Observable operator models for discrete stochastic time series[1]

Herbert Jaeger
German National Research Center
for Information Technology (GMD)
AiS.BE
Schloss Birlinghoven, D-53754 Sankt Augustin
phone +49 2241 14 2253, fax +49 2241 14 2384
email: herbert.jaeger@gmd.de

August 7, 2000

[1] Appeared in Neural Computation 12(6), 1371-1398.

Abstract

A widely used class of models for stochastic systems is Hidden Markov models. Systems which can be modeled by hidden Markov models are a proper subclass of *linearly dependent processes*, a class of stochastic systems known from mathematical investigations carried out over the last four decades. This article provides a novel, simple characterization of linearly dependent processes, called *observable operator models*. The mathematical properties of observable operator models lead to a constructive learning algorithm for the identification of linearly dependent processes. The core of the algorithm has a time complexity of $O(N + nm^3)$, where N is the size of training data, n is the number of distinguishable outcomes of observations, and m is model state space dimension.

1 Introduction

Hidden Markov models (HMMs) (Bengio, 1996) of stochastic processes have been investigated mathematically long before they became a popular tool in speech processing (Rabiner, 1990) and engineering (Elliott, Aggoun, & Moore, 1995). A basic mathematical question was to decide when two HMMs are equivalent, i.e. describe the same distribution (of a stochastic process) (Gilbert, 1959). This problem was tackled by framing HMMs within a more general class of stochastic processes, now termed *linearly dependent processes* (LDPs). Deciding the equivalence of HMMs amounts to characterising HMM-describable processes as LDPs. This strand of research came to a successful conclusion in (Ito, Amari, & Kobayashi, 1992), where equivalence of HMMs was characterised algebraically, and a decision algorithm was provided. That article also gives an overview of the work done in this area.

It should be emphasized that linearly dependent processes are unrelated to linear systems in the standard sense, i.e. systems whose state sequences are generated by some linear operator (e.g., (Narendra, 1995)). The term, "linearly dependent processes", refers to certain linear relationships between conditional distributions that arise in the study of general stochastic processes. LDP's are thoroughly "nonlinear" in the standard sense of the word.

The class of LDPs has been characterized in various ways. The most concise description was developed in (Heller, 1965), using methods from category theory and algebra. This approach was taken up and elaborated in a recent comprehensive account on LDPs and the mathematical theory of HMMs, viewed as a subclass of LDPs (Ito, 1992).

All of this work on HMMs and LDPs was mathematically oriented, and did not bear on the practical question of learning models from data.

In the present article, I develop an alternative, simpler characterization of LDPs, called *observable operator models* (OOMs). OOMs require only concepts from elementary linear algebra. The linear algebra nature of OOMs gives rise to a constructive learning procedure, which makes it possible to estimate models from data very efficiently.

The name, "observable operator models", arises from the very way in which stochastic trajectories are mathematically modeled in this approach.

Traditionally, trajectories of discrete-time systems are seen as a sequence of states, which is generated by the repeated application of a single (possibly stochastic) operator T (fig. 1a). Metaphorically speaking, a trajectory is seen as a sequence of *locations* in state space, which are visited by the system due to the action of a time step operator.

In OOM theory, trajectories are perceived in a complementary fashion. From a set of operators (say, $\{T_A, T_B\}$), one operator is stochastically selected for application at every time step. The system trajectory is then identified with the sequence of operators. Thus, an observed piece of trajectory $\ldots ABAA \ldots$ would correspond to a concatenation of operators $\ldots T_A(T_A(T_B(T_A \ldots))) \ldots$ (fig. 1b). The fact that the observables are the operators themselves, led to the naming of this kind of stochastic models. An appropriate metaphor would be to view trajectories as sequences of *actions*.

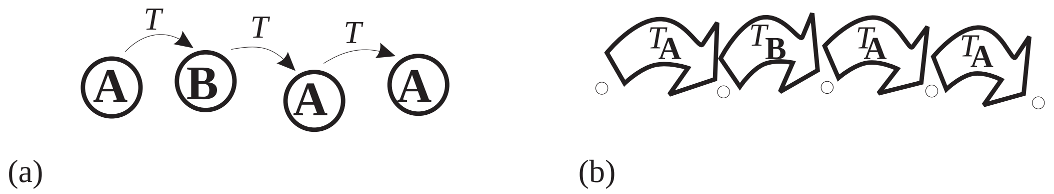

Figure 1: (a) The standard view of trajectories. A time step operator T yields a sequence ABAA of states. (b) The OOM view. Operators T_A, T_B are concatenated to yield a sequence ABAA of observables.

Stochastic sequences of operators are a well-known object of mathematical investigation (Iosifescu & Theodorescu, 1969). OOM theory grows out of the novel insight that the probability of selecting an operator can be computed *using the operator itself.*

The sections of this paper cover the following topics: (2) how a matrix representation of OOMs can be construed as a generalization of HMMs, (3) how OOMs are used to generate and predict stochastic time series, (4) how an abstract information-theoretic version of OOMs can be obtained from any stochastic process, (5) how these abstract OOMs can be used to prove a fundamental theorem which reveals when two OOMs in matrix representation are equivalent, (6) that some low-dimensional OOMs can model processes which can be modeled either only by arbitrarily high-dimensional HMMs, or by none at all; and that one can model a conditional rise and fall of probabilities in processes timed by "probability oscillators", (7) how one can use the fundamental equivalence theorem to obtain OOMs whose state space dimensions can be interpreted as probabilities of certain future events, and (8) how these interpretable OOMs directly yield a constructive procedure to estimate OOMs from data. (9) gives a brief conclusion.

2 From HMMs to OOMs

In this section, OOMs are introduced by generalization from HMMs. In this way it becomes immediately clear why the latter are a subclass of the former.

A basic HMM specifies the distribution of a discrete-time, stochastic process $(Y_t)_{t \in \mathbb{N}}$, where the random variables Y_t have finitely many outcomes from a set $\mathcal{O} = \{a_1, \ldots, a_n\}$. HMMs can be defined in several equivalent ways. In this article we will adhere to the definition which is customary in the speech recognition community and other application areas. The specification is done in two stages.

First, a Markov chain $(X_t)_{t \in \mathbb{N}}$ produces sequences of *hidden* states from a finite state set $\{s_1, \ldots, s_m\}$. Second, when the Markov chain is in state s_j

at time t, it "emits" an observable outcome, with a time-invariant probability $P[Y_t = a_i \,|\, X_t = s_j]$.

Figure 2 presents an exemplary HMM.

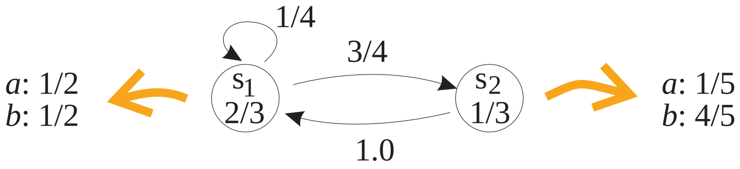

Figure 2: A HMM with two hidden states s_1, s_2 and two outcomes $\mathcal{O} = \{a, b\}$. Fine arrows indicate admissible hidden state transitions, with their corresponding probabilities marked besides. The initial state distribution $(2/3, 1/3)$ of the Markov chain is indicated inside the state circles. Emission probabilities $P[a_i \,|\, s_j]$ of outcomes are annotated besides bold grey arrows.

Formally, the state transition probabilities can be collected in a $m \times m$ matrix M which at place (i, j) contains the transition probability from state s_i to s_j (i.e., M is a Markov matrix, or *stochastic* matrix). For every $a \in \mathcal{O}$, we collect the emission probabilities $P[Y_t = a \,|\, X_t = s_j]$ in a diagonal *observation matrix* O_a of size $m \times m$. O_a contains, in its diagonal, the probabilities $P[Y_t = a \,|\, X_t = s_1], \ldots, P[Y_t = a \,|\, X_t = s_m]$. For the example from fig. 2, this gives

$$M = \begin{pmatrix} 1/4 & 3/4 \\ 1 & 0 \end{pmatrix}, \quad O_a = \begin{pmatrix} 1/2 & \\ & 1/5 \end{pmatrix}, \quad O_b = \begin{pmatrix} 1/2 & \\ & 4/5 \end{pmatrix}. \tag{1}$$

In order to fully characterize a HMM, one also must supply an initial distribution $w_0 = (P[X_0 = s_1], \ldots, P[X_0 = s_m])^\top$ (superscript $\cdot^\top$ denotes transpose of vectors and matrices. Vectors are assumed to be column vectors throughout this article, unless noted otherwise). The process described by the HMM is stationary if w_0 is an invariant distribution of the Markov chain, i.e. if it satisfies

$$M^\top w_0 = w_0. \tag{2}$$

See (Doob, 1953) for details on Markov chains.

It is well-known that the matrices M and O_a ($a \in \mathcal{O}$) can be used to compute the probability of finite-length observable sequences. Let $\mathbf{1} = (1, \ldots, 1)$ denote the m-dimensional row vector of units, and let $T_a := M^\top O_a$. Then the probability to observe the sequence $a_{i_0} \ldots a_{i_k}$ among all possible sequences of length $k+1$ is equal to the number obtained by applying $T_{a_{i_0}}, \ldots, T_{a_{i_k}}$ to w_0, and summing up the components of the resulting vector by multiplying it with $\mathbf{1}$:

$$P[a_{i_0} \ldots a_{i_k}] = \mathbf{1} T_{a_{i_k}} \cdots T_{a_{i_0}} w_0. \tag{3}$$

The term $P[a_{i_0} \dots a_{i_k}]$ in (3) is a shorthand notation for $P[X_0 = a_{i_0}, \dots, X_k = a_{i_k}]$, which will be used throughout this article.

(3) is a matrix notation of the well-known "forward algorithm" for determining probabilities of observation sequences in HMMs. Proofs of (3) may be found e.g. in (Ito et al., 1992) and (Ito, 1992).

M can be recovered from the operators T_a by observing that

$$M^\top = M^\top \cdot \mathbf{id} = M^\top(O_{a_1} + \cdots + O_{a_n}) = T_{a_1} + \cdots + T_{a_n}. \tag{4}$$

Eq. (3) shows that the distribution of the process (Y_t) is specified by the operators T_{a_i} and the vector w_0. Thus, the matrices T_{a_i} and w_0 contain the same information as the original HMM specification in terms of M, O_{a_i}, and w_0. I.e., one can rewrite a HMM as a structure $(\mathbb{R}^m, (T_a)_{a \in \mathcal{O}}, w_0)$, where $\mathbb{R}^m$ is the domain of the operators T_a. The HMM from the example, written in this way, becomes

$$\mathcal{M} = (\mathbb{R}^2, (T_a, T_b), w_0) = (\mathbb{R}^2, (\begin{pmatrix} 1/8 & 1/5 \\ 3/8 & 0 \end{pmatrix}, \begin{pmatrix} 1/8 & 4/5 \\ 3/8 & 0 \end{pmatrix}), (2/3, 1/3)^\top). \tag{5}$$

Now, one arrives at the definition of an OOM by (i) relaxing the requirement that $M^\top$ be the transpose of a stochastic matrix, to the weaker requirement that the columns of M^T each sum to 1, and by (ii) requiring from w_0 merely that it has a component sum of 1. That is, negative entries are now allowed in matrices and vectors, which are forbidden in stochastic matrices and probability vectors. Using the letter τ in OOMs in places where T appears in HMMs, and introducing $\mu = \sum_{a \in \mathcal{O}} \tau_a$ in analogy to (4), this yields:

Definition 1 *A m-dimensional OOM is a triple $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in \mathcal{O}}, w_0)$, where $w_0 \in \mathbb{R}^m$ and $\tau_a : \mathbb{R}^m \mapsto \mathbb{R}^m$ are linear operators, satisfying*

1. *$\mathbf{1}w_0 = 1$,*
2. *$\mu = \sum_{a \in \mathcal{O}} \tau_a$ has column sums equal to 1,*
3. *for all sequences $a_{i_0} \dots a_{i_k}$ it holds that $\mathbf{1}\tau_{a_{i_k}} \cdots \tau_{a_{i_0}} w_0 \geq 0$.*

Conditions 1 and 2 reflect the relaxations (i) and (ii) mentioned previously, while condition 3 ensures that one obtains non-negative values when the OOM is used to compute probabilities. Unfortunately, condition 3 is useless for deciding or constructing OOMs. An alternative to condition 3, which is suitable for constructing OOMs, will be introduced in Section 6.

Since function concatenations of the form $\tau_{a_{i_k}} \circ \cdots \circ \tau_{a_{i_0}}$ will be used very often in the sequel, we introduce a shorthand notation for handling sequences of symbols. Following the conventions of formal language theory, we shall denote the empty sequence by ε (i.e., the sequence of length 0 which "consists" of no symbol at all), the set of all sequences of length k of symbols from $\mathcal{O}$, by $\mathcal{O}^k$; $\bigcup_{k \geq 1} \mathcal{O}^k$ by $\mathcal{O}^+$; and the set $\{\varepsilon\} \cup \mathcal{O}^+$ by $\mathcal{O}^*$. We shall write $\bar{a} \in \mathcal{O}^+$ to denote any finite sequence $a_0 \dots a_n$, and $\tau_{\bar{a}}$ to denote $\tau_{a_n} \circ \cdots \circ \tau_{a_0}$.

An OOM, as defined here, specifies a stochastic process, if one makes use of an analog of (3):

Proposition 1 *Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in \mathcal{O}}, w_0)$ be an OOM according to the previous definition. Let $\Omega = \mathcal{O}^\infty$ be the set of all infinite sequences over $\mathcal{O}$, and $\mathfrak{A}$ be the σ-algebra generated by all finite-length initial events on Ω. Then, if one computes the probabilities of initial finite-length events in the following way:*

$$P_0[\bar{a}] := \mathbf{1}\tau_{\bar{a}} w_0, \tag{6}$$

the numerical function P_0 can be uniquely extended to a probability measure P on $(\Omega, \mathfrak{A})$, giving rise to a stochastic process $(\Omega, \mathfrak{A}, P, (X_t)_{t \in \mathbb{N}})$, where $X_n(a_1 a_2 \ldots) = a_n$. If w_0 is an invariant vector of μ, i.e., if $\mu w_0 = w_0$, the process is stationary.

The proof is given in appendix A.

Since we introduced OOMs here by generalizing from HMMs, it is clear that every process whose distribution can be specified by a HMM can also be characterized by an OOM.

I conclude this section with a remark on LDPs and OOMs. It is known that the distributions of LDPs can be characterized through matrix multiplications in a fashion which is very similar to (6) (cf. (Ito, 1992), theorem 1.8):

$$P[a_{i_0} \ldots a_{i_k}] = \mathbf{1} Q I_{a_{i_k}} \ldots Q I_{a_{i_0}} w_0. \tag{7}$$

The matrix Q does not depend on a, while the "projection matrices" I_a do. If one puts $Q = \mathbf{id}, I_a = \tau_a$, one easily sees that the class of processes which can be described by OOMs is the class of LDPs.

3 OOMs as generators and predictors

This section explains how to generate and predict the paths of a process $(X_t)_{t \in \mathbb{N}}$, whose distribution is specified by an OOM $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a \in \mathcal{O}}, w_0)$. We describe the procedures mathematically and illustrate them with an example.

More precisely, the generation task consists in randomly producing, at times $t = 0, 1, 2, \ldots$, outcomes $a_{i_0}, a_{i_1}, a_{i_2}, \ldots$, such that (i) at time $t = 0$, the probability of producing b is equal to $\mathbf{1}\tau_b w_0$ according to (6), and (ii) at every time step $t > 0$, the probability of producing b (after $a_{i_0}, \ldots, a_{i_{t-1}}$ have already been produced) is equal to $P[X_t = b \,|\, X_0 = a_{i_0}, \ldots, X_{t-1} = a_{i_{t-1}}]$. Using (6), the latter amounts to calculating at time t, for every $b \in \mathcal{O}$, the conditional probability

$$\begin{aligned} & P[X_t = b \,|\, X_0 = a_{i_0}, \ldots, X_{t-1} = a_{i_{t-1}}] \\ = \quad & \frac{P[X_0 = a_{i_0}, \ldots, X_{t-1} = a_{i_{t-1}}, X_t = b]}{P[X_0 = a_{i_0}, \ldots, X_{t-1} = a_{i_{t-1}}]} \\ = \quad & \mathbf{1}\tau_b \tau_{a_{i_{t-1}}} \cdots \tau_{a_{i_0}} w_0 / \mathbf{1}\tau_{a_{i_{t-1}}} \cdots \tau_{a_{i_0}} w_0 \end{aligned}$$

$$\begin{aligned} &= \mathbf{1}\tau_b(\frac{\tau_{a_{i_{t-1}}}\cdots\tau_{a_{i_0}}w_0}{\mathbf{1}\tau_{a_{i_{t-1}}}\cdots\tau_{a_{i_0}}w_0}) \\ &=: \mathbf{1}\tau_b w_t, \qquad (8) \end{aligned}$$

and producing at time t the outcome b with this conditional probability (the symbol =: denotes that the term on the rhs. is being defined by the equation). Calculations of (8) can be carried out incrementally, if one observes that for $t \geq 1$, w_t can be computed from w_{t-1}:

$$w_t = \frac{\tau_{a_{t-1}}w_{t-1}}{\mathbf{1}\tau_{a_{t-1}}w_{t-1}}. \qquad (9)$$

Note that all vectors w_t thus obtained have a component sum equal to 1.

Observe that the operation $\mathbf{1}\tau_b\cdot$ in (8) can be done effectively by pre-computing the vector $v_b := \mathbf{1}\tau_b$. Computing (8) then amounts to multiplying the (row) vector v_b with the (column) vector w_t, or, equivalently, it amounts to evaluating the inner product $< v_b, w_t >$.

The prediction task is completely analogous to the generation task. Given an initial realization $a_{i_0}, \ldots, a_{i_{t-1}}$ of the process up to time $t-1$, one has to calculate the probability by which an outcome b is going to occur at the next time step t. This is again an instance of (8), the only difference being that now the initial realization is not generated by oneself but is externally given.

Many-time-step probability predictions of collective outcomes can be calculated by evaluating inner products, too. Let the collective outcome $A = \{\bar{b}_1, \ldots, \bar{b}_n\}$ consist of n sequences of length $s+1$ of outcomes (i.e., outcome A is recorded when any of the sequences $\bar{b}_i$ occurs). Then, the probability that A is going to occur after an initial realization $\bar{a}$ of length $t-1$, can be computed as follows:

$$\begin{aligned} &P[(X_t, \ldots, X_{t+s}) \in A \,|\, (X_0, \ldots, X_{t-1}) = \bar{a}] \\ &\quad= \sum_{\bar{b}\in A} P[(X_t, \ldots, X_{t+s}) = \bar{b} \,|\, (X_0, \ldots, X_{t-1}) = \bar{a}] \\ &\quad= \sum_{\bar{b}\in A} \mathbf{1}\tau_{\bar{b}} w_t \quad =: \quad \sum_{\bar{b}\in A} < v_{\bar{b}}, w_t > \\ &\quad= < \sum_{\bar{b}\in A} v_{\bar{b}}, w_t > \quad =: \quad < v_A, w_t > . \qquad (10) \end{aligned}$$

If one wants to calculate the future probability of a collective outcome A repeatedly, utilization of (10) reduces computational load considerably because the vector v_A needs to be (pre-)computed only once.

The generation procedure shall now be illustrated using the exemplary OOM $\mathcal{M}$ from (5). We first compute the vectors v_a, v_b:

$$v_a \quad = \quad \mathbf{1}\tau_a \quad = \quad \mathbf{1}\begin{pmatrix} 1/8 & 1/5 \\ 3/8 & 0 \end{pmatrix} \quad = \quad (1/2, 1/5),$$

$$v_b \quad = \quad \mathbf{1}\tau_b \quad = \quad \mathbf{1}\begin{pmatrix} 1/8 & 4/5 \\ 3/8 & 0 \end{pmatrix} \quad = \quad (1/2, 4/5).$$

Starting with $w_0 = (2/3, 1/3)$, we obtain probabilities $< v_a, w_0 > \; = 2/5, < v_b, w_0 > \; = 3/5$ of producing a vs. b at the first time step. We make a random decision for a vs. b, weighted according to these probabilities. Let's assume the dice fall for b. We now compute $w_1 = \tau_b w_0 / \mathbf{1}\tau_b w_0 = (7/12, 5/12)^{\mathsf{T}}$. For the next time step, we repeat these computations with w_1 in place of w_0, etc., etc.

4 From stochastic processes to OOMs

This section introduces OOMs again, but this time in a top-down fashion, starting from general stochastic processes. This alternative route clarifies the fundamental nature of observable operators. Furthermore, the insights obtained in this section will yield a short and instructive proof of the central theorem of OOM equivalence, to be presented in the next section. The material presented here is not required after the next section and may be skipped by readers with not so keen an interest in probability theory.

In Section 2, we have described OOMs as structures $(\mathbb{R}^m, (\tau_a)_{a\in\mathcal{O}}, w_0)$. In this section, we will arrive at isomorphic structures $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\mathcal{O}}, \mathfrak{g}_\varepsilon)$, where again $\mathfrak{G}$ is a vector space, $(\mathfrak{t}_a)_{a\in\mathcal{O}}$ is a family of linear operators on $\mathfrak{G}$, and $\mathfrak{g}_\varepsilon \in \mathfrak{G}$. However, the vector space $\mathfrak{G}$ is now a space of certain numerical prediction functions. In order to discriminate OOMs characterized on spaces $\mathfrak{G}$ from the "ordinary" OOMs, we shall call $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\mathcal{O}}, \mathfrak{g}_\varepsilon)$ an *predictor-space OOM.*

Let $(X_t)_{t\in\mathbb{N}}$, or for short, (X_t) be a discrete-time stochastic process with values in a finite set $\mathcal{O}$. Then, the distribution of (X_t) is uniquely characterized by the probabilities of finite initial subsequences, i.e. by all probabilities of the kind $P[\bar{a}]$, where $\bar{a} \in \mathcal{O}^+$.

We introduce a shorthand for conditional probabilities, by writing $P[\bar{a} \mid \bar{b}]$ for $P[(X_n, \ldots, X_{n+s}) = \bar{a} \mid (X_0, \ldots, X_{n-1}) = \bar{b}]$. We shall formally write the unconditional probabilities as conditional probabilities, too, with the empty condition ε, i.e. we use the notation $P[\bar{a} \,|\, \varepsilon] := P[(X_0 \ldots X_s) = \bar{a}] = P[\bar{a}]$.

Thus, the distribution of (X_t) is also uniquely characterized by its *conditional continuation probabilities*, i.e. by the conditional probabilities $P[\bar{a} \mid \bar{b}]$, where $\bar{a} \in \mathcal{O}^+, \bar{b} \in \mathcal{O}^*$.

For every $\bar{b} \in \mathcal{O}^*$, we collect all conditioned continuation probabilities of $\bar{b}$ into a numerical function

$$\begin{aligned} \mathfrak{g}_{\bar{b}} : \mathcal{O}^+ \;&\rightarrow\; \mathbb{R}, \\ \bar{a} \;&\mapsto\; P[\bar{a}\,|\,\bar{b}], \text{ if } P[\bar{b}] \neq 0 \\ &\mapsto\; 0, \text{ if } P[\bar{b}] = 0. \end{aligned} \tag{11}$$

The set $\{\mathfrak{g}_{\bar{b}} \mid \bar{b} \in \mathcal{O}^*\}$ uniquely characterizes the distribution of (X_t), too. Intuitively, a function $\mathfrak{g}_{\bar{b}}$ describes the future distribution of the process after an initial realization $\bar{b}$.

Let $\mathfrak{D}$ denote the set of all functions from $\mathcal{O}^+$ into the reals, i.e. the numerical functions on non-empty sequences. $\mathfrak{D}$ canonically becomes a real vector space if one defines scalar multiplication and vector addition as follows: for $\mathfrak{d}_1, \mathfrak{d}_2 \in \mathfrak{D}$, $\alpha, \beta \in \mathbb{R}$, $\bar{a} \in \mathcal{O}^+$ put $(\alpha\mathfrak{d}_1 + \beta\mathfrak{d}_2)(\bar{a}) := \alpha(\mathfrak{d}_1(\bar{a})) + \beta(\mathfrak{d}_2(\bar{a}))$.

Let $\mathfrak{G} = \langle\{\mathfrak{g}_{\bar{b}} \mid \bar{b} \in \mathcal{O}^*\}\rangle_{\mathfrak{D}}$ denote the linear subspace spanned in $\mathfrak{D}$ by the conditioned continuations. Intuitively, $\mathfrak{G}$ is the (linear closure of the) space of future distributions of the process (X_t).

Now we are halfway done with our construction of $(\mathfrak{G}, (\mathfrak{t}_a)_{a \in \mathcal{O}}, \mathfrak{g}_\varepsilon)$: we have constructed the vector space $\mathfrak{G}$, which corresponds to $\mathbb{R}^m$ in the "ordinary" OOMs from Section 2, and we have defined the initial vector $\mathfrak{g}_\varepsilon$, which is the counterpart of w_0. It remains for us to define the family of observable operators.

In order to specify a linear operator on a vector space, it suffices to specify the values the operator takes on a basis of the vector space. Choose $\mathcal{O}_0^* \subseteq \mathcal{O}^*$ such that the set $\{\mathfrak{g}_{\bar{b}} | \bar{b} \in \mathcal{O}_0^*\}$ is a basis of $\mathfrak{G}$. Define, for every $a \in \mathcal{O}$, a linear function $\mathfrak{t}_a : \mathfrak{G} \to \mathfrak{G}$ by putting

$$\mathfrak{t}_a(\mathfrak{g}_{\bar{b}}) := P[a \,|\, \bar{b}]\mathfrak{g}_{\bar{b}a} \tag{12}$$

for all $\bar{b} \in \mathcal{O}_0^*$ ($\bar{b}a$ denotes the concatenation of the sequence $\bar{b}$ with a). It turns out that (12) carries over from basis elements $\bar{b} \in \mathcal{O}_0^*$ to all $\bar{b} \in \mathcal{O}^*$:

Proposition 2 *For all $\bar{b} \in \mathcal{O}^*$, $a \in \mathcal{O}$, the linear operator $\mathfrak{t}_a$ satisfies the condition*

$$\mathfrak{t}_a(\mathfrak{g}_{\bar{b}}) = P[a \,|\, \bar{b}]\mathfrak{g}_{\bar{b}a}. \tag{13}$$

The proof is given in appendix B. Intuitively, the operator $\mathfrak{t}_a$ describes the change of knowledge about a process due to an incoming observation of a. More precisely, assume that the process has initially been observed up to time n. That is, an initial observation $\bar{b} = b_0 \ldots b_n$ has been made. Our knowledge about the state of the process at this moment is tantamount to the predictor function $\mathfrak{g}_{\bar{b}}$. Then assume that at time $n+1$ an outcome a is observed. After that, our knowledge about the process state is then expressed by $\mathfrak{g}_{\bar{b}a}$. But this is (up to scaling by $P[a \,|\, \bar{b}]$) just the result of applying $\mathfrak{t}_a$ to the old state, $\mathfrak{g}_{\bar{b}}$.

The operators $(\mathfrak{t}_a)_{a \in \mathcal{O}}$ are the analog of the observable operators $(\tau_a)_{a \in \mathcal{O}}$ in OOMs and can likewise be used to compute probabilities of finite sequences:

Proposition 3 *Let $\{\mathfrak{g}_{\bar{b}} \mid \bar{b} \in \mathcal{O}_0^*\}$ be a basis of $\mathfrak{G}$. Let $\bar{a} := a_{i_0} \ldots a_{i_k}$ be an initial realization of (X_t) of length $k+1$. Let $\sum_{i=1,\ldots,n} \alpha_i \mathfrak{g}_{\bar{b}_i} = \mathfrak{t}_{a_{i_k}} \ldots \mathfrak{t}_{a_{i_0}} \mathfrak{g}_\varepsilon$ be the linear combination of $\mathfrak{t}_{a_{i_k}} \ldots \mathfrak{t}_{a_{i_0}} \mathfrak{g}_\varepsilon$ from basis vectors. Then it holds that*

$$P[\bar{a}] = \sum_{i=1,\ldots,n} \alpha_i. \tag{14}$$

Note that (14) is valid for any basis $\{\mathfrak{g}_{\bar{b}} \mid \bar{b} \in \mathcal{O}_0^*\}$. The proof can be found in appendix C. (14) corresponds exactly to (6), since left-multiplication of a vector with $\mathbf{1}$ amounts to summing the vector components, which in turn are the coefficients of that vector w.r.t. a vector space basis.

Due to (14), the distribution of the process (X_t) is uniquely characterized by the observable operators $(\mathfrak{t}_a)_{a \in \mathcal{O}}$. Conversely, these operators are uniquely defined by the distribution of (X_t). I.e., the following definition makes sense:

Definition 2 *Let $(X_t)_{t \in \mathbb{N}}$ be a stochastic process with values in a finite set $\mathcal{O}$. The structure $(\mathfrak{G}, (\mathfrak{t}_a)_{a \in \mathcal{O}}, \mathfrak{g}_\varepsilon)$ is called* the *predictor-space observable operator model of the process. The vector space dimension of $\mathfrak{G}$ is called the dimension of the process and is denoted by dim(X_t).*

I remarked in the introduction that stochastic processes have previously been characterized in terms of vector spaces. Although the vector spaces were constructed in other ways than $\mathfrak{G}$, they lead to equivalent notions of process dimension. (Heller, 1965) called finite-dimensional (in our sense) stochastic processes *finitary*; in (Ito et al., 1992) the process dimension (if finite) was called *minimum effective degree of freedom*.

(13) clarifies the fundamental character of observable operators: $\mathfrak{t}_a$ describes how the knowledge about the process's future after an observe past $\bar{b}$ (i.e., the predictor function $\mathfrak{g}_{\bar{b}}$) changes through an observation of a. The power of the observable operator idea lies in the fact that these operators turn out to be linear (proposition 2). I have only treated the discrete time, discrete value case here. However, predictor-space OOMs can be defined in a similar way also for continuous-time, arbitrary-valued processes (sketch in (?)). It turns out that in those cases, the resulting observable operators are linear, too. In the sense of updating predictor functions, the change of knowledge about a process due to incoming observations is a linear phenomenon.

In the remainder of this section, I describe how the dimension of a process is related to the dimensions of ordinary OOMs of that process.

Proposition 4

1. *If (X_t) is a process with finite dimension m, then an m-dimensional ordinary OOM of this process exists.*

2. *A process (X_t) whose distribution is described by a k-dimensional OOM $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a \in \mathcal{O}}, w_0)$ has a dimension $m \leq k$.*

Thus, if a process (X_t) has dimension m, and we have a k-dimensional OOM $\mathcal{A}$ of (X_t), we know that a m-dimensional OOM $\mathcal{A}'$ exists which is equivalent to $\mathcal{A}$ in the sense of specifying the same distribution. Furthermore, $\mathcal{A}'$ is minimal-dimensional in its equivalence class. A minimal-dimensional OOM $\mathcal{A}'$ can be constructively obtained from $\mathcal{A}$ in several ways, all of which amount to an implicit construction of the predictor-space OOM of the process specified by $\mathcal{A}$. Since the learning algorithm presented in later sections can be used for this construction, too, I do not present a dedicated procedure for obtaining minimal-dimensional OOMs here.

5 Equivalence of OOMs

Given two OOMs $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a \in \mathcal{O}}, w_0), \mathcal{B} = (\mathbb{R}^l, (\tau'_a)_{a \in \mathcal{O}}, w'_0)$, when are they *equivalent* in the sense that they describe the same distribution? This question can be answered using the insights gained in the previous section.

First, construct minimal-dimensional OOMs $\mathcal{A}', \mathcal{B}'$ which are equivalent to $\mathcal{A}$ and $\mathcal{B}$, respectively. If the dimensions of $\mathcal{A}', \mathcal{B}'$ are not equal, then $\mathcal{A}$ and $\mathcal{B}$ are not equivalent. We can therefore assume that the two OOMs whose equivalence we wish to ascertain have the same (and minimal) dimension. Then, the answer to our question is given in the following proposition:

Proposition 5 *Two minimal-dimensional OOMs $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in \mathcal{O}}, w_0)$, $\mathcal{B} = (\mathbb{R}^m, (\tau'_a)_{a \in \mathcal{O}}, w'_0)$ are equivalent iff there exists a bijective linear map $\varrho : \mathbb{R}^m \to \mathbb{R}^m$, satisfying the following conditions:*

1. $\varrho(w_0) = w'_0$,
2. $\tau'_a = \varrho \tau_a \varrho^{-1}$ *for all* $a \in \mathcal{O}$,
3. $\mathbf{1}v = \mathbf{1}\varrho v$ *for all (column) vectors* $v \in \mathbb{R}^m$.

Sketch of proof: $\Leftarrow$: trivial. $\Rightarrow$: We have done all the hard work in the previous section! Let $\sigma_{\mathcal{A}}, \sigma_{\mathcal{B}}$ be the canonical projections from $\mathcal{A}, \mathcal{B}$ on the predictor-space OOM of the process specified by $\mathcal{A}$ (and hence by $\mathcal{B}$). Observe that $\sigma_{\mathcal{A}}, \sigma_{\mathcal{B}}$ are bijective linear maps which preserve the component sum of vectors. Define $\varrho := \sigma_{\mathcal{B}}^{-}1 \circ \sigma_{\mathcal{A}}$. Then, *(1)* follows from $\sigma_{\mathcal{A}}(w_0) = \sigma_{\mathcal{B}}(w'_0) = \mathfrak{g}_\varepsilon$, *(2)* follows from $\forall \bar{c} \in \mathcal{O}^+ : \ \sigma(\tau_{\bar{c}} w_0) = \sigma(\tau'_{\bar{c}} w_0) = P[\bar{c}]\mathfrak{g}_{\bar{c}}$, and *(3)* from the fact that $\sigma_{\mathcal{A}}, \sigma_{\mathcal{B}}$ preserve component sum of vectors.

6 A non-HMM linearly dependent process

The question of when a LDP can be captured by a HMM has been fully answered in the literature (original result in (Heller, 1965), refinements in (Ito, 1992)), and examples of non-HMM LDPs have been given. I briefly restate the results, and then elaborate on a simple example of such a process. It was first described in a slightly different version in (Heller, 1965). The aim is to provide an intuitive insight in which sense the class of LDPs is "larger" than the class of processes which can be captured by HMMs.

Characterizing HMMs as LDPs heavily draws on the theory of convex cones and non-negative matrices. I first introduce some concepts, following the notation of a standard textbook (Berman & Plemmons, 1979).

With a set $S \subseteq \mathbb{R}^n$ we associate the set S^G, the *set generated by* S, which consists of all finite nonnegative linear combinations of elements of S. A set $K \subseteq \mathbb{R}^n$ is defined to be a *convex cone* if $K = K^G$. A convex cone K^G is called *n-polyhedral* if K has n elements. A cone K is *pointed* if for every nonzero $v \in K$, the vector $-v$ is not in K. A cone is *proper* if it is pointed, closed, and its interior is not empty.

Using these concepts, the following theorem in (*a1*), (*a2*) gives two conditions which individually are equivalent to condition 3 in definition 1, and (*b*) refines condition (*a1*) for determining when an OOM is equivalent to a HMM. Finally, (*c*) states necessary conditions which every τ_a in an OOM must satisfy.

Proposition 6 *(a1) Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\mathcal{O}}, w_0)$ be a structure consisting of linear maps $(\tau_a)_{a\in\mathcal{O}}$ on $\mathbb{R}^m$ and a vector $w_0 \in \mathbb{R}^m$. Let $\mu := \sum_{a\in\mathcal{O}} \tau_a$. Assume that the first two conditions from definition 1 hold, i.e. $\mathbf{1}w_0 = 1$ and μ has column sums equal to 1. Then $\mathcal{A}$ is an OOM if and only if there exist pointed convex cones $(K_a)_{a\in\mathcal{O}}$ satisfying the following conditions:*

1. *$\mathbf{1}v \geq 0$ for all $v \in K_a$ (where $a \in \mathcal{O}$),*
2. *$w_0 \in (\bigcup_{a\in\mathcal{O}} K_a)^G$,*
3. *$\forall a, b \in \mathcal{O} : \tau_b K_a \subseteq K_b$.*

(a2) Using the same assumptions as before, $\mathcal{A}$ is an OOM if and only if there exists a pointed convex cone K satisfying the following conditions:

1. *$\mathbf{1}v \geq 0$ for all $v \in K$,*
2. *$w_0 \in K$,*
3. *$\forall a \in \mathcal{O} : \tau_a K \subseteq K$.*

(b) Assume that $\mathcal{A}$ is an OOM. Then there exists a hidden Markov model equivalent to $\mathcal{A}$ if and only if a pointed convex cone K according to condition (a2) exists which is n-polyhedral for some n. n can be selected such that it is not greater than the minimal state number for HMMs equivalent to $\mathcal{A}$.

(c) Let $\mathcal{A}$ be a minimal-dimensional OOM, and τ_a be one of its observable operators, and K be a cone according to (a2). Then (i) the spectral radius $\varrho(\tau_a)$ of τ_a is an eigenvalue of τ_a, (ii) the degree of $\varrho(\tau_a)$ is greater or equal to the degree of any other eigenvalue λ with $\mid \lambda \mid = \varrho(\tau_a)$, and (iii) an eigenvector of corresponding to $\varrho(\tau_a)$ lies in K. (The degree of an eigenvalue λ of a matrix is the size of the largest diagonal block in the Jordan canonical form of the matrix, which contains λ).

Notes on the proof. The proof of parts (*a1*) and (*b*) go back to (Heller, 1965) and have been reformulated in (Ito, 1992)[1]. The equivalence of (*a1*) with (*a2*) is an easy exercise. The conditions collected in (*c*) are equivalent to the statement that $\tau_a K \subseteq K$ for some proper cone K (proof in theorems 3.2 and 3.5 in (Berman & Plemmons, 1979)). It is easily seen that for a minimal-dimensional OOM, the cone K required in (*a2*) is proper. Thus, (*c*) is a direct implication of (*a2*).

[1] Heller and Ito use a different definition for HMMs, which yields a different version of the minimality statement in part (*b*)

Proposition 6 has two simple but interesting implications: (i) every two-dimensional OOM is equivalent to a HMM (since all cones in two dimensions are polyhedral); (ii) every non-negative OOM (i.e., matrices τ_a have only non-negative entries) is equivalent to a HMM (since non-negative matrices map the positive orthant, which is a polyhedral cone, on itself).

Part (c) is sometimes useful to rule out a structure $\mathcal{A}$ as an OOM, by showing that some τ_a fails to satisfy the conditions given. Unfortunately, even if every τ_a of a structure $\mathcal{A}$ satisfies the conditions in (c), $\mathcal{A}$ need not be a valid OOM. Imperfect as it is, however, (c) is the strongest result available at this moment in the direction of characterising OOMs.

Proposition 6 is particularly useful to *build* OOMs from scratch, starting with a cone K and constructing observable operators satisfying $\tau_a K \subseteq K$. Note, however, that the theorem provides no means to *decide*, for a given structure $\mathcal{A}$, whether $\mathcal{A}$ is a valid OOM, since the theorem is non-constructive w.r.t. K.

More specifically, part (b) yields a construction of OOMs which are not equivalent to any HMM. This will be demonstrated in the remainder of this section.

Let $\tau_\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ be the linear mapping which right-rotates $\mathbb{R}^3$ by an angle φ around the first unit vector $e_1 = (1,0,0)$. Select some angle φ which is not a rational multiple of 2π. Then, put $\tau_a := \alpha\tau_\varphi$, where $0 < \alpha < 1$. Any 3-dimensional process described by a 3-dimensional OOM containing such a τ_a is not equivalent to any HMM: let K_a be a convex cone corresponding to a according to proposition $6(a1)$. Due to condition *3*, K_a must satisfy $\tau_a K_a \subseteq K_a$. Since τ_a rotates any set of vectors around $(1,0,0)$ by φ, this implies that K_a is rotation symmetric around $(1,0,0)$ by φ. Since φ is a non-rational multiple of 2π, K_a cannot be polyhedral. According to (b), this implies that an OOM which features this τ_a cannot be equivalent to any HMM.

I describe now such an OOM with $\mathcal{O} = \{a, b\}$. The operator τ_a is fixed, according to the previous considerations, by selecting $\alpha = 0.5$ and $\varphi = 1.0$. For τ_b, we take an operator which projects every $v \in \mathbb{R}^3$ on a multiple of $(.75, 0, .25)$, such that $\mu = \tau_a + \tau_b$ has column vectors with component sums equal to 1 (cf. definition 1(2)). The circular convex cone K whose border is obtained from rotating $(.75, 0, .25)$ around $(1,0,0)$ obviously satisfies the conditions in proposition $6(a2)$. Thus, we obtain a valid OOM provided that we select $w_0 \in K$. Using abbreviations $s := \sin(1.0), c := \cos(1.0)$, the matrices read as follows

$$\begin{aligned}
\tau_a &= 0.5 \begin{pmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{pmatrix} \\
\tau_b &= \begin{pmatrix} .75 \cdot .5 & .75(1 - .5c + .5s) & .75(1 - .5s - .5c) \\ 0 & 0 & 0 \\ .25 \cdot .5 & .25(1 - .5c + .5s) & .25(1 - .5s - .5c) \end{pmatrix}.
\end{aligned} \tag{15}$$

As starting vector w_0 we take $(.75, 0, .25)$, to obtain an OOM $\mathcal{C} = (\mathbb{R}^3, (\tau_a, \tau_b), w_0)$. I will briefly describe the phenomenology of the process generated by $\mathcal{C}$. The first observation is that every occurrence of b "resets" the process to the initial state w_0. Thus, we only have to understand what happens after uninterrupted sequences of a's. I.e., we should look at the conditional probabilities $P[\cdot \,|\, \varepsilon], P[\cdot \,|\, a], P[\cdot \,|\, aa], \ldots$, i.e., at $P[\cdot \,|\, a^t]$, where $t = 0, 1, 2 \ldots$. Figure 3 gives a plot of $P[a \,|\, a^t]$. The process could amply be called a "probability clock", or "probability oscillator"!

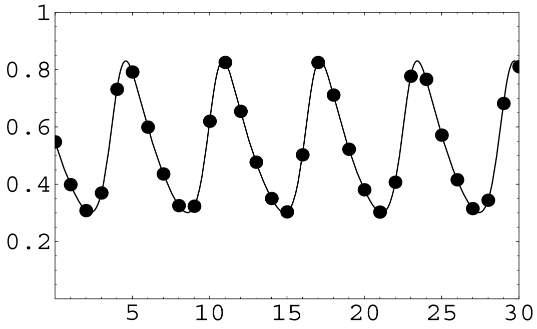

Figure 3: The rise and fall of probability to obtain a from the "probability clock" $\mathcal{C}$. Horizontal axis represents time steps t, vertical axis represents the probabilities $P[a \mid a^t] = \mathbf{1}\tau_a(\tau_a^t w_0/\mathbf{1}\tau_a^t w_0)$, which are rendered as dots. The solid line connecting the dots is given by $f(t) = \mathbf{1}\tau_a(\tau_t w_0/\mathbf{1}\tau_t w_0)$, where τ_t is the rotation by angle t described in the text.

Rotational operators can be exploited for "timing" effects. In our example, for instance, if the process would be started in the state according to $t = 4$ in fig. 3, there would be a high chance for two initial a's to occur, with a rapid drop in probability for a third or fourth. Such non-exponential-decay duration patterns for identical sequences are difficult to achieve with HMMs. Essentially, HMMs offer two possibilities for identical sequences: (i) recurrent transitions into a state where a is emitted, (ii) transitions along sequences of states, each of which can emit a. Option (i) is cumbersome because recurrent transitions imply exponential decay of state, which is unsuitable for many empirical processes; option (ii) blows up model size. A more detailed discussion of this problem in HMMs can be found in (Rabiner, 1990).

If one defines a three-dimensional OOM $\mathcal{C}'$ in a similar manner, but with a rational fraction of 2π as an angle of rotation, one obtains a process which can be modeled by a HMM. It follows from proposition $6(b)$ that the minimal HMM state number in this case is at least k, where k is the smallest integer such that $k\varphi$ is a multiple of 2π. Thus, the smallest HMM equivalent to a suitably chosen 3-dimensional OOM can have an arbitrarily great number of states. In a similar vein, any HMM which would give a reasonable approximation to the process depicted in fig. 3, would require at least 6 states, since in this process it takes approximately 6 time steps for one full rotation.

7 Interpretable OOMs

Some minimal-dimension OOMs have a remarkable property: their state space dimensions can be interpreted as probabilities of certain future outcomes. These *interpretable* OOMs will be described in this section.

First some terminology. Let $(X_t)_{t\in\mathbb{N}}$ be an m-dimensional LDP. For some suitably large k, let $\mathcal{O}^k = A_1 \cup \cdots \cup A_m$ be a partition of the set of sequences of length k into m disjoint nonempty subsets. The collective outcomes A_i are called *characteristic events* if some sequences $\bar{b}_1, \ldots, \bar{b}_m$ exist such that the $m \times m$ matrix

$$(P[A_i \,|\, \bar{b_j}])_{i,j} \tag{16}$$

is nonsingular (where $P[A_i \,|\, \bar{b_j}]$ denotes $\sum_{\bar{a}\in A_i} P[\bar{a} \,|\, \bar{b_j}]$). Every LDP has characteristic events:

Proposition 7 *Let $(X_t)_{t\in\mathbb{N}}$ be an m-dimensional LDP. Then there exists some $k \geq 1$ and a partition $\mathcal{O}^k = A_1 \cup \cdots \cup A_m$ of $\mathcal{O}^k$ into characteristic events.*

The proof is given in the appendix. Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\mathcal{O}}, w_0)$ be an m-dimensional OOM of the process (X_t). Using the characteristic events $A_1, \ldots, A_m$, we shall now construct from $\mathcal{A}$ an equivalent, *interpretable* OOM $\mathcal{A}(A_1, \ldots, A_m)$, which has the property that the m state vector components represent the probabilities of the m characteristic events to occur. More precisely, if during the generation procedure described in Section 3, $\mathcal{A}(A_1, \ldots, A_m)$ is in state $w_t = (w_t^1, \ldots, w_t^m)$ at time t, the probability $P[(X_{t+1}, \ldots, X_{t+k}) \in A_i \,|\, w_t]$ that the collective outcome A_i is generated in the next k time steps, is equal to w_t^i. In shorthand notation:

$$P[A_i \,|\, w_t] = w_t^i. \tag{17}$$

We shall use proposition 5 to obtain $\mathcal{A}(A_1, \ldots, A_m)$. Define $\tau_{A_i} := \sum_{\bar{a}\in A_i} \tau_{\bar{a}}$. Define a mapping $\varrho : \mathbb{R}^m \to \mathbb{R}^m$ by

$$\varrho(x) := (\mathbf{1}\tau_{A_1}x, \ldots, \mathbf{1}\tau_{A_m}x). \tag{18}$$

The mapping ϱ is obviously linear. It is also bijective, since the matrix $(P[A_i \,|\, \bar{b}_j]) = (\mathbf{1}\tau_{A_i}x_j)$, where $x_j := \tau_{\bar{b}_j}w_0/\mathbf{1}\tau_{\bar{b}_j}w_0$, is nonsingular. Furthermore, ϱ preserves component sums of vectors, since for $j = 1, \ldots, m$ it holds that $\mathbf{1}x_j = 1 = \mathbf{1}(P[A_1 \,|\, x_j], \ldots, P[A_m \,|\, x_j]) = \mathbf{1}(\mathbf{1}\tau_{A_1}x, \ldots, \mathbf{1}\tau_{A_m}x) = \mathbf{1}\varrho(x_j)$ (note that a linear map preserves component sums if it preserves component sums of basis vectors). Hence, ϱ satisfies the conditions of proposition 5. We therefore obtain an OOM equivalent to $\mathcal{A}$ by putting

$$\mathcal{A}(A_1, \ldots, A_m) = (\mathbb{R}^m, (\varrho\tau_a\varrho^{-1})_{a\in\mathcal{O}}, \varrho w_0) =: (\mathbb{R}^m, (\tau'_a)_{a\in\mathcal{O}}, w'_0). \tag{19}$$

In $\mathcal{A}(A_1, \ldots, A_m)$, equation (17) holds. To see this, let w'_t be a state vector obtained in a generation run of $\mathcal{A}(A_1, \ldots, A_m)$ at time t. Then conclude

$w'_t = \varrho\varrho^{-1}w'_t = (\mathbf{1}\tau_{A_1}(\varrho^{-1}w'_t), \ldots, \mathbf{1}\tau_{A_m}(\varrho^{-1}w'_t)) = (P[A_1 \mid \varrho^{-1}w'_t], \ldots, P[A_m \mid \varrho^{-1}w'_t])$ (computed in $\mathcal{A}$) $= (P[A_1 \,|\, w'_t], \ldots, P[A_m \,|\, w'_t])$ (computed in $\mathcal{A}(A_1, \ldots, A_m)$).

The $m \times m$ matrix corresponding to ϱ can easily be obtained from the original OOM $\mathcal{A}$ by observing that

$$\varrho = (\mathbf{1}\tau_{A_i}e_j), \tag{20}$$

where e_i is the i-th unit vector.

The following fact lies at the heart of the learning algorithm presented in the next section:

Proposition 8 *In an interpretable OOM $\mathcal{A}(A_1, \ldots, A_m)$ it holds that*

1. $w_0 = (P[A_1], \ldots, P[A_m])$,
2. $\tau_{\bar{b}}w_0 = (P[\bar{b}A_1], \ldots, P[\bar{b}A_m])$.

The proof is trivial.

The state dynamics of interpretable OOM can be graphically represented in a standardized fashion, which makes it possible to visually compare the dynamics of different processes. Any state vector w_t occurring in a generation run of an interpretable OOM is a probability vector. It lies in the non-negative hyperplane $H^{\geq 0} := \{(v^1, \ldots, v^m) \in \mathbb{R}^m \mid v^1 + \cdots + v^m = 1, v^i \geq 0 \text{ for } i = 1, \ldots, m\}$. Therefore, if one wishes to depict a state sequence $w_0, w_1, w_2 \ldots$, one only needs to render the bounded area $H^{\geq 0}$. Specifically, in the case $m = 3$, $H^{\geq 0}$ is the triangular surface shown in fig. 4(a). We can use it as the drawing plane, putting the point $(1/3, 1/3, 1/3)$ in the origin. For our orientation, we include the contours of $H^{\geq 0}$ into the graphical representation. This is an equilateral triangle whose edges have length $\sqrt{2}$. If $w = (w^1, w^2, w^3) \in H^{\geq 0}$ is a state vector, its components can be recovered from its position within this triangle, by exploiting $w^i = \sqrt{2/3}d_i$, where the d_i are the distances to the edges of the triangle. A similar graphical representation of states was first introduced in (Smallwood & Sondik, 1973) for HMMs.

When one wishes to graphically represent states of higher-dimensional, interpretable OOMs (i.e. where $m > 3$), one can join some of the characteristic events, until three merged events are left. State vectors can then be plotted in a way similar to the one just outlined.

To see an instance of interpretable OOMs, consider the "probability clock" example from (15). With $k = 2$, the following partition (among others) of $\{a, b\}^2$ yields characteristic events: $A_1 := \{aa\}, A_2 := \{ab\}, A_3 := \{ba, bb\}$. Using (20), one can calculate the matrix ϱ, which we omit here, and compute the interpretable OOM $\mathcal{C}(A_1, A_2, A_3)$ using (19). These are the observable operators τ_a and τ_b thus obtained:

$$\tau_a = \begin{pmatrix} 0.645 & -0.395 & 0.125 \\ 0.355 & 0.395 & -0.125 \\ 0 & 1 & 0 \end{pmatrix}, \quad \tau_b = \begin{pmatrix} 0 & 0 & 0.218 \\ 0 & 0 & 0.329 \\ 0 & 0 & 0.452 \end{pmatrix}. \tag{21}$$

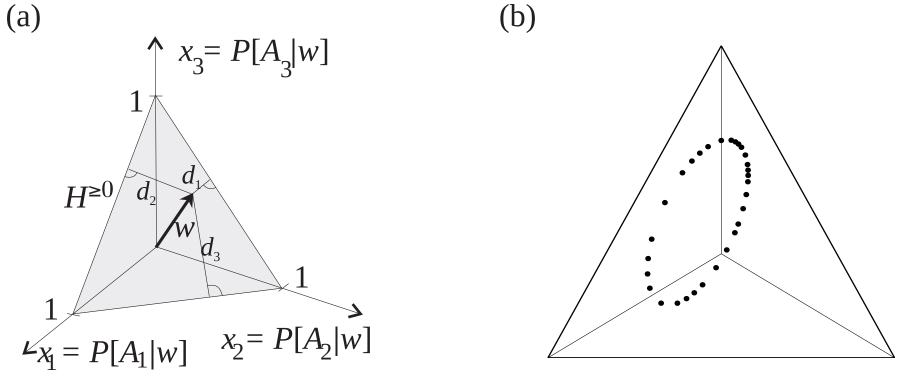

Figure 4: (a) The positioning of $H^{\geq 0}$ within state space. (b) State sequence of the probability clock, corresponding to fig. 3. For details compare text.

Fig. 4(b) shows a 30-step state sequence obtained by iterated applications of τ_a of this interpretable equivalent of the "probability clock". This sequence corresponds to fig. 3.

8 Learning OOMs

This section describes a constructive algorithm for learning OOMs from data. For demonstration, it is applied to learn the "probability clock" from data.

There are two standard situations where one wants to learn a model of a stochastic system: (i) from a single long data sequence (or few of them) one wishes to learn a model of a stationary process, and (ii) from many short sequences one wants to induce a model of a non-stationary process. OOMs can model both stationary and nonstationary processes, depending on the initial state vector w_0 (cf. proposition 1). The learning algorithm presented here are applicable in both cases. For the sake of notational convenience, we will only treat the stationary case.

We shall address the following learning task. Assume that a sequence $S = a_0 a_1 \cdots a_N$ is given, and that S is a path of an unknown stationary LDP (X_t). We assume that the dimension of (X_t) is known to be m (the question of how to assess m from S is discussed after the presentation of the basic learning algorithm). We select m characteristic events $A_1, \ldots, A_m$ of (X_t) (again, selection criteria are discussed after the presentation of the algorithm). Let $\mathcal{A}(A_1, \ldots, A_m)$ be an OOM of (X_t) which is interpretable w.r.t. $A_1, \ldots, A_m$. The learning task, then, is to induce from S an OOM $\tilde{\mathcal{A}}$ which is an estimate of $\mathcal{A}(A_1, \ldots, A_m) = (\mathbb{R}^m, (\tau_a)_{a \in \mathcal{O}}, w_0)$. We require that the estimation be consistent almost surely, i.e. for almost every infinite path $S_\infty = a_0 a_1 \cdots$ of (X_t), the sequence $(\tilde{\mathcal{A}}_n)_{n \geq n_0}$ obtained from estimating OOMs from initial sequences

$S_n = a_0 a_1 \cdots a_n$ of S_∞, converges to $\mathcal{A}(A_1, \ldots, A_m)$ (in some matrix norm).

An algorithm meeting these requirements shall now be described.

As a first step we estimate w_0. Prop. 8(*1*) states that $w_0 = (P[A_1], \ldots, P[A_m])$. Therefore, a natural estimate of w_0 is $\tilde{w}_0 = (\tilde{P}[A_1], \ldots, \tilde{P}[A_m])$, where $\tilde{P}[A_i]$ is the estimate for $P[A_i]$ obtained by counting frequencies of occurrence of A_i in S, as follows:

$$\tilde{P}[A_i] = \frac{\text{number of } \bar{a} \in A_i \text{ occurring in } S}{\text{number of } \bar{a} \text{ occurring in } S} = \frac{\text{number of } \bar{a} \in A_i \text{ occurring in } S}{N - k + 1}, \tag{22}$$

where k is the length of events A_i. In the second step, we estimate the operators τ_a. According to prop. 8(*2*), for any sequence $\bar{b}_j$ it holds that

$$\tau_a(\tau_{\bar{b}_j} w_0) \quad = \quad (P[\bar{b}_j a A_1], \ldots, P[\bar{b}_j a A_m]). \tag{23}$$

An m-dimensional linear operator is uniquely determined by the values it takes on m linearly independent vectors. This basic fact from linear algebra directly leads us to an estimation of τ_a, using (23). We estimate m linearly independent vectors $v_j := \tau_{\bar{b}_j} w_0$ by putting $\tilde{v}_j = (\tilde{P}[\bar{b}_j A_1], \ldots, \tilde{P}[\bar{b}_j A_m])$ $(j = 1, \ldots, m)$. For the estimation we use a similar counting procedure as in 22:

$$\tilde{P}[\bar{b}_j A_i] = \frac{\text{number of } \bar{b}\bar{a} \text{ (where } \bar{a} \in A_i) \text{ occurring in } S}{N - l - k + 1}, \tag{24}$$

where l is the length of $\bar{b}_j$. Furthermore, we estimate the results $v'_j := \tau_a(\tau_{\bar{b}_j} w_0)$ of applying τ_a to v_i by $\tilde{v}'_j = (\tilde{P}[\bar{b}_j a A_1], \ldots, \tilde{P}[\bar{b}_j a A_m])$, where

$$\tilde{P}[\bar{b}_j a A_i] = \frac{\text{number of } \bar{b} a \bar{a} \text{ (where } \bar{a} \in A_i) \text{ occurring in } S}{N - l - k}. \tag{25}$$

Thus we obtain estimates $(\tilde{v}_j, \tilde{v}'_j)$ of m argument-value pairs $(v_j, v'_j) = (v_j, \tau_a v_j)$ of applications of τ_a. From these estimated pairs, we can compute an estimate $\tilde{\tau}_a$ of τ_a through an elementary linear algebra construction: first collect the vectors $\tilde{v}_j$ as columns in a matrix $\tilde{V}$, and the vectors $\tilde{v}'_j$ as columns in a matrix $\tilde{W}_a$, then obtain $\tilde{\tau}_a = \tilde{W}_a \tilde{V}^{-1}$.

This basic idea can be augmented in two respects:

1. Instead of simple sequences $\bar{b}_j$, one can just as well take collective events B_j of some common lenght l to construct $\tilde{V} = (\tilde{P}[B_j A_i]), \tilde{W}_a = (\tilde{P}[B_j a A_i])$ (exercise). We will call B_j *indicative events.*

2. Instead of constructing $\tilde{V}, \tilde{W}_a$ as described above, one can also use raw count numbers, which saves the divisions on the rhs in (22),(24),(25). That is, use $V^\# = (\#_{\text{butlast}} B_j A_i), W_a^\# = (\# B_j a A_i)$, where $\#_{\text{butlast}} B_j A_i$ is the raw number of occurrences of $B_j A_i$ in $S_{\text{butlast}} := s_1 \ldots s_{N-1}$, and $\# B_j a A_i$ is the raw number of occurrences of $B_j a A_i$ in S. It is easy to see that this gives the same matrices $\tilde{\tau}_a$ as the original procedure.

Assembled in an orderly fashion, the entire procedure works as follows (assume that model dimension m, indicative events B_j, and characteristic events A_i have already been selected).

Step 1 Compute the $m \times m$ matrix $V^{\#} = (\#_{\text{butlast}} B_j A_i)$.

Step 2 Compute, for every $a \in \mathcal{O}$, the $m \times m$ matrix $W_a^{\#} = (\# B_j a A_i)$.

Step 3 Obtain $\tilde{\tau}_a = W_a^{\#} (V^{\#})^{-1}$.

The computational demands of this procedure are modest compared to today's algorithms used in HMM parameter estimation. The counting for $V^{\#}$ and $W_a^{\#}$ can be done by a single sweep of an inspection window (of length $k+l+1$) over S. Multiplying or inverting $m \times m$ matrices essentially has a computational cost of $O(m^3/p)$ (this can be slightly improved, but the effects become noticeable only for very large m), where p is the degree of parallelization. The counting and inverting/multiplying operations together give a time complexity of this core procedure of $O(N + nm^3/p)$, where n is the size of $\mathcal{O}$.

We shall now demonstrate the "mechanics" of the algorithm with an artificial toy example. Assume that the following path S of length 20 is given:

$$S = abbbaaaabaabbbabbbbb.$$

We estimate a two-dimensional OOM. We choose the simplest possible characteristic events $A_1 = \{a\}, A_2 = \{b\}$ and indicative events $B_1 = \{a\}, B_2 = \{b\}$.

First we estimate the invariant vector w_0, by putting

$$\tilde{w_0} = (\#a, \#b)/N = (8/20, 12/20).$$

Then we obtain $V^{\#}$ and $W_a^{\#}, W_b^{\#}$ by counting occurrences of subsequences in S:

$$\begin{aligned} V^{\#} &= \begin{pmatrix} \#_{\text{butlast}} aa & \#_{\text{butlast}} ba \\ \#_{\text{butlast}} ab & \#_{\text{butlast}} bb \end{pmatrix} = \begin{pmatrix} 4 & 3 \\ 4 & 7 \end{pmatrix}, \\ W_a^{\#} &= \begin{pmatrix} \#aaa & \#baa \\ \#aab & \#bab \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 1 \end{pmatrix}, \\ W_b^{\#} &= \begin{pmatrix} \#aba & \#bba \\ \#abb & \#bbb \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix}. \end{aligned}$$

From these raw counting matrices we obtain estimates of the observable operators by

$$\begin{aligned} \tilde{\tau}_a &= W_a^{\#} (V^{\#})^{-1} = \begin{pmatrix} 3/8 & 1/8 \\ 5/8 & -1/8 \end{pmatrix}, \\ \tilde{\tau}_b &= W_b^{\#} (V^{\#})^{-1} = \begin{pmatrix} -1/16 & 5/16 \\ 1/16 & 11/16 \end{pmatrix}. \end{aligned}$$

That is, we have arrived at an estimate

$$\tilde{\mathcal{A}} = (\mathbb{R}^2, \begin{pmatrix} 3/8 & 1/8 \\ 5/8 & -1/8 \end{pmatrix}, \begin{pmatrix} -1/16 & 5/16 \\ 1/16 & 11/16 \end{pmatrix}, (9/20, 11/20)). \tag{26}$$

This concludes the presentation of the learning algorithm in its core version. Obviously, before one can start the algorithm, one has to fix the model dimension, and one has to select characteristic and indicative events. These two questions shall now be briefly addressed.

In practice, the problem of determining the "true" model dimension seldom arises. Empirical systems quite likely are very high-dimensional or even infinite-dimensional. Given finite data, one cannot capture all of the true dimensions. Rather, the task is to determine m such that learning an m-dimensional model reveals m *significant* process dimensions, while any contributions of higher process dimensions are insignificant in the face of estimation error. Put bluntly, the task is to fix m such that data are neither overfitted nor underexploited.

A practical solution to this task capitalizes on the idea that a size m model does not overfit data in S if the "noisy" matrix $V^{\#}$ has full rank. Estimating the true rank of a noisy matrix is a common task in numerical linear algebra, for which heuristic solutions are known (Golub & Loan, 1996). A practical procedure for determining an appropriate model dimension, which exploits these techniques, is detailed out in (Jaeger, 1998).

Selecting optimal characteristic and indicative events is an intricate issue. The quality of estimates $\tilde{\mathcal{A}}$ varies considerably with different characteristic and indicative events. The ramifications of this question are not fully understood, and only some preliminary observations can be supplied here.

A starting point toward a principled handling of the problem are certain conditions on A_i, B_j reported in (Jaeger, 1998). If A_i, B_j are chosen according to these conditions, it is guaranteed that the resulting estimate $\tilde{\mathcal{A}}$ is itself interpretable w.r.t. the chosen characteristic events $A_1, \ldots, A_m$. I.e., it holds that $\tilde{\mathcal{A}} = \tilde{\mathcal{A}}(A_1, \ldots, A_m)$. In turn, this warrants that $\tilde{\mathcal{A}}$ yields an *unbiased* estimate of certain parameters which define (X_t).

A second observation is that the variance of estimates $\tilde{\mathcal{A}}$ depends on the selection of A_i, B_j. For instance, characteristic events should be chosen such that sequences $\bar{a}_x, \bar{a}_y \in A_i$ collected in the same characteristic event have a high correlation in the sense that the probabilities $\tilde{P}_S[\bar{a}_x \,|\, B_j], \tilde{P}_S[\bar{a}_y \,|\, B_j]$ (where $j = 1, \ldots, m$) correlate strongly. The better this condition is met, the smaller is the variance of $\tilde{P}_S[A_i]$, and accordingly the variance of $\tilde{V}$ and $\tilde{W}_a$. Several other rules of thumb similar to this one are discussed in (Jaeger, 1998).

A third and last observation is that characteristic events can be chosen such that available domain knowledge is exploited. For instance, in estimating an OOM of an written English text string S, one might collect in each characteristic or indicative event such letter strings as belong to the same linguistic (phonological or morphological) category. This would result in an interpretable model whose state dimensions represent the probabilities for these categories to be realized next in the process.

We now apply the learning procedure to the “probability clock” $\mathcal{C}$ introduced in Sections 6 and 7[2]. We provide only a sketch here. A more detailed account is contained in (Jaeger, 1998).

$\mathcal{C}$ was run to generate a path S of length $N = 30,000$. $\mathcal{C}$ was started in an invariant state (cf. prop. 1), therefore S is stationary. We shall construct from S an estimate $\tilde{\mathcal{C}}$ of $\mathcal{C}$.

Assume that we know that the process dimension is $m = 3$ (this value was also obtained by the dimension estimation heuristics mentioned above). The rules of thumb noted above lead to characteristic events $A_1 = \{aa\}, A_2 = \{ab\}, A_3 = \{ba, bb\}$ and indicative events $B_1 = \{aa\}, B_2 = \{ab, bb\}, B_3 = \{ba\}$ (details of how these events were selected are documented in (Jaeger, 1998)).

Using these characteristic and indicative events, an OOM $\tilde{\mathcal{C}} = (\mathbb{R}^3, (\tilde{\tau}_a, \tilde{\tau}_b), \tilde{w}_0)$ was estimated, using the algorithm described in this section. We will briefly highlight the quality of the model thus obtained.

First, we compare the operators $\tilde{\mathcal{C}}$ with the operators of the interpretable version $\mathcal{C}(A_1, A_2, A_3) = (\mathbb{R}^3, (\tau_a, \tau_b), w_0)$ of the probability clock (cf.(21)). The average absolute error of matrix entries in $\tilde{\tau}_a, \tilde{\tau}_b$ vs. τ_a, τ_b was found to be approximately .0038 (1.7% of average correct value).

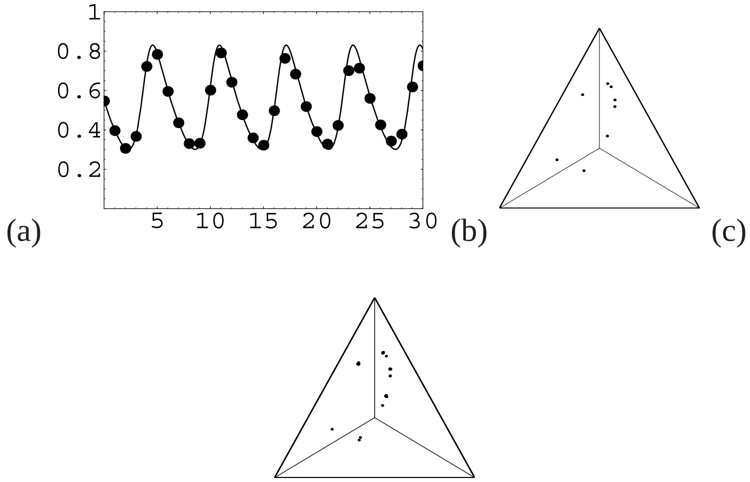

Figure 5: Comparison of estimated model vs. original ”probability clock”. (a) Probabilities $P(a|\bar{a})$ as captured by the learnt model (dots) vs. the original (solid line). (b) The most frequent states of the original. (c) States visited by the learnt model. For details see text.

[2] The calculations were done using the Mathematica software package. Data and Mathematica programs can be fetched from the author’s internet home page at www.gmd.de/People/Herbert.Jaeger/

A comparison of the matrices $\tilde{\tau}_a, \tilde{\tau}_b$ vs. τ_a, τ_b is not too illuminating, however, because some among the matrix entries have little effect on the process (i.e., varying them greatly would only slightly alter the distribution of the process). Conversely, information in S contributes only weakly to estimating these matrix entries, which are therefore likely to deviate considerably from the corresponding entries in the original matrices.

Indeed, taken as a sequence generator, the estimated model is much closer to the original process than a matrix entry deviation of 1.7% might suggest. Fig. 5 illustrates the fit of probabilities and states computed with $\tilde{\mathcal{C}}$ vs. the true values. Fig. 5(a) is a diagram similar to Figure 3, plotting the probabilities $P_{\tilde{\mathcal{C}}}[a \,|\, a^t]$ against time steps t. The solid line represents the probabilities of the original probabilitiy clock. Even after 30 time steps, the predictions are very close to the true values. Fig. 5(b) is an excerpt of Figure 4(b) and shows the eight most frequently taken states of the original probability clock. Fig. 5(c) is an overlay of (b) with a 100000-step run of the estimated model $\tilde{\mathcal{C}}$. Every 100-th step of the 100000-step run was additionally plotted into (b) to obtain (c). The original states closely agree with the states of the estimated model.

I conclude this section by emphasizing a shortcoming of the learning algorithm in its present form. An OOM is characterized by the three conditions stated in definition 1. The learning algorithm only guarantees that the estimated structure $\tilde{\mathcal{A}}$ satisfies the first two conditions (easy exercise). Unfortunately, it may happen that $\tilde{\mathcal{A}}$ does not satisfy condition 3, and thus might not be a valid OOM. That is, $\tilde{\mathcal{A}}$ might produce negative "probabilities" $P[\bar{a}]$. For instance, the structure $\tilde{\mathcal{A}}$ estimated in (26) is not a valid OOM: one obtains $\mathbf{1}\tau_b\tau_a\tau_b\tau_a\tau_b w_0 = -0.00029$. A procedure for transforming an "almost-OOM" into a nearest (in some matrix norm) valid OOM would be highly desirable. Progress is currently blocked by the lack of a decision procedure for determining whether an "OOM-like" structure $\mathcal{A}$ satisfies condition 3.

In practice, however, even an "almost-OOM" can be useful. If the model dimension is selected properly, the learning procedure yields either valid OOMs or pseudo ones which are close to valid ones. Even the latter ones yield good estimates $\tilde{P}[\bar{a}]$ if $\bar{a}$ is not too long and if $P[\bar{a}]$ is not too close to zero. Further examples and discussion can be found in (Jaeger, 1997b) and (Jaeger, 1998).

9 Conclusion

The results reported in this article are all variations on a single insight: The *change* of predictive knowledge that we have about a stochastic system, is a *linear* phenomenon. This leads to the concept of observable operators, which has been detailed out here for discrete-time, finite-valued processes, but is applicable to every stochastic process (?). The linear nature of observable operators makes it possible to cast the system identification task purely in terms of numerical linear algebra.

The proposed system identification technique proceeds in two stages: first, determine the appropriate model dimension and choose characteristic and indicative events; second, construct the counting matrices and from them, the model. While the second stage is purely mechanical, the first involves some heuristics. The situation is reminiscent of HMM estimation, where in a first stage a model structure has to be determined – usually by hand; or of neural network learning, where a network topology has to be designed before the mechanical parameter estimation can start. Although the second stage of model construction for OOMs is transparent and efficient, it remains to be seen how well the subtleties involved with the first stage can be mastered in real-life applications. Furthermore, the problem remains to be solved of how to deal with the fact that the learning algorithm may produce invalid OOMs. One reason for optimism is that these questions can be posed in terms of numerical linear algebra, which arguably is the best understood of all areas of applied mathematics.

A Proof of theorem 1

A numerical function P on the set of finite initial sequences of Ω can be uniquely extended to the probability distribution of a stochastic process, if the following two conditions are met:

1. P is a probability measure on the set of initial sequences of length $k+1$, for all $k \geq 0$. That is, (i) $P[a_{i_0} \ldots a_{i_k}] \geq 0$, and (ii) $\sum_{a_{i_0} \ldots a_{i_k} \in \mathcal{O}^{k+1}} P[a_{i_0} \ldots a_{i_k}] = 1$.
2. The values of P on the initial sequences of length $k+1$ agree with continuation of the process in the sense that $P[a_{i_0} \ldots a_{i_k}] = \sum_{b \in \mathcal{O}} P[a_{i_0} \ldots a_{i_k} b]$.

The process is stationary, if additionally the following condition holds:

3. $P[a_{i_0} \ldots a_{i_k}] = \sum_{b_0 \ldots b_s \in \mathcal{O}^{s+1}} P[b_0 \ldots b_s a_{i_0} \ldots a_{i_k}]$ for all $a_{i_0} \ldots a_{i_k} \; in \mathcal{O}^{k+1}$ and $s \geq 0$.

Point 1(i) is warranted by virtue of condition 3 from definition 1. 1(ii) is a consequence of conditions 1 and 2 from the definition (exploit that condition 2 implies that left-multiplying a vector by μ does not change the sum of components of the vector):

$$\begin{aligned} \sum_{\bar{a} \in \mathcal{O}^{k+1}} P[\bar{a}] = \sum_{\bar{a} \in \mathcal{O}^{k}} \mathbf{1} \tau_{\bar{a}} w_0 = & \\ = \; & \mathbf{1}(\sum_{a \in \mathcal{O}} \tau_a) \cdots (\sum_{a \in \mathcal{O}} \tau_a) w_0 \quad (k+1 \text{ terms } (\sum_{a \in \mathcal{O}} \tau_a)) \\ = \; & \mathbf{1}\mu \cdots \mu w_0 \quad = \quad \mathbf{1} w_0 \quad = \quad 1. \end{aligned}$$

For proving point 2, again exploit condition 2:

$$\sum_{b \in \mathcal{O}} P[\bar{a} b] = \sum_{b \in \mathcal{O}} \mathbf{1} \tau_b \tau_{\bar{a}} w_0 = \mathbf{1} \mu \tau_{\bar{a}} w_0 = \mathbf{1} \tau_{\bar{a}} w_0 = P[\bar{a}].$$

Finally, the stationarity criterium 3 is obtained by exploiting $\mu w_0 = w_0$:

$$\begin{aligned} \sum_{\bar{b} \in \mathcal{O}^{s+1}} P[\bar{b}\bar{a}] &= \sum_{\bar{b} \in \mathcal{O}^{s+1}} \mathbf{1}\tau_{\bar{a}}\tau_{\bar{b}}w_0 \\ &= \mathbf{1}\tau_{\bar{a}}\mu \dots \mu w_0 \quad (s+1 \text{ terms } \mu) \\ &= \mathbf{1}\tau_{\bar{a}} w_0 \quad = \quad P[\bar{a}]. \end{aligned}$$

B Proof of proposition 2

Let $\bar{b} \in \mathcal{O}^*$, and $\mathfrak{g}_{\bar{b}} = \sum_{i=1}^n \alpha_i \mathfrak{g}_{\bar{c}_i}$ be the linear combination of $\mathfrak{g}_{\bar{b}}$ from basis elements of $\mathfrak{G}$. Let $\bar{d} \in \mathcal{O}^+$. Then, we obtain the statement of the proposition through the following calculation:

$$\begin{aligned} (\mathfrak{t}_a(\mathfrak{g}_{\bar{b}}))(\bar{d}) \quad &= \\ &= (\mathfrak{t}_a(\sum_{i=1}^n \alpha_i \mathfrak{g}_{\bar{c}_i}))(\bar{d}) \quad = \quad (\sum \alpha_i \mathfrak{t}_a(\mathfrak{g}_{\bar{c}_i}))(\bar{d}) \\ &= (\sum \alpha_i P[a \,|\, \bar{c}_i] \; \mathfrak{g}_{\bar{c}_i a})(\bar{d}) \quad = \quad \sum \alpha_i P[a \,|\, \bar{c}_i] P[\bar{d} \,|\, \bar{c}_i a] \\ &= \sum \alpha_i P[a \,|\, \bar{c}_i] \frac{P[\bar{c}_i a \bar{d}]}{P[a \,|\, \bar{c}_i] P[\bar{c}_i]} \quad = \quad \sum \alpha_i \frac{P[\bar{c}_i] P[a\bar{d} \,|\, \bar{c}_i]}{P[\bar{c}_i]} \\ &= \mathfrak{g}_{\bar{b}}(a\bar{d}) \quad = \quad P[a\bar{d} \,|\, \bar{b}] \quad = \quad P[a \,|\, \bar{b}] \; P[\bar{d} \,|\, \bar{b}a] \\ &= P[a \,|\, \bar{b}] \; \mathfrak{g}_{\bar{b}a}(\bar{d}). \end{aligned}$$

C Proof of proposition 3

From an iterated application of (12) it follows that $\mathfrak{t}_{a_{i_k}} \dots \mathfrak{t}_{a_{i_0}} \mathfrak{g}_\varepsilon = P[a_{i_0} \dots a_{i_k}] \mathfrak{g}_{a_{i_0} \dots a_{i_k}}$. Therefore, it holds that

$$\mathfrak{g}_{a_{i_0} \dots a_{i_k}} = \sum_{i=1,\dots,n} \frac{\alpha_i}{P[a_{i_0} \dots a_{i_k}]} \mathfrak{g}_{\bar{b}_i}.$$

Interpreting the vectors $\mathfrak{g}_{\bar{b}_i}$ and $\mathfrak{g}_{a_{i_0} \dots a_{i_k}}$ as probability distributions (cf. (11)), it is easy to see that $\sum_{i=1,\dots,n} \frac{\alpha_i}{P[a_{i_0} \dots a_{i_k}]} = 1$, from which the statement immediately follows.

D Proof of proposition 4

To see *1*, let $(\mathfrak{G}, (\mathfrak{t}_a)_{a \in \mathcal{O}}, \mathfrak{g}_\varepsilon)$ be the predictor-space OOM of (X_t). Choose $\{\bar{b}_1, \dots, \bar{b}_m\} \subset \mathcal{O}^*$ such that the set $\{\mathfrak{g}_{\bar{b}_i} \,|\, i = 1, \dots, m\}$ is a basis of $\mathfrak{G}$. Then, it is an easy exercise to show that an OOM $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a \in \mathcal{O}}, w_0)$ and an isomorphism $\pi : \mathfrak{G} \to \mathbb{R}^m$ exist such that (i) $\pi(\mathfrak{g}_\varepsilon) = w_0$, (ii) $\pi(\mathfrak{g}_{\bar{b}_i}) = e_i$, where

e_i is the i-th unit vector, (iii) $\pi(\mathfrak{t}_a\mathfrak{d}) = \tau_a\pi(\mathfrak{d})$ for all $a \in \mathcal{O}, \mathfrak{d} \in \mathfrak{G}$. These properties imply that $\mathcal{A}$ is an OOM of (X_t).

In order to prove *2*, let again $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\mathcal{O}}, \mathfrak{g}_\varepsilon)$ be the predictor-space OOM of (X_t). Let Γ be the linear subspace of $\mathbb{R}^k$ spanned by the vectors $\{w_0\} \cup \{\tau_{\bar{a}}w_0 \,|\, \bar{a} \in \mathcal{O}^+\}$. Let $\{\tau_{\bar{b}_1}w_0, \ldots, \tau_{\bar{b}_l}w_0\}$ be a basis of Γ. Define a linear mapping σ from Γ to $\mathfrak{G}$ by putting $\sigma(\tau_{\bar{b}_i}w_0) := P[\bar{b}_i]\, \mathfrak{g}_{\bar{b}_i}$, where $i = 1, \ldots, l$. σ is called the *canonical projection* of $\mathcal{A}$ on the predictor-space OOM. By a straightforward calculation, it can be shown that $\sigma(w_0) = \mathfrak{g}_\varepsilon$ and that for all $\bar{c} \in \mathcal{O}^+$ it holds that $\sigma(\tau_{\bar{c}}w_0) = P[\bar{c}]\mathfrak{g}_{\bar{c}}$ (cf. (Jaeger, 1997a) for these and other properties of σ). This implies that σ is surjective, which in turn yields $m \leq k$.

E Proof of proposition 7

From the definition of process dimension (def. 2) it follows that m sequences $\bar{a}_1, \ldots, \bar{a}_m$ and m sequences $\bar{b}_1, \ldots, \bar{b}_m$ exist such that the $m \times m$ matrix $(P[\bar{a}_j \,|\, \bar{b}_i])$ is nonsingular. Let k be the maximal length occurring in the sequences $\bar{a}_1, \ldots, \bar{a}_m$. Define collective events C_j of length k by using the sequences a_j as initial sequences, i.e. put $C_j := \{\bar{a}_j\bar{c} \,|\, \bar{c} \in \mathcal{O}^{k-|\bar{a}_j|}\}$, where $|\, \bar{a}_j \,|$ denotes the length of $\bar{a}_j$. It holds that $(P[C_j \,|\, \bar{b}_i]) = (P[\bar{a}_j \,|\, \bar{b}_i])$. We transform the collective events C_j in two steps in order to obtain characteristic events.

In the first step, we make them disjoint. Observe that due to their construction, two collective events C_{j_1}, C_{j_2} are either disjoint, or one is properly included in the other. We define new, non-empty, pairwise disjoint, collective events $C'_j := C_j \setminus \bigcup_{C_x \subset C_j} C_x$ by taking away from C_j all collective events properly included in it. It is easily seen that the matrix $(P[C'_j \,|\, \bar{b}_i])$ can be obtained from $(P[C_j \,|\, \bar{b}_i])$ by subtracting certain rows from others. Therefore, this matrix is nonsingular, too.

In the second step, we enlarge the C'_j (while preserving disjointness) in order to arrive at collective events A_j which exhaust $\mathcal{O}^k$. Put $C'_0 = \mathcal{O}^k \setminus (C'_1 \cup \ldots \cup C'_m)$. If $C'_0 = \emptyset$, $A_j := C'_j$ $(j = 1, \ldots, m)$ are characteristic events. If $C'_0 \neq \emptyset$, consider the $m \times (m+1)$ matrix $(P[C'_j \,|\, \bar{b}_i])_{i=1,\ldots,m, j=0,\ldots,m}$. It has rank m, and column vectors $v_j = (P[C'_j \,|\, \bar{b}_1], \ldots, P[C'_j \,|\, \bar{b}_m])$. If v_0 is the null vector, put $A_1 := C'_0 \cup C'_1, A_2 := C'_2, \ldots, A_m := C'_m$ to obtain characteristic events. If v_0 is not the null vector, let $v_0 = \sum_{\nu=1,\ldots,m} \alpha_\nu v_\nu$ be its linear combination from the other column vectors. Since all v_ν are non-null, non-negative vectors, some α_{ν_0} must be properly greater than 0. A basic linear algebra argument (exercise) shows that the $m \times m$ matrix made from column vectors $v_1, \ldots, v_{\nu_0} + v_0, \ldots, v_m$ has rank m. Put $A_1 := C'_1, \ldots, A_{\nu_0} := C'_{\nu_0} \cup C'_0, \ldots, A_m := C'_m$ to obtain characteristic events.

Acknowledgments I feel deeply grateful toward Thomas Christaller for confidence and great support. This article has benefitted enormously from advice given by Hisashi Ito, Shun-Ichi Amari, Rafael Nuñez, Volker Mehrmann, and two extremely constructive anonymous reviewers, whom I heartily thank.

The research described in this paper was carried out under a postdoctoral grant from the German National Research Center for Information Technology (GMD).

References

Bengio, Y. (1996). *Markovian models for sequential data* (Technical Report No. 1049). Dpt. d'Informatique et Recherche Opérationelle, Université de Montréal. (http://www.iro.umontreal.ca/labs/neuro/pointeurs/hmmsTR.ps)

Berman, A., & Plemmons, R. (1979). *Nonnegative matrices in the mathematical sciences.* Academic Press.

Doob, J. (1953). *Stochastic processes.* John Wiley & Sons.

Elliott, R., Aggoun, L., & Moore, J. (1995). *Hidden markov models: Estimation and control* (Vol. 29). Springer Verlag, New York.

Gilbert, E. (1959). On the identifiability problem for functions of finite Markov chains. *Annals of Mathematical Statistics*, *30*, 688-697.

Golub, G., & Loan, C. van. (1996). *Matrix computations, third edition.* The Johns Hopkins University Press.

Heller, A. (1965). On stochastic processes derived from Markov chains. *Annals of Mathematical Statistics*, *36*, 1286-1291.

Iosifescu, M., & Theodorescu, R. (1969). *Random processes and learning* (Vol. 150). Springer Verlag.

Ito, H. (1992). *An algebraic study of discrete stochastic systems.* Phd thesis, Dpt. of Math. Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, Bunkyo-ku, Tokyo. (ftp'able from http://kuro.is.sci.toho-u.ac.jp:8080/english/D/)

Ito, H., Amari, S.-I., & Kobayashi, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE transactions on information theory*, *38*(2), 324-333.

Jaeger, H. (1997a). *Observable operator models and conditioned continuation representations* (Arbeitspapiere der GMD No. 1043). GMD, Sankt Augustin. (http://www.gmd.de/People/ Herbert.Jaeger/Publications.html)

Jaeger, H. (1997b). *Observable operator models II: Interpretable models and model induction* (Arbeitspapiere der GMD No. 1083). GMD, Sankt Augustin. (http://www.gmd.de/People/ Herbert.Jaeger/Publications.html)

Jaeger, H. (1998). *Discrete-time, discrete-valued observable operator models: a tutorial* (GMD Report No. 42). GMD, Sankt Augustin. (http://www.gmd.de/People/ Herbert.Jaeger/Publications.html)

Narendra, K. (1995). Identification and control. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (p. 477-480). MIT Press/Bradford Books.

Rabiner, L. (1990). A tutorial on hidden Markov models and selected applications in speech recognition. In A. Waibel & K.-F. Lee (Eds.), *Readings in speech recognition* (p. 267-296). Morgan Kaufmann, San Mateo. (Reprinted from Proceedings of the IEEE 77 (2), 257-286 (1989))

Smallwood, R., & Sondik, E. (1973). The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, *21*, 1071-1088.