# On Probabilistic Analog Automata

Asa Ben-Hur* Alexander Roitershtein† Hava T. Siegelmann‡

April 30, 2003; Revised February, 1, 2004

**Abstract**

We consider probabilistic automata on a general state space and study their computational power. The model is based on the concept of language recognition by probabilistic automata due to Rabin [20] and models of analog computation in a noisy environment suggested by Maass and Orponen [12], and Maass and Sontag [13]. Our main result is a generalization of Rabin's reduction theorem that implies that under very mild conditions, the computational power of such automata is limited to regular languages.

**Keywords**: probabilistic automata, probabilistic computation, noisy computational systems, regular languages, definite languages, Markov operators.

## 1 Introduction

Probabilistic automata have been studied since the early 60's [18]. Relevant to our line of interest is the work of Rabin [20] where probabilistic (finite) automata with isolated cut-points were introduced. Rabin showed that such automata recognize regular languages, and identified a condition which restricts them to definite languages, also known as "fading memory" languages. Recall that a definite language is one for which there exists an integer $r$ such that any two words coinciding on the last $r$ symbols are both or neither in the language. Paz generalized Rabin's condition for definite languages and called it *weak ergodicity*. He showed that Rabin's stability theorem holds for weakly ergodic systems as well [17, 18].

In recent years there has been much interest in analog automata and their computational properties. A model of analog computation in a noisy environment was introduced by Maass and Orponen in [12]. For a specific type of noise it recognizes only regular languages (see also [2]). Analog neural networks with Gaussian-like noise were shown by Maass and Sontag [13] to be limited in their language-recognition power to definite languages. This is in sharp

*Department of Biochemistry, B400 Beckman Center, Stanford University, CA 94305-5307, USA.

†Department of Mathematics, Technion - IIT, Haifa 32000, Israel (e-mail: roiterst@tx.technion.ac.il).

‡Department of Computer Science, University of Massachusetts at Amherst, 710 N. Pleasant Street Amherst, MA 01003-9305 USA.

contrast with the noise-free case where analog computational models are capable of simulating Turing machines, and when containing real constants, can recognize non-recursive languages [22, 23]. It is also important to note the difference between probabilistic automata and randomized Turing machines; the latter formulate the concept of probabilistic or randomized computation. A randomized Turing machine updates its state in a precise, noise-free manner, and has also access to a stream of random bits. Also note that when the underlying probability distribution has a non-recursive component, such a machine can recognize non-recursive languages (see [23] for the case of analog machines).

In this work we unravel the mechanisms that restrict the computational power of probabilistic automata. We propose a model which includes the discrete model of Rabin and the analog models suggested in [12, 13], and find general conditions related to the ergodic properties of the stochastic kernels representing the probabilistic transitions of the automaton that restrict its computational power to regular and definite languages. The results concerning definite languages first appeared (without proofs) in the conference paper by the authors [24].

The probabilistic automata we consider are homogeneous in time, in that their transitions may depend on the input, but do not depend on time. We denote the state space of the automaton by $\Omega$ and the input alphabet by $\Sigma$. We assume that a $\sigma$-algebra $\mathcal{B}$ of subsets of $\Omega$ is given, thus $(\Omega, \mathcal{B})$ is a measurable space. In our general probabilistic model, the measurable space $(\Omega, \mathcal{B})$ as well as the alphabet $\Sigma$ can be *arbitrary*.

We denote by $\mathcal{P}$ the set of probability measures on $(\Omega, \mathcal{B})$ and refer to it as the *distribution space*. When dealing with systems containing inherent elements of uncertainty (e.g., noise) we abandon the study of individual trajectories in favor of an examination of the flow of state distributions. The discrete-time dynamics we consider is defined by operators acting in a space of measures, and are called *Markov operators*.

More precisely, let $\mathcal{E}$ be the Banach space of finite signed measures on $(\Omega, \mathcal{B})$ with the total variation norm[1]

$$\|\mu\|_1 := \sup_{A \in \mathcal{B}} \mu(A) - \inf_{A \in \mathcal{B}} \mu(A),$$

and let $\mathcal{L}$ be the space of bounded linear operators in $\mathcal{E}$ with the norm[2] $\|P\|_1 = \sup_{\|\mu\|_1 = 1} \|P\mu\|_1$.

**Definition 1.1.** *An operator $P \in \mathcal{L}$ is said to be a* Markov operator *if for any probability measure $\mu$, the image $P\mu$ is again a probability measure. A* Markov system *is a set of Markov operators $T = \{P_u : u \in \Sigma\}$.*

With any Markov system $T$, one can associate a probabilistic computational system as follows. At each computation step the system receives an input signal $u \in \Sigma$ and updates its state. If the probability distribution on the initial state is given by $\mu_0 \in \mathcal{P}$, then the distribution of states after $n+1$ computational steps on inputs $w = w_0, w_1, ..., w_n$, is defined by

$$P_w \mu_0 = P_{w_n} \cdot \ldots \cdot P_{w_1} P_{w_0} \mu_0.$$

[1] A signed measure $\mu$ is a function $\mu : \mathcal{B} \to \mathbb{R}$ such that $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ for any countable collection of disjoint sets $A_i \in \mathcal{B}$, $i = 1, 2, \ldots$. It is finite if $\sup_{A \in \mathcal{B}} |\mu(A)| < \infty$. For any $\mu \in \mathcal{E}$, the state space $\Omega$ can be written as the union of disjoint sets $\Omega^+$ and $\Omega^-$, such that $\|\mu\|_1 = \mu(\Omega^+) - \mu(\Omega^-)$ (the Hahn decomposition) (see e.g. [5] or [15]).

[2] Using the Hahn decomposition it can be seen that for any $P \in \mathcal{L}$, $\|P\|_1 = \sup_{\mu \in \mathcal{P}} \|P\mu\|_1$ (see e.g. [8]).

If the probability of moving from state $x \in \Omega$ to set $A \in \mathcal{B}$ upon receiving input $u \in \Sigma$ is given by a stochastic kernel[3] $P_u(x, A)$, then $P_u\mu(A) = \int_\Omega P_u(x, A)\mu(dx)$.

Let $\mathcal{A}$ and $\mathcal{R}$ be two subsets of $\mathcal{P}$ with the property of having a $\rho$-gap

$$\operatorname{dist}(\mathcal{A}, \mathcal{R}) = \inf_{\mu\in\mathcal{A},\nu\in\mathcal{R}} \|\mu - \nu\|_1 = \rho > 0 \tag{1}$$

A Markov computational system becomes a language recognition device by agreement that an input string is accepted or rejected according to whether the distribution of states after reading the string is in $\mathcal{A}$ or in $\mathcal{R}$.

Finally, we have the definition:

**Definition 1.2.** *Let $\mu_0$ be an initial distribution and let $\mathcal{A}$ and $\mathcal{R}$ be two bounded subsets of $\mathcal{E}$ that satisfy* (1)*. Let $T = \{P_u : u \in \Sigma\}$ be a set of Markov operators on $\mathcal{E}$. We say that the* Markov Computational System *(MCS) $\mathcal{M} = \langle \mathcal{E}, \mathcal{A}, \mathcal{R}, \Sigma, \mu_0, T\rangle$* recognizes *the subset $L \subseteq \Sigma^*$ if for all $w \in \Sigma^*$:*

$$w \in L \Leftrightarrow P_w\mu_0 \in \mathcal{A}$$

$$w \notin L \Leftrightarrow P_w\mu_0 \in \mathcal{R}.$$

In the following we outline the main results of this paper. As usual, the set of all words of length $r$ is denoted by $\Sigma^r$ and $\Sigma^* := \cup_{r\in\mathbb{N}}\Sigma^r$. We recall that two words $u, v \in \Sigma^*$ are equivalent with respect to $L$ if and only if $uw \in L \Leftrightarrow vw \in L$ for all $w \in \Sigma^*$. A language $L \subseteq \Sigma^*$ is *regular* if there are finitely many equivalence classes. $L$ is *definite* if for some $r > 0$, $wu \in L \Leftrightarrow u \in L$ for all $w \in \Sigma^*$ and $u \in \Sigma^r$. If $\Sigma$ is finite, then definite languages are regular (see e.g. [20, 21]).

A quasi-compact MCS can be characterized as a system such that $\Sigma$ is finite and there is a set of compact operators[4] $\{Q_w \in \mathcal{L} : w \in \Sigma^*\}$ such that $\lim_{|w|\to\infty} \|P_w - Q_w\|_1 = 0$. Section 2 is devoted to MCSs having this property. Our main result (Theorem 2.4) states that quasi-compact MCSs can recognize regular languages only. The condition of quasi-compactness holds under very weak assumptions on the stochastic kernels $P_u$, and in particular we have:

**Theorem A.** *Let $\mathcal{M}$ be an MCS such that $\mathcal{B}$ is countably generated*[5] *and the alphabet $\Sigma$ is finite. Assume that there exist constant $K > 0$ and probability measure $\mu$ such that $P_u(x, A) \le K\mu(A)$ for all $u \in \Sigma$, $x \in \Omega$, $A \in \mathcal{B}$. Then, if a language $L \subseteq \Sigma^*$ is recognized by $\mathcal{M}$, it is a regular language.*

[3] A stochastic kernel on $(\Omega, \mathcal{B})$ is a function $P(x, A) : \Omega \times \mathcal{B} \to \mathbb{R}$, which is measurable on $x$ for each $A \in \mathcal{B}$, and such that $P(x, \cdot)$ is a probability measure for any $x \in \Omega$.

[4] An operator $Q \in \mathcal{L}$ is compact if it maps bounded subsets of $\mathcal{E}$ into compact ones. If $P$ is a bounded operator and $Q$ is compact, then $PQ$ and $QP$ are compact operators. If $\{Q_n\}_{n\ge0}$ are compact operators (e.g. have finite-dimensional ranges) and $\lim_{n\to\infty} \|Q_n - Q\|_1 = 0$ for some $Q \in \mathcal{L}$, then $Q$ is a compact operator [5].

[5] That is, $\mathcal{B}$ is generated by a countable collection of sets [16, p. 5]. The assumption is rather standard in the theory of Markov chains (see e.g. [14, p. 516] and [16, pp. 5–6]) and holds in arguably all practically interesting cases. This is the case if, for example, $\Omega$ is a Borel subset of a Polish space (separable topological space that is metrizable by a complete metric) and $\mathcal{B}$ is its Borel $\sigma$-field. The examples of Polish spaces include : countable discrete sets, $\mathbb{R}^n$, $[0, 1]^n$, $\mathbb{R}^{\mathbb{N}}$, $[0, 1]^{\mathbb{N}}$, and all compact metric spaces (e.g. [10]).

We consider another condition on the set of operators: an MCS is weakly ergodic if there is a set of probability measures $\{\nu_w \in \mathcal{P} : w \in \Sigma^*\}$ such that $\lim_{|w|\to\infty} \sup_{\mu\in\mathcal{P}} \|P_w\mu - \nu_w\|_1 = 0$. In Section 3 we carry over the theory of discrete weakly ergodic systems developed by Paz [17, 18] to our general setup. In particular, if a language $L$ is recognized by a weakly ergodic MCS, then it is a definite language. In the discrete case such automata were introduced by Rabin in [20] and in the context of analog computation were first considered in [13]. In Section 3 we discuss the connection between quasi-compact and weakly ergodic systems. We find that for a finite alphabet weak ergodicity implies quasi-compactness. This is consistent with the fact that in this case definite languages are a subclass of regular languages.

As mentioned, the model of language recognition with a gap between accepting and rejecting spaces agrees with Rabin's model of probabilistic automata with isolated cut-points [20] and the model of analog computation in a noisy environment [12, 13].

**Example 1.3.** *In [12, 13], it is assumed that $P_u(x, A) = Q\left(f(x, u), A\right)$, where the function $f : \Omega \times \Sigma \to \Omega$ is responsible for the noise-free dynamics of the computational system, and $Q(x, A)$ is a stochastic kernel representing the noise. This is interpreted as follows: upon receiving an input $u \in \Sigma$ the system jumps to $f(x, u)$ and then is dispersed by the noise into a set $A \in \mathcal{B}$ with probability $Q\left(f(x, u), A\right)$.*

We refer to [12] for a long list of examples of analog computational models, including recurrent analog neural nets (see also in [13]) and stochastic spiking neurons, where the noise effects can be modelled by stochastic kernels of the form $P_u(x, A) = Q\left(f(x, u), A\right)$. Since we do not assume that $\Omega$ is a subset of a finite-dimensional Euclidean space, our model also includes: neural networks with an unbounded number of components, networks of variable dimension (e.g., "recruiting networks"), stochastic cellular automata and stochastic coupled map lattices [24].

The stochastic kernels $P_u$ we consider are arbitrary. Thus our model addresses both "noisy computational systems" where the stochastic dynamics is a result of noise that was added to an underlying deterministic rule, and computational systems which have no underlying deterministic rule, but rather update probabilistically. The formulation of the additive noise model chosen in [12, 13], where $P_u(x, A) = Q(f(x, u), A)$, is one example of a noisy system; one can consider a more general form of additive noise that depends on the state and the input as $P_u(x, A) = Q_{x,u}\left(f(x, u), A\right)$. The abstract formulation of our results makes them directly applicable to all such systems. Moreover, it allows us to clarify the underlying mechanism leading to the restriction of the computational power of probabilistic or noisy systems, helps us reveal connections between discrete and analog systems, and relates our results to the work of Rabin [20] on probabilistic finite automata on the one hand, and to the classical theory of Markov chains in general state spaces on the other hand. Our main results in Section 2 (and in particular, Theorem A) are expressed for systems with a finite alphabet $\Sigma$, and in this case significantly improve Theorem 3.1 of Maass and Orponen [12] (see Example 2.12).

## 2 The Reduction Lemma and Quasi-compact MCSs

We prove here a general version of Rabin's reduction theorem (Lemma 2.1) which makes the connection between a measure of non-compactness of the set $\{P_w\mu_0 : w \in \Sigma^*\}$ with the computational power of MCSs. Then we introduce the notion of a quasi-compact MCS and show that these systems satisfy the conditions stated in Lemma 2.1.

If $S$ is a bounded subset of a Banach space $E$, Kuratowski's measure of non-compactness $\alpha(S)$ of $S$ is defined as follows [1]:

$$\begin{aligned}\alpha(S) &= \inf\{\varepsilon > 0 : S \text{ can be covered by a finite number of sets} \\ &\qquad \text{of diameter smaller than } \varepsilon\}. \end{aligned} \tag{2}$$

A bounded set $S$ is totally bounded if $\alpha(S) = 0$.

**Lemma 2.1.** *Let $\mathcal{M}$ be an MCS, and assume that $\alpha(\mathcal{O}) < \rho$, where $\mathcal{O} = \{P_w\mu_0 : w \in \Sigma^*\}$ is the set of all possible state distributions of $\mathcal{M}$, and $\rho$ is defined by* (1)*. Then, if a language $L \subseteq \Sigma^*$ is recognized by $\mathcal{M}$, it is a regular language.*

*Proof.* If $\|P_u\mu_0 - P_v\mu_0\|_1 < \rho$, then $u$ and $v$ are in the same equivalence class with respect to $L$. Indeed, using the contraction property of Markov operators[6], we obtain for any $w \in \Sigma^*$,

$$\|P_{uw}\mu_0 - P_{vw}\mu_0\|_1 = \|P_w\left(P_u\mu_0 - P_v\mu_0\right)\|_1 \leq \|P_u\mu_0 - P_v\mu_0\|_1 < \rho.$$

There are at most a finite number of equivalence classes, since there is a finite covering of $\mathcal{O}$ by sets with diameter less than $\rho$. □

Lemma 2.1 is a natural generalization of Rabin's reduction theorem [20], where the state space $\Omega$ is finite, and hence the whole space of probability measures $\mathcal{P}$ is compact.

Since $\mathcal{O} \subset \cup_{w\in\Sigma^r} P_w\mathcal{P}$ for any $r \in \mathbb{N}$, it follows from the lemma that if $\Sigma$ is finite and all $P_w,\ w \in \Sigma^r$, are compact operators for some $r \in \mathbb{N}$, $\mathcal{M}$ recognizes regular languages only.

**Example 2.2.** *Let $\Sigma$ be a finite alphabet. If $L \subseteq \Sigma^*$ is recognized by any one of the following systems, it is a regular language.*

*(i) Let $\mathcal{M}$ be an MCS such that $\Omega = \mathbb{Z}^n$ and for each $u \in \Sigma$ the sums $\Sigma_{|\mathbf{j}|<m}P_u(\mathbf{i},\mathbf{j})$ converge uniformly in $\mathbf{i}$ when $m$ goes to infinity. Then, the operators $P_u, u \in \Sigma$ are compact [3][7].*

*(ii) Let $\mathcal{M}$ be an MCS such that $\Omega$ is a compact metric space and $\mathcal{B}$ is its Borel $\sigma$-field. If the functions $P_u(\cdot, A)$ are continuous for every $u \in \Sigma$ and $A \in \mathcal{B}$, then $P_w$ are compact for all $w \in \Sigma^2$ (see Theorem 3.1.28 and the end of the proof of Theorem 3.1.31 in [7]).*

*(iii) Let $\mathcal{M}$ be an MCS such that $P_u(x, A) = \int_A p_u(x, y)\mu(dy)$ for some $\mu \in \mathcal{P}$ and functions $p_u(x, y) : \Omega^2 \to \mathbb{R}$, $u \in \Sigma$, are bounded and measurable in $x$ and $y$. Then, $P_w$ is compact for any $w \in \Sigma^2$ (see Lemma A.1 in the appendix).*

[6] Any Markov operator $P$ has unit norm and hence is a contraction: $\|P\|_1 = 1$ and $\|P\nu\|_1 \leq \|\nu\|_1$ for any $\nu \in \mathcal{E}$ [15].

[7] Each $P_u,\ u \in \Sigma$ is compact as the limit $P_u = \lim_{m\to\infty} P_{u,m}$ (in the $\|\cdot\|_1$ norm) of the finite-dimensional projections defined by $P_{u,m}(\mathbf{i},\mathbf{j}) = P_u(\mathbf{i},\mathbf{j})$ if $|\mathbf{j}| < m$ and $P_u(\mathbf{i},\mathbf{j}) = 0$ otherwise.

Recall that a Markov operator $P$ is called quasi-compact if there is a compact operator $Q \in \mathcal{L}$ such that $\|P - Q\|_1 < 1$ [15].

**Definition 2.3.** *An MCS $\mathcal{M}$ is called quasi-compact if the alphabet $\Sigma$ is finite, and there exist constants $r, \delta > 0$ such that for any $w \in \Sigma^r$ there is a compact operator $Q_w$ which satisfies $\|P_w - Q_w\|_1 \leq 1 - \delta$.*

If an MCS $\mathcal{M}$ is quasi-compact, then there exist a constant $M > 0$ and a collection of compact operators $\{Q_w : w \in \Sigma^*\}$ such that $\|P_w - Q_w\|_1 \leq M(1-\delta)^{|w|/r}$, for all $w \in \Sigma^*$.

The next theorem characterizes the computational power of quasi-compact MCSs.

**Theorem 2.4.** *If $\mathcal{M}$ is a quasi-compact MCS, and a language $L \subseteq \Sigma^*$ is recognized by $\mathcal{M}$, then $L$ is a regular language.*

*Proof.* Fix any $\varepsilon > 0$. There exist a number $n \in \mathbb{N}$ and compact operators $Q_w,\ w \in \Sigma^n$ such that $\|P_w - Q_w\|_1 \leq \varepsilon$ for all $w \in \Sigma^n$. For any words $v \in \Sigma^*$ and $w \in \Sigma^n$, we have $\|P_{vw}\mu_0 - Q_w(P_v\mu_0)\|_1 \leq \|P_w - Q_w\|_1 \leq \varepsilon$. Since $Q_w(P_v\mu_0)$ is an element of the totally bounded set $Q_w(\mathcal{P})$, then the last inequality implies that the set $\mathcal{O} = \{P_u\mu_0 : u \in \Sigma^*\}$ can be covered by a finite number of balls of radius arbitrarily close to $\varepsilon$. □

Doeblin's condition which follows, is a criterion for quasi-compactness (it should not be confused with its stronger version, defined in Section 3, which was used in [13]).

**Definition 2.5.** *Let $P(x, A)$ be a stochastic kernel defined on $(\Omega, \mathcal{B})$. We say that it satisfies Condition D if there exist positive constants $\theta < 1,\ \eta < 1$, and a probability measure $\mu$ on $(\Omega, \mathcal{B})$ such that*

$$\mu(A) \geq \theta \Rightarrow P(x, A) \geq \eta \text{ for all } x \in \Omega, A \in \mathcal{B}. \tag{3}$$

For a set $A \in \mathcal{B}$ let $A^c$ be its complement in $\Omega$. Since $\mu(A^c) = 1 - \mu(A)$ and $P(x, A^c) = 1 - P(x, A)$, we have the following equivalent formulation of Condition D:

$$\mu(A) \leq 1 - \theta \Rightarrow P(x, A) \leq 1 - \eta \text{ for all } x \in \Omega. \tag{4}$$

**Example 2.6.** *[4] Condition D is satisfied if one of the following conditions holds:*

*(i) $P(x, A) \leq K\mu(A)$ for some $K > 0$ and $\mu \in \mathcal{P}$. Indeed, in this case (4) holds with $1 - \theta = 1/(1 + K)$ and $1 - \eta = K/(1 + K)$.*

*(ii) $\Omega = \mathbb{R}^n$ and $\int_{|y|<m} P(x, dy)$ converges to 1 uniformly in $x$ when $m$ goes to infinity.*

*(iii) $P(x, A) \geq c\mu(A)$ for some $c > 0$ and $\mu \in \mathcal{P}$. MCSs defined by means of such stochastic kernels are considered in [13] and in Section 3 of this paper.*

Recall the definition of a countably generated $\sigma$-field from footnote 5.

**Theorem 2.7.** *Let $\mathcal{M}$ be an MCS such that $\mathcal{B}$ is countably generated and $\Sigma$ is finite. If for some $r \in \mathbb{N}$, all stochastic kernels $P_w(x, A),\ w \in \Sigma^r$, satisfy Condition D, then $\mathcal{M}$ is quasi-compact.*

This theorem together with part *(i)* of Example 2.6 yield Theorem A announced in the introduction. The proof, given in Appendix A, follows the proof in [25] that Condition D implies quasi-compactness for an *individual* Markov operator.

**Example 2.8.**

*(i) If* $\Omega = \{1, 2, \ldots, n\}$ *for some* $n \in \mathbb{N}$, *any Markov operator is compact. Moreover,* (3) *trivially holds with* $\theta = (n - 0.5)/n$, *any* $\eta \in (0, 1)$, *and the uniform probability measure* $\mu$. *Note that the stochastic matrix representing* $P$ *can be arbitrarily sparse. It is shown in [20] that finite probabilistic automata with isolated cut-points (MCSs in a finite space* $\Omega$*) can recognize any regular language (see also [12, Theorem 4.1]).*

*(ii) For some* $n \in \mathbb{N}$, *let* $\Omega = \cup_{i=1}^{n} \Omega_i$ *be a partition of the state space* $\Omega$ *into* $n$ *disjoint subsets* $\Omega_i \in \mathcal{B}$, $i = 1, 2, \ldots, n$. *Assume that for any* $i \le n$ *and for all* $x \in \Omega_i$, $P(x, A) \ge \gamma_i \mu_i(A)$ *for some* $\gamma_i \in (0, 1)$ *and a probability measure* $\mu_i$ *concentrated on* $\Omega_i$. *The linear operator* $Q \in \mathcal{L}$, *associated with the (not stochastic) kernel* $Q(x, dy) = \frac{1}{n} \sum_{i=1}^{n} \gamma_i \mu_i(dy)$ *and defined by* $Q\nu(A) = \int_\Omega Q(x, A)\nu(dx)$, $\nu \in \mathcal{E}$, *has a finite dimensional range*[8] *and hence is compact. Letting* $\gamma = \min_{1 \le 1 \le n} \gamma_i$, *we have* $\|P - Q\|_1 = 1 - \frac{1}{n} \sum_{i=1}^{n} \gamma_i \le 1 - \gamma < 1$.

*A particular case of this example is:* $\Omega = [0, 1)$, $\Omega_i = \left[\frac{i-1}{n}, \frac{i}{n}\right)$, *and for* $x \in \Omega_i$, $P(x, A) = n\lambda(A \cap \Omega_i)$, *where* $\lambda$ *is the Lebesgue measure.*

*(iii) Consider a model of* $N$ *MCSs that update asynchronously. Let* $\{\mathcal{M}_i\}_{i=1}^{N}$ *be a set of MCSs that differ only by their Markov systems. At each computational step one MCS is activated and the current state of the aggregate is represented by the state of its active component. The active component is chosen at random: the system* $\mathcal{M}_i$ *is chosen with probability* $\varepsilon_i$. *The aggregate system is then described by the stochastic kernels* $P_u(x, A) = \sum_{i=1}^{N} \varepsilon_i P_u^i(x, A)$. *It is straightforward to verify that the resulting MCS is quasi-compact if at least one set of operators* $\{P_u^1 : u \in \Sigma\}, \ldots, \{P_u^N : u \in \Sigma\}$ *is quasi-compact.*

The following lemma, whose proof is deferred to Appendix B, gives a complete characterization of quasi-compact MCSs in terms of its associated Markov operators.

**Lemma 2.9.** *If an MCS* $\mathcal{M}$ *is quasi-compact, then* $\alpha(T^*) = 0$, *where* $T^* = \{P_w : w \in \Sigma^*\}$.

It is easy to see that $\alpha(\mathcal{O}) < \sup_{u \in \Sigma} \alpha(P_u \mathcal{P}) + \alpha(T)$, where $T = \{P_u : u \in \Sigma\}$. This observation leads to the following extension of Theorem 2.4 to infinite alphabets, whose proof is included in Appendix C.

**Theorem 2.10.** *Let* $\mathcal{M}$ *be an MCS such that* $\alpha(T) = 0$. *Assume that there exist constants* $r, \delta > 0$ *such that for any* $w \in \Sigma^r$ *there is a compact operator* $Q_w$ *which satisfies* $\|P_w - Q_w\|_1 \le 1 - \delta$. *Then, if a language* $L \subseteq \Sigma^*$ *is recognized by* $\mathcal{M}$, *it is a regular language.*

The condition $\alpha(T) = 0$ holds if $\Sigma$ is a compact set and the map $P(u) = P_u : \Sigma \to \mathcal{L}$ is continuous. Consider the following example.

[8] For any $\nu \in \mathcal{E}$, $Q\nu$ can be represented as a linear combination $\sum_{i=1}^{n} f_i \mu_i$, where $f_i$ are real numbers that depend on $\nu$.

**Example 2.11.** *Let $\mathcal{M}$ be an MCS such that $\Sigma = [0,1]^m$ and $P_u(x,A) = \int_A p_u(x,y)\mu(dy)$ for a probability measure $\mu$ and a set of jointly measurable functions $p_u(x,y)$, $u \in \Sigma$, uniformly bounded by a constant $K > 0 : p_u(x,y) < K$ for all $x, y \in \Omega, u \in \Sigma$. Furthermore, assume that the family of functions $\{\mathfrak{p}_{x,y} : x, y \in \Omega\}$, where $\mathfrak{p}_{x,y}(u) = p_u(x,y) : \Sigma \to \mathbb{R}$, is equicontinuous*[9]*. By part (iii) of Example 2.2 and Theorem 2.10, if a language $L \subseteq \Sigma^*$ is recognized by $\mathcal{M}$, it is a regular language.*

We conclude this section with comparison of our result to Theorem 3.1 of Maass and Orponen [12]. In the framework of Example 1.3 they assumed that $\Omega$ is a bounded subset of $\mathbb{R}^n$ and the noise has a bounded and *piecewise uniformly continuous*[10] density $q(x,y)$ with respect to a probability measure $\mu : Q(x,A) = \int_A q(x,y)\mu(dy)$. They showed that such systems are restricted in their computational power to regular languages. If $\Sigma$ is a finite set, this result is a very particular case of Theorem A. If the alphabet is not finite, Theorem 3.1 in [12] can be modified to fit our general setup as follows. Assume that $P_u(x,A) = \int_A p_u(x,y)\mu(dy)$ for some $\mu \in \mathcal{P}$ and jointly measurable functions $p_u(x,y)$, $u \in \Sigma$ such that $p_u(x,y) < K$ for all $x, y \in \Omega$, $u \in \Sigma$, and some $K > 0$. Then, for any initial distribution $\mu_0$, there exists a family of measurable functions $\Pi = \{\pi_w : w \in \Sigma^*\}$ such that $P_w\mu_0(A) = \int_A \pi_w(y)\mu(dy)$ for all $w \in \Sigma^*$.

**Example 2.12. (A modification of [12, Theorem 3.1])** *Assume that $\Omega$ is a totally bounded metric space and, letting $\mathfrak{p}_{x,u}(y) = p_u(x,y) : \Omega \to \mathbb{R}$, $x, y \in \Omega$, $w \in \Sigma^*$, that the family $\{\mathfrak{p}_{x,u}(\cdot) : x \in \Omega, u \in \Sigma\}$ is equicontinuous (or, more generally, piecewise uniformly continuous). Then, there exist continuous densities $\pi_w(y)$ and the conditions of Lemma 2.1 can be verified by using the Ascoli-Arzela theorem [1, 5] (see also a related [5, Theorem IV.8.21]).*

Interestingly, if $\Sigma$ is finite and $\Omega$ is compact, Example 2.2 *(ii)* yields a "dual" to this one: the system is quasi-compact if the family of functions $\mathfrak{p}_{u,y}(\cdot) = p_u(\cdot, y) : \Omega \to \mathbb{R}$, $y \in \Omega, u \in \Sigma$, is equicontinuous.

# 3 Weakly Ergodic MCSs

This section is devoted to MCS with "fading memory". We adopt here the terminology introduced by Paz [17, 18] in the context of discrete automata and refer to such computational systems as weakly ergodic MCS. Following earlier works of Rabin [20], Paz [17], and Maass and Sontag [13], we show that the computational power of abstract weakly ergodic systems is limited to definite languages, and that the computational system is stable with respect to small perturbations.

[9] Let $\Omega$ be a metric space and denote its metric by $\|\cdot\|$. A family of functions $f_a(x) : \Omega \to \mathbb{R}$, $a \in A$, indexed by a set $A$, is equicontinuous if for any $\varepsilon > 0$ there exists $\delta > 0$ such that $\|x - y\| < \delta$ implies $|f_a(x) - f_a(y)| < \varepsilon$ for every $a \in A$.

[10] That is, there is a finite partition of $\Omega$ into disjoint sets $\Omega_1, \Omega_2, \ldots, \Omega_n$ such that the functions $\mathfrak{q}_{x,i}(\cdot) = q(x,\cdot) : \Omega_i \to \mathbb{R}$, $x \in \Omega, i = 1, 2, \ldots, n$, are continuous and the family $\mathcal{Q}_i = \{\mathfrak{q}_{x,i} : x \in \Omega\}$ is equicontinuous for each $i = 1, 2, \ldots, n$.

For any Markov operator $P$ define Dobrushin's coefficient

$$\delta(P) := \sup_{\mu,\nu\in\mathcal{P}} \frac{1}{2}\|P\mu - P\nu\|_1 = \sup_{x,y}\sup_{A\in\mathcal{B}} |P(x,A) - P(y,A)|. \tag{5}$$

Another characterization of $\delta(P)$ is [6, 8] :

$$\delta(P) = \sup_{\lambda\in\mathcal{N}\setminus\{0\}} \frac{\|P\lambda\|_1}{\|\lambda\|_1}, \tag{6}$$

where $\mathcal{N} = \{\lambda \in \mathcal{E} : \lambda(\Omega) = 0\}$.

**Definition 3.1.** *A Markov system* $\{P_u,\ u \in \Sigma\}$ *is called* weakly ergodic *if there exist constants* $r, \delta > 0$ *such that* $\delta(P_w) \le 1 - \delta$ *for any* $w \in \Sigma^r$. *An MCS* $\mathcal{M}$ *is called* weakly ergodic *if its associated Markov system* $\{P_u,\ u \in \Sigma\}$ *is weakly ergodic.*

It follows from the definition and (6) that $\delta(P_w) \le M(1-\delta)^{|w|/r}$, for any $w \in \Sigma^*$ and some $M > 0$. Let $\nu_0$ be any probability measure and $H_w \in \mathcal{L},\ w \in \Sigma^*$ one-dimensional (and hence compact) operators defined by $H_w\mu = P_w\nu_0$ for every $\mu \in \mathcal{P}$ [11]. Then (for the second equality see footnote 2),

$$\|P_w - H_w\|_1 = \sup_{\|\mu\|_1=1} \|P_w\mu - H_w\mu\|_1 = \sup_{\|\mu\|_1\in\mathcal{P}} \|P_w(\mu - \nu_0)\|_1 \le 2M(1-\delta)^{|w|/r}.$$

It follows that if $\Sigma$ is finite, every weakly ergodic MCS is quasi-compact. Moreover, let $n \in \mathbb{N}$ be such a large number that $\sup_{x\in\Omega,A\in\mathcal{B}} |P_w(x,A) - P_w\nu_0(A)| \le \|P_w - H_w\|_1 \le 0.1$ for every $w \in \Sigma^n$. Then, $P_w\nu_0(A) \ge 0.2$ implies that $P_w(x,A) \ge 0.1$ for all $x \in \Omega$, and hence the stochastic kernel $P_w(x,A)$ satisfies Condition D.

Maass and Sontag used a strong Doeblin's condition (see Definition 3.4 below) to bound the computational power of noisy neural networks [13]. They essentially proved (see also [18, 20]) the following result:

**Theorem 3.2.** *Let* $\mathcal{M}$ *be a weakly ergodic MCS. If a language* $L$ *can be recognized by* $\mathcal{M}$, *then it is definite.*

**Example 3.3.** *[24] Consider the aggregate MCS introduced in part (iii) of Example 2.8. It is weakly ergodic if at least one set of operators* $\{P_u^1 : u \in \Sigma\}, \ldots, \{P_u^N : u \in \Sigma\}$ *is weakly ergodic.*

The ability of a computational system to recognize only definite languages can be interpreted as saying that the system forgets all its input signals, except for the most recent ones. This property is reminiscent of human short term memory. Definite languages were introduced by Kleene [11] and studied in detail by Rabin *et al* in [19, 20, 21]. If the alphabet is finite, all definite languages are regular, but this is not always the case for an infinite alphabet[12].

[11] It follows from the Hahn decomposition (see footnote 1) that a linear operator in $\mathcal{E}$ is completely defined by its actions in the subspace of the probability measures.

[12] Consider the following example: Let $\Sigma = \mathbb{N}$, and $L = \Sigma \cup \{w \in \Sigma^{\ge 2} : w_{|w|} = w_{|w|-1} + 1\}$. In this language, each word of length one must belong to a different equivalence class, and thus the language is not regular.

**Definition 3.4.** *Let $P(x, A)$ be a stochastic kernel defined on $(\Omega, \mathcal{B})$. We say that it satisfies Condition $D_0$ if there exist a constant $c \in (0, 1)$ and a probability measure $\mu$ on $(\Omega, \mathcal{B})$ such that*

$$P(x, A) \geq c\mu(A) \text{ for all } x \in \Omega, A \in \mathcal{B}.$$

If the stochastic kernel $P(x, A)$ corresponding to a Markov operator $P$ satisfies Condition $D_0$ with a constant c, then $\delta(P) \leq 1 - c$ [4]. The following example shows that this condition is not necessary.

**Example 3.5.** *Let $\Omega = \{1, 2, 3\}$ and $P(x, y) = \frac{1}{2}$ if $x \neq y$. Then $\delta(P) = \frac{1}{2}$, but $P$ does not satisfy condition $D_0$.*

We next state a general version of the Rabin-Paz stability theorem [18, 20], which shows that all weakly ergodic MCSs are stable with respect to small perturbations of the associated Markov system, i.e. are robust with respect to architectural imprecisions and environmental noise. We first define two MCSs, $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ to be *similar* if they share the same measurable space $(\Omega, \mathcal{B})$, alphabet $\Sigma$, and sets $\mathcal{A}$ and $\mathcal{R}$, and differ only in their Markov operators.

**Theorem 3.6.** *Let $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ be two similar MCSs such that the first is weakly ergodic. Then there is $\alpha > 0$, such that if $\|P_u - \tilde{P}_u\|_1 \leq \alpha$ for all $u \in \Sigma$, then the second is also weakly ergodic. Moreover, the two MCSs recognize the same language.*

For the sake of completeness we give a proof in Appendix D.

We conclude with an example of a weakly ergodic MCS where the one-step transition probabilities $P_u(x, A)$ are localized in an arbitrarily small neighborhood of $x$ (in contrast to the results of [13], where the kernels are required to have "wide support").

**Example 3.7.** *This is a modification of Example 2.8 (ii). Let $\mathcal{M}$ be an MCS such that $\Omega = [0, 1)$ and $\Sigma = \{0, 1\}$. Further, let $\Omega_i = \left[\frac{i-1}{n}, \frac{i}{n}\right)$ and for $x \in \Omega_i$, set (with the convention that $n + 1 = 1$ and $n + 2 = 2$)*

$$P_0(x, A) = \frac{n}{2}\lambda\big(A \cap (\Omega_i \cup \Omega_{i+1})\big) \quad \text{and} \quad P_1(x, A) = \frac{n}{3}\lambda\big(A \cap (\Omega_{i-1} \cup \Omega_i \cup \Omega_{i+1})\big),$$

*where $\lambda$ is the Lebesgue measure on $[0, 1]$. That is, all the transitions are into an interval of length at most $3/n$. Since at each computational step, the system may stay in the set $\Omega_i$ where it is now located or move to the set $\Omega_{i+1}$, both with probability at least $1/3$, $P_w(x, A) \geq \left(\frac{1}{3}\right)^{n-1} \cdot \frac{n}{3}\lambda(A)$ for any $w \in \Sigma^n$. Thus, Condition $D_0$ holds for any $P_w, w \in \Sigma^n$, and hence $\mathcal{M}$ is weakly ergodic.*

# Appendices

# A Proof of Theorem 2.7

For simplicity we assume that $r = 1$. The proof for the general case is similar, with the only difference that expansion (7) below should be used for $w \in \Sigma^{rm}$ rather than for $w \in \Sigma^m$.

**Lemma A.1.** *[25] Let $K(x,A)$ and $N(x,A)$ be two stochastic kernels defined by*

$$K(x,A) = \int_A k(x,y)\mu(dx), \quad |k(x,y)| \le C_K,$$

$$N(x,A) = \int_A n(x,y)\mu(dx), \quad |n(x,y)| \le C_N,$$

*where $k(x,y)$ and $n(x,y)$ are measurable and bounded functions in $\Omega \times \Omega$, and $C_K, C_N$ are constants. Then $NK \in \mathcal{L}$ is compact.*

The proof in [25] is for a special case, so we give here an alternative proof[13].

*Proof.* Let $\{n_m(x,y) : m \in \mathbb{N}\}$ be a set of simple[14] and measurable functions such that

$$\int_\Omega \int_\Omega |n_m(x,y) - n(x,y)|\mu(dx)\mu(dy) \le \frac{1}{m},$$

and define stochastic kernels $N_m(x,A) = \int_A n_m(x,y)\mu(dy)$. Without loss of generality [9, Lemma 2.10] we can assume that these functions are finite linear combinations

$$n_m(x,y) = \sum_{k=1}^{i_m} c_k \mathbf{1}_{\{B_{m,k}\times C_{m,k}\}}(x,y)$$

of indicator functions of sets of the form $B_{m,k} \times C_{m,k}$, where $B_{m,k}, C_{m,k} \in \mathcal{B}$. Since the corresponding operators $N_m \in \mathcal{L}$ have finite dimensional ranges they are compact. On the other hand

$$\|NK - N_m K\|_1 = \sup_{\|\varphi\|_1=1} \|NK\varphi - N_m K\varphi\|_1 \le C_K/m,$$

thus, $NK = \lim_{m\to\infty} N_m K$ is a compact operator. □

Since operators $P_u,\ u \in \Sigma$ satisfy Condition D, they can be represented as $P_u = Q_u + R_u$, where the $Q_u$ are defined by stochastic kernels having bounded and measurable on $\Omega \times \Omega$ densities $q_u(x,y)$ with respect to $\mu$, and $\|R_u\|_1 \le 1-\eta$ [25][15]. Consider the expansion of $P_w = \prod_{k=1}^m (Q_{w_k} + R_{w_k}),\ w \in \Sigma^m$ in $2^m$ terms:

$$P_w = \prod_{k=1}^m Q_{w_k} + \sum_{j=1}^m \left( \prod_{k=1}^{j-1} Q_{w_k} R_{w_j} \prod_{k=j+1}^m Q_{w_k} \right) + \ldots + \prod_{k=1}^m R_{w_k}. \tag{7}$$

By Lemma A.1, the terms containing $Q_{w_i}$ at least twice as factor are all compact operators in $\mathcal{L}$. Since there are at most $m+1$ terms where $Q_{w_i}$ appear at most once, then we obtain that for any $w \in \Sigma^m$ there is a compact operator $Q_w$ such that $\|P_w - Q_w\|_1 \le (m+1)\cdot(1-\eta)^{m-1}$.

[13]The lemma follows from Theorems IV.8.9 and VI.8.12 in [5], but we prefer to give here a simple self-contained proof.

[14]That is, functions which have only a finite set of values in $\Omega^2 \backslash B$, where $B \subset \Omega^2$ is a null-set of the measure $\mu \otimes \mu$. Simple and $\mu \otimes \mu$-measurable functions are dense in $L_1(\Omega^2, \mathcal{B}\otimes\mathcal{B}, \mu\otimes\mu)$ [5, p. 125].

[15]Here the assumption that $\mathcal{B}$ is countably generated is used to ensure (by [16, Proposition 1.1]) that there exists a *jointly measurable* density.

# B Proof of Lemma 2.9

We need the following proposition suggested to us by Leonid Gurvits.

**Proposition B.1.** *Let $Q_1, Q_2 \in \mathcal{L}$ be two compact operators, and let $H = \{P_j\} \subseteq \mathcal{L}$ be a bounded set of operators. Then, the set $Q = \{Q_2 P Q_1 : P \in H\}$ is totally bounded.*

*Proof.* Let $\mathcal{K} = \{\mu \in \mathcal{E} : \|\mu\|_1 \leq 1\}$ and $X_i \subseteq \mathcal{E} : i = 1, 2$ be two compact sets such that $Q_i \mathcal{K} \subseteq X_i$. Define a bounded family $\mathcal{F} = \{f_j\}$ of continuous linear functions from $X_1$ to $X_2$ by setting $f_j = Q_2 P_j$. Since $H$ is bounded, then $\mathcal{F} \subseteq C(X_1, X_2)$ is bounded and equicontinuous, that is by the Ascoli-Arzela theorem it is conditionally compact. Fix any $\varepsilon > 0$ and consider a finite covering of $\mathcal{F}$ by balls with radii $\varepsilon$. If $f_i$ and $f_j$ are included in the same ball, then

$$\|Q_2 P_i Q_1 - Q_2 P_j Q_1\|_1 \leq \sup_{x \in X_1} \|f_i(x) - f_j(x)\|_1 \leq 2\varepsilon.$$

Therefore $\alpha(Q) \leq 2\varepsilon$. This completes the proof since $\varepsilon$ is arbitrary. □

From Proposition B.1 it follows that the set $\{Q_u P Q_v : u, v \in \Sigma^n, P \in \mathcal{L}, \|P\|_1 = 1\}$ is totally bounded.

Fix any $\varepsilon > 0$. There exist a number $n \in \mathbb{N}$ and compact operators $Q_w,\ w \in \Sigma^n$ such that $\|P_w - Q_w\|_1 \leq \varepsilon$ for all $w \in \Sigma^n$. Since any word $w \in \Sigma^{\geq 2n+1}$ can be represented in the form $w = u\hat{w}v$, where $u, v \in \Sigma^n$, and

$$\begin{aligned} \|P_w - Q_v P_{\hat{w}} Q_u\|_1 &= \|P_v P_{\hat{w}} P_u - Q_v P_{\hat{w}} Q_u\|_1 \leq \\ &\leq \|P_v P_{\hat{w}} P_u - P_v P_{\hat{w}} Q_u\|_1 + \|P_v P_{\hat{w}} Q_u - Q_v P_{\hat{w}} Q_u\|_1 \leq \\ &\leq \|P_u - Q_u\|_1 + \|P_v - Q_v\|_1 \leq 2\varepsilon, \end{aligned}$$

we can conclude that $\alpha(T^{\geq 2n+1}) \leq 2\varepsilon$, where $T^{\geq 2n+1} = \{P_w : w \in \Sigma^{\geq 2n+1}\}$. It follows that $\alpha(T^*) = \alpha(T^{\geq 2n+1}) \leq 2\varepsilon$, completing the proof since $\varepsilon > 0$ is arbitrary.

# C Proof of Theorem 2.10

The proof is by adaptation of some standard arguments for powers of individual quasi-compact operators (see Section 5.3 in [15]).

First, letting $T^n = \{P_w : w \in \Sigma^n\}$, we observe that $\alpha(T^n) = 0$ for any $n \in \mathbb{N}$. Indeed, the triangular inequality and the contraction property of Markov operators imply, by induction on $n$, that for any $v, w \in \Sigma^n$

$$\|P_v - P_w\|_1 \leq \sum_{i=1}^{n} \|P_{v_i} - P_{w_i}\|_1. \tag{8}$$

Roughly, this inequality implies that any finite covering of $T$ by sets of diameters less that $\delta$ yields a finite covering of $T^n$ by sets with diameters less that $n\delta$. More precisely, fix any

$\varepsilon > 0$ and let $\{A_j\}_{j=1}^m$ be $m$ disjoint subsets of $T$ with diameter less than $\varepsilon/n$, whose union is $T$. Such a finite covering exists, since $T$ is totally bounded. Let

$$B_{j_1,j_2,\ldots,j_n} = \{P_w \in \Sigma^n : P_{w_k} \in A_{j_k},\ k = 1, 2, \ldots, n\}, \qquad j_k = 1, 2, \ldots, m.$$

Then, by (8), $m^n$ sets $B_{j_1,j_2,\ldots,j_n}, j_k = 1, 2, \ldots, m;\ k = 1, 2, \ldots, n$, constitute a finite covering of $T^n$ by sets with diameters less than $\varepsilon$. Since $\varepsilon$ is arbitrary, it follows that $\alpha(T^n) = 0$.

Next, we will prove that for any $\varepsilon > 0$ there exist $n_\varepsilon \in \mathbb{N}$ and compact operators $Q_w,\ w \in \Sigma^{n_\varepsilon}$, such that $\|P_w - Q_w\|_1 < \varepsilon$ for every $w \in \Sigma^{n_\varepsilon}$. In particular $P_w\mathcal{P}$ can be covered by a finite number of balls of radius $\varepsilon$, and hence $\alpha(P_w\mathcal{P}) \leq 2\varepsilon$ for every $w \in \Sigma^{n_\varepsilon}$.

Let $P_u$ and $P_v$ be any two operators in $T^r$ and define $Q_{uv} = Q_vP_u + P_uQ_u - Q_vQ_u$. Since $P_u$ and $P_v$ are bounded, $Q_{uv}$ is a compact operator. Moreover,

$$\|P_{uv} - Q_{uv}\|_1 \ = \ \|(P_v - Q_v)(P_u - Q_u)\|_1 \leq \|P_v - Q_v\|_1 \cdot \|P_u - Q_u\|_1 \leq (1-\delta)^2.$$

Using the induction, we conclude that for any $w \in \Sigma^{mr},\ m \in \mathbb{N}$, there exists a compact operator $Q_w$ such that $\|P_w - Q_w\|_1 \leq (1-\delta)^m$. For $m$ large enough, $(1-\delta)^m$ will be less than $\varepsilon$.

Fix now any $\varepsilon > 0$. We are in the position to build, using a finite covering of $T^{n_\varepsilon}$ by sets with diameters at most $\varepsilon$, a finite covering of $\cup_{w\in\Sigma^{n_\varepsilon}} P_w\mathcal{P} = \cup_{w\in\Sigma^*} P_w\mathcal{P}$ by sets with diameters at most $4\varepsilon$. Clearly, this will complete the proof because $\varepsilon$ is arbitrary.

Let $\{C_j\}_{j=1}^n$ be $n$ disjoint subsets of $T^{n_\varepsilon}$ with diameter at most $\varepsilon$, whose union is $T^{n_\varepsilon}$. Suppose that $P_v \in T^{n_\varepsilon}$ and $P_w \in T^{n_\varepsilon}$ are included in the same set, say $C_1$. Consider a finite covering of $P_w\mathcal{P}$ by $\mathcal{E}$-balls of radius $\varepsilon$. Since $\|P_v - P_w\|_1 \leq \varepsilon$, the set $P_v\mathcal{P}$ can be covered by the balls with the same centers, but of radius $2\varepsilon$.

Since $v$ is arbitrary, we conclude that $\alpha\left(\cup_{P_v\in C_1} P_v\mathcal{P}\right) < 4\varepsilon$. Therefore, since $\{C_j\}_{j=1}^n$ is a finite covering of $T^{n_\varepsilon}$, $\alpha\left(\cup_{P_v\in T^{n_\varepsilon}} P_v\mathcal{P}\right) < 4\varepsilon$, completing the proof.

# D Proof of Theorem 3.6

This result is implied by the following lemma:

**Lemma D.1.** *Let $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ be two similar MCSs, such that the first is weakly ergodic and the second is arbitrary. Then, for any $\beta > 0$ there exists $\varepsilon > 0$ such that $\|P_u - \tilde{P}_u\|_1 \leq \varepsilon$ for all $u \in \Sigma$ implies $\|P_w - \tilde{P}_w\|_1 \leq \beta$ for all words $w \in \Sigma^*$.*

*Proof.* It is easy to verify by using the representation (6) that:

(i) For any Markov operators $P, Q$, and $R$, we have $\|PQ - PR\|_1 \leq \delta(P)\|Q - R\|_1$.

(ii) For any Markov operators $P, \tilde{P}$ we have $\delta(\tilde{P}) \leq \delta(P) + \|P - \tilde{P}\|_1$.

Let $r \in \mathbb{N}$ be such that $\delta(P_w) \leq \beta/7$ for any $w \in \Sigma^r$, and let $\varepsilon = \beta/r$. If $\|P_u - \tilde{P}_u\|_1 \leq \varepsilon$ for any $u \in \Sigma$, then $\|P_w - \tilde{P}_w\|_1 \leq n\varepsilon$ for any $w \in \Sigma^n$. It follows that $\|P_w - \tilde{P}_w\|_1 \leq \beta$ for any $w \in \Sigma^{\leq r}$. Moreover, for any $v \in \Sigma^r$ and $w \in \Sigma^*$, we have

$$\begin{aligned}\|P_{vw} - \tilde{P}_{vw}\|_1 &\leq \|P_{vw} - P_v\|_1 + \|P_v - \tilde{P}_v\|_1 + \|\tilde{P}_v - \tilde{P}_{vw}\|_1 \leq \\ &\leq 2\delta(P_v) + \|P_v - \tilde{P}_v\|_1 + 2\delta(\tilde{P}_v) \leq 4\delta(P_v) + 3\|P_v - \tilde{P}_v\|_1 \leq \beta,\end{aligned}$$

completing the proof. □

# Acknowledgments

We are grateful to Leonid Gurvits for valuable discussions. We also thank the referee for his very helpful comments.


# References


[1] J. Banaś and K. Goebel, *Measures of Noncompactness in Banach Spaces*, Marcel Dekker, New York, 1980.

[2] M. Casey, *The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction*, Neural Computation **8** (1996), 1135–1178.

[3] L. W. Cohen and N. Dunford, *Transformations on sequence spaces*, Duke. Math. J. **3** (1937), 689–701.

[4] J. L. Doob, *Stochastic Processes*, John Wiley and Sons, 1953.

[5] N. Dunford and J. T. Schwartz, *Linear Operators. Part I*, John Wiley and Sons, 1971.

[6] M. Iosifescu, *On two recent papers on ergodicity in non-homogeneous Markov chains*, Ann. Math. Statist. **43** (1972), 1732–1736.

[7] M. Iosifescu and S. Grigorescu, *Dependence with Complete Connections and its Applications*, Cambridge University Press, Cambridge, 1990.

[8] M. Iosifescu and R. Theodorescu, *Random Processes and Learning*, Springer-Verlag, Berlin-Heidelberg, 1969.

[9] S. Kantorovitz, *Introduction to Modern Analysis*, Oxford Graduate Texts in Mathematics. Oxford University Press, 2003.

[10] A. S. Kechris, *Classical Descriptive Set Theory*, Graduate Texts in Mathematics, 156. Springer-Verlag, New York, 1995.

[11] S. C. Kleene, *Representation of events in nerve nets and finite automata*, In Automata Studies, Annals of Mathematics Studies, vol. 34, Princeton University Press, N.J., pp. 3–41; 1956.

[12] W. Maass and P. Orponen, *On the effect of analog noise in discrete time analog computations*, Neural Computation **10** (1998), 1071–1095.

[13] W. Maass and E. Sontag, *Analog neural nets with Gaussian or other common noise distribution cannot recognize arbitrary regular languages*, Neural Computation **11** (1999), 771–782.

[14] S. P. Meyn, and R. Tweedie, *Markov Chains and Stochastic Stability* , Springer-Verlag, New York, 1993.

[15] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden Day, San Francisco, 1964.

[16] S. Orey, *Limit Theorems for Markov Chain Transition Probabilities* , Van Nostrand Reinhold, London, 1971.

[17] A. Paz, *Ergodic theorems for infinite probabilistic tables*, Ann. Math. Statist. **41** (1970), 539–550.

[18] A. Paz, *Introduction to Probabilistic Automata*, Academic Press, London, 1971.

[19] M. Perles, M. Rabin, and E. Shamir, *The theory of definite automata.* IEEE Trans. EC-**12** (1963), 233–243.

[20] M. Rabin, *Probabilistic automata*, Information and Control **3** (1963), 230–245.

[21] M. Rabin and D. Scott, *Finite automata and their decision problems.* In "Sequential Machines", E. F. Moore, ed. Addison-Wesley, Reading, Massachusetts, 1964.

[22] H. T. Siegelmann and E. D. Sontag, *Analog computation via neural networks*, JCSS **50** (1995), 132–150.

[23] H. T. Siegelmann, *Neural Networks and Analog Computation: Beyond the Turing Limit*, Birkhauser, Boston, 1999.

[24] H. T. Siegelmann, A. Roitershtein and A. Ben-Hur, *Noisy neural networks and generalizations*, Proceedings of the Annual Conference on Neural Information Systems 1999 (NIPS*99), MIT Press, 2000.

[25] K. Yosida and S. Kakutani, *Operator-theoretical treatment of Markoff's process and mean ergodic theorem*, Ann. Math. **42** (1941), 188–228.