Non-sufficient Memories That Are Sufficient for Prediction

Wolfgang Löhr[1] and Nihat Ay[1,2]

[1] Max Planck Institute for Mathematics in the Sciences,
Inselstraße 22, D-04103 Leipzig, Germany
{Wolfgang.Loehr,Nihat.Ay}@mis.mpg.de
[2] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

Abstract. The causal states of computational mechanics define the minimal sufficient (prescient) memory for a given stationary stochastic process. They induce the ε-machine which is a hidden Markov model (HMM) generating the process. The ε-machine is, however, not the minimal generative HMM and minimal internal state entropy of a generative HMM is a tighter upper bound for excess entropy than provided by statistical complexity. We propose a notion of prediction that does not require sufficiency. The corresponding models can be substantially smaller than the ε-machine and are closely related to generative HMMs.

Keywords: hidden Markov models, HMM, computational mechanics, causal states, ε-machine, prediction.

1 Introduction

Computational mechanics is a theory developed by Crutchfield, Young, Shalizi and others ([1,2]). It tackles the problem of building predictive models of stationary stochastic processes[1] and finding the minimal such model. This problem is solved by the so-called ε-machine which operates on the causal states. Although the ε-machine is a hidden Markov model (HMM) and minimal under the assumptions of computational mechanics, it is (in general) distinct from and can be much larger than the minimal HMM capable of generating the process. In the literature, this distinction is not always clear. Also, minimal entropy of a generative HMM provides a tighter upper bound for excess entropy than statistical complexity does (see Example 7).

In the present paper, we compare and highlight the difference between the approach of computational mechanics, which is based on the fundamental concept of sufficient statistics, and the construction of the minimal generative HMM. We propose a notion of predictive model that is weaker than sufficiency and thereby allows for smaller models. More specifically, we require our models to be able to generate a prediction of the future that follows the same conditional distribution as the real future (Section 4). It turns out that if a process is generated by

[1] Extensions to spatio-temporal systems exist, but we do not consider them here.

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 265–276, 2009.

an HMM, the minimal predictive model in our sense cannot be larger than the original HMM. We have already presented main idea and results of the present paper in [3]. Therefore, we omit the proofs of the propositions in this less technical review; they can be found in the appendix of [3]. Complementary to [3], we discuss the relation between excess entropy, statistical complexity and the size of generative HMMs (Corollary 6 and Example 7).

2 Sufficient Statistics and Causal States

Consider a stationary stochastic process $X_{\mathbb{Z}} = (\dots, X_{-1}, X_0, X_1, \dots)$ on a discrete alphabet D. We interpret $X_{-\mathbb{N}_0}$ as the observed past and $X_{\mathbb{N}}$ as the future, which we want to predict. Not all information of $X_{-\mathbb{N}_0}$ is necessary for predicting $X_{\mathbb{N}}$. Therefore, one tries to compress the relevant information in a memory variable M, which assumes values in a set M of memory states, via a memory kernel (transition probability) mem. This is illustrated as

$$\begin{tikzcd}
X_{-\mathbb{N}_0} \arrow[r] \arrow[dr, "\mathsf{mem}"'] & X_{\mathbb{N}} \\
 & M
\end{tikzcd}$$

Sometimes, we call both the memory variable M and the memory kernel mem simply *memory*. No confusion arises, as one determines the other. For technical simplicity, we restrict to countable M, although this restriction is not necessary (see the appendix of [3]).

The usual approach in computational mechanics is to consider the special case of deterministic functions instead of memory kernels mem, but recently an extension to stochastic maps has been considered by Still and Crutchfield ([4]). We adopt this extension and do not require mem to be deterministic, allowing for a stochastic assignment. That is

$$\mathsf{mem}\colon \mathsf{D}^{-\mathbb{N}_0} \to \mathcal{P}(\mathsf{M}) \quad \text{measurable},$$

where $\mathcal{P}(\mathsf{M})$ denotes the set of probability measures on M. Note that M is embedded in $\mathcal{P}(\mathsf{M})$ via Dirac measures and thus a (measurable) deterministic memory function $f\colon \mathsf{D}^{-\mathbb{N}_0} \to \mathsf{M}$ induces a memory kernel $\mathsf{mem}_f(x_{-\mathbb{N}_0}) = \delta_{f(x_{-\mathbb{N}_0})}$, where δ_m is the Dirac measure in m. In general, mem reduces the information about the future, which is expressed by the following inequality:

$$I(M : X_{\mathbb{N}}) \;\leq\; I(X_{-\mathbb{N}_0} : X_{\mathbb{N}}) \;=:\; E(X_{\mathbb{Z}}).$$

where I denotes the mutual information between two random variables[2] and E is the *excess entropy*, an important complexity measure also known as *effective measure complexity* and *predictive information* ([5,6]). In computational

[2] $X_{-\mathbb{N}_0}$ and $X_{\mathbb{N}}$ are not discrete-valued. Their mutual information is defined by the limit $I(X_{-\mathbb{N}_0} : X_{\mathbb{N}}) := \sup_{n,m} I(X_{[-n,0]} : X_{[1,m]}) = \lim_{n\to\infty} I(X_{[-n,0]} : X_{[1,n]})$.

mechanics, one requires that the memory preserves all information about the future. This property is called *prescient* ([2]) and formalized by

$$I(M : X_{\mathbb{N}}) \;=\; E(X_{\mathbb{Z}}). \tag{1}$$

It is this central requirement that ensures minimality of causal states (Proposition 1) and ε-machine while ruling out smaller hidden Markov models. We will relax it in Section 4 to a different notion of "predictive". Requirement (1) is equivalent to conditional independence of past and future given the memory:

$$X_{-\mathbb{N}_0} \perp\!\!\!\perp X_{\mathbb{N}} \mid M.$$

Using the language of statistics, we say that such a memory is *sufficient* for the future, or simply that M is a **sufficient memory**. Sufficient memories are the candidates for *predictive models* proposed by computational mechanics. It is natural to ask how big a sufficient memory has to be and how to obtain a minimal one. There are mainly two possibilities to measure the size of a memory: cardinality $|\mathsf{M}|$ of the set of memory states and Shannon entropy $H(M)$ of the memory variable. Both notions of size, however, yield the same notion of minimality and the unique solution is given by the causal states, which are constructed in the following way: We identify two history trajectories, $x_{-\mathbb{N}_0}, \hat{x}_{-\mathbb{N}_0} \in \mathsf{D}^{-\mathbb{N}_0}$, if they induce the same conditional probability on the future, i.e.

$$x_{-\mathbb{N}_0} \sim \hat{x}_{-\mathbb{N}_0} \quad :\Leftrightarrow \quad P(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x_{-\mathbb{N}_0}) \;=\; P(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = \hat{x}_{-\mathbb{N}_0})\,.[3]$$

The *causal state* $\mathfrak{C}(x_{-\mathbb{N}_0})$ of $x_{-\mathbb{N}_0}$ is its equivalence class,

$$\mathfrak{C}(x_{-\mathbb{N}_0}) \;:=\; \{\, \hat{x}_{-\mathbb{N}_0} \mid x_{-\mathbb{N}_0} \sim \hat{x}_{-\mathbb{N}_0} \,\},$$

and the function $\mathfrak{C}$ defines a deterministic sufficient memory (see [2]).[4] Its set of memory states is the set of causal states,[5]

$$\mathsf{M}_{\mathfrak{C}} \;:=\; \mathrm{Im}(\mathfrak{C}) \;=\; \big\{\, \mathfrak{C}(x_{-\mathbb{N}_0}) \;\big|\; x_{-\mathbb{N}_0} \in \mathsf{D}^{-\mathbb{N}_0} \,\big\},$$

and the memory kernel $\mathsf{mem}_{\mathfrak{C}}$ is defined by $\mathsf{mem}_{\mathfrak{C}}(x_{-\mathbb{N}_0}) = \delta_{\mathfrak{C}(x_{-\mathbb{N}_0})}$, the Dirac measure in the corresponding causal state. It is well-known that the set $\mathsf{M}_{\mathfrak{C}}$ of causal states is the minimal prescient partition of $\mathsf{D}^{-\mathbb{N}_0}$. Consequently, $\mathsf{mem}_{\mathfrak{C}}$ is the minimal sufficient deterministic memory. This property easily extends to the non-deterministic case:

Proposition 1 (minimality of causal states). *Any sufficient memory with set* M *of memory states and memory variable* M *satisfies*

$$|\mathsf{M}| \;\geq\; |\mathsf{M}_{\mathfrak{C}}| \qquad \textit{and} \qquad H(M) \;\geq\; H(M_{\mathfrak{C}}).$$

[3] $P(X \mid Y = y) = P(X \mid Y = \hat{y})$ means that $P(X \in B \mid Y = y) = P(X \in B \mid Y = \hat{y})$ for every measurable set (event) B.

[4] We fix a regular version of conditional probability $P(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$. Therefore, the function $\mathfrak{C}$ is measurable and the causal states are measurable subsets of $\mathsf{D}^{-\mathbb{N}_0}$.

[5] In general, $\mathsf{M}_{\mathfrak{C}}$ need not be countable. Here, we restrict to processes with a countable number of causal states. For the more general case, see the appendix of [3].

Due to the minimality of the causal states, their entropy

$$C_{\mathfrak{C}}(X_{\mathbb{Z}}) := H(M_{\mathfrak{C}})$$

is an important complexity measure called *statistical complexity.* It is evident from (1) that statistical complexity is lower bounded by excess entropy.

A memory kernel mem does not only induce a (random) memory state $M = M_0$ at time zero, but a whole stationary process $M_{\mathbb{Z}}$ of memory states. The conditional distribution of $M_{\mathbb{Z}}$ is computed as

$$P(M_{[0,T]} = m_{[0,T]} \mid X_{\mathbb{Z}} = x_{\mathbb{Z}}) = \prod_{k=0}^{T} \mathsf{mem}(x_{]-\infty,k]};\, m_k), \qquad T \in \mathbb{N}_0,$$

where we use the notation $[0,T]$ for the discrete interval $\{\, 0, \dots, T \,\}$ and $M_{[0,T]} = m_{[0,T]}$ for $M_0 = m_0, \dots, M_T = m_T$. Note that the process $M_{\mathbb{Z}}$ of a sufficient memory need not be Markovian. However, the memory process of the *minimal* sufficient memory, i.e. the process of causal states, is always Markovian ([2]).

3 Hidden Markov Models (HMMs) and ε-Machine

Sufficient memories, such as given by the causal states, contain all information about the future that is available in the past. How do we actually extract this information and justify the term "model" for sufficient memories? In computational mechanics, the ε-machine describes the mechanism of prediction. It is defined as a *stochastic output automaton*, i.e. a "machine" with the following components: It has a set S of internal states and is initialized by one of these states according to some initial probability distribution $\mu \in \mathcal{P}(\mathsf{S})$. We assume S to be countable for technical simplicity. At each time step t, depending on the current internal state S_t, an output symbol Y_{t+1} from the finite alphabet D and a new internal state S_{t+1} are (stochastically) generated. This is modeled by a joint transition probability gen from the internal states to output symbols and internal states:

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

Thus the pair (gen, μ) of generating mechanism and initial distribution induces processes $S_{\mathbb{N}_0}$, $Y_{\mathbb{N}_0}$ of internal states and output symbols. The situation is illustrated as

$$\begin{tikzcd}
S_0 \arrow[r] \arrow[dr] & S_1 \arrow[r] \arrow[d, dash] \arrow[dr] & S_2 \arrow[d, dash] & \cdots & S_{T-1} \arrow[r] \arrow[d, dash] \arrow[dr] & S_T \arrow[d, dash] \\
 & Y_1 & Y_2 & \cdots & Y_{T-1} & Y_T
\end{tikzcd}$$

The joint distribution of internal- and output process is computed according to

$$P(S_{[0,T]} = s_{[0,T]},\, Y_{[1,T]} = y_{[1,T]}) = \mu(s_0) \prod_{k=1}^{T} \mathsf{gen}(s_{k-1};\, y_k, s_k), \quad T \in \mathbb{N}\ .$$[6]

[6] $\mathsf{gen}(s;\, y, \hat{s})$ denotes the probability of the pair $(y, \hat{s})$ w.r.t. the measure $\mathsf{gen}(s)$.

Stochastic automata are also widely known as *edge-emitting hidden Markov models.*[7] We use this terminology, but for brevity we call the pair (gen, μ) hidden Markov model (*HMM*) and always mean edge-emitting HMM.

If the initial distribution μ is gen-invariant, i.e. if

$$\mu(s) \;=\; \sum_{\hat{s}\in\mathsf{S},\, d\in\mathsf{D}} \mu(\hat{s})\,\mathsf{gen}(\hat{s};\, d, s) \qquad \forall s \in \mathsf{S},$$

then the processes $S_{\mathbb{N}_0}$ and $Y_{\mathbb{N}_0}$ are (jointly) stationary and uniquely extended to processes $S_{\mathbb{Z}}$ and $Y_{\mathbb{Z}}$ respectively. We then call the HMM *stationary.* Because our aim is to investigate given processes $X_{\mathbb{Z}}$ with time set $\mathbb{Z}$, we assume in this section that μ is gen-invariant. If the law of the output process $Y_{\mathbb{Z}}$ of a stationary HMM (gen, μ) coincides with the law of the given process $X_{\mathbb{Z}}$, the HMM is a *generative model* for the process of interest: we can easily simulate and investigate statistical properties of $X_{\mathbb{Z}}$ by means of the HMM. We call such an HMM **generative**. A generative HMM is a possibility, how the process $X_{\mathbb{Z}}$ might have been produced, although it is of course highly non-unique. The question about a *minimal* generative HMM suggests itself. With "minimal" we either mean minimal cardinality $|\mathsf{S}|$ of the set of internal states or minimal entropy $H(\mu) = H(S_0)$ of the invariant initial distribution. Unlike in the situation of sufficient memories, these two notions do not coincide (see Example 7).

Finding the minimal generative HMM is intrinsically difficult,[8] but every sufficient memory M induces an HMM, thus providing an upper bound. In general, the process of internal states of the associated HMM cannot have the same distribution as the process $M_{\mathbb{Z}}$ of memory states, because the latter process need not be Markovian. The (first order) Markov approximation of the joint process $(M_{\mathbb{Z}}, X_{\mathbb{Z}})$, however, yields the desired HMM:

Proposition 2 (sufficient memories induce generative HMMs). *Let* mem *be a sufficient memory kernel and* $M_{\mathbb{Z}}$ *its process of memory states. Then a generative HMM is given by* $\mathsf{S} := \mathsf{M}$, $\mu(s) := P(M_0 = s)$ *and*

$$\mathsf{gen}(s;\, d, \hat{s}) \;:=\; P(X_1 = d,\, M_1 = \hat{s} \mid M_0 = s).$$

Example 3 (ε-machine). If we take the causal states as sufficient memory, the HMM $(\mathsf{gen}_{\mathfrak{C}}, \mu_{\mathfrak{C}})$ constructed in Proposition 2 is the ε-machine of computational mechanics. As the process of causal states is already Markovian, the ε-machine fully describes the (statistics of the) time evolution of the causal states. ◇

The causal states provide the minimal sufficient memory and induce the ε-machine. But is the latter also the minimal generative HMM? In general, the answer is "no". The ε-machine may be arbitrarily much bigger than the minimal HMM. It can be infinite or even uncountable, while there is a generative HMM

[7] "Edge-emitting" means that in visualizations as transition graphs the output symbols appear as edge labels.

[8] A geometric condition for minimality in terms of cardinality was specified by Heller in [7], but no constructive algorithm is known to us.

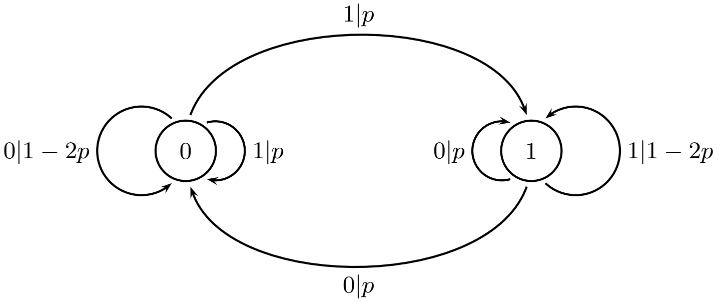


Fig. 1. Transition graph of the generator defined by (2). Circled nodes are internal states and edges are transitions, labeled with output symbol x and transition probability q as "$x|q$".

with only two internal states. This was already mentioned by Crutchfield in [8], but not everyone who applies computational mechanics seems to be aware of the fact. In the following, we give an example of this phenomenon.

Example 4 (uncountable ε-machine). We define the observable process $X_{\mathbb{Z}}$ by a stationary HMM with $\mathsf{D} := \mathsf{S} := \{0, 1\}$. It is clear that there is a generative HMM with two internal states, namely the original one. Nevertheless, the number of causal states (and thus the ε-machine) turns out to be uncountable. The initial distribution μ of the HMM is the uniform distribution. With a parameter $0 < p < \frac{1}{4}$, we define the generator by

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

See Figure 1 for an illustration of the transition graph. It is easy to check that μ is gen-invariant. One can show (see [3]) that the conditional probability for the internal state given a finite history behaves as follows: There exist intervals $I_n(x_{[-n,0]})$, disjoint for fixed n, such that

$$P(S_0 = 1 \mid X_{[-n,0]} = x_{[-n,0]}) \;\in\; I_n(x_{[-n,0]})$$

and the intervals are nested for increasing n, i.e.

$$I_{n+1}(x_{[-n-1,0]}) \;\subset\; I_n(x_{[-n,0]}).$$

Note that $P(S_0 \mid X_{-\mathbb{N}_0}) = \lim_{n\to\infty} P(S_0 \mid X_{[-n,0]})$ (a.s.) and that histories inducing different expectations on S_0 also induce different expectations on $X_{\mathbb{N}}$. Consequently, every causal state contains at most two (infinite) histories.[9] $\diamondsuit$

[9] This is true for the canonical version of conditional probability $P(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x_{-\mathbb{N}_0}) = \lim_{n\to\infty} \sum_{s=0}^{1} P(S_0 = s \mid X_{[-n,0]} = x_{[-n,0]}) P(X_{\mathbb{N}} \mid S_0 = s)$. Note that this limit always (not only a.s.) exists. Other choices may produce identifications on sets of measure zero, but still lead to uncountably many causal states.

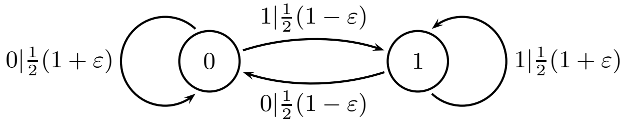

$1|\frac{1}{2}(1-\varepsilon)$
$0|\frac{1}{2}(1+\varepsilon)$
0
1
$1|\frac{1}{2}(1+\varepsilon)$
$0|\frac{1}{2}(1-\varepsilon)$

Fig. 2. ε-machine for a "nearly i.i.d." Markov process

Analogously to statistical complexity, one can consider the minimal internal-state entropy of a generative HMM:

Definition 5. Let $X_{\mathbb{Z}}$ be a stationary process. We call the quantity

$$C_{\mathrm{hmm}}(X_{\mathbb{Z}}) \;:=\; \inf_{(\mathsf{gen},\mu)} H(\mu),$$

where the infimum is taken over all generative HMMs, **generative complexity**.

For any generative HMM, $Y_{-\mathbb{N}_0} \to S_0 \to Y_{\mathbb{N}}$ is a Markov chain and $I(Y_{-\mathbb{N}_0} : Y_{\mathbb{N}}) = I(X_{-\mathbb{N}_0} : X_{\mathbb{N}})$. Therefore, the internal state entropy $H(\mu) = H(S_0)$ is lower-bounded by the excess entropy. Together with Proposition 2 we obtain

Corollary 6. $E(X_{\mathbb{Z}}) \;\leq\; C_{\mathrm{hmm}}(X_{\mathbb{Z}}) \;\leq\; C_{\mathfrak{C}}(X_{\mathbb{Z}})$

The following example demonstrates that for some processes both inequalities in Corollary 6 are strict. It also illustrates that HMMs with minimal entropy need not have the minimal number of internal states.

Example 7. Let $\mathsf{D} := \{0,1\}$ and consider the stationary Markov process $X^{\varepsilon}_{\mathbb{Z}}$ defined by

$$P(X^{\varepsilon}_0 = d) \;:=\; \tfrac{1}{2} \qquad \text{and} \qquad P(X^{\varepsilon}_{n+1} {=} \hat{d} \mid X^{\varepsilon}_n = d) \;:=\; \begin{cases} \frac{1}{2}(1+\varepsilon), & \text{if } d = \hat{d} \\ \frac{1}{2}(1-\varepsilon), & \text{if } d \neq \hat{d} \end{cases}.$$

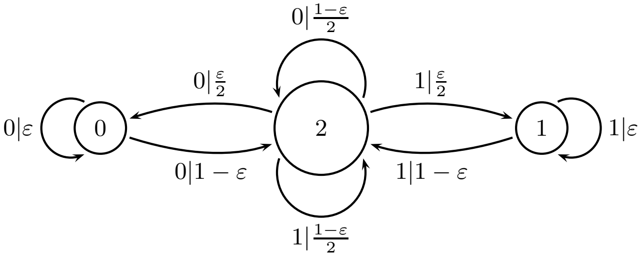
$X^{\varepsilon}_{\mathbb{Z}}$ is a disturbed i.i.d. process with disturbance of magnitude ε towards a constant process: For $\varepsilon = 0$, it is i.i.d. and for $\varepsilon = 1$ it is constantly 0 or 1, each with equal probability. For $0 < \varepsilon \leq 1$, there are two causal states which correspond to the last observed symbol. The ε-machine is visualised in Figure 2 and statistical complexity is given by

$$C_{\mathfrak{C}}(X^{\varepsilon}_{\mathbb{Z}}) \;=\; H(X^{\varepsilon}_0) \;=\; \ln(2),$$

regardless how small the parameter ε is. At $\varepsilon = 0$, $\varepsilon \mapsto C_{\mathfrak{C}}(X^{\varepsilon}_{\mathbb{Z}})$ has a discontinuity and assumes the value 0. The excess entropy behaves differently: at $\varepsilon = 1$ and $\varepsilon = 0$ it coincides with statistical complexity, but it is continuous in ε and behaves like $\frac{1}{2}\varepsilon^2$ for small ε. It can easily be calculated:

$$E(X^{\varepsilon}_{\mathbb{Z}}) \;=\; I(X^{\varepsilon}_1 : X^{\varepsilon}_0) \;=\; \tfrac{1}{2}\big((1+\varepsilon)\ln(1+\varepsilon) + (1-\varepsilon)\ln(1-\varepsilon)\big).$$

Now we show that for sufficiently small $\varepsilon > 0$, the generative complexity is strictly greater than excess entropy and strictly smaller than statistical complexity, i.e. $E(X^{\varepsilon}_{\mathbb{Z}}) < C_{\mathrm{hmm}}(X^{\varepsilon}_{\mathbb{Z}}) < C_{\mathfrak{C}}(X^{\varepsilon}_{\mathbb{Z}})$.



$0|\frac{1-\varepsilon}{2}$
$0|\frac{\varepsilon}{2}$
$1|\frac{\varepsilon}{2}$
$0|\varepsilon$
0
2
1
$1|\varepsilon$
$0|1-\varepsilon$
$1|1-\varepsilon$
$1|\frac{1-\varepsilon}{2}$

Fig. 3. HMM for the same Markov process as in Figure 2. The internal state entropy is lower, as node 2 carries nearly all weight: $\mu^{\varepsilon}(2) = 1 - \varepsilon$.

It is clear that no HMM can do with less than two internal states, but we can construct an HMM with lower internal state entropy on three states. The idea is to have one state corresponding to the i.i.d. process and getting most of the invariant measure when ε is small. The other two states correspond to the disturbances towards constantly 0 and 1 respectively. More precisely, let $\mathsf{S} := \{0, 1, 2\}$ and consider the stationary HMM given by the generator visualized in Figure 3 together with the invariant initial distribution $\mu^{\varepsilon}(s) = \begin{cases} 1-\varepsilon, & \text{if } s = 2 \\ \frac{\varepsilon}{2}, & \text{if } s \in \{0,1\} \end{cases}$. It is straightforward to verify that this HMM indeed generates $X^{\varepsilon}_{\mathbb{Z}}$. The internal state entropy is given by

$$H(\mu^{\varepsilon}) \;=\; -(1-\varepsilon)\ln(1-\varepsilon) - \varepsilon \ln(\tfrac{\varepsilon}{2}) \quad \xrightarrow{\varepsilon \to 0} \quad 0.$$

Thus it is smaller than $C_{\mathfrak{C}}(X_{\mathbb{Z}})$ for sufficiently small ε. On the other hand, any generative HMM has to take the "disturbance of magnitude ε" into account: It is easy to see that no single internal state can get greater invariant measure than $1 - \frac{\varepsilon}{2}$. Thus the generative complexity is lower bounded as follows:

$$C_{\text{hmm}}(X^{\varepsilon}_{\mathbb{Z}}) \;\geq\; L \;:=\; -(1-\tfrac{\varepsilon}{2})\ln(1-\tfrac{\varepsilon}{2}) - \tfrac{\varepsilon}{2}\ln(\tfrac{\varepsilon}{2}) \;\geq\; -\tfrac{\varepsilon}{2}\ln(\tfrac{\varepsilon}{2}).$$

This bound converges to zero slower than linearly in ε. Consequently, for sufficiently small ε, excess entropy cannot be achieved or approximated by entropies of generative HMMs. The different entropies are plotted in Figure 4. $\diamondsuit$

4 Predictive Interpretation of HMMs

We have seen that there can be a huge discrepancy between minimal sufficient memory and minimal generative HMM. The requirement of sufficiency is based on a certain understanding of "prediction". Here, we propose an alternative, weaker notion of prediction that allows for a predictive interpretation of all HMMs.

We model prediction by two steps: First the past $X_{-\mathbb{N}_0}$ is processed by a memory kernel mem, like in Section 2 but without the sufficiency assumption.

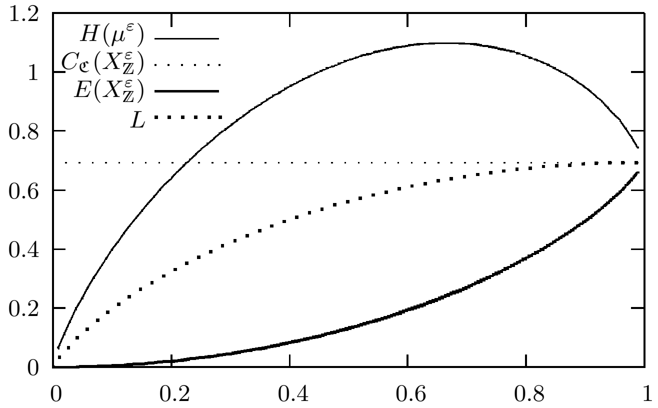


Fig. 4. Internal state entropy of the HMM of Figure 3, statistical complexity, excess entropy and the lower bound L for the generative complexity are plotted against the parameter ε. For $\varepsilon = 0$, all values are 0 and for $\varepsilon = 1$, all values are $\ln(2)$.

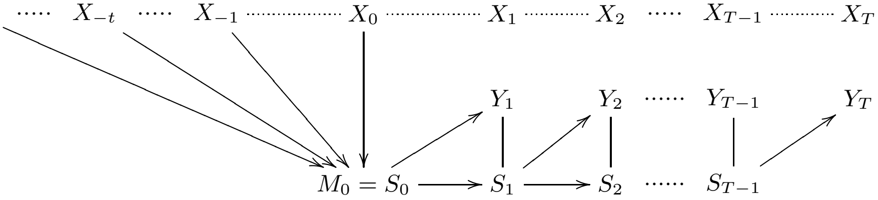


Fig. 5. The process of generating $Y_{\mathbb{N}}$ as prediction of $X_{\mathbb{N}}$. The dotted lines symbolize that $X_{\mathbb{Z}}$ may have arbitrary dependencies and need not be Markovian.

Then the actual prediction is done by generating a predicted future $Y_{\mathbb{N}}$. To this end we assume a generator gen, which uses the set of memory states as internal states and is initialized by the random memory state M_0 produced by mem. Thus gen, or rather the (non-invariant) HMM $\big(\mathsf{gen}, \mathsf{mem}(X_{-\mathbb{N}_0})\big)$, generates the prediction $Y_{\mathbb{N}}$ as in Section 3. The situation is illustrated as

$$\begin{tikzcd}
X_{-\mathbb{N}_0} \arrow[rr] \arrow[dr, "\mathsf{mem}"'] & & X_{\mathbb{N}} \\
& M_0 \arrow[r, "\mathsf{gen}'"] & Y_{\mathbb{N}}
\end{tikzcd}$$

where gen' is the kernel from $\mathsf{M} = \mathsf{S}$ to $\mathsf{D}^{\mathbb{N}}$ obtained by iterating gen and projecting to the output. Figure 5 shows the situation in more detail. The resulting joint conditional distribution is given by

$$P(S_{[0,T]} = s_{[0,T]},\ Y_{[1,T]} = y_{[1,T]} \mid X_{-\mathbb{N}_0} = x_{-\mathbb{N}_0}) \\ = \mathsf{mem}(x_{-\mathbb{N}_0};\, s_0) \prod_{k=1}^{T} \mathsf{gen}(s_{k-1};\, y_k, s_k), \qquad T \in \mathbb{N}.$$

Due to the intrinsic stochasticity, we cannot expect the prediction $Y_{\mathbb{N}}$ and the actual future $X_{\mathbb{N}}$ to coincide. But we require that the *distributions*, conditioned on the known past $X_{-\mathbb{N}_0}$, are identical. This is the best one can possibly do and means that actual and predicted future cannot be distinguished statistically, based on the observed past.

Definition 8. The pair $(\mathsf{mem}, \mathsf{gen})$ is called **predictive model** of $X_{\mathbb{Z}}$ if $\mathsf{mem}\colon \mathsf{D}^{-\mathbb{N}_0} \to \mathcal{P}(\mathsf{M})$ is measurable, $\mathsf{gen}\colon \mathsf{M} \to \mathcal{P}(\mathsf{D} \times \mathsf{M})$, and the process $Y_{\mathbb{N}}$ generated by the HMM $\big(\mathsf{gen}, \mathsf{mem}(X_{-\mathbb{N}_0})\big)$ satisfies

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

A memory kernel mem (resp. generator gen) is called **predictive** if there exists a generator gen (resp. memory mem) such that $(\mathsf{mem}, \mathsf{gen})$ is a predictive model.

If mem is a sufficient memory kernel and gen the associated generator constructed in Proposition 2, it is straightforward to see that $(\mathsf{mem}, \mathsf{gen})$ is a predictive model. Thus we obtain:

Proposition 9. *Sufficient memory kernels are predictive.*

One could say that a predictive memory is *sufficient for prediction*. Then, however, sufficiency for prediction does not imply sufficiency in the sense of statistics. In fact, predictive memories can be much smaller than any sufficient memory: Assume any generative HMM (gen, μ). We know from Example 4 that for certain processes the number of internal states can be substantially smaller than the number of causal states. But now we show that gen is predictive, i.e. the HMM induces a predictive model and thus in particular a predictive memory kernel mem. Of course mem is in general not sufficient but has only as few memory states as the generative HMM.

Proposition 10 (generative HMMs are predictive). *Let* (gen, μ) *be a generative HMM. Then* gen *is predictive, i.e. there is a memory kernel* mem*, such that* $(\mathsf{mem}, \mathsf{gen})$ *is a predictive model. More specifically, we can choose*

$$\mathsf{mem}(x_{-\mathbb{N}_0}) \;:=\; P(S_0 \mid Y_{-\mathbb{N}_0} = x_{-\mathbb{N}_0}).$$

If $(\mathsf{gen}_{\mathfrak{C}}, \mu_{\mathfrak{C}})$ is the ε-machine, then the memory kernel mem constructed in Proposition 10 recovers the causal state projection, i.e. $\mathsf{mem} = \mathsf{mem}_{\mathfrak{C}}$. In particular, this memory mem is deterministic. Of course, for general generative HMM, the associated memory need not be deterministic. Even more, it cannot be deterministic whenever the HMM is smaller than the corresponding ε-machine: In the following proposition we see that determinism implies sufficiency.

Proposition 11 (determinism implies sufficiency). *If* mem *is a predictive memory kernel and deterministic, i.e.* $\mathsf{mem} = \mathsf{mem}_f$ *for some measurable* $f\colon \mathsf{D}^{-\mathbb{N}_0} \to \mathsf{M}$*, then* mem *is sufficient. In particular,* $|\mathsf{M}| \geq |\mathsf{M}_{\mathfrak{C}}|$ *and* $H(M) \geq H(M_{\mathfrak{C}})$*.*

5 Summary and Discussion

There are two aspects of prediction: a memory which compresses the past to a set of memory states (such as the causal states) and an encoding of the mechanism of prediction (such as the ε-machine). Looking at the memory part, sufficiency is a natural requirement, which leads to minimality of causal states and ε-machine. Sufficiency is the central assumption of computational mechanics. It has to be stressed, however, that the ε-machine is not the minimal generative hidden Markov model. Analogously to statistical complexity $C_{\mathfrak{C}}(X_{\mathbb{Z}})$, we defined generative complexity $C_{\mathrm{hmm}}(X_{\mathbb{Z}})$ as size in terms of entropy of the minimal generative HMM and obtained that $E(X_{\mathbb{Z}}) \leq C_{\mathrm{hmm}}(X_{\mathbb{Z}}) \leq C_{\mathfrak{C}}(X_{\mathbb{Z}})$, where $E(X_{\mathbb{Z}})$ is the excess entropy. Furthermore, we gave an example, where both inequalities are strict. We proposed a different notion of "predictive" and compared it to the sufficiency requirement used in computational mechanics. According to our notion, it has to be possible to generate a prediction $Y_{\mathbb{N}}$ for the future with the same statistical properties as the real future $X_{\mathbb{N}}$, conditioned on the observed past, i.e. $P(Y_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) = P(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$. It turned out that predictive in this sense is strictly weaker than sufficient and that any generative HMM can be interpreted as predictive in our sense.

Extending the model class from sufficient to predictive includes models that are substantially smaller than the ε-machine. At the same time, it preserves a notion of predictive power which we consider quite natural. Nevertheless, we have to point out two drawbacks of our approach: Firstly, constructing a minimal HMM is intrinsically difficult, whereas efficient algorithms are available for the construction of the ε-machine from data. Secondly, and conceptually more important, the memory state is no longer a complete substitute for the past. Given a sufficient memory, the complete conditional future distribution that corresponds to an observation $x_{-\mathbb{N}_0}$ is encoded in a single memory state $m \in \mathsf{M}$. On the other hand, assume that we have observed a particular past $x_{-\mathbb{N}_0}$ and want to use a predictive model for sampling the conditional future distribution several times. We first choose a memory state m according to $\mathsf{mem}(x_{-\mathbb{N}_0})$ and then initialize gen with m for generating a prediction $y_{\mathbb{N}}$. We repeat this sampling procedure and obtain the correct future distribution $P(Y_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x_{-\mathbb{N}_0}) = P(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x_{-\mathbb{N}_0})$. But if we "forget" the history state $x_{-\mathbb{N}_0}$ and, instead of sampling new m's according to $\mathsf{mem}(x_{-\mathbb{N}_0})$, initialize gen always with the same m, the resulting distribution of $Y_{\mathbb{N}}$ can be different from $P(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x_{-\mathbb{N}_0})$. Thus, we have to memorize the *distribution* (the information state) $\mathsf{mem}(x_{-\mathbb{N}_0})$ of the initial memory states m. It is easy to show that the number of these information states is lower bounded by the number of causal states, because the map from history to information state defines a predictive deterministic memory, which is sufficient according to Proposition 11.

Currently, we do not know which of the two notions of prediction is more natural in which situations, and further steps towards revealing and comparing operational aspects of prediction are subject of our research.

References

1. Crutchfield, J.P., Young, K.: Inferring statistical complexity. Phys. Rev. Let. 63, 105–108 (1989)
2. Shalizi, C.R., Crutchfield, J.P.: Computational mechanics: Pattern and prediction, structure and simplicity. Journal of Statistical Physics 104, 817–879 (2001)
3. Löhr, W., Ay, N.: On the generative nature of prediction. Accepted for publication in Advances in Complex Systems (preprint, 2008), `http://www.mis.mpg.de/publications/preprints/2008/prepr2008-8.html`
4. Still, S., Crutchfield, J.P.: Optimal causal inference. Informal publication (2007), `http://arxiv.org/abs/0708.1580`
5. Grassberger, P.: Toward a quantitative theory of self-generated complexity. Int. J. Theor. Phys. 25, 907–938 (1986)
6. Bialek, W., Nemenman, I., Tishby, N.: Predictability, complexity, and learning. Neural Computation 13, 2409–2463 (2001)
7. Heller, A.: On stochastic processes derived from Markov chains. Annals of Mathematical Statistics 36, 1286–1291 (1965)
8. Crutchfield, J.P.: The calculi of emergence: Computation, dynamics and induction. Physica D 75, 11–54 (1994)