A short introduction to observable operator models of stochastic processes*

Herbert Jaeger
German National Research Center
for Information Technology (GMD), FIT.KI
Schloss Birlinghoven, D-53754 Sankt Augustin
email: herbert.jaeger@gmd.de

Abstract

The article describes a new formal approach to model discrete stochastic processes, called observable operator models (OOMs). It is shown how hidden Markov models (HMMs) can be properly generalized to OOMs. These OOMs afford both mathematical simplicity and algorithmic efficiency, where HMMs exhibit neither. The observable operator idea also leads to an abstract, information-theoretic representation of stationary stochastic processes. It is shown how any such process can be uniquely characterized by linear, observable operators, yielding an abstract OOM of the process. All in all, observable operators open a lucid, general, and computationally extremely powerful avenue to stochastic processes.

1 Introduction

In the theory of systems and control, trajectories of discrete-time dynamical systems are usually seen as a sequence of states [Zadeh, 1969] [Stengel, 1986]. They are 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 is visited by the system due to the effects of a time step operator.

In this article, trajectories are perceived in a complementary fashion. From a set of operators (say, $\{A, B\}$), 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 $\ldots A(A(B(A \ldots))) \ldots$ (fig. 1b). Since in this perspective, the observables are the operators themselves, I have named this kind of stochastic models, "observable operator models" (OOMs). An appropriate metaphor would be to view trajectories as sequences of actions.

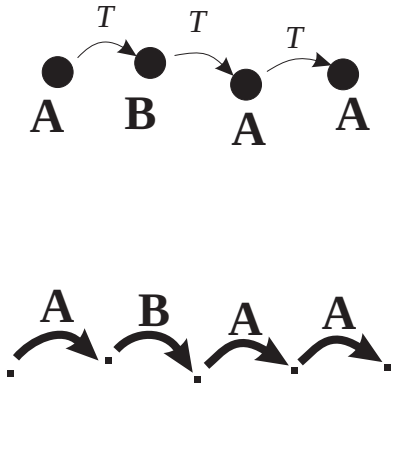

Figure 1: (a) The standard view of trajectories. A time step operator T yields a sequence ABAA of states. (b) The OOM view. Operators A,B are concatenated to yield a sequence ABAA of observables.

Stochastic sequences of operators are a well-known object of mathematical investigation [Iosifescu and Theodorescu, 1969]. The results reported in the present article arise from a crucial new insight: the probability of selecting an observable operator at a given time can be computed *using the operator itself.*

This twist in perspective has enormous practical and theoretical impact. On the more practical side (section 2), I show how hidden Markov models (HMMs) can be recast, and generalized, to become OOMs. These OOMs are mathematically very transparent, and have exciting algorithmic properties. On the theoretical side (section 3), I develop an abstract, information-theoretic version of OOMs. It turns out that the "concrete" OOMs obtained in section 2 as a generalization of HMMs coincide (up to isomorphism) with the finite-dimensional abstract OOMs.

*presented at the Fourteenth European Meeting on Cybernetics and Systems Research (EMCSR 98), Vienna, April 1998

The article gives an informal overview of results which are documented in mathematical rigour in [Jaeger, 1997a] [Jaeger, 1997b].

2 OOMs as a generalization of HMMs

Today, hidden Markov models (HMMs) [Rabiner, 1990] [Elliott *et al.*, 1995] [Bengio, 1996] provide the state-of-the-art techniques for analyzing discrete stochastic sequences of observations. They are standardly put to tasks as diverse as protein classification, ion channel activity measurements, speech recognition, or the detection of gestures in computer vision. In spite of their wide use, HMMs possess neither an elegant mathematical theory nor pleasant algorithmic properties. In this section I show how HMMs can be generalized to arrive at OOMs, and compare the properties of HMMs vs. OOMs.

2.1 From HMMs to OOMs

I will indicate how OOMs can be generalized from HMMs by re-writing a simple HMM as an OOM. I assume that the reader is familiar with HMMs. Consider the HMM depicted in fig. 2. It has two hidden states $\{s_1, s_2\}$ and two observable events $\Sigma = \{a, b\}$.

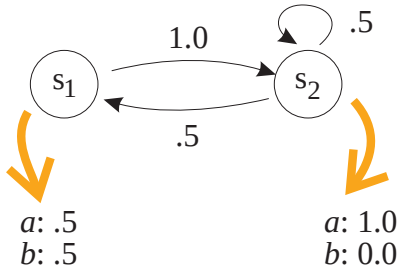

Figure 2: An exemplary HMM. Thin arrows indicate state transitions, with annotated probabilities. Bold grey arrows denote probabilistic emission of observable events.

Formally, the state transition probabilities can be collected in a stochastic matrix M which at place (i, j) contains the transition probability from state s_i to s_j. The emission probabilities $P[a_i \,|\, s_j]$ can be sorted into diagonal *observation matrices* O_a, O_b. For instance, O_a contains, in its diagonal, the probabilities $P[a \,|\, s_1]$ and $P[a \,|\, s_2]$:

$$M = \begin{pmatrix} 0.0 & 1.0 \\ .5 & .5 \end{pmatrix}$$

$$O_a = \begin{pmatrix} .5 & \\ & 1.0 \end{pmatrix} \quad O_b = \begin{pmatrix} .5 & \\ & 0.0 \end{pmatrix}$$

In order to fully characterize this HMM, one also must supply an initial distribution $w_0 = (P[s_1], P[s_2])$, where $P[s_i]$ is the probability that the system starts in state s_i. If one investigates stationary processes, w_0 is taken to be a stationary distribution which is uniquely determined in most cases of interest by the condition

$$M' w_0 = w_0, \tag{1}$$

where M' denotes the transform of M. This paper only deals with stationary processes, and w_0 always denotes the stationary distribution. In our example, $w_0 = (1/3, 2/3)$.

The first step toward an OOM is to multiply M' with the matrices O_a, O_b to obtain $T_a = M'O_a, T_b = M'O_b$. Note that no information is lost by this merging. M', O_a, O_b can be recovered from T_a, T_b by observing that

$$M' = M' \cdot \mathbf{1} = M'(O_a + O_b) = T_a + T_b. \tag{2}$$

The operators T_a, T_b can be used to compute the probability $P[c_1 \dots c_k]$ by which the finite event sequence $c_1 \dots c_k$ (where $c_i \in \{a, b\}$) occurs among all other sequences of equal length in the process described by the HMM. It can be shown that

$$P[c_1 \dots c_k] = \sigma(T_{c_k} \circ T_{c_{k-1}} \circ \dots \circ T_{c_1} w_0) \tag{3}$$

In this equation, $\sigma(x_1, \dots, x_n) := x_1 + \dots + x_n$ denotes the *internal sum* of a vector. (3) provides a more transparent way of computing sequence probabilities than does the "forward-backward" algorithm, which is traditionally used by HMM practicians for the same purpose [Rabiner, 1990].

Now, one arrives at the definition of an OOM by (i) relaxing the requirement that M' be the transpose of a stochastic matrix, to the weaker requirement that the columns of M' each sum to 1, and by (ii) requiring from w_0 merely that it has an internal sum of 1. Using the letter τ in OOMs in places where T appears in HMMs, and introducing $\mu := \sum_{c \in \Sigma} \tau_c$ in analogy to (2), this yields:

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

1. *$\sigma w_0 = 1$,*
2. *μ has column sums equal to 1,*
3. *$\mu w_0 = w_0$,*
4. *for all sequences $c_1 \dots c_k$ it holds that $\sigma(\tau_{c_k} \cdots \tau_{c_1} w_0) \in [0, 1]$.*

Conditions 1 and 2 reflect the relaxations (i) and (ii) mentioned previously, condition 3 corresponds to (1), while condition 4 ensures that one obtains a valid analogue of (3).

After carrying out the multiplications $T_a = M'O_a, T_b = M'O_b$, our exemplary HMM is formally written in OOM format, as follows:

$$\mathcal{A} = (\mathbb{R}^2, (\begin{pmatrix} 0.0 & .5 \\ .5 & .5 \end{pmatrix}, \begin{pmatrix} 0.0 & 0.0 \\ .5 & 0.0 \end{pmatrix}), (1/3, 2/3)). \tag{4}$$

OOMs specify stationary stochastic processes, if one puts

$$P[c_1 \dots c_k] = \sigma(\tau_{c_k} \cdots \tau_{c_1} w_0), \tag{5}$$

which corresponds to (3) and is the fundamental equation of OOM theory.

Processes which can be described by HMMs clearly are a subclass of those describable by OOMs. The inclusion is proper: stochastic processes exist which can be modeled by some OOM but not by any HMM.

2.2 OOMs as generators

OOMs can be used to algorithmically generate stochastic sequences. The procedure is completely different from the way how sequences are produced with HMMs. This shall now be demonstrated with the exemplary OOM (4).

Consider the state space $\mathbb{R}^2$ of (4). At time $t = 0$, the system is in state $w_0 = (1/3, 2/3)$ (fig. 3a). It must now be computed by which probability the operator τ_a (vs. τ_b) is to be applied on w_0 in the first time step. This is done by computing both $\tau_a w_0$ and $\tau_b w_0$. The conditions from definition 1 imply that $(\sigma(\tau_a w_0), \sigma(\tau_b w_0))$ is a probability vector (!). In our example, we find $(\sigma(\tau_a w_0), \sigma(\tau_b w_0)) = (\sigma((1/3, 1/2)), \sigma((1/6, 0))) = (5/6, 1/6)$. These probabilities are used to make a random choice between τ_a vs. τ_b. Let us assume the dice fall for τ_a. This means that the next state, w_1, is obtained by applying τ_a on w_0. An event $c_1 = a$ is observed at this moment.

Actually, w_1 is not $\tau_a w_0$ but $\tau_a w_0 / \sigma(\tau_a w_0)$, i.e. $\tau_a w_0$ is "renormalized" to an internal sum of 1 (fig. 3b). The reason is that internal sums stand for probabilities; but, after the decision for τ_a, the probability $\sigma(\tau_a w_0)$ has turned into certainty, i.e. an internal sum of 1.

This procedure can now be iterated, using w_1 instead of w_0, obtaining a next observable event c_2 and state w_2, etc.

Somewhat surprisingly, this generation procedure requires only a single random decision per time step. In a HMM, by contrast, two such decisions are necessary.

2.3 Theoretical and algorithmic properties of OOMs

In this subsection, I give a quick overview of the most important properties of OOMs, and compare them with the corresponding properties of HMMs.

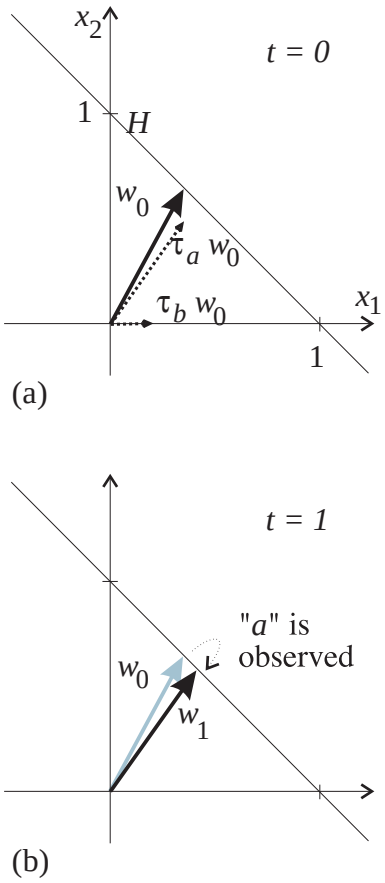

Figure 3: Using an OOM as a generator. Compare text for detail.

Model equivalence

Despite repeated efforts since the late 50ies, the question when two HMMs describe the same processe has only recently been fully answered [Ito *et al.*, 1992]. The proof given there is rather involved. The question of minimal-state HHMs, which is intimately related to the issue of model equivalence, is posed but not answered.

Model equivalence can be characterized much more transparently for OOMs. The central theorem states that two minimal-dimensional OOMs $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$, $\mathcal{B} = (\mathbb{R}^m, (\tau'_a)_{a\in\Sigma}, w'_0)$ are equivalent if and only if there exists an regular, internal-sum-preserving, linear mapping $\varrho : \mathbb{R}^m \to \mathbb{R}^m$ which maps w_0 on w'_0, and which transports each τ_a to τ'_a in the sense that $\tau'_a = \varrho \tau_a \varrho^{-1}$. A second, largely similar theorem states how any OOM can be transformed into a minimal-dimensional one. Taken together, both theorems fully characterize OOM equivalence, and answer the question of minimal models.

Interpretable OOMs

HMMs are appealing in that they can be interpreted in terms of two intuitively understandable subprocesses,

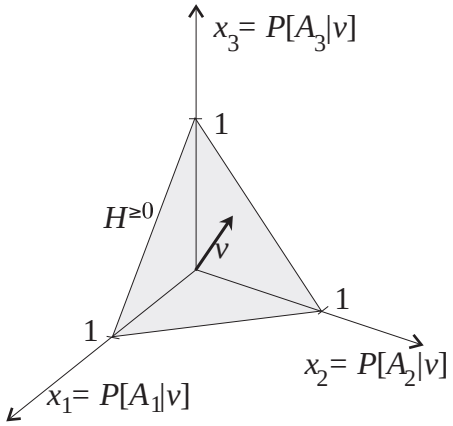

Figure 4: The positioning of $H^{\geq 0}$ within the state space.

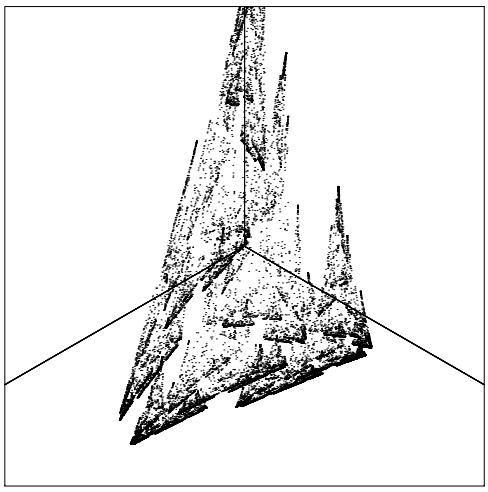

Figure 5: State plot of a 3-dimensional OOM. The drawing plane is $H^{\geq 0}$.

namely, a (hidden) Markov chain and probabilistic "measurement" of states.

OOMs affort an intuitive interpretation, too, albeit of a completely different kind. For an m-dimensional OOM and arbitrary k consider an arbitrary disjoint partitioning $\Sigma^k = A_1 \dot\cup \cdots \dot\cup A_m$ of the sequences of length k into m events. For instance, using again our example (4), consider the partitioning $\Sigma^2 = A_1 \dot\cup A_2 := \{aa\} \dot\cup \{ab, ba, bb\}$ into the events "no b occurs" and "at least one b occurs".

Then, there exists an OOM $\mathcal{A}(A_1, A_2) = (\mathbb{R}^2, (\tau'_a)_{a\in\Sigma}, w'_0)$ which (i) is equivalent to (4), and (ii) in which the two dimensions of the state space $\mathbb{R}^2$ can be interpreted in terms of the future probabilities of events A_1, A_2. I.e., when the OOM $\mathcal{A}(A_1, A_2)$ is in state $v = (x_1, x_2)$ at some time t, then the probability that during the next two time steps A_1 will occur (i.e. "no b") is x_1. Stated in formal terms, an interpretable OOM $\mathcal{A}(A_1, \ldots, A_m)$ is characterized by the fact $P[A_i \mid (x_1, \ldots, x_m)] = x_i$.

If we consider a 3-dimensional, interpretable OOM $\mathcal{A}(A_1, A_2, A_3)$, all system states fall into the triangular hyperplane region $H^{\geq 0}$ depicted in fig. 4. This gives rise to a standardized graphical "fingerprint" of 3-dimensional OOMs, if one plots state sequences that occur during a long run of such an OOMs. Typically, states lie on a fractal attractor, like the one depicted in fig. 5. OOMs of dimensions greater than 3 can be projected on three events and "fingerprinted" similarly.

Inducing models from data

The first and fundamental step in stochastic sequence analysis is model induction: given an empirical stochastic data sequence S, find a model that in some sense optimally accounts for it. The induction algorithms most commonly used for HMMs is the Baum-Welch algorithm [Rabiner, 1990]. A number of variants and alternatives have been developed [Elliott *et al.*, 1995] [Smyth *et al.*, 1997] [Baldi and Chauvin, 1994]. All of these techniques are gradient-descent algorithms, and may need simulated annealing schemes and manual pre-estimation of model structure for good results.

By contrast, OOMs can be induced from data with a constructive one-step procedure. The basic idea is to exploit that in an interpretable OOM $\mathcal{A}(A_1, \ldots, A_m)$, the invariant vector w_0 is $(P[A_1], \ldots, P[A_m])$. This vector can be estimated from an empirical sequence S by counting the frequencies of the events $A_1, \ldots, A_m$. Similarly, it holds that $\tau_{c_i} w_0 = (P[c_i A_1], \ldots, P[c_i A_m])$, $\tau_{c_j}\tau_{c_i} w_0 = (P[c_i c_j A_1], \ldots, P[c_i c_j A_m])$, etc. These vectors can likewise be estimated from data via simple frequency counting. I.e., one can estimate both certain vectors, and their images under the maps τ_{c_i}. Elementary linear algebra then provides simple means to reconstruct the τ_{c_i}. This involves mainly the inversion of a matrix of size $m \times m$. Thus, the computational cost of inducing an OOM from S essentially is the cost of a single, fixed-window inspection of S (which is necessary to count frequencies of certain events), plus the cost of a matrix inversion.

This estimation procedure is optimal in the following sense. If S (of length N) has been generated by some OOM $\mathcal{A}$, the OOM $\mathcal{B}$ recovered from S becomes exactly equivalent to $\mathcal{A}$ in the limit of $N \to \infty$.

A variant of this this algorithm allows to follow shifting sources. One may expect progress where HMMs have hitherto faced limitations, e.g. adapting on-line to changing speakers in automated speech recognition.

3 Abstract observable operator models

The idea of observable operators is *universally* applicable to stationary, stochastic processes, yielding a lucid linear-algebra framework for the investigation of such processes. I will first show in some detail how an abstract OOM can be obtained for any stationary, discrete-time, symbolic process, and then present the most general form of an OOM.

Let $(\Omega, \mathfrak{A}, P, (X_t)_{t\in\mathbb{Z}})$, or for short, (X_t) be a stationary, discrete-time stochastic process with values in a finite set Σ (notation for stochastic processes is taken from [Bauer, 1978]). Then, (X_t) is uniquely characterized by the distribution of finite subsequences, i.e. by all probabilities of the kind $P[d_1 \dots d_n]$, where $n \in \mathbb{N}$ and $d_1 \dots d_n \in \Sigma^n$. I shall use the shorthand notation $\bar{d}$ to denote such sequences.

By consequence, (X_t) is uniquely characterized by its *conditioned continuations*, i.e. by the conditioned probabilities $P[\bar{d} \mid \bar{c}] = P[\bar{c}\bar{d}]/P[\bar{c}]$, by which a sequence $\bar{c}$ is followed by $\bar{d}$. Note that the empty sequence $\bar{c} = \varepsilon$ is included here, which means that the unconditioned probabilities $P[\bar{d}] = P[\bar{d} \mid \varepsilon]$ are just a special case of conditioned continuations.

Collect all conditioned continuations of $\bar{c}$ into a numerical function

$$\begin{aligned} \mathfrak{g}_{\bar{c}} : \Sigma^* &\to \mathbb{R}, \\ \bar{d} &\mapsto P[\bar{d}\,|\,\bar{c}], \text{ if } P[\bar{c}] \neq 0 \\ &\mapsto 0, \text{ if } P[\bar{c}] = 0 \end{aligned} \tag{6}$$

The set $\{\mathfrak{g}_{\bar{c}} \mid \bar{c} \in \Sigma^*\}$ uniquely characterizes (X_t). Intuitively, the functions contained in this set specify the information obtainable from finite (possibly even empty) pasts $\bar{c}$ about finite futures $\bar{d}$.

Let $\mathfrak{D}$ denote the set of all functions from Σ^* into the reals, i.e. the numerical functions on words. $\mathfrak{D}$ can be viewed as a real vector space in a canonical fashion. Let $\mathfrak{G} = [\{\mathfrak{g}_{\bar{c}} \mid \bar{c} \in \Sigma\}]_{\mathfrak{D}}$ denote the linear subspace spanned by the conditioned continuations in $\mathfrak{D}$.

Let $\mathfrak{G}_0$ be a basis of $\mathfrak{G}$. Define, for every $a \in \Sigma$, a linear function $\mathfrak{t}_a : \mathfrak{G} \to \mathfrak{G}$ by putting $\mathfrak{t}_a(\mathfrak{g}_{\bar{c}}) := P[a \mid \bar{c}]\mathfrak{g}_{\bar{c}a}$ for all $\bar{c} \in \mathfrak{G}_0$. A straightforward calculation shows that in fact this definition carries over to any arguments for $\mathfrak{t}_a$, i.e., it holds that

$$\mathfrak{t}_a(\mathfrak{g}_{\bar{c}}) = P[a\,|\,\bar{c}]\mathfrak{g}_{\bar{c}a} \text{ for all } \bar{c} \in \mathfrak{G}. \tag{7}$$

The vector space $\mathfrak{G}$ takes the role of the space of internal states $\mathbb{R}^m$ known from previous sections, the family $(\mathfrak{t}_a)_{a\in\Sigma}$ is the abstract analogue of observable operators, and $\mathfrak{g}_\varepsilon$ corresponds to w_0. We say, $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\Sigma}, \mathfrak{g}_\varepsilon)$ is the abstract OOM of the process (X_t).

So far, this is just a lengthy definition. The usefulness of this construction lies in the fact that an analogue of (3) resp. (5) can be found. To this end, we recur to the representation of any $\mathfrak{d} \in \mathfrak{G}$ as its unique linear combination from basis vectors, i.e. $\mathfrak{d} = \sum_{\mathfrak{g}_{\bar{c}}\in\mathfrak{G}_0} \alpha_{\mathfrak{d}}^{\mathfrak{g}_{\bar{c}}} \mathfrak{g}_{\bar{c}}$, where only finitely many of the $\alpha_{\mathfrak{d}}^{\mathfrak{g}_{\bar{c}}}$ are nonzero. We define the internal sum of vectors from $\mathfrak{G}$ by putting

$$\sigma\mathfrak{d} := \sum_{\mathfrak{g}_{\bar{c}}\in\mathfrak{G}_0} \alpha_{\mathfrak{d}}^{\mathfrak{g}_{\bar{c}}} \tag{8}$$

It can be shown that this definition does not depend on the choice of the basis $\mathfrak{G}_0$. Then, it holds that

$$P[a_1, \dots, a_n] = \sigma(\mathfrak{t}_{a_n} \circ \dots \circ \mathfrak{t}_{a_1}\mathfrak{g}_\varepsilon) \tag{9}$$

The class of processes which we "concretely" described in the preceding sections, by abstracting away from HMMs, can now alternatively be characterized as the class of processes whose abstract OOMs are finite-dimensional.

The construction of abstract OOMs for stationary processes is not confined to discrete time or discrete-valued processes. Generalizing the above construction offers no difficulties. I conclude this section with the most general theorem currently available (the proof is an exercise of modest difficulty):

Proposition 1 *Let $(\Omega, \mathfrak{A}, P, (X_t)_{t\in\mathbb{R}})$ be a stationary process with values in a measure space $(B, \mathfrak{B})$. Then there exists an* observable operator model *$(V, (\tau_A^r)_{r\in\mathbb{R}, A\in\mathfrak{B}}, w_0)$ of this process, where V is a real vector space, τ_A^r are linear operators on V obeying*

1. $\tau^r_{\bigcup A_i} = \sum \tau^r_{A_i}$
2. $\tau_A^{r+s} = \tau_A^s \circ \tau_B^r$,

and $w_0 \in V$ obeying $\tau_B^r w_0 = w_0$ for all $r \in \mathbb{R}$, such that for all $t_1 < \dots < t_n$ it holds that

$$\begin{aligned} P[X_{t_1} \in A_1, \dots, X_{t_n} \in A_n] \quad = \\ \sigma(\tau_{A_n}^{t_n - t_{n-1}} \circ \dots \circ \tau_{A_2}^{t_2 - t_1} \circ \tau_{A_1}^{t_1} w_0). \end{aligned} \tag{10}$$

Loosely speaking, this theorem states that stationary stochastic processes can be described in terms of observable operators, where *(1)* disjoint unions of observed events translate to sums of operators, and *(2)* progression in time translates to concatenating them. In this way, the theory of stationary processes becomes part of linear algebra.

4 Discussion

In the introduction, I tried to give an intuitive account of observable operators by drawing fig.1(b). In fact, there is redundancy even in that simple sketch. The system states, which I rendered as little dots, can be

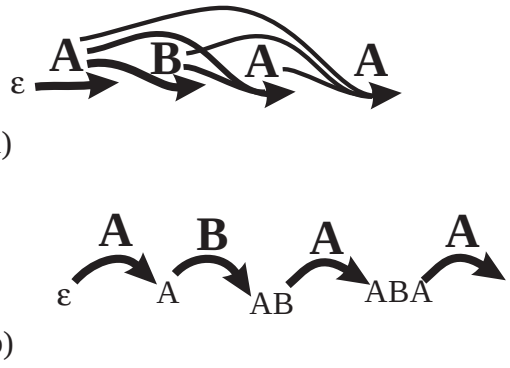

Figure 6: The really appropriate view on observable operator trajectories.

taken away! The story told by abstract OOMs is that observable operators need no "system states" other than the trajectory's own past. In order to see this, note that $\mathfrak{g}_{c_1 \ldots c_n}$, which is a system state on which the $\mathfrak{t}_a$ operate at time n, is nothing but $P[\,\cdot\,|\,c_1 \ldots c_n]$, i.e. the *state is the information given by the past about the future*. Graphically, therefore, an OOM trajectory should best be rendered without the little system state dots, as in fig. 6a. Alternatively, the dots could be interpreted as the system's memory of its own past (fig. 6b).

OOM models intimately blend linear algebra with probability theory. Furthermore, I believe that modern dynamical systems theory can be made to bear on OOM trajectories. The (apparently) fractal dynamics of "information states", exemplified in fig. 5, is a fascinating object for further studies.

Acknowledgments I feel deeply grateful toward Thomas Christaller for confidence and great support. My notion of trajectories grew richer by the minute during intense discussions with Rafael Nuñez. The research described in this paper was carried out while the author obtained a postdoctoral grant from the German National Research Center for Information Technology (GMD).

References

[Baldi and Chauvin, 1994] P. Baldi and Y. Chauvin. Smooth on-line learning algorithms for hidden Markov models. *Neural Computation*, 6:307–318, 1994.

[Bauer, 1978] H. Bauer. *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie*. de Gruyter, Berlin/New York, 3 edition, 1978. English translation: Probability Theory and Elements of Measure Theory, New York: Holt, Rinehart & Winston, 1972.

[Bengio, 1996] Y. Bengio. Markovian models for sequential data. Technical Report 1049, Dpt. d'Informatique et Recherche Opérationelle, Université de Montréal, 1996. http://www.iro.umontreal.ca/labs/neuro/pointeurs/hmmsTR.ps.

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

[Iosifescu and Theodorescu, 1969] M. Iosifescu and R. Theodorescu. *Random Processes and Learning*, volume 150 of *Die Grundlagen der mathematischen Wissenschaften in Einzeldarstellungen*. Springer Verlag, 1969.

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

[Jaeger, 1997a] H. Jaeger. Observable operator models and conditioned continuation representations. Arbeitspapiere der GMD 1043, GMD, Sankt Augustin, 1997. http://www.gmd.de/People/ Herbert.Jaeger/Publications.html.

[Jaeger, 1997b] H. Jaeger. Observable operator models II: Interpretable models and model induction. Arbeitspapiere der GMD 1083, GMD, Sankt Augustin, 1997. http://www.gmd.de/People/ Herbert.Jaeger/Publications.html.

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

[Smyth *et al.*, 1997] P. Smyth, D. Heckerman, and M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–270, 1997.

[Stengel, 1986] R.F. Stengel. *Stochastic Optimal Control*. John Wiley and Sons, 1986.

[Zadeh, 1969] L.A. Zadeh. The concept of system, aggregate, and state in system theory. In L.A. Zadeh and E. Polak, editors, *System Theory*, volume 8 of *Inter-University Electronics Series*, pages 3–42. McGraw-Hill, New York, 1969.