine).** In computational mechanics, the *$\varepsilon$-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 this case, 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[6]

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

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

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

As memory set, we take the set $\mathsf{M}_{\mathfrak{C}} := \mathrm{Im}(\mathfrak{C}) := \big\{\, \mathfrak{C}(x_{[-t+1,0]}) \,\big\}$ of causal states. The deterministic memory map $\mathsf{mem}_{\mathfrak{C}}$ 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 concatenation of $x_{[-t+1,0]}$ and $x$, i.e. $y_0 = x$ and $y_{-k} = x_{-k+1}$ for $k = 1, \dots, t$. 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 $\varepsilon$-machine transition) is well defined:

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

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

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

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

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 $\mathsf{gen}$ 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 $t = 0$, i.e. by $\mathsf{mem}(\emptyset)$, where $\emptyset$ is the empty history. In the following proposition, we show the converse of this statement: Every generator that generates the positive time restriction $X_{\mathbb{N}}$ can be used in a predictive model with an appropriate memory map.[7] In particular, if the number of memory states in $\mathsf{M}$ 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* $\mathsf{mem}\colon \mathsf{D}^* \to \mathcal{P}(\mathsf{M})$*, such that* $(\mathsf{mem}, \mathsf{gen})$ *is a predictive model of* $X_{\mathbb{Z}}$*.*

*Proof.* Let the initial distribution for $\mathsf{gen}$ 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 $t > 0$:

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

[7] Thus, generative and predictive models are essentially equivalent. Note, however, that the memory map need not be unique.

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 $M_t$. □

**Remark.** Proposition 2.5 is the main source of examples of predictive models: We can take an arbitrary (edge-emitting) HMM with stationary output process and automatically obtain a predictive model of the output. This model uses the same memory states and generating mechanism as the original HMM. Example 3.6 below is of that type.

# 3 What does "prediction" really mean?

## 3.1 Predictive versus prescient 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 variable $M$ and the future, similar to the requirement (2) 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 $M^t$.[8] Simply assuming the information equality for every length $t$ separately, i.e.

$$I(M^t : X_{[1,T]}) \;=\; I(X_{[-t+1,0]} : X_{[1,T]}) \qquad \text{for all } T \text{ and all } t, \tag{9}$$

is a weak requirement which does not provide the correct correspondence to (2) in the context of computational mechanics for finite but varying observation lengths. For memory maps satisfying (9), 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 (2).

First, we simply assume, in addition to (9), 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 (9) 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 (2):

$$\begin{aligned} &\mathbb{P}(X_{[-t+1,0]} = x_{[-t+1,0]},\; M^t = m) \;>\; 0, \quad \mathbb{P}(M^s = m) \;>\; 0 \\ &\qquad\Longrightarrow \qquad \mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \;=\; \mathbb{P}(X_{[1,T]} \mid M^s = m) \quad \text{for all } T. \end{aligned} \tag{10}$$

[8] This technicality does not appear in the infinite-history case discussed in the appendix.

As a second definition, which is equivalent to (10), we provide a $t$-independent version of (9). To this end, we imagine that $t$ is determined randomly by a $\mathbb{N}_0$-valued random variable $\tau$ which is assumed to be independent of all other variables. We call such a variable $\tau$ *random time*. Combining the family of memory variables $M^t$, $t \in \mathbb{N}_0$, with a random time $\tau$ we get a new variable $M^\tau$ which is equal to $M^t$ precisely when $\tau = t$. We require that, for all random times $\tau$, the corresponding $M^\tau$ contains maximal information about the future, even if we don't know the value of $\tau$. More precisely,

$$I(M^\tau : X_{[1,T]}) \;=\; I(X_{[-\tau+1,0]} : X_{[1,T]}) \quad \text{for all } T \in \mathbb{N} \text{ and all random times } \tau. \tag{11}$$

Note that (11) contains (9) as the special case of constant random times. It is straightforward to show (but omitted here) that (11) is equivalent to (10). We illustrate the difference between (9) and (11) 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 (9), whereas $M$ also satisfies (11) and $\widehat{M}$ does not. This is because the information $\widehat{M}^\tau = m$ is useless if we don't know whether $\tau$ is odd or even. $\Diamond$

**Definition 3.2 (prescient).** We call a memory *prescient*, if it satisfies the equivalent conditions (11) and (10).

The following example shows that predictive memories need not be prescient and can have fewer memory states than the $\varepsilon$-machine (which turns out to be the minimal prescient memory in Proposition 3.5). It is an extension of Example 1.1 to the setting of stochastic processes.

**Example 3.3.** Let $\mathsf{D} := \{\, 0, 1, 2 \,\}$ and $X_{\mathbb{Z}}$ the Markov process defined by

$$\mathbb{P}(X_0 = x_0) \;=\; \tfrac{1}{3} \qquad \text{and} \qquad \mathbb{P}(X_{k+1} = x_{k+1} \mid X_k = x_k) \;=\; \begin{cases} \frac{2}{3}, & \text{if } x_{k+1} = x_k \neq 2 \\ \frac{1}{3}, & \text{if } x_k = 2 \text{ or } x_{k+1} = 2 \\ 0, & \text{otherwise} \end{cases} .$$

As the process is Markov and different last symbols of the history yield different expectations of the next symbol, there are three causal states as defined in Example 2.4 (the empty history induces the same expectation as the histories ending in the symbol 2). On the other hand, we obtain a predictive memory map with two memory states in the same way as in Example 1.1: Let $\mathsf{M} := \{\, 0, 1 \,\}$ and

$$\mathsf{mem}(x_{[-t+1,0]};\, m) \;:=\; \begin{cases} 1, & \text{if } t \neq 0 \text{ and } m = x_0 \neq 2 \\ \frac{1}{2}, & \text{if } t = 0 \text{ or } x_0 = 2 \\ 0, & \text{otherwise} \end{cases} .$$

Just as in the situation of Example 1.1, we see that $\mathsf{mem}$ is indeed predictive. $\Diamond$

The property "prescient" only refers to the map $\mathsf{mem}$ and not to the operational aspects of prediction given by the generating "mechanism" of a generator $\mathsf{gen}$. Therefore, it does not refer to the predicted process $\widetilde{X}^t_{[1,T]}$. In the following proposition, we show that nevertheless we can associate to any prescient memory a generative mechanism: Every prescient memory is predictive. Furthermore, in the case of deterministic memory, predictive and prescient are equivalent.

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

*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 $\mathsf{M}$). Let $\widehat{M}^t$ be constructed from $X_{[-t+2,1]}$ with $\mathsf{mem}$, 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 $T > 0$:

$$\begin{aligned}
&\mathbb{P}(\widetilde{X}^t_{[1,T]} = x_{[1,T]} \mid M^t_0 = m) \qquad (12)\\
&\quad = \quad \sum_{\hat{m}} \mathbb{P}(M^t_1 = \hat{m},\; \widetilde{X}^t_1 = x_1 \mid M^t_0 = m)\, \mathbb{P}(\widetilde{X}^t_{[2,T]} = x_{[2,T]} \mid M^t_1 = \hat{m})\\
&\quad = \quad \sum_{\hat{m}} \mathsf{gen}(m;\, x_1, \hat{m})\, \mathbb{P}(\widetilde{X}^{t_m+1}_{[1,T-1]} = x_{[2,T]} \mid M^{t_m+1}_0 = \hat{m})\\
&\quad \overset{\text{(ind. as.)}}{=} \quad \sum_{\hat{m}} \mathbb{P}(\widehat{M}^{t_m+1} = \hat{m},\; X_1 = x_1 \mid M^{t_m}_0 = m)\, \mathbb{P}(X_{[1,T-1]} = x_{[2,T]} \mid M^{t_m+1}_0 = \hat{m}).
\end{aligned}$$

Now using stationarity of $X_{\mathbb{Z}}$ and (10), 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

$$\begin{aligned}
\mathbb{P}(X_{[1,T-1]} = x_{[2,T]} \mid M^{t_m+1} = \hat{m}) \;&=\; \mathbb{P}(X_{[2,T]} = x_{[2,T]} \mid \widehat{M}^{t_m+1} = \hat{m}) \qquad (13)\\
&=\; \mathbb{P}(X_{[2,T]} = x_{[2,T]} \mid \widehat{M}^{t_m+1} = \hat{m},\; M^{t_m}_0 = m,\; X_1 = x_1).
\end{aligned}$$

In total we obtain the required equality:

$$\mathbb{P}(\widetilde{X}^t_{[1,T]} \mid M^t_0 = m) \overset{(12)+(13)}{=} \mathbb{P}(X_{[1,T]} \mid M^{t_m}_0 = m) \overset{(10)}{=} \mathbb{P}(X_{[1,T]} \mid M^t_0 = m).$$

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 (10). □

**Remark.** Let $\mathsf{mem}$ be a predictive memory map. Consider the set $\mathsf{M}' := \{\, \mathsf{mem}(x) \mid x \in \mathsf{D}^* \,\} \subseteq \mathcal{P}(\mathsf{M})$ of *information states* and the deterministic $\mathsf{M}'$-valued memory map defined by $\mathsf{mem}'\big(x; \mathsf{mem}(x)\big) = 1$. Then $\mathsf{mem}'$ is easily seen to be a prescient memory map (i.e. a sufficient statistic on the past for the future). Note that while $\mathsf{mem}'$ is directly linked to $\mathsf{mem}$, $\mathsf{M}'$ may be much bigger than $\mathsf{M}$. In Proposition 3.5 we show that the cardinality of $\mathsf{M}'$ is bounded below by the number of causal states.

### 3.2 Implications on minimality of predictive models

Figure 6 illustrates the situation in view of Proposition 3.4 with the abbreviations "DM = deterministic memory," "PsM = prescient memory," and "PdM = predictive memory." In

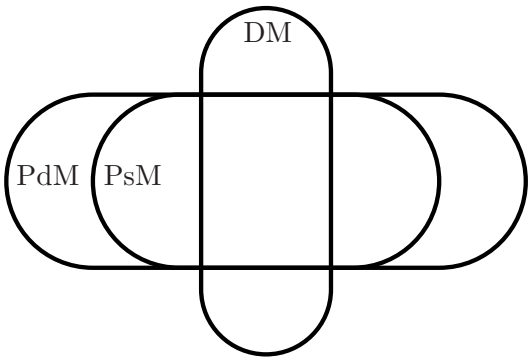


Figure 6: The extension of the memory class suggested by predictive models

computational mechanics, prescient deterministic memories have been studied, that is the intersection of DM and PsM. 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:

$$\mathrm{DM} \cap \mathrm{PsM} \;=\; \mathrm{DM} \cap \mathrm{PdM} \;\subsetneq\; \mathrm{PsM} \;\subsetneq\; \mathrm{PdM}. \tag{14}$$

According to the equality in (14), considering predictive memory maps without dropping the determinism requirement does not enlarge the class. Only recently, an extension to the class PsM 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 PsM to PdM and show in Example 3.6 that this extension is indeed effective.

**Proposition 3.5 ($\varepsilon$-machine minimality in PsM).** *The causal state projection $\mathfrak{C}$ of Example 2.4 defines a prescient deterministic memory map to $\mathsf{M}_{\mathfrak{C}}$. Further, it has minimal number of memory states amongst the prescient (not necessarily deterministic) memories. More precisely:*

$$\mathsf{mem}\colon \mathsf{D}^* \to \mathcal{P}(\mathsf{M}) \text{ prescient} \qquad \Rightarrow \qquad |\mathsf{M}| \;\geq\; |\mathsf{M}_{\mathfrak{C}}|.$$

*Proof.* We show (10). 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 prescient. Now assume that $\mathsf{mem}$ is another prescient 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 (10), 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 memories in PdM not captured within PsM. It is an example of an HMM with two hidden states that, consequently, admits a predictive model with two memory states. The corresponding minimal number of memory states within PsM, which according to Proposition 3.5 is realized by the $\varepsilon$-machine, turns out to be infinite.[9] Moreover, the causal states are singletons, so that the causal state projection does not achieve any compression. Here, different histories lead to different expectations on the hidden states (different information states) of the HMM. When considering infinite histories, the causal states of the $\varepsilon$-machine turn out to be even uncountable, whereas the HMM can still be turned into a predictive model with two internal states (see the appendix).

[9] The existence of finite HMMs leading to infinitely many causal states was already mentioned in [Crutchfield, 1994].

**Example 3.6 (predictive model smaller than the $\varepsilon$-machine).** In order to specify this example, we consider an HMM $(\mathsf{gen}, \mu)$ with internal state set $\mathsf{M}$. 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 $\mathsf{D}$ and $\mathsf{M}$ 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 $\mathsf{D} := \mathsf{M} := \{\,0, 1\,\}$ and the uniform distribution on $\mathsf{M}$ as initial distribution. With a parameter $p$, $0 < p < \frac{1}{4}$, we define the generator by

$$\mathsf{gen}(m;\, x, \hat{m}) \;:=\; \begin{cases} 1-2p, & \text{if } \hat{m} = x = m \\ p, & \text{if } x \neq m \\ 0, & \text{otherwise} \end{cases} . \tag{15}$$

See Figure 7 for an illustration of the transition graph. It is easy to check that the stationarity

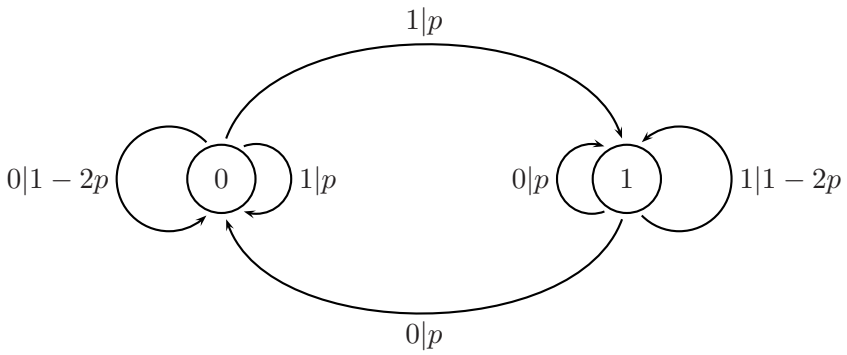


Figure 7: Transition graph of the generator defined by (15). Circled nodes are internal states and edges are transitions, labeled with output symbol $x$ and transition probability $q$ as "$x|q$".

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:

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

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

$$\begin{aligned} \mathbb{P}(X_{[1,T]} \mid X_{[-t+1,0]} = x_{[-t+1,0]}) \;&=\; \mathbb{P}(X_{[1,T]} \mid X_{[-s+1,0]} = x'_{[-s+1,0]}) \\ &\Leftrightarrow \qquad f_{x_0} \circ \cdots \circ f_{x_{-t+1}}(\tfrac{1}{2}) \;=\; f_{x'_0} \circ \cdots \circ f_{x'_{-s+1}}(\tfrac{1}{2}). \end{aligned} \tag{16}$$

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 (16), the causal states are singletons. $\Diamond$

# 4 Conclusions

There are two natural and well-known concepts of constructing an HMM as model of a given stochastic process. The first one is based on the idea that the memory variable (hidden variable of the HMM) is prescient in the sense of being a sufficient statistic on the past for the future. Minimizing the size of such a prescient memory leads to the so-called $\varepsilon$-machine which can be obtained in a constructive and efficient way. The second way of assigning an HMM to a stochastic process requires the HMM to be capable of generating the given process. The problem of finding a minimal such HMM is an intrinsically hard problem and can not be solved as easily as in the $\varepsilon$-machine case. Both approaches provide different ways of understanding prediction which is a central theme of computational mechanics. We feel that the current literature does not highlight the difference between these two approaches in sufficient detail. This, unfortunately, leaves space for a misperception of $\varepsilon$-machines as minimal HMMs capable of generating the underlying process, which corresponds to the second approach and not to the $\varepsilon$-machine approach. Based on this perception, there is an apparent inconsistency between the $\varepsilon$-machine minimality and the existence of (substantially) smaller hidden Markov models that generate the same underlying process. In this article, we clarified the formal relation between the two approaches and thereby identified the $\varepsilon$-machine approach as the more restrictive one. We argued that the assumptions made within the second approach are related to a generative understanding of prediction which leads to the notion of a predictive memory and a corresponding predictive model. Based on this generative understanding of prediction we could show that the underlying HMM can be significantly smaller than the corresponding $\varepsilon$-machine. We compared the operational aspects of the two concepts of prediction and discussed some drawbacks of our approach in this regard. Currently, we do not know which of the two notions of prediction is the more natural one and further steps towards revealing and comparing operational aspects of prediction are subject of our research.

# A Appendix: Prediction with infinite histories

In this appendix, we translate our setting and results to the case of infinite histories. Essentially all results remain true, but we have to deal with measure theoretic issues because the set of infinite histories is uncountable and may enforce uncountably many internal states of predictive models including the $\varepsilon$-machine. As it does not require much additional effort, we also consider more general, non-discrete state spaces for the (observable) stochastic process.

## A.1 Generating a process

We assume that the state space $\mathsf{D}$ of the generated process is a Polish space, the space $\mathsf{M}$ of memory (or internal) states is a Souslin space,[10] and the generator $\mathsf{gen}$ is a Markov kernel, i.e. a *measurable* map

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

Here, the topological spaces $\mathsf{M}$ and $\mathsf{D}$ are equipped with their Borel $\sigma$-algebras and $\mathcal{P}(\mathsf{D} \times \mathsf{M})$ with the $\sigma$-algebra induced by the evaluations $e_A\colon \mu \mapsto \mu(A)$ which coincides with the Borel $\sigma$-algebra of the weak-$*$ topology. We use the notation $\mathsf{gen}(m;\, D \times B)$ to denote the probability of a measurable set (event) $D \times B$ with respect to $\mathsf{gen}(m)$. Like in the discrete case, $\mathsf{gen}$ together with an initial distribution $\mu$ on $\mathsf{M}$ generates stochastic processes $\widetilde{X}_{\mathbb{N}}$ and $M_{\mathbb{N}_0}$ on $\mathsf{D}$ and $\mathsf{M}$ (the Kolmogorov extension theorem holds for Souslin spaces). The finite dimensional joint distributions are computed according to

$$\begin{aligned}&\mathbb{P}(\widetilde{X}_{[1,T]} \in D_{[1,T]},\ M_{[0,T]} \in B_{[0,T]}) \\ &\quad = \int_{B_0} \int_{D_1 \times B_1} \cdots \int_{D_T \times B_T} 1\ \mathsf{gen}(m_{T-1};\, \mathrm{d}(d_T, m_T)) \ \cdots\ \mathsf{gen}(m_0;\, \mathrm{d}(d_1, m_1))\ \mu(\mathrm{d}m_0),\end{aligned}$$

where we use the notation $D_I = \prod_{i \in I} D_i$ for measurable sets $D_i$. If the initial distribution $\mu$ is $\mathsf{gen}$-invariant, the processes $M_{\mathbb{N}_0}$ and $\widetilde{X}_{\mathbb{N}}$ are extended to stationary processes $M_{\mathbb{Z}}$ and $\widetilde{X}_{\mathbb{Z}}$.

**Definition A.1 (generating a process).** Let $X_{\mathbb{N}}$ (resp. $X_{\mathbb{Z}}$) be a stochastic process on $\mathsf{D}$. We say that $\mathsf{gen}$ *generates* $X_{\mathbb{N}}$ (resp. $X_{\mathbb{Z}}$) if there exists an initial distribution (resp. *invariant* distribution) for $\mathsf{gen}$ such that $\widetilde{X}_{\mathbb{N}}$ (resp. $\widetilde{X}_{\mathbb{Z}}$) has the same distribution as $X_{\mathbb{N}}$ (resp. $X_{\mathbb{Z}}$).

## A.2 Our prediction setting

Let $X_{\mathbb{Z}}$ be a $\mathsf{D}$-valued, stationary stochastic process and $\mathsf{D}$ Polish.

**Definition A.2 (memory).** A *memory* (*map*) $\mathsf{mem}$ is a Markov kernel from the set $\mathsf{D}^{-\mathbb{N}_0}$ of history trajectories to a Souslin space $\mathsf{M}$ of *memory states*:

$$\mathsf{mem}\colon \mathsf{D}^{-\mathbb{N}_0} \ \to\ \mathcal{P}(\mathsf{M}).$$

As in Section 2.2, $\mathsf{mem}$ induces a random variable $M = M_0$ and, together with a generator $\mathsf{gen}$, we obtain processes $\widetilde{X}_{\mathbb{N}}$, $M_{\mathbb{N}_0}$. Note that here we have a single memory state at time 0 and there is no need for an upper index, as the history length is fixed (and infinite). In the following, we will use conditional probabilities and make the general assumption that they are *regular* versions. In Souslin spaces, this is always possible. Also note that the $\sigma$-algebra of a Souslin space is countably generated.

[10] Souslin spaces are slightly more general than Polish spaces ensuring that images of Souslin spaces under measurable maps, such as the set of causal states, are again Souslin spaces (see [Bourbaki, 1989]).

**Definition A.3 (predictive model).** We call the memory map $\mathsf{mem}$ *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

$$\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(\widetilde{X}_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \qquad \text{a.s. (almost surely)}$$

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

**Proposition A.4 (generator as predictive model).** *Let* $\mathsf{gen}\colon \mathsf{M} \to \mathcal{P}(\mathsf{D} \times \mathsf{M})$ *be a generator that generates the process* $X_{\mathbb{Z}}$. *Then there is a memory map* $\mathsf{mem}\colon \mathsf{D}^{-\mathbb{N}_0} \to \mathcal{P}(\mathsf{M})$, *such that* $(\mathsf{mem}, \mathsf{gen})$ *is a predictive model of* $X_{\mathbb{Z}}$.

*Proof.* Let the invariant initial distribution for $\mathsf{gen}$ be such that $\widetilde{X}_{\mathbb{Z}}$ has the same distribution as $X_{\mathbb{Z}}$. Define $\mathsf{mem}$ by

$$\mathsf{mem} \circ \widetilde{X}_{-\mathbb{N}_0} \;:=\; \mathbb{P}(M_0 \mid \widetilde{X}_{-\mathbb{N}_0}).$$

We denote the processes generated by $\mathsf{gen}$ started in $\mathsf{mem} \circ X_{-\mathbb{N}_0}$ by $\widehat{X}_{\mathbb{N}}$ and $\widehat{M}_{\mathbb{N}_0}$, respectively. We obtain

$$\begin{aligned}
\mathbb{P}(\widehat{X}_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x) \;&=\; \int \mathbb{P}(\widehat{X}_{\mathbb{N}} \mid \widehat{M}_0 = m)\, \mathsf{mem}(x;\, \mathrm{d}m) \\
&=\; \int \mathbb{P}(\widetilde{X}_{\mathbb{N}} \mid M_0 = m)\, \mathbb{P}(M_0 \in \mathrm{d}m \mid \widetilde{X}_{-\mathbb{N}_0} = x) \\
&=\; \mathbb{P}(\widetilde{X}_{\mathbb{N}} \mid \widetilde{X}_{-\mathbb{N}_0} = x) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x),
\end{aligned}$$

$\mathbb{P} \circ X_{-\mathbb{N}_0}^{-1}$-a.s. in $x \in \mathsf{D}^{-\mathbb{N}_0}$. Thus $(\mathsf{mem}, \mathsf{gen})$ is a predictive model. □

**Remark.** Note that in Proposition A.4, in contrast to the finite-history case (see Proposition 2.5), it is not sufficient to assume that $\mathsf{gen}$ generates only the positive time process $X_{\mathbb{N}}$. Instead, we have to assume that the whole process $X_{\mathbb{Z}}$ is generated. This is, however, satisfied in our main Example 3.6. Thus it also admits a predictive model for infinite histories.

## A.3 Comparison to computational mechanics

**Definition A.5 (prescient).** We call a memory map $\mathsf{mem}$ *prescient*, if the resulting $M$ satisfies

$$\mathbb{P}(X_{\mathbb{N}} \mid M) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \qquad \text{a.s.}$$

Predictive and prescient memories are related in the same way as in the finite-history case:

**Proposition A.6 (predictive versus prescient).**
1. *Every prescient memory map is predictive.*
2. *If a memory map is deterministic and predictive, then it is also prescient.*

*Proof.*
1. Let $\mathsf{mem}$ be a prescient memory and $M^k = M_0^k$ the resulting memory state at time $k$ (calculated by $\mathsf{mem}$ from $X_{]-\infty,k]}$). Define the generator $\mathsf{gen}\colon \mathsf{M} \to \mathcal{P}(\mathsf{D} \times \mathsf{M})$ by

$$\mathsf{gen}(m;\, D \times B) \;:=\; \mathbb{P}(X_1 \in D,\; M^1 \in B \mid M^0 = m), \tag{17}$$

and let $M_l = M_l^0$ be the memory state produced by $\mathsf{gen}$ at time $l$ when started in $M^0$ at time 0. We show that $\mathbb{P}(\widetilde{X}_{[1,T]} \mid M^0) = \mathbb{P}(X_{[1,T]} \mid M^0)$ a.s. by induction over $T$. We obtain

$$\begin{aligned}\mathbb{P}(\widetilde{X}_{[2,T]} \mid M_1 = m) \quad &= \quad \mathbb{P}(\widetilde{X}_{[1,T-1]} \mid M_0 = m) \overset{\text{(ind. as.)}}{=} \mathbb{P}(X_{[1,T-1]} \mid M_0 = m) \\ &\overset{(X_{\mathbb{Z}} \text{ stat.})}{=} \mathbb{P}(X_{[2,T]} \mid M_1 = m), \end{aligned} \tag{18}$$

which leads to

$$\begin{aligned}\mathbb{P}(\widetilde{X}_{[1,T]} \in D_{[1,T]} \mid M^0) \quad &= \quad \int_{D_1 \times \mathsf{M}} \mathbb{P}(\widetilde{X}_{[2,T]} \in D_{[2,T]} \mid M_1^0 = \,\cdot\,) \,\mathrm{d}\,\mathsf{gen}(M^0) \\ &\overset{(17)+(18)}{=} \int_{X_1^{-1}(D_1)} \mathbb{P}(X_{[2,T]} \in D_{[2,T]} \mid M^1) \,\mathrm{d}\mathbb{P}(\,\cdot \mid M^0) \\ &\overset{\text{(prescience)}}{=} \mathbb{P}(X_{[1,T]} \in D_{[1,T]} \mid M^0).\end{aligned}$$

This finishes the induction and in total we obtain

$$\mathbb{P}(\widetilde{X}_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \int \overbrace{\mathbb{P}(\widetilde{X}_{\mathbb{N}} \mid M^0)}^{=\mathbb{P}(X_{\mathbb{N}} \mid M^0)} \,\mathrm{d}\mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0}) \overset{\text{(prescience)}}{=} \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}).$$

2. Now let $(\mathsf{mem}, \mathsf{gen})$ be a predictive model and $\mathsf{mem}$ deterministic, i.e. $M = M^0 = f \circ X_{-\mathbb{N}_0}$ for some measurable function $f\colon \mathsf{D}^{-\mathbb{N}_0} \to \mathsf{M}$. Then $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) = \mathbb{P}(\widetilde{X}_{\mathbb{N}} \mid M)$ is $\sigma(M)$-measurable modulo $\mathbb{P}$. Since obviously $\sigma(M) \subseteq \sigma(X_{-\mathbb{N}_0})$, it holds a.s. that

$$\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M). \qquad \square$$

The causal states are defined as equivalence classes of histories inducing the same expectation on the future (with a fixed, regular version of conditional probability). Within the measure theoretic setting of this appendix, we find it more convenient to use a different but equivalent definition. To this end, consider the *causal state projection* $\mathfrak{C}\colon \mathsf{D}^{-\mathbb{N}_0} \to \mathcal{P}(\mathsf{D}^{\mathbb{N}})$ defined by

$$x \;\mapsto\; \mathfrak{C}(x) \;:=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x).$$

The set of *causal states* is the image of this map:

$$\mathsf{M}_{\mathfrak{C}} \;:=\; \mathrm{Im}(\mathfrak{C}) \;\subseteq\; \mathcal{P}(\mathsf{D}^{\mathbb{N}}).$$

Note that $\mathsf{M}_{\mathfrak{C}}$ inherits a Souslin topology and thus a measurable structure from $\mathcal{P}(\mathsf{D}^{\mathbb{N}})$. The map $\mathfrak{C}$ is measurable, and it is easy to see that the corresponding deterministic memory is prescient. Before we prove that $\mathfrak{C}$ corresponds to, as in the discrete case, a minimal prescient memory, we consider the process $X_{\mathbb{Z}}$ of Example 3.6. With a slight modification of the arguments there, one can show that the corresponding set $\mathsf{M}_{\mathfrak{C}}$ of causal states is uncountable (regardless of the version of conditional probability). On the other hand, there is a predictive model (for infinite histories) with two memory states.

**Proposition A.7 ($\varepsilon$-machine minimaltiy).** *Let* $\mathsf{mem}$ *be a prescient memory map with set* $\mathsf{M}$ *of memory states. Then there exist* disjoint *measurable subsets* $\mathsf{N}_z \subseteq \mathsf{M}$, $z \in \mathsf{M}_{\mathfrak{C}}$, *such that*

$$\mathsf{mem}(x;\, \mathsf{N}_{\mathfrak{C}(x)}) \;=\; 1 \qquad a.s.$$

*In particular,* $\mathsf{M}$ *cannot be essentially (i.e. up to zero-sets) smaller than* $\mathsf{M}_{\mathfrak{C}}$.

*Proof.* Let $M = M_0$ be the memory variable obtained by $\mathsf{mem}$. Define $P_M \colon \mathsf{M} \to \mathcal{P}(\mathsf{D}^{\mathbb{N}})$ by

$$P_M \circ M \;:=\; \mathbb{P}(X_{\mathbb{N}} \mid M), \qquad \text{and} \qquad \mathsf{N}_z \;:=\; P_M^{-1}(z) \;\text{ for } z \in \mathsf{M}_{\mathfrak{C}}.$$

The $\mathsf{N}_z$ are obviously measurable and disjoint. Due to prescience, $P_M \circ M = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$ a.s., and consequently

$$\mathsf{mem}(x;\, \mathsf{N}_{\mathfrak{C}(x)}) \;=\; \mathbb{P}(M \in \mathsf{N}_{\mathfrak{C}(x)} \mid X_{-\mathbb{N}_0} = x) \;=\; \mathbb{P}(P_M \circ M = \mathfrak{C}(x) \mid X_{-\mathbb{N}_0} = x) \;=\; 1,$$

where we used $\mathfrak{C}(x) = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x)$. □

## Acknowledgments

The authors are grateful for discussions with Jim Crutchfield, Cosma Shalizi, and Susanne Still. They also thank the anonymous referees for their constructive comments which substantially improved the paper. 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.

[Bourbaki, 1989] Bourbaki, N. (1989). *General Topology, Chapters 5-10*. Springer-Verlag.

[Bukharaev, 1995] Bukharaev, R. G. (1995). *Theorie der stochastischen Automaten*. B.G. Teubner.

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

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

[Faigle and Schönhuth, 2007] Faigle, U. and Schönhuth, A. (2007). Asymptotic mean stationarity of sources with finite evolution dimension. *IEEE Transactions on Information Theory*, 53(7):2342–2348.

[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.

[Kulhavý, 1996] Kulhavý, R. (1996). *Recursive Nonlinear Estimation: A Geometric Approach*, volume 216 of *Lecture Notes in Control and Information Sciences*. Springer.

[Littman et al., 2001] Littman, M. L., Sutton, R. S., and Singh, S. (2001). Predictive representations of state. *Advances in Neural Information Processing Systems*, 14.

[Schönhuth and Jaeger, 2007] Schönhuth, A. and Jaeger, H. (2007). Characterization of ergodic hidden markov sources. Technical report, Zentrum für Angewandte Informatik Köln. Submitted to IEEE Transactions on Information Theory.

[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.

[Shalizi et al., 2002] Shalizi, C. R., Shalizi, K. L., and Crutchfield, J. P. (2002). An algorithm for pattern discovery in time series. Informal publication, `http://arxiv.org/abs/cs.LG/0210025`.

[Singh et al., 2004] Singh, S., James, M. R., and Rudary, M. R. (2004). Predictive state representations: a new theory for modeling dynamical systems. In *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 512–519, Arlington, Virginia, United States. AUAI Press.

[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.