Spectral Learning from a Single Trajectory under Finite-State Policies

Borja Balle [1] **Odalric-Ambrym Maillard** [2]

Abstract

We present spectral methods of moments for learning sequential models from a single trajectory, in stark contrast with the classical literature that assumes the availability of multiple i.i.d. trajectories. Our approach leverages an efficient SVD-based learning algorithm for weighted automata and provides the first rigorous analysis for learning many important models using dependent data. We state and analyze the algorithm under three increasingly difficult scenarios: probabilistic automata, stochastic weighted automata, and reactive predictive state representations controlled by a finite-state policy. Our proofs include novel tools for studying mixing properties of stochastic weighted automata.

1. Introduction

Spectral methods of moments are a powerful tool for designing provably correct learning algorithms for latent variable models. Successful applications of this approach include polynomial-time algorithms for learning topic models (Anandkumar et al., 2012; 2014), hidden Markov models (Hsu et al., 2012; Siddiqi et al., 2010; Anandkumar et al., 2014), mixtures of Gaussians (Anandkumar et al., 2014; Hsu & Kakade, 2013), predictive state representations (Boots et al., 2011; Hamilton et al., 2014; Bacon et al., 2015; Langer et al., 2016), weighted automata (Bailly, 2011; Balle et al., 2011; Balle & Mohri, 2012; Balle et al., 2014a;b; Glaude & Pietquin, 2016), and weighted context-free grammars (Bailly et al., 2010; Cohen et al., 2013; 2014). All these methods can be split into two classes depending on which spectral decomposition they rely on. The first class includes algorithms based on an Singular Value Decomposition (SVD) decomposition of a matrix containing (estimated) moments of the target distribution (e.g. Hsu et al. (2012); Boots et al. (2011); Balle et al. (2014a)). The other class includes algorithms relying on symmetric tensor decomposition methods (e.g. Anandkumar et al. (2014); Hsu & Kakade (2013)). The advantage of tensor methods is that their output is always a proper probabilistic model. On the other hand, SVD methods, which do not always output a probabilistic model, provide learning algorithms for models which are provably non-learnable using tensor methods. A notable example is the class of stochastic weighted automata that do not admit a probabilistic parametrization (Jaeger, 2000; Denis & Esposito, 2008).

Natural language processing (NLP) and reinforcement learning (RL) are the two main application domains of spectral learning. For example, SVD methods for learning weighted context-free grammars have proved very successful in language-related problems (Cohen et al., 2013; Luque et al., 2012). In the context of RL, efficient SVD methods for learning predictive state representations were proposed in (Boots et al., 2011; Hamilton et al., 2014). A recent application of tensor methods to RL is given in (Azizzadenesheli et al., 2016), where the authors use a spectral algorithm to obtain a PAC-RL learning result for POMDP under memory-less policies. All these results have in common that they provide learning algorithms for models over sequences. However, there is a fundamental difference between the nature of data in NLP and RL. With the exception of a few problems, most of NLP "safely" relies on the assumption that i.i.d. data from the target distribution is available. In RL, however, the most general scenario assumes that the learner can only collect a single continuous trajectory of data while all existing analysis of the SVD method for sequential models[1] rely on the i.i.d. assumption (Hsu et al., 2012; Balle & Mohri, 2015; Glaude & Pietquin, 2016). Regarding tensor methods, (Azizzadenesheli et al., 2016) gave the first analysis under dependent data satisfying certain mixing conditions.

The purpose of this paper is to provide the first rigorous analyses of spectral SVD methods for learning sequential models from non-i.i.d. data. We provide efficient algorithms with *provable guarantees* for learning several sequential models from a *single trajectory*. Specifically, we

[1]Amazon Research, Cambridge, UK (work done at Lancaster University) [2]Inria Lille - Nord Europe, Villeneuve d'Ascq, France. Correspondence to: Borja Balle <pigem@amazon.co.uk>, Odalric-Ambrym Maillard <odalric.maillard@inria.fr>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]See (Thon & Jaeger, 2015) for a survey of sequential models learnable with SVD methods.

consider three models: probabilistic automata, stochastic weighted automata, and PSRs under control from a finite-state policy. The first two results extend existing results in the literature for i.i.d. data (Hsu et al., 2012; Balle et al., 2014a). The last result is an analog of the environment learning result for POMDP – not the whole RL result – of (Azizzadenesheli et al., 2016), with the difference that our analysis provides guarantees under a much larger set of policies (finite-state, as opposed to memory-less). This result can also be interpreted as an extension of the batch-based PSR learning algorithm from (Boots et al., 2011) to the non-i.i.d. case, although they do not provide finite-sample guarantees. Our analysis is especially relevant since the single trajectory spectral algorithm we analyze has been used previously without an explicit instantiation or analysis. For example, (Kulesza et al., 2015; Shaban et al., 2015) present experiments with datasets containing single or few long trajectories which are broken into short subsequences and given as input to an spectral algorithm designed for i.i.d. data. A more detailed list of our contributions is as follows:

(1) A single-trajectory spectral-learning algorithm for probabilistic automata whose sample complexity depends on the *mixing* properties of the target automaton (Section 3).

(2) An extension of this result showing that the same algorithm also learns stochastic weighted automata (Section 4). In this case the analysis is more involved, and requires a novel notion of mixing for stochastic weighted automata and tools from the theory of linear-invariant cones (Berman & Plemmons, 1994).

(3) A generalization of the algorithm that learns reactive PSR controlled by a finite-state stochastic policy (Section 5). We provide for this algorithm finite-sample bounds under a simple exploration assumption.

The most important tool in our analysis is a concentration inequality for functions of dependent random variables. These inequalities depend on the mixing coefficients of the underlying process. We provide technical estimates for the relevant mixing coefficients in each of the three cases listed above. Our goal for future work is to extend (3) to prove a PAC-RL for PSR under finite-state policies. We also think that the tools we develop to prove (2) can be used to improve the sample complexity of algorithms for learning stochastic weighted automata in the i.i.d. case.

In Section 2, we start by recalling several facts about weighted automata, spectral learning, and mixing that will play a role in the sequel. For space reasons, most of our proofs are deferred to the Supplementary Material.

2. Background

Let Σ be a finite alphabet, $\Sigma^\star$ denote the set of words of finite length on Σ, Σ^ω the set of all infinite words on Σ, and ϵ be the empty word. Given two sets of words $\mathcal{U}, \mathcal{V} \subset \Sigma^\star$ we write $\mathcal{U} \cdot \mathcal{V}$ to denote the set of words $\{uv | u \in \mathcal{U}, v \in \mathcal{V}\}$ obtained by concatenating all words in $\mathcal{U}$ with all words in $\mathcal{V}$. Let $\mathcal{P}(\Sigma^\omega)$ be the set of probability distributions over Σ^ω. A member $\rho \in \mathcal{P}(\Sigma^\omega)$ is called a *stochastic process* and a random infinite word $\xi \sim \rho$ is called a *trajectory*.

Weighted and probabilistic automata A *weighted finite automaton* (WFA) with n states is a tuple $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\}_{\sigma \in \Sigma} \rangle$ where $\alpha, \beta \in \mathbb{R}^n$ are vectors of initial and final weights respectively, and $A_\sigma \in \mathbb{R}^{n \times n}$ are matrices of transition weights. A weighted automaton $\mathbb{A}$ computes a function $f_\mathbb{A} : \Sigma^\star \to \mathbb{R}$ given by $f_\mathbb{A}(w) = \alpha^\top A_w \beta$ where $A_w = A_{w_1} \cdots A_{w_t}$ for $w = w_1 \cdots w_t$. A WFA is minimal if there does not exist another WFA with less states computing the same function. A WFA $\mathbb{A}$ is *stochastic* if there exists a stochastic process ρ such that for every $w \in \Sigma^\star$, $f_\mathbb{A}(w) = \mathbb{P}[\xi \in w\Sigma^\omega]$; that is, $\mathbb{A}$ provides a representation for the probabilities of prefixes under the distribution of ρ. A weighted automaton is *irreducible* if the labelled directed graph with n vertices obtained by adding a transition from i to j with label σ whenever $A_\sigma(i,j) \neq 0$ is strongly connected. It can be shown that irreducibility implies minimality, and that almost all WFA are irreducible in the sense that the set of irreducible WFA are dense on the set of all WFA (Balle et al., 2017).

A *probabilistic finite automaton* (PFA) is a stochastic WFA $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\} \rangle$ where the weights have a probabilistic interpretation. Namely, α is a probability distribution over $[n]$, $A_\sigma(i,j)$ is the probability of emitting symbol σ and transitioning to state j starting from state i, and $\beta(i) = 1$ for all $i \in [n]$. It is immediate to check that a PFA satisfying these conditions induces a stochastic process. However not all stochastic WFA admit an equivalent PFA (Jaeger, 2000; Denis & Esposito, 2008).

If $\mathbb{A}$ is a PFA, then the matrix $A = \sum_{\sigma \in \Sigma} A_\sigma$ yields the Markov kernel $A(i,j) = \mathbb{P}[j \mid i]$ on the state space $[n]$ after marginalizing over the observations. It is easily checked that A is row-stochastic, and thus $A\beta = \beta$. Furthermore, for every distribution $\alpha_0 \in \mathbb{R}^n$ over $[n]$ we have $\alpha_0^\top A = \alpha_1$ for some other probability distribution α_1 over $[n]$. In the case of PFA irreducibility coincides with the usual concept of irreducibility of the Markov chain induced by A.

Hankel matrices and spectral learning The Hankel matrix of a function $f : \Sigma^\star \to \mathbb{R}$ is the infinite matrix $H_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ with entries $H_f(u,v) = f(uv)$. Given finite sets $\mathcal{U}, \mathcal{V} \subset \Sigma^\star$, $H_f^{\mathcal{U},\mathcal{V}} \in \mathbb{R}^{\mathcal{U} \times \mathcal{V}}$ denotes the restriction of matrix H_f to prefixes in $\mathcal{U}$ and suffixes in $\mathcal{V}$.

Fliess' theorem (Fliess, 1974) states that a Hankel matrix

H_f has finite rank n if and only if there exists a WFA $\mathbb{A}$ with n states such that $f = f_{\mathbb{A}}$. This implies that a WFA $\mathbb{A}$ with n states is minimal if and only if $n = \mathrm{rank}(H_{f_{\mathbb{A}}})$. The spectral learning algorithm for WFA (Balle et al., 2014a) provides a mechanism for recovering such a WFA from a finite sub-block $H_f^{\mathcal{U},\mathcal{V}}$ of H_f such that: 1) $\epsilon \in \mathcal{U} \cap \mathcal{V}$, 2) there exists a set $\mathcal{U}'$ such that $\mathcal{U} = \mathcal{U}' \cup (\cup_{\sigma\in\Sigma}\mathcal{U}'\sigma)$, 3) $\mathrm{rank}(H_f) = \mathrm{rank}(H_f^{\mathcal{U}',\mathcal{V}})$. A pair $(\mathcal{U},\mathcal{V})$ that satisfies these conditions is called a *complete basis* for f. The pseudo-code of this algorithm is given below:

Algorithm 1: Spectral Learning for WFA

Input: number of states n, Hankel matrix $H^{\mathcal{U},\mathcal{V}}$

Find $\mathcal{U}'$ such that $\mathcal{U} = \mathcal{U}' \cup (\cup_{\sigma\in\Sigma}\mathcal{U}'\sigma)$
Let $H_\epsilon = H^{\mathcal{U}',\mathcal{V}}$
Compute the rank n SVD $H_\epsilon \approx UDV^\top$
Let $h_{\mathcal{V}} = H^{\{\epsilon\},\mathcal{V}}$ and take $\alpha = V^\top h_{\mathcal{V}}$
Let $h_{\mathcal{U}'} = H^{\mathcal{U}',\{\epsilon\}}$ and take $\beta = D^{-1}U^\top h_{\mathcal{U}'}$
foreach $\sigma \in \Sigma$ **do**
  Let $H_\sigma = H^{\mathcal{U}'\sigma,\mathcal{V}}$ and take $A_\sigma = D^{-1}U^\top H_\sigma V$
return $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\}\rangle$

The main strength of Algorithm 1 is its robustness to noise. Specifically, if only an approximation $\widehat{H}^{\mathcal{U},\mathcal{V}}$ of the Hankel matrix is known, then the error between the target automaton $\mathbb{A}$ and the automaton $\widehat{\mathbb{A}}$ learned from $\widehat{H}^{\mathcal{U},\mathcal{V}}$ can be controlled in terms of the error $\|H^{\mathcal{U},\mathcal{V}} - \widehat{H}^{\mathcal{U},\mathcal{V}}\|_2$; see (Hsu et al., 2012) for a proof in the HMM case and (Balle, 2013) for a proof in the general WFA case. These tedious but now standard arguments readily reduce the problem of learning WFA via spectral learning to that of estimating the corresponding Hankel matrix.

Classical applications of spectral learning assume one has access to i.i.d. samples from a stochastic process ρ. In this setting one can obtain a sample $S = (\xi^{(1)}, \ldots, \xi^{(N)})$ containing N finite-length trajectories from ρ, and use them to estimate a Hankel matrix $\widehat{H}_S^{\mathcal{U},\mathcal{V}}$ as follows:

$$\widehat{H}_S^{\mathcal{U},\mathcal{V}}(u,v) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{\xi^{(i)} \in uv\Sigma^\omega\}\,.$$

If $\rho = \rho_{\mathbb{A}}$ for some stochastic WFA, then obviously $\mathbb{E}_S[\widehat{H}_S^{\mathcal{U},\mathcal{V}}] = H_{f_{\mathbb{A}}}^{\mathcal{U},\mathcal{V}}$ and a large sample size N will provide a good approximation $\widehat{H}_S^{\mathcal{U},\mathcal{V}}$ of $H_{f_{\mathbb{A}}}^{\mathcal{U},\mathcal{V}}$. Explicit concentration bounds for Hankel matrices bounding the error $\|H_{f_{\mathbb{A}}}^{\mathcal{U},\mathcal{V}} - \widehat{H}_S^{\mathcal{U},\mathcal{V}}\|_2$ can be found in (Denis et al., 2016).

In this paper we consider the more challenging setup where we only have access to a sample $S = \{\xi\}$ of size $N = 1$ from ρ. In particular, we show it is possible to replace the empirical average above by a Césaro average and still use the spectral learning algorithm to recover the transition matrices of a stochastic WFA. To obtain a finite-sample analysis of this *single-trajectory* learning algorithm we prove concentration results for Césaro averages of Hankel matrices. Our analysis relies on concentration inequalities for functions of dependent random variables which depend on mixing properties of the underlying process.

Mixing and concentration Let $\rho \in \mathcal{P}(\Sigma^\omega)$ be a stochastic process and $\xi = x_1 x_2 \cdots$ a random word drawn from ρ. For $1 \leqslant s < t \leqslant T$ and $u \in \Sigma^s$ we let $\rho_{t:T}(\cdot|u)$ denote the distribution of $x_t \cdots x_T$ conditioned on $x_1 \cdots x_s = u$. With this notation we define the quantity

$$\eta_t(u,\sigma,\sigma') = \|\rho_{t:T}(\cdot|u\sigma) - \rho_{t:T}(\cdot|u\sigma')\|_{TV}$$

for any $u \in \Sigma^{s-1}$, and $\sigma,\sigma' \in \Sigma$. Then the *$\eta$-mixing coefficients* of ρ at horizon T are given by

$$\eta_{s,t} = \sup_{u\in\Sigma^{s-1},\sigma,\sigma'\in\Sigma} \eta_t(u,\sigma,\sigma')\ .$$

Mixing coefficients are useful in establishing concentration properties of functions of dependent random variables. The Lipschitz constant of a function $g : \Sigma^T \to \mathbb{R}$ with respect to the Hamming distance is defined as

$$\|g\|_{Lip} = \sup |g(w) - g(w')|\ ,$$

where the supremum is taken over all pairs of words $w, w' \in \Sigma^T$ differing in exactly one symbol. The following theorem proved in (Chazottes et al., 2007; Kontorovich et al., 2008) provides a concentration inequality for Lipschitz functions of weakly dependent random variables.

Theorem 1 *Let $\rho \in \mathcal{P}(\Sigma^\omega)$ and $\xi = x_1 x_2 \cdots \sim \rho$. Suppose $g : \Sigma^T \to \mathbb{R}$ satisfies $\|g\|_{Lip} \leqslant 1$ and let $Z = g(x_1, \ldots, x_T)$. Let $\eta_\rho = 1 + \max_{1<s<T} \sum_{t=s+1}^{T} \eta_{s,t}$, where $\eta_{s,t}$ are the η-mixing coefficients of ρ at horizon T. Then the following holds for any $\varepsilon > 0$:*

$$\mathbb{P}_\xi\left[Z - \mathbb{E}Z > \varepsilon T\right] \leqslant \exp\left(\frac{-2\varepsilon^2 T}{\eta_\rho^2}\right)\ ,$$

with an identical bound for the other tail.

Theorem 1 shows that the mixing coefficient η_ρ is a key quantity in order to control the concentration of a function of dependent variables. In fact, upper-bounding η_ρ in terms of geometric ergodicity coefficients of a latent variable stochastic process enables (Kontorovich & Weiss, 2014) to analyze the concentration of functions of HMMs and (Azizzadenesheli et al., 2016) to provide PAC guarantees for an RL algorithm for POMDP based on spectral tensor decompositions. Our Lemma 2 uses a similar but more refined bounding strategy that directly applies when the transition and observation processes are *not* conditionally independent. Lemma 4 refines this strategy further to control η_ρ for stochastic WFA (for which there may be no underlying Markov stochastic process in general). To the best of our knowledge this yields the first concentration results for the challenging setting of stochastic WFA.

3. Single-Trajectory Spectral Learning of PFA

In this section we focus on the problem of learning the transition structure of a PFA $\mathbb{A}$ using single trajectory is generated by $\mathbb{A}$. We provide a spectral learning algorithm for this problem and a finite-sample analysis consisting of a concentration bound for the error on the Hankel matrix estimated by the algorithm. We assume the learner has access to a single infinite-length trajectory $\xi \sim \rho_{\mathbb{A}}$ that is progressively uncovered. The algorithm uses a length t prefix from ξ to estimate a Hankel matrix whose entries are Césaro averages. This Hankel matrix is then processed by the usual spectral learning algorithm to recover an approximation to an automaton with transition weights equivalent to those of $\mathbb{A}$. We want to analyze the quality of the model learned by the algorithm after observing the first t symbols from ξ.

We start by showing that Césaro averages provide a consistent mechanism for learning the transition structure of $\mathbb{A}$. Then we proceed to analyze the accuracy of the Hankel estimation step. As discussed in Section 2, this is enough to obtain finite-sample bounds for learning PFA. The general case of stochastic WFA is considered in Section 4.

3.1. Learning with Césaro Averages is Consistent

Let $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\} \rangle$ be a PFA computing a function $f_{\mathbb{A}} : \Sigma^\star \to \mathbb{R}$ and defining a stochastic process $\rho_{\mathbb{A}} \in \mathcal{P}(\Sigma^\omega)$. For convenience we drop the subscript and just write f and ρ. Since we only have access to a single trajectory ξ from ρ we cannot obtain an approximation of the Hankel matrix for f by averaging over multiple i.i.d. trajectories. Instead, we compute Césaro averages over the trajectory ξ to obtain a Hankel matrix whose expectation is related to $\mathbb{A}$ as follows.

For any $t \in \mathbb{N}$ let $\bar{f}_t : \Sigma^\star \to \mathbb{R}$ be the function given by $\bar{f}_t(w) = (1/t)\sum_{s=0}^{t-1} f(\Sigma^s w)$, where $f(\Sigma^s w) = \sum_{u \in \Sigma^s} f(uw)$. We shall sometimes write $f_s(w) = f(\Sigma^s w)$. Using the definition of the function computed by a WFA it is easy to see that

$$\sum_{u \in \Sigma^s} f(uw) = \sum_{u \in \Sigma^s} \alpha^\top A_u A_w \beta = \alpha^\top A^s A_w \beta \ ,$$

where $A = \sum_\sigma A_\sigma$ is the Markov kernel on the state space of A. Thus, introducing $\bar{\alpha}_t^\top = (1/t)\sum_{s=0}^{t-1} \alpha^\top A^s$ we get $\bar{f}_t(w) = \bar{\alpha}_t^\top A_w \beta$. Since α is a probability distribution, A is a Markov kernel, and probability distributions are closed by convex combinations, then $\bar{\alpha}_t$ is also a probability distribution over $[n]$. Thus, we have just proved the following:

Lemma 1 (Consistency) *The Césaro average of f over t steps, $\bar{f}_t$, is computed by the probabilistic automaton $\bar{\mathbb{A}}_t = \langle \bar{\alpha}_t, \beta, \{A_\sigma\} \rangle$. In particular, $\mathbb{A}$ and $\bar{\mathbb{A}}_t$ have the same number of states and the same transition probability matrices. Furthermore, if $\mathbb{A}$ is irreducible then $\bar{\mathbb{A}}_t$ is minimal.*

The irreducibility claim follows from (Balle et al., 2017).

For convenience, in the sequel we write $\bar{H}_t^{\mathcal{U},\mathcal{V}}$ for the $(\mathcal{U},\mathcal{V})$-block of the Hankel matrix $H_{f_{\bar{\mathbb{A}}_t}}$.

Remark 1 *The irreducible condition simply ensures there is a unique stationary distribution, and that the Hankel matrix of $\bar{\mathbb{A}}_t$ has the same rank as the Hankel matrix of $\mathbb{A}$ (otherwise it could be smaller).*

3.2. Spectral Learning Algorithm

Algorithm 2 describes the estimation of the empirical Hankel matrix $\widehat{H}_{t,\xi}^{\mathcal{U},\mathcal{V}}$ from the first $t+L$ symbols of a single trajectory using the corresponding Césaro averages. To avoid cumbersome notations, in the sequel we may drop super and subscripts when not needed and write $\widehat{H}_t$ or $\widehat{H}$ when $\mathcal{U}$, $\mathcal{V}$, and ξ are clear from the context. Note that by Lemma 1 the expectation $\mathbb{E}[\widehat{H}]$ over $\xi \sim \rho$ is equal to the Hankel matrix $\bar{H}_t$ of the function $\bar{f}_t$ computed by the PFA $\bar{\mathbb{A}}_t$.

Algorithm 2: Single Trajectory Spectral Learning (Generative Case)

Input: number of states n, length t, prefixes $\mathcal{U} \subset \Sigma^\star$, suffixes $\mathcal{V} \subset \Sigma^\star$

Let $L = \max_{w \in \mathcal{U}\cdot\mathcal{V}} |w|$

Sample trajectory $\xi = x_1 x_2 \cdots x_{t+L} \cdots \sim \rho$

foreach $u \in \mathcal{U}$ *and* $v \in \mathcal{V}$ **do**

  Let $\widehat{H}(u,v) = \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{I}\{x_{s+1:s+|uv|} = uv\}$

Apply the spectral algorithm to $\widehat{H}$ with rank n

3.3. Concentration Results

Now we proceed to analyze the error $\widehat{H}_t - \bar{H}_t$ in the Hankel matrix estimation inside Algorithm 2. In particular, we provide concentration bounds that depend on the length t, the mixing coefficient η_ρ of the process ρ, and the structure of the basis $(\mathcal{U},\mathcal{V})$. The main result of this section is the matrix-wise concentration bound Theorem 3 where we control the spectral norm of the error matrix. For comparison we also provide a simpler entry-wise bound and recall the equivalent matrix-wise bound in the i.i.d. setting.

Before trying to bound the concentration of the errors using Theorem 1 we need to analyze the mixing coefficient of the process generated by a PFA. This is the goal of the following result, whose proof is provided in Appendix A.

Lemma 2 (η-mixing for PFA) *Let $\mathbb{A}$ be PFA and assume that it is (C,θ)-geometrically mixing in the sense that for some constants $C > 0, \theta \in (0,1)$ we have*

$$\forall t \in \mathbb{N}, \quad \mu_t^{\mathbb{A}} = \sup_{\alpha,\alpha'} \frac{\|\alpha A^t - \alpha' A^t\|_1}{\|\alpha - \alpha'\|_1} \leqslant C\theta^t \ ,$$

where the supremum is over all probability vectors. Then we have $\eta_{\rho_{\mathbb{A}}} \leqslant C/(\theta(1-\theta))$.

Remark 2 *A sufficient condition for the geometric control of $\mu_t^{\mathbb{A}}$ is that A admits a spectral gap. In this case θ can be chosen to be the modulus of the second eigenvalue $|\lambda_2(A)| < 1$ of the transition kernel A.*

Before the main result of this section we provide a concentration result for each individual entry of the estimated Hankel matrix as a warmup (see Appendix D).

Theorem 2 (Single-trajectory, entry-wise) *Let $\mathbb{A}$ be a (C,θ)-geometrically mixing PFA and $\xi \sim \rho_{\mathbb{A}}$ a trajectory of observations. Then for any $u \in \mathcal{U}, v \in \mathcal{V}$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left[\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi}(u,v) - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}(u,v) \geqslant \frac{|uv|C}{\theta(1-\theta)}\sqrt{\left(1+\frac{|uv|-1}{t}\right)\frac{\log(1/\delta)}{2t}}\right] \leqslant \delta \ ,$$

with an identical bound for the other tail.

A naive way to handle the concentration of the whole Hankel matrix is to control the Frobenius norm $\|\widehat{H}_t - \bar{H}_t\|_F$ by taking a union bound over all entries using Theorem 2. However, the resulting concentration bound would scale as $\sqrt{|\mathcal{U}||\mathcal{V}|}$. To have better dependency with the dimension (the matrix has dimension $|\mathcal{U}| \times |\mathcal{V}|$) can split the empirical Hankel matrix $\widehat{H}$ into blocks containing strings of the same length (as suggested by the dependence of the bound above on $|uv|$). We thus introduce the maximal length $L = \max_{w \in \mathcal{U}\cdot\mathcal{V}} |w|$, and the set $\mathcal{U}_\ell = \{u \in \mathcal{U} : |u| = \ell\}$ for any $\ell \in \mathbb{N}$. We use these to define the quantity $n_{\mathcal{U}} = |\{\ell \in [0,L] : |\mathcal{U}_\ell| > 0\}|$, and introduce likewise $\mathcal{V}_\ell, n_{\mathcal{V}}$ with obvious definitions. With this notation we can now state the main result of this section.

Theorem 3 (Single-trajectory, matrix-wise) *Let $\mathbb{A}$ be as in Theorem 2. Let $m = \sum_{u\in\mathcal{U},v\in\mathcal{V}} \bar{f}_t(uv)$ be the probability mass and $d = \min\{|\mathcal{U}||\mathcal{V}|, 2n_{\mathcal{U}}n_{\mathcal{V}}\}$ be the effective dimension. Then, for all $\delta \in (0,1)$ we have*

$$\mathbb{P}\left[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}\|_2 \geqslant \left(\sqrt{L}+\sqrt{\frac{2C}{1-\theta}}\right)\sqrt{\frac{2m}{t}} + \frac{2LC}{\theta(1-\theta)}\sqrt{\left(1+\frac{L-1}{t}\right)\frac{d\ln(1/\delta)}{2t}}\right] \leqslant \delta \ .$$

Remark 3 *Note that quantity $n_{\mathcal{U}}n_{\mathcal{V}}$ in d can be exponentially smaller than $|\mathcal{U}||\mathcal{V}|$. Indeed, for $\mathcal{U} = \mathcal{V} = \Sigma^{\leqslant L/2}$ we have $|\mathcal{U}||\mathcal{V}| = \Theta(|\Sigma|^L)$ while $n_{\mathcal{U}}n_{\mathcal{V}} = \Theta(L^2)$.*

For comparison, we recall a state-of-the-art concentration bound for estimating the Hankel matrix of a stochastic language[2] from N i.i.d. trajectories.

[2] A stochastic language is a probability distribution over $\Sigma^\star$.

Theorem 4 (Theorem 7 in (Denis et al., 2014)) *Let $\mathbb{A}$ be a stochastic WFA with stopping probabilities and $S = (\xi^{(1)}, \dots, \xi^{(N)})$ be an i.i.d. sample of size N from the distribution $\rho_{\mathbb{A}} \in \mathcal{P}(\Sigma^\star)$. Let $m = \sum_{u\in\mathcal{U},v\in\mathcal{V}} f_{\mathcal{A}}(uv)$. Then, for all $c > 0$ we have*

$$\mathbb{P}\left[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{S} - H^{\mathcal{U},\mathcal{V}}_{f_{\mathbb{A}}}\|_2 > \sqrt{\frac{2cm}{N}} + \frac{2c}{3N}\right] \leqslant \frac{2c}{e^c - c - 1} \ .$$

3.4. Sketch of the Proof of Theorem 3

In this section we sketch the main steps of the proof of Theorem 3 (the full proof is given in Appendices A and D). We focus on highlighting the main difficulties and paving the path for the extension of Theorem 3 to stochastic WFA given in Section 4.

The key of the proof is to study the function $g(\xi) = \|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}\|_2$, in view of applying Theorem 1. To this end, we first control the η-mixing coefficients using Lemma 2. The next step is to control the Lipschitz constant $\|g\|_{Lip}$. This part is not very difficult and we derive after a few careful steps the bound $\|g\|_{Lip} \leqslant L\sqrt{d}/t$.

The second and most interesting part of the proof is about the control of $\mathbb{E}[g(\xi)]$. Let us give some more details.

Decomposition step We control $\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t} - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}\|_2$ by its Frobenius norm and get

$$\mathbb{E}[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t} - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}\|_2]^2 \leqslant \sum_{w\in\mathcal{U}\cdot\mathcal{V}} |w|_{\mathcal{U},\mathcal{V}}\mathbb{E}[(\widehat{f}_t(w) - \bar{f}_t(w))^2] \ ,$$

where we introduced $|w|_{\mathcal{U},\mathcal{V}} = |\{(u,v) \in \mathcal{U}\times\mathcal{V} : uv = w\}|$, and $\widehat{f}_t(w) = \frac{1}{t}\sum_{s=1}^t b_s(w)$ using the shorthand notation $b_s(w) = \mathbb{I}\{x_s \dots x_{s+|w|-1} = w\}$. Also, $\bar{f}_t(w) = \mathbb{E}[\widehat{f}_t(w)] = \frac{1}{t}\sum_{s=1}^t f_s(w)$, where $f_s(w) = \rho_{\mathbb{A}}(\Sigma^{s-1}w\Sigma^\omega)$. This implies that we have a sum of variances, where each of the terms can be written as

$$\mathbb{E}[(\widehat{f}_t(w) - \bar{f}_t(w))^2] = \frac{1}{t^2}\mathbb{E}\left[\left(\sum_{s=1}^t b_s(w)\right)^2\right] - \frac{1}{t^2}\left(\sum_{s=1}^t f_s(w)\right)^2 \ .$$

Slicing step An important observation is that each probability term satisfies $f_s(w) = \alpha^\top A^{s-1}A_w\beta$ because of the PFA assumption on $\rho_{\mathbb{A}}$. Furthermore, it follows from $\mathbb{A}$ being a PFA that $\sum_{|w|=l} f_s(w) = 1$ for all s and l. This suggests that we group the terms in the sum over $\mathcal{W} = \mathcal{U}\cdot\mathcal{V}$ by length, so we write $\mathcal{W}_l = \mathcal{W}\cap\Sigma^l$ and define $L_l = \max_{w\in\mathcal{W}_l} |w|_{\mathcal{U},\mathcal{V}}$ the maximum number of ways to write a string of length l in $\mathcal{W}$ as a concatenation of a prefix in $\mathcal{U}$ and a suffix in $\mathcal{V}$. Note that we always have $L_l \leqslant l+1$.

A few more steps lead to the following bound

$$\mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2]^2 \leqslant \quad (1)$$
$$\frac{1}{t^2}\sum_{l=0}^{\infty}\sum_{w\in\mathcal{W}_l}|w|_{\mathcal{U},\mathcal{V}}\Big[\sum_{s=1}^{t}(1-f_s(w))f_s(w)$$
$$+2\sum_{1\leqslant s<s'\leqslant t}\Big(\mathbb{E}[b_s(w)b_{s'}(w)]-f_s(w)f_{s'}(w)\Big)\Big] .$$

We control the first term in (1) using

$$\sum_{l=0}^{\infty}\sum_{w\in\mathcal{W}_l}|w|_{\mathcal{U},\mathcal{V}}\sum_{s=1}^{t}(1-f_s(w))f_s(w)\leqslant t\sum_{u\in\mathcal{U},v\in\mathcal{V}}\bar{f}_t(uv) .$$

Cross terms Regarding the remaining "cross"-term in (1) we fix $w\in\mathcal{W}_l$ and obtain the equation

$$\mathbb{E}[b_s(w)b_{s'}(w)] - f_s(w)f_{s'}(w)$$
$$= \alpha_{s-1}^{\top}\Big(A_w^{s'-s} - A_w\beta\alpha_{s'-1}^{\top}A_w\Big)\beta \ , \quad (2)$$

where we introduced the vectors $\alpha_s^{\top}=\alpha^{\top}A^s$ and transition matrix $A_w^{s'-s}=\sum_{x\in\Sigma_w^{s'-s}}A_x$ corresponding to the "event" $\Sigma_w^{s'-s}=w\Sigma^{s'-s}\cap\Sigma^{s'-s}w$. We now discuss two cases.

First control If $s'-s<l$, we use the simplifying fact that $\Sigma_w^{s'-s}\subset w\Sigma^{s'-s}$ to upper bound (2) by

$$\alpha_{s-1}^{\top}A_w\Big(A^{s'-s}-\beta\alpha_{s'-1}^{\top}A_w\Big)\beta$$
$$= f_s(w)(1-f_{s'}(w))\leqslant f_s(w) \ .$$

Second control When $s'-s\geqslant|w|=l$ we have $\Sigma_w^{s'-s}=w\Sigma^{s'-s-l}w$ and $A_w^{s'-s}=A_wA^{s'-s-l}A_w$. Thus, we rewrite (2) and bound it using Hölder's inequality as follows:

$$\alpha_{s-1}^{\top}A_w\Big(A^{s'-s-l}-\beta\alpha_{s'-1}^{\top}\Big)A_w\beta\leqslant \quad (3)$$
$$\|\alpha_{s-1}^{\top}A_w\|_1\|A^{s'-s-l}-\beta\alpha_{s'-1}^{\top}\|_{\infty}\|A_w\beta\|_{\infty} .$$

Using Lemma 6 in Appendix A we bound the induced norm as $\|A^{s'-s-l}-\beta\alpha_{s'-1}^{\top}\|_{\infty}\leqslant 2\mu_{s'-s-l}^{\mathbb{A}}$, where $\mu_t^{\mathbb{A}}$ is the mixing coefficient defined in Lemma 2. Also, it holds that $\|A_w\beta\|_{\infty}\leqslant 1$. Finally, since $\alpha_{s-1}^{\top}A_w$ is a sub-distribution over states, we have the key equalities

$$\sum_{w\in\mathcal{W}_l}|w|_{\mathcal{U},\mathcal{V}}\|\alpha_{s-1}^{\top}A_w\|_1=\sum_{w\in\mathcal{W}_l}|w|_{\mathcal{U},\mathcal{V}}\alpha_{s-1}^{\top}A_w\beta \quad (4)$$
$$=\sum_{w\in\mathcal{W}_l}|w|_{\mathcal{U},\mathcal{V}}f_s(w)=\sum_{u\in\mathcal{U},v\in\mathcal{V}:uv\in\mathcal{W}_l}f_s(uv) \ .$$

The proof is concluded by collecting the previous bounds, plugging them into (1), and using Lemma 2 to get

$$\mathbb{E}[g(\xi)]^2\leqslant\left(2L-1+\frac{4C}{1-\theta}\right)\frac{m}{t} \ . \quad (5)$$

4. Extension to Stochastic WFA

We now generalize the results in previous section to the case where the distribution over ξ is generated by a *stochastic weighted automaton* that might not have a probabilistic representation. The key observation is that Algorithm 2 can learn stochastic WFA without any change, and the consistency result in Lemma 1 extends verbatim to stochastic WFA. However, the proof of the concentration bound in Theorem 3 requires further insights into the mixing properties of stochastic WFA. Before describing the changes required in the proof, we discuss some important geometric properties of stochastic WFA.

4.1. The State-Space Geometry of SWFA

Recall that a stochastic WFA (SWFA) $\mathbb{A}=\langle\alpha,\beta,\{A_\sigma\}\rangle$ defines a stochastic process $\rho_{\mathbb{A}}$ and computes a function $f_{\mathbb{A}}$ such that $f_{\mathbb{A}}(w)=\mathbb{P}[\xi\in w\Sigma^{\omega}]$, where $\xi\sim\rho_{\mathbb{A}}$. It is immediate to check that this implies that the weights of $\mathbb{A}$ satisfy the properties: (i) $\alpha^{\top}A_x\beta\geqslant 0$ for all $x\in\Sigma^{\star}$, and (ii) $\alpha^{\top}A^t\beta=\sum_{|w|=t}\alpha^{\top}A_w\beta=1$ for all $t\geqslant 0$, where $A=\sum_{\sigma\in\Sigma}A_\sigma$. Without loss of generality we assume throughout this section that $\mathbb{A}$ is a minimal SWFA of dimension n, meaning that any SWFA computing the same probability distribution than $\mathbb{A}$ must have dimension at least n. Importantly, the weights in α, β, and A_σ are *not* required to be non-negative in this definition. Nonetheless, it follows from these properties that β is an eigenvector of A of eigenvalue 1 exactly like in the case of PFA. We now introduce further facts about the geometry of SWFA.

A minimal SWFA $\mathbb{A}$ is naturally associated with a proper (i.e. pointed, closed, and solid) cone in $\mathcal{K}\subset\mathbb{R}^n$ called the *backward cone* (Jaeger, 2000), and characterized by the following properties: 1) $\beta\in\mathcal{K}$, 2) $A_\sigma\mathcal{K}\subseteq\mathcal{K}$ for all $\sigma\in\Sigma$, and 3) $\alpha^{\top}v\geqslant 0$ for all $v\in\mathcal{K}$. Condition 2) says that every transition matrix A_σ leaves $\mathcal{K}$ invariant, and in particular the backward vector $A_w\beta$ belongs to $\mathcal{K}$ for all $w\in\Sigma^{\star}$.

The vector of final weights β plays a singular role in the geometry of the state space of a SWFA. This follows from facts about the theory of invariant cones (Berman & Plemmons, 1994) which provides a generalization of the classical Perron–Frobenius theory of non-negative matrices to arbitrary matrices. We recall from (Berman & Plemmons, 1994) that a norm on $\mathbb{R}^n$ can be associated with every vector in the interior of $\mathcal{K}$. In particular, we will take the norm associated with the final weights $\beta\in\mathcal{K}$. This norm, denoted by $\|\cdot\|_\beta$, is completely determined by its unit ball $B_\beta=\{v\in\mathbb{R}^n:-\beta\leqslant_{\mathcal{K}}v\leqslant_{\mathcal{K}}\beta\}$, where $u\leqslant_{\mathcal{K}}v$ means $v-u\in\mathcal{K}$. In particular, $\|v\|_\beta=\inf\{r\geqslant 0:v\in rB_\beta\}$. Induced and dual norms are derived from $\|\cdot\|_\beta$ as usual. When $\mathbb{A}$ is a PFA one can take $\mathcal{K}$ to be the cone of vectors in $\mathbb{R}^n$ with non-negative entries, in which case $\beta=(1,\ldots,1)$ and $\|\cdot\|_\beta$ reduces to $\|\cdot\|_\infty$ (Berman & Plemmons, 1994). The following result shows that $\|\cdot\|_\beta$ provides the right

generalization to SWFA of the norm $\|\cdot\|_\infty$ used in the **Second control** step of the proof for PFA (see Appendix B).

Lemma 3 *For any $w \in \Sigma^\star$: (i) $\|A_w\beta\|_\beta \leqslant 1$, and (ii) $\|\alpha^\top A_w\|_{\beta,*} = \alpha^\top A_w\beta$.*

It is also natural to consider mixing coefficients for stochastic processes generated by SWFA in terms of the dual β-norm. This provides a direct analog to Lemma 2 for PFA:

Lemma 4 (η-mixing for SWFA) *Let $\mathbb{A}$ be SWFA and assume that it is (C,θ)-geometrically mixing in the sense that for some $C \geqslant 0, \theta \in (0,1)$,*

$$\mu_t^{\mathbb{A}} = \sup_{\alpha_0,\alpha_1:\alpha_0^\top\beta=\alpha_1^\top\beta=1} \frac{\|\alpha_0^\top A^t - \alpha_1^\top A^t\|_{\beta,\star}}{\|\alpha_0-\alpha_1\|_{\beta,\star}} \leqslant C\theta^t \ .$$

Then the η-mixing coefficient satisfies $\eta_{\rho_{\mathbb{A}}} \leqslant C/\theta(1-\theta)$.

Remark 4 *A sufficient condition for the geometric control of $\mu_t^{\mathbb{A}}$ is that A admits a spectral gap. In this case θ can be chosen to be the modulus of the second eigenvalue $|\lambda_2(A)|<1$ of A. Another sufficient condition is that $\theta=\gamma_\beta(A)<1$, where*

$$\gamma_\beta(A) = \sup\left\{\frac{||A\nu||_{\beta,\star}}{||\nu||_{\beta,\star}} : \nu \text{ s.t. } ||\nu||_{\beta,\star} \neq 0, \nu^\top\beta = 0\right\}.$$

4.2. Concentration of Hankel Matrices for SWFA

We are now ready to extend the proof of Theorem 3 to SWFA. Using that both PFA and SWFA define probability distributions over prefixes it follows that any argument in Section 3.4 that only appeals to the function computed by the automaton can remain unchanged. Therefore, the only arguments that need to be revisited are described in the **Second control** step. In particular, we must provide versions of (3) and (4) for SWFA.

Recalling that Hölder's inequality can be applied with any pair of dual norms, we start by replacing the norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$ in (3) with the cone-norms $\|\cdot\|_\beta$ and $\|\cdot\|_{\beta,\star}$ respectively. Next we use Lemma 3 to obtain, for any $w \in \Sigma^\star$, the bound $\|A_w\beta\|_\beta \leqslant 1$ and the equation $\|\alpha^\top A_w\|_{\beta,*} = \alpha^\top A_w\beta$ which are direct analogs of the results used for PFA. Then it only remains to relate the β-norm of $A^{s'-s-l} - \beta\alpha_{s'-1}^\top$ to the mixing coefficients $\mu_t^{\mathbb{A}}$. Applying Lemma 8 in Appendix A yields $\|A^{s'-s-l} - \beta\alpha_{s'-1}^\top\|_\beta \leqslant 2\mu_{s'-s-l}^{\mathbb{A}}$. Thus we obtain for SWFA exactly the same concentration result that we obtained for empirical Hankel matrices estimated from a single trajectory of observations generated by a PFA.

Theorem 5 (Single-trajectory, SWFA) *Let $\mathbb{A}$ be a SWFA that is (C,θ)-geometrically mixing with the definition in Lemma 4. Then the concentration bound in Theorem 3 also holds for trajectories $\xi \sim \rho_{\mathbb{A}}$.*

5. The Controlled Case

This section describes the final contribution of the paper: a generalization of our analysis of spectral learning from a single trajectory the case of dynamical systems under finite-state control. We consider discrete-time dynamical systems with finite set of observations $\mathcal{O}$ and finite set of actions $\mathcal{A}$, and let $\Sigma = \mathcal{O}\times\mathcal{A}$. We assume the learner has access to a single trajectory $\xi = (o_t,a_t)_{t\geqslant 1}$ in Σ^ω. The trajectory is generated by coupling an environment defining a distribution over observations conditioned on actions and a policy defining a distribution over actions conditioned on observations. Assuming the joint action-observation distribution can be represented by a stochastic WFA is equivalent to saying that the environment corresponds to a POMDP or PSR, and the policy has finite memory.

To fix some notation we assume the environment is represented by a conditional[3] stochastic WFA $\mathbb{A} = \langle\alpha,\beta,\{A_\sigma\}\rangle$ with n states. This implies the semantics $f_{\mathbb{A}}(w) = \mathbb{P}[o_1\cdots o_t|a_1\cdots a_t]$ for the function computed by $\mathbb{A}$, where $w = w_1\cdots w_t$ with $w_i = (o_i,a_u)$. For any $w \in \Sigma^\star$ we shall write $w^{\mathcal{A}} = a_1\cdots a_t$ and $w^{\mathcal{O}} = o_1\cdots o_t$. We also assume there is a stochastic policy π represented by a conditional PFA $\mathbb{A}_\pi = \langle\alpha_\pi,\beta_\pi,\{\Pi_\sigma\}\rangle$ with k states; that is, $f_{\mathbb{A}_\pi}(w) = \pi(w^{\mathcal{A}}|w^{\mathcal{O}}) = \mathbb{P}[a_1\cdots a_t|o_1\cdots o_t]$. In particular, $\mathbb{A}_\pi$ represents a stochastic policy that starts in a state $s_1 \in [k]$ sampled according to $\alpha_\pi(i) = \mathbb{P}[s_1 = i]$, and at each time step samples an action and changes state according to $\Pi_{o,a}(i,j) = \mathbb{P}[s_{t+1} = j, a_t = a|o_t = o, s_t = i]$.

The trajectory ξ observed by the learner is generated by the stochastic process $\rho \in \mathcal{P}(\Sigma^\omega)$ obtained by coupling $\mathbb{A}$ and $\mathbb{A}_\pi$. A standard construction in the theory of weighted automata (Berstel & Reutenauer, 1988) shows that this process can be computed by the product automaton $\mathbb{B} = \mathbb{A}\otimes\mathbb{A}_\pi = \langle\alpha_\otimes,\beta_\otimes,\{B_\sigma\}\rangle$, where $\alpha_\otimes = \alpha\otimes\alpha_\pi$, $\beta_\otimes = \beta\otimes\beta_\pi$, and $B_{o,a} = A_{o,a}\otimes\Pi_{o,a}$. It is easy to verify that $\mathbb{B}$ is a stochastic WFA with nk states computing the function $f_{\mathbb{B}}(w) = f_{\mathbb{A}}(w)f_{\mathbb{A}_\pi}(w) = \mathbb{P}[\xi \in w\Sigma^\omega]$.

At this point, the spectral algorithm from Section 4 could be used to learn $\mathbb{B}$ directly from a trajectory $\xi \sim \rho_{\mathbb{B}}$. However, since the agent interacting with environment $\mathbb{A}$ knows the policy π, we would like to leverage this information to learn directly a model of the environment. This approach is formalized in Algorithm 3, which provides a single-trajectory version of the algorithm in (Bowling et al., 2006) for learning PSR from non-blind policies with i.i.d. data.

The main difference with Algorithm 2 is that in the reactive case we need a smoothing parameter κ that will prevent the entries in the empirical Hankel matrix $\widehat{H}$ to grow unboundedly plus that the policy π satisfies an exploration assumption. κ plays a similar role in our analysis as the smoothing parameter introduced in (Denis et al., 2014) for learning

[3] Such WFA are also called reactive predictive state representations in the RL literature.

Algorithm 3: Single Trajectory Spectral Learning (Reactive Case)

Input: number of states n, length t, $\mathcal{U},\mathcal{V} \subset (\mathcal{A}\times\mathcal{O})^\star$, policy π, smoothing coefficient κ

Let $L = \max_{w\in\mathcal{U}\cdot\mathcal{V}} |w|$
Sample trajectory $o_1 a_1 o_2 a_2 \cdots o_{t+L} a_{t+L}$ using π
foreach $u \in \mathcal{U}$ *and* $v \in \mathcal{V}$ **do**
  Let $\widehat{H}(u,v) = \frac{1}{t}\sum_{s=0}^{t-1} \frac{\mathbb{I}\{o_{s+1}a_{s+1}\cdots o_{s+|uv|}a_{s+|uv|}=uv\}}{\kappa^s \pi(a_1\cdots a_{s+|uv|}|o_1\cdots o_{s+|uv|})}$
Apply the spectral algorithm to $\widehat{H}$ with rank n

stochastic languages from factor estimates in the i.i.d. case. The difference is that in our case the smoothing parameter must satisfy $\kappa\varepsilon > 1$, where ε is the exploration probability of the policy π provided by the following assumption.

Assumption 1 (Exploration) *There exists some $\varepsilon > 0$ such that for each $w \in \Sigma^\star$ the policy π satisfies $\pi(w^{\mathcal{A}}|w^{\mathcal{O}}) \geqslant \varepsilon^{|w|}$. In particular, at every time step each action $a \in \mathcal{A}$ is picked with probability at least ε.*

Before moving to the next section, where we provide finite-sample concentration results for the Hankel matrix estimated by Algorithm 3, we show that Algorithm 3 is consistent, that is it learns in expectation a WFA whose transition matrices are equivalent to those of the environment $\mathbb{A}$. The proof of the following lemma is provided in Appendix E.

Lemma 5 *The Hankel matrix $\widehat{H} = \widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi}$ computed in Algorithm 3 satisfies $\mathbb{E}[\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi}] = \tilde{H}^{\mathcal{U},\mathcal{V}}_t$, where $\tilde{H}^{\mathcal{U},\mathcal{V}}_t$ is a block of the Hankel matrix corresponding to the stochastic WFA $\tilde{\mathbb{A}}_t = \langle \tilde{\alpha}_t, \beta, \{A_\sigma\}\rangle$ where we introduced the modified vector $\tilde{\alpha}_t = (1/t)\sum_{s=0}^{t-1} \alpha^\top (A/\kappa)^s$. We denote by $\tilde{f}_t$ the function computed by $\tilde{\mathbb{A}}_t$.*

5.1. Concentration Results

Broadly speaking, a concentration bound for the estimation error $\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2$ can be obtained by following a proof strategy similar to the ones used in Theorems 3 and 5. However, almost all the bounds used in the previous proofs need to be reworked to account for (i) the effect of the extra dependencies introduced by the policy π, and (ii) the fact that the target automaton $\mathbb{A}$ to be learned is not a stochastic WFA in the sense of Section 4 but rather a conditional stochastic WFA.

Point (i) is addressed in our proof by introducing a "normalized" reference process $\rho_{\bar{\mathbb{A}}}$ corresponding to the coupling $\bar{\mathbb{A}} = \mathbb{A} \otimes \mathbb{A}_{unif}$ between the environment $\mathbb{A}$ and the uniform random policy that at each step takes each action independently with probability $1/|\mathcal{A}|$. Assuming the smoothing parameter satisfies $\kappa\varepsilon > 1$ for some exploration parameter ε (cf. Assumption 1), then $1/\kappa \leqslant 1/|\mathcal{A}|$. This observation is used, for example, to bound some variance terms in $\mathbb{E}[g(\xi)]$ by replace occurrences of $\tilde{f}_t$ with $\tilde{f}^{unif}_t$, the function computed by taking the Césaro average of the first t steps of $\bar{\mathbb{A}}$. Ultimately, this makes our bound depend not only on the mixing properties of $\rho_{\mathbb{B}}$, but also on those of the normalized process $\rho_{\bar{\mathbb{A}}}$ induced by the SWFA $\bar{\mathbb{A}}$. Incidentally, this argument is also used to address point (ii): by bounding quantities involving κ by quantities computed by a SWFA we can use again the arguments sketched in Section 4.1.

Pursuing the ideas above, and assuming that $\bar{\mathbb{A}}$ is $(\bar{C},\bar{\theta})$-geometrically mixing, we obtain the following bound which can be compared to the one in (5):

$$\mathbb{E}[g(\xi)]^2 \leqslant \frac{\tilde{m}}{t\varepsilon^L(1-1/(\kappa\varepsilon)^2)} + \frac{2\bar{m}}{t\varepsilon^{2L}}\left(L + \frac{\bar{C}}{1-\bar{\theta}}\right) ,$$

where $L = \max_{w\in\mathcal{U}\cdot\mathcal{V}}|w|$, $\tilde{m} = \sum_{u\in\mathcal{U},v\in\mathcal{V}} \tilde{f}_t(uv)$, and $\bar{m} = \sum_{u\in\mathcal{U},v\in\mathcal{V}} \tilde{f}^{unif}_t(uv)$.

Theorem 6 *Suppose that $\mathbb{B}$ is (C,θ)-geometrically mixing and $\bar{\mathbb{A}}$ is $(\bar{C},\bar{\theta})$-geometrically mixing. Suppose π satisfies Assumption 1 and the smoothing coefficient κ satisfies $\kappa\varepsilon > 1$. Let $d = \sum_{w\in\mathcal{U}\cdot\mathcal{V}}|w|_{\mathcal{U},\mathcal{V}}$, and define L, $\tilde{m}$, $\bar{m}$ as above. Then for any $\delta \in (0,1)$ we have*

$$\mathbb{P}\left[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2 > \sqrt{\frac{\tilde{m}}{t\varepsilon^L(1-\kappa^{-2}\varepsilon^{-2})}} + \sqrt{\frac{2\bar{m}}{t\varepsilon^{2L}}\left(L+\frac{\bar{C}}{1-\bar{\theta}}\right)} + \frac{C}{\theta(1-\theta)\varepsilon^L}\sqrt{\frac{2d\ln(1/\delta)}{t}}\right] \leqslant \delta .$$

On a final note we remark that the dependence on ε^L might be unavoidable due to inherent increase in variance produced by importance sampling estimators.

6. Conclusion

We present the first rigorous analysis of single-trajectory SVD-based spectral learning algorithms for sequential models with latent variables. Our analysis highlights the role of mixing properties of WFA and their relation with the geometry of the underlying state space. In the controlled case we obtain a result for control with *finite-state* policies, a much more general class than previously considered *memoryless* policies. In future work we will use our results to get upper confidence bounds on the predictions made by the learned environment with the goal of solving the full RL problem for PSR with complex control policies.

Acknowledgements

O.-A. M. acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-16-CE40-0002 (project BADASS).

References

Anandkumar, Anima, Foster, Dean P, Hsu, Daniel J, Kakade, Sham M, and Liu, Yi-Kai. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pp. 917–925, 2012.

Anandkumar, Animashree, Ge, Rong, Hsu, Daniel J, Kakade, Sham M, and Telgarsky, Matus. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Azizzadenesheli, Kamyar, Lazaric, Alessandro, and Anandkumar, Animashree. Reinforcement learning of pomdps using spectral methods. In *29th Annual Conference on Learning Theory*, pp. 193–256, 2016.

Bacon, Pierre-Luc, Balle, Borja, and Precup, Doina. Learning and planning with timing information in markov decision processes. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 111–120. AUAI Press, 2015.

Bailly, R., Habrard, A., and Denis, F. A spectral approach for probabilistic grammatical inference on trees. In *Algorithmic Learning Theory*, pp. 74–88, 2010.

Bailly, Raphaël. Quadratic weighted automata: Spectral algorithm and likelihood maximization. In *Proceedings of the 3rd Asian Conference on Machine Learning, ACML 2011, Taoyuan, Taiwan, November 13-15, 2011*, pp. 147–163, 2011.

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

Balle, Borja and Mohri, Mehryar. Spectral learning of general weighted automata via constrained matrix completion. In *Advances in neural information processing systems*, pp. 2168–2176, 2012.

Balle, Borja and Mohri, Mehryar. Learning weighted automata. In *International Conference on Algebraic Informatics*, pp. 1–21. Springer, 2015.

Balle, Borja, Quattoni, Ariadna, and Carreras, Xavier. A spectral learning algorithm for finite state transducers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 156–171. Springer, 2011.

Balle, Borja, Carreras, Xavier, Luque, Franco M, and Quattoni, Ariadna. Spectral learning of weighted automata. *Machine learning*, 96(1-2):33–63, 2014a.

Balle, Borja, Hamilton, William L, and Pineau, Joelle. Methods of moments for learning stochastic languages: Unified presentation and empirical comparison. In *ICML*, pp. 1386–1394, 2014b.

Balle, Borja, Gourdeau, Pascale, and Panangaden, Prakash. Bisimulation metrics for weighted automata. In *44rd International Colloquium on Automata, Languages, and Programming, ICALP*, 2017.

Berman, Abraham and Plemmons, Robert J. *Nonnegative matrices in the mathematical sciences*. SIAM, 1994.

Berstel, Jean and Reutenauer, Christophe. *Rational Series and Their Languages*. Springer, 1988.

Boots, Byron, Siddiqi, Sajid M, and Gordon, Geoffrey J. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(7):954–966, 2011.

Bowling, Michael, McCracken, Peter, James, Michael, Neufeld, James, and Wilkinson, Dana. Learning predictive state representations using non-blind policies. In *Proceedings of the 23rd international conference on Machine learning*, pp. 129–136. ACM, 2006.

Chazottes, J-R, Collet, Pierre, Külske, Christof, and Redig, Frank. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137 (1-2):201–225, 2007.

Cohen, S. B., Stratos, K., Collins, M., Foster, D. P., and Ungar, L. Experiments with spectral learning of latent-variable PCFGs. In *Proceedings of NAACL*, 2013.

Cohen, S. B., Stratos, K., Collins, M., Foster, D. P., and Ungar, L. Spectral learning of latent-variable PCFGs: Algorithms and sample complexity. *Journal of Machine Learning Research*, 2014.

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

Denis, François and Esposito, Yann. On rational stochastic languages. *Fundamenta Informaticae*, 86(1, 2):41–77, 2008.

Denis, François, Gybels, Mattias, and Habrard, Amaury. Dimension-free concentration bounds on hankel matrices for spectral learning. In *ICML*, pp. 449–457, 2014.

Fliess, M. Matrices de Hankel. *Journal de Mathématiques Pures et Appliquées*, 53, 1974.

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

Hamilton, William L, Fard, Mahdi Milani, and Pineau, Joelle. Efficient learning and planning with compressed predictive states. *Journal of Machine Learning Research*, 15(1):3395–3439, 2014.

Hsu, Daniel and Kakade, Sham M. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science*, 2013.

Hsu, Daniel, Kakade, Sham M, and Zhang, Tong. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Jaeger, Herbert. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

Kontorovich, Aryeh and Weiss, Roi. Uniform chernoff and dvoretzky-kiefer-wolfowitz-type inequalities for markov chains and related processes. *Journal of Applied Probability*, 51(04):1100–1113, 2014.

Kontorovich, Leonid Aryeh, Ramanan, Kavita, et al. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

Kontoyiannis, Ioannis and Meyn, Sean P. Geometric ergodicity and the spectral gap of non-reversible markov chains. *Probability Theory and Related Fields*, 154(1-2):327–339, 2012.

Kulesza, Alex, Jiang, Nan, and Singh, Satinder. Low-rank spectral learning with weighted loss functions. In *Artificial Intelligence and Statistics*, pp. 517–525, 2015.

Langer, Lucas, Balle, Borja, and Precup, Doina. Learning multi-step predictive state representations. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 1662–1668, 2016.

Luque, Franco M, Quattoni, Ariadna, Balle, Borja, and Carreras, Xavier. Spectral learning for non-deterministic dependency parsing. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 409–419. Association for Computational Linguistics, 2012.

Rosenthal, Jeffrey S. Convergence rates for markov chains. *Siam Review*, 37(3):387–405, 1995.

Shaban, Amirreza, Farajtabar, Mehrdad, Xie, Bo, Song, Le, and Boots, Byron. Learning latent variable models by improving spectral solutions with exterior point method. In *UAI*, pp. 792–801, 2015.

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

Thon, Michael and Jaeger, Herbert. Links between multiplicity automata, observable operator models and predictive state representations–a unified learning framework. *Journal of Machine Learning Research*, 16:103–147, 2015.

Spectral Learning from a Single Trajectory under Finite-State Policies (Supplementary Material)

Borja Balle **Odalric-Ambrym Maillard**

A. Mixing Properties of Probabilistic Automata

Lemma 2 (η-mixing for PFA) *Let $\mathbb{A}$ be PFA and assume that it is (C,θ)-geometrically mixing in the sense that for some constants $C > 0, \theta \in (0,1)$ we have*

$$\forall t \in \mathbb{N}, \quad \mu_t^{\mathbb{A}} = \sup_{\alpha,\alpha'} \frac{\|\alpha A^t - \alpha' A^t\|_1}{\|\alpha - \alpha'\|_1} \leqslant C\theta^t \,,$$

where the supremum is over all probability vectors. Then we have $\eta_{\rho_{\mathbb{A}}} \leqslant C/(\theta(1-\theta))$.

Proof of Lemma 2:

We start by controlling the term η, defined by

$$\eta_{\rho_{\mathbb{A}}} = 1 + \max_{1<i<t} \sum_{j=i+1}^{t} \eta_{i,j} \,,$$

We proceed similarly to Lemma 7 of (Kontorovich & Weiss, 2014). By definition of the total variation norm $\|\cdot\|_{TV}$,

$$\eta_{i,j} = \frac{1}{2} \sup_{u \in \Sigma^{i-1}, \sigma, \sigma' \in \Sigma} \sup_{Z \subseteq \Sigma^{t-j+1}} \left| \frac{\alpha^\top A_u A_\sigma A^{j-i-1} A_Z \beta}{\alpha^\top A_u A_\sigma \beta} - \frac{\alpha^\top A_u A_{\sigma'} A^{j-i-1} A_Z \beta}{\alpha^\top A_u A_{\sigma'} \beta} \right| ,$$

where $A_Z = \sum_{z \in Z} A_z$. At this point, it is convenient to introduce the vector $\alpha_{u,\sigma}^\top = \frac{\alpha^\top A_u A_\sigma}{\alpha^\top A_u A_\sigma \beta}$. Indeed, we then have the rewriting

$$\begin{aligned} \eta_{i,j} &= \frac{1}{2} \sup_{u \in \Sigma^{i-1}, \sigma, \sigma' \in \Sigma} \sup_{Z \subseteq \Sigma^{t-j+1}} \left| (\alpha_{u,\sigma} - \alpha_{u,\sigma'})^\top A^{j-i-1} A_Z \beta \right| \\ &\leqslant \frac{1}{2} \sup_{u \in \Sigma^{i-1}, \sigma, \sigma' \in \Sigma} \sup_{Z \subseteq \Sigma^{t-j+1}} \|(\alpha_{u,\sigma} - \alpha_{u,\sigma'})^\top A^{j-i-1}\|_1 \|A_Z \beta\|_\infty \end{aligned}$$

where we used a simple application of Hölder inequality. Since $\mathbb{A}$ is a PFA, we note that $\|A_Z\beta\|_\infty \leqslant 1$ because $\|\sum_{|z|=t-j+1} A_z \beta\|_\infty = 1$ and all the entries are non-negative. Also note that $\alpha_{u,\sigma}^\top \beta = \|\alpha_{u,\sigma}^\top\|_1 = 1$. Thus $\|\alpha_{u,\sigma} - \alpha_{u,\sigma'}\| \leqslant 2$. We deduce from these steps that

$$\eta_{i,j} \leqslant \sup_{\alpha,\alpha'} \frac{\|(\alpha - \alpha')^\top A^{j-i-1}\|_1}{\|\alpha - \alpha'\|_1} \,,$$

where the supremum is taken over all α, α' that are probability vectors. We note that the later quantity is precisely the definition of the coefficient $\mu_{j-i-1}^{\mathbb{A}}$. Assuming (C,θ)-geometrically mixing, that is $\mu_j^{\mathbb{A}} \leqslant C\theta^j$ for all j, this implies that

$$\eta_{i,j} \leqslant C\theta^{j-i-1} \,.$$

We then deduce that

$$\eta_{\rho_{\mathbb{A}}} \leqslant 1 + C \max_{1<i<t} \sum_{j=i+1}^{t} \theta^{j-i-1} \leqslant \frac{C}{\theta}(1 + \sum_{j=1}^{t-2} \theta^j) = \frac{C}{\theta}\frac{1-\theta^{t-1}}{1-\theta} \leqslant \frac{C}{\theta(1-\theta)}\,. \qquad \square$$

The following result provides a control of the $\eta_{\rho_{\mathbb{A}}}$ coefficients, and shows this can be made explicit in specific cases.

Corollary 1 *Let $\mathbb{A}$ be PFA with n states and assume that its matrix A has a spectral gap, that is $|\lambda_2(A)| < 1$. then there exists C such that $\eta_{\rho_{\mathbb{A}}} \leqslant \frac{C}{|\lambda_2(A)|(1-|\lambda_2(A)|)}$. When the corresponding chain is further aperiodic, irreducible and reversible, we further have $C \leqslant \sqrt{n}$.*

Proof of Corollary 1:

The first part of the result is folklore, and can be proven using some tedious steps involving the Jordan decomposition of the matrix see e.g. Fact 3 in (Rosenthal, 1995).

When the chain is irreducible, aperiodic and more importantly reversible, the spectral gap admits the following characterization, see Lemma 2.2 from (Kontoyiannis & Meyn, 2012):

$$\gamma_2(A) = \lambda_2(A) = \sup\left\{ \frac{\|A\nu\|_2}{\|\nu\|_2} : \nu \text{ s.t. } \|\nu\|_2 \neq 0, \nu^\top \mathbf{1} = 0 \right\}.$$

Thus, from $\lambda_2(A) < 1$ together with a change of norm from $\|\cdot\|_1$ to $\|\cdot\|_2$ and a standard argument (closely following that of Lemma 7), we obtain that

$$\mu_j^{\mathbb{A}} \leqslant C|\lambda_2|^j\,,$$

where $C = \max_{x\in\mathbb{R}^n} \frac{||x||_1}{||x||_2} = \sqrt{n}$. $\square$

We end this section with a more technical lemma, that is useful to decompose terms in the proof of Theorem 3.

Lemma 6 (Mixing times of PFA) *Let $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\}\rangle$ be a PFA. Then, for any $s \geqslant s' \in \mathbb{N}$ it holds*

$$\|A^{s'} - \beta\alpha^\top A^s\|_\infty \leqslant 2\mu_{s'}^{\mathbb{A}}\,.$$

Proof of Lemma 6:

Let denote $\alpha_s^\top = \alpha^\top A^s$. We need to bound $\|A^{s'} - \beta\alpha_s^\top\|_\infty$. Recall that for any matrix M the $\|\cdot\|_\infty$-induced norm is given by $\|M\|_\infty = \max_i \sum_j |M(i,j)| = \max_i \|M(i,:)\|_1$. The ith row of $A^{s'} - \beta\alpha_s^\top$ is given by $e_i^\top A^{s'} - \alpha_s^\top$, where e_i is the ith column of the identity matrix. In particular, $e_i^\top A^{s'}$ is the distribution over states after starting in state i and running the chain for s' steps, and $\alpha_s^\top$ is the distribution over states starting from the distribution given by α and running the chain for s steps. The latter can also be rewritten as $\alpha_s^\top = \alpha^\top A^s = \alpha^\top A^{s-s'} A^{s'} = \alpha_{s-s'}^\top A^{s'}$, where $\alpha_{s-s'}^\top$ is again a distribution over states. Therefore we obtain the desired bound, since:

$$\begin{aligned}
\|A^{s'} - \beta\alpha_s^\top\|_\infty &= \max_{i\in[n]} \|e_i^\top A^{s'} - \alpha_{s-s'}^\top A^{s'}\|_1 \\
&\leqslant \sup_{\alpha_1,\alpha_2} \frac{\|\alpha_1^\top A^{s'} - \alpha_2^\top A^{s'}\|_1}{\|\alpha_1 - \alpha_2\|_1} \|e_i - \alpha_{s-s'}\|_1 \\
&\leqslant 2\mu_{s'}^{\mathbb{A}}\,.
\end{aligned}$$

$\square$

B. Geometry of Stochastic Weighted Automata

Lemma 3 (claim (i)) *For any $w \in \Sigma^\star$ we have $\|A_w\beta\|_\beta \leqslant 1$.*

Proof of Lemma 3 (claim (i)):

We shall use the cone monotonicity property of $\|\cdot\|_\beta$, which says that $0 \leqslant_{\mathcal{K}} u \leqslant_{\mathcal{K}} v$ implies $\|u\|_\beta \leqslant \|v\|_\beta$. First note that by construction of $\mathcal{K}$ we have $0 \leqslant_{\mathcal{K}} A_w\beta$. If we show that $A_w\beta \leqslant_{\mathcal{K}} \beta$ also holds, then cone monotonicity implies $\|A_w\beta\|_\beta \leqslant \|\beta\|_\beta = 1$.

To prove the claim note that because β is an eigenvector of A of eigenvalue 1 we have $\beta = A^t\beta = \sum_{|w|=t} A_w\beta$. Therefore, $\beta - A_w\beta = \sum_{|w'|=|w|, w'\neq w'} A_{w'}\beta$ which is a vector in $\mathcal{K}$ because convex cones are closed under non-negative linear combinations, and we conclude that $A_w\beta \leqslant_{\mathcal{K}} \beta$. □

Lemma 3 (claim (ii)) *For any $w \in \Sigma^\star$ we have $\|\alpha^\top A_w\|_{\beta,*} = \alpha^\top A_w\beta$.*

Proof of Lemma 3 (claim (ii)):

By unrolling the definitions of the dual norm and B_β we get

$$\|\alpha^\top A_w\|_{\beta,*} = \sup_{-\beta \leqslant_{\mathcal{K}} v \leqslant_{\mathcal{K}} \beta} \alpha^\top A_w v \ .$$

Now note that for any v such that $\beta - v \in \mathcal{K}$ we have

$$\alpha^\top A_w v = \alpha^\top A_w\beta - \alpha^\top A_w(\beta - v) \leqslant \alpha^\top A_w\beta \ ,$$

where we used that $\beta - v \in \mathcal{K}$ implies $A_w(\beta - v) \in \mathcal{K}$ implies $\alpha^\top A_w(\beta - v) \geqslant 0$. Since $-\beta \leqslant_{\mathcal{K}} \beta \leqslant_{\mathcal{K}} \beta$, the supremum in the definition of $\|\alpha^\top A_w\|_{\beta,*}$ is attained at $v = \beta$ and the result follows. □

C. Mixing Properties of Stochastic Weighted Automata

Lemma 4 (η-mixing for SWFA) *Let $\mathbb{A}$ be SWFA and assume that it is (C, θ)-geometrically mixing in the sense that for some $C \geqslant 0, \theta \in (0, 1)$,*

$$\mu_t^{\mathbb{A}} = \sup_{\alpha_0, \alpha_1 : \alpha_0^\top \beta = \alpha_1^\top \beta = 1} \frac{\|\alpha_0^\top A^t - \alpha_1^\top A^t\|_{\beta,*}}{\|\alpha_0 - \alpha_1\|_{\beta,*}} \leqslant C\theta^t \ .$$

Then the η-mixing coefficient satisfies

$$\eta_{\rho_{\mathbb{A}}} \leqslant \frac{C}{\theta(1-\theta)} \, .$$

Proof of Lemma 4:

The proof follows the same initial steps as for Lemma 2. Introducing the vector $\alpha_{u,\sigma}^\top = \frac{\alpha^\top A_u A_\sigma}{\alpha^\top A_u A_\sigma \beta}$, we then have the rewriting

$$\begin{aligned}\eta_{i,j} &= \frac{1}{2} \sup_{u\in\Sigma^{i-1},\sigma,\sigma'\in\Sigma} \sup_{Z\subseteq\Sigma^{t-j+1}} \left|(\alpha_{u,\sigma}-\alpha_{u,\sigma'})^\top A^{j-i-1} A_Z \beta\right| \\ &\leqslant \frac{1}{2} \sup_{u\in\Sigma^{i-1},\sigma,\sigma'\in\Sigma} \sup_{Z\subseteq\Sigma^{t-j+1}} \|(\alpha_{u,\sigma}-\alpha_{u,\sigma'})^\top A^{j-i-1}\|_{\beta,*} \|A_Z\beta\|_\beta\end{aligned}$$

where we used a simple application of Hölder inequality and the norm induced by β. Since $\mathbb{A}$ is a SWFA, the same argument in the proof of Lemma 3 (i) can be used to show that $\|A_Z\beta\|_\beta \leqslant 1$ for any $Z \subseteq \Sigma^{t-j+1}$. On the other hand, from Lemma 3 (ii) we have $1 = \alpha_{u,\sigma}^\top \beta = \|\alpha_{u,\sigma}^\top\|_{\beta,*}$. Thus $\|\alpha_{u,\sigma} - \alpha_{u,\sigma'}\|_{\beta,*} \leqslant 2$. We deduce from these steps that

$$\eta_{i,j} \leqslant \sup_{\alpha,\alpha'} \frac{\|(\alpha-\alpha')^\top A^{j-i-1}\|_{\beta,*}}{\|\alpha-\alpha'\|_{\beta,*}}\,,$$

where the supremum is taken over all α, α' that satisfy $\alpha^\top\beta = 1$ We note that the later quantity is precisely the definition of the coefficient $\mu_{j-i-1}^{\mathbb{A}}$. We then conclude similarly to the proof of Lemma 2. □

Lemma 7 (Geometrical mixing of weighted automata) *Let $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\}\rangle$ be a stochastic WFA, $A = \sum_\sigma A_\sigma$, and*

$$\gamma_\beta(A) = \sup\left\{ \frac{\|A\nu\|_{\beta,*}}{\|\nu\|_{\beta,*}} : \nu \text{ s.t. } \|\nu\|_{\beta,*} \neq 0, \nu^\top\beta = 0 \right\}.$$

be its spectral gap with respect to β. It holds that

$$\mu_t^{\mathbb{A}} = \sup_{\alpha_0,\alpha_1:\alpha_0^\top\beta=\alpha_1^\top\beta=1} \frac{\|\alpha_0^\top A^t - \alpha_1^\top A^t\|_{\beta,*}}{\|\alpha_0-\alpha_1\|_{\beta,*}} \leqslant \gamma_\beta(A)^t \ .$$

Proof of Lemma 7:

To this end, note that if α_0, α_1 are are such that $\alpha_0^\top\beta = \alpha_1^\top\beta = 1$, then $v = \alpha_0 - \alpha_1$ is such that $v^\top\beta = 0$. A crucial remark is that since A is a weighted automaton matrix, $\alpha_0^\top A\beta = \alpha_1^\top A\beta = 1$ and thus $w = A(\alpha_0-\alpha_1)$ also satisfies $w^\top\beta = 0$. Likewise, $(\alpha_0-\alpha_1)^\top A^t\beta = 0$ for all $t \in \mathbb{N}$.

A second remark is that if $\|A^s v\|_{\beta,*} = 0$ for some $s < t$, then $\|A^t v\|_{\beta,*} = 0$. Thus, we can restrict to v such that $\|A^s v\|_{\beta,*} \neq 0$ for all $s \leqslant t$. Then, it comes for such $v = \alpha_0 - \alpha_1$,

$$\frac{\|A^t\nu\|_{\beta,*}}{\|\nu\|_{\beta,*}} = \frac{\|AA^{t-1}\nu\|_{\beta,*}}{\|A^{t-1}\nu\|_{\beta,*}} \cdots \frac{\|A\nu\|_{\beta,*}}{\|\nu\|_{\beta,*}} \leqslant \gamma_\beta(A)^t\,.$$

For the last inequality, we used the fact that since A is a weighted automaton matrix, and $v = \alpha_0 - \alpha_1$, then $v^\top A^s\beta = 0$ for all s. This guarantees that indeed $\frac{\|AA^s\nu\|_{\beta,*}}{\|A^s\nu\|_{\beta,*}} \leqslant \gamma_\beta(A)$ for all s. □

Lemma 8 (Mixing times of SWA) *Let $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\}\rangle$ be a SWFA. Then, for all $s \geqslant s' \in \mathbb{N}$ it holds*

$$\|A^{s'} - \beta\alpha^\top A^s\|_\beta \leqslant 2\mu_{s'}^{\mathbb{A}}\,.$$

Proof of Lemma 8:

Let $\alpha_s^\top = \alpha^\top A^s$. To prove this we proceed as follows:

$$\begin{aligned}\|A^{s'} - \beta\alpha_s^\top\|_\beta &= \sup_{\|v\|_\beta \leqslant 1} \|(A^{s'} - \beta\alpha_s^\top)v\|_\beta \\ &= \sup_{\|v\|_\beta \leqslant 1} \sup_{\|u\|_{\beta,*} \leqslant 1} u^\top (A^{s'} - \beta\alpha_s^\top)v \\ &= \sup_{\|u\|_{\beta,*} \leqslant 1} \|u^\top (A^{s'} - \beta\alpha_s^\top)\|_{\beta,*} \\ &= \sup_{\|u\|_{\beta,*} \leqslant 1} \|u^\top (A^{s'} - \beta\alpha_{s-s'}^\top A^{s'})\|_{\beta,*} \ .\end{aligned}$$

Next we note that for any u such that $\|u\|_{\beta,*} \leqslant 1$ we have $|u^\top \beta| \leqslant 1$, so:

$$\begin{aligned}\|u^\top \beta\alpha_t^\top\|_{\beta,*} &= |u^\top \beta| \|\alpha_t^\top\|_{\beta,*} \\ &\leqslant \|\alpha_t^\top\|_{\beta,*} \ .\end{aligned}$$

Furthermore, the same argument we used to show that $\|\alpha^\top A_x\|_{\beta,*} = \alpha^\top A_x \beta$ implies that $\|\alpha_t^\top\|_{\beta,*} = \|\alpha^\top A^t\|_{\beta,*} = \alpha^\top A^t \beta = 1$. Therefore, we see that $\|u\|_{\beta,*} \leqslant 1$ implies $\|u^\top \beta\alpha_t^\top\|_{\beta,*} \leqslant 1$, and we get the inequality

$$\begin{aligned}\|A^{s'} - \beta\alpha_s^\top\|_\beta &\leqslant \sup_{\|u_1\|_{\beta,*} \leqslant 1} \sup_{\|u_2\|_{\beta,*} \leqslant 1} \|u_1^\top A^{s'} - u_2^\top A^{s'}\|_{\beta,*} \\ &\leqslant \mu_{s'}^{\mathbb{A}} \sup_{\|u_1\|_{\beta,*} \leqslant 1} \sup_{\|u_2\|_{\beta,*} \leqslant 1} \|u_1 - u_2\|_{\beta,*} \\ &\leqslant 2\mu_{s',\beta}^{\mathbb{A}} \ .\end{aligned}$$

□

D. Single-Trajectory Concentration Inequalities for Probabilistic Automata

Theorem 2 (Single-trajectory, entry-wise concentration) *Let $\mathbb{A}$ be a PFA that is (C, θ)-geometrically mixing, and $\xi \sim \rho_{\mathbb{A}} \in \mathcal{P}(\Sigma^\omega)$ a trajectory of observations. Then for any $u \in \mathcal{U}, v \in \mathcal{V}$ and $\delta \in (0,1)$ it holds*

$$\mathbb{P}\left(\widehat{H}_{t,\xi}^{\mathcal{U},\mathcal{V}}(u,v) - \bar{H}_t^{\mathcal{U},\mathcal{V}})u,v) > \frac{|uv|C}{\theta(1-\theta)}\sqrt{\left(1 + \frac{|uv|-1}{t}\right)\frac{\ln(1/\delta)}{2t}}\right) \leqslant \delta\,.$$

Proof of Theorem 2:

We control $\eta_{\rho_{\mathbb{A}}}$ by a direct application of Lemma 2.

Control of $\|g\|_{Lip}$: Let us fix $u \in \mathcal{U}, v \in \mathcal{V}$ define $g(\xi) = t\widehat{H}_{t,\xi}^{\mathcal{U},\mathcal{V}}(u,v)$. We first control the regularity of f.

To this end, let ξ' be a trajectory $\xi' = x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_\ell$ that only differs by one element from ξ, say at position k. Then, we get for any u, v

$$\begin{aligned}g(\xi) - g(\xi') &= \sum_{s=1}^{t} (\mathbb{I}\{x_s \ldots x_{s+|uv|-1} = uv\} - \mathbb{I}\{x_s \ldots x'_k \ldots x_{s+|uv|-1} = uv\}) \\ &\leqslant |\{s \in [1,t] : k \in [s : s + |uv| - 1]\}|\,.\end{aligned}$$

Now, in order to bound $|\{s \in [1,t] : k \in [s : s+|uv|-1]\}|$ note that $k \in [s : s+|uv|-1]$ if and only if $s \leqslant k \leqslant s+|uv|-1$. From the first inequality we see that $s \leqslant k$, and from the second one $s \geqslant k-|uv|+1$. Combined with the restrictions on s, this means that

$$|\{s \in [1,t] : k \in [s : s+|uv|-1]\}| = |[\max\{1, k-|uv|+1\}, \min\{k,t\}]| \leqslant |uv|\,,$$

which show that $\|g\|_{Lip} \leqslant |uv|$.

Combining the two quantities Combining these two results, and noting that $t+|uv|-1$ symbols appears in $g(\xi)$, we deduce that $\forall \varepsilon > 0$,

$$\mathbb{P}\bigg(t(\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi}(u,v) - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}(u,v)) > |uv|(t+|uv|-1)\varepsilon\bigg) \leqslant \exp\bigg(-\frac{2(t+|uv|-1)\theta^2(1-\theta)^2\varepsilon^2}{C^2}\bigg)\,,$$

or equivalently, for all $\delta \in (0,1)$,

$$\mathbb{P}\bigg(\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi}(u,v) - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}(u,v) > \frac{\sqrt{t+|uv|-1}|uv|}{t}\frac{C}{\theta(1-\theta)}\sqrt{\frac{\ln(1/\delta)}{2t}}\bigg) \leqslant \delta\,. \qquad \square$$

The proof of following result is more challenging.

Theorem 3 (Single-trajectory, matrix-wise) *Let $\rho_{\mathbb{A}} \in \mathcal{P}(\Sigma^\omega)$ be as in Theorem 2 and define the probability mass $m^{\mathcal{U},\mathcal{V}} = \sum_{u\in\mathcal{U},v\in\mathcal{V}} \bar{f}_t(uv)$. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\bigg(\|\widehat{H}^{\mathcal{U},\mathcal{V}}_t - \bar{H}^{\mathcal{U},\mathcal{V}}_t\|_2 \geqslant \bigg(\sqrt{L}+\sqrt{\frac{2C}{1-\theta}}\bigg)\sqrt{\frac{2m^{\mathcal{U},\mathcal{V}}}{t}}$$
$$+\frac{2LC}{\theta(1-\theta)}\sqrt{\Big(1+\frac{L-1}{t}\Big)\frac{\min\{|\mathcal{U}||\mathcal{V}|, 2n_{\mathcal{U}}n_{\mathcal{V}}\}\ln(1/\delta)}{2t}}\bigg) \leqslant \delta\,.$$

Proof of Theorem 3:

Let us introduce the function $g(\xi) = \|\widehat{H}^{\mathcal{U},\mathcal{V}}_t - \bar{H}^{\mathcal{U},\mathcal{V}}_t\|_2$. We first control $\|g\|_{Lip}$ then $\mathbb{E}[g(\xi)]$, before applying Theorem 1.

Step 1: Control of $\|g\|_{Lip}$. In this step, we show that

$$\|g\|_{Lip} \leqslant \frac{L}{t}\sqrt{\min\{|\mathcal{U}||\mathcal{V}|, 2n_{\mathcal{U}}n_{\mathcal{V}}\}}$$

where $L = \max_{u\in\mathcal{U},v\in\mathcal{V}}|uv|$ denote the maximal length of words in $U\cdot V$ and $n_{\mathcal{U}} = |\ell \in [0,L] : |U_\ell| > 0|$, $n_{\mathcal{V}} = |\ell \in [0,L] : |V_\ell| > 0|$, denote the number lengths such that the set $U_\ell = \{u \in \mathcal{U} : |u| = \ell\}$ (respectively $V_\ell = \{v \in \mathcal{V} : |v| = \ell\}$) is non empty. Note that the second term in the min can be exponentially smaller than the first. For example, taking $U = V = \Sigma^{\leqslant L/2}$ we have $|U||V| = \Theta(|\Sigma|^L)$ while $n_{\mathcal{U}}n_{\mathcal{V}} = \Theta(L^2)$.

Step 1.1. Let $\xi' \sim p$ be a trajectory $\xi' = x_1,\dots,x_{k-1},x'_k,x_{k+1},\dots,x_\ell$ that only differs by one

element from ξ, say at position k. We note that

$$\begin{aligned}
&\left| \|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}\|_2 - \|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi'} - \bar{H}^{\mathcal{U},\mathcal{V}}_{t}\|_2 \right| \leqslant \|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi'}\|_2 \\
&= \sup_{q\in\mathbb{R}^V} \frac{1}{t\|q\|_2} \sqrt{\sum_{u\in\mathcal{U}} \Big(\sum_{v\in\mathcal{V}} \sum_{s=1}^{t} \big(\mathbb{I}\{x_s \dots x_{s+|uv|-1} = uv\} - \mathbb{I}\{x_s \dots x'_k \dots x_{s+|uv|-1} = uv\} \big) q_v \Big)^2} \\
&\leqslant \sup_{q\in\mathbb{R}^V} \frac{1}{t\|q\|_2} \sqrt{\sum_{u\in\mathcal{U}} \Big(\sum_{v\in\mathcal{V}} |uv| q_v \Big)^2} \leqslant \sup_{q\in\mathbb{R}^V} \frac{1}{t\|q\|_2} \sqrt{\sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} |uv|^2} \sqrt{\sum_{v\in\mathcal{V}} q_v^2} \,.
\end{aligned}$$

Let $L = \max_{u\in\mathcal{U}, v\in\mathcal{V}} |uv|$. A simple bound is then $\|g\|_{Lip} \leqslant \frac{L}{t}\sqrt{|\mathcal{U}||\mathcal{V}|}$, which is essentially optimal if all words uv, $u \in \mathcal{U}, v \in \mathcal{V}$ have same length.

Step 1.2. A more refined bound may be helpful in case many words have length $|uv|$ much smaller than L. To his end, let us write $\widehat{H}^{\mathcal{U},\mathcal{V}}_t = \frac{1}{t}\sum_{s=1}^t M_s$ with $M_s(u,v) = b_{s,uv} = \mathbb{I}\{x_s \dots x_{s+|uv|-1} = uv\}$. Similarly, let $H^{\mathcal{U},\mathcal{V}}_{t,\xi'} = \frac{1}{t}\sum_{s=1}^t M'_s$ with the obvious definition. Now, by the same argument we used to bound $\|g\|_{Lip}$ in the entry-wise case, we have

$$t\left(\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi'}\right) = \sum_{s=k-L+1}^{k} M_s - M'_s = \sum_{s=k-L+1}^{k} \Delta_s \,,$$

since for $s < k - L + 1$ or $s > k$ we must have $M_s = M'_s$.

Now let us partition the sets U and V as disjoint unions of sets with strings of the same length. That is, we write $U = \cup_{\ell=0}^L U_\ell$ with $U_\ell = U \cap \Sigma^\ell$, and $V = \cup_{\ell=0}^L V_\ell$ with analogous definitions. This allows us to write $M_s \in \{0,1\}^{U\times V}$ as a block matrix $M_s = (M_s^{i,j})_{0\leqslant i,j\leqslant L}$ with $M_s^{i,j} \in \{0,1\}^{U_i\times V_j}$.

For simplicity of notation, in the sequel we are assuming that $U_\ell, V_\ell \neq \emptyset$ for all $0 \leqslant \ell \leqslant L$, but the argument remains the same after we remove the empty sets of rows and columns. Note that by definition we have $M_s^{i,j}(u,v) = \mathbb{I}\{x_s \dots x_{s+i+j-1} = uv\}$ for any $u \in \mathcal{U}_i$ and $v \in \mathcal{V}_j$. This implies that each of the block matrices $M_s^{i,j}$ contains *at most one non-zero entry*.

If we make analogous definitions and write $M'_s = (M_s'^{i,j})_{0\leqslant i,j\leqslant L}$, then we obtain a block decomposition for $\Delta_s = M_s - M'_s = (\Delta_s^{i,j})_{0\leqslant i,j\leqslant L}$ where each block is either:

1. zero,
2. a $\{0,1\}$-matrix with a single 1,
3. a $\{0,-1\}$-matrix with a single -1,
4. a $\{0,1,-1\}$-matrix with a single 1 and a single -1.

In any of these cases one can see that the bound $\|\Delta_s^{i,j}\|_2 \leqslant \|\Delta_s^{i,j}\|_F \leqslant \sqrt{2}$ is always satisfied. Therefore, we have $\|\Delta_s\|_2^2 \leqslant \|\Delta_s\|_F^2 = \sum_{i,j} \|\Delta_s^{i,j}\|_F^2 \leqslant 2n_{\mathcal{U}}n_{\mathcal{V}}$, where

$$\begin{aligned}
n_{\mathcal{U}} &= |\ell \in [0,L] : |U_\ell| > 0| \,, \\
n_{\mathcal{V}} &= |\ell \in [0,L] : |V_\ell| > 0| \,.
\end{aligned}$$

By plugging these estimates into $\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi'}$ we finally get

$$\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi'}\|_2 \leqslant \frac{L\sqrt{2n_{\mathcal{U}}n_{\mathcal{V}}}}{t} \,.$$

Therefore we obtain the bound $\|g\|_{Lip} \leqslant (L\sqrt{2n_{\mathcal{U}}n_{\mathcal{V}}})/t$.

Control of $\mathbb{E}[g(\xi)]$. We now want to control the following quantity $\mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2]$. More precisely, we show in this step that

$$\mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2]^2 \;\leqslant\; \frac{O\left(L^2 + \frac{1}{1-\theta}\right)\left(\sum_{u\in\mathcal{U},v\in\mathcal{V}} \bar{f}_t(uv)\right)}{t}\;,$$

where $L = \max_{w\in\mathcal{U}\cdot V}|w|$, and $\bar{f}_t(w) = \frac{1}{t}\sum_{s=1}^t f_s(w)$, where $f_s(w) = \mathbb{P}[\xi \in \Sigma^{s-1}w\Sigma^\omega]$.

Step 2.1. Let $q \in \mathbb{R}^V$ be a unit vector ($\|q\|_2 = 1$.) Then, by Jensen's inequality, the norm of $\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}$ is controlled by its Frobenius norm

$$\begin{aligned}
\mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2]^2 &\leqslant \mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2^2] \\
&\leqslant \mathbb{E}[\sum_{u\in\mathcal{U}}\Big(\sum_{v\in\mathcal{V}}(\widehat{H}_t^{\mathcal{U},\mathcal{V}}(u,v) - \bar{H}_t^{\mathcal{U},\mathcal{V}}(u,v))q_v\Big)^2] \\
&\leqslant \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}}\mathbb{E}[(\widehat{H}_t^{\mathcal{U},\mathcal{V}}(u,v) - \bar{H}_t^{\mathcal{U},\mathcal{V}}(u,v))^2] \\
&= \sum_{w\in\mathcal{U}\cdot V}|w|_{\mathcal{U},\mathcal{V}}\mathbb{E}[(\widehat{f}_t(w) - \bar{f}_t(w))^2]\;,
\end{aligned}$$

where $U\cdot V$ is the set of all words of the form $u\cdot v$ with $u\in\mathcal{U}$ and $v\in\mathcal{V}$; $|w|_{\mathcal{U},\mathcal{V}} = |(u,v)\in \mathcal{U}\times V : u\cdot v = w|$, and $\widehat{f}_t(w) = \frac{1}{t}\sum_{s=1}^t b_{s,w}$ with the notation defined above. We also use $\bar{f}_t(w) = \mathbb{E}[\widehat{f}_t(w)] = \frac{1}{t}\sum_{s=1}^t f_s(w)$, where $f_s(w) = \mathbb{P}[\xi\in\Sigma^{s-1}w\Sigma^\omega]$. This implies that we have a sum of variances, and each of them can be written as

$$\mathbb{E}[(\widehat{f}_t(w) - \bar{f}_t(w))^2] = \mathbb{E}[\widehat{f}_t(w)^2] - \bar{f}_t(w)^2\;.$$

An important first observation is that we can write $f_s(w) = \alpha^\top A^{s-1}A_w\beta$. Furthermore, it follows from A being a probabilistic automaton that $\sum_{|w|=l} f_s(w) = 1$ for all s and l. This suggests that we group the terms in the sum over $W = U\cdot V$ by length, so we write $W_l = W\cap\Sigma^l$ and define $L_l = \max_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}$ the maximum number of ways to write a string of length l in W as a product of a prefix in U and a suffix in V. Note that we always have $L_l \leqslant l+1$. Henceforth, we want to control the following terms for all possible values of l:

$$\sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}\Big(\mathbb{E}[\widehat{f}_t(w)^2] - \bar{f}_t(w)^2\Big) = \frac{1}{t^2}\sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}\left[\mathbb{E}\left[\left(\sum_{s=1}^t b_{s,w}\right)^2\right] - \left(\sum_{s=1}^t f_s(w)\right)^2\right].$$

Step 2.2. Let us focus on each of the quadratic terms. On the one hand, it holds

$$\left(\sum_{s=1}^t f_s(w)\right)^2 \;=\; \sum_{s=1}^t f_s(w)^2 + 2\sum_{1\leqslant s<s'\leqslant t} f_s(w)f_{s'}(w)\,,$$

while other on the other hand, we get

$$\mathbb{E}\left[\left(\sum_{s=1}^t b_{s,w}\right)^2\right] \;=\; \sum_{s=1}^t \mathbb{E}[b_{s,w}^2] + 2\sum_{1\leqslant s<s'\leqslant t}\mathbb{E}[b_{s,w}b_{s',w}]\,.$$

Hence this enables to derive the following bound

$$\begin{aligned}
\mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2]^2 \;\leqslant\; &\frac{1}{t^2}\sum_{l=0}^\infty\sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}\Big[\sum_{s=1}^t(1-f_s(w))f_s(w) \\
&+2\sum_{1\leqslant s<s'\leqslant t}\Big(\mathbb{E}[b_{s,w}b_{s',w}] - f_s(w)f_{s'}(w)\Big)\Big]\,. \qquad (6)
\end{aligned}$$

Step 2.3. In order to control the first term in (6), we remark that

$$\begin{aligned}\sum_{l=0}^{\infty}\sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}\sum_{s=1}^{t}(1-f_s(w))f_s(w) &= \sum_{u\in\mathcal{U},v\in\mathcal{V}}\sum_{s=1}^{t}(1-f_s(uv))f_s(uv)\\ &\leqslant \sum_{u\in\mathcal{U},v\in\mathcal{V}}\sum_{s=1}^{t}f_s(uv)\\ &= t\sum_{u\in\mathcal{U},v\in\mathcal{V}}\bar{f}_t(uv)\,. \end{aligned}\tag{7}$$

Step 2.4. We thus focus on controlling the remaining "cross"-term in (6) and to this end we study, for $w\in W_l$, the quantity

$$\mathbb{E}[b_{s,w}b_{s',w}]-f_s(w)f_{s'}(w) \;=\; \mathbb{P}[\xi\in\Sigma^{s-1}\Sigma_w^{s'-s}\Sigma^\omega]-(\alpha^\top A^{s-1}A_w\beta)(\alpha^\top A^{s'-1}A_w\beta)\,,$$

where we introduced for convenience the set $\Sigma_w^{s'-s}=w\Sigma^{s'-s}\cap\Sigma^{s'-s}w$. Introducing as well the vectors $\alpha_{s-1}^\top=\alpha^\top A^{s-1}$, $\alpha_{s'-1}^\top=\alpha^\top A^{s'-1}$ and the transition matrix $A_w^{s'-s}=\sum_{x\in\Sigma_w^{s'-s}}A_x$ corresponding to the "event" $\Sigma_w^{s'-s}$, it comes

$$\mathbb{E}[b_{s,w}b_{s',w}]-f_s(w)f_{s'}(w) \;=\; \alpha_{s-1}^\top\Big(A_w^{s'-s}-A_w\beta\alpha_{s'-1}^\top A_w\Big)\beta\,.$$

We now discuss two cases. First the case when $s'-s\geqslant l$, then the case when $s'-s<l$.

Note that if $s'-s\geqslant|w|=l$, then $\Sigma_w^{s'-s}$ simplifies to $\Sigma_w^{s'-s}=w\Sigma^{s'-s-l}w$ and thus $A_w^{s'-s}=A_wA^{s'-s-l}A_w$. For such words, we thus obtain

$$\begin{aligned}\alpha_{s-1}^\top\Big(A_w^{s'-s}-A_w\beta\alpha_{s'-1}^\top A_w\Big)\beta &= \alpha_{s-1}^\top A_w\Big(A^{s'-s-l}-\beta\alpha_{s'-1}^\top\Big)A_w\beta\\ &\leqslant \|\alpha_{s-1}^\top A_w\|_1\|A^{s'-s-l}-\beta\alpha_{s'-1}^\top\|_\infty\|A_w\beta\|_\infty\,.\end{aligned}$$

Moreover, from Lemma 6, it holds $\|A^{s'-s-l}-\beta\alpha_{s'-1}^\top\|_\infty\leqslant 2\mu^{\mathbb{A}}_{s'-s-l}$. Also, it holds that $\|A_w\beta\|_\infty\leqslant 1$. Finally, since $\alpha_{s-1}^\top A_w$ is a sub-distribution over states, we have

$$\begin{aligned}\sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}\|\alpha_{s-1}^\top A_w\|_1 &= \sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}\alpha_{s-1}^\top A_w\beta\\ &= \sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}f_s(w)=\sum_{u\in\mathcal{U},v\in\mathcal{V}:uv\in W_l}f_s(uv)\;.\end{aligned}$$

Now, on the other hand if $s'-s<l$, using the fact that $\Sigma_w^{s'-s}\subset w\Sigma^{s'-s}$, then

$$\begin{aligned}\alpha_{s-1}^\top\Big(A_w^{s'-s}-A_w\beta\alpha_{s'-1}^\top A_w\Big)\beta &\leqslant \alpha_{s-1}^\top A_w\Big(A^{s'-s}-\beta\alpha_{s'-1}^\top A_w\Big)\beta\\ &= f_s(w)(1-f_{s'}(w))\leqslant f_s(w)\;.\end{aligned}$$

So in this case we again see that $\sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}f_s(w)=\sum_{u\in\mathcal{U},v\in\mathcal{V}:uv\in W_l}f_s(uv)$.

Step 2.5. Therefore, combining the above steps, so far we have seen that for a fixed $l\geqslant 0$, the sum $\sum_{w\in W_l}|w|_{\mathcal{U},\mathcal{V}}\sum_{1\leqslant s<s'\leqslant t}(\mathbb{E}[b_{s,w}b_{s',w}]-f_s(w)f_{s'}(w))$ is upper bounded by:

$$\begin{aligned}&\sum_{1\leqslant s<s'\leqslant t}\;\sum_{u\in\mathcal{U},v\in\mathcal{V}:|uv|=l}f_s(uv)(2\mu^{\mathbb{A}}_{s'-s-l}\mathbb{I}\{s'-s\geqslant l\}+\mathbb{I}\{s'-s<l\})\\ =\;&\sum_{u\in\mathcal{U},v\in\mathcal{V}:|uv|=l}\;\sum_{s=1}^{t-1}f_s(uv)\Big[\sum_{s'=s+1}^{t}2\mu^{\mathbb{A}}_{s'-s-l}\mathbb{I}\{s'-s\geqslant l\}+\mathbb{I}\{s'-s<l\}\Big]\;.\end{aligned}$$

Now note that $\sum_{s'=s+1}^{t} \mathbb{I}\{s'-s<l\} = \min\{l-1, t-s\} \leqslant l-1$. Furthermore, using that $\mu_t^{\mathbb{A}} \leqslant C\theta^t$ we get

$$\begin{aligned}\sum_{s'=s+1}^{t} \mu^{\mathbb{A}}_{s'-s-l}\mathbb{I}\{s'-s \geqslant l\} &= \mathbb{I}\{t \geqslant s+l\} \sum_{k=0}^{t-s-l} \mu_k^{\mathbb{A}} \\ &\leqslant C\mathbb{I}\{t \geqslant s+l\}\frac{1-\theta^{t-s-l+1}}{1-\theta} \leqslant \frac{C}{1-\theta}\ .\end{aligned}$$

In conclusion, we get

$$\begin{aligned}&\sum_{u\in\mathcal{U},v\in\mathcal{V}:|uv|=l}\sum_{s=1}^{t-1} f_s(uv)\Big[\sum_{s'=s+1}^{t} 2\mu^{\mathbb{A}}_{s'-s-l}\mathbb{I}\{s'-s\geqslant l\} + \mathbb{I}\{s'-s<l\}\Big] \\ \leqslant\ &\left(l-1+\frac{2C}{1-\theta}\right)\sum_{u\in\mathcal{U},v\in\mathcal{V}:|uv|=l}\sum_{s=1}^{t-1} f_s(uv) \\ \leqslant\ &t\left(l-1+\frac{2C}{1-\theta}\right)\sum_{u\in\mathcal{U},v\in\mathcal{V}:|uv|=l} \bar{f}_t(uv)\ .\end{aligned}$$

Finally, putting all the pieces together and introducing $L = \max_{w\in\mathcal{U}\cdot V}|w|$, we get from equations (6), (7), (8),

$$\begin{aligned}\mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2]^2 &\leqslant \frac{\sum_{u\in\mathcal{U},v\in\mathcal{V}}\bar{f}_t(uv)}{t} + \frac{2}{t}\sum_{l=0}^{\infty}\sum_{u\in\mathcal{U},v\in\mathcal{V}:|uv|=l}\bar{f}_t(uv)(l-1+\frac{2C}{1-\theta}) \\ &\leqslant \left[2L-1+\frac{4C}{1-\theta}\right]\frac{\sum_{u\in\mathcal{U},v\in\mathcal{V}}\bar{f}_t(uv)}{t}\ .\end{aligned}$$

Step 3. Application of Theorem 1. It remains to apply Theorem 1 with

$$\begin{aligned}\|g\|_{Lip} &\leqslant \frac{L}{t}\sqrt{\min\{|\mathcal{U}||\mathcal{V}|, 2n_{\mathcal{U}}n_{\mathcal{V}}\}}\,, \\ \mathbb{E}[\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2] &\leqslant \left(\sqrt{L}+\sqrt{\frac{2C}{1-\theta}}\right)\sqrt{\frac{2\sum_{u\in\mathcal{U},v\in\mathcal{V}}\bar{f}_t(uv)}{t}}\ ,\end{aligned}$$

for some constant C. After some rewriting, it comes

$$\begin{aligned}\mathbb{P}\Bigg(\|\widehat{H}_t^{\mathcal{U},\mathcal{V}} - \bar{H}_t^{\mathcal{U},\mathcal{V}}\|_2 > &\left(\sqrt{L}+\sqrt{\frac{2C}{1-\theta}}\right)\sqrt{\frac{2\sum_{u\in\mathcal{U},v\in\mathcal{V}}\bar{f}_t(uv)}{t}} \\ &+ \frac{LC}{(1-\theta)}\sqrt{\left(1+\frac{L-1}{t}\right)\frac{\min\{|\mathcal{U}||\mathcal{V}|, 2n_{\mathcal{U}}n_{\mathcal{V}}\}\ln(1/\delta)}{2t}}\Bigg) \leqslant \delta\,.\end{aligned}$$ □

E. Single-Trajectory Hankel Concentration Inequalities with Finite-State Control

Lemma 5 *The Hankel matrix $\widehat{H} = \widehat{H}_{t,\xi}^{\mathcal{U},\mathcal{V}}$ computed in Algorithm 3 satisfies $\mathbb{E}[\widehat{H}_{t,\xi}^{\mathcal{U},\mathcal{V}}] = \tilde{H}_t^{\mathcal{U},\mathcal{V}}$, where $\tilde{H}_t^{\mathcal{U},\mathcal{V}}$ is a block of the Hankel matrix corresponding to the stochastic WFA $\tilde{\mathbb{A}}_t = \langle\tilde{\alpha}_t, \beta, \{A_\sigma\}\rangle$ where we introduced the modified vector $\tilde{\alpha}_t = (1/t)\sum_{s=0}^{t-1}\alpha^\top(A/\kappa)^s$. We denote by $\tilde{f}_t$ the function computed by $\tilde{\mathbb{A}}_t$.*

Proof of Lemma 5:

For any $t \geqslant 0$ and $w \in \Sigma^\star$ let us define the function $\varphi_{s,w} : \Sigma^\omega \to \mathbb{R}$ given by

$$\varphi_{s,w}(x) = \frac{\mathbb{I}\{o_{s+1}a_{s+1}\cdots o_{s+|uv|}a_{s+|uv|} = w\}}{\kappa^s \pi(a_1\cdots a_{s+|w|}|o_1\cdots o_{s+|w|})} \ ,$$

where $x = (o_1, a_1)(o_2, a_2)\cdots$. Thus, the entries of the Hankel matrix computed in Algorithm 3 can be written as $\widehat{H}(u, v) = (1/t)\sum_{s=0}^{t-1} \varphi_{s,uv}(\xi)$. Now note that the expectation $\mathbb{E}[\varphi_{s,w}]$ with respect to a trajectory $\xi \sim \rho_{\mathbb{B}}$ can be written as

$$\sum_{w'\in\Sigma^s} \frac{\mathbb{P}[\xi \in w'w\Sigma^\omega]}{\kappa^s \pi(w'^{\mathcal{A}}w^{\mathcal{A}}|w'^{\mathcal{O}}w^{\mathcal{O}})} = \sum_{w'\in\Sigma^s} \frac{f_{\mathbb{B}}(w'w)}{\kappa^s f_{\mathbb{A}_\pi}(w'w)}$$
$$= \sum_{w'\in\Sigma^s} \frac{f_{\mathbb{A}}(w'w)}{\kappa^s} = \frac{\alpha^\top A^s A_w \beta}{\kappa^s} \ .$$

Therefore, the Hankel matrix $\widehat{H} = \widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi}$ computed in Algorithm 3 satisfies $\mathbb{E}[\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi}] = \tilde{H}^{\mathcal{U},\mathcal{V}}_t$, where $\tilde{H}^{\mathcal{U},\mathcal{V}}_t$ is a block of the Hankel matrix corresponding to the stochastic WFA $\tilde{\mathbb{A}}_t = \langle \tilde{\alpha}_t, \beta, \{A_\sigma\}\rangle$ with modified vector $\tilde{\alpha}_t = (1/t)\sum_{s=0}^{t-1} \alpha^\top (A/\kappa)^s$. We denote by $\tilde{f}_t$ the function computed by $\tilde{\mathbb{A}}_t$. □

Theorem 6 (Controlled case, single-trajectory, matrix-wise) *Let $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\}\rangle$ be a stochastic environment and π a stochastic policy induced by a probabilistic automaton $\mathbb{A}_\pi$, both over $\Sigma = \mathcal{A} \times \mathcal{O}$. Let $\mathbb{B} = \mathbb{A} \otimes \mathbb{A}_\pi$ be the stochastic WFA obtained by coupling the environment and the policy and $\rho_{\mathbb{B}} \in \mathcal{P}(\Sigma^\omega)$ the corresponding stochastic process. Suppose that $\mathbb{B}$ is (C, θ)-geometrically mixing. Suppose π satisfies the exploration Assumption 1 with parameter ε. Suppose the importance sampling constant κ in Algorithm 3 satisfies $\kappa\varepsilon > 1$. Let $\tilde{\mathbb{A}}_t = \langle \tilde{\alpha}_t, \beta, \{A_\sigma\}\rangle$ be the WFA defined in Section 5, where the initial vector is $\tilde{\alpha}_t = (1/t)\sum_{s=0}^{t-1} \alpha^\top (A/\kappa)^s$. Let $\bar{\mathbb{A}} = \mathbb{A} \otimes \mathbb{A}_{unif}$ be the stochastic WFA $\langle \alpha, \beta, A_\sigma/|\mathcal{A}|\rangle$ obtained by coupling the environment $\mathbb{A}$ with the uniform random policy. Suppose $\bar{\mathbb{A}}$ is $(\bar{C}, \bar{\theta})$-geometrically mixing. Let $L = \max_{w\in\mathcal{U}\cdot\mathcal{V}} |w|$, $\tilde{m} = \sum_{u\in\mathcal{U},v\in\mathcal{V}} \tilde{f}_t(uv)$, and $\bar{m} = \sum_{u\in\mathcal{U},v\in\mathcal{V}} \tilde{f}^{unif}_t(uv)$, where $\tilde{f}_t = f_{\tilde{\mathbb{A}}_t}$ and $\tilde{f}^{unif}_t$ is the function computed by the stochastic WFA obtained by Césaro averaging $\bar{\mathbb{A}}$ over t steps. Let $d = \sum_{w\in\mathcal{U}\cdot\mathcal{V}} |w|_{\mathcal{U},\mathcal{V}}$. Then for any $\delta \in (0, 1)$ we have*

$$\mathbb{P}\left(\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2 > \sqrt{\frac{\tilde{m}}{t\varepsilon^L(1-\kappa^{-2}\varepsilon^{-2})}} + \sqrt{\frac{2\bar{m}}{t\varepsilon^{2L}}\left(L + \frac{\bar{C}}{1-\bar{\theta}}\right)} + \frac{C}{\theta(1-\theta)\varepsilon^L}\sqrt{\frac{2d\ln(1/\delta)}{t}} \right) \leqslant \delta \ .$$

Proof of Theorem 6:

Let us introduce the function $g(\xi) = \|\widehat{H}^{\mathcal{U},\mathcal{V}}_t - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2$. We first control $\|g\|_{Lip}$ then $\mathbb{E}[g(\xi)]$, before applying Theorem 1.

Step 1: Control of $\|g\|_{Lip}$.

Let $\xi, \xi' \in \Sigma^\omega$ be trajectories $\xi = x_1x_2\cdots$ and $\xi' = x'_1x'_2\cdots$ differing by one element, say at

position ℓ. That is, $x_s = x'_s$ for all $s \neq \ell$. We note that

$$\begin{aligned}\left| \|\widehat{H}^{U,V}_{t,\xi} - \tilde{H}^{U,V}_t\|_2 - \|\widehat{H}^{U,V}_{t,\xi'} - \tilde{H}^{U,V}_t\|_2 \right| &\leqslant \|\widehat{H}^{U,V}_{t,\xi} - \widehat{H}^{U,V}_{t,\xi'}\|_2 \\ &\leqslant \sqrt{\sum_{u\in U}\sum_{v\in V}(\widehat{f}_{t,\xi}(uv) - \widehat{f}_{t,\xi'}(uv))^2} \\ &= \frac{1}{t}\sqrt{\sum_{u\in U}\sum_{v\in V}\left(\sum_{s=0}^{t-1}\varphi_{s,uv}(\xi) - \varphi_{s,uv}(\xi')\right)^2} \,.\end{aligned}$$

Next we take any $w \in \mathcal{U}\cdot\mathcal{V}$ and use $x_i = (o_i, a_i)$ to write

$$\begin{aligned}|\varphi_{s,w}(\xi) - \varphi_{s,w}(\xi')| &= \left| \frac{\mathbb{I}\{o_{s+1}a_{s+1}\cdots o_{s+|uv|}a_{s+|uv|} = w\}}{\kappa^s \pi(a_1\cdots a_{s+|w|}|o_1\cdots o_{s+|w|})} - \frac{\mathbb{I}\{o'_{s+1}a'_{s+1}\cdots o'_{s+|uv|}a'_{s+|uv|} = w\}}{\kappa^s \pi(a'_1\cdots a'_{s+|w|}|o'_1\cdots o'_{s+|w|})} \right| \\ &\leqslant \frac{1}{\kappa^s}\left(\frac{1}{\pi(a_1\cdots a_{s+|w|}|o_1\cdots o_{s+|w|})} + \frac{1}{\pi(a'_1\cdots a'_{s+|w|}|o'_1\cdots o'_{s+|w|})}\right) \\ &\leqslant \frac{2}{\kappa^s\varepsilon^{s+|w|}} \,,\end{aligned}$$

where we used the exploration assumption $\pi(u^{\mathcal{A}}|u^{\mathcal{O}}) \geqslant \varepsilon^{|u|}$ for all $u \in \Sigma^\star$.

From the expression above we see that for any $w \in \mathcal{U}\cdot\mathcal{V}$ we have

$$\sum_{s=0}^{t-1}\varphi_{s,w}(\xi) - \varphi_{s,w}(\xi') \leqslant \frac{2}{(1-1/(\kappa\varepsilon))\varepsilon^{|w|}} \,,$$

where we used that $\kappa\varepsilon > 1$. Thus, we can conclude that

$$\|g\|_{Lip} \leqslant \frac{2}{t(1-1/(\kappa\varepsilon))}\sqrt{\sum_{w\in\mathcal{U}\cdot\mathcal{V}}\frac{|w|_{\mathcal{U},\mathcal{V}}}{\varepsilon^{2|w|}}} \leqslant \frac{2}{t\varepsilon^L(1-1/(\kappa\varepsilon))}\sqrt{\sum_{w\in\mathcal{U}\cdot\mathcal{V}}|w|_{\mathcal{U},\mathcal{V}}} \,.$$

Note that $d = \sum_{w\in\mathcal{U}\cdot\mathcal{V}}|w|_{\mathcal{U},\mathcal{V}}$ is the quantity defined in the statement of Theorem 3.

Step 2: Control of $\mathbb{E}[g(\xi)]$. We now want to control the following quantity $\mathbb{E}[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2]$. We start in the same way as in the proof of Theorem 3.

Step 2.1. By Jensen's inequality, the norm of $\widehat{H}^{\mathcal{U},\mathcal{V}}_t - \tilde{H}^{\mathcal{U},\mathcal{V}}_t$ is controlled by its Frobenius norm

$$\begin{aligned}\mathbb{E}[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2]^2 &\leqslant \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}}\mathbb{E}\left[\left(\widehat{f}_{t,\xi}(uv) - \tilde{f}_t(uv)\right)^2\right] \\ &= \sum_{w\in\mathcal{U}\cdot\mathcal{V}}|w|_{\mathcal{U},\mathcal{V}}\mathbb{E}\left[\left(\widehat{f}_{t,\xi}(w) - \tilde{f}_t(w)\right)^2\right] \,.\end{aligned}$$

Recall that in Section 5 we showed that $\mathbb{E}[\widehat{f}_{t,\xi}(w)] = \tilde{f}_t(w)$ for any $w \in \Sigma^\star$. Hence the expression above is a sum of variances, each of which can be written as

$$\mathbb{E}\left[\left(\widehat{f}_{t,\xi}(w) - \tilde{f}_t(w)\right)^2\right] = \mathbb{E}\left[\widehat{f}_{t,\xi}(w)^2\right] - \tilde{f}_t(w)^2 \,. \tag{8}$$

Now we recall the definitions of the quantities appearing in this expression:

$$\begin{aligned}
\widehat{f}_{t,\xi}(w) &= \frac{1}{t}\sum_{s=0}^{t-1}\varphi_{s,w}(\xi) \\
&= \frac{1}{t}\sum_{s=0}^{t-1}\frac{\mathbb{I}\{o_{s+1}a_{s+1}\cdots o_{s+|w|}a_{s+|w|}=w\}}{\kappa^s\pi(a_1\cdots a_{s+|w|}|o_1\cdots o_{s+|w|})} \ , \\
\tilde{f}_t(w) &= \frac{1}{t}\sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)}{\kappa^s} \\
&= \frac{1}{t}\sum_{s=0}^{t-1}\alpha^\top\left(\frac{A}{\kappa}\right)^s A_w\beta \ .
\end{aligned}$$

Therefore, we can expand the squares in (8) as follows:

$$\begin{aligned}
\mathbb{E}\left[\widehat{f}_{t,\xi}(w)^2\right] &= \frac{1}{t^2}\left(\sum_{s=0}^{t-1}\mathbb{E}\left[\varphi_{s,w}(\xi)^2\right] + 2\sum_{0\leqslant s<s'\leqslant t-1}\mathbb{E}\left[\varphi_{s,w}(\xi)\varphi_{s',w}(\xi)\right]\right) \ , \\
\tilde{f}_t(w)^2 &= \frac{1}{t^2}\left(\sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)^2}{\kappa^{2s}} + 2\sum_{0\leqslant s<s'\leqslant t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)f_{\mathbb{A}}(\Sigma^{s'} w)}{\kappa^{s+s'}}\right) \ .
\end{aligned}$$

Using these expression we now bound the difference in (8) by considering the "squared" and the "cross" terms separately.

Step 2.2. We start with the "squared" terms and note that for any $0 \leqslant s \leqslant t-1$ and $w \in \mathcal{U}\cdot\mathcal{V}$ we have

$$\begin{aligned}
\mathbb{E}\left[\varphi_{s,w}(\xi)^2\right] &= \sum_{w'\in\Sigma^s}\frac{f_{\mathbb{B}}(w'w)}{\kappa^{2s}\pi(w'^{\mathcal{A}}w^{\mathcal{A}}|w'^{\mathcal{O}}w^{\mathcal{O}})^2} \\
&= \sum_{w'\in\Sigma^s}\frac{f_{\mathbb{A}}(w'w)}{\kappa^{2s}\pi(w'^{\mathcal{A}}w^{\mathcal{A}}|w'^{\mathcal{O}}w^{\mathcal{O}})} \\
&\leqslant \frac{f_{\mathbb{A}}(\Sigma^s w)}{\kappa^{2s}\varepsilon^{s+|w|}} \\
&= \frac{f_{\mathbb{A}}(\Sigma^s w)}{\kappa^s(\kappa\varepsilon)^s\varepsilon^{|w|}} \ .
\end{aligned}$$

Using Cauchy–Schwartz to sum these terms over t we obtain:

$$\begin{aligned}
\sum_{s=0}^{t-1}\mathbb{E}\left[\varphi_{s,w}(\xi)^2\right] &\leqslant \sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)}{\kappa^s(\kappa\varepsilon)^s\varepsilon^{|w|}} \\
&\leqslant \frac{1}{(1-1/(\kappa^2\varepsilon^2))\varepsilon^{|w|}}\left(\sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)^2}{\kappa^{2s}}\right)
\end{aligned}$$

Using this bound we can now see that the contribution of the "squared" terms to (8) is at most

$$\begin{aligned}
\frac{1}{t^2}\left(\sum_{s=0}^{t-1}\mathbb{E}\left[\varphi_{s,w}(\xi)^2\right] - \sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)^2}{\kappa^{2s}}\right) &\leqslant \frac{1}{t^2}\left(\frac{1}{(1-1/(\kappa^2\varepsilon^2))\varepsilon^{|w|}} - 1\right)\left(\sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)^2}{\kappa^{2s}}\right) \\
&\leqslant \frac{1}{t^2(1-1/(\kappa^2\varepsilon^2))\varepsilon^{|w|}}\left(\sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)^2}{\kappa^{2s}}\right) \ .
\end{aligned}$$

This expression can be further simplified by noting that $\varepsilon \leqslant 1/|\mathcal{A}|$ implies $\kappa > |\mathcal{A}|$ and therefore $f_{\mathbb{A}}(\Sigma^s w)/\kappa^s \leqslant f_{\mathbb{A}}(\Sigma^s w)/|\mathcal{A}|^s \leqslant 1$ since this corresponds to the probability of observing $w^{\mathcal{O}}$ when taking the actions in $w^{\mathcal{A}}$ after the first s actions have been chosen by a uniform random policy. Thus, we get

$$\frac{1}{t^2}\left(\sum_{s=0}^{t-1}\mathbb{E}\left[\varphi_{s,w}(\xi)^2\right] - \sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)^2}{\kappa^{2s}}\right) \leqslant \frac{1}{t^2(1-1/(\kappa^2\varepsilon^2))\varepsilon^{|w|}}\left(\sum_{s=0}^{t-1}\frac{f_{\mathbb{A}}(\Sigma^s w)}{\kappa^s}\right)$$
$$= \frac{\tilde{f}_t(w)}{t(1-1/(\kappa^2\varepsilon^2))\varepsilon^{|w|}} \ .$$

To complete this step we sum this bound for all $w \in \mathcal{U}\cdot\mathcal{V}$ to control the contribution of the "squared" terms in (8):

$$\sum_{w\in\mathcal{U}\cdot\mathcal{V}} |w|_{\mathcal{U},\mathcal{V}}\frac{\tilde{f}_t(w)}{t(1-1/(\kappa^2\varepsilon^2))\varepsilon^{|w|}} \leqslant \frac{1}{t(1-1/(\kappa^2\varepsilon^2))\varepsilon^{L}}\sum_{w\in\mathcal{U}\cdot\mathcal{V}} |w|_{\mathcal{U},\mathcal{V}}\tilde{f}_t(w)$$
$$= \frac{1}{t(1-1/(\kappa^2\varepsilon^2))\varepsilon^{L}}\sum_{u\in\mathcal{U},v\in\mathcal{V}}\tilde{f}_t(uv) \ ,$$

where $L = \max_{w\in\mathcal{U}\cdot\mathcal{V}}|w|$.

Step 2.3. We now focus on controlling the "cross" terms in (8) of the form

$$\mathbb{E}\left[\varphi_{s,w}(\xi)\varphi_{s',w}(\xi)\right] - \frac{f_{\mathbb{A}}(\Sigma^s w)f_{\mathbb{A}}(\Sigma^{s'} w)}{\kappa^{s+s'}} \ . \tag{9}$$

Using the same notation $\Sigma_w^{s'-s} = w\Sigma^{s'-s} \cap \Sigma^{s'-s}w$ as in the proof of Theorem 3, we first note that

$$\begin{aligned}
\mathbb{E}\left[\varphi_{s,w}(\xi)\varphi_{s',w}(\xi)\right] &= \sum_{x\in\Sigma^s\Sigma_w^{s'-s}} \frac{f_{\mathbb{B}}(x)}{\kappa^{s+s'}\pi(x^{\mathcal{A}}_{1:s+|w|}|x^{\mathcal{O}}_{1:s+|w|})\pi(x^{\mathcal{A}}_{1:s'+|w|}|x^{\mathcal{O}}_{1:s'+|w|})} \\
&= \sum_{x\in\Sigma^s\Sigma_w^{s'-s}} \frac{f_{\mathbb{A}}(x)}{\kappa^{s+s'}\pi(x^{\mathcal{A}}_{1:s+|w|}|x^{\mathcal{O}}_{1:s+|w|})} \\
&\leqslant \sum_{x\in\Sigma^s\Sigma_w^{s'-s}} \frac{f_{\mathbb{A}}(x)}{\kappa^{s+s'}\varepsilon^{s+|w|}} \\
&= \frac{f_{\mathbb{A}}(\Sigma^s\Sigma_w^{s'-s})}{\kappa^{s+s'}\varepsilon^{s+|w|}} \\
&= \frac{\alpha^\top A^s A_w^{s'-s}\beta}{\kappa^{s+s'}\varepsilon^{s+|w|}} \ ,
\end{aligned}$$

where we used the notation $A_w^{s'-s} = \sum_{x\in\Sigma_w^{s'-s}} A_x$. We also define $\tilde{A} = A/\kappa$ and $\alpha_s^\top = \alpha^\top\tilde{A}^s$. Then we can write (9) as

$$\frac{\alpha^\top A^s A_w^{s'-s}\beta}{\kappa^{s+s'}\varepsilon^{s+|w|}} - \frac{(\alpha^\top A^s A_w\beta)(\alpha^\top A^{s'} A_w\beta)}{\kappa^{s+s'}} = \alpha_s^\top\left(\frac{A_w^{s'-s}}{\kappa^{s'}\varepsilon^{s+|w|}} - A_w\beta\alpha_{s'}^\top A_w\right)\beta \ . \tag{10}$$

To bound this quantity we proceed by considering two cases.

Step 2.4. First suppose that $s' - s \geqslant l = |w|$. In this case we have $A_w^{s'-s} = A_w A^{s'-s-l}A_w$ and (10) equals to

$$\alpha_s^\top A_w\left(\frac{A^{s'-s-l}}{\kappa^{s'}\varepsilon^{s+l}} - \beta\alpha_{s'}^\top\right)A_w\beta = \alpha_s^\top A_w\left(\frac{\tilde{A}^{s'-s-l}}{\kappa^{s+l}\varepsilon^{s+l}} - \beta\alpha_{s+l}^\top\tilde{A}^{s'-s-l}\right)A_w\beta \ .$$

Now we apply the same argument we used to bound the "cross" terms in the case of stochastic WFA using cone norms. In particular, we consider the stochastic WFA $\bar{\mathbb{A}} = \langle \alpha, \beta, \bar{A}_\sigma \rangle$, where $\bar{A}_\sigma = A_\sigma/|\mathcal{A}|$. Note this is the stochastic WFA obtained by coupling environment $\mathbb{A}$ with the random policy that at each step chooses each action independently with probability $1/|\mathcal{A}|$. Now we let $\|\cdot\|_\beta$ and $\|\cdot\|_{\beta,\star}$ denote the cone norms corresponding to $\bar{\mathbb{A}}$. Using Lemma 3 we see that the following hold for all $w \in \Sigma^\star$:

$$\|A_w\beta\|_\beta = |\mathcal{A}|^l \|\bar{A}_w\beta\|_\beta \leqslant |\mathcal{A}|^l$$
$$\|\alpha_s^\top A_w\|_{\beta,\star} = \frac{|\mathcal{A}|^{s+l}}{\kappa^s}\|\alpha^\top \bar{A}^s \bar{A}_w\|_{\beta,\star} = \frac{|\mathcal{A}|^{s+l}}{\kappa^s}\alpha^\top \bar{A}^s \bar{A}_w\beta \ ,$$

where we used the notation $\bar{A} = A/|\mathcal{A}|$. We also note that for any vector satisfying $\|u\|_{\beta,\star} \leqslant 1$ we have

$$\|u^\top \beta \alpha_s^\top\|_{\beta,\star} \leqslant \|\alpha_s^\top\|_{\beta,\star} = \frac{|\mathcal{A}|^s}{\kappa^s}\|\alpha^\top \bar{A}^s\|_{\beta,\star} \leqslant \frac{|\mathcal{A}|^s}{\kappa^s} \leqslant \frac{1}{\kappa^s \varepsilon^s} \ .$$

This last bound can now be combined with the argument used in the case of stochastic WFA to show that

$$\begin{aligned}
\left\| \frac{\tilde{A}^{s'-s-l}}{\kappa^{s+l}\varepsilon^{s+l}} - \beta\alpha_{s+l}^\top \tilde{A}^{s'-s-l} \right\|_\beta &= \sup_{\|u\|_{\beta,\star}\leqslant 1} \left\| u^\top \frac{\tilde{A}^{s'-s-l}}{\kappa^{s+l}\varepsilon^{s+l}} - u^\top \beta\alpha_{s+l}^\top \tilde{A}^{s'-s-l} \right\|_{\beta,\star} \\
&\leqslant \sup_{\|u_1\|_{\beta,\star}\leqslant 1} \sup_{\|u_2\|_{\beta,\star}\leqslant 1} \left\| u_1^\top \frac{\tilde{A}^{s'-s-l}}{\kappa^{s+l}\varepsilon^{s+l}} - u_2^\top \frac{\tilde{A}^{s'-s-l}}{\kappa^{s+l}\varepsilon^{s+l}} \right\|_{\beta,\star} \\
&= \frac{|\mathcal{A}|^{s'-s+l}}{\kappa^{s'}\varepsilon^{s+l}} \sup_{\|u_1\|_{\beta,\star}\leqslant 1} \sup_{\|u_2\|_{\beta,\star}\leqslant 1} \left\| u_1^\top \bar{A}^{s'-s-l} - u_2^\top \bar{A}^{s'-s-l} \right\|_{\beta,\star} \\
&\leqslant \frac{|\mathcal{A}|^{s'-s+l}}{\kappa^{s'}\varepsilon^{s+l}} \mu_{s'-s-l}^{\bar{\mathbb{A}}} \ ,
\end{aligned}$$

where we used the definition of the mixing coefficient $\mu_{s'-s-l}^{\bar{\mathbb{A}}}$ for stochastic WFA $\bar{\mathbb{A}}$.

We now observe that $|\mathcal{A}| \leqslant 1/\varepsilon < \kappa$ implies $|\mathcal{A}|^{s'+l}/\kappa^{s+s'}\varepsilon^{s+l} \leqslant 1/\kappa^s\varepsilon^{s+2l}$. Finally, by plugging all these bounds together on an application of Hölder's inequality yields:

$$\left| \alpha_s^\top A_w \left(\frac{A^{s'-s-l}}{\kappa^{s'}\varepsilon^{s+l}} - \beta\alpha_{s'}^\top \right) A_w\beta \right| \leqslant \frac{\mu_{s'-s-l}^{\bar{\mathbb{A}}}}{\kappa^s\varepsilon^{s+2l}} \alpha^\top \bar{A}^s \bar{A}_w\beta \ .$$

Step 2.5. Now we consider the case $s' - s < l = |w|$. Using the fact that this implies $\Sigma_w^{s'-s} \subset w\Sigma^{s'-s}$, then

$$\alpha_s^\top A_w^{s'-s}\beta \leqslant \alpha_s^\top A_w A^{s'-s}\beta = |\mathcal{A}|^{s'-s}\alpha_s^\top A_w \bar{A}^{s'-s}\beta = |\mathcal{A}|^{s'-s}\alpha_s^\top A_w\beta \ ,$$

where we used $\bar{A}\beta = \beta$. Therefore, we can bound the expression in (10) as

$$\begin{aligned}
\alpha_s^\top \left(\frac{A_w^{s'-s}}{\kappa^{s'}\varepsilon^{s+l}} - A_w\beta\alpha_{s'}^\top A_w \right) \beta &\leqslant \alpha_s^\top A_w\beta \left(\frac{|\mathcal{A}|^{s'-s}}{\kappa^{s'}\varepsilon^{s+l}} - \alpha_{s'}^\top A_w\beta \right) \leqslant \frac{|\mathcal{A}|^{s'-s}}{\kappa^{s'}\varepsilon^{s+l}} \alpha_s^\top A_w\beta \\
&= \frac{|\mathcal{A}|^{s'+l}}{\kappa^{s'+s}\varepsilon^{s+l}} \alpha^\top \bar{A}^s \bar{A}_w\beta \leqslant \frac{1}{\kappa^s\varepsilon^{s+2l}} \alpha^\top \bar{A}^s \bar{A}_w\beta \ .
\end{aligned}$$

Step 2.6. Finally, we can combine the bounds above by summing over all $w \in \mathcal{U} \cdot \mathcal{V}$ and all $0 \leqslant s < s' \leqslant t-1$ in the same way we did for PFA. We first note that from Steps 2.4 and 2.5 we obtain the

following bound for (10):

$$\alpha_s^\top \left(\frac{A_w^{s'-s}}{\kappa^{s'}\varepsilon^{s+|w|}} - A_w \beta \alpha_{s'}^\top A_w \right) \beta \leqslant \frac{\bar{f}_s(w)}{\kappa^s \varepsilon^{s+2|w|}} \left(\mu^{\bar{\mathbb{A}}}_{s'-s-|w|} \mathbb{I}\{s'-s \geqslant |w|\} + \mathbb{I}\{s'-s < |w|\} \right) \ .$$

Now let $l = |w|$ and note that $\mu^{\bar{\mathbb{A}}}_{s'-s-l} \leqslant \bar{C}\bar{\theta}^{s'-s-l}$, where $\bar{C}$ and $\bar{\theta}$ are the geometric mixing constants for stochastic WFA $\bar{\mathbb{A}}$. Thus, summing first over s' we get

$$\sum_{s'=s+1}^{t-1} \mu^{\bar{\mathbb{A}}}_{s'-s-|w|} \mathbb{I}\{s'-s \geqslant |w|\} + \mathbb{I}\{s'-s < |w|\} \leqslant l + \frac{\bar{C}}{1-\bar{\theta}} \ .$$

Therefore, writing $\mathcal{W}_l$ for all words of length l in $\mathcal{W} = \mathcal{U} \cdot \mathcal{V}$ we get:

$$\begin{aligned}
&\frac{2}{t^2} \sum_{w \in \mathcal{U}\cdot\mathcal{V}} |w|_{\mathcal{U},\mathcal{V}} \sum_{0 \leqslant s < s' \leqslant t-1} \left(\mathbb{E}\left[\varphi_{s,w}(\xi)\varphi_{s',w}(\xi)\right] - \frac{f_{\mathbb{A}}(\Sigma^s w) f_{\mathbb{A}}(\Sigma^{s'} w)}{\kappa^{s+s'}} \right) \\
&\leqslant \frac{2}{t^2} \sum_{l=0}^{\infty} \sum_{w \in \mathcal{W}_l} \frac{|w|_{\mathcal{U},\mathcal{V}}}{\varepsilon^{2l}} \left(l + \frac{\bar{C}}{1-\bar{\theta}} \right) \sum_{s=0}^{t-2} \frac{\bar{f}_s(w)}{\kappa^s \varepsilon^s} \\
&\leqslant \frac{2}{t} \sum_{l=0}^{\infty} \frac{1}{\varepsilon^{2l}} \left(l + \frac{\bar{C}}{1-\bar{\theta}} \right) \sum_{w \in \mathcal{W}_l} |w|_{\mathcal{U},\mathcal{V}} \sum_{s=0}^{t-1} \frac{\bar{f}_s(w)}{t} \\
&\leqslant \frac{2}{t\varepsilon^{2L}} \left(L + \frac{\bar{C}}{1-\bar{\theta}} \right) \sum_{u \in \mathcal{U}, v \in \mathcal{V}} \tilde{f}_t^{unif}(uv) \ ,
\end{aligned}$$

where we used that $\kappa\varepsilon > 1$ and $\tilde{f}_t^{unif}(w) = (1/t)\sum_{s=0}^{t-1} \bar{f}_s(w)$.

Step 2.7. Our final bound for $\mathbb{E}[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2]$ is now obtained by combining the results from Step 2.2 and 2.6:

$$\mathbb{E}[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2]^2 \leqslant \frac{1}{t\varepsilon^L(1-1/(\kappa\varepsilon)^2)} \sum_{u \in \mathcal{U}, v \in \mathcal{V}} \tilde{f}_t(uv) + \frac{2}{t\varepsilon^{2L}} \left(L + \frac{\bar{C}}{1-\bar{\theta}} \right) \sum_{u \in \mathcal{U}, v \in \mathcal{V}} \tilde{f}_t^{unif}(uv) \ .$$

Note that $\tilde{m} = \sum_{u \in \mathcal{U}, v \in \mathcal{V}} \tilde{f}_t(uv)$ and $\bar{m} = \sum_{u \in \mathcal{U}, v \in \mathcal{V}} \tilde{f}_t^{unif}(uv)$ are the quantities defined in the statement of Theorem 6.

Step 3. Application of Theorem 1 It follows directly from Theorem 1 that with probability at least $1-\delta$ we have

$$\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2 \leqslant \mathbb{E}[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2] + \eta_{\rho_{\mathbb{B}}} \|g\|_{Lip} \sqrt{\frac{t\ln(1/\delta)}{2}} \ .$$

Using that $\rho_{\mathbb{B}}$ is (C,θ)-geometrically mixing and Lemma 4 we can bound the η-mixing coefficient as $\eta_{\rho_{\mathbb{B}}} \leqslant C/(\theta(1-\theta))$. Thus, by plugging our estimates for $\|g\|_{Lip}$ and $\mathbb{E}[\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2]$ we obtain that with probability at least $1-\delta$:

$$\|\widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \tilde{H}^{\mathcal{U},\mathcal{V}}_t\|_2 \leqslant \sqrt{\frac{\tilde{m}}{t\varepsilon^L(1-\kappa^{-2}\varepsilon^{-2})}} + \sqrt{\frac{2\bar{m}}{t\varepsilon^{2L}} \left(L + \frac{\bar{C}}{1-\bar{\theta}} \right)} + \frac{C}{\theta(1-\theta)\varepsilon^L} \sqrt{\frac{2d\ln(1/\delta)}{t}} \ .$$

□
