Observable Operator Processes and Conditioned Continuation Representations[1]

Herbert Jaeger
GMD, St. Augustin
herbert.jaeger@gmd.de

February 19, 1997

[1]This paper appeared in the series "Arbeitspapiere der GMD", Nr. 1043, January 1997, GMD, Sankt Augustin

Abstract: This article introduces *observable operator models* (OOM) and *conditioned continuation representations* (CCR). They are tightly interrelated models of certain stationary, finite-valued, discrete, stochastic processes. Both OOM's and CCR's are vector spaces equipped with a set of linear operators, which correspond to the observables of the process. However, while the vectors of OOM's are ordinary vectors from $\mathbb{R}^k$, the vectors of a CCR are probability distributions of continuations of the process given finite information about its past history. The CCR of a process (X_t) is uniquely determined, whereas there are many different OOM's that generate it. By mapping these OOM's on the CCR, a full characterization of minimal-dimension OOM's is derived. Hidden Markov models (HMM) are a proper subclass of OOM's. The results derived for OOM's can be adapted to yield characterization results for minimal-state HMM's.

Zusammenfassung: Dieser Aufsatz führt *observable operator models* (OOM) und *conditioned continuation representations* (CCR) ein. Es handelt sich dabei um zwei eng verwandte Modelle bestimmter stationärer, endlichwertiger, diskreter, stochastischer Prozesse. Sowohl OOMs als auch CCRs sind Vektorräume mit einer Menge linearer Operatoren, welche den beobachtbaren Ereignissen des Prozesses entsprechen. Die Vektoren von OOMs sind gewöhnliche Vektoren aus $\mathbb{R}^k$, wohingegen die Vektoren einer CCR jeweils Mengen bedingter Verteilungen von Fortsetzungen des Prozesses bei gegebener Information über eine endliche Vorgeschichte sind. Ist ein Prozess (X_t) gegeben, so ist seine CCR eindeutig bestimmt, wohingegen es viele verschiedene OOMs gibt, welche denselben Prozess generieren. Indem diese OOMs auf die CCR abgebildet werden, kann eine vollständige Charakterisierung der OOMs minimaler Dimension erreicht werden. Hidden Markov Modelle (HMM) sind eine echte Teilklasse von OOMs. Die Ergebnisse über OOMs führen auch zu vertieften Einsichten in HMMs, u.a. bezüglich der Charakterisierung äquivalenter HMMs mit minimaler Zustandsmenge.

1 Introduction

This article introduces *observable operator models* (OOMs). They are models of certain stationary, finite-valued, discrete, stochastic processes, among them hidden Markov processes, which they properly include as a subclass.

Hidden Markov models (HMM) are widely used, and it is not necessary here to emphasize their practical importance (cf. [4]). However, from a mathematical point of view, HMMs are not particularly well-behaved objects. Almost nothing is known, for instance, about equivalence of HMMs (in the sense of two such models generating the same observable process).

OOMs, by contrast, are mathematically quite transparent objects. An OOM $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a\in\Sigma}, w_0)$ consists of a (hidden) *state space*, which is taken to be $\mathbb{R}^k$, a family of linear operators $(\tau_a)_{a\in\Sigma}$, which is indexed with the (discrete) observables of the process, and a vector w_0 which essentially is a stationary distribution of the process. The operators are called *observable operators*, since sequences of applications of these operators (i.e., $\tau_{a_n} \circ \tau_{a_{n-1}} \circ \ldots \circ \tau_{a_0}$) correspond to finite-time observations of the process (i.e., to $a_0 a_1 \ldots a_n$).

This simple setup makes possible a transparent and effective characterization of equivalence in terms of certain linear mappings between the state spaces of different OOMs, which convey the observable operators of one OOM into the observable operators of the other.

Deriving these results is the first main theme of this paper. The derivation relies heavily on a certain unique vector space representation of stationary stochastic processes, its *conditioned continuation representation* (CCR). They are introduced in section 3, and they are applied to the characterization of OOM equivalence in section 4.

The second main theme of this article is to elucidate how HMMs can be characterized as a subclass of OOMs. Section 5 is devoted to this topic. It turns out that HMMs are special cases of OOMs in that they must satisfy a large number of numerical constraints on certain matrix entries. This finding helps to explain why a "nice" mathematical theory of HMMs does not seem to exist.

Since in fact I have been led to OOMs through thinking about HMMs, I will motivate and develop the definition of OOMs by abstracting away from HMMs (section 2). However, before proceeding with this programme, it may be helpful to informally describe a simple OOM.

Consider a stochastic process with values in $\Sigma = \{a, b\}$. The paths of this process are sequences of a's and b's. We will now see how an OOM can be used to *generate* such paths. Consider the OOM $\mathcal{A} = (\mathbb{R}^2, \{\tau_a, \tau_b\}, w_0)$. We skip the role of w_0 in this introduction. The observable operators τ_a and τ_b are linear operators on $\mathbb{R}^2$, i.e. they can be identified with 2×2-matrices.

These matrices must have certain properties, which I skip here, too. The following would be admissible observable operators:

$$\tau_a = \begin{pmatrix} 1/4 & 1 \\ 1/4 & 0 \end{pmatrix} \quad \tau_b = \begin{pmatrix} 1/4 & 0 \\ 1/4 & 0 \end{pmatrix}$$

These matrices can be used to generate stochastic sequences of a's and b's. The general idea is to apply τ_a, τ_b in a stochastic sequence on vectors from $\mathbb{R}^2$, thus generating a trajectory of a (hidden) dynamical system. The sequence of operators will then correspond directly to the sequence of observable events, while the (hidden) system states determine the probabilities by which one of the operators is selected at a given time step. More concretely, this works as follows.

Let us call the sum of components of a vector its *internal sum* σ, i.e., $\sigma((x, y)) = x + y$. Consider in $\mathbb{R}^2$ the hyperplane $H = \{(x, y) \in \mathbb{R}^2 \mid x + y = 1\}$ which consists of all vectors of internal sum 1 (cf. fig. 1). These vectors will be the possible (hidden) *states* of our OOM. In OOM theory, internal sums of vectors correspond to probabilities, thus the vectors in H can be interpreted as "fully ascertained" system states.

Assume that at time t_0, the system is in state $v_0 = (2/3, 1/3)$. Compute the vectors $\tau_a(v_0) = (1/2, 1/6), \tau_b(v_0) = (1/6, 1/6)$ (cf. fig.1(a)). Note that the internal sums of these vectors are values from $[0, 1]$ (namely, 4/6 and 2/6), and they sum to 1. Thus, the internal sums can be interpreted as probabilities (we will learn in section 3 that this is something natural and fundamental, not merely a technical trick). We put the probability that the operator τ_a will be selected for application, equal to $\sigma(\tau_a(v_0)) = 4/6$. Similarly, we determine the probability for an application of b to be 2/6.

Now we select at random one of the operators τ_a, τ_b, according to the probabilities just computed. In this example, the odds are for τ_a, so let's say the dice decide for this operator.

This selection makes certain that the system state at time t_1 will be, in a sense, $\tau_a(v_0)$. However, a system state which is determined with certainty should have internal sum 1. This is why we do not take $\tau_a(v_0)$ for the next system state, but we rather put $v_1 = \frac{\tau_a(v_0)}{\sigma(\tau_a(v_0))}$. I.e., we "renormalize" to internal sum 1 (cf. fig.1(b)). Intuitively, this renormalization corresponds to the fact that something has changed from a probability to certainty, due to the outcome of a random *decision*.

The same procedure is now repeated, with v_1 taking the role of v_0. We get $\tau_a(v_1) = (7/16, 3/16)$, $\tau_b(v_1) = (3/16, 3/16)$. Again, we randomly select one the operators with biases according to the internal sums of these vectors (note again that they are from $[0, 1]$ and sum to 1). Let's say the dice fell for τ_b. This would give us $v_2 = \frac{\tau_b(v_1)}{\sigma(\tau_b(v_1))} = (1/2, 1/2)$ at time t_2 (fig.1(c)).

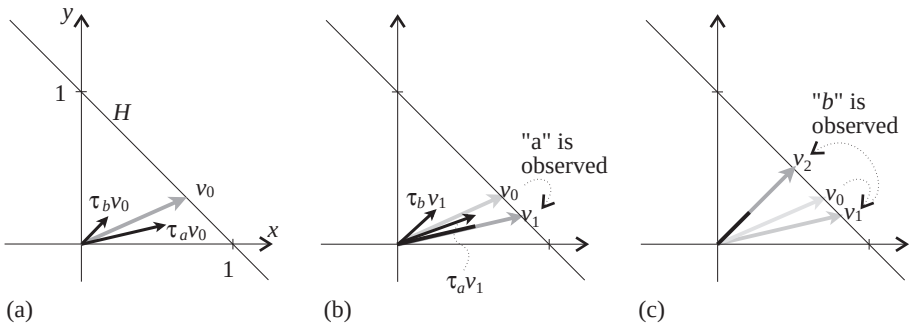

Figure 1: How an OOM generates a stochastic sequence. (a) Time t_0: probabilities of next obeservable event are computed. (b) From time t_0 to t_1: one of the events (a in this case) is chosen according to its probability, and a new state vector v_1 is produced. (c) From t_1 to t_2: procedure is repeated, in this example yielding event b.

So far, we have generated the sequence ab. It is obvious how this process could be continued indefinitely.

Incidentally, the process described in this example can be generated by a HMM, too.

In a nutshell, an OOM is characterized by two features:

1. The observed events $a \in \Sigma$ are interpreted as (linear) operators τ_a which change the state v of a dynamical system.
2. The probability $P[a \mid v]$ that in some state v the observation a will be made is equal to the contraction in internal sum of v induced by an application of τ_a.

With respect to the first point, OOM's are similar to *random systems with complete connections*, as described in [3]. However, the second point is absent from that theory. It is the dual nature of observable operators of (i) being the process observables and (ii) coding (in the state vector contraction) their own probability of getting selected, which makes OOM theory "crisp".

2 From HMMs to OOMs

In this section, we will motivate and develop the precise definition of OOMs, by some simple abstraction steps which start from hidden Markov models

(HMMs) and finally lead to OOMs.

It is assumed that the reader is familiar with the basic theory of Markov processes and linear algebra.

We start by fixing some HMM notation. There are various, more or less equivalent ways to formally define HMM's. We will stick to the kind of presentation which is probably the most widespread. For an alternative approach, which frames HMMs in Martingale theory, see [2].

Throughout this paper, we restrict the subject to processes with a finite-dimensional state space, discrete time, and a finite set of observables (i.e., an alphabet of symbols $\Sigma = a_1, \ldots, a_n$).

Definition 1 *A* hidden Markov model *is a quadruple* $(S, M, (O_a)_{a\in\Sigma}, P_0)$, *where* $S = \{s_1, \ldots, s_k\}$ *is a set of (hidden) states,* $M : S \times S \to [0, 1]$ *gives the state transition probabilities (satisfying* $\sum_{i=1}^{k} M(s, s_i) = 1$ *for all* $s \in S$*),* $O_a : S \to [0, 1]$ *specifies the probability that the event* a *is observed when the process is in hidden state* s *(satisfying* $\sum_{a\in\Sigma} O_a(s) = 1$ *for all* $s \in S$*), and* $P_0 : S \to [0, 1]$ *gives the starting distribution (satisfying* $\sum_{s\in S} P_0(s) = 1$*).*

It is customary to notate M as a stochastic matrix $M = (m_{ij})$, where $m_{ij} = M(s_i, s_j)$.

S can be interpreted as the k-dimensional real vector space $\mathbb{R}^k$, where vectors $(p_1, \ldots, p_k)$ from the subspace $[0, 1]^k$ are the state probability distributions of the HMM.

M gives rise to a Markov process (for details, consult e.g. [1]). M has a stationary distribution $P_{stat} : S \to [0, 1]$, i.e. a distribution vector which is a fixed point of the linear mapping given by M', where M' is the transpose of M:

$$M' P_{stat} = P_{stat} \tag{1}$$

In most cases of interest, P_{stat} will be unique (if it isn't, then the process essentially consists of decoupled subprocesses, which can be investigated in isolation. For details cf. [1], Theorem 2.4).

If we take P_0 to be such a (in most cases: *the*) stationary distribution P_{stat}, then $(S, M, (O_a)_{a\in\Sigma}, P_0)$ gives rise to a stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$, where the X_t are random variables with values in Σ. In this article, we will deal exclusively with stationary processes, i.e., we will assume that P_0 obeys (1).

The cylinder distributions of the process generated by $(S, M, (O_a)_{a\in\Sigma}, P_0)$,

$$\{P[X_{i+1} \in A_1, \ldots, X_{i+n} \in A_n] \mid i, n \in \mathbb{N}, A_1, \ldots, A_n \subseteq \Sigma\},$$

can be derived from the distributions of finite observations sequences,

$$\{P[X_{i+1} = a_1, \ldots, X_{i+n} = a_n] \mid i, n \in \mathbb{N}, a_1, \ldots, a_n \in \Sigma\},$$

which in turn can be computed from M, O_{a_i}, and P_0 in the following way (cf. [4]):

$$P[X_{i+1} = a_1, \ldots, X_{i+n} = a_n] = \sum_{s_1 \in S} P_0(s_1) O_{a_1}(s_1) \sum_{s_2 \in S} M(s_1, s_2) O_{a_2}(s_2) \ldots \sum_{s_n \in S} M(s_{n-1}, s_n) O_{a_n}(s_n). \tag{2}$$

Henceforward, we will use the simplified notation $P[a_1 \ldots a_n]$ instead of $P[X_{i+1} = a_1, \ldots, X_{i+n} = a_n]$, which is justified in stationary processes. We will also sometimes abbreviate a sequence $a_1, \ldots, a_n$ or a word $a_1 \ldots a_n$ to $\bar{a}$.

We will now describe a more transparent way of computing $P[a_1 \ldots a_n]$, which will turn out to be the decisive step toward OOM's.

First, we interpret each O_a as a linear mapping $O_a : \mathbb{R}^k \to \mathbb{R}^k$ of the state vector space onto itself, which is specified by the diagonal matrix

$$O_a = \begin{pmatrix} O_a(s_1) & & 0 \\ & \ddots & \\ 0 & & O_a(s_k) \end{pmatrix} \tag{3}$$

We will not distinguish in notation between the matrix and the linear mapping specified by it, i.e. both will be written as O_a. Intuitively, O_a simply weighs each state with the corresponding observation probability of a.

Now we define a new linear operator T_a on $\mathbb{R}^k$ by the concatenation of O_a with M':

$$\begin{aligned} T_a &: S^k \to S^k \\ T_a(v) &= M' O_a(v), \end{aligned} \tag{4}$$

where v is a vector from $\mathbb{R}^k$ in column notation. The mapping T_a is specified by the matrix $M'O_a$. Again, we will not distinguish in notation between the mapping and the matrix.

Definition 2 *For a vector* $v = (v_1, \ldots, v_k) \in \mathbb{R}^k$ *let*

$$\sigma(v) := v_1 + \ldots + v_k \tag{5}$$

denote its internal sum.

Using the internal sum σ, we can compute $P[a_1 \ldots a_n]$ in the following way:

$$P[a_1 \ldots a_n] \;=\; \sigma(T_{a_n} T_{a_{n-1}} \cdots T_{a_1} P_0) \tag{6}$$

That (6) gives indeed the same value as (2), can be verified by elementary transformations, using that (2) is equivalent to the following expression:

$$\begin{aligned} P[a_1 \ldots a_n] = \\ \sum_{s_1 \in S} P_0(s_1) \Big(O_{a_1}(s_1) \sum_{s_2 \in S} M(s_1, s_2) \Big(O_{a_2}(s_2) \ldots \\ \ldots \sum_{s_n \in S} M(s_{n-1}, s_n) \Big(O_{a_n}(s_n) \sum_{s_{n+1} \in S} M(s_n, s_{n+1}) \Big) \cdots \Big) \Big) \end{aligned} \tag{7}$$

The matrix M' can be recovered from the T_a. Observing that $\sum_{a \in \Sigma} O_a = 1$ yields the identity matrix, we find:

$$M' \;=\; \sum_{a \in \Sigma} T_a \tag{8}$$

The observations made so far motivate the following definition, which generalizes from HMM's by retaining only those properties of the T_a which are necessary to guarantee that some process (X_t) can ultimately be specified:

Definition 3 *An* observable operator model *is a triple* $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a \in \Sigma}, w_0)$, *where* $\mathbb{R}^k$ *the OOM's* state space, Σ *is a finite alphabet, the* τ_a *are linear mappings on* $\mathbb{R}^k$, *and* $w_0 \in \mathbb{R}^k$, *such that the following conditions are satisfied:*

1. *the matrix* $\mu = \sum_{a \in \Sigma} \tau_a$ *has column sums equal to 1, i.e.* $\mu_{1j} + \cdots + \mu_{kj} = 1$ *for all columns* j,
2. $\sigma(w_0) = 1$ *and* $\mu w_0 = w_0$,
3. *for every finite sequence* $a_1, \ldots, a_n$ *of elements of* Σ, *it holds that* $\sigma(\tau_{a_n} \cdots \tau_{a_1} w_0) \in [0, 1]$.

The mappings τ_a *are called the* observable operators *of* $\mathcal{A}$.

In this definition, the τ_a's are the OOM versions of the T_a, w_0 corresponds to P_0, and μ to M'.

The name, *observable operator model*, is meant to characterize the fact that a sequence of observed events $a_1, \ldots, a_n$ can be interpreted as a sequence of linear operations. I will further comment on the intuitive interpretation of these operators in the next section.

We do not require μ to be (the transpose of) a stochastic matrix. We don't restrict the elements of μ to values in $[0, 1]$ but allow $\mu_{i,j} \in \mathbb{R}$. In fact, this generality will turn out to be essential for the theory to be developed in this article. In a similar vein, we do not require the vector w_0 (which corresponds to P_0) to be a probability distribution. This vector is allowed to have negative components, too.

Generally, we use capital latin letters in the "world" of HMM's, and equivalent greek letters in the "world" of OOM's (like M vs. μ etc.).

Note that the OOM counterpart of the transposed matrix M' is the matrix μ, which is not transposed. The reason for this slight discrepancy lies in the tradition to notate stochastic matrices as transposes of the matrices which actually coorrespond to the linear mapping of the state transitions. I did not wish to perpetuate this somewhat unfortunate state of affair into the OOM world.

Proposition 1 *An OOM* $(\mathbb{R}^k, (\tau_a)_{a\in\Sigma}, w_0)$ *generates a stationary stochastic process* (X_t) *with values in* Σ*, if we define (like in (6)):*

$$P[a_1 \ldots a_n] \;=\; \sigma(\tau_{a_n} \tau_{a_{n-1}} \cdots \tau_{a_1} w_0) \tag{9}$$

Proof. We first have to show that the probabilities $P[a_1 \ldots a_n]$ of sequences of a given length add up to 1, i.e. that $\sum_{a_1,\ldots,a_n\in\Sigma} P[a_1 \ldots a_n] = 1$. This follows from $\sigma(w_0) = 1$ and the fact that columns in μ sum to 1, which implies that $\sigma(\mu v) = \sigma(v)$ for every vector v in $\mathbb{R}^k$.

Furthermore, we have to show that the probabilities of sequences of different length, computed according to (9), agree with each other in the following sense:

1. For all sequences $a_1, \ldots, a_n$, for all $m \in \mathbb{N}$, it holds that $P[a_1 \ldots a_n] = \sum_{b_1,\ldots,b_m\in\Sigma} P[b_1 \ldots b_m a_1 \ldots a_n]$.
2. For all sequences $b_1, \ldots, b_m$, for all $n \in \mathbb{N}$, it holds that $P[b_1 \ldots b_m] = \sum_{a_1,\ldots,a_n\in\Sigma} P[b_1 \ldots b_m a_1 \ldots a_n]$.

Both claims are easily verified. For the first, use that w_0 is a fixed point of μ; for the second, again exploit that columns in μ sum to 1. $\square$

Since in the sequel we will frequently use terms like on the rhs. in (9) or in the rhs. of (6), we introduce the following shorthand notation:

Definition 4 *Let $\tau_{a_1 \ldots a_n}$ denote the concatenation of mappings $\tau_{a_n} \circ \ldots \circ \tau_{a_1}$, and $T_{a_1 \ldots a_n}$ the concatenation of mappings $T_{a_n} \circ \ldots \circ T_{a_1}$.*

Note that the ordering of a_i's is reversed in the shorthand with respect to the full concatenation notation.

An (X_t)-OOM $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a \in \Sigma}, w_0)$ can be used to effectively generate left-finite paths $c_0, c_1, c_2, \ldots$, which are distributed according to (X_t), by the following inductive procedure.

1. Initialization:
   (a) For all $a \in \Sigma$, compute $P[a]$.
   (b) Randomly select one $a \in \Sigma$, with probability $P[a]$. Put $c_0 := a$.
2. Induction step:
   (a) If the finite sequence $c_0, c_1, \ldots, c_n =: \bar{c}$ has already been generated, compute, for all $a \in \Sigma$, the conditioned probability that a appears in the next step, $P[a \mid \bar{c}]$.
   (b) Randomly select one $a \in \Sigma$, with probability $P[a \mid \bar{c}]$. Put $c_{n+1} := a$.

This procedure obviousy yields paths which are distributed according to (X_t). The probabilities $P[a]$ and $P[a \mid \bar{c}]$ can be effectively computed from the operators $(\tau_a)_{a \in \Sigma}$ if we observe (9) and the fact that

$$P[a \mid \bar{c}] = \frac{P[\bar{c}a]}{P[\bar{c}]}.$$

Note that the case $P[\bar{c}] = 0$ cannot occur in this procedure.

These conditional probabilities can be computed iteratively in a way which requires only one matrix multiplication per time step, plus the computation of $\mid \Sigma \mid$ many scalar products, in the following way. Let $\tau_a =: (\tau_a^{ij})$. For every $a \in \Sigma$, define a vector

$$x_a := \begin{pmatrix} \sum_{i=1,\ldots,k} \tau_a^{i1} \\ \vdots \\ \sum_{i=1,\ldots,k} \tau_a^{ik} \end{pmatrix}$$

which contains the column sums of τ_a. Then, it is easy to see that for every $v \in \mathbb{R}^k$, it holds that $\sigma(\tau_a(v)) = \langle x_a, v \rangle$. Exploiting this fact, the above procedure can be detailed as follows:

1. Initialization:
   (a) Put $s_0 := w_0$.
   (b) For all $a \in \Sigma$, compute $P[a] = \langle x_a, s_0 \rangle$.
   (c) Randomly select one $a \in \Sigma$, with probability $P[a]$. Put $c_0 := a$.
2. Induction step:
   (a) If the finite sequence $c_0, c_1, \ldots, c_n =: \bar{c}$ has already been generated, and s_n has been computed, compute, for all $a \in \Sigma$, the conditioned probability that a appears in the next step, $P[a \mid \bar{c}] = \langle x_a, s_n \rangle$.
   (b) Randomly select one $a \in \Sigma$, with probability $P[a \mid \bar{c}]$. Put $c_{n+1} := a$.
   (c) Put $s_{n+1} := \frac{\tau_a(s_n)}{\sigma(\tau_a(s_n))}$.

It is noteworthy that this procedure makes do with a single random selection operation per time step, whereas in classical HMM methods two such operations are needed (one for the the hidden state transition, another for the determination of an observable event in the current state).

An immediate question that arises is whether different OOM's exist which generate the same process (X_t), and if so, how they can be transformed into each other. This kind of question, although it is central for a mathematical understanding of any kind of sequence generating algorithms, seems not to have been investigated for HMMs. The remainder of this article is mainly concerned with answering this question for OOMs (sections 3 and 4), and in using the results obtained for OOMs in order to partially answer the question for HMMs (in section 5).

First we equip ourselves with a definition of equivalence:

Definition 5 *Two OOM's* $(\mathbb{R}^k, (\tau_a)_{a\in\Sigma}, w_0)$ *and* $(\mathbb{R}^l, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ *are* equivalent *if they generate the same process* (X_t).

Equivalence in the sense of this definition obviously is an equivalence relation on the class of OOM's. We will sometimes refer to an OOM which generates (X_t), as an "(X_t)-OOM".

The question, then, is to characterize the class of (X_t)-OOMs for a given (X_t). A way to tackle this question is to find some "normal form" representation of OOM's, such that OOM's that generate the same (X_t) yield the same "normal form" representation. The latter can then be used as a means of comparison to establish the connections between equivalent OOM's.

In the next section, we will provide ourselves with such a "normal form" representation.

3 Conditioned continuation representations of stationary processes

In this section, we develop an essentially unique representation of a stationary process (X_t), its *conditioned continuation representation* (CCR). It will be a vector space whose vectors are constructed from conditioned probability distributions, which are derived directly from (X_t), and therefore are uniquely determined. Since CCR's can be considered to be the most important theoretical contribution of this article, with potential applications beyond OOM theory, we will define them in a slightly more general way than would be required for a treatment of OOM equivalence.

We will reserve $\mathfrak{Fraktur}$ letters for denoting CCR-related entities.

Assume that some stationary, discrete process $(X_t)_{t\in\mathbb{Z}}$ with values in a finite alphabet Σ is given. Recall that Σ^* customarily denotes the set of all finite words made from symbols from Σ, including the empty word ε, and that the length of a word w is denoted by $\mid w \mid$. We will now introduce, in several steps, the vectors which we will need for building CCRs.

Definition 6 *A* generalized word distribution *is any mapping* $\mathfrak{d} : \Sigma^* \to \mathbb{R}$ *which satisfies*

$$\exists\, r \in \mathbb{R}\ \forall n \geq 0 \sum_{w\in\Sigma^*, |w|=n} \mathfrak{d}(w) = r \tag{10}$$

The set of all generalized word distributions is denoted by $\mathfrak{D}$.

The term, "generalized word distribution", is motivated by (10), since it reminds one of probability distributions if $r = 1$.

Proposition 2 *$\mathfrak{D}$ can be interpreted as a real vector space (of infinite dimension) if we define scalar multiplication by $(x\mathfrak{d})(w) := x(\mathfrak{d}(w))$ and addition by $(\mathfrak{d} + \mathfrak{d}')(w) := \mathfrak{d}(w) + \mathfrak{d}'(w)$ for every $w \in \Sigma^*$.*

Proof. Obvious.

We will henceforth let $\mathfrak{D}$ denote this vector space.
The following is an analogue to definition 2:

Definition 7 *For $\mathfrak{d} \in \mathfrak{D}$, let*

$$\sigma(\mathfrak{d}) \quad := \quad r,$$

where r is defined as in definition 6, be the internal sum *of $\mathfrak{d}$.*

Next we introduce a metric d on $\mathfrak{D}$:

Definition 8 *For $\mathfrak{d}_1, \mathfrak{d}_2 \in \mathfrak{D}$, define the* distance *$d$ by*

$$d(\mathfrak{d}_1, \mathfrak{d}_2) = \sup_{w \in \Sigma^*} \min\{1, \|\mathfrak{d}_1(w) - \mathfrak{d}_2(w)\|\},$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}$.

Proposition 3 *d is a metric.*

Proof. It is obvious that $d(\mathfrak{d}_1, \mathfrak{d}_2) = 0$ iff $\mathfrak{d}_1 = \mathfrak{d}_2$, and it is also obvious that d is symmetric. Thus it remains to show that the triangle inequality holds, i.e. that $d(\mathfrak{d}_1, \mathfrak{d}_3) \leq d(\mathfrak{d}_1, \mathfrak{d}_2) + d(\mathfrak{d}_2, \mathfrak{d}_3)$ for all $\mathfrak{d}_1, \mathfrak{d}_2, \mathfrak{d}_3 \in \mathfrak{D}$:

$$\begin{aligned}
& d(\mathfrak{d}_1, \mathfrak{d}_3) \\
&= \sup_{w \in \Sigma^*} \min\{1, \|\mathfrak{d}_1(w) - \mathfrak{d}_3(w)\|\} \\
&\leq \sup_{w \in \Sigma^*} \min\{1, \|\mathfrak{d}_1(w) - \mathfrak{d}_2(w)\| + \|\mathfrak{d}_2(w) - \mathfrak{d}_3(w)\|\} \\
&\leq \sup_{w \in \Sigma^*} \min\{1, \|\mathfrak{d}_1(w) - \mathfrak{d}_2(w)\|\} + \sup_{w \in \Sigma^*} \min\{1, \|\mathfrak{d}_2(w) - \mathfrak{d}_3(w)\|\} \\
&\leq d(\mathfrak{d}_1, \mathfrak{d}_2) + d(\mathfrak{d}_2, \mathfrak{d}_3) \qquad \square
\end{aligned}$$

For later use, we note the following inequality:

Proposition 4 *For all $\mathfrak{d}_1, \ldots, \mathfrak{d}_4 \in \mathfrak{D}$ it holds that*

$$d(\mathfrak{d}_1 + \mathfrak{d}_2, \mathfrak{d}_3 + \mathfrak{d}_4) \leq d(\mathfrak{d}_1, \mathfrak{d}_3) + d(\mathfrak{d}_2, \mathfrak{d}_4)$$

Proof.

$$\begin{aligned} & d(\mathfrak{d}_1+\mathfrak{d}_2,\mathfrak{d}_3+\mathfrak{d}_4) \quad = \\ & = \sup_{w\in\Sigma^*} \min\{1, \|(\mathfrak{d}_1+\mathfrak{d}_2)(w)-(\mathfrak{d}_3+\mathfrak{d}_4)(w)\|\} \\ & = \sup_{w\in\Sigma^*} \min\{1, \|\mathfrak{d}_1(w)+\mathfrak{d}_2(w)-\mathfrak{d}_3(w)-\mathfrak{d}_4(w)\|\} \\ & \leq \sup_{w\in\Sigma^*} \min\{1, \|\mathfrak{d}_1(w)-\mathfrak{d}_3(w)\|+\|\mathfrak{d}_2(w)-\mathfrak{d}_4(w)\|\} \\ & \leq d(\mathfrak{d}_1,\mathfrak{d}_3)+d(\mathfrak{d}_2,\mathfrak{d}_4) \qquad \square \end{aligned}$$

We will now single out a certain linear subspace $\mathfrak{G}$ of $\mathfrak{D}$ as the vector space for our desired CCR.

Definition 9 *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary, discrete process with values in Σ. The* conditioned continuation distribution *of $(X_t)_{t\in\mathbb{Z}}$ is the mapping $\mathfrak{g}:\Sigma^*\to\mathfrak{D}$ defined by*

$$\begin{aligned} & (\mathfrak{g}(a_1\ldots a_n))(b_1\ldots b_m) = \\ & \quad P[X_{n+1}=b_1,\ldots,X_{n+m}=b_m \mid X_1=a_1,\ldots,X_n=a_n], \end{aligned} \tag{11}$$

where $P[X_{n+1}=b_1,\ldots,X_{n+m}=b_m \mid X_1=a_1,\ldots,X_n=a_n]$ is the conditioned probability that we observe the sequence $b_1,\ldots,b_m$ directly after the sequence $a_1,\ldots,a_n$. We will briefly write $\mathfrak{g}_{a_1\ldots a_n}$ or even $\mathfrak{g}_{\bar{a}}$ for $\mathfrak{g}(a_1\ldots a_n)$.

In cases where $P[\bar{a}]=0$ (where (11) is not well-defined), we put

$$\forall \bar{b}\in\Sigma^*: \quad \mathfrak{g}_{\bar{a}}(\bar{b}) := 0.$$

Note that $\mathfrak{g}(\varepsilon)(b_1\ldots b_m)=P[b_1,\ldots,b_m]$, where ε denotes the empty word.

The image of the conditioned continuation distribution of $(X_t)_{t\in\mathbb{Z}}$ spans a linear subspace in $\mathfrak{D}$, which is our desired vector space $\mathfrak{G}$.

Note that $\mathfrak{G}$ is completely determined by $(X_t)_{t\in\mathbb{Z}}$. We are therefore entitled to call $\mathfrak{G}$ the *conditioned continuation space of* (X_t), and to write $\mathfrak{G}(X_t)$ if we want to make explicit from which stationary process $\mathfrak{G}$ is derived.

The following proposition follows directly from the definitions of $\mathfrak{g}_{\bar{c}}$ and σ:

Proposition 5 *1.* $\forall \bar{c}\in\Sigma^* \quad \sigma(\mathfrak{g}_{\bar{c}})=1$

2. $\forall \bar{c}_1,\ldots,\bar{c}_n\in\Sigma^*\ \forall \alpha_1,\ldots,\alpha_n\in\mathbb{R} \quad \sigma(\sum_{i=1,\ldots,n}\alpha_i\mathfrak{g}_{\bar{c}_i})=\sum_{i=1,\ldots,n}\alpha_i$ $\square$

We now define CCR analogues of the observable operators $(\tau_a)_{a\in\Sigma}$.

Definition 10 *Let $\mathfrak{G}$ be the conditioned continuation space of (X_t). For each $a \in \Sigma$, let $\mathfrak{t}_a : \mathfrak{G} \to \mathfrak{G}$ be a mapping which satisfies*

$$\forall \bar{c} \in \Sigma^* : \qquad \mathfrak{t}_a(\mathfrak{g}_{\bar{c}}) = P[a \mid \bar{c}]\, \mathfrak{g}_{\bar{c}a} \tag{12}$$

A pair $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\Sigma})$ is called a conditioned continuation representation *(CCR) of (X_t). A family $(\mathfrak{t}_a)_{a\in\Sigma}$ of operators satisfying (12) is called a family of* observable operators.

A CCR of (X_t) is not uniquely determined, since (12) fixes $\mathfrak{t}_a$ only on a subset of $\mathfrak{G}$. However, in many instances where CCR's would be practically used, a canonical extension of $\mathfrak{t}_a$ on the entire space $\mathfrak{G}$ will exist, which will lead to a unique CCR. (In this article, we'll see that the requirement of linearity gives us such a canonical extension).

We adapt definition 4 to CCR terminology:

Definition 11 *Let $\mathfrak{t}_{a_1\ldots a_n}$ denote the concatenation of mappings $\mathfrak{t}_{a_n} \circ \cdots \circ \mathfrak{t}_{a_1}$.*

We end the technical part of this section with a proposition which is not in itself very interesting, but which we will later need.

Proposition 6 *Let $\mathfrak{g}_{\bar{c}} = \sum_{i=1,\ldots,m} \gamma_i \mathfrak{g}_{\bar{c}_i}$. Then for all $a \in \Sigma$ it holds that*

$$P[a \mid \bar{c}]\mathfrak{g}_{\bar{c}a} = \sum_{i=1,\ldots,m} \gamma_i P[a \mid \bar{c}_i]\mathfrak{g}_{\bar{c}_i a}.$$

Proof. We have to show that for all $\bar{d} \in \Sigma^*$, it holds that

$$P[a \mid \bar{c}]\mathfrak{g}_{\bar{c}a}(\bar{d}) = \sum_{i=1,\ldots,m} \gamma_i P[a \mid \bar{c}_i]\mathfrak{g}_{\bar{c}_i a}(\bar{d})$$

This is revealed by the following transformations:

$$\begin{aligned} & P[a \mid \bar{c}]\mathfrak{g}_{\bar{c}a}(\bar{d}) \quad = \\ & = \quad P[a \mid \bar{c}]P[\bar{d} \mid \bar{c}a] \;=\; P[a\bar{d} \mid \bar{c}] \\ & = \quad \mathfrak{g}_{\bar{c}}(a\bar{d}) \;=\; \sum_{i=1,\ldots,m} \gamma_i \mathfrak{g}_{\bar{c}_i}(a\bar{d}) \\ & = \quad \sum_{i=1,\ldots,m} \gamma_i P[a \mid \bar{c}_i]\mathfrak{g}_{\bar{c}_i a}(\bar{d}) \qquad \square \end{aligned}$$

The vectors $\mathfrak{g}_{\bar{c}}$, and the operators $\mathfrak{t}_a$, can each be interpreted in two complementary ways. I shall briefly explain them, although these "meta" considerations are not needed in the sequel.

First, $\mathfrak{g}_{c_1 \ldots c_n}$ can be considered as the *information* that an observer has about the future of the process after he has observed the finite sequence $c_1, \ldots, c_n$. The operators $(\mathfrak{t}_a)_{a \in \Sigma}$ should be interpreted accordingly as "information gain operators" which specify how additional information is gained through an observation of a after some $\bar{c}$.

Second, it is possible with certain processes (X_t) to interpret $\mathfrak{g}_{c_1 \ldots c_n}$ as an approximation of the state of a system which realizes (X_t). In order to make this interpretation valid, sequences of the kind $\mathfrak{g}_{\varepsilon}, \mathfrak{g}_{c_0}, \mathfrak{g}_{c_{-1} c_0}, \mathfrak{g}_{c_{-2} c_{-1} c_0}, \ldots$ should converge to a definite vector $\mathfrak{g}_{\ldots c_{-2} c_{-1} c_0}$ almost certainly (It is a natural guess that ergodic systems show this kind of convergence). Intuitively, then, $\mathfrak{g}_{\ldots c_{-2} c_{-1} c_0}$ can be interpreted as the "true" system state at time t_0, after it has gone through the (infinite) past history $\ldots, c_{-2}, c_{-1}, c_0$. This interpretation is in perfect agreement with a standard notion in theoretical physics of a system state, which says that the state of a system is what "in" the system determines its future development. Accordingly, $\mathfrak{t}_a$ could be interpreted as a time step operator on system states. The finite-past-history vectors $\mathfrak{g}_{c_1 \ldots c_n}$ can be considered as approximations to the "true" system state.

4 On the equivalence of OOM's

In this section, we will investigate the equivalence of OOM's. It turns out that the classes of minimal-dimension, equivalent OOM's can be completely, and very simply, characterized, and that non-minimal-dimension OOM's can be mapped by a internal sum-preserving map on minimal-dimensional, equivalent ones. These are quite strong results. Their derivation relies heavily on mapping OOM's on CCRs and vice versa. This is why we start by investigating mappings of OOM's on CCR's.

Definition 12 *Let $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a \in \Sigma}, w_0)$ be an OOM which (according to proposition 1) generates a process (X_t). Let $(\mathfrak{G}, (\mathfrak{t}_a)_{a \in \Sigma})$ be a CCR of (X_t). Let $\mathfrak{,} \ = \mathfrak{,} \ (\mathcal{A})$ be the linear subspace of $\mathbb{R}^k$ spanned by the vectors $\{\tau_{a_1 \ldots a_n} w_0 \mid n \in \mathbb{N}, a_1 \ldots a_n \in \Sigma^n\}$. Let $B(\mathfrak{,}\ ) = \{\tau_{\bar{b}_1} w_0, \ldots, \tau_{\bar{b}_l} w_0\}$ be a basis of $\mathfrak{,}$. Define a linear mapping π from $\mathfrak{,}$ to $\mathfrak{G}$ by putting*

$$\pi(\tau_{\bar{b}_i} w_0) := P[\bar{b}_i]\, \mathfrak{g}_{\bar{b}_i} \qquad i = 1, \ldots, l$$

This mapping is called the canonical projection *of $(\mathbb{R}^k, (\tau_a)_{a \in \Sigma}, w_0)$ onto $(\mathfrak{G}, (\mathfrak{t}_a)_{a \in \Sigma})$.*

We will now collect some properties of this canonical projection which show that it deserves its name.

Proposition 7 *Given the terminology and assumptions of definition 12, the canonical projection has the following properties:*

1. $\forall \bar{c} \in \Sigma^* \quad \pi(\tau_{\bar{c}} w_0) = P[\bar{c}]\, \mathfrak{g}_{\bar{c}}$.

2. π *is surjective.*

3. π *preserves* σ*, i.e.*

$$\forall v \in \, , \, : \sigma(\pi(v)) \quad = \quad \sigma(v)$$

4. π *is continuous with respect to the ordinary Euclidean metric in* $\, , \,$ *and the metric* d *(cf. definition 8) in* $\mathfrak{G}$.

Proof. 1. We have to show that for all $\bar{d} \in \Sigma^*$, it holds that

$$(\pi(\tau_{\bar{c}} w_0))(\bar{d}) = P[\bar{c}]\, \mathfrak{g}_{\bar{c}}(\bar{d}) \tag{13}$$

Let $\tau_{\bar{c}} w_0 = \sum_{i=1,\ldots,l} \alpha_i \tau_{\bar{b}_i} w_0$ be the linear combination of $\tau_{\bar{c}} w_0$ from basis vectors.

Then, a sequence of elementary transformations yields the desired result:

$$\begin{aligned}
&(\pi(\tau_{\bar{c}} w_0))(\bar{d}) \quad = \\
&= \quad \left(\pi\left(\sum \alpha_i \tau_{\bar{b}_i} w_0\right)\right)(\bar{d}) \quad = \quad \sum \alpha_i\, P[\bar{b}_i]\, \mathfrak{g}_{\bar{b}_i}(\bar{d}) \\
&= \quad \sum \alpha_i\, P[\bar{b}_i]\, P[\bar{d} \mid \bar{b}_i] \quad = \quad \sum \alpha_i\, P[\bar{b}_i \bar{d}] \\
&= \quad \sum \alpha_i \sigma(\tau_{\bar{b}_i \bar{d}} w_0) \quad = \quad \sigma(\tau_{\bar{d}} \left(\sum \alpha_i \tau_{\bar{b}_i} w_0\right)) \\
&= \quad \sigma(\tau_{\bar{d}}(\tau_{\bar{c}} w_0)) \quad = \quad P[\bar{c}\bar{d}] \quad = \quad P[\bar{c}]\, P[\bar{d} \mid \bar{c}] \\
&= \quad P[\bar{c}]\, \mathfrak{g}_{\bar{c}}(\bar{d})
\end{aligned}$$

2. Is a direct consequence of 1.

3. $\sigma(\pi(\tau_{\bar{b}_i} w_0)) = \sigma(P[\bar{b}_i]\, \mathfrak{g}_{\bar{b}_i}) = P[\bar{b}_i] = \sigma(\tau_{\bar{b}_i} w_0)$.

4. We have to show that $(\pi(v_j))_{j \in \mathbb{N}}$ converges in $\mathfrak{G}$ to $\pi(v)$ w.r.t. d if $(v_j)_{j \in \mathbb{N}}$ converges in $\, , \,$ w.r.t. $\| \cdot \|$.

Let $v_j = \sum \alpha_i^j \tau_{\bar{b}_i} w_0, v = \alpha_i \tau_{\bar{b}_i} w_0$ be the linear combinations of v_j, v from basis vectors of $\, , \,$. Then, $\lim_{j \to \infty} v_j = v$ is equivalent to $\lim_{j \to \infty} \alpha_i^j = \alpha_i$.

We consider for each $i = 1, \ldots, l$ the sequence

$$(\pi(\alpha_i^j \tau_{\bar{b}_i} w_0))_{j\in\mathbb{N}} = (\alpha_i^j P[\bar{b}_i]\mathfrak{g}_{\bar{b}_i})_{j\in\mathbb{N}}.$$

It is easy to see that

$$\lim_{j\to\infty} d(\alpha_i^j P[\bar{b}_i]\mathfrak{g}_{\bar{b}_i}, \alpha_i P[\bar{b}_i]\mathfrak{g}_{\bar{b}_i}) = 0.$$

An obvious application of proposition 4 finishes the proof. □

The following proposition is an immediate consequence of proposition 7:

Proposition 8 *Given the terminology and assumptions from definition 12, the dimension of* $\mathfrak{G}$ *is finite, and* $dim(\mathfrak{G}) \leq k$. □

Next we see how the observable operators τ_a from $\mathcal{A}$ carry over via the canonical projection into observable operators $\mathfrak{t}_a$ on $\mathfrak{G}$. We proceed in two steps. First, we use a version of τ_a restricted to the kernel of π to show the existence of linear observable operators $\mathfrak{t}_a$ on $\mathfrak{G}$. Then we show that these $\mathfrak{t}_a$ can be understood as the image of τ_a on the full space , .

Given the terminology and assumptions from definition 12, let $\{\mathfrak{g}_{\bar{c}_1}, \ldots, \mathfrak{g}_{\bar{c}_m}\}$ be a basis of $\mathfrak{G}$. Using proposition 7(1), we see that $\{\tau_{\bar{c}_1} w_0, \ldots, \tau_{\bar{c}_m} w_0\}$ is a basis of a linear subspace $G \subseteq$, which is complementary to the kernel $ker\,\pi$ of π. It holds that the reduct π_G of π on G yields an isomorphism of vector spaces $\pi_G : G \simeq \mathfrak{G}$. In particular, π_G is bijective. Thus, for all $a \in \Sigma$ we can define mappings $(\pi\tau)_a : \mathfrak{G} \to \mathfrak{G}$ by putting

$$\forall v \in G: \quad (\pi\tau)_a(\pi(v)) := \pi(\tau_a(v)). \tag{14}$$

Proposition 9 *1.* $(\pi\tau)_a$ *is a linear mapping.*

2. $((\pi\tau)_a)_{a\in\Sigma}$ *is a family of observable operators on* $\mathfrak{G}$.

Proof. 1. Multiplication with scalars: $(\pi\tau)_a(\alpha\pi(v)) = (\pi\tau)_a(\pi(\alpha v)) = \pi(\tau_a(\alpha v)) = \alpha\pi(\tau_a(v)) = \alpha(\pi\tau)_a$.

Additivity: $(\pi\tau)_a(\pi(v_1) + \pi(v_2)) = (\pi\tau)_a(\pi(v_1 + v_2)) = \pi(\tau_a(v_1 + v_2)) = \pi(\tau_a(v_1) + \tau_a(v_2)) = \pi(\tau_a(v_1)) + \pi(\tau_a(v_2)) = (\pi\tau)_a(\pi(v_1)) + (\pi\tau)_a(\pi(v_2))$.

2. We have to show: for all $\bar{c} \in \Sigma^*$, it holds that $(\pi\tau)_a \mathfrak{g}_{\bar{c}} = P[a \mid \bar{c}]\mathfrak{g}_{\bar{c}a}$. Let $\{\mathfrak{g}_{\bar{c}_1}, \ldots, \mathfrak{g}_{\bar{c}_m}\}$ be the same basis of $\mathfrak{G}$ as the one used in the definition of $(\pi\tau)_a$, and let $\mathfrak{g}_{\bar{c}} = \sum_{i=1,\ldots,m} \gamma_i \mathfrak{g}_{\bar{c}_i}$.

$$
\begin{aligned}
(\pi\tau)_a \mathfrak{g}_{\bar{c}} \quad &= \\
&= (\pi\tau)_a \sum_{i=1,\ldots,m} \gamma_i \mathfrak{g}_{\bar{c}_i} \\
&= \sum \gamma_i (\pi\tau)_a \mathfrak{g}_{\bar{c}_i} \qquad \text{(since } (\pi\tau)_a \text{ is linear)} \\
&= \sum \gamma_i (\pi\tau)_a \left(\pi \left(\frac{1}{P[\bar{c}_i]} \tau_{\bar{c}_i} w_0 \right) \right) \quad \text{(use prop. 7(1))} \\
&= \sum \gamma_i \pi \left(\tau_a \left(\frac{1}{P[\bar{c}_i]} \tau_{\bar{c}_i} w_0 \right) \right) \quad \text{(definition of } (\pi\tau)_a) \\
&= \sum \gamma_i \frac{1}{P[\bar{c}_i]} \pi(\tau_a \tau_{\bar{c}_i} w_0) \\
&= \sum \gamma_i \frac{1}{P[\bar{c}_i]} P[\bar{c}_i a] \mathfrak{g}_{\bar{c}_i a} \\
&= \sum \gamma_i P[a \mid \bar{c}_i] \mathfrak{g}_{\bar{c}_i a} \\
&= P[a \mid \bar{c}] \mathfrak{g}_{\bar{c}a} \quad \text{(apply prop. 6)}
\end{aligned}
$$

□

It is easy to see that there can exist at most one family of observable operators on $\mathfrak{G}$ which are linear. Since in this article we are exclusively concerned with processes (X_t) generated by some OOM, we will henceforth speak of *the* CCR $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\Sigma})$ of (X_t), and take $\mathfrak{t}_a$ to be the observable operator which is linear.

The subspace G introduced above is a somewhat arbitrary subspace of , , since it depends on the choice of $\{\mathfrak{g}_{\bar{c}_1}, \ldots, \mathfrak{g}_{\bar{c}_m}\}$. Therefore, it would be nice to use , instead of G in the definition of $(\pi\tau)_a$. The next proposition shows that this is in fact possible.

Proposition 10 *Given the terminology and assumptions from definition 12, it holds that*

$$\forall v_1, v_2 \in , \ : \quad \pi(v_1) = \pi(v_2) \rightarrow \pi(\tau_a(v_1)) = \pi(\tau_a(v_2))$$

Proof. Let $v_1 = \sum_{i=1,\ldots,l} \alpha_i \tau_{\bar{b}_i} w_0, v_2 = \sum_{i=1,\ldots,l} \beta_i \tau_{\bar{b}_i} w_0$ be linear combinations from basis vectors of , . Now conclude

$$\pi(v_1) = \pi(v_2)$$

$$
\begin{aligned}
&\Rightarrow \quad \pi(\sum_{i=1,\ldots,l} \alpha_i \tau_{\bar{b}_i} w_0) = \pi(\sum_{i=1,\ldots,l} \beta_i \tau_{\bar{b}_i} w_0) \\
&\Rightarrow \quad \sum \alpha_i P[\bar{b}_i] \mathfrak{g}_{\bar{b}_i} = \sum \beta_i P[\bar{b}_i] \mathfrak{g}_{\bar{b}_i} \\
&\Rightarrow \quad \mathfrak{t}_a \left(\sum \alpha_i P[\bar{b}_i] \mathfrak{g}_{\bar{b}_i} \right) = \mathfrak{t}_a \left(\sum \beta_i P[\bar{b}_i] \mathfrak{g}_{\bar{b}_i} \right) \\
&\Rightarrow \quad \sum \alpha_i P[\bar{b}_i] P[a \mid \bar{b}_i] \mathfrak{g}_{\bar{b}_i a} = \sum \beta_i P[\bar{b}_i] P[a \mid \bar{b}_i] \mathfrak{g}_{\bar{b}_i a} \quad \text{(since } \mathfrak{t}_a \text{ is linear)} \\
&\Rightarrow \quad \sum \alpha_i P[\bar{b}_i a] \mathfrak{g}_{\bar{b}_i a} = \sum \beta_i P[\bar{b}_i a] \mathfrak{g}_{\bar{b}_i a} \\
&\Rightarrow \quad \sum \alpha_i \pi(\tau_{\bar{b}_i a} w_0) = \sum \beta_i \pi(\tau_{\bar{b}_i a} w_0) \quad (\textit{prop. } 7(1)) \\
&\Rightarrow \quad \pi(\sum \alpha_i (\tau_{\bar{b}_i a} w_0)) = \pi(\sum \beta_i (\tau_{\bar{b}_i a} w_0)) \\
&\Rightarrow \quad \pi(\tau_a(\sum \alpha_i \tau_{\bar{b}_i} w_0)) = \pi(\tau_a(\sum \beta_i \tau_{\bar{b}_i} w_0)) \\
&\Rightarrow \quad \pi(\tau_a(v_1)) = \pi(\tau_a(v_2))
\end{aligned}
$$

□

The following proposition further clarifies the nature of the canonical mapping. In intuitive terms, it states that π maps two vectors v, v' on the same generalized word distribution iff both vectors, taken as starting distributions, generate the same "futures".

Proposition 11 *For $v, v' \in$, , the two following conditions are equivalent:*

1. $\pi v = \pi v'$
2. $\forall \bar{c} \in \Sigma^* : \ \sigma(\tau_{\bar{c}}(v)) = \sigma(\tau_{\bar{c}}(v'))$

Proof. 1. $\Rightarrow$ 2.: Let $v = \sum_{i=1,\ldots,l} \alpha_i \tau_{\bar{b}_i} w_0$, $v' = \sum_{i=1,\ldots,l} \beta_i \tau_{\bar{b}_i} w_0$ be combinations of v, v' from basis vectors of , . Conclude

$$
\begin{aligned}
&\sigma(\tau_{\bar{c}}(v)) \quad = \\
&\quad = \quad \sigma(\pi(\tau_{\bar{c}}(v))) \qquad (\pi \text{ preserves internal sum}) \\
&\quad = \quad \sigma(\pi(\tau_{\bar{c}} \circ (\sum \alpha_i \tau_{\bar{b}_i} w_0))) \\
&\quad = \quad \sigma(\pi(\sum \alpha_i \tau_{\bar{b}_i \bar{c}} w_0)) \quad = \quad \sigma(\sum \alpha_i \pi(\tau_{\bar{b}_i \bar{c}} w_0)) \\
&\quad = \quad \sigma(\sum \alpha_i P[\bar{b}_i \bar{c}] \mathfrak{g}_{\bar{b}_i \bar{c}}) \quad = \quad \sigma(\sum \alpha_i P[\bar{b}_i \bar{c}] \frac{\mathfrak{t}_{\bar{c}} \mathfrak{g}_{\bar{b}_i}}{P[\bar{c} \mid \bar{b}_i]}) \\
&\quad = \quad \sigma(\sum \alpha_i P[\bar{b}_i] \mathfrak{t}_{\bar{c}} \mathfrak{g}_{\bar{b}_i}) \quad = \quad \sigma(\mathfrak{t}_{\bar{c}}(\sum \alpha_i P[\bar{b}_i] \mathfrak{g}_{\bar{b}_i}))
\end{aligned}
$$

$$
\begin{aligned}
&= \sigma(\mathfrak{t}_{\bar{c}}(\sum \beta_i P[\bar{b}_i]\mathfrak{g}_{\bar{b}_i})) \qquad (\text{since } \pi(v) = \pi(v')) \\
&= \ldots \qquad (\text{same transformations backwards}) \\
&= \sigma(\tau_{\bar{c}}(v'))
\end{aligned}
$$

2. $\Rightarrow$ 1.: First we observe (using prop. 7 and eq. (14)) that 2. is obviously equivalent to

$$\forall \bar{c} \in \Sigma^* : \quad \sigma(\mathfrak{t}_{\bar{c}}(\pi(v))) \;=\; \sigma(\mathfrak{t}_{\bar{c}}(\pi(v')))$$

Therefore, in order to show "2. $\Rightarrow$ 1.", it suffices to show that for all $\mathfrak{d}, \mathfrak{d}' \in \mathfrak{G}$ it holds that

$$\forall \bar{c} \in \Sigma^* : \quad \sigma(\mathfrak{t}_{\bar{c}}\mathfrak{d}) = \sigma(\mathfrak{t}_{\bar{c}}\mathfrak{d}') \quad \rightarrow \quad \mathfrak{d} = \mathfrak{d}' \tag{15}$$

Let $\mathfrak{g}_{\bar{b}_i}, i = 1, \ldots, m$ be a basis of $\mathfrak{G}$, and $\mathfrak{d} = \sum \alpha_i \mathfrak{g}_{\bar{b}_i}, \mathfrak{d}' = \sum \beta_i \mathfrak{g}_{\bar{b}_i}$. We have to show that $\alpha_i = \beta_i$ for all i.

$$
\begin{aligned}
&\forall \bar{c} \in \Sigma^* : \quad \sigma(\mathfrak{t}_{\bar{c}}\mathfrak{d}) = \sigma(\mathfrak{t}_{\bar{c}}\mathfrak{d}') \\
&\Rightarrow \quad \forall \bar{c} \quad \sigma(\sum \alpha_i P[\bar{c} \mid \bar{b}_i]\mathfrak{g}_{\bar{b}_i\bar{c}}) = \sigma(\sum \beta_i P[\bar{c} \mid \bar{b}_i]\mathfrak{g}_{\bar{b}_i\bar{c}}) \\
&\Rightarrow \quad \forall \bar{c} \quad \sum \alpha_i P[\bar{c} \mid \bar{b}_i]\sigma(\mathfrak{g}_{\bar{b}_i\bar{c}}) = \sum \beta_i P[\bar{c} \mid \bar{b}_i]\sigma(\mathfrak{g}_{\bar{b}_i\bar{c}}) \\
&\Rightarrow \quad \forall \bar{c} \quad \sum \alpha_i P[\bar{c} \mid \bar{b}_i] = \sum \beta_i P[\bar{c} \mid \bar{b}_i] \quad (\text{since } \sigma(\mathfrak{g}_{\bar{b}_i\bar{c}}) = 1) \\
&\Rightarrow \quad \forall \bar{c} \quad \sum \alpha_i \mathfrak{g}_{\bar{b}_i}(\bar{c}) = \sum \beta_i \mathfrak{g}_{\bar{b}_i}(\bar{c}) \\
&\Rightarrow \quad \sum \alpha_i \mathfrak{g}_{\bar{b}_i} = \sum \beta_i \mathfrak{g}_{\bar{b}_i} \\
&\Rightarrow \quad \forall i : \quad \alpha_i = \beta_i
\end{aligned}
$$

□

Next we investigate a bit the CCR equivalent of the matrix/mapping μ in OOM's (cf. definition 3), which can be defined as the sum of the observable operators:

Definition 13 *For a CCR* $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\Sigma})$, *let* $\mathfrak{m} := \sum_{a\in\Sigma} \mathfrak{t}_a$ *be the* transition mapping *of the CCR.*

The following proposition collects propertes of $\mathfrak{m}$ which indicate how a CCR relates to an OOM (cf. 3.).

Proposition 12 *Given the notation and assumptions from definition 12, the following statements hold:*

1. $\forall v \in ,\quad \pi(\mu(v)) = \mathfrak{m}(\pi(v))$

2. $\forall \mathfrak{d} \in \mathfrak{G} \quad \sigma(\mathfrak{d}) = \sigma(\mathfrak{m}(\mathfrak{d}))$

3. $\mathfrak{g}_\varepsilon = \pi(w_0)$

4. $\mathfrak{g}_\varepsilon$ *is a fixed point of* $\mathfrak{m}$ *and has internal sum 1.*

Proof. 1. Follows from propositions 9(2), 10, and the definition of $\mathfrak{m}$.

2. is a consequence of (i) the fact that $\sigma(\mu(v)) = \sigma(v)$ for all $v \in \mathbb{R}^k$, (ii) surjectivity of π, and (iii) preservation of σ by π (cf. proposition 7).

3. Is a special case of proposition 7(1).

4.

$$\begin{aligned}
&\mathfrak{m}\mathfrak{g}_\varepsilon = \\
&= \sum_{a\in\Sigma} \mathfrak{t}_a \mathfrak{g}_\varepsilon \;=\; \sum_{a\in\Sigma} \pi\tau_a w_0 \qquad \text{(use (14))} \\
&= \pi \sum_{a\in\Sigma} \tau_a w_0 \;=\; \pi\mu w_0 \\
&= \pi w_0 \;=\; \mathfrak{g}_\varepsilon
\end{aligned}$$

□

Since the definition of $\mathfrak{G}$ depends entirely on the process (X_t) (and not on a particular OOM of this process), we are entitled to the following

Definition 14 *For a stationary process* $(X_t)_{t\in\mathbb{Z}}$ *with CCR* $(\mathfrak{G}, (\mathfrak{t}_a)_{a\in\Sigma})$*, we call* $dim(X_t) := dim(\mathfrak{G})$ *the* dimension *of the process.*

Since the canonical projection π of an OOM on its CCR is surjective, every OOM $(\mathbb{R}^k, (\tau_a)_{a\in\Sigma}, w_0)$ which generates (X_t) must at least have dimension $k \geq dim(X_t)$. We now turn to the question whether a minimal-dimension OOM $(\mathbb{R}^{dim(X_t)}, (\tau_a)_{a\in\Sigma}, w_0)$ always exists. This question is answered in the affirmative by the following proposition:

Proposition 13 *Using the assumptions and terminology of definition 12, let* $\mathfrak{g}_{\bar{b}_1}, \ldots, \mathfrak{g}_{\bar{b}_m}$ *be a basis of* $\mathfrak{G}$*. Let* $u_1, \ldots, u_m$ *be a basis of* $\mathbb{R}^m$*, where each* u_i *has internal sum 1 (for instance, let* u_i *be the i-th unit vector).*

Define a linear mapping

$$\begin{aligned}
\tilde{\pi}^{-1} \;&:\; \mathfrak{G} \to \mathbb{R}^m \\
&\quad\; \tilde{\pi}^{-1}(\mathfrak{g}_{\bar{b}_i}) = u_i,
\end{aligned}$$

Note that $\tilde{\pi}^{-1}$ is an isomorphism of vector spaces, which allows us to define, for every $a \in \Sigma$, a linear mapping $\tilde{\tau}_a$ on $\mathbb{R}^m$ by putting

$$\tilde{\tau}_a = \tilde{\pi}^{-1} \mathfrak{t}_a \tilde{\pi},$$

where $\tilde{\pi}$ denotes the inverse of $\tilde{\pi}^{-1}$.
Put $\tilde{w}_0 := \tilde{\pi}^{-1} \mathfrak{g}_\varepsilon$.
Then, $\mathcal{A} = (\mathbb{R}^m, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ is an OOM which generates the process (X_t).

Proof. First we show that $(\mathbb{R}^m, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ is an OOM, i.e. that it satisfies the conditions 1. – 3. from definition 3.

Condition 1: We have to show that $\tilde{\mu} := \sum_{a\in\Sigma} \tilde{\tau}_a$ has column sums 1 (as a matrix), i.e. that it is internal sum-preserving (as a mapping on $\mathbb{R}^m$).

This follows from the facts (i) that $\tilde{\mu}(\tilde{\pi}^{-1}(\mathfrak{d})) = \tilde{\pi}^{-1}(\mathfrak{m}(\mathfrak{d}))$, (ii) $\mathfrak{m}$ preserves internal sums in $\mathfrak{G}$, and (iii) that $\tilde{\pi}^{-1}$ preserves vector internal sums (which holds because $\tilde{\pi}^{-1}$ maps a set of basis vectors of internal sum 1 in $\mathfrak{G}$ on basis vectors in $\mathbb{R}^m$ that are also of internal sum 1).

Condition 2: We have to show that $\sigma(\tilde{w}_0) = 1$ and that $\tilde{\mu}\tilde{w}_0 = \tilde{w}_0$. The former follows from the obvious fact that $\tilde{\pi}^{-1}$ preserves internal sums, and that $\sigma(\mathfrak{g}_\varepsilon) = 1$. The latter can be seen as follows:

$$\begin{aligned}
\tilde{\mu}\tilde{w}_0 \quad &= \\
&= \sum_{a\in\Sigma} \tilde{\tau}_a \tilde{w}_0 \;=\; \sum \tilde{\pi}^{-1} \mathfrak{t}_a \tilde{\pi} \tilde{\pi}^{-1} \tilde{\pi}^{-1} \mathfrak{g}_\varepsilon \\
&= \sum \tilde{\pi}^{-1} \mathfrak{t}_a \mathfrak{g}_\varepsilon \;=\; \tilde{\pi}^{-1} \sum \mathfrak{t}_a \mathfrak{g}_\varepsilon \\
&= \tilde{\pi}^{-1} \mathfrak{g}_\varepsilon \;=\; \tilde{w}_0
\end{aligned}$$

Condition 3: We have to show that $\sigma(\tilde{\tau}_{\bar{c}} \tilde{w}_0) \in [0, 1]$ for all $\bar{c} \in \Sigma^*$. This follows from

$$\sigma(\tilde{\tau}_{\bar{c}} \tilde{w}_0) = \sigma(\tilde{\pi}^{-1} \mathfrak{t}_{\bar{c}} \mathfrak{g}_\varepsilon) = \sigma(\tilde{\pi}^{-1} P[\bar{c}] \mathfrak{g}_{\bar{c}}) = P[\bar{c}]$$

Second, we have to show that the OOM $(\mathbb{R}^m, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ generates the same process (X_t) as the original OOM $(\mathbb{R}^k, (\tau_a)_{a\in\Sigma}, w_0)$. This follows from the fact that π and $\tilde{\pi}^{-1}$ are internal sum preserving, and (9).

□

Note that the mapping $\tilde{\pi}$ used in this proof actually is the canonical projection of $\mathcal{B}$ onto the CCR.

Now we have everything in place for answering a big part of the question asked at the end of section 2, i.e. a classification of equivalent OOM's.

Proposition 14 *Let $dim(X_t) = m$. Let $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a\in\Sigma}, w_0)$ and $\mathcal{B} = (\mathbb{R}^m, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ be two equivalent OOM's, which generate (X_t). Then there exists a linear, surjective, internal sum-preserving mapping $\varrho : , (\mathcal{A}) \to \mathbb{R}^m$ such that $\varrho(w_0) = \tilde{w}_0$, and for all $a \in \Sigma$ it holds that*

$$\forall v \in , (\mathcal{A}) \quad \tilde{\tau}_a(\varrho(v)) = \varrho(\tau_a(v)). \tag{16}$$

We write $\varrho : \mathcal{A} \to \mathcal{B}$ for such ϱ.

Proof. Let π be the canonical projection of $\mathcal{A}$ on the CCR of (X_t), and let $\tilde{\pi}$ be the canonical projection of $\mathcal{B}$ on this CCR. Then, apply proposition 13 on $\tilde{\pi}^{-1}$ and conclude that $\varrho = \tilde{\pi}^{-1} \circ \pi : \mathcal{A} \to \mathcal{B}$. □

Proposition 15 *Let $dim(X_t) = m$, and $\varrho : \mathbb{R}^m \to \mathbb{R}^m$ be a bijective, linear, internal sum-preserving mapping. Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$ be an OOM. Define for every $a \in \Sigma$ a mapping $\tilde{\tau}_a : \mathbb{R}^m \to \mathbb{R}^m$ by putting*

$$\tilde{\tau}_a(\varrho(v)) := \varrho(\tau_a(v)). \tag{17}$$

Then $\mathcal{B} = (\mathbb{R}^m, (\tilde{\tau}_a)_{a\in\Sigma}, \varrho(w_0))$ is an OOM equivalent to $\mathcal{A}$.

Proof. Rewrite ϱ as $\varrho = \tilde{\pi}^{-1} \circ \pi$, where π is the canonical projection of $\mathcal{A}$, and $\tilde{\pi}^{-1}$ is a mapping as in proposition 13. Then apply proposition 13. □

A combination of propositions 14 and 15 yields the following characterization of equivalence of minimal-dimension OOMs:

Proposition 16 *Two minimal-dimension OOMs $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$, $\mathcal{B} = (\mathbb{R}^m, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ are equivalent iff there exists a internal sum-preserving isomorphism $\varrho : \mathcal{A} \to \mathcal{B}$ of vector spaces which maps w_0 on $\tilde{w}_0$, and which transports $(\tau_a)_{a\in\Sigma}$ into $(\tilde{\tau}_a)_{a\in\Sigma}$ in the sense of (17).* □

5 From OOMs back to HMMs (halfways)

In this section we investigate in more detail how HMM's relate to OOM's.

We start with some preparatory remarks.

HMM's can be seen as OOM's with two special properties:

1. There exist diagonal matrices o_a such that $\tau_a = \mu o_a$.

2. All entries in the matrices o_a and μ are probability values, i.e. are from $[0, 1]$.

Only the first condition fits nicely with the linear algebra perspective taken in this article. Diagonal matrices are a much-sought commodity when one investigates linear mappings. By contrast, the second condition introduces interval restrictions which are alien to methods of linear algebra.

In this section, we will learn that HMMs are truly special among OOMs only with respect to the second condition. While for every OOM there exists an equivalent one featuring diagonal matrices o_a, there exist OOMs for which no equivalent one exists whose matrix entries are all probabilities. An immediate consequence of the second finding is that HMMs are a proper subclass of OOMs.

We will first investigate a bit the topic of OOMs which are in "diagonal form":

Definition 15 *An OOM is called* diagonal-form *if there exist diagonal matrices o_a such that $\tau_a = \mu o_a$ (where $\mu = \sum \tau_a$, as usual). An OOM is simply called a HMM if it also has property 2 from above.*

As we have seen in the preceding section, OOM theory is particularly convenient for minimal-dimension OOMs. In the following proposition, we develop the picture of how equivalent, minimal-dimension, diagonal-form OOMs are mutually related. These results may be of some interest for applications of HMMs, since often one will have minimal-dimension HMMs. Then, this proposition affords some insight in the existence or non-existence of equivalent, but different, HMMs.

Proposition 17 Part 1. *Let $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0), \mathcal{B} = (\mathbb{R}^m, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ be two equivalent, minimal-dimension, diagonal-form OOMs with diagonal matrices $(o_a)_{a\in\Sigma}, (\tilde{o}_a)_{a\in\Sigma}$. Let the transition matrix $\mu = \sum_{a\in\Sigma} \tau_a$ be regular. Let $\varrho : \mathcal{A} \to \mathcal{B}$ be the isomorphism according to proposition 16. Then, modolo permutations π_1, π_2 of coordinates in $\mathcal{A}$ and $\mathcal{B}$ (i.e. change from from ϱ to $\pi_2 \varrho \pi_1^{-1}$, from o_a to $\pi_1 o_a \pi_1^{-1}$, and from $\tilde{o}_a$ to $\pi_2 \tilde{o}_a \pi_2^{-1}$ for some permutations π_1, π_2 applied to the coordinates of $\mathcal{A}$ and $\mathcal{B}$, respectively), ϱ, $(o_a)_{a\in\Sigma}$, and $(\tilde{o}_a)_{a\in\Sigma}$ obey the following restrictions:*

1. *The matrix ϱ consists of quadratic submatrices R_ν aligned on the diagonal, where $\nu = 1, \ldots, m'$ with $m' \leq m$:*

$$\varrho = \begin{pmatrix} R_1 & & 0 \\ & \ddots & \\ 0 & & R_{m'} \end{pmatrix} \tag{18}$$

2. *Let for any* $a \in \Sigma$

$$o_a = \begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_m \end{pmatrix} \qquad \tilde{o}_a = \begin{pmatrix} \tilde{a}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{a}_m \end{pmatrix}$$

Let R_ν *be a matrix of size* $n_\nu \times n_\nu$*, and let* $i_\nu := n_1 + \cdots + n_\nu$*, i.e.* i_ν *is the coordinate index of the last column/row of* R_ν *in* ϱ*. Put* $i_0 := 0$*. Then, it holds that for* $i_{\nu-1} < i, j \leq i_\nu$ *(where* $\nu = 1, \ldots, m'$*)*

$$a_i = a_j = \tilde{a}_i = \tilde{a}_j.$$

Said in less formal terms, the second condition states that in each matrix $o_a, \tilde{o}_a$ *the elements that lie in the same region as the region covered by some* B_ν *in* ϱ*, are equal.*

Part 2 (inverse of part 1). *Let* $\mathcal{A}$ *be a diagonal-form, minimal-dimension OOM, and* $(o_a)_{a\in\Sigma}$ *its diagonal matrices, let* $\varrho : \mathcal{A} \to \mathcal{B}$*, and let* $\tilde{o}_a := \varrho o_a \varrho^{-1}$*.*

If a permutation of coordinates of $\mathcal{A}$ *and* $\mathcal{B}$ *exists (analogous to the one in part 1) such that*

1. ϱ *is of the form (18),*
2. *for each* $a \in \Sigma$*, diagonal entries in* o_a *which lie in the region of the same* R_ν *are equal to each other,*

then, $\mathcal{B}$ *is also diagonal-form, and it holds that* $\tilde{o}_a = o_a$ *for all* $a \in \Sigma$*.*

Proof. Part 1. Let $\varrho : \mathcal{A} \to \mathcal{B}$ be a transformation of $\mathcal{A}$ into $\mathcal{B}$ according to proposition 14. Since ϱ is regular, the matrix (ϱ_{ij}) in every row and in every column has at least one nonzero element. Let

$$o_a = \begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_m \end{pmatrix} \qquad \tilde{o}_a = \begin{pmatrix} \tilde{a}_1 & & 0 \\ & \ddots & \\ 0 & & \tilde{a}_m \end{pmatrix}$$

Using that $\tilde{\tau}_a = \varrho\tau_a\varrho^{-1}$, $\tilde{\mu} = \varrho\mu\varrho^{-1}$, $\tilde{\mu}\tilde{o}_a = \tilde{\tau}_a$, $\mu o_a = \tau_a$, and that μ is regular, conclude that $\tilde{o}_a \circ \varrho = \varrho \circ o_a$. A multiplication of matrices $\tilde{o}_a\varrho = \varrho o_a$ yields equations $\tilde{a}_i\varrho_{ij} = \varrho_{ij}a_j$ for $i, j = 1, \ldots, m$. Since for every i (and for every j, respectively) some $\varrho_{ij} \neq 0$, this implies that every $\tilde{a}_i$ is equal to at least one a_j. This in turn implies that modolo a permutation of coordinates (as described in the proposition), it holds that $o_a = \tilde{o}_a$.

Since an identical inference can be made for all $a \in \Sigma$, this means that (modolo the same permutations π_1, π_2 for all a) $(o_a)_{a\in\Sigma} = (\tilde{o}_a)_{a\in\Sigma}$. We will in the remainder of this proof assume that such a permutation has been carried out.

Now we take one of the o_a and see what we can infer from it about the structure of ϱ. We may assume without loss of generality (modolo a further permutation of coordinates, applied uniformly in $\mathcal{A}$ and $\mathcal{B}$) that if some of the entries a_i on the diagonal of o_a are equal, they follow each other directly on the diagonal.

Observing $o_a = \tilde{o}_a$ and $\varrho \circ o_a = \tilde{o}_a \circ \varrho$, two multiplication of matrices ϱo_a and $o_a \varrho$, and a subsequent comparison of matrix entries, yields us equations

$$a_j \varrho_{ij} = a_i \varrho_{ij} \quad \text{for } 1 \leq i, j \leq m',$$

from which it follows that for $a_i \neq a_j$, it holds that $\varrho_{ij} = 0$. Since the a_i have been ordered such that equal ones succeed each other on the diagonal, this implies that ϱ consists of quadradic submatrices R_{ν_a} on the diagonal and is zero otherwise, where each R_{ν_a} covers the part of the diagonal in which consecutive, equal a_i lie.

A similar argument can be made for the remaining $o_{a'}$. It is easy to see how this leads to the statements made in part 1 of the proposition.

Part 2: Easy exercise. $\Box$.

A special case of part 1 of this proposition, which will occur often in applications, deserves to be mentioned explicitly:

Proposition 18 *Given the assumptions of proposition 17, part 1, if in one of the o_a all diagonal entries a_i are mutually unequal, then ϱ is uniquely determined. In other words, the miminal-dimension, diagonal-form OOM is then unique (modolo permuation of coordinates).* $\Box$

We consider a little example which illustrates proposition 17.

Let $\Sigma = \{a, b\}$, and consider the 3-dimensional HMM $\mathcal{A}$ which is given by

$$\mu = \begin{pmatrix} 1/2 & 0 & 1 \\ 1/2 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

and

$$o_a = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \qquad o_b = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \tag{19}$$

$\mathcal{A}$ is minimal-dimensional (this can be seen by showing that $\mathfrak{g}_a, \mathfrak{g}_b, \mathfrak{g}_{bb}$ are linearly independent – exercise).

Now proposition 17 tells us that modolo coordinate permutations we get all equivalent minimal-dimension, diagonal-form OOM's $\mathcal{B}$ by mappings $\varrho : \mathcal{A} \to \mathcal{B}$, where ϱ is of the form

$$\varrho = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-r & s \\ 0 & r & 1-s \end{pmatrix} \tag{20}$$

with $r, s \in \mathbb{R}, 1 - r \neq s$.

Which of these diagonal-form OOMs is a HMM, i.e. for which ϱ does the matrix $\varrho\mu\varrho^{-1}$ contain only entries from $[0, 1]$? The answer is that this is ony the case for $\varrho = id$, which means that in this example the minimal-dimension HMM is unique (modolo permutation of coordinates). Proving this is straigthforward but tedious (requires the algebraic computation of ϱ^{-1}, the algebraic multiplication $\varrho\mu\varrho^{-1}$, and the subsequent exploitation of the constraint that $\varrho\mu\varrho^{-1}$ be a stochastic matrix). We skip this exercise here. The intuitive reason why we get $\varrho = id$ in this example is that μ contains many 0's and 1's, and even slight deviations of ϱ from the identity matrix (i.e. even small values of r and s) either push some 0's below zero, or some 1's above unity.

The uniqueness of the minimal-dimension HMM in this example is somewhat accidental, since it results from the 0's and 1's in μ. I put in these 0's and 1's to facilitate the (manual) computation of $\varrho\mu\varrho^{-1}$. In empirically derived HMMs, one will rarely find transition matrices with 0's and 1's. If we adapt our example a bit and consider the matrix

$$\mu' = \begin{pmatrix} 1/2 - \varepsilon & \varepsilon & 1 - 2\varepsilon \\ 1/2 - \varepsilon & \varepsilon & \varepsilon \\ 2\varepsilon & 1 - 2\varepsilon & \varepsilon \end{pmatrix} \tag{21}$$

instead of our original μ (where ε is a small positive number), and keep o_a and o_b, then a continuity argument teaches us that indeed there exist $\varrho \neq id$ of the form (20) such that $\varrho\mu\varrho^{-1}$ is a stochastic matrix. I.e., the minimal-dimension HMM specified by (21) and (19) is not unique.

A complete characterization of all minimal-dimension HMMs equivalent to (21) would essentially consist of $m^2 = 9$ two-sided inequalities of the form

$$0 \leq (\varrho\mu\varrho^{-1})_{ij} \leq 1,$$

where the $(\varrho\mu\varrho^{-1})_{ij}$ are the matrix entries of $\varrho\mu\varrho^{-1}$. Obviously this exercise would be neither entertaining nor very enlightening.

Now we turn to the question of whether for every OOM there exists some (possibly higher-dimensional) equivalent, diagonal-form OOM. In more casual terms, can every OOM be "diagonalized"?

Proposition 19 *For every OOM $\mathcal{A}$ there exists an equivalent diagonal-form OOM $\mathcal{B}$.*

The proof requires some work. Before we explain the basic idea, we give a technical lemma which we'll need.

Proposition 20 *Let $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$ such that $\sigma(z) = m$. Then there exists m vectors $v^1, \ldots, v^m \in \mathbb{R}^m$, where $v^i = (v^i_1, \ldots, v^i_m)$, such that*

1. *The vectors v^i are linearly independent.*
2. *$\sigma(v^i) = 1$ for all v^i.*
3. *$z = v^1 + \cdots + v^m$.*

Proof of proposition 20. There are many ways to construct vectors v^i satisfying the conditions 1–3. First note that we may assume $m \geq 2$, since for $m = 1$ we can simply take $v^1 = z$. Now define x^i as the m-dimensional vector which is zero everywhere, excepted at the i-th component, which is 1, and the $i+1$-th (modolo m) component, which is -1. Then, put $v^i := m^{-1}z + x^i$. It is easily verified that these v^i satisfy the requirements stated in the proposition. □

Proof of proposition 19. Outline: We may assume that $\mathcal{A} = (\mathbb{R}^m, (\tau_a)_{a\in\Sigma}, w_0)$ is minimal-dimensional. The linear mappings from $\mathbb{R}^m$ to $\mathbb{R}^m$ can be interpreted as a vector space whose dimension is m^2. We embed $\mathcal{A}$ in an m^2-dimensional OOM $\mathcal{B} = (\mathbb{R}^{m^2}, (\tilde{\tau}_a)_{a\in\Sigma}, \tilde{w}_0)$ such that, essentially, $\mathcal{A} = , (\mathcal{B})$. We arrange things in a way such that there exist m^2 linear mappings $\tilde{e}^{ij} : \mathbb{R}^{m^2} \to \mathbb{R}^{m^2}$, which (i) leave , $(\mathcal{B})$ invariant, (ii) are linearly independent on , $(\mathcal{B})$, and (iii) can be written as $\tilde{e}^{ij} = \tilde{\mu} d^{ij}$, where d^{ij} is a diagonal $m^2 \times m^2$ matrix. Properties (i) and (ii) ensure that these $\tilde{e}^{ij}$ can be interpreted as a basis of the vector space of all linear mappings from , $(\mathcal{B})$ to , $(\mathcal{B})$. In particular, every $\tilde{\tau}_a :$, $(\mathcal{B}) \to$, $(\mathcal{B})$ can be linearly combined from them. Property (iii) then says that $\mathcal{B}$ is diagonal-form.

As a preliminary to working this sketch out in detail, we introduce a shorthand notation for indices ranging from 1 to m^2 in "blocks of length m" by defining

$$[ij] := (i-1)m + j,$$

where $1 \leq i, j \leq m$.

Now we embed $\mathcal{A}$ in $\mathcal{B}$. First we define an m-dimensional subspace U of $\mathbb{R}^{m^2}$, which we will use as the image of $\mathcal{A}$ in $\mathcal{B}$. Let $e_i^{m^2}$ be the m^2-dimensional vector which is $1/m$ at dimensions $[i1], \ldots, [im]$, and is 0 otherwise. The vectors $(e_i^{m^2})_{i=1,\ldots,m}$ span an m-dimensional subspace U of $\mathbb{R}^{m^2}$.

Now we linearly map U isomorphically on $\mathbb{R}^m$ by

$$\varrho : U \to \mathbb{R}^m, \quad e_i^{m^2} \mapsto e_i^m,$$

where e_i^m is the i-th unit vector of $\mathbb{R}^m$. Note that ϱ is internal sum-preserving, as is the inverse mapping

$$\varrho^{-1} : \mathbb{R}^m \to U, \quad e_i^m \mapsto e_i^{m^2}.$$

Next we define a $m^2 \times m^2$-matrix $\tilde{\mu}$, which will be the transition matrix of $\mathcal{B}$. The idea is to essentially "blow up" μ by a factor of m, such that each entry μ_{ij} of μ corresponds to a $m \times m$-submatrix of $\tilde{\mu}$. Let $v_{[ij]}$ denote the $[ij]$-th column in $\tilde{\mu}$. Use proposition 20 to verify that column vectors $v_{[ij]}$ can be constructed such that the following specifications are met:

1. Each $v_{[ij]}$ is in U.
2. Each $v_{[ij]}$ has internal sum 1.
3. For $i = 1, \ldots, m$, the vectors $v_{[i1]}, \ldots, v_{[im]}$ are linearly independent.
4. For all $i = 1, \ldots, m$, it holds that the summed column vectors in the i-th "block" of length m are the image of the i-th column vector v_i of μ, i.e. $\sum_{j=1,\ldots,m} v_{[ij]} = \varrho^{-1} v_i$.

Obviously the mapping $\tilde{\mu} : \mathbb{R}^{m^2} \to \mathbb{R}^{m^2}$ leaves U invariant, i.e. $im(\tilde{\mu}) \subseteq U$. Furthermore, $\tilde{\mu}$ can be seen as a version of μ in the sense that $\tilde{\mu}\varrho^{-1}v = \varrho^{-1}\mu v$ for all $v \in \mathbb{R}^m$. In particular, $\varrho^{-1}w_0 =: \tilde{w}_0$ is the (unique) fixed point of internal sum 1 of $\tilde{\mu}$.

Let $e^{ij} : \mathbb{R}^m \to \mathbb{R}^m$ be defined by the $m \times m$-matrix which is zero everywhere with the exception of the entry at column i and row j, where it is 1. I.e., e^{ij} maps the unit vector e_i^m on the unit vector e_j^m. The family of mappings $(e^{ij})_{i,j=1,\ldots,m}$ is a basis of the vector space of all linear mappings on $\mathbb{R}^m$.

We now construct a corresponding basis for linear mappings on U. Let $\tilde{e}^{ij}$ denote the $m^2 \times m^2$-matrix whose column vectors are all zero with the exception of columns in positions $[i1], \ldots, [im]$, where the column vectors are possibly nonzero multiples of the column vectors of $\tilde{\mu}$, namely, $\alpha_1^{ij} v_{[i1]}, \ldots, \alpha_m^{ij} v_{[im]}$.

Recall that the vectors $v_{[i1]}, \ldots, v_{[im]}$ are in U and linearly independent, i.e. form a basis of U. Using this fact, it is easy to show that $\alpha_1, \ldots, \alpha_m$ can be selected in a way such that $\varrho(\sum_{j=1,\ldots,m} v_{[ij]}) = e_i^m$. I.e., $\tilde{e}^{ij}$ can be considered a "blown-up" version of e^{ij}.

Furthermore, $\tilde{e}^{ij}$ can be written as the product $\tilde{e}^{ij} = \tilde{\mu} d^{ij}$ of $\tilde{\mu}$ with a diagonal matrix d^{ij}, namely, the diagonal matrix which has 0's everywhere on its diagonal with the exception of positions $[i1], \ldots, [im]$, where the diagonal reads $\alpha_1^{ij}, \ldots, \alpha_m^{ij}$.

Now let $\tau_a = \sum_{i,j=1,\ldots,m} \beta^{ij} e^{ij}$ be the linear combination of τ_a from basis functions. Define $\tilde{\tau}_a := \sum_{i,j=1,\ldots,m} \beta^{ij} \tilde{e}^{ij}$. Then, it is easy to show that for all $v \in \mathbb{R}^m$, it holds that $\tilde{\tau}_a \varrho^{-1} v = \varrho^{-1} \tau_a v$. Furthermore, $\tilde{\tau}_a$ is the product of $\tilde{\mu}$ with a diagonal matrix $\tilde{o}_a$, namely,

$$\begin{aligned}
\tilde{\tau}_a &= \sum_{i,j=1,\ldots,m} \beta^{ij} \tilde{e}^{ij} \\
&= \sum \beta^{ij} \tilde{\mu} d^{ij} \\
&= \tilde{\mu} \sum \beta^{ij} d^{ij} \\
&=: \tilde{\mu} \tilde{o}_a
\end{aligned}$$

Therefore, $\mathcal{B}$ is diagonal-form. It is also equivalent to $\mathcal{A}$, as the following transformations reveal:

$$\begin{aligned}
&\sigma(\tilde{\tau}_{a_n} \cdots \tilde{\tau}_{a_1} \tilde{w}_0) \quad = \\
&\quad = \quad \sigma(\tilde{\tau}_{a_n} \cdots \tilde{\tau}_{a_1} \varrho^{-1} w_0) \quad = \quad \sigma(\tilde{\tau}_{a_n} \cdots \tilde{\tau}_{a_2} \varrho^{-1} \tau_{a_1} w_0) \\
&\quad = \quad \ldots \quad = \quad \sigma(\varrho^{-1} \tau_{a_n} \ldots \tau_{a_1} w_0) \quad = \quad \sigma(\tau_{a_n} \ldots \tau_{a_1} w_0)
\end{aligned}$$

□

Recall that HMMs are OOMs which are special in that they (i) are diagonal- form and (ii) have entries from $[0, 1]$ in their μ and o_a. The previous proposition says that property (i) cannot in fact be used to discern HMMs from OOMs which have no equivalent HMMs. In the remainder of this section, we will show that it is property (ii) which distinguishes HMMs as a proper subclass of OOMs.

We proceed in two steps. First, we describe a property which is inherited from any HMM two every of its equivalent, minimal-dimension OOMs. Second, we present an example of a miminal-dimension OOM which lacks this property. But before we deal with either point, we prove the following auxiliary proposition from linear algebra:

Proposition 21 *Let $U \subseteq \mathbb{R}^k$ be a linear subspace of $\mathbb{R}^k$. Let $N \subset U$ be the set of vectors in U which have only non-negative entries and are not the 0 vector. Then there exist finitely many vectors $v^1, \ldots, v^n \in N$ such that every $u \in N$ can be written as a linear combination $u = \sum_{i=1,\ldots,n} \alpha_i v^i$, where all α_i are non-negative.*

Proof. Let $V := \{v = (v_1, \ldots, v_k) \in N \mid \forall v' = (v'_1, \ldots, v'_k) \in N\ (\forall i = 1, \ldots, m :\ v_i = 0 \rightarrow v'_i = 0) \rightarrow (\forall i :\ v_i \neq 0 \rightarrow v'_i \neq 0)\}$ be the subset of N of vectors with 0's in maximally many places.

We call two vectors from V "related" iff they have 0's in the same places. It holds that if v, w are related, then one is a positive multiple of the other, i.e. $\exists \alpha > 0 : v = \alpha w$. The reason is because if this would not hold, then w and v would be linearly independent, and it is easy to see that then some positive β would exist such that $v - \beta w$ would lie in N and have at least one 0 more than v, which contradicts $v \in V$.

Relatedness obviously is an equivalence relation on V. We take from every equivalence class one representative and get a collection $\{v^1, \ldots, v^n\}$. We will now show that this collection satisfies the requirements of the proposition.

Let $u \in N, u \neq 0$. u has 0's at x places, where $x \geq 0$. We have to show that u can be written as a non-negative linear combination from vectors from $\{v^1, \ldots, v^n\}$.

Case 1: $u \in V$. Then $u = \alpha v^i$ for some positive α and the representative v^i related to u, and we are done.

Case 2: $u \notin V$. Then, some $v \in N$ exists which has 0's at all places where u has 0's, plus at least one 0 at a place where u has a positive component. v and u are linearly independent. Therefore, some positive α exists such that $u - \alpha v$ is in N and has at least one 0 more than u. We put $u_1 := u - \alpha v, u_2 := \alpha v$. Then, $u = u_1 + u_2$ is a decomposition of u into vectors of N which each have properly more 0's than u. This argument can be iterated on u_1 and u_2, and on the resulting decompositions, etc., until one has reached a decomposition of u into vectors from V. Then use case 1 to conclude that u is a positive linear combination of vectors from $\{v^1, \ldots, v^n\}$. □

Now we are equipped for describing a property which is inherited from a HMM to its minimal-dimension, equivalent OOMs.

Proposition 22 *Let $\mathcal{A} = (\mathbb{R}^k, (\tau_a)_{a \in \Sigma}, w_0)$ be a HMM, and let $\mathcal{B} = (\mathbb{R}^m, (\tilde{\tau}_a)_{a \in \Sigma}, \tilde{w}_0)$ be an equivalent, minimal-dimension OOM. Then there exists a finite set of vectors $\{v^1, \ldots, v^n\} \subset \mathbb{R}^m$ of positive internal sum such that for all $\bar{c} \in \Sigma^*$, for all $i = 1, \ldots, n$, $\tilde{\tau}_{\bar{c}} v^i$ can be written as a linear combination of the vectors $\{v^1, \ldots, v^n\}$ with non-negative coefficients, i.e.*

$$\exists v^1, \ldots, v^n \; \forall i = 1, \ldots, n : \quad (\sigma(v^i) > 0) \quad \wedge$$
$$(\forall \bar{c} \in \Sigma \; \exists \alpha_1, \ldots, \alpha_n \geq 0 : \; \tilde{\tau}_{\bar{c}} v^i = \sum_{j=1,\ldots,n} \alpha_j v^j) \tag{22}$$

Proof. We first show that (22) holds for the special case where $\bar{c} = a, (a \in \Sigma)$.

The transition matrix μ and the matrices o_a of $\mathcal{A}$ have non- negative entries since $\mathcal{A}$ is a HMM. Therefore, the matrices $\tau_a = \mu o_a$ have only non-negative entries, too. Let $N \subset$, $(\mathcal{A})$ be the vectors in , $(\mathcal{A})$ which have only non-negative entries. Then, for every $u \in N$, $a \in \Sigma$ it holds that $\tau_a u \in N$, since τ_a only has non-negative entries.

According to the auxiliary proposition, a subset $\{w^1, \ldots, w^n\} \subset N$ of vectors with positive internal sum exists such that every element from N can be written as a non-negative linear combination from this subset. From this it follows that for $i = 1, \ldots, n$ and $a \in$, , it holds that $\tau_a w^i = \sum_{i=1,\ldots,n} \alpha_i w^i$, where all α_i are non-negative.

Now let $\varrho : \mathcal{A} \to \mathcal{B}$ be a internal sum-preserving, linear mapping according to proposition 14. Put $v^i := \varrho(w^i)$. Apply property (16) to conclude (22).

Now we have to show that (22) holds for all $\bar{c} \in \Sigma^*$. This is an easy exercise (induction over length of $\bar{c}$). $\square$

In order to complete our argument that (loosely speaking) HMMs are a proper subclass of OOMs, we now present an example of a minimal-dimension OOM $\mathcal{A} = (\mathbb{R}^3, \{\tau_a, \tau_b, \tau_c\}, w_0)$ which does not satisfy (22).

The basic idea is to use for τ_a a rotation mapping, which rotates $\mathbb{R}^3$ by a non-rational multiple of 2π, thus ensuring that iterations of τ_a never run into a period. As we will see, such a rotation mapping cannot be obtained in a HMM. The operators τ_b and τ_c are less remarkable; their role is mainly to ensure that the dimension of the process generated by $\mathcal{A}$ is indeed 3.

Before I further explain the "logics" of this example, it will be helpful to get a clear picture of τ_a (cf. fig. 2). Consider a rotation ϱ of $\mathbb{R}^3$ around an axis A which is given by the vector $(1/3, 1/3, 1/3)$. This rotation leaves invariant the hyperplane H of vectors of internal sum 1, and within H, it leaves invariant the circle C which is fixed by the unit vectors. Every radius r is turned counterclockwise by the angle of rotation, φ. Now, define $\tau_a := 1/2\varrho$. I.e., τ_a rotates $\mathbb{R}^3$ like ϱ, but additionally contracts vectors in their internal sum by 1/2. Thus, for instance, e_2 is mapped on $\tau_a e_2$, as indicated in fig. 2. The algebraic representation of τ_a is rather unhandy, therefore I

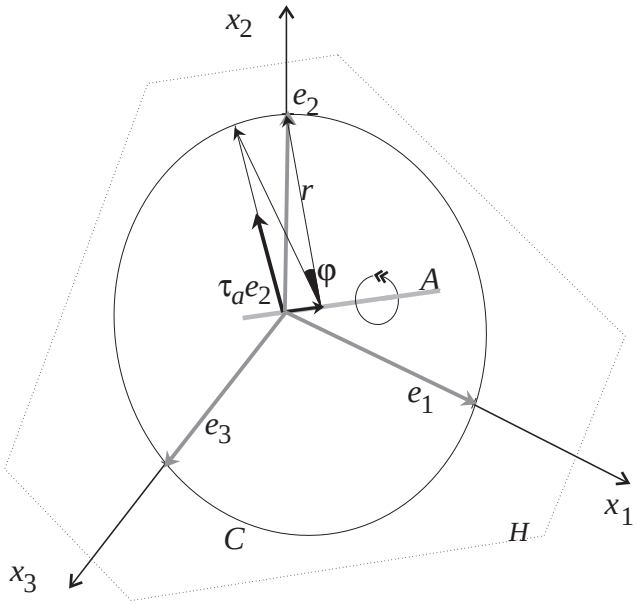

Figure 2: How τ_a is defined. For explanation see text.

give only an approximate numerical version for the case $\varphi = .1$ (measured in rad):

$$\tau_a = \begin{pmatrix} .498 & -.139 & .141 \\ .141 & .498 & -.139 \\ -.139 & .141 & .498 \end{pmatrix} \tag{23}$$

For τ_b and τ_c, we use operators which map every vector on varying fractions of e_2:

$$\tau_b = \begin{pmatrix} 0 & 0 & 0 \\ 1/8 & 2/8 & 3/8 \\ 0 & 0 & 0 \end{pmatrix} \quad \tau_c = \begin{pmatrix} 0 & 0 & 0 \\ 3/8 & 2/8 & 1/8 \\ 0 & 0 & 0 \end{pmatrix}$$

For the sake of completeness, I note that this selection of observable operators leads to $w_0 \approx (-.161, .872, .289)$.

The OOM defined in this way is indeed of minimal dimension. This can be shown e.g. by computing the three vectors $(P[ab], P[b], P[c]), (P[ab \mid a], P[b \mid a], P[c \mid a]), (P[ab \mid aa], P[b \mid aa], P[c \mid aa])$. One will find that they are linearly independent, from which it follows that $\mathfrak{g}_\varepsilon, \mathfrak{g}_a, \mathfrak{g}_{aa}$ are linearly independent.

Now assume that some HMM $\mathcal{B}$ exists which is equivalent to $\mathcal{A}$. Then, proposition 22 says that a set of vectors $V = \{v^1, \ldots, v^n\}$ of positive internal

sum exists, such that for all $\bar{d} \in \{a, b, c\}^*$, for all $i = 1, \ldots, n$, $\tilde{\tau}_{\bar{d}} v^i$ can be written as a linear combination of the vectors $\{v^1, \ldots, v^n\}$ with non-negative coefficients.

We can assume without loss of generality that the vectors $\{v^1, \ldots, v^n\}$ all have internal sum 1, i.e., they all lie in H. Let P be the smallest convex subset of H which contains all vectors from V. P is a polygon. Furthermore, P consists exactly of those vectors from H which can be linearly combined from V with non-negative coefficients.

We observe that V must contain at least one vector which is different from the vector through the rotation axis, i.e., $V \neq \{(1/3, 1/3, 1/3)\}$, since $\tau_b(1/3, 1/3, 1/3) = (0, 1/4, 0)$, and $(0, 1/4, 0)$ is no linear combination of vectors from $\{(1/3, 1/3, 1/3)\}$.

Therefore, V must contain vectors which have a positive angle with the axis A of rotation. Let $w \in V$ be a vector whose angle with A is maximal in V. Now, consider the trajectory $T = ((2\tau_a)^i w)_{i \in \mathbb{N}}$. It lies on a circle $C(w)$ in H, which is concentric with C. The trajectory even lies densely on $C(w)$, since φ was chosen to be a non-rational multiple of 2π.

All vectors from V lie on or within $C(w)$, since w was taken to have a maximal angle with A. Therefore, no point of P lies outside $C(w)$.

Now, since P is a polygon with finitely many vertices, some (indeed almost all) points of T lie outside P. This implies that some j exists such that $(2\tau_a)^j w$ cannot be linearly combined from V with non-negative coefficients, which also implies that $\tau_a^j w$ cannot be combined from V with non-negative coefficients, which contradicts proposition 22. Therefore, no HMM $\mathcal{B}$ equivalent to $\mathcal{A}$ exists.

It took me more time and effort to find this example of an OOM which is not a HMM, than it took me to develop all the rest of the material contained in this article. Still, this example is unsatisfactory since it seems to exploit a rather particular effect. It would be desirable to gain more general insights into the difference between OOMs and HMMs. In particular, it would be nice if we could effectively tell, given a particular OOM $\mathcal{A}$, whether an equivalent HMM exists.

Let me conclude this section with the remark that from a linear algebra perspective, HMMs are not a "natural" subclass of OOMs. The two-sided numerical inequalities of the kind "$0 \leq$ matrix entry ≤ 1", which single out HMMs among OOMs, do not go well with a linear algebra framework. We had occasion to note (in example (21)) how unpleasant it is to handle such inequalities. Maybe this helps to explain why "nice" mathematical results about HMMs are scarce.

6 Discussion

In this article, the foundations have been laid for describing certain discrete-valued, discrete-time, stationary stochastic processes (among them hidden Markov processes), in a transparent fashion which is characterized by two points:

1. The observed events $a \in \Sigma$ are interpreted as linear operators τ_a (or $\mathfrak{t}_a$, resp.) which change the state v of a dynamical system.

2. The probability $P[a \mid v]$ that in some state v an operator τ_a (or $\mathfrak{t}_a$) is observed is equal to the contraction in internal sum of v, i.e. $P[a \mid v] = \frac{\sigma(\tau_a(v))}{\sigma(v)}$ (or $\frac{\sigma(\mathfrak{t}_a(v))}{\sigma(v)}$).

These two points characterize both the various OOM's and the unique CCR of a process (X_t).

The OOM's and the CCR of a process are tightly interrelated. It appears that the former are more suited for practical applications, given the simple and explicit nature of their vectors, whereas the rather abstract CCR reveals more clearly the fundamental nature of observable operators and may thus be more helpful in theoretical investigations. In my personal opinion, the definition of CCRs is the most important contribution of this article. I feel that this concept opens interesting prospects for a deeper understanding of stationary stochastic processes beyond the linear case. By interpreting observables a_i as operators $\mathfrak{t}_{a_i}$, CCRs may allow to exploit techniques even from nonlinear system theory for stochastic processes research.

From a linear algebra perspective, HMM's are a somewhat less natural class of models of stochastic processes than OOM's are. However, it remains to be seen e.g. whether OOM-based system identification algorithms can be found which are in any respect better than the derivates of the Baum-Welch algorithm customarily used in the HMM field (cf.[4]). Furthermore, there may exist other mathematical perspectives (besides the one of linear algebra), which might make HMMs look more natural. Also, the epistemological connotations of what a (hidden) system state is are quite different between HMMs (as they are currently used) and OOMs. The question of how to properly handle the concept of hidden system states requires a careful consideration which is beyond the scope of this article. I think that all that can be said at the present time about OOMs vs. HMMs is that the former are an alternative to HMMs which is worth to be further investigated.

Acknowledgments I feel very grateful toward Thomas Christaller for encouragement and intellectual freedom. Special thanks go to my brother Manfred for drawing my attention to one bummer of a high cholesterol christmas theorem in an early version of this paper. The results described in this article were achieved under a postdoctoral grant donated by GMD, Sankt Augustin.

References

[1] J.L. Doob. *Stochastic Processes*. John Wiley & Sons, 1953.

[2] 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.

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

[4] 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.