Bisimulation Metrics for Weighted Automata *

Borja Balle[1], Pascale Gourdeau[2], and Prakash Panangaden[3]

1 **Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.**
`b.deballepigem@lancaster.ac.uk`

2 **School of Computer Science, McGill University, Montreal, Quebec, Canada**
`pascale.gourdeau@mail.mcgill.ca`

3 **School of Computer Science, McGill University, Montreal, Quebec, Canada**
`prakash@cs.mcgill.ca`

Abstract

We develop a new bisimulation (pseudo)metric for weighted finite automata (WFA) that generalizes Boreale's linear bisimulation relation. Our metrics are induced by seminorms on the state space of WFA. Our development is based on spectral properties of sets of linear operators. In particular, the joint spectral radius of the transition matrices of WFA plays a central role. We also study continuity properties of the bisimulation pseudometric, establish an undecidability result for computing the metric, and give a preliminary account of applications to spectral learning of weighted automata.

1998 ACM Subject Classification F.1.1 Models of Computation

Keywords and phrases Weighted automata, bisimulation, metrics, spectral theory, learning

Digital Object Identifier 10.4230/LIPIcs.ICALP.2017.

1 Introduction

Weighted finite automata (WFA) form a fundamental computational model that subsumes probabilistic automata and various other types of quantitative automata. They are much used in machine learning and natural language processing, and are certainly relevant to quantitative verification and to the theory of control systems [16]. The theory of minimization of weighted finite automata goes back to Schützenberger [30] which implicitly exploits duality as made explicit in [9]. In [6] we began studying *approximate* minimization of WFA by using spectral methods. The idea there was to obtain automata for a given weighted language, smaller than the minimal possible which, of course, means that the automaton constructed does not *exactly* recognize the given weighted language but comes "close enough."

In [6] the notion of proximity to the desired language was captured by an ℓ_2 distance. However, a powerful technique for understanding approximate behavioural equivalence is by using more general *behavioural metrics*. In particular, with a behavioural pseudometric we recover bisimulation as the kernel. Such behavioural metrics for Markov processes were proposed by Giacalone et al. [20] and the first successful pseudometric that has bisimulation as its kernel is due to Desharnais et al. [14, 15]; see [28] for an expository account. The subject was greatly developed by van Breugel and Worrell [32] among others. For WFA, a beautiful treatment of linear bisimulation relations was given by Boreale [10]. We were motivated to develop a metric analogue of Boreale's linear bisimulation with the eventual goal

* The full version of this paper, including the appendix, can be found at https://arxiv.org/abs/1702.08017.

44th International Colloquium on Automata, Languages, and Programming (ICALP 2017).
Editors: Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl; Article No. ; pp. :1–:14
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of using it to analyze approximate minimization. In the present paper we develop the general theory of bisimulation (pseudo)metrics for WFA (and for weighted languages) deferring the application to approximate minimization to future work.

It turns out that in the linear algebraic setting appropriate to WFA it is a (semi)norm rather than a (pseudo)metric that is the fundamental quantity of interest. Indeed, as one might expect, in a vector space setting norms and seminorms are the natural objects from which metrics and pseudometrics can be derived. The bisimulation metric that we construct actually comes from a bisimulation seminorm which is obtained, as usual, using the Banach fixed-point theorem. Interestingly, we also provide a closed-form expression for the fixed point bisimulation seminorm and use it to study several of its properties.

Our main contributions are:

1. The construction of bisimulation seminorms and the associated pseudometric on WFA (Section 3). The existence of the fixed point depends on some delicate applications of spectral theory, specifically the joint spectral radius of a set of matrices.
2. We obtain metrics on the space of weighted languages from the metrics on WFA (Section 3).
3. We show two continuity properties of the metric; one using definitions due to Jaeger et al. [24] and the other developed here (Section 4).
4. We show undecidability results for computing our metrics (Section 5).
5. Nevertheless, we show that one can successfully exploit these metrics for applications in machine learning (Section 6).

The metric of the present paper led naturally to some sophisticated topological and spectral theory arguments which one would not have anticipated from the treatment of linear bisimulation in [10].

2 Background

In this section we recall preliminary definitions and results that will be used throughout the rest of the paper. We assume the reader is familiar with norms and vector spaces; these topics are reviewed in the appendix [4]. Here we discuss Boreale's linear bisimulation relations for weighted automata and provide a short primer on the joint spectral radius of a set of linear operators.

2.1 Strings and Weighted Automata

Given a finite alphabet Σ we let $\Sigma^\star$ denote the set of all finite strings with symbols in Σ and let Σ^∞ denote the set of all infinite strings with symbols in Σ and we write $\Sigma^\omega = \Sigma^\star \cup \Sigma^\infty$. The length of a string $x \in \Sigma^\omega$ is denoted by $|x|$; $|x| = \infty$ whenever $x \in \Sigma^\infty$. Given a string $x \in \Sigma^\omega$ and an integer $0 \leq t \leq |x|$ we write $x_{\leq t}$ to denote the prefix containing the first t symbols from x, with $x_{\leq 0} = \epsilon$. Given an integer $t \geq 0$ we will write Σ^t (resp. $\Sigma^{\leq t}$) for the set of all strings with length equal to (resp. at most) t. The reverse of a finite string $x = x_1 x_2 \cdots x_t$ is given by $\bar{x} = x_t x_{t-1} \cdots x_1$.

We only consider automata with weights in the real field $\mathbb{R}$. We will mostly be concerned with properties of weighted automata that are invariant under change of basis. Accordingly, our presentation uses weighted automata whose state space is an abstract real vector space.

A weighted finite automaton (WFA) is a tuple $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma \in \Sigma} \rangle$ where Σ is a finite alphabet, V is a finite-dimensional vector space, $\alpha \in V$ is a vector representing the initial weights, $\beta \in V^*$ is a linear form representing the final weights, and $\tau_\sigma : V \to V$ is a linear map representing the transition indexed by $\sigma \in \Sigma$. The vectors in V are called states of A. We shall denote by $n = \dim(A) = \dim(V)$ the dimension of A. The transition maps τ_σ can be extended to arbitrary finite strings in the obvious way.

A weighted automaton $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$ computes the function $f_A : \Sigma^\star \to \mathbb{R}$ (sometimes also referred to as the weighted language in $\mathbb{R}^{\Sigma^\star}$ recognized by A) given by $f_A(x) = \beta(\tau_x(\alpha))$. Given a WFA $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$ and a state $v \in V$ we define the weighted automaton $A_v = \langle \Sigma, V, v, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$ obtained from A by taking v as the initial state. We call f_{A_v} the function realized by state v. Similarly, give a linear form $w \in V^*$ we define the weighted automaton $A^w = \langle \Sigma, V, \alpha, w, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$ where the final weights are replaced by w. The reverse of a weighted automaton A is $\bar{A} = \langle \Sigma, V^*, \beta, \alpha, \{\tau_\sigma^\top\}_{\sigma\in\Sigma} \rangle$, where $\tau_\sigma^\top : V^* \to V^*$ is the transpose of τ_σ. It is easy to check that the function computed by $\bar{A}$ satisfies $f_{\bar{A}}(x) = f_A(\bar{x})$ for all $x \in \Sigma^\star$.

2.2 Linear Bisimulations

Linear bisimulations for weighted automata were introduced by Boreale in [10]. Here we recall the key definition and several important facts.

▶ **Definition 1.** A *linear bisimulation* for a weighted automaton $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$ on a vector space V is a linear subspace $W \subseteq V$ satisfying the following two conditions:

1. $\beta(v) = 0$ for all $v \in W$; that is, $W \subseteq \ker(\beta)$, and
2. W is invariant by each τ_σ; that is, $\tau_\sigma(W) \subseteq W$ for all $\sigma \in \Sigma$.

Furthermore, two states $u, v \in V$ are called W-*bisimilar* if $u - v \in W$.

In particular, the trivial subspace $W = \{0\}$ is always a linear bisimulation. The notion of W-bisimilarity induces an equivalence relation on V which we will denote by $\sim_W$. The kernel of an equivalence relation $\sim$ on a vector space V is the set of vectors in the equivalence class of the null vector: $\ker(\sim) = \{v \in V : v \sim 0\}$. It is immediate from the definition that for any bisimulation relation $\sim_W$ we have $\ker(\sim_W) = W$.

Given a weighted automaton A we say that $u, v \in V$ are A-bisimilar if there exists a bisimulation W for A such that $u \sim_W v$. The corresponding equivalence relation is denoted by $\sim_A$. Boreale showed in [10] that for every WFA A there exists a bisimulation W_A such that $\sim_{W_A}$ exactly coincides with $\sim_A$, and the bisimulation can be obtained as $W_A = \ker(\sim_A)$. He also showed that W_A is in fact the largest linear bisimulation for A in the sense that any other linear bisimulation W for A must be a subspace of W_A. Accordingly, we shall refer to the relation $\sim_A$ and the subspace W_A as A-bisimulation.

Note that the subspaces considered in Definition 1 are independent of the initial state α of A. In fact, A-bisimilarity can be understood as a relation between possible initial states for A. Indeed, using the definition of $\sim_A$ it is immediate to check that for any states $u, v \in V$ we have $u \sim_A v$ if and only if $f_{A_u} = f_{A_v}$. This implies that in a WFA where the bisimulation W_A corresponding to $\sim_A$ satisfies $W_A = \{0\}$ every state realizes a different function. Such an automaton is called observable. A weighted automaton is called reachable if the reverse $\bar{A}$ is observable.

A weighted automaton A is minimal if for any other weighted automaton A' over the same alphabet such that $f_A = f_{A'}$ we have $\dim(A) \leq \dim(A')$. It is also shown in [10] that linear bisimulations can be used to characterize minimality, in the sense that A is minimal if and only if it is observable and reachable.

2.3 Joint Spectral Radius

The joint spectral radius of a set of linear operators is a natural generalization of the spectral radius of a single linear operator. The joint spectral radius and several equivalent notions have been thoroughly studied since the 1960's. These radiuses arise in many fundamental problems in operator theory, control theory, and computational complexity. See [25] for an introduction to their properties and applications. Here we recall the basic definitions and some important facts related to quasi-extremal norms.

▶ **Definition 2.** The *joint spectral radius* of a collection $M = \{\tau_i\}_{i \in I}$ of linear maps $\tau_i : V \to V$ on a normed vector space $(V, \|\cdot\|)$ is defined as

$$\rho(M) = \limsup_{t\to\infty} \left(\sup_{T \in I^t} \left\| \prod_{i \in T} \tau_i \right\| \right)^{1/t} = \lim_{t\to\infty} \left(\sup_{T \in I^t} \left\| \prod_{i \in T} \tau_i \right\| \right)^{1/t} .$$

The second equality above is a generalization of Gelfand's formula for the spectral radius of a single operator due to Daubechies and Lagarias [11, 12]. An important fact about the joint spectral radius is that $\rho(M)$ is independent of the norm $\|\cdot\|$, i.e. one obtains the same radius regardless of the norm given to the vector space V. The joint spectral radius behaves nicely with respect to direct sums, in the sense that given two sets of operators $M = \{\tau_i\}_{i\in I}$ and $M' = \{\tau'_i\}_{i\in I}$, then $\rho(\{\tau_i \oplus \tau'_i\}_{i\in I}) = \max\{\rho(M), \rho(M')\}$.

The notion of joint spectral radius can be readily extended to weighted automata. Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ be a weighted automaton with states on a normed vector space $(V, \|\cdot\|)$. Then the spectral radius of A is defined as $\rho(A) = \rho(M)$ where $M = \{\tau_\sigma\}_{\sigma\in\Sigma}$. In this case the definition above can be rewritten as

$$\rho(A) = \lim_{t\to\infty} \left(\sup_{x\in\Sigma^t} \|\tau_x\| \right)^{1/t} .$$

Now we discuss several fundamental properties of the joint spectral radius that will play a role in the rest of the paper. Like in the case of the classic spectral radius, the joint spectral radius is upper bounded by the norms of the operators in M: $\rho(M) \le \sup_{i\in I} \|\tau_i\|$. Obtaining lower bounds for $\rho(M)$ is a major problem directly related to the hardness of computing approximations to $\rho(M)$. An approach often considered in the literature is to search for extremal norms. A norm $\|\cdot\|$ on V is extremal for M if the corresponding induced norm satisfies $\|\tau_i\| \le \rho(M)$ for all $i \in I$. This immediately implies that given an extremal norm for M we have $\rho(M) = \sup_{i\in I} \|\tau_i\|$. Conditions on M guaranteeing the existence of an extremal norm have been derived by Barabanov and others; see [33] and references therein. However, most of these conditions are quite technical and algorithmically hard to verify. On the other hand, if one only insists on approximate extremality, the following result, due to Rota and Strang, guarantees the existence of such norms for any set of matrices M that is compact with respect to the topology generated by the operator norm in V.

▶ **Theorem 3** ([29]). *Let $M = \{\tau_i\}_{i\in I}$ be a compact set of linear maps on V. For any $\eta > 0$ there exists a norm $\|\cdot\|$ on V that satisfies $\|\tau_i(v)\| \le (\rho(M) + \eta)\|v\|$ for every $i \in I$ and every $v \in V$.*

The statement above is in fact a special case of Proposition 1 in [29]; a proof for finite sets M can be found in [8]. An important result due to Barabanov [7] states that the function $M \mapsto \rho(M)$ defined on compact sets of operators is continuous (see also [22]). Another result that we will need was again proved by Barbanov in [7] and it states that if M is a bounded set of linear operators and $\bar{M}$ denotes its closure then $\rho(M) = \rho(\bar{M})$. Note that if M is bounded then its closure $\bar{M}$ is compact by the Heine–Borel theorem.

A special case which makes the joint spectral radius easier to work with is when the set of matrices M is irreducible. A set of linear maps M is called *irreducible* if the only subspaces $W \subseteq V$ such that $\tau_i(W) \subseteq W$ for all $i \in I$ are $W = \{0\}$ and $W = V$. If there exists a non-trivial subspace $W \subset V$ invariant by all τ_i we say that M is reducible. In fact, almost all sets of matrices are irreducible in following sense. The Hausdorff distance between two sets of linear maps M and M' on the same normed vector space $(V, \|\cdot\|)$ is given by

$$d_H(M, M') = \max\left\{\sup_{\tau\in M}\inf_{\tau'\in M'}\|\tau - \tau'\|, \sup_{\tau'\in M'}\inf_{\tau\in M}\|\tau-\tau'\|\right\} .$$

It is possible to show that irreducible sets of matrices are dense among compact sets of matrices with respect to the topology induced by the Haussdorff distance. Furthermore, Wirth showed in [33] that the joint spectral radius is locally Lipschitz continuous around irreducible sets of matrices with respect to the Hausdorff topology (see also [26] for explicit expressions for the Lipschitz constants). This can be seen as an extension of Barabanov's continuity result providing extra information about the behaviour of the function $M \mapsto \rho(M)$.

Again, the concept of irreducibility can be readily extended to WFA. We say that the weighted automaton $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ is irreducible if $M = \{\tau_\sigma\}_{\sigma\in\Sigma}$ is irreducible. This concept will play a role in Section 6. The following result provides a characterization of irreducibility for weighted automata in terms of minimality. In particular, the result shows that irreducibility is a stronger condition than minimality. A proof is provided in the appendix [4].

▶ **Theorem 4.** *A weighted automaton $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ is irreducible if and only if A_v^w is minimal for all $v \in V$ and $w \in V^*$ with $v \neq 0$ and $w \neq 0$.*

3 Bisimulation Seminorms and Pseudometrics for WFA

In the same way that the largest bisimulation relation in many settings can be obtained as a fixed point of a certain operator on equivalence relations, a possible way to define bisimulation (pseudo)metrics is via a similar fixed-point construction. See [17] for an example in the case of Markov decision processes. In this section, the fixed-point construction is used to obtain a bisimulation seminorm on states of a given WFA. Given two WFA we can build their difference automaton A and compute the corresponding seminorm of the initial state of A. This construction yields a bisimulation pseudometric between weighted automata.

Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ be a weighted automaton over the vector space V. Let $\mathcal{S}$ denote the set of all seminorms on V. Given $\gamma > 0$ we define the map $F_{A,\gamma} : \mathcal{S} \to \mathcal{S}$ between seminorms given by

$$F_{A,\gamma}(s)(v) = |\beta(v)| + \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(v)) \ . \tag{1}$$

Note that this definition is independent of the initial state α, as is the linear bisimulation for A described in Section 2.2. In the sequel we shall write F instead of $F_{A,\gamma}$ whenever A and γ are clear from the context.

To verify that $F : \mathcal{S} \to \mathcal{S}$ is well defined we must check that the image $F(s)$ of any seminorm s is also a seminorm. Absolute homogeneity is immediate by the linearity of β and τ_σ and the absolute homogeneity of s. For the subadditivity we have

$$\begin{aligned} F(s)(u+v) &= |\beta(u+v)| + \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(u+v)) \\ &= |\beta(u) + \beta(v)| + \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(u) + \tau_\sigma(v)) \\ &\leq |\beta(u)| + |\beta(v)| + \gamma \max_{\sigma\in\Sigma} (s(\tau_\sigma(u)) + s(\tau_\sigma(v))) \\ &\leq F(s)(u) + F(s)(v) \ , \end{aligned}$$

where the last inequality uses subadditivity of the maximum.

To construct bisimulation seminorms for the states of a weighted automaton A we shall study the fixed points of $F_{A,\gamma}$. We start by showing that $F_{A,\gamma}$ has a unique fixed point whenever γ is small enough.

▶ **Theorem 5.** *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$. If $\gamma < 1/\rho(A)$, then $F_{A,\gamma}$ has a unique fixed point.*

Proof. For simplicity, let $F = F_{A,\gamma}$. By the assumption on γ there exists some $\delta > 0$ such that $\gamma \leq 1/(\rho(A)+\delta)$. Now take $M = \{\tau_\sigma\}_{\sigma\in\Sigma}$ and $\eta = \delta/2$ and let $\|\cdot\|$ be the corresponding quasi-extremal norm on V obtained from Theorem 3. Using this norm we can endow $\mathcal{S}$ with the metric given by $d(s, s') = \sup_{\|v\|\leq 1} |s(v) - s'(v)|$ to obtain a complete metric space $(\mathcal{S}, d)$. To see this, note that for a fixed v with $\|v\| \leq 1$ the sequence $(s_n(v))$ is Cauchy, hence convergent. Call this limit $s(v)$; it is straightforward to see that this defines a seminorm. Thus, if we show that F is a contraction on $\mathcal{S}$ with respect to this metric, then by Banach's fixed point theorem F has a unique fixed point. To see that F is indeed a contraction we start by observing that:

$$d(F(s), F(s')) = \sup_{\|v\|\leq 1} |F(s)(v) - F(s')(v)| = \gamma \sup_{\|v\|\leq 1} \left|\max_\sigma s(\tau_\sigma(v)) - \max_{\sigma'} s'(\tau_{\sigma'}(v))\right| \ . \quad (2)$$

Fix any $v \in V$ with $\|v\| \leq 1$ and suppose without loss of generality (otherwise we exchange s and s') that $\max_\sigma s(\tau_\sigma(v)) \geq \max_{\sigma'} s'(\tau_{\sigma'}(v))$. Then, letting $\sigma_* = \arg\max_\sigma s(\tau_\sigma(v))$ and using the absolute homogeneity of s and s', it can be shown that:

$$\left|\max_\sigma s(\tau_\sigma(v)) - \max_{\sigma'} s'(\tau_{\sigma'}(v))\right| \leq \|\tau_{\sigma_*}(v)\| d(s, s') \ . \quad (3)$$

We refer the reader to the appendix [4], for a full derivation. Finally, we use the definition of $\|\cdot\|$ and the choices of δ and η to see that

$$\gamma\|\tau_{\sigma_*}(v)\| \leq \gamma(\rho(A) + \eta)\|v\| \leq \frac{\rho(A) + \delta/2}{\rho(A) + \delta} < 1 \ ,$$

from which we conclude by combining (2) with (3) that $d(F(s), F(s')) < d(s, s')$. ◀

We now exhibit the fixed point of $F_{A,\gamma}$ in closed form. This provides a useful formula for studying properties of the resulting seminorm.

▶ **Theorem 6.** *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$. Suppose $\gamma < 1/\rho(A)$ and let $s_{A,\gamma} \in \mathcal{S}$ be the fixed point of $F_{A,\gamma}$. Then for any $v \in V$ we have*

$$s_{A,\gamma}(v) = \sup_{x\in\Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t |\beta(\tau_{x_{\leq t}}(v))| = \sup_{x\in\Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t |f_{A_v}(x_{\leq t})| \ . \quad (4)$$

The proof can be found in the appendix [4]. The next theorem is the main result of this section. It shows that any seminorm arising as a fixed point of $F_{A,\gamma}$ captures the notion of A-bisimulation through its kernel for any γ. Namely, two states $u, v \in V$ are A-bisimilar if and only if $s_{A,\gamma}(u - v) = 0$. Note that this result is independent of the choice of γ, as long as the fixed point of $F_{A,\gamma}$ is guaranteed to exist.

▶ **Definition 7.** Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ be a weighted automaton with A-bisimulation $\sim_A$. We say that a seminorm s over V is a *bisimulation seminorm* for A if $\ker(s) = \ker(\sim_A)$.

▶ **Theorem 8.** *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma \in \Sigma} \rangle$. For any $0 < \gamma < 1/\rho(A)$ the fixed point $s_{A,\gamma} \in \mathcal{S}$ of $F_{A,\gamma}$ is a bisimulation seminorm for A.*

Proof. For simplicity, let $F = F_{A,\gamma}$ and $s = s_{A,\gamma}$. Since $W_A = \ker(\sim_A)$ is the largest bisimulation for A, it suffices to show that $\ker(s)$ is a bisimulation for A with $W_A \subseteq \ker(s)$. For the first property we recall that $\ker(s)$ is a linear subspace of V and note that for any $v \in \ker(s)$ we have, using Theorem 6,

$$0 = s(v) = |\beta(v)| + \sup_{x \in \Sigma^\infty} \sum_{t=1}^{\infty} \gamma^t |\beta(\tau_{x_{\leq t}}(v))| \geq |\beta(v)| \geq 0 \ .$$

Therefore $\ker(s) \subseteq \ker(\beta)$. Using the fact that $\beta(v) = 0$, we can also verify the invariance of $\ker(s)$ under all τ_σ, namely $s(\tau_\sigma(v)) = 0$ for all $v \in \ker(s)$ and $\sigma \in \Sigma$ (the full derivation is shown in the appendix [4]). Therefore $\ker(s)$ is a bisimulation for A.

Now let $v \in W_A$. Since W_A is contained in the kernel of β and is invariant for all τ_σ, we see that $\beta(\tau_x(v)) = 0$ for all $x \in \Sigma^\star$. Therefore, using the expression for s given in Theorem 6 we obtain $s(v) = 0$. This concludes the proof. ◀

Because every fixed point of $F_{A,\gamma}$ is a seminorm whose kernel agrees with that of Boreale's bisimulation relation $\sim_A$, we shall call them *γ-bisimulation seminorms* for A. Interestingly, we can show that when A is observable then every γ-bisimulation seminorm is in fact a norm.

▶ **Corollary 9.** *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma \in \Sigma} \rangle$ and $\gamma < 1/\rho(A)$. If A is observable then the γ-bisimulation seminorm $s_{A,\gamma}$ is a norm.*

Proof. By Theorem 8 and the observability of A we have $\ker(s_{A,\gamma}) = \ker(\sim_A) = \{0\}$. Thus, $s_{A,\gamma}$ is a norm. ◀

Given an automaton A, and state vectors $v, w \in V$, the pseudometric between states of A induced by $s_{A,\gamma}$ is $d_{A,\gamma}(v, w) = s_{A,\gamma}(v - w)$. Pseudometrics of this form will be called *γ-bisimulation pseudometrics*. By Corollary 9, if A is observable then $d_{A,\gamma}$ is in fact a metric.

To conclude this section we show how to use our γ-bisimulation pseudometrics to define a pseudometric between weighted automata. In order to capture the idea of distance between two WFA let us build the automaton computing the difference between their functions. Given weighted automata $A_i = \langle \Sigma, V_i, \alpha_i, \beta_i, \{\tau_{i,\sigma}\}_{\sigma \in \Sigma} \rangle$ for $i = 1, 2$, we define their *difference automaton* as $A = A_1 - A_2 = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma \in \Sigma} \rangle$ where $V = V_1 \oplus V_2$, $\alpha = \alpha_1 \oplus (-\alpha_2)$, $\beta = \beta_1 \oplus \beta_2$, and $\tau_\sigma = \tau_{1,\sigma} \oplus \tau_{2,\sigma}$ for all $\sigma \in \Sigma$. Note that A satisfies $f_A(x) = f_{A_1}(x) - f_{A_2}(x)$ for all $x \in \Sigma^\star$ and that $\rho(A) = \max\{\rho(A_1), \rho(A_2)\}$. Then, letting $s_{A,\gamma}$ be the bisimulation seminorm for A we are ready to define our bisimulation distance between weighted automata.

▶ **Definition 10.** Let A_1 and A_2 be two weighted automata and let A be their difference automaton. For any $\gamma < 1/\rho(A)$ we define the *γ-bisimulation distance* between A_1 and A_2 as $d_\gamma(A_1, A_2) = s_{A,\gamma}(\alpha)$.

By exploiting the closed form expression for $s_{A,\gamma}$ given in Theorem 6 we can provide a closed form expression for d_γ.

▶ **Corollary 11.** *Let A_1 and A_2 two weighted automata and $\gamma < 1/\max\{\rho(A_1), \rho(A_2)\}$. Then the γ-bisimulation distance between A_1 and A_2 is given by*

$$d_\gamma(A_1, A_2) = \sup_{x \in \Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t \left| f_{A_1}(x_{\leq t}) - f_{A_2}(x_{\leq t}) \right| \ . \tag{5}$$

Using the properties of our bisimulation seminorms one can immediately see that d_γ is indeed a pseudometric between all pairs of WFA such that $\gamma < 1/\rho(A_1 - A_2)$. It is also easy to see that d_γ captures the notion of equivalence between weighted automata, in the sense that $d_\gamma(A_1, A_2) = 0$ if and only if $f_{A_1} = f_{A_2}$. Therefore, since minimal weighted automata are unique up to a change of basis, the only way to have $d_\gamma(A_1, A_2) = 0$ when A_1 is minimal is to have either $A_1 = A_2$ or A_2 is a non-minimal WFA recognizing the same weighted language as A_1. In particular, this implies that d_γ is a metric on the set of all minimal WFA A with $\gamma < 1/\rho(A)$.

4 Continuity Properties

In this section we study several continuity properties of our bisimulation pseudometrics between weighted automata. The continuity notions we consider are adapted from those presented by Jaeger et al. in [24], which are developed for labelled Markov chains. Here we extend their definitions of parameter continuity and property continuity to the case of weighted automata. Such notions can be motivated by applications of metrics between transition systems to problems in machine learning [15, 19, 18]; see Section 6 for a discussion on how to use our bisimulation pseudometrics in the analysis of learning algorithms.

4.1 Parameter Continuity

Given a sequence of weighted automata A_i converging to a weighted automaton A, parameter continuity captures the notion that, as the weights in A_i converge to the weights in A, the behavioural distance between A_i and A tends to zero. To make this formal we first define convergence for a sequence of automata and then parameter continuity.

▶ **Definition 12.** Let $(A_i)_{i\in\mathbb{N}}$ be a sequence of WFA $A_i = \langle \Sigma, V, \alpha_i, \beta_i, \{\tau_{i,\sigma}\}_{\sigma\in\Sigma}\rangle$ over the same alphabet Σ and normed vector space $(V, \|\cdot\|)$. We say that the sequence (A_i) *converges* to $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ if $\lim_{i\to\infty} \|\alpha_i - \alpha\| = 0$, $\lim_{i\to\infty} \|\beta_i - \beta\|_* = 0$, and $\lim_{i\to\infty} \|\tau_{i,\sigma} - \tau_\sigma\| = 0$ for all $\sigma \in \Sigma$.

▶ **Definition 13.** A pseudometric d between weighted automata is *parameter continuous* if for any sequence $(A_i)_{i\in\mathbb{N}}$ converging to some weighted automaton A we have $\lim_{i\to\infty} d(A, A_i) = 0$.

The main result of this section is the following theorem stating that our bisimulation pseudometric d_γ is parameter continuous.

▶ **Theorem 14.** *The γ-bisimulation distance between weighted automata is parameter continuous for any sequence of weighted automata $(A_i)_{i\in\mathbb{N}}$ converging to a weighted automaton A with $\gamma < 1/\rho(A)$.*

The proof of this result is quite technical and combines the following two tools:

1. A technical estimate of $d_\gamma(A, A_i)$ in terms of the distance between the weights of A and A_i with respect to a certain norm (Lemma 25 in the appendix [4]). This result also plays a prominent result in Section 6.
2. Several topological properties of the joint spectral radius discussed in Section 2.3.

These proofs are given in the appendix [4].

4.2 Input Continuity

Inspired by the notion of property continuity presented in [24], input g-continuity encapsulates the idea that an upper bound on the behavioural distance between two systems should entail an upper bound on the difference between their outputs on any input $x \in \Sigma^\star$.

▶ **Definition 15.** Let $g : \mathbb{N} \to \mathbb{R}$ be such that $g(l) > 0$ for all $l \in \mathbb{N}$. A distance function d between weighted automata is *input g-continuous* when the following holds: if $(A_i)_{i\in\mathbb{N}}$ is a sequence of weighted automata such that $\lim_{i\to\infty} d(A, A_i) = 0$ for some weighted automaton A, then one has

$$\lim_{i\to\infty} \sup_{x\in\Sigma^\star} \frac{|f_A(x) - f_{A_i}(x)|}{g(|x|)} = 0 \ . \tag{6}$$

Note the special case $g(l) = 1$ is tightly related to the notion of property continuity presented in [24]. The authors of that paper consider differences between the probabilities of the same event under different labelled Markov chains, and therefore always have numbers between 0 and 1. However, for general weighted automata the quantity $|f_A(x) - f_{A'}(x)|$ can grow unboundedly with $|x|$. Thus, in some cases we will need to have a $g(|x|)$ growing with $|x|$ in order to guarantee that (6) stays bounded. The next two results, whose proofs are deferred to the appendix [4], show that essentially $g(|x|) = \gamma^{-|x|}$ is the threshold between input continuity and input non-continuity in our γ-bisimulation pseudometrics.

▶ **Theorem 16.** *The pseudometric d_γ from Definition 10 is input g-continuous for any $g(l) = \Omega(\gamma^{-l})$.*

Note that when $\gamma > 1$ (i.e. when dealing with weighted automata with $\rho(A) \leq 1$) we have $g(l) = 1 \in \Omega(\gamma^{-l})$. This shows that in the case of weighted automata A where every transition operator τ_σ can be represented by a stochastic matrix—a fact that implies $\rho(A) = 1$—our γ-bisimulation pseudometric is property continuous with respect to the definition in [24].

Further, if g does not grow fast enough as a function of the size of $x \in \Sigma^\star$, then our bisimulation pseudometric is not input g-continuous. In particular, the proof of Theorem 17 provides simple examples of cases where d_γ is not input g-continuous.

▶ **Theorem 17.** *Let $0 < \gamma < 1$. The pseudometric d_γ from Definition 10 is not input g-continuous for any $g(l) = c^{o(l)}$ with $c > 1$.*

5 An Undecidability Result

In this section we will prove that given a weighted automaton $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$, a discount factor $\gamma < 1/\rho(A)$, and a threshold $\nu > 0$, it is undecidable to check whether $s_{A,\gamma}(\alpha) > \nu$. This implies that in general the seminorms and pseudometrics studied in the previous sections are not computable.

The proof of our undecidability result involves a reduction from an undecidable planning problem. Partially observable Markov decision processes (POMDPs) are a generalization of Markov Decision Processes (MDPs) where we have a set of observations Ω and conditional observation probabilities O. Each state emits some observation $o \in \Omega$ with a certain probability, and so we have a belief over which state we are in after taking an action and observing o. An MDP is a special case of a POMDP where each state has a unique observation, and an unobservable Markov decision process (UMDP) is a special case of a POMDP where all the states emit the same observation. While planning for infinite-horizon UMDPs is undecidable [27], planning for finite-horizon POMDPs is decidable.

Formally, a UMDP is a tuple $U = \langle \Sigma, Q, \alpha, \{\beta_\sigma\}_{\sigma\in\Sigma}, \{T_\sigma\}_{\sigma\in\Sigma}, \gamma\rangle$ where Σ is a finite set of actions, Q is a finite set of states, $\alpha : Q \to [0, 1]$ is a probability distribution over initial states in Q, $\beta_\sigma : Q \to \mathbb{R}$ represents the rewards obtained by taking action σ from every state in Q, $T_\sigma : Q \times Q \to [0, 1]$ is the transition kernel between states for action σ (i.e. $T_\sigma(q, q')$ is the probability of transitioning to q' given that action σ is taken in q), and $0 < \gamma < 1$ is a discount factor. The value $V_U(x)$ of an infinite sequence of actions $x \in \Sigma^\infty$ in U is the expected discounted cumulative reward collected by executing the actions in x in U starting from a state drawn from α. This can be obtained as follows:

$$V_U(x) = \sum_{t=1}^{\infty} \gamma^{t-1} \alpha^\top T_{x_{\leq t-1}} \beta_{x_t} \enspace , \tag{7}$$

where $T_y = T_{y_1} \cdots T_{y_t}$ for any finite string $y = y_1 \cdots y_t$ and $T_\epsilon = I$. The following undecidability result was proved by Madani et al. in [27].

▶ **Theorem 18** (Theorem 4.4 in [27]). *The following problem is undecidable: given a UMDP U and a threshold ν decide whether there exists a sequence of actions $x \in \Sigma^\infty$ such that $V_U(x) > \nu$.*

Given a UMDP $U = \langle \Sigma, Q, \alpha, \{\beta_\sigma\}_{\sigma\in\Sigma}, \{T_\sigma\}_{\sigma\in\Sigma}, \gamma \rangle$, we say that U has action-independent rewards if $\beta_\sigma = \beta$ for all $\sigma \in \Sigma$. We say that U has non-negative rewards if $\beta_\sigma(q) \geq 0$ for all $q \in Q$ and $\sigma \in \Sigma$. A careful inspection of the proof in [27] reveals that in fact the reduction provided in the paper always produces as output a UMDP with non-negative action-independent rewards. Thus, we have the following corollary, which forms the basis of our reduction showing that s_γ is not computable.

▶ **Corollary 19.** *The problem in Theorem 18 remains undecidable when restricted to UMDP with non-negative action-independent rewards.*

▶ **Theorem 20.** *The following problem is undecidable: given a weighted automaton $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$, a discount factor $\gamma < 1/\rho(A)$, and a threshold $\nu > 0$, decide whether $s_{A,\gamma}(\alpha) > \nu$.*

Proof. Let $U = \langle \Sigma, Q, \alpha, \beta, \{T_\sigma\}_{\sigma\in\Sigma}, \gamma \rangle$ be a UMDP with non-negative action-independent rewards. With each UMDP of this form we associate the weighted automaton $A = \langle \Sigma, \mathbb{R}^Q, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$. Here we assume that the linear form $\beta : \mathbb{R}^Q \to \mathbb{R}$ is given by $\beta(v) = v^\top \beta$, and that the linear operators $\tau_\sigma : \mathbb{R}^Q \to \mathbb{R}^Q$ are given by $\tau_\sigma(v) = v^\top T_\sigma$.

Note that the matrices T_σ are row-stochastic and therefore we have $\rho(A) \leq \max_\sigma \|\tau_\sigma\|_\infty = 1$. Thus, the discount factor in U satisfies $\gamma < 1 \leq 1/\rho(A)$ and the bisimulation seminorm $s_{A,\gamma}$ associate with A is defined. Using that U has non-negative action-independent rewards we can write for any $x \in \Sigma^\infty$:

$$V_U(x) = \sum_{t=1}^{\infty} \gamma^{t-1} \alpha^\top T_{x_{\leq t-1}} \beta = \sum_{t=0}^{\infty} \gamma^{t} \alpha^\top T_{x_{\leq t}} \beta = \sum_{t=0}^{\infty} \gamma^{t} |\alpha^\top T_{x_{\leq t}} \beta| = \sum_{t=0}^{\infty} \gamma^{t} |\beta(\tau_{x_{\leq t}}(\alpha))| \enspace .$$

Therefore we have the relation $s_{A,\gamma}(\alpha) = \sup_{x\in\Sigma^\infty} V_U(x)$ between the bisimulation seminorm of A and the value of U. Since deciding whether $V_U(x) > \nu$ for some $x \in \Sigma^\infty$ is undecidable, the theorem follows. ◀

6 Application: Spectral Learning for WFA

An important problem in machine learning is that of finding a weighted automaton $\hat{A}$ approximating an unknown automaton A given only access to data generated by A. A variety of algorithms in different learning frameworks have been considered in the literature; see [5] for an introductory survey. In most learning scenarios it is impossible to exactly recover the target automaton A from a finite amount of data. In that case one aims for algorithms with formal guarantees of the form "the output $\hat{A}$ automaton gets closer to A as the amount of training data grows". To prove such a result one obviously needs a way to measure the distance between two WFA. In this section we show how our γ-bisimulation pseudometric can be used to provide formal learning guarantees for a family of learning algorithms widely referred to as spectral learning. We also briefly discuss the case for behavioural metrics in automata learning problems and compare our metric to other metrics used in the spectral learning literature.

Generally speaking, spectral learning algorithms for WFA work in two phases: the first phase uses the data obtained from the target automaton A to estimate a finite sub-block of the Hankel matrix of f_A; the second phase computes the singular value decomposition of this Hankel matrix and uses the corresponding singular vectors to solve a set of systems of linear equations yielding the weights of the output WFA $\hat{A}$. The Hankel matrix of a function $f : \Sigma^\star \to \mathbb{R}$ is an infinite matrix $H_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ with entries given by $H_f(x, y) = f(xy)$, where xy denotes the string obtained by concatenating the prefix x with the suffix y. Spectral learning algorithms work with a finite sub-block $H \in \mathbb{R}^{P \times S}$ of this Hankel matrix indexed by a set of prefixes $P \subset \Sigma^\star$ and a set of suffixes $S \subset \Sigma^\star$. The pair $B = (P, S)$ is usually an input to the algorithm, in which case formal learning guarantees can be provided under the assumption that B is complete for H_{f_A}. This assumption essentially states that the sub-block of H_{f_A} indexed by B contains enough information to recover a WFA equivalent to A, and is composed of a syntactic condition ensuring B contains a set of prefixes and their extensions by any symbol in Σ, and an algebraic condition ensuring the rank of the Hankel sub-matrix indexed by B has the same rank as the full Hankel matrix H_{f_A}. We refer the reader to [5, 3] for further details about the spectral learning algorithm and a discussion of the completeness property for B. In the sequel we focus on the analysis of the error in the output of the spectral learning algorithm, and show how to provide learning guarantees in terms of our distance d_γ.

The following lemma encapsulates the first step of the analysis of spectral learning algorithms. It shows how the error between the operators of A and $\hat{A}$ depends on the error between the true and the approximated Hankel matrix as measured by the standard operator ℓ_2-norm.

▶ **Lemma 21.** *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma} \rangle$ be a WFA and let H be a finite sub-block of the Hankel matrix H_{f_A} indexed by $B = (P, S)$. Suppose $\hat{A} = \langle \Sigma, V, \hat{\alpha}, \hat{\beta}, \{\hat{\tau}_\sigma\}_{\sigma\in\Sigma} \rangle$ is the WFA returned by the spectral learning algorithm using an estimation $\hat{H}$ of H. Let $\|\cdot\|$ be any norm on V. If B is complete, then we have $\|\alpha - \hat{\alpha}\|, \|\beta - \hat{\beta}\|_*, \max_{\sigma\in\Sigma} \|\tau_\sigma - \hat{\tau}_\sigma\| \le O(\|H - \hat{H}\|_2)$ as $\|H - \hat{H}\|_2 \to 0$. Furthermore, the constants hidden in the big-O notation only depend on the norm $\|\cdot\|$, the Hankel sub-block indices $B = (P, S)$, and the size of the alphabet $|\Sigma|$.*

Proof. Combine Lemma 9.3.5 and Lemma 6.3.2 from [2]. ◀

The results from [2] also provide explicit expressions for the constants hidden in the big-O notation. Concentration of measure for random matrices can be used to show that as the amount of training data increases then the distance between H and $\hat{H}$ converges to zero with high probability (see e.g. [13]). Thus, Lemma 21 implies that as more training data becomes available, spectral learning will output a WFA $\hat{A}$ converging to A.

The last step in the analysis involves showing that as the weights of $\hat{A}$ get closer to the weights of A, the behaviour of the two automata also gets closer. Invoking the parameter continuity of d_γ (Theorem 14) one readily sees that $d_\gamma(A, \hat{A}) \to 0$ as $\|H - \hat{H}\|_2 \to 0$. This provides a proof of consistency of spectral learning with respect to the γ-bisimulation pseudometric. However, machine learning applications often require more precise information about the convergence rate of $d_\gamma(A, \hat{A})$ in order to, for example, compute the amount of data required to achieve a certain error. The following result provides such rate of convergence in the case where the target automaton is irreducible.

▶ **Theorem 22.** *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ be an irreducible WFA and let H be a finite sub-block of the Hankel matrix H_{f_A} indexed by $B = (P, S)$. Suppose $\hat{A} = \langle \Sigma, V, \hat{\alpha}, \hat{\beta}, \{\hat{\tau}_\sigma\}_{\sigma\in\Sigma}\rangle$ is the WFA returned by the spectral learning algorithm using an estimation $\hat{H}$ of H. Suppose B is complete. Then for any $\gamma < 1/\rho(A)$ we have $d_\gamma(A, \hat{A}) \leq O(\|H - \hat{H}\|_2)$ as $\|H - \hat{H}\|_2 \to 0$. Furthermore, the hidden constants in the big-O notation only depend on A, γ, the Hankel block indices $B = (P, S)$, and the size of the alphabet $|\Sigma|$.*

The local Lipschitz continuity of ρ around irreducible sets of matrices plays an important role in the proof of this result (see the appendix [4]). Nonetheless, the irreducibility constraint is not a stringent one since the sets of irreducible matrices are known to be dense among compact sets of matrices with respect to the Hausdorff metric.

We conclude this section by comparing Theorem 22 with analyses of spectral learning based on other error measures. We start by noting that all finite-sample analyses of spectral learning for WFA we are aware of in the literature provide error bounds in terms of some finite variant of the ℓ_1 distance. In particular, the analyses in [23, 31] bound $\sum_{x\in\Sigma^t} |f_A(x) - f_{\hat{A}}(x)|$ for a fixed $t \geq 0$, while the analyses in [1, 2, 21] extend the bounds to $\sum_{x\in\Sigma^{\leq t}} |f_A(x) - f_{\hat{A}}(x)|$ for a fixed $t \geq 0$. This approach poses several drawbacks, including:

1. Finite ℓ_1-norms provide a pseudo-metric between WFA whose kernel includes pairs of non-equivalent WFA.
2. The number of samples required to achieve a certain error increase with the horizon t, meaning that more data is required to get the same error on longer strings, and that existing bounds become vacuous in the case $t \to \infty$.

In contrast, our result in terms of d_γ establishes a bound on the discrepancy between A and $\hat{A}$ on strings of arbitrary length and will never assign zero distance to a pair of automata realizing different functions. Furthermore, our bisimulation metric still makes sense outside the setting of spectral learning of probabilistic automata where most of the techniques mentioned above have been developed.

7 Conclusion

The metric developed in this paper was very much motivated and informed by spectral ideas. Not surprisingly it was well suited for analyzing spectral learning algorithms for weighted automata. Two obvious directions for future work are:

1. Approximation algorithms for the bisimulation metric.
2. Exploring the relation to approximate minimization.

Both of these are well underway. It seems that some recent ideas from non-linear optimization are very useful in developing approximation algorithms and we hope to be able to report our results soon. Exploring the relation to approximate minimization is less further along, but the spectral ideas at the heart of the approximate minimization algorithm in [6] should be well adapted to the techniques of the present paper.

Acknowledgements. We would like to thank Doina Precup who was actively involved in the approximate minimization work. This research has been supported by a grant from NSERC (Canada).

References

1 R. Bailly. *Méthodes spectrales pour l'inférence grammaticale probabiliste de langages stochastiques rationnels*. PhD thesis, Aix-Marseille Université, 2011.

2 Borja Balle. *Learning Finite-State Machines: Algorithmic and Statistical Aspects*. PhD thesis, Universitat Politècnica de Catalunya, 2013.

3 Borja Balle, Xavier Carreras, Franco M. Luque, and Ariadna Quattoni. Spectral learning of weighted automata: A forward-backward perspective. *Machine Learning*, 2014.

4 Borja Balle, Pascale Gourdeau, and Prakash Panangaden. Bisimulation metrics for weighted automata. 2017. URL: `https://arxiv.org/abs/1702.08017`.

5 Borja Balle and Mehryar Mohri. Learning weighted automata. In *Conference on Algebraic Informatics*, 2015.

6 Borja Balle, Prakash Panangaden, and Doina Precup. A canonical form for weighted automata and applications to approximate minimization. In *Proceedings of the Thirtieth Annual ACM-IEEE Symposium on Logic in Computer Science*, July 2015.

7 Nikita E. Barabanov. On the Lyapunov indicator of discrete inclusions, part I, II, and III. *Avtomatika i Telemekhanika*, 2:40–46, 1988.

8 Vincent D. Blondel, Yurii Nesterov, and Jacques Theys. On the accuracy of the ellipsoid norm approximation of the joint spectral radius. *Linear Algebra and its Applications*, 394:91–107, 2005.

9 Filippo Bonchi, Marcello M. Bonsangue, Helle Hvid Hansen, Prakash Panangaden, Jan Rutten, and Alexandra Silva. Algebra-coalgebra duality in Brzozowski's minimization algorithm. *ACM Transactions on Computational Logic*, 2014.

10 Michele Boreale. Weighted bisimulation in linear algebraic form. In *CONCUR 2009-Concurrency Theory*, pages 163–177. Springer, 2009.

11 Ingrid Daubechies and Jeffrey C. Lagarias. Sets of matrices all infinite products of which converge. *Linear algebra and its applications*, 161:227–263, 1992.

12 Ingrid Daubechies and Jeffrey C. Lagarias. Corrigendum/addendum to: Sets of matrices all infinite products of which converge. *Linear Algebra and its Applications*, 327(1-3):69–83, 2001.

13 François Denis, Mattias Gybels, and Amaury Habrard. Dimension-free concentration bounds on hankel matrices for spectral learning. *Journal of Machine Learning Research*, 17(31):1–32, 2016.

14 J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Metrics for labeled Markov systems. In *Proceedings of CONCUR99*, number 1664 in Lecture Notes in Computer Science. Springer-Verlag, 1999.

15 Josée Desharnais, Vineet Gupta, Radhakrishnan Jagadeesan, and Prakash Panangaden. A metric for labelled Markov processes. *Theoretical Computer Science*, 318(3):323–354, June 2004.

16 Manfred Droste, Werner Kuich, and Heiko Vogler, editors. *Handbook of weighted automata*. EATCS Monographs on Theoretical Computer Science. Springer, 2009.

17 Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, pages 162–169. AUAI Press, 2004.

18 Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for Markov decision processes with infinite state spaces. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 201–208, July 2005.

19 Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

20 A. Giacalone, C. Jou, and S. Smolka. Algebraic reasoning for probabilistic concurrent systems. In *Proceedings of the Working Conference on Programming Concepts and Methods*, IFIP TC2, 1990.

21 Hadrien Glaude and Olivier Pietquin. Pac learning of probabilistic automaton based on the method of moments. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 820–829, 2016.

22 Christopher Heil and Gilbert Strang. Continuity of the joint spectral radius: application to wavelets. In *Linear Algebra for Signal Processing*, pages 51–61. Springer, 1995.

23 Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5), 2012.

24 Manfred Jaeger, Hua Mao, Kim Guldstrand Larsen, and Radu Mardare. Continuity properties of distances for Markov processes. In *Proceedings of QEST 2014 Quantitative Evaluation of Systems: 11th International Conference*, pages 297–312. Springer International Publishing, 2014.

25 Raphaël Jungers. *The joint spectral radius: theory and applications*, volume 385. Springer Science and Business Media, 2009.

26 Victor Kozyakin. An explicit lipschitz constant for the joint spectral radius. *Linear Algebra and its Applications*, 433(1):12–18, 2010.

27 Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1-2):5–34, 2003.

28 Prakash Panangaden. *Labelled Markov Processes*. Imperial College Press, 2009.

29 Gian-Carlo Rota and W. Strang. A note on the joint spectral radius. *Indag. Math.*, 22:379–381, 1960.

30 Marcel Paul Schützenberger. On the definition of a family of automata. *Information and control*, 4(2):245–270, 1961.

31 S. M. Siddiqi, B. Boots, and G. Gordon. Reduced-rank hidden Markov models. In *AISTATS*, 2010.

32 Franck van Breugel and James Worrell. Towards quantitative verification of probabilistic systems. In *Proceedings of the Twenty-eighth International Colloquium on Automata, Languages and Programming*. Springer-Verlag, July 2001.

33 Fabian Wirth. The generalized spectral radius and extremal norms. *Linear Algebra and its Applications*, 342(1-3):17–40, 2002.