Bisimulation Metrics and Norms for Real-Weighted Automata

Borja Balle* Pascale Gourdeau
University of Oxford
Prakash Panangaden
McGill University

17th November 2020

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.

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 [37]. In [5, 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 [5, 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 *behavioural metrics* which are crafted to capture behaviour. Bisimulation, or more precisely probabilistic bisimulation [31, 33], was defined to capture the notion of probabilistic processes with indistinguishable behaviour. It was soon realized that an equivalence relation

*Now at DeepMind. Based on work completed while at Lancaster University and Amazon Research Cambridge.

was too unstable a concept in a quantitative setting and bisimulation pseudometrics were invented [14] to provide a concept appropriate for quantitative settings. In particular, with a behavioural pseudometric we recover bisimulation as the kernel. Such behavioural metrics for Markov processes were proposed by Giacalone et al. [23] and the first successful pseudometric that has bisimulation as its kernel is due to Desharnais et al. [14, 15]; see [33] for an expository account. The subject was greatly developed by van Breugel and Worrell [39] among others.

For WFA, a beautiful treatment of linear bisimulation relations was given by Boreale in [11] and by Bonchi et al. in its subsequent journal version [9]. We were motivated to develop a metric analogue of Boreale's linear bisimulation with the eventual goal of using it to analyze approximate minimization. As it turns out, our treatment of norms and metrics in the present paper is not well adapted to the spectral algorithm of [5, 6] but other interesting connections emerged in the present work. In the present paper we develop the general theory of bisimulation (pseudo)metrics for WFA (and for weighted languages).

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. [28] 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 7).

6. We investigated the connection between our methods and previous bisimulation metrics for probabilistic automata, thus establishing a number of relations between our metric and the bisimulation metric of Feng and Zhang [17] (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 [9, 11]. We have chosen to work with WFA defined over fields rather than over semi-rings. This is a limitation but it still includes all the applications to probabilistic situations where one naturally

works with the real numbers. The examples that are ruled out are situations where one is interested in combinatorial applications. By staying with fields and vector spaces we have many basic properties: existence of a basis, spectral theory results and other crucial mathematical features that are lost in the semi-ring setting.

2 Background

In this section we recall preliminary definitions and results that will be used throughout the rest of the paper. 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, which will play an important technical role in the remainder of the paper.

2.1 Norms, Seminorms, and Pseudometrics

Seminorms (resp. pseudometrics) are generalizations of norms (resp. metrics) often used in analysis. The key difference is that seminorms (resp. pseudometrics) are allowed to assign zero value to non-zero vectors (resp. zero distance to pairs of distinct vectors). This section recalls their definitions and main properties.

A seminorm s on a vector space V is a function $s : V \to \mathbb{R}$ satisfying the following two axioms:

1. (absolute homogeneity) $s(cv) = |c|s(v)$ for all $c \in \mathbb{R}$ and $v \in V$, and
2. (subadditivity) $s(u + v) \leq s(u) + s(v)$ for all $u, v \in V$.

Jointly, these two conditions imply $s(v) \geq 0$ for all $v \in V$. Furthermore, the first condition implies $s(0) = 0$, but unlike in the case of norms we do not require that 0 is the only vector with $s(v) = 0$. The kernel of a seminorm s is defined as $\ker(s) = \{v \in V : s(v) = 0\}$. Therefore, a seminorm s is a norm if and only if $\ker(s) = \{0\}$. It can be readily verified that $\ker(s)$ is always a linear subspace of V.

Given a finite-dimensional normed real vector space $(V, \|\cdot\|)$ we let V^* denote the dual vector space equipped with the dual norm $\|w\|_* = \sup_{\|v\|\leq 1} w(v)$ for any $w \in V^*$. The induced norm of a linear operator $\tau : V \to V$ is defined as $\|\tau\| = \sup_{\|v\|\leq 1} \|\tau(v)\|$. We recall that on a finite-dimensional vector space all norms are equivalent. Namely, given two norms $\|\cdot\|$ and $\|\cdot\|'$ on V there exists a pair of constants $0 < c \leq C$ such that $c\|v\| \leq \|v\|' \leq C\|v\|$ holds for all $v \in V$. It is immediate to check that the inequalities $C^{-1}\|w\|_* \leq \|w\|'_* \leq c^{-1}\|w\|_*$ hold for the corresponding dual norms.

A pseudometric on a set V is a function $d : V \times V \to \mathbb{R}$ satisfying the following axioms:

1. (non-negativity) $d(v, w) \geq 0$ for all $v, w \in V$,
2. (indiscernibility of identicals) $d(v, v) = 0$ for all $v \in V$,
3. (symmetry) $d(v, w) = d(w, v)$ for all $v, w \in V$, and
4. (triangle inequality) $d(v, u) \leq d(v, w) + d(w, u)$ for all $u, v, w \in V$.

Note that the only difference between a metric and a pseudometric is that in the latter case we do not require that $d(v, w) = 0$ implies $v = w$. Therefore, a pseudometric might not be able to distinguish between every pair of points in V. Seminorms provide a convenient way to build pseudometrics: if V is a real vector space and $s : V \to \mathbb{R}$ is a seminorm on V, then $d(v, w) = s(v - w)$ is a pseudometric on V. We shall say that d is the pseudometric induced by s.

2.2 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 $x_{\leq 0} = \epsilon$, the empty string. 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: $\tau_{x_1 \cdots x_t} = \tau_{x_t} \circ \cdots \circ \tau_{x_1}$.

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 map 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 $x \in \Sigma^\star$.

2.3 Linear Bisimulations

Linear bisimulations for weighted automata were introduced by Boreale in [11] and by Bonchi et al. in [9], where its characterisation was given from a coalgebraic perspective. Here we recall the key definition and several important facts. 09

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 under the action of 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 [9, 11] 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, as presented in [10]. 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 [9, 11] 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.

2.4 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 [29] 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 [12, 13].

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: $\rho(M) \leq \sup_{i\in I} \|\tau_i\|$. 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 $\|\tau_i\| \leq \rho(M)$ 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 [40] 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. We remark that, unfortunately, the proof of this result is non-constructive.

Theorem 3 ([36])**.** *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 $\|\tau_i(v)\| \leq (\rho(M) + \eta)\|v\|$ for every $i \in I$ and every $v \in V$.*

The statement above is in fact a special case of Proposition 1 in [36]; 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 [26]). 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 under 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 Hausdorff distance. Furthermore, Wirth showed in [40] that the joint spectral radius is locally Lipschitz continuous around irreducible sets of matrices with respect to the Hausdorff topology (see also [30] 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 7. 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.

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 $w \neq 0$.*

Before proving the above theorem, we introduce the following characterizations of reachability and observability, which will be used in the proof.

Lemma 5. *Given a weighted automaton $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ the following hold:*

1. *A is observable if and only if $f_{A_v} \neq 0$ for all $v \in V \setminus \{0\}$.*
2. *A is reachable if and only if $f_{A^w} \neq 0$ for all $w \in V^* \setminus \{0\}$.*

Proof. To prove the first claim we note that if A is not observable then there exist two different states $u, v \in V$ such that $f_{A_u} = f_{A_v}$. Therefore, we see that $w = u - v \neq 0$ and A_w computes the function $f_{A_w} = f_{A_u} - f_{A_v} = 0$. On the other hand, if $v \in V \setminus \{0\}$ is such that $f_{A_v} = 0$, then A_v and A_0 compute the zero function and A is not observable.

The second claims follows from applying the first claim to the reverse automaton $\bar{A}$. □

Proof of Theorem 4. To prove the "only if" part assume that the set of linear maps $M = \{\tau_\sigma\}_{\sigma\in\Sigma}$ is reducible. Then there exists a non-trivial subspace $W \subset V$ that is left invariant by all the τ_σ. Using this subspace we can find a non-zero vector $v \in W$ and a non-zero linear form $w \in V^*$ such that $W \subseteq \ker(w)$. We claim that $A' = A_v^w$ is not minimal. Indeed, since W is invariant under the action of every τ_σ we have $\tau_x(v) \in W$ for all $x \in \Sigma^\star$, which implies $f_{A'}(x) = w(\tau_x(v)) = 0$ for all $x \in \Sigma^\star$. Therefore we have $f_{A'} = 0$ which is also computed by the weighted automaton A_0^w with initial weights $0 \in V$, so by Lemma 5 A' is not observable.

For the "if" part we assume that A_v^w is not minimal for some $v \in V \setminus \{0\}$ and $w \in V^* \setminus \{0\}$. Since A is irreducible if and only if $\bar{A}$ is irreducible, we can assume without loss of generality that A_v^w is not observable. Furthermore, by Lemma 5 we can further assume that (replacing v by a different state if necessary) A_v^w computes the zero function. Now let us take the subspace $W = span\{\tau_x(v) : x \in \Sigma^\star\} \subseteq V$ and show that it is a witness for the reducibility of M. Note that by construction we immediately have $\tau_\sigma(W) \subseteq W$ for any $\sigma \in \Sigma$, so we only need to check that W is not trivial. On the one hand we have $0 \neq v \in W$, so

$\dim(W) \geq 1$. On the other hand, since A_v^w computes the zero function we must have $W \subseteq \ker(w)$, which implies $\dim(W) \leq \dim(\ker(w)) = n - 1$ since w is not zero. □

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 [18] 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 2.3. 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 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 6. *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 $(\mathcal{S}, d)$. 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:

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

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, using the absolute homogeneity of s and s', it can be shown that:

$$\begin{aligned}
\left| \max_{\sigma} s(\tau_\sigma(v)) - \max_{\sigma'} s'(\tau_{\sigma'}(v)) \right| &= \max_{\sigma} s(\tau_\sigma(v)) - \max_{\sigma'} s'(\tau_{\sigma'}(v)) \\
&= s(\tau_{\sigma_*}(v)) - \max_{\sigma'} s'(\tau_{\sigma'}(v)) \\
&\leq s(\tau_{\sigma_*}(v)) - s'(\tau_{\sigma_*}(v)) \\
&= \|\tau_{\sigma_*}(v)\| \left(s\left(\frac{\tau_{\sigma_*}(v)}{\|\tau_{\sigma_*}(v)\|} \right) - s'\left(\frac{\tau_{\sigma_*}(v)}{\|\tau_{\sigma_*}(v)\|} \right) \right) \\
&\leq \|\tau_{\sigma_*}(v)\| \sup_{\|v'\| \leq 1} |s(v') - s'(v')| \\
&= \|\tau_{\sigma_*}(v)\| d(s, s') \ .
\end{aligned} \tag{3}$$

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 7. *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*

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

Proof of Theorem 7. For simplicity, let $F = F_{A,\gamma}$ and $s = s_{A,\gamma}$. In the first place we note that s clearly satisfies the seminorm axioms. However, this is not enough to guarantee that s is a seminorm because the supremum over Σ^∞ could be unbounded while the definition of seminorm requires the image by s of every element in V to be in $\mathbb{R}$. To guarantee that s is a seminorm we must show that $s(v)$ is always finite. Let $\|\cdot\|$ be the norm on V constructed in the proof of Theorem 6. Then we can use Hölder's inequality and the submultiplicativity of induced norms to show that for any $v \in V$ and $x \in \Sigma^\star$ we have

$$|\beta(\tau_x(v))| \leq \|\tau_x(v)\| \|\beta\|_* \leq (\rho(A) + \eta)^{|x|} \|v\| \|\beta\|_* \ ,$$

where $\eta = \delta/2$ for some $\delta > 0$ such that $\gamma \leq 1/(\rho(A) + \delta)$. Thus, for any $v \in V$ we can bound the expression in (4) as

$$s(v) \leq \|v\| \|\beta\|_* \sum_{t=0}^{\infty} \gamma^t (\rho(A) + \eta)^t \leq \|v\| \|\beta\|_* \sum_{t=0}^{\infty} \left(\frac{\rho(A) + \delta/2}{\rho(A) + \delta} \right)^t < \infty \ .$$

Now that we know that s is a seminorm and F has a unique fixed point in $\mathcal{S}$, we only need to verify that the expression in (4) is a fixed point of F. To see that this is the case we just note the following holds for any $v \in V$:

$$\begin{aligned} F(s)(v) &= |\beta(v)| + \gamma \max_{\sigma\in\Sigma} |s(\tau_\sigma(v))| \\ &= |\beta(v)| + \gamma \max_{\sigma\in\Sigma} \left| \sup_{x\in\Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t |\beta(\tau_{x_{\leq t}}(\tau_\sigma(v)))| \right| \\ &= |\beta(v)| + \max_{\sigma\in\Sigma} \sup_{x\in\Sigma^\infty} \sum_{t=0}^{\infty} \gamma^{t+1} |\beta(\tau_{(\sigma x)_{\leq t+1}}(v))| \\ &= |\beta(v)| + \sup_{x\in\Sigma^\infty} \sum_{t=1}^{\infty} \gamma^t |\beta(\tau_{x_{\leq t}}(v))| \\ &= s(v) \ . \end{aligned}$$

Finally, note that the second equality follows from the identity $|\beta(\tau_y(v))| = f_{A_v}(y)$ for all $y \in \Sigma^\star$. □

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 8. *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 9. *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 7,

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

Therefore $\ker(s) \subseteq \ker(\beta)$. To verify the invariance of $\ker(s)$ under all τ_σ let

$v \in \ker(s)$ and note that using $\beta(v) = 0$ we can write

$$\begin{aligned}
0 \leq s(\tau_\sigma(v)) &= \sup_{x\in\Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t |\beta(\tau_{x_{\leq t}}(\tau_\sigma(v)))| \\
&= \sup_{x\in\Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t |\beta(\tau_{(\sigma x)_{\leq t+1}}(v))| \\
&= \frac{1}{\gamma} \sup_{x\in\Sigma^\infty} \sum_{t=0}^{\infty} \gamma^{t+1} |\beta(\tau_{(\sigma x)_{\leq t+1}}(v))| \\
&\leq \frac{1}{\gamma} \sup_{x\in\Sigma^\infty} \sum_{t=1}^{\infty} \gamma^t |\beta(\tau_{x_{\leq t}}(v))| \\
&= \frac{1}{\gamma} \left(|\beta(v)| + \sup_{x\in\Sigma^\infty} \sum_{t=1}^{\infty} \gamma^t |\beta(\tau_{x_{\leq t}}(v))| \right) \\
&= \frac{1}{\gamma} s(v) = 0 \ .
\end{aligned}$$

This implies $\tau_\sigma(v) \in \ker(s)$ for all $v \in \ker(s)$ and $\sigma \in \Sigma$. 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 7 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 now show that when A is observable then every γ-bisimulation seminorm is in fact a norm.

Corollary 10. *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 9 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 10, 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 11. *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 7 we can provide a closed form expression for d_γ.

Corollary 12. *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*

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

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 $\gamma < 1/\rho(A)$. Equivalently, we see that d_γ can be interpreted as the metric

$$d_\gamma(f_1, f_2) = \sup_{x \in \Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t \left| f_1(x_{\leq t}) - f_2(x_{\leq t}) \right|$$

on the set of weighted languages

$$\left\{ f : \Sigma^\star \to \mathbb{R} \ : \ \sup_{x \in \Sigma^\infty} \sum_{t=0}^{\infty} \gamma^t |f(x_{\leq t})| < \infty \right\} \quad .$$

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 [28], 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, 20, 19]; see Section 7 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 13. *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 14. *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 15. *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 17). This result also plays a prominent result in Section 7.

2. Several topological properties of the joint spectral radius discussed in Section 2.4.

We first state an elementary lemma that we need in order to prove an upper bound on d_γ. This estimate also plays an important role in the application of our bismulation pseudometric to spectral learning presented in Section 7.

Lemma 16. *Let $(s_l)_{l\in\mathbb{N}}$ be a sequence such that there exists a constant a and a sequence $(b_l)_{l\in\mathbb{N}}$ satisfying $s_{l+1} \leq as_l + b_l$ for all $l \geq 0$. Then for all $l \geq 0$ we have $s_{l+1} \leq a^{l+1}s_0 + \sum_{i=0}^{l} a^{l-i}b_i$.*

Proof. Simple proof by induction on l. □

Lemma 17. *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ and $A' = \langle \Sigma, V, \alpha', \beta', \{\tau'_\sigma\}_{\sigma\in\Sigma}\rangle$ be two weighted automata over the same alphabet Σ and the same vector space V. Let $M = \{\tau_\sigma\} \cup \{\tau'_\sigma\}$ and $\rho = \rho(M)$. Suppose $\gamma < 1/\rho$ and $\|\cdot\|$ is a norm on V such that for all $\sigma \in \Sigma$ we have $\|\tau_\sigma\|, \|\tau'_\sigma\| \leq \theta$ for some θ such that $\nu = \gamma\theta < 1$. Then we have the following:*

$$d_\gamma(A, A') \leq \frac{\|\alpha\|\|\beta - \beta'\|_* + \|\beta'\|_*\|\alpha - \alpha'\|}{1-\nu} + \frac{\gamma\|\alpha\|\|\beta'\|_* \max_\sigma \|\tau_\sigma - \tau'_\sigma\|}{(1-\nu)^2} \quad . \quad (6)$$

Proof. Fix $x \in \Sigma^\infty$ and given $l \geq 0$ define $D_l(x) = \sum_{t=0}^{l} \gamma^t |f_A(x_{\leq t}) - f_{A'}(x_{\leq t})|$. By applying the triangle and Hölder inequalities to any term in the summation $D_l(x)$ we get

$$|f_A(x_{\leq t}) - f_{A'}(x_{\leq t})| \leq \|\beta - \beta'\|_*\|\tau_{x_{\leq t}}(\alpha)\| + \|\beta'\|_*\|\tau_{x_{\leq t}}(\alpha) - \tau'_{x_{\leq t}}(\alpha')\| \quad . \quad (7)$$

Using the assumption on $\|\cdot\|$ we can see that $\|\tau_{x_{\leq t}}(\alpha)\| \leq \theta^t\|\alpha\|$ for any $t \geq 0$. Now let $\varepsilon_\beta = \|\beta - \beta'\|_*$ and $\varepsilon_t = \|\tau_{x_{\leq t}}(\alpha) - \tau'_{x_{\leq t}}(\alpha')\|$. Plugging these definitions and the bound (7) in D_l we get

$$D_l(x) \leq \varepsilon_\beta\|\alpha\| \left(\sum_{t=0}^{l} \gamma^t\theta^t\right) + \|\beta'\|_* \left(\sum_{t=0}^{l} \gamma^t\varepsilon_t\right) \quad . \quad (8)$$

Now we shall bound the term $s_l = \sum_{t=0}^{l} \gamma^t \varepsilon_t$. Suppose $x_{\leq t+1} = y\sigma$, where $y \in \Sigma^t$ and $\sigma \in \Sigma$. Let $\varepsilon_\tau = \max_\sigma \|\tau_\sigma - \tau'_\sigma\|$. Using the triangle inequality we can show the following:

$$\begin{aligned}
\varepsilon_{t+1} &= \|\tau_{y\sigma}(\alpha) - \tau'_{y\sigma}(\alpha')\| \\
&= \|\tau_\sigma(\tau_y(\alpha)) - \tau'_\sigma(\tau'_y(\alpha'))\| \\
&\leq \|\tau_\sigma(\tau_y(\alpha)) - \tau'_\sigma(\tau_y(\alpha))\| + \|\tau'_\sigma(\tau_y(\alpha) - \tau'_y(\alpha'))\| \\
&\leq \|\tau_\sigma - \tau'_\sigma\|\|\tau_y(\alpha)\| + \|\tau'_\sigma\|\|\tau_y(\alpha) - \tau'_y(\alpha')\| \\
&\leq \varepsilon_\tau \theta^t \|\alpha\| + \theta \varepsilon_t \ .
\end{aligned}$$

We will now use the inequality above to show that s_l satisfies a recurrence of the form considered in Lemma 16 for all $l \geq 0$:

$$\begin{aligned}
s_{l+1} &= \varepsilon_0 + \sum_{t=1}^{l+1} \gamma^t \varepsilon_t \\
&= \varepsilon_0 + \gamma \sum_{t=0}^{l} \gamma^t \varepsilon_{t+1} \\
&\leq \varepsilon_0 + \gamma \sum_{t=0}^{l} \gamma^t \left(\varepsilon_\tau \theta^t \|\alpha\| + \theta \varepsilon_t\right) \\
&= \gamma\theta s_l + \varepsilon_0 + \gamma \varepsilon_\tau \|\alpha\| \sum_{t=0}^{l} (\gamma\theta)^t \ .
\end{aligned}$$

Let $\varepsilon_\alpha = \|\alpha - \alpha'\|$ and note that $s_0 = \varepsilon_0 = \varepsilon_\alpha$. Thus, applying Lemma 16 with $a = \gamma\theta$ and $b_l = \varepsilon_\alpha + \gamma\varepsilon_\tau\|\alpha\| \sum_{t=0}^{l} (\gamma\theta)^t$ to the sequence s_l we get:

$$\begin{aligned}
s_l &\leq (\gamma\theta)^l \varepsilon_\alpha + \sum_{i=0}^{l-1} (\gamma\theta)^{l-1-i} \left(\varepsilon_\alpha + \gamma\varepsilon_\tau\|\alpha\| \sum_{t=0}^{i} (\gamma\theta)^t\right) \\
&= \varepsilon_\alpha \sum_{t=0}^{l} (\gamma\theta)^t + \gamma\varepsilon_\tau\|\alpha\| \sum_{i=0}^{l-1} \left((\gamma\theta)^{l-1-i} \sum_{t=0}^{i} (\gamma\theta)^t\right) \\
&= \varepsilon_\alpha \frac{1 - (\gamma\theta)^{l+1}}{1 - \gamma\theta} + \frac{\gamma\varepsilon_\tau\|\alpha\|}{1 - \gamma\theta} \sum_{i=0}^{l-1} \left((\gamma\theta)^{l-1-i} - (\gamma\theta)^l\right) \\
&= \varepsilon_\alpha \frac{1 - (\gamma\theta)^{l+1}}{1 - \gamma\theta} + \frac{\gamma\varepsilon_\tau\|\alpha\|}{1 - \gamma\theta} \left(\frac{1 - (\gamma\theta)^l}{1 - \gamma\theta} - l(\gamma\theta)^l\right) \\
&= \frac{\varepsilon_\alpha}{1 - \gamma\theta} + \frac{\gamma\varepsilon_\tau\|\alpha\|}{(1 - \gamma\theta)^2} - (\gamma\theta)^l \left(\frac{\varepsilon_\alpha\gamma\theta + l\gamma\varepsilon_\tau\|\alpha\|}{1 - \gamma\theta} + \frac{\gamma\varepsilon_\tau\|\alpha\|}{(1 - \gamma\theta)^2}\right) \ .
\end{aligned}$$

Plugging this bound into (8) and grouping the terms multiplied by $(\gamma\theta)^l$ into R_l we get

$$D_l(x) \leq \frac{\varepsilon_\beta\|\alpha\| + \varepsilon_\alpha\|\beta'\|_*}{1 - \gamma\theta} + \frac{\gamma\varepsilon_\tau\|\alpha\|\|\beta'\|_*}{(1 - \gamma\theta)^2} - (\gamma\theta)^l R_l \ . \tag{9}$$

Finally, observing that $R_l = O(l)$ and using that $\gamma\theta = \nu < 1$, we take the limit $l \to \infty$ and obtain the desired bound using the closed form expression for $d_\gamma(A, A')$ given in Corollary 12. □

Now we proceed to the proof of Theorem 15. The main ingredient of this proof is the construction of a norm on V satisfying the conditions of Lemma 17 uniformly for all A_i with $i \geq j_0$ for some $j_0 \in \mathbb{N}$.

Proof of Theorem 15. Let $A_i = \langle \Sigma, V, \alpha_i, \beta_i, \{\tau_{i,\sigma}\}_{\sigma\in\Sigma}\rangle$ be a sequence of weighted automata converging to $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ with respect to some norm $\|\cdot\|$ on V and suppose $\gamma < 1/\rho(A)$. For any $j \in \mathbb{N}$ we define the set

$$M_j = \{\tau_\sigma\}_{\sigma\in\Sigma} \cup \bigcup_{i\geq j} \{\tau_{i,\sigma}\}_{\sigma\in\Sigma} \ .$$

Since $\lim_{i\to\infty} \tau_{i,\sigma} = \tau_\sigma$ for all $\sigma \in \Sigma$, the set M_j is bounded for all $j \in \mathbb{N}$. Let $\rho_j = \rho(M_j) = \rho(\bar{M}_j)$, where $\bar{M}_j$ is the compact set obtained as the closure of M_j. Using the continuity of the joint spectral radius on compact sets of operators we see that $\lim_{j\to\infty} \rho_j = \rho(A)$. Thus, letting $\delta = 1-\gamma\rho(A) > 0$, there exists a constant $j_0 \in \mathbb{N}$ such that $|\rho_j - \rho(A)| < \delta/(4\gamma)$ is satisfied for all $j \geq j_0$. Now we can apply Theorem 3 to $\bar{M}_{j_0}$ with $\eta = \delta/(4\gamma)$ to find a norm $\|\cdot\|'$ on V such that $\|\tau_\sigma\|' \leq \rho(A)+\delta/(2\gamma)$ and $\|\tau_{i,\sigma}\|' \leq \rho(A)+\delta/(2\gamma)$ for all $\sigma \in \Sigma$ and all $i \geq j_0$. Taking $\theta = \rho(A)+\delta/(2\gamma)$ we see that $\gamma\theta = \gamma\rho(A)+\delta/2 < \gamma\rho(A)+\delta = 1$. Hence, we are under the hypotheses of Lemma 17 and we have that the following holds for all $i \geq j_0$:

$$d_\gamma(A, A_i) \leq \frac{\|\alpha\|'\|\beta-\beta_i\|'_* + \|\beta_i\|'_*\|\alpha-\alpha_i\|'}{1-\nu} + \frac{\gamma\|\alpha\|'\|\beta_i\|'_* \max_\sigma \|\tau_\sigma - \tau_{i,\sigma}\|'}{(1-\nu)^2} \ , \tag{10}$$

where $\nu = \gamma\theta = \gamma\rho(A) + \delta/2$.

Now recall that all norms in a finite dimensional vector space are equivalent. Therefore, we can find a pair constants $0 < c \leq C$ such that $c\|v\| \leq \|v\|' \leq C\|v\|$ holds for all $v \in V$ and $C^{-1}\|w\|_* \leq \|w\|'_* \leq c^{-1}\|w\|_*$ for all $w \in V^*$. Plugging these inequalities in (10) we see that for all $i \geq j_0$ we have

$$d_\gamma(A, A_i) \leq \frac{C(\|\alpha\|\|\beta-\beta_i\|_* + \|\beta_i\|_*\|\alpha-\alpha_i\|)}{c(1-\nu)} + \frac{C^2\gamma\|\alpha\|\|\beta_i\|_* \max_\sigma \|\tau_\sigma - \tau_{i,\sigma}\|}{c(1-\nu)^2} \ .$$

Since the sequence of automata (A_i) converges to A with respect to $\|\cdot\|$, we conclude that $\lim_{i\to\infty} d_\gamma(A, A_i) = 0$. □

4.2 Input Continuity

Inspired by the notion of property continuity presented in [28], we define a new notion of input g-continuity encapsulating 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 18. *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*

$$\lim_{i\to\infty} \sup_{x\in\Sigma^\star} \frac{|f_A(x) - f_{A_i}(x)|}{g(|x|)} = 0 \ . \tag{11}$$

Figure 1: Two weighted automata with $\Sigma = \{a\}$ and initial weight $\alpha = 1$.

Note the special case $g(l) = 1$ is tightly related to the notion of property continuity presented in [28]. 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 in the numerator of (11). 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 (11) stays bounded. The next two results show that essentially $g(|x|) = \gamma^{-|x|}$ is the threshold between input continuity and input non-continuity in our γ-bisimulation pseudometrics.

Theorem 19. *The pseudometric d_γ from Definition 11 is input g-continuous for any $g(l) = \Omega(\gamma^{-l})$.*

Proof. Let A be weighted automaton such that $\gamma < 1/\rho(A)$ and let $(A_i)_{i\in\mathbb{N}}$ be a sequence of weighted automata converging to A with respect to d_γ. Note that for any $i \in \mathbb{N}$ we have the following:

$$\begin{aligned}\sup_{x\in\Sigma^\star} \frac{|f_A(x) - f_{A_i}(x)|}{g(|x|)} &= \sup_{x\in\Sigma^\star} \frac{|f_A(x) - f_{A_i}(x)|\gamma^{|x|}}{g(|x|)\gamma^{|x|}} \\ &\leq \sup_{x\in\Sigma^\star} \frac{d_\gamma(A, A_i)}{g(|x|)\gamma^{|x|}} \\ &= \sup_{l\in\mathbb{N}} \frac{d_\gamma(A, A_i)}{g(l)\gamma^l} \ .\end{aligned}$$

Now note that $g(l) > 0$ and $g(l) = \Omega(\gamma^{-l})$ implies $\inf_{l\in\mathbb{N}} g(l)\gamma^l > 0$. Using the assumption that $\lim_{i\to\infty} d_\gamma(A, A_i) = 0$ we now see that (11) is satisfied. □

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 [28].

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 20 provides simple examples of cases where d_γ is not input g-continuous.

Theorem 20. *Let $0 < \gamma < 1$. The pseudometric d_γ from Definition 11 is not input g-continuous for any $g(l) = c^{o(l)}$ with $c > 1$.*

Proof. Let $\Sigma = \{a\}$ be an alphabet with one symbol and let $A_i = \langle \Sigma, V, \alpha, \beta, \tau_i\rangle$ with $\tau_i = 1 + 2^{-i}$ and $\alpha = \beta = 1$ be the weighted automaton shown on the left of Figure 1, and let $A = \langle \Sigma, V, \alpha, \beta, \tau\rangle$ with $\tau = 1$ be the weighted automaton

shown on the right of Figure 1. For any $i > \log_2(\gamma/(1-\gamma))$ we have $\gamma\tau_i < 1$. Hence, we can write

$$\begin{aligned} d_\gamma(A, A_i) &= \sup_{x\in\Sigma^\infty} \sum_{t\geq 0} \gamma^t |\tau^t - \tau_i^t| \\ &= \sum_{t\geq 0} \gamma^t \left((1+2^{-i})^t - 1\right) \\ &= \frac{1}{1-\gamma(1+2^{-i})} - \frac{1}{1-\gamma} \ . \end{aligned}$$

Therefore we see that $\lim_{i\to\infty} d_\gamma(A, A_i) = 0$. Now let us show that for these automata the limit in (11) is not zero for any $g(l) = c^{o(l)}$ with $c > 1$. Indeed, we can write

$$\begin{aligned} \sup_{x\in\Sigma^\star} \frac{|f_A(x) - f_{A_i}(x)|}{g(|x|)} &= \sup_{x\in\Sigma^\star} \frac{(1+2^{-i})^{|x|} - 1}{c^{o(|x|)}} \\ &= \sup_{l\in\mathbb{N}} \frac{(1+2^{-i})^l - 1}{c^{o(l)}} \\ &\geq \sup_{l\in\mathbb{N}} \frac{(1+2^{-i})^l}{c^{o(l)}} - \sup_{l\in\mathbb{N}} \frac{1}{c^{o(l)}} \\ &= \infty \ , \end{aligned}$$

where the last equality uses that $\frac{(1+2^{-i})^l}{c^{o(l)}} = \omega(1)$ and $\frac{1}{c^{o(l)}} = O(1)$ with respect to $l \to \infty$. Therefore d_γ is not input g-continuous for these choices of g. □

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 $\mathcal{O}$. 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 [32], 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:

$$V_U(x) = \sum_{t=1}^{\infty} \gamma^{t-1} \alpha^\top T_{x_{\leq t-1}} \beta_{x_t} \ , \tag{12}$$

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 [32].

Theorem 21 (Theorem 4.4 in [32])**.** *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 [32] 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 22. *The problem in Theorem 21 remains undecidable when restricted to UMDP with non-negative action-independent rewards.*

Theorem 23. *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 $x \in \Sigma^\infty$:

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

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 A Coarser Bisimulation Metric

In Section 3, we constructed bisimulation pseudo-metrics via a the following operator on seminorms:

$$F_{A,\gamma}(s)(v) = |\beta(v)| + \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(v)) \ . \tag{13}$$

Obviously, this operator was chosen with a specific goal in mind: constructing a pseudometric that captures bisimulation in its kernel. But this is not the only choice of operator from which one can obtain bisimulation pseudometrics. We show here that a slight modification of this operator yields a coarser pseudometric, and show how this new construction is related to a bisimulation pseudometric for probabilistic automata proposed by Feng and Zhang [17].

Letting $F^+ = F^+_{A,\gamma}$ be the operator defined in (13), we denote by $F^\vee = F^\vee_{A,\gamma}$ the operator obtained by replacing the sum by a maximum:

$$F^\vee_{A,\gamma}(s)(v) := \max\left\{|\beta(v)|, \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(v))\right\} \ .$$

This new operator $F^\vee$ shares a number of properties with the operator F^+. For example, invoking a similar argument as in Section 3, we can obtain a unique fixed point, as well as a closed-form expression for it whenever $\gamma < 1/\rho(A)$. First, we check again that $F^\vee(s)$ is a seminorm for $s \in \mathcal{S}$. Absolute homogeneity is immediate, while for subadditivity we have that

$$\begin{aligned}
F^\vee(s)(u+v) &= \max\left\{|\beta(u+v)|, \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(u+v))\right\} \\
&= \max\left\{|\beta(u)+\beta(v)|, \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(u)+\tau_\sigma(v))\right\} \\
&\leq \max\left\{|\beta(u)|+|\beta(v)|, \gamma\left(\max_{\sigma\in\Sigma} s(\tau_\sigma(u)) + \max_{\sigma'\in\Sigma} s(\tau_{\sigma'}(v))\right)\right\} \\
&\leq \max\left\{|\beta(u)|, \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(u))\right\} + \max\left\{|\beta(v)|, \gamma \max_{\sigma\in\Sigma} s(\tau_\sigma(v))\right\} \\
&\leq F^\vee(s)(u) + F^\vee(s)(v) \ .
\end{aligned}$$

Next, we show that $F^\vee$ has a unique fixed point, provided γ is smaller than the critical value $1/\rho(A)$, thus obtaining an analog of Theorem 6 for the new operator.

Theorem 24. *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$. If $\gamma < 1/\rho(A)$, then $F^\vee_{A,\gamma}$ has a unique fixed point.*

Proof. For simplicity, let $F^\vee = F^\vee_{A,\gamma}$. Let $\|\cdot\|$ and d be as in Theorem 6. We only need to show the contractivity of $F^\vee$ in the metric space (S, d). To see that $F^\vee$ is indeed a contraction, we note that:

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

To show that $d(F(s), F(s')) < d(s, s')$, the remaining of the argument is exactly the same as in Theorem 6. □

Similarly, we can obtain a closed-form expression for the unique fixed point defined via $F^\vee$. The argument is again nearly identical to the proof for the closed form of the fixed point defined via the operator F^+, and is omitted here.

Theorem 25. *Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$. Suppose $\gamma < 1/\rho(A)$ and let $s^\vee_{A,\gamma} \in \mathcal{S}$ be the fixed point of $F^\vee_{A,\gamma}$. Then for any $v \in V$ we have*

$$s^\vee_{A,\gamma}(v) = \max_{x\in\Sigma^\star} \gamma^{|x|}|\beta(\tau_x(v))| = \max_{x\in\Sigma^\star} \gamma^{|x|}|f_{A_v}(x)| \ . \tag{14}$$

It is immediate to see from the closed forms defined in Equations (4) and (14) that $s^\vee_{A,\gamma}$ is a lower bound for $s^+_{A,\gamma}$. Similarly, given any two automata A_1, A_2 and defining $d^\vee_\gamma(A_1, A_2)$ as $s^\vee_{A,\gamma}(\alpha)$ on the difference automaton $A = A_1 - A_2$, we have that $d^\vee_\gamma(A_1, A_2)$ is a lower bound for $d^+_\gamma(A_1, A_2)$, provided $\gamma < 1/\rho(A)$. Then, $s^\vee_{A,\gamma}$ and $d^\vee_\gamma$ are bisimulation seminorms and pseudometrics, respectively. Consequently, the continuity properties defined in Section 4 that hold for d^+_γ also hold for $d^\vee_\gamma$. Moreover, by a simple reduction to the emptiness of stochastic languages [24, 34], we can see that the threshold problem for the seminorm $s^\vee_{A,\gamma}(\alpha)$ is also undecidable, obtaining an analog of Theorem 23 for the new seminorm.

6.1 Comparison with Probabilistic Bisimulation Metrics

Now we can show that when restricted to probabilistic automata (in the sense of Rabin) interpreted as WFA, the pseudometric $d^\vee_\gamma$ coincides with the bisimulation pseudometric d^{FZ}_γ developed by Feng and Zhang in [17]. We start by recalling that a *probabilistic automaton* (PA) in the sense of Rabin (also sometimes known as a *reactive* probabilistic automaton) is defined by taking $P = \langle \Sigma, Q, \alpha, B, \{T_\sigma\}_{\sigma\in\Sigma}\rangle$, 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, $B \subseteq Q$ is the set of accepting states, and $T_\sigma : Q \times Q \to [0, 1]$ 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). The automaton P defines a language $f_P : \Sigma^\star \to [0, 1]$ where $f_P(x)$ is the probability of reaching an accepting state when the initial state q_0 is sampled according to α and for each symbol x_i in x a next state q_i is sampled from the distribution induced by $T_{x_i}(q_{i-1}, \cdot)$. One can construct a WFA $A_P = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma\in\Sigma}\rangle$ recognizing the same language by taking $V = \mathbb{R}^Q$, interpreting α as a vector in $[0, 1]^Q \subseteq \mathbb{R}^Q$, letting $\beta \in \{0, 1\}^Q$ with $\beta(q) = 1$ if and only if $q \in B$, and representing $\tau_\sigma : \mathbb{R}^Q \to \mathbb{R}^Q$ by the matrix with entries $\tau_\sigma(q, q') = T_\sigma(q, q')$. An immediate calculation shows that indeed $f_{A_P} = f_P$.

Equipped with this conversion from PA to WFA we can establish the following equivalence between $d^\vee$ and the probabilitistic bisimulation pseudometric d^{FZ}_γ. This result shows we can interpret the bisimulation metric $d^\vee$ as a generalization for WFA of the probabilitistic bisimulation pseudometric d^{FZ}_γ.

Theorem 26. *For any probabilistic automata P_1 and P_2 over the same alphabet and any $\gamma < 1$ we have $d^{FZ}_\gamma(P_1, P_2) = d^\vee(A_{P_1}, A_{P_2})$.*

Proof. Follows immediately by comparing the definition of our operator $F^\vee$ with the operator that Feng and Zhang define in [17, Definition 14] (when specialized to Rabin probabilistic automata). □

From the point of view of the bisimulation metric d_γ^{FZ}, our results have two interesting consequences in the case $\gamma < 1$. First, since d^+ is an upper bound for $d^\vee$, the continuity properties proved in Section 4 are immediately inherited by d^{FZ}. Furthermore, the uniqueness of the fixed-point for $F^\vee$ implies that the pseudometric d^{FZ} obtained as a least fixed-point of an operator on a complete lattice by virtue of the Knaster–Tarski fixed-point theorem is in fact unique.

In view of the results in [17], one can also ask what happens at the critical value $\gamma = 1$ for probabilistic automata, and, more generally, $\gamma = 1/\rho$. This setting is a priori excluded by our results, while it is allowed in the theory of Feng and Zhang in the case of probabilistic automata. At this critical value the arguments we used in Theorem 6 to prove the contractivity of $F^\vee$ no longer work. However, the argument in [17] still yields a bisimulation pseudometric as a least fixed-point of a certain operator. Interestingly, for the case of $F^\vee$ one can show that even at the critical value $\gamma = 1/\rho(A)$ the seminorm

$$s^\vee_{A,\gamma}(v) = \sup_{x \in \Sigma^\star} \gamma^{|x|} |f_{A_v}(x)| \tag{15}$$

is again a fixed-point of $F^\vee_{A,\gamma}$, although this fixed-point is no longer guaranteed to be unique. By comparing the pseudometric $d^\vee_\gamma$ obtained from this fixed-point of $F^\vee$ with the closed-form expression for the pseudometric d^{FZ} at the critical value $\gamma = 1$ given in [17, Proposition 1] we see that we have identified the fixed-point that is also obtained by Knaster–Tarski.

To illustrate that the strict inequality $\gamma < 1/\rho$ we used throughout the paper is not the result of a technical artifact, we now prove that, in general, the pseudometric $d^\vee_\gamma$ obtained at the critical value is not parameter continuous. In particular, this shows that the behavior at the critical value $\gamma = 1/\rho$ is qualitatively different from the behavior in the regime $\gamma < 1/\rho$.

Theorem 27. *There exists a WFA A and a sequence of WFA $(A_i)_{i \geq 1}$ such that $\rho(A) = \rho(A_i) = 1$, (A_i) converges to A, but $\lim_{i \to \infty} d_1^\vee(A, A_i) \neq 0$.*

Proof. Let $\Sigma = \{a\}$ be an alphabet with one symbol and let $A = \langle \Sigma, V, \alpha, \beta, \tau \rangle$ with $\alpha = [1, 0]$, $\beta = [1, 1]$ and $\tau = I$ being the identity matrix. For $i \geq 1$ define the WFA $A_i = \langle \Sigma, V, \alpha, \beta, \tau_i \rangle$ with

$$\tau_i = \begin{bmatrix} \cos(\theta_i) & \sin(\theta_i) \\ -\sin(\theta_i) & \cos(\theta_i) \end{bmatrix} ,$$

where $\theta_i = \pi/i$. Since $\lim_{i \to \infty} \cos(\theta_i) = 1$ and $\lim_{i \to \infty} \sin(\theta_i) = 0$, we have $\lim_{i \to \infty} \|\tau - \tau_i\| = 0$ so (A_i) converges to A. We also note that the spectral radius of A and A_i satisfy $\rho(A) = \rho(A_i) = 1$ for all i.

Now we shall show that $d_1^\vee(A, A_i) \geq 2$ for all $i \geq 1$, thus implying that $d^\vee_\gamma$ is not parameter continuous at the critical point $\gamma = 1/\max\{\rho(A), \rho(A_i)\}$. To see this we first note that a standard trigonometric calculation yields the following identity for all $l \in \mathbb{N}$:

$$\tau_i^l = \begin{bmatrix} \cos(l\theta_i) & \sin(l\theta_i) \\ -\sin(l\theta_i) & \cos(l\theta_i) \end{bmatrix} .$$

In particular, taking $l = i$ we get $\tau_i^i = -I$. Thus, for every i we have $f_{A_i}(a^i) = -1$, while $f_A(a^i) = 1$. This shows that for any $i \geq 1$ we have

$$d_1^\vee(A, A_i) = \sup_{x \in \Sigma^\star} |f_{A_i}(x) - f_A(x)| \geq 2 .$$

Thus $\lim_{i\to\infty} d_1^{\vee}(A, A_i) \neq 0$. $\square$

7 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 [4] 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. In particular, it is interesting to see how – despite being unable to compute the γ-bisimulation pseudometric between two WFA (cf. Section 5) – we can still obtain bounds showing that algorithms for learning WFA from data will produce hypothesis that approximate an unknown WFA in terms of our bisimulation pseudometric.

7.1 Background on Spectral Learning

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 $\hat{A}$. Here we provide a brief outline of how the algorithm works. We refer the reader to [4, 3] for further details about the rationale behind this spectral learning algorithm, which rests on a powerful theorem by Fliess about the connection between WFA and the rank of infinite Hankel matrices [21].

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 rows and columns indexed by finite strings and 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. Fliess theorem [21] states that the rank of H_f is exactly the number of states of a minimal WFA computing f. In particular, the infinite matrix H_f has finite rank if and only if f can be computed by a WFA. The "only if" part of this theorem is constructive and can be interpreted as an algorithm to recover a WFA computing f from the Hankel matrix H_f. The main idea behind spectral learning is to turn this proof strategy into an efficient algorithm that works with a finite sub-block H of H_f and is robust to small perturbation of H. The success of this strategy is contingent on the sub-block H containing enough information to recover a WFA for f, and the perturbation on H being small enough. To ensure the sub-block H contains enough information one needs to impose two conditions, one of which is syntactic and the other is

algebraic. We will now give a formal description of these conditions and then proceed to sketch the main steps in the spectral learning algorithm.

A finite sub-block of an infinite Hankel matrix H_f can be obtained all the entries corresponding to a given set of rows (ie. prefixes) $P \subset \Sigma^\star$ and columns (ie. suffixes) $S \subset \Sigma^\star$. We write $H \in \mathbb{R}^{P \times S}$ to denote the resulting sub-block, whose entries are obtained from H_f in the natural way: $H(p,s) = H_f(p,s)$. The pair $B = (P,S)$ identifying the sub-block is usually called a *mask* and the success of the learning algorithm depends on this mask being a *complete basis*. Suppose H_f has finite rank. We say that B is a *basis* for H_f if the sub-block H corresponding to B satisfies $rank(H_f) = rank(H)$. In addition, we say that a basis $B = (P,S)$ is *complete* if there exists a set of prefixes $P' \subset P$ such that the following are satisfied: (i) $\epsilon \in P' \cap S$; (ii) $p \in P'$ implies $p\sigma \in P$ for all $\sigma \in \Sigma$; and (iii) $B' = (P', S)$ is a basis.

The *spectral learning* algorithm for WFA takes as input a sub-block H of H_f defined by a complete basis $B = (P,S)$ and proceeds as follows:

1. Find a set $P' \subset P$ witnessing that B is a complete basis.[1]
2. Take the sub-block $H_{B'}$ of H defined by B' and finds its singular value decomposition[2](SVD) $H_{B'} = LDR^\top$.
3. For each $\sigma \in \Sigma$ let H_σ be the sub-block of H corresponding to the mask[3] $B_\sigma = (P'\sigma, S)$ and let τ_σ be given by the matrix $T_\sigma = D^{-1}L^\top H_\sigma R$.
4. Define the masks $B_\alpha = (\{\epsilon\}, S)$ and $B_\beta = (P', \{\epsilon\})$ and identify α and β with the vectors $H_{B_\alpha}R$ and $D^{-1}LH_{B_\beta}$ respectively.

As was already mentioned above, one can view this algorithm as a proof of the "only if" part of Fliess' theorem. In particular, if $f = f_A$ for some WFA A, then the algorithm above will return a minimal WFA A' computing function f. We note that in general the algorithm works with a fixed basis representation for A', but if A is minimal then we know that A and A' are equivalent.

In the form given above, the algorithm requires an exact sub-block H for the Hankel matrix H_f. However, a slight modification of this algorithm can also learn from a perturbed version $\hat{H}$ of the sub-block H. Roughly speaking, the modification entails taking as input the desired number of states n in the target WFA and taking an *approximate* singular value decomposition of the sub-block $\hat{H}_{B'}$ of rank n. We shall call this version of the algorithm *robust spectral learning*; we refer the reader to [4, 3] for further details. In the sequel we focus on the analysis of the error in the output of the spectral learning algorithm under perturbations, and show how to provide learning guarantees in terms of our distance d_γ.

7.2 Bisimulation-based Learning Guarantees

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 $\hat{A}$ de-

[1] Note that because B is a basis, the condition of B' being a basis is equivalent to $rank(H) = rank(H_{B'})$, which can be efficiently checked.

[2] Any real matrix M of rank r admits an SVD $M = LDR^\top$ where $L^\top L = I$, $R^\top R = I$, and D is a diagonal matrix with entries $s_1 \geq \ldots \geq s_r > 0$ in the diagonal. This decomposition can be computed efficiently.

[3] Here $P'\sigma = \{p\sigma : p \in P'\}$.

pends on the error between the true and the approximated Hankel matrix as measured by the standard operator ℓ_2-norm.

Lemma 28. *Suppose H_f is a finite-rank infinite Hankel matrix and $B = (P, S)$ is a complete basis defining the sub-block $H \in \mathbb{R}^{P \times S}$. Let $A = \langle \Sigma, V, \alpha, \beta, \{\tau_\sigma\}_{\sigma \in \Sigma} \rangle$ be the WFA with $f_A = f$ produced by the spectral learning algorithm on input H. Let $\hat{H} \in \mathbb{R}^{P \times S}$ be another Hankel sub-block on the same mask B and let $\hat{A} = \langle \Sigma, V, \hat{\alpha}, \hat{\beta}, \{\hat{\tau}_\sigma\}_{\sigma \in \Sigma} \rangle$ be the output of robust spectral learning on inputs $\hat{H}$ and $n = rank(H_f)$. Then the following error estimate holds as $\|H - \hat{H}\|_2 \to 0$:*

$$\max\{\|\alpha - \hat{\alpha}\|, \|\beta - \hat{\beta}\|_*, \max_{\sigma \in \Sigma} \|\tau_\sigma - \hat{\tau}_\sigma\|\} \leq O(\|H - \hat{H}\|_2) \ ,$$

where 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. In the case of stochastic WFA[4], concentration of measure for random matrices can be used to show that as the amount of training data m increases then the distance between H and $\hat{H}$ converges to zero at a rate $O(1/\sqrt{m})$ with high probability (see e.g. [22]). Thus, Lemma 28 implies that as more training data becomes available, spectral learning will output a WFA $\hat{A}$ converging to A.

The last step in the usual analysis of spectral learning 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. In the learning literature, this step is usually provided for pseudo-metrics arising from truncated ℓ_1 norms (more details are provided at the end of this section). By invoking the parameter continuity of d_γ (Theorem 15) 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, which for some applications might be more appealing than the usual truncated ℓ_1 guarantees. But 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 29. *Suppose H_f is a finite-rank infinite Hankel matrix and $B = (P, S)$ is a complete basis defining the sub-block $H \in \mathbb{R}^{P \times S}$. Let H, A, $\hat{H}$ and $\hat{A}$ be as in Lemma 28. If A is irreducible, 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 $|\Sigma|$.*

Proof. Let $M = \{\tau_\sigma\}_{\sigma \in \Sigma}$ and let $\|\cdot\|$ be a norm on V obtained from Theorem 3 with M and a small enough constant $\eta > 0$. Let $\hat{M} = \{\tau_\sigma\}_{\sigma \in \Sigma} \cup \{\hat{\tau}_\sigma\}_{\sigma \in \Sigma}$. Let d_H denote the Hausdorff distance between sets of linear operators induced by $\|\cdot\|$. Since M is irreducible we can use the local Lipschitz continuity of the joint

[4] A WFA A is stochastic if the language f_A defines a probability distribution over $\Sigma^\star$.

spectral radius to see that there exists a constant $c_M > 0$ depending only on M such that the following holds:

$$\begin{aligned}|\rho(M) - \rho(\hat{M})| &\leq c_M d_H(M, \hat{M}) \\ &= c_M \max\left\{ \sup_{\tau \in M} \inf_{\tau' \in \hat{M}} \|\tau - \tau'\|, \sup_{\tau' \in \hat{M}} \inf_{\tau \in M} \|\tau - \tau'\| \right\} \\ &\leq c_M \max_{\sigma \in \Sigma} \|\tau_\sigma - \hat{\tau}_\sigma\| \ .\end{aligned}$$

Note that by Lemma 28 we have $\max_{\sigma \in \Sigma} \|\tau_\sigma - \hat{\tau}_\sigma\| = O(\|H - \hat{H}\|_2)$. Thus, by making $\|H - \hat{H}\|_2$ small enough we can assume that $\gamma\rho(\hat{M}) < 1$. Using this fact and our choice of η we can apply Lemma 17 to see that $d_\gamma(A, \hat{A}) \leq O(\|H - \hat{H}\|_2)$. Furthermore, the hidden constants in the big-O notation depend on the Hankel block indices $B = (P, S)$, the size of the alphabet $|\Sigma|$, the automaton A, the norm $\|\cdot\|$, and the constant c_M through Lemma 28; and on γ through Lemma 17. □

The local Lipschitz continuity of ρ around irreducible sets of matrices plays an important role in the proof of this result. 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 29 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 [27, 38] 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, 25] 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.

8 Conclusion

8.1 Extension to Vector-Valued Weighted Automata

Throughout this paper, we considered weighted automata that compute functions of the form $f : V \to \mathbb{R}$. It is also possible to work with a class of

weighted automata that output a vector, which subsumes the class of WFA presented here. Indeed, we can replace the final map by a linear map of the form $\beta : V \to W$, where W is a real vector space. This form was presented in [35] for multitask learning. By replacing the absolute values on the output by a norm $\|\cdot\|$ of our choice (as they are all equivalent up to a constant factor), we can still define pseudometrics for this new class of automata.

Extending the definitions to vector-valued outputs, we get the following fixed-point operator:

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

which gives rise to the following seminorm under the usual condition that $\gamma\rho(A) < 1$:

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

Similarly, we also get a closed-form expression for the pseudometric:

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

We can also recover the continuity properties presented in Section 4 for these pseudometrics; the proofs remain the same in essence. In particular, the continuity properties could be used to prove convergence of the target automaton and the automaton output by the spectral learning algorithm for vector-valued WFA as we get access to more data, with respect to d_γ.

8.2 Future Work

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.

For the first one we suspect some recent ideas from non-linear optimization might be useful in developing approximation algorithms. To explore the relation to approximate minimization it would be interesting to extend the spectral ideas at the heart of the approximate minimization algorithm in [5, 6] with respect to the ℓ_2 metric to the pseudometric developed in the present paper.

Acknowledgements. We would like to thank Doina Precup who was actively involved in the approximate minimization work. One of us (PP) has benefitted from discussions with Clare Lyle who independently worked out the results in Section 8.1. This research has been supported by a grant from NSERC (Canada).

References

[1] Raphaël 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 and Mehryar Mohri. Learning weighted automata. In *Conference on Algebraic Informatics*, 2015.

[5] 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.

[6] Borja Balle, Prakash Panangaden, and Doina Precup. Singular value automata and approximate minimization. *Mathematical Structures in Computer Science*, 29(9):1444–1478, October 2019.

[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 Bonsangue, Michele Boreale, Jan Rutten, and Alexandra Silva. A coalgebraic perspective on linear weighted automata. *Information and Computation*, 211:77–105, 2012.

[10] 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.

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

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

[13] 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.

[14] Josée Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash 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] Yuan Feng and Lijun Zhang. When equivalence and bisimulation join forces in probabilistic automata. In *International Symposium on Formal Methods*, pages 247–262. Springer, 2014.

[18] 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.

[19] 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.

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

[21] Michel Fliess. Matrices de Hankel. *J. Math. Pures Appl*, 53(9):197–222, 1974.

[22] Denis François, 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.

[23] 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.

[24] Hugo Gimbert and Youssouf Oualhadj. Probabilistic automata on finite words: Decidable and undecidable problems. In *International Colloquium on Automata, Languages, and Programming*, pages 527–538. Springer, 2010.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

[31] K. G. Larsen and A. Skou. Bisimulation through probablistic testing. *Information and Computation*, 94:1–28, 1991.

[32] 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.

[33] Prakash Panangaden. *Labelled Markov Processes*. Imperial College Press, 2009.

[34] Azaria Paz. *Probabilistic automata*. Academic Press, Inc., 1971.

[35] Guillaume Rabusseau, Borja Balle, and Joelle Pineau. Multitask spectral learning of weighted automata. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2588–2597. Curran Associates, Inc., 2017.

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

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

[38] Sajid M. Siddiqi, Byron Boots, and Geoffrey Gordon. Reduced-rank hidden Markov models. In *AISTATS*, 2010.

[39] 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.

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