Characterization of ergodic hidden Markov sources

Alexander Schönhuth[1 ⋆], Member IEEE and Herbert Jaeger[2], Member IEEE

[1] Pacific Institute for the Mathematical Sciences
School of Computing Science
Simon Fraser University
8888 University Drive
Burnaby, BC, V5A 1S6, Canada
schoenhuth@cs.sfu.ca

[2] Jacobs University Bremen
School of Engineering and Science
Campus Ring
D-28759 Bremen, Germany

Abstract. An algebraic criterium for the ergodicity of discrete random sources is presented. For finite-dimensional sources, which contain hidden Markov sources as a subclass, the criterium can be effectively computed. This result is obtained on the background of a novel, elementary theory of discrete random sources, which is based on linear spaces spanned by word functions, and linear operators on these spaces. An outline of basic elements of this theory is provided.

Keywords. Asymptotic mean, dimension, entropy, ergodic, evolution operator, hidden Markov model, linearly dependent process, Markov chain, observable operator model, random source, stable, state generating function, stationary

1 Introduction

The theory of finite-valued Markov chains is fundamental for probability and information theory. By identifying states with the vertices of a graph and edge weights with transition probabilities one can conveniently infer a variety of statistical properties by inspecting combinatorial properties of the graph. A prevalent example is that (a special form of) ergodicity is equivalent to the underlying graph being irreducible and aperiodic (e.g. th. 6.4.17, [8]).

However, in case of hidden Markov chains (HMCs)—we subsequently speak of hidden Markov sources (HMSs) when we want to address the random source associated to an HMC—the inspection of combinatorial properties of the underlying Markov chain is of limited use to demonstrate ergodicity. In the general case, only sufficient, but not necessary conditions could be established, namely, the hidden Markov chain inherits ergodicity from the underlying Markov chain. For related work see [16, 7, 17] and also the excellent review [15] and citations therein. The main result of this paper is a novel—and to the best of our knowledge, the first—sufficient and necessary condition for the ergodicity of an arbitrary hidden Markov chain.

⋆ corresponding author

The criterium can be naturally established within a general theory of discrete-time, discrete-valued stochastic processes, which interprets processes as vectors in certain functional vector spaces. The first author has developed this theory in [3]. Since this work was written in German, the present paper also serves to make this line of research more accessible to an English-reading audience, while at the same time simplifying some aspects of the original theory as given in [3].

In sum, the original contributions of this paper are

(i) making accessible basic parts of the general algebraic theory of random sources given in [3], with improvements in simplicity and clarity of the theoretical account, including and up to a general algebraic criterium for ergodicity of discrete random sources,
(ii) to provide a criterion that characterizes ergodicity for the class of finite-dimensional sources (which include HMMs), which is based on standard spectral properties of a matrix and can be computationally tested
(iii) and, as a minor contribution, to sketch a general theory of classification of ergodic random sources.

The general framework within which we work branches from the theory of *observable operator models (OOMs)* which has been developed in the field of machine learning by the second author as a generalization of HMMs [13]. OOMs, in turn, can be seen as the culmination of a long series of investigations into the equivalence of HMMs (e.g., [6] [10] [12], survey in [13]), which has led to a generalization of hidden Markov sources termed *linearly dependent processes* [6] or *finitary sources* [10].

2 Random sources and word functions

As usual, $\Sigma^* = \cup_{k\geq 0}\Sigma^k$ denotes the set of all strings of finite length over the finite alphabet Σ together with the concatenation operation:

$$w \in \Sigma^t, v \in \Sigma^k \quad \Longrightarrow \quad wv \in \Sigma^{t+k}$$

where the word $\Box \in \Sigma^0$ of length $|\Box| = 0$ is the *empty string*. We denote the *length* of $w \in \Sigma^t$ by $|w| = t$ and write $a^t \in \Sigma^t$ for the concatenation of t times the letter a. Given a random source (X_t) we write

$$p_X(v = v_0...v_t) = \Pr(\{X_0 = v_0, ..., X_t = v_t\})$$

for the probability that the associated random source emits the string $v_0v_1...v_t$ at periods $s = 0, ..., t$. Accordingly, we think of random sources (X_t) as being specified by word functions

$$p_X : \Sigma^* \to [0,1] \subseteq \mathbb{R} \quad \text{such that} \sum_{a\in\Sigma} p(wa) = p(w) \quad \text{for all } w \in \Sigma^*, \tag{1}$$

assuming $p(\Box) = 1$, which implies

$$\sum_{w\in\Sigma^t} p(w) = 1 \quad \text{for all } t = 0, 1, \ldots. \tag{2}$$

Note that this class of word functions fully describe the class of one-sided random processes with values in Σ. To discern them from arbitrary word functions we refer to them as *stochastic word functions (SWFs)* in the following.

If convenient from a technical point of view, we identify one-sided random sources and the associated SWFs with probability measures on the measurable space of one-sided sequences

$$\Omega = \Sigma^{\mathbb{N}} = \bigotimes_{t=0}^{\infty} \Sigma$$

equipped with the σ-algebra $\mathcal{B}$ generated by the cylinder sets. In this vein, we sometimes identify subsets of words $A \subset \Sigma^t$ with cylinder sets $C[A] \in \mathcal{B}$ with where $C[A]$ is the set of all sequences whose prefixes are strings from A. In the special case of $A = \{v\}$ for a single word $v = v_0...v_t$ we have that $C[v] := C[\{v\}] = \{X_0 = v_0, ..., X_t = v_t\}$. In this vein, if p is an SWF and P is the probability measure associated with p then

$$P(C[A]) = p(A) := \sum_{v \in A} p(v)$$

for A a subset of words of equal length.

2.1 Operators

Upon having seen the string $w = w_0...w_t$ at time t, we think of the random source (X_t) as being in a *state* that depends only on w and completely describes the probabilities for the symbols to be produced at times $t+1, t+2, ...$ This is reflected by a transformation of the SWF p into an SWF p_w where

$$p_w(v) := p(v|w) = \Pr\{X_{t+1} = v_1, \ldots, X_{t+k} = v_k|w\} = p(wv)/p(w).$$

for $v = v_1...v_k\Sigma^k$.

This transformation can be described by an *observable operator* [13] τ_w which, in a more general fashion, acts as a linear operator on the linear space of word functions $\mathbb{R}^{\Sigma^*} = \{f : \Sigma^* \to \mathbb{R}\}$ and is defined by

$$(\tau_w f)(v) := f(wv)$$

for all $v \in \Sigma^*$. Note further that

$$\tau_{w_1...w_t} = \tau_{w_t} \circ ... \circ \tau_{w_1}. \tag{3}$$

If τ_w is applied to an SWF p with $p(w) > 0$ then $1/p(w)\tau_w p = p_w$ and $\tau_w p = 0$ in case of $p(w) = 0$. Accordingly, we define we $p_w = 0$ in case of $p(w) = 0$. We call p_w a *predictor function* of p. We extend the definitions of observable operators and predictor functions from words w to subsets of words of equal length $A \subset \Sigma^t$ by setting

$$\tau_A f := \sum_{w \in A} \tau_w f$$

that is, $(\tau_A f)(v) = \sum_{w \in A} f(wv)$, and ($p(A) := \sum_{v \in A} p(v)$) $p_A := 1/p(A)\tau_A p$

We further introduce the *evolution operator* μ on $\mathbb{R}^{\Sigma^*}$ which is defined by

$$(\mu f)(v) := \sum_{a \in \Sigma} (\tau_a f)(v) = \sum_{a \in \Sigma} f(av).$$

By multinomial expansion we obtain

$$\mu^t f = \tau_{\Sigma^t} f = \sum_{v \in \Sigma^t} \tau_v f. \tag{4}$$

2.2 Spaces and norms

We consider the set of word functions $\mathbb{R}^{\Sigma^*}$ as a vector space and define

$$\mathcal{S} := \text{span}\{f \in \mathbb{R}^{\Sigma^*} \mid f \text{ is stochastic}\}$$

which is the linear subspace of finite linear combinations of SWFs. Note that $\mathcal{S}$ can be identified with the linear space of finite, signed measures on $(\Omega, \mathcal{B})$. Therefore, we can make it a normed space by equipping it with the norm of total variation which we denote by $||.||$ (see appendix A for a brief compilation of the theory of finite, signed measures). Furthermore, in [19] it was shown that

$$||p|| = \sup_{t \in \mathbb{N}} \sum_{v \in \Sigma^t} |p(v)| = \lim_{t \in \mathbb{N}} \sum_{v \in \Sigma^t} |p(v)| \tag{5}$$

for $p \in \mathcal{S}$ which is a more handy characterisation of the norm of total variation in case of the measurable space at hand.

Clearly, $\tau_w(\mathcal{S}) \subset \mathcal{S}$ for all $w \in \Sigma^*$. Hence $\tau_A(\mathcal{S}) \subset \mathcal{S}$ as well as $\mu(\mathcal{S}) \subset \mathcal{S}$.

Lemma 1. *Let $A \subset \Sigma^t$ be a subset of words of equal length. Then it holds that*

$$||\mu|| = ||\tau_A|| = 1 \tag{6}$$

where here $||.||$ refers to the operator norm of endomorphisms on $\mathcal{S}$.

Proof. ¿From

$$\begin{aligned} \sum_{v \in \Sigma^s} |\tau_A p(v)| &= \sum_{v \in \Sigma^t} |\sum_{w \in A} p(wv)| \leq \sum_{w \in \Sigma^t} \sum_{v \in \Sigma^s} |p(wv)| \\ &= \sum_{u \in \Sigma^{t+s}} |p(u)| \leq ||p|| \end{aligned} \tag{7}$$

we obtain $||\tau_A|| \leq 1$. Further choose a sequence $\omega \in \Omega = \bigotimes_{t=0}^{\infty} \Sigma$ such that w is a prefix of ω for a $w \in A$. Let p_ω be the SWF associated with the random source that emits the sequence ω with probability one, that is

$$p_\omega(v) = \begin{cases} 1 & v \text{ is a prefix of } \omega \\ 0 & \text{else} \end{cases}.$$

It follows that both $||p_\omega|| = 1$ and $||\tau_A p_\omega|| = 1$ from which we obtain $||\tau_A|| = 1$. From $\mu = \tau_\Sigma$ we infer the left equation of (6). $\diamond$

2.3 Dimension

Given an SWF p, we consider the *predictor space*

$$\mathcal{V}_p := \text{span}\{p_w \mid w \in \Sigma^*\} = \text{span}\{\tau_w p \mid w \in \Sigma^*\} \subset \mathcal{S} \subset \mathbb{R}^{\Sigma^*}$$

that is, the linear subspace of finite linear combinations of predictor functions. This subspace can be identified with the column space of the infinite *prediction matrix*

$$\mathcal{P}_p = [p(v|w)_{v,w\in\Sigma^*}] \in \mathbb{R}^{\Sigma^*\times\Sigma^*} \ . \tag{8}$$

Analogously we define the *evolution space*

$$\mathcal{E}_p := \text{span}\{\mu^t p \mid t \in \mathbb{N}\} \subset \mathcal{S} \subset \mathbb{R}^{\Sigma^*}$$

which, because of (4), is a subspace of $\mathcal{V}_p$.

The dimension of $\mathcal{V}_p$ for an SWF p is referred to as the *dimension* of p resp. as the dimension of the random source associated with p. Accordingly, a random source is said to be *finite-dimensional* iff $\dim \mathcal{V}_p < \infty$. Analogously, the dimension of $\mathcal{E}_p$ is referred to as the *evolution dimension* of p resp. of the random source associated with p and p is said to be *finite-evolutiondimensional* iff $\dim \mathcal{E}_p < \infty$.

As finite dimension implies finite evolution dimension, the class of finite-dimensional sources is contained in that of the finite-evolutiondimensional sources. It can be shown that there are infinite-dimensional sources of finite evolution dimension [5].

If the dimension of an SWF p is finite there is a practicable way for reading it off the prediction matrix. Therefore, we set $\Sigma^{\leq t}$ to be the set of strings of length at most t and define

$$\mathcal{V}_p^t := \text{span}\{p_w \,|\, w \in \Sigma^{\leq t}\}.$$

Obviously $\mathcal{V}_p^t \subset \mathcal{V}_p^{t+1}$ for all $t \in \mathbb{N}$.

Lemma 2.

$$\forall t \in \mathbb{N}: \qquad \mathcal{V}_p^t = \mathcal{V}_p^{t+1} \quad \Rightarrow \quad \dim p = \dim \mathcal{V}_p^t. \tag{9}$$

Proof. It suffices to show that $\mathcal{V}_p^{t+n} = \mathcal{V}_p^t$ for all $n \in \mathbb{N}$. We will do that by induction on n where $n = 0$ is trivial. Let $n > 0$. Note that, because of (3),

$$\mathcal{V}_p^{t+n} = \text{span}\, \mathcal{V}_p^{t+n-1} \cup (\bigcup_{a\in\Sigma} \tau_a(\mathcal{V}_p^{t+n-1})). \tag{10}$$

Therefore, the left hand side of (9) translates to

$$\tau_a(\mathcal{V}_p^t) \subset \mathcal{V}_p^t \tag{11}$$

for all $a \in \Sigma$. To finish the proof we compute

$$\begin{aligned}\mathcal{V}_p^{t+n} &\overset{(10)}{=} \text{span}\, (\mathcal{V}_p^{t+n-1} \cup (\bigcup_{a\in\Sigma} \tau_a(\mathcal{V}_p^{t+n-1})) \\ &\overset{(*)}{=} \text{span}\, (\mathcal{V}_p^t \cup (\bigcup_{a\in\Sigma} \tau_a(\mathcal{V}_p^t)) \overset{(11)}{=} \mathcal{V}_p^t.\end{aligned}$$

where $(*)$ follows from the induction hypothesis. $\diamond$

Corollary 1.

$$\dim p = n \quad \Rightarrow \quad \mathcal{V}_p = \mathcal{V}_p^{n-1}. \tag{12}$$

Proof. Consider

$$\text{span}\{p\} = \mathcal{V}_p^0 \subset \mathcal{V}_p^1 \subset ... \subset \mathcal{V}_p^{n-1} \subset \mathcal{V}_p^n$$

which is a chain of vector spaces of length $n+1$. Because of (9) any equality in this chain will establish the desired result. Because of n being the dimension of $\mathcal{V}_p$ we will not find more than $n-1$ proper inclusions in this chain. So, at the latest, $\mathcal{V}_p^{n-1} = \mathcal{V}_p^n$. $\diamond$

In an analogous fashion we study the row space of the predictor matrix. Therefore we set

$$\mathcal{P}_{p,t} := [p(v|w)]_{v \in \Sigma^{\leq t}, w \in \Sigma^*} \in \mathbb{R}^{\Sigma^{\leq t} \times \Sigma^*}$$

that is, the rows of $\mathcal{P}_p$ which refer to strings of length at most t. We further write

$$f_v := [p(v|w)]_{w \in \Sigma^*}$$

for the v-row of $\mathcal{P}$. Note that for $u, v, w \in \Sigma^*$

$$f_u(wv) = p(u|wv) = \frac{1}{p(wv)} p(wvu) = \frac{p(w)}{p(wv)} p(vu|w) = \frac{p(w)}{p(wv)} f_{vu}(w). \tag{13}$$

Lemma 3.

$$\forall t \in \mathbb{N}: \qquad \text{rk } \mathcal{P}_{p,t} = \text{rk } \mathcal{P}_{p,t+1} \quad \Rightarrow \quad \dim p = \text{rk } \mathcal{P}_{p,t}. \tag{14}$$

Proof. We show that rk $\mathcal{P}_{p,t+2} =$ rk $\mathcal{P}_{p,t+1}$ from which the claim follows by induction on t. By assumption, for each $v \in \Sigma^{t+1}$

$$f_v = \sum_{u \in \Sigma^{\leq t}} \alpha_{v,u} f_u$$

that is, the v-row is a linear combination of u-rows where $|u| \leq t$. Let now $v = v_1 ... v_{t+2} \in \Sigma^{t+2}$. Writing $v' = v_2 ... v_{t+2} \in \Sigma^{t+1}$ we find that

$$\begin{aligned} f_v(w) = p(v|w) &= \frac{1}{p(w)} p(wv) = \frac{1}{p(w)} p(wv_1 v') \\ &= \frac{p(wv_1)}{p(w)} f_{v'}(wv_1) = \sum_{u \in \Sigma^{\leq r}} \frac{p(wv_1)}{p(w)} \alpha_{v',u} f_u(wv_1) \\ &\overset{(13)}{=} \sum_{u \in \Sigma^{\leq r}} \alpha_{v',u} f_{uv_1}(w) \end{aligned}$$

which shows that f_v is a linear combination of vectors from $\mathcal{P}_{p,t+1}$. $\diamond$

Corollary 2.

$$\dim p = n \quad \Longrightarrow \quad \text{rk } \mathcal{P}_p = \text{rk } \mathcal{P}_{p,n-1}. \tag{15}$$

Proof. This follows from considerations which are completely analogous to that of corollary 1. $\diamond$

Gathering the results from corollaries 1,2 the following lemma is obvious.

Lemma 4. *Let* p *be an SWF such that* $\dim p \leq n$*. Then*

$$\dim p = \text{rk } [p(v|w)]_{v,w\in\Sigma^{\leq n-1}} = \text{rk } [p(wv)]_{v,w\in\Sigma^{\leq n-1}}.$$

That is, n *is the rank of the finite submatrix of* $\mathcal{P}_p$ *whose entries refer to words up to length* $n-1$ *only.*

Proof. The left equation follows straightforwardly from corollaries 1,2 and the right one comes from $p(wv) = p(w)p(v|w)$. $\diamond$

2.4 Conditional SWFs

If p is an SWF of a random source (X_t) associated with a probability measure P on $(\Omega, \mathcal{B})$ and $B \in \mathcal{B}$ is an event for which $P(B) > 0$ we define an SWF p^B by

$$p^B(v = v_0...v_t) := \frac{1}{P(B)}P(C[v] \ \cap \ B) = \frac{1}{P(B)}P(\{X_0 = v_0, ..., X_t = v_t\} \ \cap \ B)$$

that is $p^B(v)$ reflects our knowledge about seeing the word v when we already know that B is to happen. We refer to p^B as a *conditional SWF*. We can establish the following relationship between conditional SWFs and predictor functions.

Lemma 5. *Let* p *be an SWF and* $A \subset \Sigma^t$ *where* $P(C[A]) = p(A) = \sum_{v\in A} p(v) > 0$ *for the probability measure* P *associated with* p*. It holds that*

$$\tau_A p^{C[A]} = \mu^t p^{C[A]} = p_A = \frac{1}{p(A)}\tau_A p. \tag{16}$$

Proof. Let $v \in \Sigma^*$. We compute

$$(\mu^t p^{C[A]})(v) = \sum_{w\in\Sigma^t} p^{C[A]}(wv) \overset{p^{C[A]}(wv)=0, w\notin A}{=} \sum_{w\in A} p^{C[A]}(wv) = (\tau_A p^{C[A]})(v)$$

which establishes the first equation of (16). Furthermore,

$$\begin{aligned}(\tau_A p^{C[A]})(v) &= \sum_{w\in A} p^{C[A]}(wv) \\ &= \sum_{w\in A} \frac{1}{P(C[A])}P(C[A]\cap C[wv]) = \sum_{w\in A} \frac{1}{P(C[A])}P(C[wv]) \\ &= \sum_{w\in A} \frac{1}{p(A)}p(wv) = \frac{1}{p(A)}(\tau_A p)(v)\end{aligned}$$

where the third equation follows from $C[wv] \subset C[A]$ which in turn is implied by $w \in A$. $\diamond$

Lemma 6. *Let p be an SWF and $B \in \mathcal{B}$ such that $P(B) > 0$ for the probability measure P associated to p. There is a sequence of subsets of words $F_n \subset \Sigma^n$ such that*

$$\lim_{n\to\infty} ||p^{C[F_n]} - p^B|| = 0. \tag{17}$$

Proof. ¿From the approximation theorem ([9]) we obtain a sequence of cylinder sets $C[F_n]$ such that

$$P(C[F_n] \triangle B) \underset{n\to\infty}{\longrightarrow} 0$$

where $A \triangle B$ is the symmetric set difference of two events A, B. Without loss of generality, these cylinder sets can be chosen such that $F_n \subset \Sigma^n$. Because of $|P(F_n) - P(B)| \leq P(F_n \triangle B)$ this in particular yields $P(F_n) \to_{n\to\infty} P(B)$. Therefore without loss of generality, $P(F_n) > 0$ for all n. It is well known (e.g. [7],?) that

$$||P - Q|| = 2 \sup_{B\in\mathcal{B}} |P(B) - Q(B)| \tag{18}$$

for arbitrary probability measures P, Q. Therefore

$$||p^{F_n} - p^B|| = 2 \sup_{C\in\mathcal{B}} |P(C|F_n) - P(C|B)| = |\frac{1}{P(F_n)} P(F_n \cap C) - \frac{1}{P(B)} P(B \cap C)|.$$

Knowing on one hand that $1/P(F_n) \to_{n\to\infty} 1/P(B)$ and on the other hand, by standard arguments from measure theory, that $|P(F_n \cap C) - P(B \cap C)| \leq P((F_n \cap C) \triangle (B \cap C)) \leq P(F_n \triangle B) \to_{n\to\infty} 0$ we obtain the claim of the lemma. ⋄

3 Ergodic Properties

3.1 Stationarity

We call $p \in \mathcal{S}$ *stationary* if $\mu p = p$. For an SWF p this is equivalent to $\dim \mathcal{E}_p = 1$, that is, p has evolution dimension 1. This straightforwardly translates to stationarity of the associated random source P as stationarity needs to be checked on generating events alone (here we immediately get $P(T^{-1}C[v]) = P(C[v])$ for all strings $v \in \Sigma^*$, where T is the familiar shift operator). Vice versa, $\mu p = p$ for the SWF p of a stationary random source P. As eigenvectors of a linear operator, the stationary random sources span a linear subspace

$$\mathcal{S}_\mu := \text{span}\{p \text{ SWF} \mid \mu p = p\} = \{p \in \mathcal{S} \mid \mu p = p\}.$$

3.2 Asymptotic Mean Stationarity

A random source P is called *asymptotically mean stationary* (AMS) if there is a stationary $\bar{P}$ such that

$$\forall B \in \mathcal{B}: \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(T^{-i}B) = \bar{P}(B). \tag{19}$$

$\bar{P}$ is called the *stationary mean* of P. A SWF p is called *asymptotically mean stationary* (AMS) if its associated random source P is. Furthermore, we denote an SWF p for which there is a stationary SWF $\bar{p} \in \mathcal{S}_\mu$ such that

$$\lim_{n\to\infty} ||\frac{1}{n}\sum_{i=0}^{n-1}\mu^i p - \bar{p}|| = 0 \tag{20}$$

as *strongly asymptotically mean stationary (strongly AMS)*. It can be shown that strong asymptotic mean stationarity is equivalent to asymptotic mean stationarity [18]. Here, we restrict ourselves to noting that strong asymptotic mean stationarity straightforwardly implies asymptotic mean stationarity as (20) translates to that the convergence of (19) is uniform in $B \in \mathcal{B}$, see (18). However, the reverse implication requires an involved ergodic theorem.

As it was shown in [5], finite evolution dimension implies asymptotic mean stationarity.

Theorem 1. *Let p be an SWF with $\dim \mathcal{E}_p < \infty$. Then it holds that*

$$\lim_{n\to\infty} ||\frac{1}{n}\sum_{i=0}^{n-1}\mu^i p - \bar{p}|| = 0$$

for a stationary SWF $\bar{p}$. Hence p is (strongly) AMS.

Proof. See [5], cor. 3.3. $\diamond$

As finite dimension implies finite evolution dimension this implies that finite-dimensional random sources are AMS. Note further the following lemma.

Lemma 7. *Let p be a strongly AMS SWF. Then it holds that*

$$\dim(\overline{\mathcal{E}_p} \cap \mathcal{S}_\mu) = 1 \tag{21}$$

where $\overline{\mathcal{E}_p}$ is the closure of the evolution space of p in $\mathcal{S}$.

Proof. The definition of the stationary mean $\bar{p}$ as the limit of the $1/n\sum_{i=0}^{n-1}\mu^i p \in \mathcal{E}_p$ immediately implies that $\bar{p} \in \overline{\mathcal{E}_p}$. Hence $\dim(\overline{\mathcal{E}_p} \cap \mathcal{S}_\mu) \geq 1$. Let $p^* \in \overline{\mathcal{E}_p} \cap \mathcal{S}_\mu$. We will show that

$$\text{dist}\,(p^*, \text{span}\{\bar{p}\}) = \inf_{q\,\in\,\text{span}\{\bar{p}\}} ||p^* - q|| = 0$$

from which the assertion follows. Therefore let $\epsilon \in \mathbb{R}_+$ and $(q_k)_{k\in\mathbb{N}}$ be a sequence from $\mathcal{E}_p$ which converges to p^*. By definition of $\mathcal{E}_p$ we can write

$$q_k = \sum_{j\in J_k} \alpha_{j,k}\mu^j p$$

for suitable finite $J_k \subset \mathbb{N}$ and $\alpha_{j,k} \in \mathbb{R}$. Therefore

$$\frac{1}{n}\sum_{i=0}\mu^i q_k = \sum_{j\in J_k}\alpha_{j,k}(\frac{1}{n}\sum_{i=0}^{n-1}\mu^{i+j}p) \longrightarrow_{n\to\infty} \sum_{j\in J_k}\alpha_{j,k}\bar{p} \in \text{span}\{\bar{p}\}.$$

Choose $K \in \mathbb{N}$ such that

$$||p^* - q_K|| < \frac{\epsilon}{2} \tag{22}$$

and, according to the considerations from above, $N_K \in \mathbb{N}$ such that for $q^* := \sum_{j \in J_K} \alpha_{j,K}\bar{p} \in$ span$\{\bar{p}\}$

$$||\frac{1}{N_K}\sum_{i=0}^{N_K-1}\mu^i q_K - q^*|| < \frac{\epsilon}{2}. \tag{23}$$

It follows that

$$\begin{aligned}
\text{dist}\,(p^*, \text{span}\{\bar{p}\}) &\leq ||p^* - q^*|| \\
&= ||p^* - \frac{1}{N_K}\sum_{i=0}^{N_K-1}\mu^i q_K + \frac{1}{N_K}\sum_{i=0}^{N_K-1}\mu^i q_K - q^*|| \\
&\leq ||p^* - \frac{1}{N_K}\sum_{i=0}^{N_K-1}\mu^i q_K|| + ||\frac{1}{N_K}\sum_{i=0}^{N_K-1}\mu^i q_K - q^*|| \\
&\overset{\mu p^* = p^*, (23)}{<} ||\frac{1}{N_K}\sum_{i=0}^{N_K-1}\mu^i p^* - \frac{1}{N_K}\sum_{i=0}^{N_K-1}\mu^i q_K|| + \frac{\epsilon}{2} \\
&\leq \frac{1}{N_K}\sum_{i=0}^{N_K-1}||\mu^i|| \cdot ||p^* - q_K|| + \frac{\epsilon}{2} \\
&\overset{(6)}{\leq} ||p^* - q_K|| + \frac{\epsilon}{2} \overset{(22)}{<} \epsilon.
\end{aligned}$$

$\diamond$

3.3 Invariant Events

An event $I \in \mathcal{B}$ is called *invariant* if $T^{-1}I = I$. The set of invariant events $\mathcal{I}$ is a sub-σ-algebra of $\mathcal{B}$.

Stationary probability measures can be identified by their values on invariant events alone. This is a consequence of the following lemma.

Lemma 8. *Let P be a* stationary finite signed measure *on $(\Omega, \mathcal{B})$, that is*

$$\forall B \in \mathcal{B} : \quad P(T^{-1}B) = P(B).$$

Then

$$P = 0 \quad \Longleftrightarrow \quad \forall I \in \mathcal{I} : \; P(I) = 0.$$

Proof. We have deferred the measure-theoretical proof to appendix A. $\diamond$

Note further that for SWFs p

$$\mu p = p \quad \Rightarrow \quad \forall I \in \mathcal{I} : \; \mu p^I = p^I \tag{24}$$

meaning that conditioning stationary SWFs on invariant events results in stationary SWFs which, when translated back to random sources, is a well-known result.

The following lemma is a key insight of this paper.

Lemma 9. *Let p be a stationary SWF and $I \in \mathcal{I}$ be an invariant event. Then it holds that*

$$p^I \in \overline{\mathcal{V}_p}. \tag{25}$$

That is, p^I lies in the closure of p's predictor space in $\mathcal{S}$.

Proof. For technical convenience, we subsequently identify p with its associated probability measure P. The case $p(I) = 0$ is trivial. For $p(I) > 0$ choose a sequence of subsets of strings $F_n \subset \Sigma^n$ such that $||p^{C[F_n]} - p^I|| \longrightarrow 0$ according to lemma 6. Without loss of generality $p(C[F_n]) > 0$ for all n. We compute

$$\begin{aligned} ||\tau_{F_n} p - p^I|| &\overset{(16),(24)}{=} ||\mu^n p^{C[F_n]} - \mu^n p^I|| \\ &\leq ||\mu^n|| \cdot ||p^{C[F_n]} - p^I||_{TV} \overset{(6)}{\leq} ||p^{C[F_n]} - p^I||. \end{aligned}$$

Therefore, the $\tau_{F_n} \in \mathcal{V}_p$ converge to p^I. Hence $p^I \in \overline{\mathcal{V}_p}$. ⋄

3.4 Ergodicity

A SWF p is said to be *ergodic* if its associated probability measure P is. That is,

$$\forall I \in \mathcal{I} : \quad P(I) \in \{0, 1\}. \tag{26}$$

For technical convenience, we will identify p with P and write $p(I)$ in the following.

REMARK If p is induced by a Markov chain then ergodicity, as given by this definition, is, in terms of the Markov chain, characterized by that there is only one closed, irreducible set of states (see th. 6.3.4, [8]).

Clearly, if p is AMS then p is ergodic if and only if its stationary mean $\bar{p}$ is. Moreover, if $A \in \Sigma^t$ is a subset of words and p is ergodic, then

$$p_A(I) \overset{(16)}{=} \mu^t p^A(I) = p^A(T^{-t} I) = p^A(I) = \frac{1}{p(A)} p(A \cap I) = \begin{cases} 1 & p(I) = 1 \\ 0 & p(I) = 0 \end{cases}. \tag{27}$$

Hence, p_A is itself ergodic as it agrees on the invariant sets with p. The main result of this paper is that in case of AMS SWFs p the concepts of ergodicity and predictor space can be coupled.

Theorem 2. *Let p be an AMS SWF and $\overline{\mathcal{V}_p}$ be the closure of its predictor space in $\mathcal{S}$. Then the following statements are equivalent:*

(i) p *is ergodic.*
(ii) $\overline{\mathcal{V}_p} \cap \mathcal{S}_\mu = \text{span}\{\bar{p}\}$.
(iii) $\dim(\overline{\mathcal{V}_p} \cap \mathcal{S}_\mu) = 1$.

Roughly speaking, the theorem tells that there is only one stationary word function in the boundary of the predictor space of an ergodic AMS SWF p and that is the stationary mean of p.

Proof. The equivalence of (ii) and (iii) is immediate as, by definition of the stationary mean $\bar{p}$, it always holds that

$$\bar{p} \in \overline{\mathcal{E}_p} \subset \overline{\mathcal{V}_p} \tag{28}$$

$(i) \Rightarrow (ii)$: Let p be ergodic. Because of (28), we have span$\{\bar{p}\} \subset \overline{\mathcal{V}_p} \cap \mathcal{S}_\mu$ for any choice of AMS p. Therefore it suffices to show

$$\overline{\mathcal{V}_p} \cap \mathcal{S}_\mu \subset \text{span}\{\bar{p}\}.$$

Assume the contrary, that is the existence of a $q \in \overline{\mathcal{V}_p}$ with $\mu q = q$ which is linearly independent of $\bar{p}$. Let p_n be a sequence in $\mathcal{V}_p$ that converges to q. Choose a basis of predictor functions (p_{v_i}) and represent p_n over this basis:

$$p_n = \sum_i \alpha_{i,n} p_{v_i}.$$

Because of (27) we know that the p_{v_i} agree with p on the invariant sets. Therefore $p_n(I) \in \{0, \sum \alpha_{i,n}\}$ for all invariant I. Convergence of the p_n to q in norm of total variation further implies

$$\forall I \in \mathcal{I}: \quad p_n(I) \to_{n\to\infty} q(I).$$

Hence the limes

$$K := \lim_{n\to\infty} \sum_i \alpha_{i,n}$$

exists and

$$q(I) = \begin{cases} K & \text{if } p(I) = \bar{p}(I) = 1 \\ 0 & \text{if } p(I) = \bar{p}(I) = 0 \end{cases}.$$

Assuming $K = 0$ would mean that $q(I) = 0$ for all invariant I. As a consequence of lemma 8 we would obtain $q = 0$ in this case which is a contradiction to the linear independence of q. In case of $K \neq 0$ we obtain that $(1/K)q$ is a stationary finite signed measure which agrees with $\bar{p}$ on the invariant sets. Hence (again because of lemma 8)

$$(1/K)q = \bar{p}$$

which again is a contradiction to the linear independence of q.

$(iii) \Rightarrow (i)$: Let p be not ergodic. Hence there is an invariant I with

$$\bar{p}(I) = p(I) = \alpha \in]0,1[. \tag{29}$$

As $\bar{p} \in \overline{\mathcal{V}_p}$ we know from the definition of predictor space that

$$\overline{\mathcal{V}_{\bar{p}}} \subset \overline{\mathcal{V}_p}.$$

¿From lemma 9 we further know that

$$\bar{p}^I, \bar{p}^{\complement I} \in \overline{\mathcal{V}_{\bar{p}}}.$$

Because of (29)

$$\bar{p}^I(I) = 1 \neq 0 = \bar{p}^{\complement I}(I)$$
$$\bar{p}^I(\complement I) = 0 \neq 1 = \bar{p}^{\complement I}(\complement I)$$

which implies that $\bar{p}^I, \bar{p}^{\complement I}$ are linearly independent as finite signed measures. This immediately reveals them as linearly independent word functions. $\diamond$

This theorem becomes particularly useful in case of finite-dimensional SWFs p.

Corollary 3. *Let p be a finite-dimensional SWF. Then p is ergodic if and only if*

$$\dim(\mathcal{V}_p \cap \mathcal{S}_\mu) = 1. \tag{30}$$

Proof. As p is AMS (see th. 1) theorem 2 applies for p. It remains to notice that $\overline{\mathcal{V}_p} = \mathcal{V}_p$ for finite-dimensional $\mathcal{V}_p$. $\diamond$

It is this corollary that the algorithm for deciding ergodicity of hidden Markov sources is based on. We will expand on this issue in section 5.1.

4 Classification of ergodic sources

We conclude our general treatment of ergodic sources this section with some remarks on how the different classes of such sources, as introduced by this wor,k are related to one another. Writing $\mathcal{S}_{e,AMS}$ resp. $\mathcal{S}_{e,edim}$ resp. $\mathcal{S}_{e,dim}$ resp. $\mathcal{S}e,\mu$ for the classes of ergodic AMS resp. ergodic finite-evolutiondimensional resp. ergodic finite-dimensional resp. ergodic stationary sources it holds that

$$\mathcal{S}_{e,AMS} \;\supset\; \mathcal{S}_{e,edim} \;\supset \begin{cases} \mathcal{S}_{e,dim} \\ \mathcal{S}_{e,\mu} \end{cases} \tag{31}$$

where the first inclusion is theorem 1 and the second one immediately follows from the definitions of stationarity, dimension and evolution dimension. We also know that

$$\mathcal{S}_{e,dim} \;\not\subset\; \mathcal{S}_{e,\mu}$$

as, for example, it is known that hidden Markov sources are finite-dimensional (see [10, 12, 5]) and there are non-stationary ergodic hidden Markov sources. Furthermore,

$$\mathcal{S}_{e,AMS} \;\supsetneq\; \mathcal{S}_{e,edim}$$

because of the following lemma.

Lemma 10. *There is an ergodic AMS source of infinite evolution dimension.*

Proof. Let $\Sigma = \{a, b\}$ and $\alpha \in]0, 1[$. We consider the SWF p which is recursively defined by

$$p(v) = \begin{cases} 1 & v = \square \\ \alpha^{|w|+1} p(w) & \exists w \in \Sigma^* :\; v = wa \;. \\ (1-\alpha^{|w|+1}) p(w) & \exists w \in \Sigma^* :\; v = wb \end{cases} \tag{32}$$

For example, $p(abab) = \alpha(1-\alpha^2)\alpha^3(1-\alpha^4)$. It is straightforward to show that p is indeed an SWF. It encodes the independent process $(X_t)_{t\in\mathbb{N}}$ with values in Σ given by

$$P(X_t = a) = \alpha^{t+1}, P(X_t = b) = 1 - \alpha^{t+1}$$

and

$$P(X_0 = a_0, \cdots, X_{t-1} = a_{t-1}) = P(X_0 = a_0) \times \cdots \times P(X_{t-1} = a_{t-1}).$$

Note first that ($v \in \Sigma^*$)

$$\mu^k p(v) = \begin{cases} 1 & v = \square \\ \alpha^{|v|+k}\mu^k p(w) & \exists w \in \Sigma^* :\ v = wa \ , \\ (1-\alpha^{|v|+k})\mu^k p(w) & \exists w \in \Sigma^* :\ v = wb \end{cases} \tag{33}$$

which can straightforwardly inferred by induction on k.

Infinite evolution dimension: For showing that $\dim \mathcal{E}_p = \infty$ we consider the matrices

$$A_n := (\mu^{k-1}p(a^i))_{1\le i,k\le n} \in \mathbb{R}^{n\times n}.$$

¿From (33) we infer

$$\mu^k p(a^i) = \alpha^{\sum_{t=1}^{i}(k+t)}.$$

Hence

$$\begin{aligned} det(A_n) &= det \begin{pmatrix} \alpha & \alpha^2 & \dots & \alpha^n \\ \alpha^{1+2} & \alpha^{2+3} & \dots & \alpha^{n+n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{1+\ldots+n-1} & \alpha^{2+\ldots+n} & \dots & \alpha^{n+\ldots+2n-1} \end{pmatrix} \\ &= \prod_{k=1}^{n} \alpha^{2n-1} det \begin{pmatrix} 1 & \alpha & \dots & \alpha^{n-1} \\ 1 & \alpha^2 & \dots & \alpha^{2(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha^n & \dots & \alpha^{n(n-1)} \end{pmatrix} \\ &= \prod_{k=1}^{n} \alpha^{2n-1} \prod_{1\le i,j\le n, i<j} (\alpha^i - \alpha^j) \neq 0, \end{aligned}$$

where the last equation follows from that the matrix is a *Vandermonde matrix* (see [11], sec. 6.1). Therefore, the rank of the infinite set $(p, \mu p..., \mu^{n-1}p, ...)$ is not bounded which translates to $\dim \mathcal{E}_p = \infty$.

Asymptotic mean stationarity: We define a vector $\bar{p}$ by

$$\bar{p}(v) = \begin{cases} 1 & \text{if } v = b^{|v|} = b...b \in \Sigma^{|v|} \\ 0 & \text{else} \end{cases} \tag{34}$$

and prove that

$$\lim_{n\to\infty} ||\mu^n p - \bar{p}||_{TV} \overset{(5)}{=} \lim_{n\to\infty} \sup_{t\in\mathbb{N}} \sum_{v\in\Sigma^t} |\mu^n p(v) - \bar{p}(v)| = 0 \tag{35}$$

from which we can clearly infer that p is AMS. Consider

$$\sum_{v\in\Sigma^t} |\mu^n p(v) - \bar{p}(v)| = 1 - \mu^n p(b^t) + \sum_{v\in\Sigma^t\setminus\{b^t\}} \mu^n p(v). \tag{36}$$

To order to show (35) we will show that $\mu^n p(b^t)$ converges to 1 *uniformly* in t. Therefore, we will prove that (let $\log$ be the natural logarithm)

$$\log \frac{1}{\mu^n p(b^t)} \le \frac{\alpha^{n+1}}{(1-\alpha)^2}, \tag{37}$$

as this implies

$$1 \ge \mu^n p(b^t) \ge (\exp(\frac{\alpha^{n+1}}{(1-\alpha)^2}))^{-1} \underset{n\to\infty}{\longrightarrow} 1$$

and with it the assertion. To do this we first note that, because of the mean value theorem, for all $r > 1$ there is $\xi \in [r-1, r]$ such that

$$\log(r) - \log(r-1) = \frac{\log(r) - \log(r-1)}{r-(r-1)} = (\log)'(\xi) = \frac{1}{\xi} \le \frac{1}{r}. \tag{38}$$

In order to establish (37) we finally compute

$$\begin{aligned}
\log\frac{1}{\mu^n p(b^t)} &\overset{(33)}{=} \log(\prod_{l=1}^{t} \frac{1}{1-\alpha^{l+n}}) = \log(\prod_{l=1}^{t} \frac{(1/\alpha)^{l+n}}{(1/\alpha)^{l+n}-1}) \\
&= \sum_{l=1}^{t} \log((1/\alpha)^{l+n}) - \log((1/\alpha)^{l+n}-1) \\
&\overset{(38)}{\le} \sum_{l=1}^{t} \frac{1}{(1/\alpha)^{l+n}-1} = \sum_{l=1}^{t} \frac{\alpha^{l+n}}{1-\alpha^{l+n}} \le \sum_{l=1}^{t} \frac{\alpha^{l+n}}{1-\alpha} \\
&= (1-\alpha)\sum_{l=1}^{t} \alpha^{l+n} = (1-\alpha)\alpha^{n+1}\sum_{l=1}^{t}\alpha^{l-1} \\
&\le \frac{\alpha^{n+1}}{(1-\alpha)^2}.
\end{aligned}$$

Therefore, p is AMS.

Ergodicity: As a preparation, we consider that for $v \in \Sigma^*$

$$\begin{aligned}
\tau_a\mu^k p(v) &= \mu^k p(av) = \alpha^{k+1}\cdot\mu^{k+1}p(v), \\
\tau_b\mu^k p(v) &= \mu^k p(bv) = (1-\alpha^{k+1})\cdot\mu^{k+1}p(v)
\end{aligned} \tag{39}$$

where the equations on the left are just the definition of τ_a, τ_b and the equations on the right follow by induction on the word length $|v|$. This implies

$$\tau_a\mu^k p,\ \tau_b\mu^k p \in \text{span}\{\mu^{k+1}p\} \subset \mathcal{E}_p.$$

from which we immediately get $\tau_a(\mathcal{E}_p) \subset \mathcal{E}_p, \tau_b(\mathcal{E}_p) \subset \mathcal{E}_p$. Hence, because of (3),

$$\tau_w(\mathcal{E}_p) \subset \mathcal{E}_p$$

for all $w \in \Sigma^*$ which further translates to $\mathcal{V}_p \subset \mathcal{E}_p$. As always $\mathcal{E}_p \subset \mathcal{V}_p$ we finally obtain

$$\dim(\overline{\mathcal{V}_p} \cap \mathcal{S}_\mu) = \dim(\overline{\mathcal{E}_p} \cap \mathcal{S}_\mu) \stackrel{(21)}{=} 1$$

and theorem 2 implies the ergodicity of p. ◇

FINAL REMARK: The relationship between the classes of stationary and finite-dimensional ergodic sources has not been fully explored yet. Unlike in the case of arbitrary non-ergodic sources, the question of existence of an infinite-dimensional, stationary source has not been answered for the class of ergodic sources. As is easily checked, the aforementioned example source p (see [5], lemma 6) has the remarkable property that $\mathcal{V}_p \subset \mathcal{S}_\mu$ which further translates to $\dim(\mathcal{V}_p \cap \mathcal{S}_\mu) = \infty$. This is quite the opposite of being ergodic according to theorem 2.

5 Observable Operator Models

Finite-dimensional random sources p can be parameterized by identifying the finite-dimensional $\mathcal{V}_p$ with an $\mathbb{R}^n$ where $n = \dim \mathcal{V}_p$ and providing matrix representations T_v for the observable operators τ_v. The crucial point is that such a parameterization is finite as, by providing matrix representations T_a for $a \in \Sigma$ only we obtain the remaining matrices by

$$T_{v=v_t \dots v_1} = T_{v_t} \cdot \dots \cdot T_{v_1}$$

which holds because of (3). To put it more concrete, we choose a basis of predictor functions $p_{w_j}, j = 1, ..., n$ that are identified with $e_i = (0, .., 0, \underset{i}{1}, 0, ..., 0) \in \mathbb{R}^n$ and set e_p to be the coordinate representation of p according to this basis. If $\sum_{j=0}^n \alpha_{a,i,j} e_j$ is a representation of $\tau_a p_{w_i}$ on this basis then corresponding matrix representations T_a of τ_a are obtained by setting

$$(T_a)_{ij} := \alpha_{a,i,j}.$$

Observe further that probabilities $p(v = v_1 ... v_t)$ can be read off the coefficients of $T_v e_p \in \mathbb{R}^n$ (which represents $\tau_v p$) the following way:

$$e_v = \sum_{i=1}^n \beta_i e_i \quad \Rightarrow \quad p(v) = \sum_{i=1}^n \beta_i.$$

This follows from the translation

$$p(v) = \tau_v p(\Box) = \sum_{i=1}^n \beta_i \underbrace{p_{w_i}(\Box)}_{=1}$$

back to the world of word functions. These observations are summarized within the following theorem.

Theorem 3. *A SWF p is finite-dimensional if and only if there is $n \in \mathbb{N}$ such that on $\mathbb{R}^n$ there are $e_p \in \mathbb{R}^n$ and $T_a \in \mathbb{R}^{n\times n}, a \in \Sigma$ for which*

$$p(v = v_1 ... v_t) = \mathbf{1}_n^T T_{v_t} ... T_{v_1} e_p \tag{40}$$

where $\mathbf{1}_n = (1, ..., 1) \in \mathbb{R}^n$.

Proof. See [13, 5] for variants of the following. By identifying $\mathcal{V}_p$ with $\mathbb{R}^n$ for $n = \dim \mathcal{V}_p$ and, accordingly, e_p with a coordinate vector of p and T_a with matrix representations of the observable operators $\tau_a : \mathcal{V}_p \to \mathcal{V}_p$, the first direction follows from the considerations from above. For the inverse direction define

$$g_v := \mathbf{1}_v^T = \mathbf{1}^T T_{v_t} ... T_{v_1}$$

for all $v = v_1 ... v_t \in \Sigma^*$. Define word functions $p_i, i = 1, ..., n$ by

$$p_i(v) := g_v e_i$$

for all $v \in \Sigma^*$. Now consider the w-row of the prediction matrix $\mathcal{P}$, that is

$$\mathcal{P}_w := (p(v|w))_{w\in\Sigma^*}$$

in case of $p(w) \neq 0$, see (8). Let $T_w e_p = \sum_i \alpha_i e_i$. According to (40) we compute

$$\begin{aligned} p(v|w) &= \frac{1}{p(w)} p(wv) \\ &= \frac{1}{p(w)} \mathbf{1}^T T_{wv} e_p = \frac{1}{p(w)} \mathbf{1}^T T_v T_w e_p = \frac{1}{p(w)} f_v T_w e_p \\ &= \sum_{i=1}^n \frac{1}{p(w)} \alpha_i f_v e_i = \sum_{i=1}^n \frac{1}{p(w)} \alpha_i p_i(v). \end{aligned}$$

This translates to that $\mathcal{P}_w$ is a linear combination of the p_i. Hence

$$\dim p = \text{rk}\, \mathcal{P} = \dim \text{span}\{\mathcal{P}_w \mid w \in \Sigma^*\} \leq \dim \text{span}\{p_i \mid i = 1, ..., n\} \leq n.$$

$\diamond$

Note immediately that for an SWF p given by a representation from the theorem, the SWF's dimension does not necessarily have to coincide with that of the underlying $\mathbb{R}^n$. Indeed it is easy to come up with examples where $n > \dim p$.

Definition 1 ([13]). *Tuples $(\mathbb{R}^n, (T_a)_{a\in\Sigma}, e_p)$ encoding finite-dimensional SWFs p have been termed* Observable Operator Models *(OOMs). If $n = \dim p$ we speak of a* minimal-dimensional *OOM:*

The investigation of OOMs has led to a class of learning algorithms which, on a variety of natural instances, outperform their classical counterpart, the EM algorithm, for HMCs [14]. Therefore note that HMCs can be canonically transformed to OOMs which, above all, reveals them as finite-dimensional. We will draw the connection between HMCs and OOMs in subsection 5.1.

5.1 HMCs to OOMs

In its most prevalent form, a finite-valued HMM is given by a set of hidden states $Q = \{1, ..., n\}$ and a finite set Σ of output symbols. The hidden states form a Markov chain and corresponding transition probabilities a_{ij} of changing from state i to state j are collected in a matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$. We further have an *emission probability distribution* for each hidden state over the output symbols which are given by an emission matrix $E = (e_{ia})_{1 \leq i \leq n, a \in \Sigma}$ where e_{ia} is the probability that symbol $a \in \Sigma$ is emitted from state $i \in Q$. Finally, there is an *initial probability distribution* $\pi = (\pi_1, ..., \pi_n)$ over the hidden states. The probability that the HMM emits a string of symbols $v = v_1...v_t \in \Sigma^t$ is then computed as

$$P_{HMM}(v = v_1...v_t) = \sum_{i_1...i_t \in Q^t} \pi_{i_1} e_{i_1 v_1} a_{i_1 i_2} e_{i_2 v_2} ... a_{i_{t-1} i_t} e_{i_t v_t}.$$

To identify the HMM as finite-dimensional, we define matrices $O_a \in \mathbb{R}^{n \times n}$ for each output symbol $a \in \Sigma$ through

$$(O_a)_{ij} = \begin{cases} e_{ia} & i = j \\ 0 & i \neq j \end{cases}$$

and further

$$T_a := A^T O_a \in \mathbb{R}^{n \times n}.$$

It then turns out that

$$P_{HMM}(v) = \mathbf{1}_n^T T_{v_t} ... T_{v_1} \pi$$

which, because of theorem 3, shows that the random source encoded by the HMM has dimension of at most n.

5.2 Ergodicity of OOMs

If an OOM is minimal-dimensional the theorems from earlier sections can be applied to it by identifying the OOM as a coordinate representation of the finite-dimensional SWF encoded by it. This provides us with a way to check minimal-dimensional OOMs for ergodicity.

Theorem 4. *Let $(T_a \in \mathbb{R}^{n \times n})_{a \in \Sigma}, e_p \in \mathbb{R}^n$ be a minimal-dimensional OOM. Let $M := \sum_{a \in \Sigma} T_a$ be the sum of the matrices T_a. Then the finite-dimensional SWF p encoded by the OOM is ergodic if and only if*

$$\dim \mathrm{Eig}\ (M; 1) = 1$$

that is, M's eigenspace of the eigenvalue 1 is one-dimensional.

Proof. This is straightforwardly established by identifying the parameterization with a coordinate representation of the finite-dimensional SWF p where it turns out that M is a matrix representation of the evolution operator μ. Subsequent application of corollary 3 yields the result. ◇

6 Computationally Testing HMCs for Ergodicity

Based on the insights from section 5 we can come up with an algorithm for checking HMCs for ergodicity.

1. Produce a matrix representation M of the evolution operator in an equivalent minimal-dimensional OOM.
2. Check the dimension d of the eigenspace of the matrix $M = \sum_{a\in\Sigma} \tilde{T}_a$ for the eigenvalue 1.
3. Output yes, if $d = 1$ and no else.

As checking the dimension of eigenspaces is routine, the second point poses no major problems. The first point, though, needs to be illustrated.

We cast the first point's problem in a more general fashion and consider arbitrary SWFs p such that $\dim p \leq n$. According to lemma 4

$$m := \dim p = \text{rk}\,[p(wv)]_{v,w\in\Sigma^{\leq n-1}} \leq n.$$

We choose words $v_i, w_j \in \Sigma^{\leq n-1}, i, j = 1, ..., m$ such that the matrix

$$V := [p(v_i|w_j)]_{i,j=1,\dots,m}$$

is regular. As a consequence we know that $p_{w_j}, j = 1, ..., m$ is a basis of $\mathcal{V}_p$.

Lemma 11. *Let p be an SWF of finite dimension. Let $w_j, v_i, i, j = 1, ..., m$ and V be chosen by the procedure from above. Define matrices*

$$W_a := [p(av_i|w_j)]_{i,j=1,\dots,m}$$

for all $a \in \Sigma$. Then (p_{w_j}) is a basis of $\mathcal{V}_p$ and

$$T_a := V^{-1}W_a$$

is a matrix representation corresponding to the coordinate representation

$$\begin{array}{rccc} \Phi: & \mathcal{V}_p & \longrightarrow & \mathbb{R}^m \\ & p_{w_j} & \mapsto & e_j \end{array}.$$

Hence $M := \sum_{a\in\Sigma} T_a$ is a matrix representation of the evolution operator.

Proof. Consider the alternative coordinate representation

$$\begin{array}{rccc} \Phi': & \mathcal{V}_p & \longrightarrow & \mathbb{R}^m \\ & p_{w_j} & \mapsto & V^j \end{array}$$

where $V^j := (p(v_1|w_j), ..., p(v_m|w_j))$ is the j-th column of V. From $\tau_a p_{w_j}(v_i) = p(av_i|w_j)$ we know that for a matrix representation T'_a of τ_a according to Φ'

$$T'_a(V^j) = W^j_a \tag{41}$$

where W_a^j is the j-th column of W_a. Note that $\Phi' \circ \Phi^{-1}(e_j) = V^j$. So $\Phi' \circ \Phi$ is precisely described by the matrix representation V. Therefore we obtain a commutative diagram

$$\begin{array}{ccc} \mathbb{R}^m & \xrightarrow{T_a} & \mathbb{R}^m \\ V\big\downarrow & & \big\downarrow V. \\ \mathbb{R}^m & \xrightarrow{T'_a} & \mathbb{R}^m \end{array}$$

which translates to $VT_a = T'_aV$. Because of (41) $T'_aV = W_a$ from which the lemma's assertion follows. ◇

REMARK As spectra of linear operators do not change under similarity transformations we could have directly chosen $M' := \sum_{a\in\Sigma} T'_a$ as a choice for the evolution operator where T'_a would have been defined by the equations $T'_a(V^j) = W_a^j$. However we wanted to provide a basis such that the matrix representations give rise to an OOM.

6.1 Example

We conclude with an example of an ergodic HMM whose underlying Markov chain is not ergodic. Let $\mathcal{M}$ be a 3-state HMM over the alphabet $\{0, 1\}$ parameterized by

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } E = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

where A is the transition matrix of the underlying Markov chain and E is the emission matrix of the hidden states over the symbols $\{0.1\}$. At the beginning, state no. 1 is entered with probability one. The underlying Markov chain has two closed, irreducible sets of states (states no. 2 and 3 each make up one of them) hence is not ergodic. Indeed, a somewhat closer second look immediately reveals the ergodicity of the HMC as a stochastic process that almost surely generates sequences with only finitely many 0s.

According to the procedure above, we find that the dimension is 2 and that

$$V = \begin{bmatrix} p(\Box) & p(\Box|0) \\ p(0) & p(0|0) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & \frac{1}{2} \end{bmatrix}$$

is regular. Further

$$W_0 = \begin{bmatrix} p(0) & p(0|0) \\ p(00) & p(00|0) \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

and

$$W_1 = \begin{bmatrix} p(1) & p(1|0) \\ p(10) & p(10|0) \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{bmatrix}$$

According to lemma 11 a matrix representation of the evolution operator is

$$M = V^{-1}(W_0 + W_1) = \begin{bmatrix} -1 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ \frac{1}{2} & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & \frac{3}{2} \end{bmatrix}.$$

One can then straightforwardly check that M's eigenvalues are 1 and $1/2$, from which $\dim Eig(M; 1) = 1$ follows. Hence the HMC $\mathcal{M}$ is ergodic.

7 Discussion

In this paper, we have developed a criterion for an HMC to be ergodic that can be algorithmically tested. However, the algorithm presented here is exponential in the number of hidden states of the HMC and thus of limited use. It is an open question whether there is an efficient algorithm which outputs matrix representations of the operators involved in the theory presented here and we conjecture that, indeed, there is. A hint to this is that Balasubramanian [2] claims to have one to solve the identifiability problem whose solution in [12] as well is based on an exponential algorithm and the solutions presented in these works are relatives of ours.

In a subsequent paper, we shall explore the spectrum of the evolution operator to expand on the issue of classification of finite-dimensional sources which is justified by that they do not only include HMMs, but also quantum random walks [1], [4], a statistical model that serves the emulation of Markov chain Monte Carlo methods on quantum computers.

A Finite signed measures

A finite, signed measure on $(\Omega, \mathcal{B}(\Sigma))$ is a σ-additive but not necesarily positive, finite set function on $\mathcal{B}(\Sigma)$. The most relevant properties of finite signed measures are summarized in the following theorem (see [9], ch. VI for proofs).

Theorem 5.

(i) *The* Jordan decomposition *theorem tells that for every $P \in \mathcal{P}$ there are finite measures P_+, P_- such that*

$$P = P_+ - P_-$$

and for every other decomposition $P = P_1 - P_2$ with measures P_1, P_2 it holds that $P_1 = P_+ + \delta, P_2 = P_- + \delta$ for another measure δ. In this sense, P_+ and P_- are unique and called positive *resp.* negative variation*. The measure $|P| := P_+ + P_-$ is called* total variation.

(ii) *In parallel to the Jordan decomposition we have the* Hahn decomposition *of Ω into two disjoint events Ω_+, Ω_-*

$$\Omega = \Omega_+ \,\dot{\cup}\, \Omega_-$$

such that $P_-(\Omega_+) = 0$ and $P_+(\Omega_-) = 0$. Ω_+, Ω_- are uniquely determined up to $|P|$-null-sets.

(iii) *The* norm of total variation $||.||_{TV}$ *on $\mathcal{P}$ is given by*

$$||P||_{TV} := |P|(\Omega) = P_+(\Omega) + P_-(\Omega) = P_+(\Omega_+) + P_-(\Omega_-).$$

Obviously $||\,|P|\,||_{TV} = ||P||_{TV}$.

A.1 Proof of lemma 8

Before it comes to proving the lemma, we provide us with a preparatory result.

Lemma 12. *Let P be a finite, signed measure on $(\Omega, \mathcal{B})$. Then $P \circ T^{-1} = P$ if and only if both $P_+ \circ T^{-1} = P_+$ and $P_- \circ T^{-1} = P_-$ are.*

Proof. The inverse direction is obvious as $P = P_+ - P_-$. For the other direction first note that for an arbitrary measure Q, by definition of the norm of total variation (th. 5, (iii))

$$||Q \circ T^{-1}|| = Q(T^{-1}\Omega) = Q(\Omega) = ||Q||. \tag{42}$$

Further observe that $P = P \circ T^{-1} = P_+ \circ T^{-1} - P_- \circ T^{-1}$. Hence

$$\begin{aligned} ||P|| = ||P \circ T^{-1}|| &= ||(P_+ - P_-) \circ T^{-1}|| \\ &\le ||P_+ \circ T^{-1}|| + ||P_- \circ T^{-1}|| \\ &\overset{(42)}{=} ||P_+|| + ||P_-|| = ||P||. \end{aligned}$$

Therefore $||P|| = ||P_+ \circ T^{-1}|| + ||P_- \circ T^{-1}||$. As $P = P_+ \circ T^{-1} - P_- \circ T^{-1}$ the lemma's claim follows from the uniqueness property of the Jordan deocmposition (see th. 5, (i)). ⋄

We are now in position to prove lemma 8.

Proof. "$\Longrightarrow$" is trivial. For the inverse direction we assume the existence of a finite signed measure $P \neq 0$ with $P(I) = 0$ for $I \in \mathcal{I}$. Because of lemma 12 P_+, P_- are stationary and so, without loss of generality $P_+ \neq 0$. Let Ω_+, Ω_- the Hahn decomposition of P, that is, $\Omega = \Omega_+ \,\dot\cup\, \Omega_-$ and $P_+(\Omega_+) = P_+(\Omega), P_-(\Omega_-) = P_-(\Omega)$. As $P_+ > 0$ we obtain $P_+(\Omega_+) > 0$. We now define

$$I_+ := \limsup_n T^{-n}\Omega_+ = \bigcap_{n \ge 0} \bigcup_{m \ge n} T^{-m}\Omega_+ \ \subset \ \bigcup_{n \ge 0} T^{-n}\Omega_+.$$

Clearly, I_+ is invariant. Further

$$\begin{aligned} P_-(I_+) &\le P_-(\bigcup_{n \ge 0} T^{-n}\Omega_+) \\ &\le \sum_{n \ge 0} P_-(T^{-n}\Omega_+) \overset{(*)}{=} \sum_{n \ge 0} P_-(\Omega_+) = 0 \end{aligned}$$

as well as

$$\begin{aligned} P_+(I_+) &= P_+(\limsup_n T^{-n}\Omega_+) \\ &\overset{(**)}{\ge} \limsup_{n \to \infty} P_+(T^{-n}\Omega_+) \overset{(*)}{=} P_+(\Omega_+) > 0, \end{aligned}$$

where $(*)$ follows from lemma 12 and $(**)$ is a consequence of Fatou's lemma Herewith

$$P(I_+) = P_+(I_+) - P_-(I_+) = P_+(I_+) > 0.$$

which is a contradiction to that P vanishes on the invariant events. ⋄

References

1. D. Aharonov, A. Ambainis, J. Kempe, U. Vazirani, ”Quantum walks on graphs”, in *Proc. of 33rd ACM STOC, New York*, 2001, pp. 50-59.
2. V. Balasubramanian, “Equivalence and reduction of hidden Markov models”, *Technical Report 1370, MIT AI Lab*, 1993.
3. A. Schönhuth, “Discrete-valued stochastic vector spaces” (German), PhD thesis, University Cologne, 2006.
4. U. Faigle, A. Schönhuth, ”Quantum predictor models”, *Electonic Notes in Discrete Mathematics*, vol. 25, pp. 149-155, 2006.
5. U. Faigle and A. Schoenhuth, “Asymptotic mean stationarity of sources with finite evolution dimension”, *IEEE Trans. Inf. Theory*, vol. 53(7), pp. 2342-2348, 2007
6. E.J. Gilbert, ”On the identifiability problem for functions of finite Markov chains”, *Ann. Math. Stat.*, vol. 30, pp. 688-697, 1959.
7. Robert M. Gray, *Entropy and Information Theory*. Springer Verlag, 1990.
8. G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford University Press, 2001.
9. P.R. Halmos, *Measure Theory*. Van Nostrand, 1964.
10. A. Heller “On stochastic processes derived from Markov chains”, *Annals of Mathematical Statistics*, vol. 36(4), pp. 1286-1291, 1965
11. R.A. Horn and C.A.Johnson, *Topics in matrix analysis*. Cambridge University Press, 1991.
12. H. Ito, S.-I. Amari and K. Kobayashi “Identifiability of hidden Markov information sources and their minimum degrees of freedom”, *IEEE Trans. Inf. Theory*, vol. 38(2), pp. 324-333, 1992.
13. H. Jaeger. “Observable operator models for discrete stochastic time series”, *Neural Computation*, vol. 12(6), pp. 1371-1398, 2000.
14. H. Jaeger, M. Zhao, K. Kretzschmar, T. Oberstein, D. Popovici, and A. Kolling, “Learning observable operator models via the ES algorithm”, in: S. Haykin, J. Principe, T. Sejnowski, J. McWhirter (eds.), *New Directions in Statistical Signal Processing: from Systems to Brain*, MIT Press, Cambridge, MA., pp. 417-464, 2006.
15. Y. Ephraim, N. Merhav, ”Hidden Markov processes”, *IEEE Trans. on Information Theory*, vol. 48(6), pp. 1518-1569, 2002.
16. L. E. Baum, T. Petrie, ”Statistical inference for probabilistic functions of finite state Markov chains”, *Annals of Mathematical Statistics*, vol. 37, pp. 1554-1563, 1966.
17. B. G. Leroux, ”Maximum-likelihood estimation for hidden Markov models”, *Stochastic Processes and Their Applications*, vol. 40, pp. 127-143, 1992.
18. A. Schönhuth, ”The ergodic decomposition of asymptotically mean stationary systems”, submitted to *Ergodic Theory and Dynamical Systems*, 2007.
19. A. Schönhuth, “On analytic properties of entropy rate of discrete random sources”, submitted to *IEEE Trans. Inf. Theory*, 2007.