JMLR: Workshop and Conference Proceedings vol 49:1–70, 2016

Reinforcement Learning of POMDPs using Spectral Methods

Kamyar Azizzadenesheli * KAZIZZAD@UCI.EDU
University of California, Irvine

Alessandro Lazaric † ALESSANDRO.LAZARIC@INRIA.FR
Institut National de Recherche en Informatique et en Automatique, (Inria)

Animashree Anandkumar‡ A.ANANDKUMAR@UCI.EDU
University of California, Irvine

Abstract

We propose a new reinforcement learning algorithm for partially observable Markov decision processes (POMDP) based on spectral decomposition methods. While spectral methods have been previously employed for consistent learning of (passive) latent variable models such as hidden Markov models, POMDPs are more challenging since the learner interacts with the environment and possibly changes the future observations in the process. We devise a learning algorithm running through episodes, in each episode we employ spectral techniques to learn the POMDP parameters from a trajectory generated by a fixed policy. At the end of the episode, an optimization oracle returns the optimal memoryless planning policy which maximizes the expected reward based on the estimated POMDP model. We prove an order-optimal regret bound with respect to the optimal memoryless policy and efficient scaling with respect to the dimensionality of observation and action spaces.

Keywords: Spectral Methods, Method of Moments, Partially Observable Markov Decision Process, Latent Variable Model, Upper Confidence Reinforcement Learning.

1. Introduction

Reinforcement Learning (RL) is an effective approach to solve the problem of sequential decision–making under uncertainty. RL agents learn how to maximize long-term reward using the experience obtained by direct interaction with a stochastic environment (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). Since the environment is initially unknown, the agent has to balance between *exploring* the environment to estimate its structure, and *exploiting* the estimates to compute a policy that maximizes the long-term reward. As a result, designing a RL algorithm requires three different elements: **1)** an estimator for the environment's structure, **2)** a planning algorithm to compute the optimal policy of the estimated envi-

* K. Azizzadenesheli is supported in part by NSF Career award CCF-1254106 and ONR Award N00014-14-1-0665

† A. Lazaric is supported in part by a grant from CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, CRIStAL (Centre de Recherche en Informatique et Automatique de Lille), and the French National Research Agency (ANR) under project ExTra-Learn n.ANR-14-CE24-0010-01.

‡ A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, ONR Award N00014-14-1-0665, ARO YIP Award W911NF-13-1-0084 and AFOSR YIP FA9550-15-1-0221

ronment (LaValle, 2006), and **3)** a strategy to make a trade off between exploration and exploitation to minimize the *regret*, i.e., the difference between the performance of the exact optimal policy and the rewards accumulated by the agent over time.

Most of RL literature assumes that the environment can be modeled as a Markov decision process (MDP), with a Markovian state evolution that is fully observed. A number of exploration–exploitation strategies have been shown to have strong performance guarantees for MDPs, either in terms of regret or sample complexity (see Sect. 1.2 for a review). However, the assumption of full observability of the state evolution is often violated in practice, and the agent may only have noisy observations of the true state of the environment (e.g., noisy sensors in robotics). In this case, it is more appropriate to use the partially-observable MDP or POMDP (Sondik, 1971) model.

Many challenges arise in designing RL algorithms for POMDPs. Unlike in MDPs, the estimation problem (element 1) involves identifying the parameters of a latent variable model (LVM). In an MDP the agent directly observes (stochastic) state transitions, and the estimation of the generative model is straightforward via empirical estimators. On the other hand, in a POMDP the transition and reward models must be inferred from noisy observations and the Markovian state evolution is hidden. The planning problem (element 2), i.e., computing the optimal policy for a POMDP with known parameters, is PSPACE-complete (Papadimitriou and Tsitsiklis, 1987), and it requires solving an augmented MDP built on a continuous belief space (i.e., a distribution over the hidden state of the POMDP). Finally, integrating estimation and planning in an exploration–exploitation strategy (element 3) with guarantees is non-trivial and no no-regret strategies are currently known (see Sect. 1.2).

1.1. Summary of Results

The main contributions of this paper are as follows: (i) We propose a new RL algorithm for POMDPs that incorporates spectral parameter estimation within a exploration-exploitation framework, (ii) we analyze regret bounds assuming access to an optimization oracle that provides the best memoryless planning policy at the end of each learning episode, (iii) we prove order optimal regret and efficient scaling with dimensions, thereby providing the first guaranteed RL algorithm for a wide class of POMDPs.

The estimation of the POMDP is carried out via spectral methods which involve decomposition of certain moment tensors computed from data. This learning algorithm is interleaved with the optimization of the planning policy using an exploration–exploitation strategy inspired by the UCRL method for MDPs (Ortner and Auer, 2007; Jaksch et al., 2010). The resulting algorithm, called SM-UCRL (*Spectral Method for Upper-Confidence Reinforcement Learning*), runs through episodes of variable length, where the agent follows a fixed policy until enough data are collected and then it updates the current policy according to the estimates of the POMDP parameters and their accuracy. Throughout the paper we focus on the estimation and exploration–exploitation aspects of the algorithm, while we assume access to a *planning oracle* for the class of memoryless policies (i.e., policies directly mapping observations to a distribution over actions).[1]

1. This assumption is common in many works in bandit and RL literature (see e.g., Abbasi-Yadkori and Szepesvári (2011) for linear bandit and Chen et al. (2013) in combinato-

Theoretical Results. We prove the following learning result. For the full details see Thm. 3 in Sect. 3.

Theorem ***(Informal Result on Learning POMDP Parameters)*** *Let M be a POMDP with X states, Y observations, A actions, R rewards, and $Y > X$, and characterized by densities $f_T(x'|x,a)$, $f_O(y|x)$, and $f_R(r|x,a)$ defining state transition, observation, and the reward models. Given a sequence of observations, actions, and rewards generated by executing a memoryless policy where each action a is chosen $N(a)$ times, there exists a spectral method which returns estimates $\widehat{f}_T$, $\widehat{f}_O$, and $\widehat{f}_R$ that, under suitable assumptions on the POMDP, the policy, and the number of samples, satisfy*

$$\|\widehat{f}_O(\cdot|x) - f_O(\cdot|x)\|_1 \le \widetilde{O}\Big(\sqrt{\frac{YR}{N(a)}}\Big),$$

$$\|\widehat{f}_R(\cdot|x,a) - f_R(\cdot|x,a)\|_1 \le \widetilde{O}\Big(\sqrt{\frac{YR}{N(a)}}\Big),$$

$$\|\widehat{f}_T(\cdot|x,a) - f_T(\cdot|x,a)\|_2 \le \widetilde{O}\Big(\sqrt{\frac{YRX^2}{N(a)}}\Big),$$

with high probability, for any state x and any action a.

This result shows the consistency of the estimated POMDP parameters and it also provides explicit confidence intervals.

By employing the above learning result in a UCRL framework, we prove the following bound on the regret Reg_N w.r.t. the optimal memoryless policy. For full details see Thm. 4 in Sect. 4.

Theorem ***(Informal Result on Regret Bounds)*** *Let M be a POMDP with X states, Y observations, A actions, and R rewards, with a diameter D defined as*

$$D := \max_{x,x' \in \mathcal{X}, a,a' \in \mathcal{A}} \min_{\pi} \mathbb{E}\big[\tau(x', a'|x, a; \pi)\big],$$

i.e., the largest mean passage time between any two state-action pairs in the POMDP using a memoryless policy π mapping observations to actions. If SM-UCRL *is run over N steps using the confidence intervals of Thm. 3, under suitable assumptions on the POMDP, the space of policies, and the number of samples, we have*

$$Reg_N \le \widetilde{O}\Big(DX^{3/2}\sqrt{AYRN}\Big),$$

with high probability.

The above result shows that despite the complexity of estimating the POMDP parameters from noisy observations of hidden states, the regret of SM-UCRL is similar to the case of

rial bandit), where the focus is on the exploration–exploitation strategy rather than the optimization problem.

MDPs, where the regret of UCRL scales as $\widetilde{O}(D_{\text{MDP}} X\sqrt{AN})$. The regret is order-optimal, since $\widetilde{O}(\sqrt{N})$ matches the lower bound for MDPs.

Another interesting aspect is that the diameter of the POMDP is a natural extension of the MDP case. While D_{MDP} measures the mean passage time using state–based policies (i.e., a policies mapping *states* to actions), in POMDPs policies cannot be defined over states but rather on observations and this naturally translates into the definition of the diameter D. More details on other problem-dependent terms in the bound are discussed in Sect. 4.

The derived regret bound is with respect to the best memoryless (stochastic) policy for the given POMDP. Indeed, for a general POMDP, the optimal policy need not be memoryless. However, finding the optimal policy is uncomputable for infinite horizon regret minimization (Madani, 1998). Instead memoryless policies have shown good performance in practice (see the Section on related work). Moreover, for the class of so-called *contextual MDP*, a special class of POMDPs, the optimal policy is also memoryless (Krishnamurthy et al., 2016).

Analysis of the learning algorithm. The learning results in Thm. 3 are based on spectral tensor decomposition methods, which have been previously used for consistent estimation of a wide class of LVMs (Anandkumar et al., 2014). This is in contrast with traditional learning methods, such as expectation-maximization (EM) (Dempster et al., 1977), that have no consistency guarantees and may converge to local optimum which is arbitrarily bad.

While spectral methods have been previously employed in sequence modeling such as in HMMs (Anandkumar et al., 2014), by representing it as multiview model, their application to POMDPs is not trivial. In fact, unlike the HMM, the consecutive observations of a POMDP are no longer conditionally independent, when conditioned on the hidden state of middle *view*. This is because the decision (or the action) depends on the observations themselves. By limiting to memoryless policies, we can control the range of this dependence, and by conditioning on the actions, we show that we can obtain conditionally independent *views*. As a result, starting with samples collected along a trajectory generated by a fixed policy, we can construct a multi-view model and use the tensor decomposition method on each action separately, estimate the parameters of the POMDP, and define confidence intervals.

While the proof follows similar steps as in previous works on spectral methods (e.g., HMMs Anandkumar et al., 2014), here we extend concentration inequalities for dependent random variables to matrix valued functions by combining the results of Kontorovich et al. (2008) with the matrix Azuma's inequality of Tropp (2012). This allows us to remove the usual assumption that the samples are generated from the stationary distribution of the current policy. This is particularly important in our case since the policy changes at each episode and we can avoid discarding the initial samples and waiting until the corresponding Markov chain converged (i.e., the *burn-in* phase).

The condition that the POMDP has more observations than states ($Y > X$) follows from standard non-degeneracy conditions to apply the spectral method. This corresponds to considering POMDPs where the underlying MDP is defined over a few number of states (i.e., a low-dimensional space) that can produce a large number of noisy observations.

This is common in applications such as spoken-dialogue systems (Atrash and Pineau, 2006; Png et al., 2012) and medical applications (Hauskrecht and Fraser, 2000). We also show how this assumption can be relaxed and the result can be applied to a wider family of POMDPs.

Analysis of the exploration–exploitation strategy. SM-UCRL applies the popular *optimism-in-face-of-uncertainty* principle[2] to the confidence intervals of the estimated POMDP and compute the optimal policy of the most optimistic POMDP in the admissible set. This *optimistic* choice provides a smooth combination of the exploration encouraged by the confidence intervals (larger confidence intervals favor uniform exploration) and the exploitation of the estimates of the POMDP parameters.

While the algorithmic integration is rather simple, its analysis is not trivial. The spectral method cannot use samples generated from different policies and the length of each episode should be carefully tuned to guarantee that estimators improve at each episode. Furthermore, the analysis requires redefining the notion of diameter of the POMDP. In addition, we carefully bound the various perturbation terms in order to obtain efficient scaling in terms of dimensionality factors.

Finally, in the Appendix F, we report preliminary synthetic experiments that demonstrate superiority of our method over existing RL methods such as Q-learning and UCRL for MDPs, and also over purely exploratory methods such as random sampling, which randomly chooses actions independent of the observations. SM-UCRL converges much faster and to a better solution. The solutions relying on the MDP assumption, directly work in the (high) dimensional observation space and perform poorly. In fact, they can even be worse than the random sampling policy baseline. In contrast, our method aims to find the lower dimensional latent space to derive the policy and this allows UCRL to find a much better memoryless policy with vanishing regret.

It is worth noting that, in general, with slight changes on the learning set up, one can come up with new algorithms to learn different POMDP models with, slightly, same upper confidence bounds. Moreover, after applying memoryless policy and collecting sufficient number of samples, when the model parameters are learned very well, one can do the planing on the belief space, and get memory dependent policy, therefore improve the performance even further.

1.2. Related Work

In last few decades, MDP has been widely studied (Kearns and Singh, 2002; Brafman and Tennenholtz, 2003; Bartlett and Tewari, 2009; Jaksch et al., 2010) in different setting. Even for the large state space MDP, where the classical approaches are not scalable, Kocsis and Szepesvári (2006) introduces MDP Monte-Carlo planning tree which is one of the few viable approaches to find the near-optimal policy. In addition, for special class of MDPs, Markov Jump Affine Model, when the action space is continuous, (Baltaoglu et al., 2016) proposes an order optimal learning policy.

2. This principle has been successfully used in a wide number of exploration–exploitation problems ranging from multi-armed bandit (Auer et al., 2002), linear contextual bandit (Abbasi-Yadkori et al., 2011), linear quadratic control (Abbasi-Yadkori and Szepesvári, 2011), and reinforcement learning (Ortner and Auer, 2007; Jaksch et al., 2010).

While RL in MDPs has been widely studied, the design of effective exploration–exploration strategies in POMDPs is still relatively unexplored. Ross et al. (2007) and Poupart and Vlassis (2008) propose to integrate the problem of estimating the belief state into a model-based Bayesian RL approach, where a distribution over possible MDPs is updated over time. The proposed algorithms are such that the Bayesian inference can be done accurately and at each step, a POMDP is sampled from the posterior and the corresponding optimal policy is executed. While the resulting methods implicitly balance exploration and exploitation, no theoretical guarantee is provided about their regret and their algorithmic complexity requires the introduction of approximation schemes for both the inference and the planning steps. An alternative to model-based approaches is to adapt model-free algorithms, such as Q-learning, to the case of POMDPs. Perkins (2002) proposes a Monte-Carlo approach to action-value estimation and it shows convergence to locally optimal memoryless policies. While this algorithm has the advantage of being computationally efficient, local optimal policies may be arbitrarily suboptimal and thus suffer a linear regret.

An alternative approach to solve POMDPs is to use policy search methods, which avoid estimating value functions and directly optimize the performance by searching in a given policy space, which usually contains memoryless policies (see e.g., (Ng and Jordan, 2000),(Baxter and Bartlett, 2001),(Poupart and Boutilier, 2003; Bagnell et al., 2004)). Beside its practical success in offline problems, policy search has been successfully integrated with efficient exploration–exploitation techniques and shown to achieve small regret (Gheshlaghi-Azar et al., 2013, 2014). Nonetheless, the performance of such methods is severely constrained by the choice of the policy space, which may not contain policies with good performance. Another approach to solve POMDPs is proposed by (Guo et al., 2016). In this work, the agent randomly chooses actions independent of the observations and rewards. The agent executes random policy until it collects sufficient number of samples and then estimates the model parameters given collected information. The authors propose Probably Approximately Correct (PAC) framework for RL in POMDP setting and shows polynomial sample complexity for learning of the model parameters. During learning phase, they defines the induced Hidden Markov Model and applies random policy to capture different aspects of the model, then in the planing phase, given the estimated model parameters, they compute the optimum policy so far. In other words, the proposed algorithm explores the environment sufficiently enough and then exploits this exploration to come up with a optimal policy given estimated model. In contrast, our method considers RL of POMDPs in an episodic learning framework.

Matrix decomposition methods have been previously used in the more general setting of predictive state representation (PSRs) (Boots et al., 2011) to reconstruct the structure of the dynamical system. Despite the generality of PSRs, the proposed model relies on strong assumptions on the dynamics of the system and it does not have any theoretical guarantee about its performance. Gheshlaghi azar et al. (2013) used spectral tensor decomposition methods in the multi-armed bandit framework to identify the hidden generative model of a sequence of bandit problems and showed that this may drastically reduce the regret. Recently, (Hamilton et al., 2014) introduced compressed PSR (CPSR) method to reduce the computation cost in PSR by exploiting the advantages in dimensionality reduction, incremental matrix decomposition, and compressed sensing. In this work, we take these ideas further by considering more powerful tensor decomposition techniques.

Krishnamurthy et al. (2016) recently analyzed the problem of learning in contextual-MDPs and proved sample complexity bounds polynomial in the capacity of the policy space, the number of states, and the horizon. While their objective is to minimize the regret over a finite horizon, we instead consider the infinite horizon problem. It is an open question to analyze and modify our spectral UCRL algorithm for the finite horizon problem. As stated earlier, contextual MDPs are a special class of POMDPs for which memoryless policies are optimal. While they assume that the samples are drawn from a contextual MDP, we can handle a much more general class of POMDPs, and we minimize regret with respect to the best memoryless policy for the given POMDP.

Finally, a related problem is considered by Ortner et al. (2014), where a series of possible representations based on observation histories is available to the agent but only one of them is actually Markov. A UCRL-like strategy is adopted and shown to achieve near-optimal regret.

In this paper, we focus on the learning problem, while we consider access to an optimization oracle to compute the optimal memoryless policy. The problem of planning in general POMDPs is intractable (PSPACE-complete for finite horizon (Papadimitriou and Tsitsiklis, 1987) and uncomputable for infinite horizon (Madani, 1998)).

Many exact, approximate, and heuristic methods have been proposed to compute the optimal policy (see Spaan (2012) for a recent survey). An alternative approach is to consider memoryless policies which directly map observations (or a finite history) to actions (Littman, 1994; Singh et al., 1994; Li et al., 2011). While deterministic policies may perform poorly, stochastic memoryless policies are shown to be near-optimal in many domains (Barto et al., 1983; Loch and Singh, 1998; Williams and Singh, 1998) and even optimal in the specific case of contextual MDPs (Krishnamurthy et al., 2016). Although computing the optimal stochastic memoryless policy is still NP-hard (Littman, 1994), several model-based and model-free methods are shown to converge to nearly-optimal policies with polynomial complexity under some conditions on the POMDP (Jaakkola et al., 1995; Li et al., 2011). In this work, we employ memoryless policies and prove regret bounds for reinforcement learning of POMDPs. The above works suggest that focusing to memoryless policies may not be a restrictive limitation in practice.

1.3. Paper Organization

The paper is organized as follows. Sect. 2 introduces the notation (summarized also in a table in Sect. 6) and the technical assumptions concerning the POMDP and the space of memoryless policies that we consider. Sect. 3 introduces the spectral method for the estimation of POMDP parameters together with Thm. 3. In Sect. 4, we outline SM-UCRL where we integrate the spectral method into an exploration–exploitation strategy and we prove the regret bound of Thm. 4. Sect. 5 draws conclusions and discuss possible directions for future investigation. The proofs are reported in the appendix together with preliminary empirical results showing the effectiveness of the proposed method.

2. Preliminaries

A POMDP M is a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{Y}, \mathcal{R}, f_T, f_R, f_O \rangle$, where $\mathcal{X}$ is a finite state space with cardinality $|\mathcal{X}| = X$, $\mathcal{A}$ is a finite action space with cardinality $|\mathcal{A}| = A$, $\mathcal{Y}$ is a finite

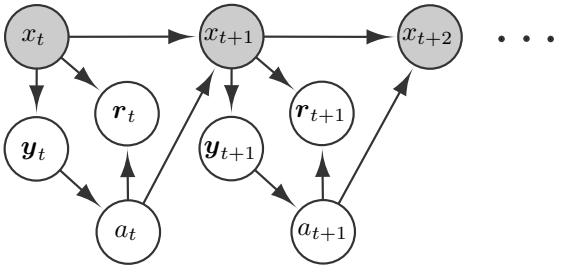


Figure 1: Graphical model of a POMDP under memoryless policies.

observation space with cardinality $|\mathcal{Y}| = Y$, and $\mathcal{R}$ is a finite reward space with cardinality $|\mathcal{R}| = R$ and largest reward $r_{\max}$. For notation convenience, we use a vector notation for the elements in $\mathcal{Y}$ and $\mathcal{R}$, so that $\boldsymbol{y} \in \mathbb{R}^Y$ and $\boldsymbol{r} \in \mathbb{R}^R$ are indicator vectors with entries equal to 0 except a 1 in the position corresponding to a specific element in the set (e.g., $\boldsymbol{y} = \boldsymbol{e}_n$ refers to the n-th element in $\mathcal{Y}$). We use $i, j \in [X]$ to index states, $k, l \in [A]$ for actions, $m \in [R]$ for rewards, and $n \in [Y]$ for observations. Finally, f_T denotes the transition density, so that $f_T(x'|x, a)$ is the probability of transition to x' given the state-action pair (x, a), f_R is the reward density, so that $f_R(\boldsymbol{r}|x, a)$ is the probability of receiving the reward in $\mathcal{R}$ corresponding to the value of the indicator vector $\boldsymbol{r}$ given the state-action pair (x, a), and f_O is the observation density, so that $f_O(\boldsymbol{y}|x)$ is the probability of receiving the observation in $\mathcal{Y}$ corresponding to the indicator vector $\boldsymbol{y}$ given the state x. Whenever convenient, we use tensor forms for the density functions such that

$$\begin{aligned} T_{i,j,l} &= \mathbb{P}[x_{t+1} = j | x_t = i, a_t = l] = f_T(j|i, l), && s.t.\ \ T \in \mathbb{R}^{X \times X \times A} \\ O_{n,i} &= \mathbb{P}[\boldsymbol{y} = \boldsymbol{e}_n | x = i] = f_O(\boldsymbol{e}_n|i), && s.t.\ \ O \in \mathbb{R}^{Y \times X} \\ \Gamma_{i,l,m} &= \mathbb{P}[\boldsymbol{r} = \boldsymbol{e}_m | x = i, a = l] = f_R(\boldsymbol{e}_m|i, l), && s.t.\ \ \Gamma \in \mathbb{R}^{X \times A \times R}. \end{aligned}$$

We also denote by $T_{:,j,l}$ the fiber (vector) in $\mathbb{R}^X$ obtained by fixing the arrival state j and action l and by $T_{:,:,l} \in \mathbb{R}^{X \times X}$ the transition matrix between states when using action l. The graphical model associated to the POMDP is illustrated in Fig. 1.

We focus on stochastic memoryless policies which map observations to actions and for any policy π we denote by $f_\pi(a|\boldsymbol{y})$ its density function. We denote by $\mathcal{P}$ the set of all stochastic memoryless policies that have a non-zero probability to explore all actions:

$$\mathcal{P} = \{\pi : \min_{\boldsymbol{y}} \min_{a} f_\pi(a|\boldsymbol{y}) > \pi_{\min}\}.$$

Acting according to a policy π in a POMDP M defines a Markov chain characterized by a transition density

$$f_{T,\pi}(x'|x) = \sum_a \sum_{\boldsymbol{y}} f_\pi(a|\boldsymbol{y}) f_O(\boldsymbol{y}|x) f_T(x'|x, a),$$

and a stationary distribution ω_π over states such that $\omega_\pi(x) = \sum_{x'} f_{T,\pi}(x'|x)\omega_\pi(x')$. The expected average reward performance of a policy π is

$$\eta(\pi; M) = \sum_x \omega_\pi(x) \overline{r}_\pi(x),$$

where $\overline{r}_\pi(x)$ is the expected reward of executing policy π in state x defined as

$$\overline{r}_\pi(x) = \sum_a \sum_{\boldsymbol{y}} f_O(\boldsymbol{y}|x) f_\pi(a|\boldsymbol{y}) \overline{r}(x,a),$$

and $\overline{r}(x,a) = \sum_r r f_R(r|x,a)$ is the expected reward for the state-action pair (x,a). The best stochastic memoryless policy in $\mathcal{P}$ is $\pi^+ = \arg\max_{\pi\in\mathcal{P}} \eta(\pi; M)$ and we denote by $\eta^+ = \eta(\pi^+; M)$ its average reward.[3] Throughout the paper we assume that we have access to an optimization oracle returning the optimal policy π^+ in $\mathcal{P}$ for any POMDP M. We need the following assumptions on the POMDP M.

Assumption 1 (Ergodicity) *For any policy $\pi \in \mathcal{P}$, the corresponding Markov chain $f_{T,\pi}$ is ergodic, so $\omega_\pi(x) > 0$ for all states $x \in \mathcal{X}$.*

We further characterize the Markov chains that can be generated by the policies in $\mathcal{P}$. For any ergodic Markov chain with stationary distribution ω_π, let $f_{1\to t}(x_t|x_1)$ by the distribution over states reached by a policy π after t steps starting from an initial state x_1. The inverse mixing time $\rho_{\text{mix},\pi}(t)$ of the chain is defined as

$$\rho_{\text{mix},\pi}(t) = \sup_{x_1} \|f_{1\to t}(\cdot|x_1) - \omega_\pi\|_{\text{TV}},$$

where $\|\cdot\|_{\text{TV}}$ is the total-variation metric. Kontorovich et al. (2014) show that for any ergodic Markov chain the mixing time can be bounded as

$$\rho_{\text{mix},\pi}(t) \le G(\pi)\theta^{t-1}(\pi),$$

where $1 \le G(\pi) < \infty$ is the *geometric ergodicity* and $0 \le \theta(\pi) < 1$ is the *contraction coefficient* of the Markov chain generated by policy π.

Assumption 2 (Full Column-Rank) *The observation matrix $O \in \mathbb{R}^{Y\times X}$ is full column rank.*

This assumption guarantees that the distribution $f_O(\cdot|x)$ in a state x (i.e., a column of the matrix O) is not the result of a linear combination of the distributions over other states. We show later that this is a sufficient condition to recover f_O since it makes all states *distinguishable* from the observations and it also implies that $Y \ge X$. Notice that POMDPs have been often used in the opposite scenario ($X \gg Y$) in applications such as robotics, where imprecise sensors prevents from distinguishing different states. On the other hand, there are many domains in which the number of observations may be much larger than the set of states that define the dynamics of the system. A typical example is the case of spoken dialogue systems (Atrash and Pineau, 2006; Png et al., 2012), where the observations (e.g., sequences of words uttered by the user) is much larger than the state of the conversation (e.g., the actual meaning that the user intended to communicate). A similar scenario is found in medical applications (Hauskrecht and Fraser, 2000), where the state of a patient (e.g., sick or healthy) can produce a huge body of different (random) observations. In these problems it is crucial to be able to reconstruct the underlying small state space and the actual dynamics of the system from the observations.

3. We use π^+ rather than π^* to recall the fact that we restrict the attention to $\mathcal{P}$ and the actual optimal policy for a POMDP in general should be constructed on the belief-MDP.

Assumption 3 (Invertible) *For any action $a \in [A]$, the transition matrix $T_{:,:,a} \in \mathbb{R}^{X \times X}$ is invertible.*

Similar to the previous assumption, this means that for any action a the distribution $f_T(\cdot|x,a)$ cannot be obtained as linear combination of distributions over other states, and it is a sufficient condition to be able to recover the transition tensor. Both Asm. 2 and 3 are strictly related to the assumptions introduced by Anandkumar et al. (2014) for tensor methods in HMMs. In Sect. 4 we discuss how they can be partially relaxed.

3. Learning the Parameters of the POMDP

In this section we introduce a novel spectral method to estimate the POMDP parameters f_T, f_O, and f_R. A stochastic policy π is used to generate a trajectory $(\boldsymbol{y}_1, a_1, \boldsymbol{r}_1, \dots, \boldsymbol{y}_N, a_N, \boldsymbol{r}_N)$ of N steps. We need the following assumption that, together with Asm. 1, guarantees that all states and actions are constantly visited.

Assumption 4 (Policy Set) *The policy π belongs to $\mathcal{P}$.*

Similar to the case of HMMs, the key element to apply the spectral methods is to construct a multi-view model for the hidden states. Despite its similarity, the spectral method developed for HMM by Anandkumar et al. (2014) cannot be directly employed here. In fact, in HMMs the state transition and the observations only depend on the current state. On the other hand, in POMDPs the probability of a transition to state x' not only depends on x, but also on action a. Since the action is chosen according to a memoryless policy π based on the current observation, this creates an indirect dependency of x' on observation $\boldsymbol{y}$, which makes the model more intricate.

3.1. The multi-view model

We estimate POMDP parameters for each action $l \in [A]$ separately. Let $t \in [2, N-1]$ be a step at which $a_t = l$, we construct three views $(a_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{r}_{t-1})$, $(\boldsymbol{y}_t, \boldsymbol{r}_t)$, and $(\boldsymbol{y}_{t+1})$ which all contain observable elements. As it can be seen in Fig. 1, all three views provide some information about the hidden state x_t (e.g., the observation $\boldsymbol{y}_{t-1}$ triggers the action a_{t-1}, which influence the transition to x_t). A careful analysis of the graph of dependencies shows that conditionally on x_t, a_t all the views are independent. For instance, let us consider $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t+1}$. These two random variables are clearly dependent since $\boldsymbol{y}_t$ influences action a_t, which triggers a transition to x_{t+1} that emits an observation $\boldsymbol{y}_{t+1}$. Nonetheless, it is sufficient to condition on the action $a_t = l$ to break the dependency and make $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t+1}$ independent. Similar arguments hold for all the other elements in the views, which can be used to recover the latent variable x_t. More formally, we encode the triple $(a_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{r}_{t-1})$ into a vector $\boldsymbol{v}_{1,t}^{(l)} \in \mathbb{R}^{A \cdot Y \cdot R}$, so that view $\boldsymbol{v}_{1,t}^{(l)} = \boldsymbol{e}_s$ whenever $a_{t-1} = k$, $\boldsymbol{y}_{t-1} = \boldsymbol{e}_n$, and $\boldsymbol{r}_{t-1} = \boldsymbol{e}_m$ for a suitable mapping between the index $s \in \{1, \dots, A \cdot Y \cdot R\}$ and the indices (k, n, m) of the action, observation, and reward. Similarly, we proceed for $\boldsymbol{v}_{2,t}^{(l)} \in \mathbb{R}^{Y \cdot R}$ and $\boldsymbol{v}_{3,t}^{(l)} \in \mathbb{R}^{Y}$. We introduce the three view matrices $V_\nu^{(l)}$ with $\nu \in \{1,2,3\}$ associated with

action l defined as $V_1^{(l)} \in \mathbb{R}^{A\cdot Y\cdot R\times X}$, $V_2^{(l)} \in \mathbb{R}^{Y\cdot R\times X}$, and $V_3^{(l)} \in \mathbb{R}^{Y\times X}$ such that

$$[V_1^{(l)}]_{s,i} = \mathbb{P}\big(\boldsymbol{v}_1^{(l)} = \boldsymbol{e}_s | x_2 = i\big) = [V_1^{(l)}]_{(n,m,k),i} = \mathbb{P}\big(\boldsymbol{y}_1 = \boldsymbol{e}_n, \boldsymbol{r}_1 = \boldsymbol{e}_m, a_1 = k | x_2 = i\big),$$
$$[V_2^{(l)}]_{s,i} = \mathbb{P}\big(\boldsymbol{v}_2^{(l)} = \boldsymbol{e}_s | x_2 = i, a_2 = l\big) = [V_2^{(l)}]_{(n',m'),i} = \mathbb{P}\big(\boldsymbol{y}_2 = \boldsymbol{e}_{n'}, \boldsymbol{r}_2 = \boldsymbol{e}_{m'} | x_2 = i, a_2 = l\big),$$
$$[V_3^{(l)}]_{s,i} = \mathbb{P}\big(\boldsymbol{v}_3^{(l)} = \boldsymbol{e}_s | x_2 = i, a_2 = l\big) = [V_3^{(l)}]_{n'',i} = \mathbb{P}\big(\boldsymbol{y}_3 = \boldsymbol{e}_{n''} | x_2 = i, a_2 = l\big).$$

In the following we denote by $\mu_{\nu,i}^{(l)} = [V_\nu^{(l)}]_{:,i}$ the ith column of the matrix $V_\nu^{(l)}$ for any $\nu \in \{1,2,3\}$. Notice that Asm. 2 and Asm. 3 imply that all the view matrices are full column rank. As a result, we can construct a multi-view model that relates the spectral decomposition of the second and third moments of the (modified) views with the columns of the third view matrix.

Proposition 1 (Thm. 3.6 in (Anandkumar et al., 2014)) *Let $K_{\nu,\nu'}^{(l)} = \mathbb{E}\big[\boldsymbol{v}_\nu^{(l)} \otimes \boldsymbol{v}_{\nu'}^{(l)}\big]$ be the correlation matrix between views ν and ν' and $K^\dagger$ is its pseudo-inverse. We define a modified version of the first and second views as*

$$\widetilde{\boldsymbol{v}}_1^{(l)} := K_{3,2}^{(l)}(K_{1,2}^{(l)})^\dagger \boldsymbol{v}_1^{(l)}, \quad \widetilde{\boldsymbol{v}}_2^{(l)} := K_{3,1}^{(l)}(K_{2,1}^{(l)})^\dagger \boldsymbol{v}_2^{(l)}. \tag{1}$$

Then the second and third moment of the modified views have a spectral decomposition as

$$M_2^{(l)} = \mathbb{E}\big[\widetilde{\boldsymbol{v}}_1^{(l)} \otimes \widetilde{\boldsymbol{v}}_2^{(l)}\big] = \sum_{i=1}^X \omega_\pi^{(l)}(i)\mu_{3,i}^{(l)} \otimes \mu_{3,i}^{(l)}, \tag{2}$$

$$M_3^{(l)} = \mathbb{E}\big[\widetilde{\boldsymbol{v}}_1^{(l)} \otimes \widetilde{\boldsymbol{v}}_2^{(l)} \otimes \boldsymbol{v}_3^{(l)}\big] = \sum_{i=1}^X \omega_\pi^{(l)}(i)\mu_{3,i}^{(l)} \otimes \mu_{3,i}^{(l)} \otimes \mu_{3,i}^{(l)}, \tag{3}$$

where $\otimes$ is the tensor product and $\omega_\pi^{(l)}(i) = \mathbb{P}[x = i | a = l]$ is the state stationary distribution of π conditioned on action l being selected by policy π.

Notice that under Asm. 1 and 4, $\omega_\pi^{(l)}(i)$ is always bounded away from zero. Given $M_2^{(l)}$ and $M_3^{(l)}$ we can recover the columns of the third view $\mu_{3,i}^{(l)}$ directly applying the standard spectral decomposition method of Anandkumar et al. (2012). We need to recover the other views from $V_3^{(l)}$. From the definition of modified views in Eq. 1 we have

$$\begin{aligned}
\mu_{3,i}^{(l)} &= \mathbb{E}\big[\widetilde{\boldsymbol{v}}_1 | x_2 = i, a_2 = l\big] = K_{3,2}^{(l)}(K_{1,2}^{(l)})^\dagger \mathbb{E}\big[\boldsymbol{v}_1 | x_2 = i, a_2 = l\big] = K_{3,2}^{(l)}(K_{1,2}^{(l)})^\dagger \mu_{1,i}^{(l)}, \\
\mu_{3,i}^{(l)} &= \mathbb{E}\big[\widetilde{\boldsymbol{v}}_2 | x_2 = i, a_2 = l\big] = K_{3,1}^{(l)}(K_{2,1}^{(l)})^\dagger \mathbb{E}\big[\boldsymbol{v}_2 | x_2 = i, a_2 = l\big] = K_{3,1}^{(l)}(K_{2,1}^{(l)})^\dagger \mu_{2,i}^{(l)}.
\end{aligned} \tag{4}$$

Thus, it is sufficient to invert (pseudo invert) the two equations above to obtain the columns of both the first and second view matrices. This process could be done in any order, e.g., we could first estimate the second view by applying a suitable symmetrization step (Eq. 1) and recovering the first and the third views by reversing similar equations to Eq. 4. On the other hand, we cannot repeat the symmetrization step multiple times and estimate the views independently (i.e., without inverting Eq. 4). In fact, the estimates returned by the spectral method are consistent "up to a suitable permutation" on the indexes of the states.

While this does not pose any problem in computing one single view, if we estimated two views independently, the permutation may be different, thus making them non-consistent and impossible to use in recovering the POMDP parameters. On the other hand, estimating first one view and recovering the others by inverting Eq. 4 guarantees the consistency of the labeling of the hidden states.

3.2. Recovery of POMDP parameters

Once the views $\{V_\nu^{(l)}\}_{\nu=2}^3$ are computed from $M_2^{(l)}$ and $M_3^{(l)}$, we can derive f_T, f_O, and f_R. In particular, all parameters of the POMDP can be obtained by manipulating the second and third view as illustrated in the following lemma.

Lemma 2 *Given the views $V_2^{(l)}$ and $V_3^{(l)}$, for any state $i \in [X]$ and action $l \in [A]$, the POMDP parameters are obtained as follows. For any reward $m \in [R]$ the reward density is*

$$f_R(\boldsymbol{e}_{m'}|i,l) = \sum_{n'=1}^{Y} [V_2^{(l)}]_{(n',m'),i}; \tag{5}$$

for any observation $n' \in [Y]$ the observation density is

$$f_O^{(l)}(\boldsymbol{e}_{n'}|i) = \sum_{m'=1}^{R} \frac{[V_2^{(l)}]_{(n',m'),i}}{f_\pi(l|\boldsymbol{e}_{n'})\rho(i,l)}, \tag{6}$$

with

$$\rho(i,l) = \sum_{m'=1}^{R}\sum_{n'=1}^{Y} \frac{[V_2^{(l)}]_{(n',m'),i}}{f_\pi(l|\boldsymbol{e}_{n'})} = \frac{1}{\mathbb{P}(a_2 = l|x_2 = i)}.$$

Finally, each second mode of the transition tensor $T \in \mathbb{R}^{X\times X\times A}$ is obtained as

$$[T]_{i,:,l} = O^\dagger [V_3^{(l)}]_{:,i}, \tag{7}$$

where $O^\dagger$ is the pseudo-inverse of matrix observation O and $f_T(\cdot|i,l) = [T]_{i,:,l}$.

In the previous statement we use $f_O^{(l)}$ to denote that the observation model is recovered from the view related to action l. While in the exact case, all $f_O^{(l)}$ are identical, moving to the empirical version leads to A different estimates, one for each action view used to compute it. Among them, we will select the estimate with the better accuracy.

Empirical estimates of POMDP parameters. In practice, $M_2^{(l)}$ and $M_3^{(l)}$ are not available and need to be estimated from samples. Given a trajectory of N steps obtained executing policy π, let $\mathcal{T}(l) = \{t \in [2, N-1] : a_t = l\}$ be the set of steps when action l is played, then we collect all the triples $(a_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{r}_{t-1})$, $(\boldsymbol{y}_t, \boldsymbol{r}_t)$ and $(\boldsymbol{y}_{t+1})$ for any $t \in \mathcal{T}(l)$ and construct the corresponding views $\boldsymbol{v}_{1,t}^{(l)}$, $\boldsymbol{v}_{2,t}^{(l)}$, $\boldsymbol{v}_{3,t}^{(l)}$. Then we symmetrize the views using

4. Each column of $O^{(l)}$ corresponds to $\ell 1$-closest column of $O^{(l^*)}$

Algorithm 1 Estimation of the POMDP parameters. The routine TensorDecomposition refers to the spectral tensor decomposition method of Anandkumar et al. (2012).

Input:

Policy density f_π, number of states X

Trajectory $\langle (\boldsymbol{y}_1, a_1, \boldsymbol{r}_1), (\boldsymbol{y}_2, a_2, \boldsymbol{r}_2), \dots, (\boldsymbol{y}_N, a_N, \boldsymbol{r}_N) \rangle$

Variables:

Estimated second and third views $\widehat{V}_2^{(l)}$, and $\widehat{V}_3^{(l)}$ for any action $l \in [A]$

Estimated observation, reward, and transition models $\widehat{f}_O$, $\widehat{f}_R$, $\widehat{f}_T$

for $l = 1, \dots, A$ **do**

Set $\mathcal{T}(l) = \{t \in [N-1] : a_t = l\}$ and $N(l) = |\mathcal{T}(l)|$

Construct views $\boldsymbol{v}_{1,t}^{(l)} = (a_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{r}_{t-1})$, $\boldsymbol{v}_{2,t}^{(l)} = (\boldsymbol{y}_t, \boldsymbol{r}_t)$, $\boldsymbol{v}_{3,t}^{(l)} = \boldsymbol{y}_{t+1}$ for any $t \in \mathcal{T}(l)$

Compute covariance matrices $\widehat{K}_{3,1}^{(l)}$, $\widehat{K}_{2,1}^{(l)}$, $\widehat{K}_{3,2}^{(l)}$ as

$$\widehat{K}_{\nu,\nu'}^{(l)} = \frac{1}{N(l)} \sum_{t \in \mathcal{T}(l)} \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)}; \;\; \nu, \nu' \in \{1,2,3\}$$

Compute modified views $\widetilde{\boldsymbol{v}}_{1,t}^{(l)} := \widehat{K}_{3,2}^{(l)} (\widehat{K}_{1,2}^{(l)})^\dagger \boldsymbol{v}_1$, $\widetilde{\boldsymbol{v}}_{2,t}^{(l)} := \widehat{K}_{3,1}^{(l)} (\widehat{K}_{2,1}^{(l)})^\dagger \boldsymbol{v}_{2,t}^{(l)}$ for any $t \in \mathcal{T}(l)$

Compute second and third moments

$$\widehat{M}_2^{(l)} = \frac{1}{N(l)} \sum_{t \in \mathcal{T}_l} \widetilde{\boldsymbol{v}}_{1,t}^{(l)} \otimes \widetilde{\boldsymbol{v}}_{2,t}^{(l)}, \quad \widehat{M}_3^{(l)} = \frac{1}{N(l)} \sum_{t \in \mathcal{T}_l} \widetilde{\boldsymbol{v}}_{1,t}^{(l)} \otimes \widetilde{\boldsymbol{v}}_{2,t}^{(l)} \otimes \boldsymbol{v}_{3,t}^{(l)}$$

Compute $\widehat{V}_3^{(l)} =$ TensorDecomposition$(\widehat{M}_2^{(l)}, \widehat{M}_3^{(l)})$

Compute $\widehat{\mu}_{2,i}^{(l)} = \widehat{K}_{1,2}^{(l)} (\widehat{K}_{3,2}^{(l)})^\dagger \widehat{\mu}_{3,i}^{(l)}$ for any $i \in [X]$

Compute $\widehat{f}(\boldsymbol{e}_m | i, l) = \sum_{n'=1}^{Y} [\widehat{V}_2^{(l)}]_{(n',m),i}$ for any $i \in [X]$, $m \in [R]$

Compute $\rho(i,l) = \sum_{m'=1}^{R} \sum_{n'=1}^{Y} \frac{[V_2^{(l)}]_{(n',m'),i}}{f_\pi(l | \boldsymbol{e}_{n'})}$ for any $i \in [X]$, $n \in [Y]$

Compute $\widehat{f}_O^{(l)}(\boldsymbol{e}_n | i) = \sum_{m'=1}^{R} \frac{[V_2^{(l)}]_{(n,m'),i}}{f_\pi(l | \boldsymbol{e}_n) \rho(i,l)}$ for any $i \in [X]$, $n \in [Y]$

end for

Compute bounds $\mathcal{B}_O^{(l)}$

Set $l^* = \arg\min_l \mathcal{B}_O^{(l)}$, $\widehat{f}_O = \widehat{f}_O^{l^*}$ and construct matrix $[\widehat{O}]_{n,j} = \widehat{f}_O(\boldsymbol{e}_n | j)$

Reorder columns of matrices $\widehat{V}_2^{(l)}$ and $\widehat{V}_3^{(l)}$ such that matrix $O^{(l)}$ and $O^{(l^*)}$ match, $\forall l \in [A]$[4]

for $i \in [X]$, $l \in [A]$ **do**

Compute $[T]_{i,:,l} = \widehat{O}^\dagger [\widehat{V}_3^{(l)}]_{:,i}$

end for

Return: $\widehat{f}_R$, $\widehat{f}_T$, $\widehat{f}_O$, $\mathcal{B}_R$, $\mathcal{B}_T$, $\mathcal{B}_O$

empirical estimates of the covariance matrices and build the empirical version of Eqs. 2 and 3 using $N(l) = |\mathcal{T}(l)|$ samples, thus obtaining

$$\widehat{M}_2^{(l)} = \frac{1}{N(l)} \sum_{t \in \mathcal{T}_l} \widetilde{\boldsymbol{v}}_{1,t}^{(l)} \otimes \widetilde{\boldsymbol{v}}_{2,t}^{(l)}, \qquad \widehat{M}_3^{(l)} = \frac{1}{N(l)} \sum_{t \in \mathcal{T}_l} \widetilde{\boldsymbol{v}}_{1,t}^{(l)} \otimes \widetilde{\boldsymbol{v}}_{2,t}^{(l)} \otimes \boldsymbol{v}_{3,t}^{(l)}. \tag{8}$$

Given the resulting $\widehat{M}_2^{(l)}$ and $\widehat{M}_3^{(l)}$, we apply the spectral tensor decomposition method to recover an empirical estimate of the third view $\widehat{V}_3^{(l)}$ and invert Eq. 4 (using estimated covariance matrices) to obtain $\widehat{V}_2^{(l)}$. Finally, the estimates $\widehat{f}_O$, $\widehat{f}_T$, and $\widehat{f}_R$ are obtained by plugging the estimated views $\widehat{V}_\nu$ in the process described in Lemma 2.
Spectral methods indeed recover the factor matrices up to a permutation of the hidden states. In this case, since we separately carry out spectral decompositions for different actions, we recover permuted factor matrices. Since the observation matrix O is common to all the actions, we use it to align these decompositions. Let's define d_O

$$d_O =: \min_{x,x'} \|f_O(\cdot|x) - f_O(\cdot|x')\|_1$$

Actually, d_O is the minimum separability level of matrix O. When the estimation error over columns of matrix O are less than $4d_O$, then one can come over the permutation issue by matching columns of O^l matrices. In T condition is reflected as a condition that the number of samples for each action has to be larger some number.

The overall method is summarized in Alg. 1. The empirical estimates of the POMDP parameters enjoy the following guarantee.

Theorem 3 (Learning Parameters) *Let $\widehat{f}_O$, $\widehat{f}_T$, and $\widehat{f}_R$ be the estimated POMDP models using a trajectory of N steps. We denote by $\sigma^{(l)}_{\nu,\nu'} = \sigma_X(K^{(l)}_{\nu,\nu'})$ the smallest non-zero singular value of the covariance matrix $K_{\nu,\nu'}$, with $\nu,\nu' \in \{1,2,3\}$, and by $\sigma_{\min}(V^{(l)}_\nu)$ the smallest singular value of the view matrix $V^{(l)}_\nu$ (strictly positive under Asm. 2 and Asm. 3), and we define $\omega^{(l)}_{\min} = \min_{x\in\mathcal{X}} \omega^{(l)}_\pi(x)$ (strictly positive under Asm. 1). If for any action $l \in [A]$, the number of samples $N(l)$ satisfies the condition*

$$N(l) \geq \max\left\{ \frac{4}{(\sigma^{(l)}_{3,1})^2}, \frac{16C_O^2 YR}{{\lambda^{(l)}}^2 d_O^2}, \left(\frac{G(\pi)\frac{2\sqrt{2}+1}{1-\theta(\pi)}}{\omega^{(l)}_{\min} \min_{\nu\in\{1,2,3\}} \{\sigma^2_{\min}(V^{(l)}_\nu)\}} \right)^2 \Theta^{(l)} \right\} \log\left(\frac{2(Y^2+AYR)}{\delta}\right), \tag{9}$$

with $\Theta^{(l)}$, defined in Eq 27[5], and $G(\pi), \theta(\pi)$ are the geometric ergodicity and the contraction coefficients of the corresponding Markov chain induced by π, then for any $\delta \in (0,1)$ and for any state $i \in [X]$ and action $l \in [A]$ we have

$$\|\widehat{f}^{(l)}_O(\cdot|i) - f_O(\cdot|i)\|_1 \leq \mathcal{B}^{(l)}_O := \frac{C_O}{\lambda^{(l)}} \sqrt{\frac{YR\log(1/\delta)}{N(l)}}, \tag{10}$$

$$\|\widehat{f}_R(\cdot|i,l) - f_R(\cdot|i,l)\|_1 \leq \mathcal{B}^{(l)}_R := \frac{C_R}{\lambda^{(l)}} \sqrt{\frac{YR\log(1/\delta)}{N(l)}}, \tag{11}$$

5. We do not report the explicit definition of $\Theta^{(l)}$ here because it contains exactly the same quantities, such as $\omega^{(l)}_{\min}$, that are already present in other parts of the condition of Eq. 9.

$$\|\widehat{f}_T(\cdot|i,l) - f_T(\cdot|i,l)\|_2 \le \mathcal{B}_T^{(l)} := \frac{C_T}{\lambda^{(l)}} \sqrt{\frac{YRX^2 \log(1/\delta)}{N(l)}}, \tag{12}$$

with probability $1-6(Y^2+AYR)A\delta$ *(w.r.t. the randomness in the transitions, observations, and policy), where* C_O*,* C_R*, and* C_T *are numerical constants and*

$$\lambda^{(l)} = \sigma_{\min}(O)(\pi_{\min}^{(l)})^2 \sigma_{1,3}^{(l)}(\omega_{\min}^{(l)} \min_{\nu\in\{1,2,3\}} \{\sigma_{\min}^2(V_\nu^{(l)})\})^{3/2}. \tag{13}$$

Finally, we denote by $\widehat{f}_O$ *the most accurate estimate of the observation model, i.e., the estimate* $\widehat{f}_O^{(l^*)}$ *such that* $l^* = \arg\min_{l\in[A]} \mathcal{B}_O^{(l)}$ *and we denote by* $\mathcal{B}_O$ *its corresponding bound.*

Remark 1 (consistency and dimensionality). All previous errors decrease with a rate $\widetilde{O}(1/\sqrt{N(l)})$, showing the consistency of the spectral method, so that if all the actions are repeatedly tried over time, the estimates converge to the true parameters of the POMDP. This is in contrast with EM-based methods which typically get stuck in local maxima and return biased estimators, thus preventing from deriving confidence intervals.

The bounds in Eqs. 10, 11, 12 on $\widehat{f}_O$, $\widehat{f}_R$ and $\widehat{f}_T$ depend on X, Y, and R (and the number of actions only appear in the probability statement). The bound in Eq. 12 on $\widehat{f}_T$ is worse than the bounds for $\widehat{f}_R$ and $\widehat{f}_O$ in Eqs. 10, 11 by a factor of X^2. This seems unavoidable since $\widehat{f}_R$ and $\widehat{f}_O$ are the results of the manipulation of the matrix $V_2^{(l)}$ with $Y \cdot R$ columns, while estimating $\widehat{f}_T$ requires working on both $V_2^{(l)}$ and $V_3^{(l)}$. In addition, to come up with upper bound for $\widehat{f}_T$, more complicated bound derivation is needed and it has one step of Frobenious norms to $\ell 2$ norm transformation. The derivation procedure for $\widehat{f}_T$ is more complicated compared to $\widehat{f}_O$ and $\widehat{f}_R$ and adds the term X to the final bound. (Appendix. C)

Remark 2 (POMDP parameters and policy π). In the previous bounds, several terms depend on the structure of the POMDP and the policy π used to collect the samples:

- $\lambda^{(l)}$ captures the main problem-dependent terms. While $K_{1,2}$ and $K_{1,3}$ are full column-rank matrices (by Asm. 2 and 3), their smallest non-zero singular values influence the accuracy of the (pseudo-)inversion in the construction of the modified views in Eq. 1 and in the computation of the second view from the third using Eq. 4. Similarly the presence of $\sigma_{\min}(O)$ is justified by the pseudo-inversion of O used to recover the transition tensor in Eq. 7. Finally, the dependency on the smallest singular values $\sigma_{\min}^2(V_\nu^{(l)})$ is due to the tensor decomposition method (see App. J for more details).
- A specific feature of the bounds above is that they do not depend on the state i and the number of times it has been explored. Indeed, the inverse dependency on $\omega_{\min}^{(l)}$ in the condition on $N(l)$ in Eq. 9 implies that if a state j is poorly visited, then the empirical estimate of any other state i may be negatively affected. This is in striking contrast with the fully observable case where the accuracy in estimating, e.g., the reward model in state i and action l, simply depends on the number of times that state-action pair has been explored, even if some other states are never explored at all. This difference is intrinsic in the partial observable nature of the

POMDP, where we reconstruct information about the states (i.e., reward, transition, and observation models) only from indirect observations. As a result, in order to have accurate estimates of the POMDP structure, we need to rely on the policy π and the ergodicity of the corresponding Markov chain to guarantee that the whole state space is covered.

- Under Asm. 1 the Markov chain $f_{T,\pi}$ is ergodic for any $\pi \in \mathcal{P}$. Since no assumption is made on the fact that the samples generated from π being sampled from the stationary distribution, the condition on $N(l)$ depends on how fast the chain converge to ω_π and this is characterized by the parameters $G(\pi)$ and $\theta(\pi)$.

- If the policy is deterministic, then some actions would not be explored at all, thus leading to very inaccurate estimations (see e.g., the dependency on $f_\pi(l|\boldsymbol{y})$ in Eq. 6). The inverse dependency on $\pi_{\min}$ (defined in $\mathcal{P}$) accounts for the amount of exploration assigned to every actions, which determines the accuracy of the estimates. Furthermore, notice that also the singular values $\sigma_{1,3}^{(l)}$ and $\sigma_{1,2}^{(l)}$ depend on the distribution of the views, which in turn is partially determined by the policy π.

Notice that the first two terms are basically the same as in the bounds for spectral methods applied to HMM (Song et al., 2013), while the dependency on $\pi_{\min}$ is specific to the POMDP case. On the other hand, in the analysis of HMMs usually there is no dependency on the parameters G and θ because the samples are assumed to be drawn from the stationary distribution of the chain. Removing this assumption required developing novel results for the tensor decomposition process itself using extensions of matrix concentration inequalities for the case of Markov chain (not yet in the stationary distribution). The overall analysis is reported in App. I and J. It worth to note that, Kontorovich et al. (2013), without stationary assumption, proposes new method to learn the transition matrix of HMM model given factor matrix O, and it provides theoretical bound over estimation errors.

4. Spectral UCRL

The most interesting aspect of the estimation process illustrated in the previous section is that it can be applied when samples are collected using any policy π in the set $\mathcal{P}$. As a result, it can be integrated into any exploration-exploitation strategy where the policy changes over time in the attempt of minimizing the regret.

The algorithm. The SM-UCRL algorithm illustrated in Alg. 2 is the result of the integration of the spectral method into a structure similar to UCRL (Jaksch et al., 2010) designed to optimize the exploration-exploitation trade-off. The learning process is split into episodes of increasing length. At the beginning of each episode $k > 1$ (the first episode is used to initialize the variables), an estimated POMDP $\widehat{M}^{(k)} = (X, A, Y, R, \widehat{f}_T^{(k)}, \widehat{f}_R^{(k)}, \widehat{f}_O^{(k)})$ is computed using the spectral method of Alg. 1. Unlike in UCRL, SM-UCRL cannot use all the samples from past episodes. In fact, the distribution of the views $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ depends on the policy used to generate the samples. As a result, whenever the policy changes, the spectral method should be re-run using only the samples collected by that specific policy. Nonetheless we can exploit the fact that the spectral method is applied to each action separately. In SM-UCRL at episode k for each action l we use the samples coming from the past episode

Algorithm 2 The SM-UCRL algorithm.

Input: Confidence δ'
Variables:
Number of samples $N^{(k)}(l)$
Estimated observation, reward, and transition models $\widehat{f}_O^{(k)}$, $\widehat{f}_R^{(k)}$, $\widehat{f}_T^{(k)}$
Initialize: $t=1$, initial state x_1, $\delta=\delta'/N^6$, $k=1$
while $t<N$ **do**
Compute the estimated POMDP $\widehat{M}^{(k)}$ with the Alg. 1 using $N^{(k)}(l)$ samples per action
Compute the set of admissible POMDPs $\mathcal{M}^{(k)}$ using bounds in Thm. 3
Compute the optimistic policy $\widetilde{\pi}^{(k)}=\arg\max_{\pi\in\mathcal{P}}\max_{M\in\mathcal{M}^{(k)}}\eta(\pi;M)$
Set $v^{(k)}(l)=0$ for all actions $l\in[A]$
while $\forall l\in[A], v^{(k)}(l)<2N^{(k)}(l)$ **do**
Execute $a_t\sim f_{\widetilde{\pi}^{(k)}}(\cdot|\boldsymbol{y}_t)$
Obtain reward $\boldsymbol{r}_t$, observe next observation $\boldsymbol{y}_{t+1}$, and set $t=t+1$
end while
Store $N^{(k+1)}(l)=\max_{k'\le k}v^{(k')}(l)$ samples for each action $l\in[A]$
Set $k=k+1$
end while

which returned the largest number of samples for that action. Let $v^{(k)}(l)$ be the number of samples obtained during episode k for action l, we denote by $N^{(k)}(l)=\max_{k'<k}v^{(k')}(l)$ the largest number of samples available from past episodes for each action separately and we feed them to the spectral method to compute the estimated POMDP $\widehat{M}^{(k)}$ at the beginning of each episode k.

Given the estimated POMDP $\widehat{M}^{(k)}$ and the result of Thm. 3, we construct the set $\mathcal{M}^{(k)}$ of *admissible* POMDPs $\widetilde{M}=\langle\mathcal{X},\mathcal{A},\mathcal{Y},\mathcal{R},\widetilde{f}_T,\widetilde{f}_R,\widetilde{f}_O\rangle$ whose transition, reward, and observation models belong to the confidence intervals (e.g., $\|\widehat{f}_O^{(k)}(\cdot|i)-\widetilde{f}_O(\cdot|i)\|_1\le\mathcal{B}_O$ for any state i). By construction, this guarantees that the true POMDP M is included in $\mathcal{M}^{(k)}$ with high probability. Following the *optimism in face of uncertainty* principle used in UCRL, we compute the optimal memoryless policy corresponding to the most optimistic POMDP within $\mathcal{M}^{(k)}$. More formally, we compute[6]

$$\widetilde{\pi}^{(k)}=\arg\max_{\pi\in\mathcal{P}}\max_{M\in\mathcal{M}^{(k)}}\eta(\pi;M). \tag{14}$$

Intuitively speaking, the optimistic policy implicitly balances exploration and exploitation. Large confidence intervals suggest that $\widehat{M}^{(k)}$ is poorly estimated and further exploration is needed. Instead of performing a purely explorative policy, SM-UCRL still exploits the current estimates to construct the set of admissible POMDPs and selects the policy that

6. The computation of the optimal policy (within $\mathcal{P}$) in the optimistic model may not be trivial. Nonetheless, we first notice that given an horizon N, the policy needs to be recomputed at most $O(\log N)$ times (i.e., number of episodes). Furthermore, if an optimization oracle to $\eta(\pi;M)$ for a given POMDP M is available, then it is sufficient to randomly sample multiple POMDPs from $\mathcal{M}^{(k)}$ (which is a computationally cheap operation), find their corresponding best policy, and return the best among them. If *enough* POMDPs are sampled, the additional regret caused by this approximately optimistic procedure can be bounded as $\widetilde{O}(\sqrt{N})$.

maximizes the performance $\eta(\pi; M)$ over all POMDPs in $\mathcal{M}^{(k)}$. The choice of using the optimistic POMDP guarantees the $\widetilde{\pi}^{(k)}$ explores more often actions corresponding to large confidence intervals, thus contributing to improve the estimates over time. After computing the optimistic policy, $\widetilde{\pi}^{(k)}$ is executed until the number of samples for one action is doubled, i.e., $v^{(k)}(l) \geq 2N^{(k)}(l)$. This stopping criterion avoids switching policies too often and it guarantees that when an episode is terminated, enough samples are collected to compute a new (better) policy. This process is then repeated over episodes and we expect the optimistic policy to get progressively closer to the best policy $\pi^+ \in \mathcal{P}$ as the estimates of the POMDP get more and more accurate.

Regret analysis. We now study the regret SM-UCRL w.r.t. the best policy in $\mathcal{P}$. While in general π^+ may not be optimal, $\pi_{\min}$ is usually set to a small value and oftentimes the optimal memoryless policy itself is stochastic and it may actually be contained in $\mathcal{P}$. Given an horizon of N steps, the regret is defined as

$$\mathrm{Reg}_N = N\eta^+ - \sum_{t=1}^{N} r_t, \tag{15}$$

where r_t is the random reward obtained at time t according to the reward model f_R over the states traversed by the policies performed over episodes on the actual POMDP. To restate, similar to the MDP case, the complexity of learning in a POMDP M is partially determined by its diameter, defined as

$$D := \max_{x,x' \in \mathcal{X}, a,a' \in \mathcal{A}} \min_{\pi \in \mathcal{P}} \mathbb{E}\big[\tau(x', a'|x, a; \pi)\big], \tag{16}$$

which corresponds to the expected passing time from a state x to a state x' starting with action a and terminating with action a' and following the most effective memoryless policy $\pi \in \mathcal{P}$. The main difference w.r.t. to the diameter of the underlying MDP (see e.g., Jaksch et al. (2010)) is that it considers the distance between state-action pairs using memoryless policies instead of state-based policies.

Before stating our main result, we introduce the worst-case version of the parameters characterizing Thm. 3. Let $\overline{\sigma}_{1,2,3} := \min_{l \in [A]} \min_{\pi \in \mathcal{P}} \omega^{(l)}_{\min} \min_{\nu \in \{1,2,3\}} \sigma^2_{\min}(V^{(l)}_\nu)$ be the worst smallest non-zero singular value of the views for action l when acting according to policy π and let $\overline{\sigma}_{1,3} := \min_{l \in [A]} \min_{\pi \in \mathcal{P}} \sigma_{\min}(K^{(l)}_{1,3}(\pi))$ be the worst smallest non-zero singular value of the covariance matrix $K^{(l)}_{1,3}(\pi)$ between the first and third view for action l when acting according to policy π. Similarly, we define $\overline{\sigma}_{1,2}$. We also introduce $\overline{\omega}_{\min} := \min_{l \in [A]} \min_{x \in [X]} \min_{\pi \in \mathcal{P}} \omega^{(l)}_\pi(x)$ and

$$\overline{N} := \max_{l \in [A]} \max_{\pi \in \mathcal{P}} \max \left\{ \frac{4}{(\overline{\sigma}^2_{3,1})}, \frac{16 C_O^2 Y R}{{\lambda^{(l)}}^2 d_O^2}, \left(\frac{G(\pi) \frac{2\sqrt{2}+1}{1-\theta(\pi)}}{\overline{\omega}_{\min} \overline{\sigma}_{1,2,3}} \right)^2 \overline{\Theta}^{(l)} \right\} \log \left(2 \frac{(Y^2 + AYR)}{\delta} \right), \tag{17}$$

which is a sufficient number of samples for the statement of Thm. 3 to hold for any action and any policy. Here $\overline{\Theta}^{(l)}$ is also model related parameter which is defined in Eq. 36. Then we can prove the following result.

Theorem 4 (Regret Bound) *Consider a POMDP M with X states, A actions, Y observations, R rewards, characterized by a diameter D and with an observation matrix $O \in \mathbb{R}^{Y\times X}$ with smallest non-zero singular value $\sigma_X(O)$. We consider the policy space $\mathcal{P}$, such that the worst smallest non-zero value is $\overline{\sigma}_{1,2,3}$ (resp. $\overline{\sigma}_{1,3}$) and the worst smallest probability to reach a state is $\overline{\omega}_{\min}$. If* SM-UCRL *is run over N steps and the confidence intervals of Thm. 3 are used with $\delta = \delta'/N^6$ in constructing the plausible POMDPs $\widetilde{\mathcal{M}}$, then under Asm. 1, 2, and 3 it suffers from a total regret*

$$Reg_N \leq C_1 \frac{r_{\max}}{\overline{\lambda}} D X^{3/2} \sqrt{AYRN \log(N/\delta')} \tag{18}$$

with probability $1-\delta'$, where C_1 is numerical constants, and $\overline{\lambda}$ is the worst-case equivalent of Eq. 13 defined as

$$\overline{\lambda} = \sigma_{\min}(O) \pi_{\min}^2 \overline{\sigma}_{1,3} \overline{\sigma}_{1,2,3}^{3/2} \tag{19}$$

Remark 1 (comparison with MDPs). If UCRL could be run directly on the underlying MDP (i.e., as if the states where directly observable), then it would obtain a regret (Jaksch et al., 2010)

$$\mathrm{Reg}_N \leq C_{\mathrm{MDP}} D_{\mathrm{MDP}} X \sqrt{AN \log N},$$

where

$$D_{\mathrm{MDP}} := \max_{x,x' \in \mathcal{X}} \min_{\pi} \mathbb{E}[\tau(x'|x;\pi)],$$

with high probability. We first notice that the regret is of order $\widetilde{O}(\sqrt{N})$ in both MDP and POMDP bounds. This means that despite the complexity of POMDPs, SM-UCRL has the same dependency on the number of steps as in MDPs and it has a vanishing per-step regret. Furthermore, this dependency is known to be minimax optimal. The diameter D in general is larger than its MDP counterpart D_{MDP}, since it takes into account the fact that a memoryless policy, that can only work on observations, cannot be as efficient as a state-based policy in moving from one state to another. Although no lower bound is available for learning in POMDPs, we believe that this dependency is unavoidable since it is strictly related to the partial observable nature of POMDPs.

Remark 2 (dependency on POMDP parameters). The dependency on the number of actions is the same in both MDPs and POMDPs. On the other hand, moving to POMDPs naturally brings the dimensionality of the observation and reward models (Y,X, and R respectively) into the bound. The dependency on Y and R is directly inherited from the bounds in Thm. 3. The term $X^{3/2}$ is indeed the results of two terms; X and $X^{1/2}$. The first term is the same as in MDPs, while the second comes from the fact that the transition tensor is derived from Eq. 7. Finally, the term $\overline{\lambda}$ in Eq. 18 summarizes a series of terms which depend on both the policy space $\mathcal{P}$ and the POMDP structure. These terms are directly inherited from the spectral decomposition method used at the core of SM-UCRL and, as discussed in Sect. 3, they are due to the partial observability of the states and the fact that all (unobservable) states need to be visited often enough to be able to compute accurate estimate of the observation, reward, and transition models.

Remark 3 (computability of the confidence intervals). While it is a common assumption that the dimensionality X of the hidden state space is known as well as the number of actions, observations, and rewards, it is not often the case that the terms $\lambda^{(l)}$ appearing in Thm. 3 are actually available. While this does not pose any problem for a *descriptive* bound as in Thm. 3, in SM-UCRL we actually need to compute the bounds $\mathcal{B}_O^{(l)}$, $\mathcal{B}_R^{(l)}$, and $\mathcal{B}_T^{(l)}$ to explicitly construct confidence intervals. This situation is relatively common in many exploration–exploitation algorithms that require computing confidence intervals containing the range of the random variables or the parameters of their distributions in case of sub-Gaussian variables. In practice these values are often replaced by parameters that are tuned by hand and set to much smaller values than their theoretical ones. As a result, we can run SM-UCRL with the terms $\lambda^{(l)}$ replaced by a fixed parameter. Notice that any inaccurate choice in setting $\lambda^{(l)}$ would mostly translate into bigger multiplicative constants in the final regret bound or in similar bounds but with smaller probability.
In general, computing confidence bound is a hard problem, even for simpler cases such as Markov chains Hsu et al. (2015). Therefore finding upper confidence bounds for POMDP is challenging if we do not know its mixing properties. As it mentioned, another parameter is needed to compute upper confidence bound is $\lambda^{(l)}$ 13. As it is described in, in practice, one can replace the coefficient $\lambda^{(l)}$ with some constant which causes bigger multiplicative constant in final regret bound. Alternatively, one can estimate $\lambda^{(l)}$ from data. In this case, we add a lower order term to the regret which decays as $\frac{1}{N}$.

Remark 4 (relaxation on assumptions). Both Thm. 3 and 4 rely on the observation matrix $O \in \mathbb{R}^{Y \times X}$ being full column rank (Asm. 2). As discussed in Sect. 2 may not be verified in some POMDPs where the number of states is larger than the number of observations ($X > Y$). Nonetheless, it is possible to correctly estimate the POMDP parameters when O is not full column-rank by exploiting the additional information coming from the reward and action taken at step $t+1$. In particular, we can use the triple $(a_{t+1}, \boldsymbol{y}_{t+1}, r_{t+1})$ and redefine the third view $V_3^{(l)} \in \mathbb{R}^{d \times X}$ as

$$\begin{aligned}[V_3^{(l)}]_{s,i} &= \mathbb{P}(\boldsymbol{v}_3^{(l)} = \boldsymbol{e}_s | x_2 = i, a_2 = l) = [V_3^{(l)}]_{(n,m,k),i} \\ &= \mathbb{P}(\boldsymbol{y}_3 = \boldsymbol{e}_n, \boldsymbol{r}_3 = \boldsymbol{e}_m, a_3 = k | x_2 = i, a_2 = l),\end{aligned}$$

and replace Asm. 2 with the assumption that the view matrix $V_3^{(l)}$ is full column-rank, which basically requires having rewards that jointly with the observations are informative enough to reconstruct the hidden state. While this change does not affect the way the observation and the reward models are recovered in Lemma 2, (they only depend on the second view $V_2^{(l)}$), for the reconstruction of the transition tensor, we need to write the third view $V_3^{(l)}$

as

$$
\begin{aligned}
&[V_3^{(l)}]_{s,i} = [V_3^{(l)}]_{(n,m,k),i} \\
&= \sum_{j=1}^{X} \mathbb{P}\big(\boldsymbol{y}_3 = \boldsymbol{e}_n, \boldsymbol{r}_3 = \boldsymbol{e}_m, a_3 = k | x_2 = i, a_2 = l, x_3 = j\big) \mathbb{P}\big(x_3 = j | x_2 = i, a_2 = l\big) \\
&= \sum_{j=1}^{X} \mathbb{P}\big(\boldsymbol{r}_3 = \boldsymbol{e}_m | x_3 = j, a_3 = k\big) \mathbb{P}(a_3 = k | \boldsymbol{y}_3 = \boldsymbol{e}_n) \mathbb{P}\big(\boldsymbol{y}_3 = \boldsymbol{e}_n | x_3 = j\big) \mathbb{P}\big(x_3 = j | x_2 = i, a_2 = l\big) \\
&= f_\pi(k|\boldsymbol{e}_n) \sum_{j=1}^{X} f_R(\boldsymbol{e}_m | j, k) f_O(\boldsymbol{e}_n | j) f_T(j | i, l),
\end{aligned}
$$

where we factorized the three components in the definition of $V_3^{(l)}$ and used the graphical model of the POMDP to consider their dependencies. We introduce an auxiliary matrix $W \in \mathbb{R}^{d \times X}$ such that

$$
[W]_{s,j} = [W]_{(n,m,k),j} = f_\pi(k|\boldsymbol{e}_n) f_R(\boldsymbol{e}_m | j, k) f_O(\boldsymbol{e}_n | j),
$$

which contain all known values, and for any state i and action l we can restate the definition of the third view as

$$
W[T]_{i,:,l} = [V_3^{(l)}]_{:,i}, \tag{20}
$$

which allows computing the transition model as $[T]_{i,:,l} = W^\dagger [V_3^{(l)}]_{:,i}$, where $W^\dagger$ is the pseudo-inverse of W. While this change in the definition of the third view allows a significant relaxation of the original assumption, it comes at the cost of potentially worsening the bound on $\widehat{f}_T$ in Thm. 3. In fact, it can be shown that

$$
\|\widetilde{f}_T(\cdot|i,l) - f_T(\cdot|i,l)\|_F \leq \mathcal{B}'_T := \max_{l'=1,\dots,A} \frac{C_T A Y R}{\lambda^{(l')}} \sqrt{\frac{X A \log(1/\delta)}{N(l')}}. \tag{21}
$$

Beside the dependency on multiplication of Y, R, and R, which is due to the fact that now $V_3^{(l)}$ is a larger matrix, the bound for the transitions triggered by an action l scales with the number of samples from the least visited action. This is due to the fact that now the matrix W involves not only the action for which we are computing the transition model but all the other actions as well. As a result, if any of these actions is poorly visited, W cannot be accurately estimated is some of its parts and this may negatively affect the quality of estimation of the transition model itself. This directly propagates to the regret analysis, since now we require all the actions to be repeatedly visited enough. The immediate effect is the introduction of a different notion of diameter. Let $\tau_{M,\pi}^{(l)}$ the mean passage time between two steps where action l is chosen according to policy $\pi \in \mathcal{P}$, we define

$$
D_{\text{ratio}} = \max_{\pi \in \mathcal{P}} \frac{\max_{l \in \mathcal{A}} \tau_{M,\pi}^{(l)}}{\min_{l \in \mathcal{A}} \tau_{M,\pi}^{(l)}} \tag{22}
$$

as the diameter ratio, which defines the ratio between maximum mean passing time between choosing an action and choosing it again, over its minimum. As it mentioned above, in order to have an accurate estimate of f_T all actions need to be repeatedly explored. The D_{ratio} is small when each action is executed frequently enough and it is large when there is at least one action that is executed not as many as others. Finally, we obtain

$$\text{Reg}_N \leq \widetilde{O}\Big(\frac{r_{\max}}{\overline{\lambda}}\sqrt{YRD_{\text{ratio}}N\log N}X^{3/2}A(D+1)\Big).$$

While at first sight this bound is clearly worse than in the case of stronger assumptions, notice that $\overline{\lambda}$ now contains the smallest singular values of the newly defined views. In particular, as $V_3^{(l)}$ is larger, also the covariance matrices $K_{\nu,\nu'}$ are bigger and have larger singular values, which could significantly alleviate the inverse dependency on $\overline{\sigma}_{1,2}$ and $\overline{\sigma}_{2,3}$. As a result, relaxing Asm. 2 may not necessarily worsen the final bound since the bigger diameter may be compensated by better dependencies on other terms. We leave a more complete comparison of the two configurations (with or without Asm. 2) for future work.

5. Conclusion

We introduced a novel RL algorithm for POMDPs which relies on a spectral method to consistently identify the parameters of the POMDP and an optimistic approach for the solution of the exploration–exploitation problem. For the resulting algorithm we derive confidence intervals on the parameters and a minimax optimal bound for the regret.

This work opens several interesting directions for future development. **1)** SM-UCRL cannot accumulate samples over episodes since Thm. 3 requires samples to be drawn from a fixed policy. While this does not have a very negative impact on the regret bound, it is an open question how to apply the spectral method to all samples together and still preserve its theoretical guarantees. **2)** While memoryless policies may perform well in some domains, it is important to extend the current approach to bounded-memory policies. **3)** The POMDP is a special case of the predictive state representation (PSR) model Littman et al. (2001), which allows representing more sophisticated dynamical systems. Given the spectral method developed in this paper, a natural extension is to apply it to the more general PSR model and integrate it with an exploration–exploitation algorithm to achieve bounded regret.

6. Table of Notation

POMDP Notation (Sect. 2)	
$\boldsymbol{e}$	indicator vector
M	POMDP model
$\mathcal{X}, X, x, (i, j)$	state space, cardinality, element, indices
$\mathcal{Y}, Y, \boldsymbol{y}, n$	observation space, cardinality, indicator element, index
$\mathcal{A}, A, a, (l, k)$	action space, cardinality, element, indices
$\mathcal{R}, R, r, \boldsymbol{r}, m, r_{\max}$	reward space, cardinality, element, indicator element, index, largest value
$f_T(x'\|x, a), T$	transition density from state x to state x' given action a and transition tensor
$f_O(\boldsymbol{y}\|x), O$	observation density of indicator $\boldsymbol{y}$ given state x and observation matrix
$f_R(\boldsymbol{r}\|x, a), \Gamma$	reward density of indicator $\boldsymbol{r}$ given pair of state-action and reward tensor
$\pi, f_\pi(a\|\boldsymbol{y}), \Pi$	policy, policy density of action a given observation indicator $\boldsymbol{y}$ and policy matrix
$\pi_{\min}$, $\mathcal{P}$	smallest element of policy matrix and set of stochastic memoryless policies
$f_{\pi,T}(x'\|x)$	Markov chain transition density for policy π on a POMDP with transition density f_T
ω_π, $\omega_\pi^{(l)}$	stationary distribution over states given policy π and conditional on action l
$\eta(\pi, M)$	expected average reward of policy π in POMDP M
η^+	best expected average reward over policies in $\mathcal{P}$
POMDP Estimation Notation (Sect. 3)	
$\nu \in \{1, 2, 3\}$	index of the views
$\boldsymbol{v}_{\nu,t}^{(l)}$, $V_\nu^{(l)}$	νth view and view matrix at time t given $a_t = l$
$K_{\nu,\nu'}^{(l)}$, $\sigma_{\nu,\nu'}^{(l)}$	covariance matrix of views ν, ν' and its smallest non-zero singular value given action l
$M_2^{(l)}$, $M_3^{(l)}$	second and third order moments of the views given middle action l
$\widehat{f}_O^{(l)}$, $\widehat{f}_R^{(l)}$, $\widehat{f}_T^{(l)}$	estimates of observation, reward, and transition densities for action l
N, $N(l)$	total number of samples and number of samples from action l
C_O, C_R, C_T	numerical constants
$\mathcal{B}_O$, $\mathcal{B}_R$, $\mathcal{B}_T$	upper confidence bound over error of estimated f_O, f_R, f_T
SM-UCRL (Sect. 4)	
Reg_N	cumulative regret
D	POMDP diameter
k	index of the episode
$\widehat{f}_T^{(k)}, \widehat{f}_R^{(k)}, \widehat{f}_O^{(k)}, \widehat{M}^{(k)}$	estimated parameters of the POMDP at episode k
$\mathcal{M}^{(k)}$	set of plausible POMDPs at episode k
$v^{(k)}(l)$	number of samples from action l in episode k
$N^{(k)}(l)$	maximum number of samples from action l over all episodes before k
$\widetilde{\pi}^{(k)}$	optimistic policy executed in episode k
$\overline{N}$	min. number of samples to meet the condition in Thm. 3 for any policy and any action
$\overline{\sigma}_{\nu,\nu'}$	worst smallest non-zero singular value of covariance $K_{\nu,\nu'}^{(l)}$ for any policy and action
$\overline{\omega}_{\min}$	smallest stationary probability over actions, states, and policies

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, pages 1–26, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 - NIPS*, pages 2312–2320, 2011.

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012.

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

A. Atrash and J. Pineau. Efficient planning and tracking in pomdps with large observation spaces. In *AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems*, 2006.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

J. A. Bagnell, Sham M Kakade, Jeff G. Schneider, and Andrew Y. Ng. Policy search by dynamic programming. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 831–838. MIT Press, 2004.

Sevi Baltaoglu, Lang Tong, and Qing Zhao. Online learning and optimization of markov jump affine models. *arXiv preprint arXiv:1605.02213*, 2016.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009.

A.G. Barto, R.S. Sutton, and C.W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-13(5):834–846, Sept 1983. ISSN 0018-9472. doi: 10.1109/TSMC.1983.6313077.

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *J. Artif. Int. Res.*, 15(1):319–350, November 2001. ISSN 1076-9757.

D. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

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

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3: 213–231, 2003.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 151–159. JMLR Workshop and Conference Proceedings, 2013.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

M. Gheshlaghi-Azar, A. Lazaric, and E. Brunskill. Regret bounds for reinforcement learning with policy advice. In *Proceedings of the European Conference on Machine Learning (ECML'13)*, 2013.

M. Gheshlaghi-Azar, A. Lazaric, and E. Brunskill. Resource-efficient stochastic optimization of a locally smooth function under correlated bandit feedback. In *Proceedings of the Thirty-First International Conference on Machine Learning (ICML'14)*, 2014.

Mohammad Gheshlaghi azar, Alessandro Lazaric, and Emma Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2220–2228. Curran Associates, Inc., 2013.

Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A pac rl algorithm for episodic pomdps. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 510–518, 2016.

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

Milos Hauskrecht and Hamish Fraser. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial Intelligence in Medicine*, 18(3): 221 – 244, 2000. ISSN 0933-3657.

Daniel J Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. In *Advances in Neural Information Processing Systems*, pages 1459–1467, 2015.

Tommi Jaakkola, Satinder P. Singh, and Michael I. Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In *Advances in Neural Information Processing Systems 7*, pages 345–352. MIT Press, 1995.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010. ISSN 1532-4435.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*, pages 282–293. Springer, 2006.

Aryeh Kontorovich, Boaz Nadler, and Roi Weiss. On learning parametric-output hmms. *arXiv preprint arXiv:1302.6009*, 2013.

Aryeh Kontorovich, Roi Weiss, et al. Uniform chernoff and dvoretzky-kiefer-wolfowitz-type inequalities for markov chains and related processes. *Journal of Applied Probability*, 51 (4):1100–1113, 2014.

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

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Contextual-mdps for pac-reinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722v1*, 2016.

Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

Yanjie Li, Baoqun Yin, and Hongsheng Xi. Finding optimal memoryless policies of pomdps under the expected average reward criterion. *European Journal of Operational Research*, 211(3):556–567, 2011.

Michael L. Littman. Memoryless policies: Theoretical limitations and practical results. In *Proceedings of the Third International Conference on Simulation of Adaptive Behavior : From Animals to Animats 3: From Animals to Animats 3*, SAB94, pages 238–245, Cambridge, MA, USA, 1994. MIT Press. ISBN 0-262-53122-4.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *In Advances In Neural Information Processing Systems 14*, pages 1555–1561. MIT Press, 2001.

John Loch and Satinder P Singh. Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. In *ICML*, pages 323–331, 1998.

Omid Madani. On the computability of infinite-horizon partially observable markov decision processes. In *AAAI98 Fall Symposium on Planning with POMDPs, Orlando, FL*, 1998.

Lingsheng Meng and Bing Zheng. The optimal perturbation bounds of the moore–penrose inverse under the frobenius norm. *Linear Algebra and its Applications*, 432(4):956–963, 2010.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Andrew Y. Ng and Michael Jordan. Pegasus: A policy search method for large mdps and pomdps. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, pages 406–415, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-709-9.

P Ortner and R Auer. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, 19:49, 2007.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In Peter Auer, Alexander Clark, Thomas Zeugmann, and Sandra Zilles, editors, *Algorithmic Learning Theory*, volume 8776 of *Lecture Notes in Computer Science*, pages 140–154. Springer International Publishing, 2014. ISBN 978-3-319-11661-7.

Christos Papadimitriou and John N. Tsitsiklis. The complexity of markov decision processes. *Math. Oper. Res.*, 12(3):441–450, August 1987. ISSN 0364-765X.

Theodore J. Perkins. Reinforcement learning for POMDPs based on action values and stochastic optimization. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI 2002)*, pages 199–204. AAAI Press, 2002.

Shaowei Png, J. Pineau, and B. Chaib-draa. Building adaptive dialogue systems via bayes-adaptive pomdps. *Selected Topics in Signal Processing, IEEE Journal of*, 6(8):917–927, Dec 2012. ISSN 1932-4553. doi: 10.1109/JSTSP.2012.2229962.

P. Poupart and N. Vlassis. Model-based bayesian reinforcement learning in partially observable domains. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2008.

Pascal Poupart and Craig Boutilier. Bounded finite state controllers. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS*, pages 823–830. MIT Press, 2003.

Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive pomdps. In *Advances in neural information processing systems*, pages 1225–1232, 2007.

Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Learning without state-estimation in partially observable markovian decision processes. In *ICML*, pages 284–292. Citeseer, 1994.

E. J. Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Stanford University, 1971.

Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. *arXiv preprint arXiv:1311.3287*, 2013.

Matthijs T.J. Spaan. Partially observable markov decision processes. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning*, volume 12 of *Adaptation,*

Learning, and Optimization, pages 387–414. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-27644-6.

Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. MIT Press, 1998.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

John K. Williams and Satinder P. Singh. Experimental results on learning stochastic memoryless policies for partially observable markov decision processes. In Michael J. Kearns, Sara A. Solla, and David A. Cohn, editors, *NIPS*, pages 1073–1080. The MIT Press, 1998.

Appendix A. Organization of the Appendix

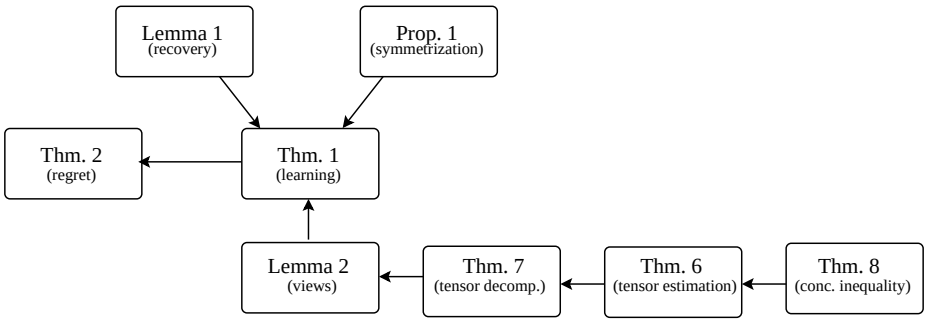


Figure 2: Organization of the proofs.

We first report the proofs of the main results of the paper in sections B, C, D, E and we postpone the technical tools used to derive them from Section I on right after preliminary empirical results in Sect. F. In particular, the main lemmas and theorems of the paper are organized as in Fig. 2.

Furthermore, we summarize the additional notation used throughout the appendices in the following table.

$\Delta_{n^{(l)}}$	Concentration matrix
$\eta_{i,j}^{(l)}(\cdot,\cdot,\cdot)$	mixing coefficient
$p(i,l)$	translator of $i'th$ element in sequence of samples given middle action l to the actual sequence number
$S^{i\|l}$	$i'th$ quadruple consequence of states random variable given second action l
$s^{i\|l}$	$i'th$ quadruple consequence of states given second action l
$S_i^{j\|l}$	sequence of all $S^{i'\|l}$ for $i' \in \{i,\dots,j\}$
$s^{i\|l}$	sequence of all $s^{i'\|l}$ for $i' \in \{i,\dots,j\}$
$B^{i\|l}$	$i'th$ triple consequence of views random variable given second action l
$b^{i\|l}$	$i'th$ triple consequence of observation given second action l
$B_i^{j\|l}$	sequence of all $B^{i'\|l}$ for $i' \in \{i,\dots,j\}$
$b^{i\|l}$	sequence of all $b^{i'\|l}$ for $i' \in \{i,\dots,j\}$

For the tensor $A \in \mathbb{R}^{d_1 \times d_2 \dots \times d_p}$, and matrices $\{V_i \in \mathbb{R}^{d_i, n_i} : i \in \{1,\dots,p\}\}$, the tensor multi-linear operator is defined as follows
For the $i_1, i_2, \dots, i_p - th$ element

$$[A(V_1, V_2, \dots, V_p)]_{i_1,i_2,\dots,i_p} \sum_{j_1,j_2,\dots,j_p \in \{1,2,\dots,p\}} A_{j_1,j_2,\dots,j_p \in \{1,2,\dots,p\}} [V_1]_{j_1,i_1} [V_2]_{j_2,i_2} \cdots [V_p]_{j_p,i_p}$$

Appendix B. Proof of Lemma 2

The proof proceeds by construction. First notice that the elements of the second view can be written as

$$\begin{aligned}[V_2^{(l)}]_{s,i} &= [V_2^{(l)}]_{(n',m'),i} \\ &= \mathbb{P}(\boldsymbol{y}_2 = \boldsymbol{e}_{n'}|x_2 = i, a_2 = l)\mathbb{P}(\boldsymbol{r}_2 = \boldsymbol{e}_{m'}|x_2 = i, a_2 = l) \\ &= \mathbb{P}(\boldsymbol{y}_2 = \boldsymbol{e}_{n'}|x_2 = i, a_2 = l) f_R(\boldsymbol{e}_{m'}|i, l),\end{aligned}$$

where we used the independence between observations and rewards. As a result, summing up over all the observations n', we can recover the reward model as

$$f_R(\boldsymbol{e}_{m'}|i, l) = \sum_{n'=1}^{Y} [V_2^{(l)}]_{(n',m'),i}, \tag{23}$$

for any combination of states $i \in [X]$ and actions $l \in [A]$. In order to compute the observation model, we have to further elaborate the definition of $V_2^{(l)}$ as

$$\begin{aligned}&[V_2^{(l)}]_{s,i} = [V_2^{(l)}]_{(n',m'),i} \\ &= \frac{\mathbb{P}(a_2 = l|x_2 = i, \boldsymbol{y}_2 = \boldsymbol{e}_{n'})\mathbb{P}(\boldsymbol{y}_2 = \boldsymbol{e}_{n'}|x_2 = i)}{\mathbb{P}(a_2 = l|x_2 = i)} \cdot \mathbb{P}(\boldsymbol{r}_2 = \boldsymbol{e}_{m'}|x_2 = i, a_2 = l) \\ &= \frac{f_\pi(l|\boldsymbol{e}_{n'}) f_O(\boldsymbol{e}_{n'}|i) f_R(\boldsymbol{e}_{m'}|i, l)}{\mathbb{P}(a_2 = l|x_2 = i)}.\end{aligned}$$

Since the policy f_π is known, if we divide the previous term by $f_\pi(l|\boldsymbol{e}_{n'})$ and sum over observations and rewards, we obtain the denominator of the previous expression as

$$\sum_{m'=1}^{R}\sum_{n'=1}^{Y} \frac{[V_2^{(l)}]_{(n',m'),i}}{f_\pi(l|\boldsymbol{e}_{n'})} = \frac{1}{\mathbb{P}(a_2 = l|x_2 = i)}.$$

Let $\rho(i, l) = 1/\mathbb{P}(a_2 = l|x_2 = i)$ as computed above, then the observation model is

$$f_O^{(l)}(\boldsymbol{e}_{n'}|i) = \sum_{m'=1}^{R} \frac{[V_2^{(l)}]_{(n',m'),i}}{f_\pi(l|\boldsymbol{e}_{n'})\rho(i, l)}. \tag{24}$$

Repeating the procedure above for each n' gives the full observation model $f_O^{(l)}$. We are left with the transition tensor, for which we need to resort to the third view $V_3^{(l)}$, that can be

written as

$$
\begin{aligned}
[V_3^{(l)}]_{s,i} &= [V_3^{(l)}]_{n'',i} \\
&= \sum_{j=1}^{X} \mathbb{P}(\boldsymbol{y}_3 = \boldsymbol{e}_{n''} | x_2 = i, a_2 = l, x_3 = j) \cdot \mathbb{P}(x_3 = j | x_2 = i, a_2 = l) \\
&= \sum_{j=1}^{X} \mathbb{P}(\boldsymbol{y}_3 = \boldsymbol{e}_{n''} | x_3 = j) \mathbb{P}(x_3 = j | x_2 = i, a_2 = l) \\
&= \sum_{j=1}^{X} f_O(\boldsymbol{e}_n | j) f_T(j | i, l), \qquad (25)
\end{aligned}
$$

where we used the graphical model of the POMDP to introduce the dependency on x_3. Since the policy f_π is known and the observation model is obtained from the second view with Eq. 6, it is possible to recover the transition model. We recall that the observation matrix $O \in \mathbb{R}^{Y \times X}$ is such that $[O]_{n,j} = f_O(\boldsymbol{e}_n | j)$, then we can restate Eq. 25 as

$$
O[T]_{i,:,l} = [V_3^{(l)}]_{:,i} \qquad (26)
$$

where $[T]_{i,:,l}$ is the second mode of the transition tensor $T \in \mathbb{R}^{X \times X \times A}$. Since all the terms in O are known, we finally obtain $[T]_{i,:,l} = O^\dagger [V_3^{(l)}]_{:,i}$, where $O^\dagger$ is the pseudo-inverse of O. Repeating for all states and actions gives the full transition model f_T.

Appendix C. Proof of Thm. 3

The proof builds upon previous results on HMM by Anandkumar et al. (2012), Song et al. (2013), Thm. 10, Appendix I, . All the following statements hold under the assumption that the samples are drawn from the stationary distribution induced by the policy π on the POMDP (i.e., $f_{T,\pi}$). In proving Thm. 4, we will consider the additional error coming from the fact that samples are not necessarily drawn from $f_{T,\pi}$.

We denote by $\sigma_1(A) \geq \sigma_2(A) \geq \dots$ the singular values of a matrix A and we recall that the covariance matrices $K^{(l)}_{\nu,\nu'}$ have rank X under Asm. 2 and we denote by $\sigma^{(l)}_{\nu,\nu'} = \sigma_X(K^{(l)}_{\nu,\nu'})$ its smallest non-zero singular value, where $\nu, \nu' \in \{1,2,3\}$. Adapting the result by Song et al. (2013), we have the following performance guarantee when the spectral method is applied to recover each column of the third view.

Lemma 5 *Let $\widehat{\mu}^{(l)}_{3,i} \in \mathbb{R}^{d_3}_3$ and $\widehat{\omega}^{(l)}_\pi(i)$ be the estimated third view and the conditional distribution computed in state $i \in \mathcal{X}$ using the spectral method in Sect. 3 using $N(l)$ samples. Let $\omega^{(l)}_{\min} = \min_{x \in \mathcal{X}} \omega^{(l)}_\pi(x)$ and the number of samples $N(l)$ is such that*

$$N(l) > \left(\frac{G(\pi)\frac{2\sqrt{2}+1}{1-\theta(\pi)}}{\omega_{\min}^{(l)} \min_{\nu\in\{1,2,3\}} \{\sigma_{\min}^2(V_\nu^{(l)})\}} \right)^2 \log(2\frac{(d_1 d_2 + d_3)}{\delta})\Theta^{(l)} \tag{27}$$

$$\Theta^{(l)} = \max\left\{ \frac{16X^{\frac{1}{3}}}{C_1^{\frac{2}{3}}(\omega_{\min}^{(l)})^{\frac{1}{3}}}, 4, \frac{2\sqrt{2}X}{C_1^2\omega_{\min}^{(l)} \min_{\nu\in\{1,2,3\}} \{\sigma_{\min}^2(V_\nu^{(l)})\}} \right\}, \tag{28}$$

where C_1 is numerical constants and d_1, d_2 are dimensions of first and second views. Then under Thm. 16 for any $\delta \in (0,1)$ we have [7]

$$\left\| [\widehat{V}_3^{(l)}]_{:,i} - [V_3^{(l)}]_{:,i} \right\|_2 \leq \epsilon_3$$

with probability $1-\delta$ (w.r.t. the randomness in the transitions, observations, and policy), where[8]

$$\epsilon_3(l) := G(\pi)\frac{4\sqrt{2}+2}{(\omega_{\min}^{(l)})^{\frac{1}{2}}(1-\theta(\pi))}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}} + \frac{8\widetilde{\epsilon}_M}{\omega_{\min}^{(l)}} \tag{29}$$

and

$$\widetilde{\epsilon}_M(l) \leq \frac{2\sqrt{2}G(\pi)\frac{2\sqrt{2}+1}{1-\theta(\pi)}\sqrt{\frac{\log(\frac{2(d_1 d_2+d_3)}{\delta})}{N(l)}}}{((\omega_{\min}^{(l)})^{\frac{1}{2}} \min_{\nu\in\{1,2,3\}} \{\sigma_{\min}(V_\nu^{(l)})\})^3} + \frac{\left(64G(\pi)\frac{2\sqrt{2}+1}{1-\theta(\pi)}\right)}{\min_{\nu\in\{1,2,3\}} \{\sigma_{\min}^2(V_\nu^{(l)})\}(\omega_{\min}^{(l)})^{1.5}}\sqrt{\frac{\log(2\frac{(d_1 d_2+d_3)}{\delta})}{N(l)}},$$

Notice that although not explicit in the notation, $\epsilon_3(l)$ depends on the policy π through the term $\omega_{\min}^{(l)}$.

Proof We now proceed with simplifying the expression of $\epsilon_3(l)$. Rewriting the condition on $N(l)$ in Eq. 27 we obtain

$$\frac{\log(2\frac{(d_1 d_2+d_3)}{\delta})}{N(l)} \leq \left(\frac{\omega_{\min}^{(l)} \min_{\nu\in\{1,2,3\}} \{\sigma_{\min}^2(V_\nu^{(l)})\}}{G(\pi)\frac{2\sqrt{2}+1}{1-\theta(\pi)}} \right)^2$$

7. More precisely, the statement should be phrased as "there exists a suitable permutation on the label of the states such that". This is due to the fact that the spectral method cannot recover the exact *identity* of the states but if we properly relabel them, then the estimates are accurate. In here we do not make explicit the permutation in order to simplify the notation and readability of the results.
8. Notice that $\epsilon_3(l)$ does not depend on the specific state (column) i.

Substituting this bound on a factor $\log(2\frac{Y^2+YAR}{\delta})/N(l)$ in the second term of Eq. 29, we obtain

$$\widetilde{\epsilon}_M(l) \leq \frac{2\sqrt{2}G(\pi)\frac{2\sqrt{2}+1}{1-\theta(\pi)}\sqrt{\frac{\log(\frac{2(d_1d_2+d_3)}{\delta})}{N(l)}}}{((\omega_{\min}^{(l)})^{\frac{1}{2}}\min_{\nu\in\{1,2,3\}}\{\sigma_{\min}(V_\nu^{(l)})\})^3} + \frac{\left(64G(\pi)\frac{2\sqrt{2}+1}{1-\theta(\pi)}\right)}{(\min_{\nu\in\{1,2,3\}}\{\sigma^2_{\min}(V_\nu^{(l)})\})(\omega_{\min}^{(l)})^{1.5}}\sqrt{\frac{\log(2\frac{(d_1d_2+d_3)}{\delta})}{N(l)}},$$

which leads to the final statement after a few trivial bounds on the remaining terms. ■

While the previous bound does hold for both the first and second views when computed independently with a suitable symmetrization step, as discussed in Section 3, this leads to inconsistent state indexes. As a result, we have to compute the other views by inverting Eq. 4. Before deriving the bound on the accuracy of the corresponding estimates, we introduce two propositions which will be useful later.

Proposition 6 *Fix $\boldsymbol{\varsigma} = (\varsigma_1, \varsigma_2, \ldots, \varsigma_{(Y^2RA)})$ a point in $(Y^2)RA-1$ simplex.*[9] *Let $\boldsymbol{\xi}$ be a random one-hot vector such that $\mathbb{P}(\boldsymbol{\xi} = \boldsymbol{e}_i) = \varsigma_i$ for all $i \in \{1, \ldots, (Y)^2RA\}$ and let $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N$ be N i.i.d. copies of $\boldsymbol{\xi}$ and $\hat{\varsigma} = \frac{1}{N}\sum_j^N \boldsymbol{\xi}_j$ be their empirical average, then*

$$\|\hat{\varsigma} - \boldsymbol{\varsigma}\|_2 \leq \sqrt{\frac{\log(1/\delta)}{N}},$$

with probability $1-\delta$.

Proof See Lemma F.1. in Anandkumar et al. (2012). ■

Proposition 7 *Let $\widehat{K}_{3,1}^{(l)}$ be an empirical estimate of $K_{3,1}^{(l)}$ obtained using $N(l)$ samples. Then if*

$$N(l) \geq 4\frac{\log(1/\delta)}{(\sigma_{3,1}^{(l)})^2}, \tag{30}$$

then

$$\|(K_{3,1}^{(l)})^\dagger - (\widehat{K}_{3,1}^{(l)})^\dagger\|_2 \leq \frac{\sqrt{\frac{\log(1/\delta)}{N(l)}}}{\sigma_{3,1}^{(l)} - \sqrt{\frac{\log(1/\delta)}{N(l)}}} \leq \frac{2}{\sigma_{3,1}^{(l)}}\sqrt{\frac{\log(\frac{1}{\delta})}{N(l)}},$$

with probability $1-\delta$.

Proof Since $K_{3,1}^{(l)} = \mathbb{E}[\boldsymbol{v}_3^{(l)} \otimes \boldsymbol{v}_1^{(l)}]$ and the views are one-hot vectors, we have that each entry of the matrix is indeed a probability (i.e., a number between 0 and 1) and the sum

9. Such that $\forall i = 1, \ldots, d^2$, $\varsigma_i > 0$ and $\sum_i \varsigma_i = 1$.

of all the elements in the matrix sums up to 1. As a result, we can apply Proposition 6 to $K_{3,1}^{(l)}$ and obtain

$$\|K_{3,1}^{(l)} - \widehat{K}_{3,1}^{(l)}\|_2 \leq \sqrt{\frac{\log(1/\delta)}{N(l)}}, \tag{31}$$

with probability $1-\delta$. Then the statement follows by applying Lemma E.4. in Anandkumar et al. (2012). ■

The previous proposition holds for $K_{2,1}^{(l)}$ as well with $\sigma_{2,1}^{(l)}$ replacing $\sigma_{3,1}^{(l)}$. We are now ready to state and prove the accuracy of the estimate of the second view (a similar bound holds for the first view).

Lemma 8 *Let $\widehat{V}_2^{(l)}$ be the second view estimated inverting Eq. 4 using estimated covariance matrices K and $V_3^{(l)}$, then if $N(l)$ satisfies the conditions in Eq. 27 and Eq. 30 with probability $1-3\delta$*

$$\left\|[\widehat{V}_2^{(l)}]_{:,i} - [V_2^{(l)}]_{:,i}\right\|_2 = \epsilon_2(l) := \frac{21}{\sigma_{3,1}^{(l)}}\epsilon_3(l).$$

Proof For any state $i \in \mathcal{X}$ and action $l \in A$, we obtain the second view by inverting Eq. 4, that is by computing

$$[V_2^{(l)}]_{:,i} = K_{2,1}^{(l)}(K_{3,1}^{(l)})^\dagger [V_3^{(l)}]_{:,i}.$$

To derive a confidence bound on the empirical version of $\mu_{2,i}^{(l)}$, we proceed by first upper bounding the error as

$$\begin{aligned}\left\|[\widehat{V}_2^{(l)}]_{:,i} - [V_2^{(l)}]_{:,i}\right\|_2 \leq & \|K_{2,1}^{(l)} - \widehat{K}_{2,1}^{(l)}\|_2 \|(K_{3,1}^{(l)})^\dagger\|_2 \|[V_3^{(l)}]_{:,i}\|_2 \\ & + \|K_{2,1}^{(l)}\|_2 \|(K_{3,1}^{(l)})^\dagger - (\widehat{K}_{3,1}^{(l)})^\dagger\|_2 \|[V_3^{(l)}]_{:,i}\|_2 \\ & + \|K_{2,1}^{(l)}\|_2 \|(K_{3,1}^{(l)})^\dagger\|_2 \left\|[\widehat{V}_3^{(l)}]_{:,i} - [V_3^{(l)}]_{:,i}\right\|_2.\end{aligned}$$

The error $\|K_{2,1}^{(l)} - \widehat{K}_{2,1}^{(l)}\|_2$ can be bounded by a direct application of Proposition 6 (see also Eq. 31). Then we can directly use Proposition 7 to bound the second term and Lemma 5 for the third term, and obtain

$$\left\|[\widehat{V}_2^{(l)}]_{:,i} - [V_2^{(l)}]_{:,i}\right\|_2 \leq \frac{3}{\sigma_{3,1}^{(l)}}\sqrt{\frac{\log(\frac{1}{\delta})}{N(l)}} + \frac{18\epsilon_3(l)}{\sigma_{3,1}^{(l)}} \leq \frac{21\epsilon_3(l)}{\sigma_{3,1}^{(l)}},$$

where we used $\|(K_{3,1}^{(l)})^\dagger\|_2 \leq 1/\sigma_{3,1}^{(l)}$, $\|K_{2,1}^{(l)}\|_2 \leq 1$ and $\|[V_3^{(l)}]_{:,i}\|_2 \leq 1$. Since each of the bounds we used hold with probability $1-\delta$, the final statement is valid with probability at least $1-3\delta$. ■

We are now ready to derive the bounds in Thm. 3.

Proof [Proof of Thm. 3] We first recall that the estimates $\widehat{f}_R$, $\widehat{f}_O$, and $\widehat{f}_T$ are obtained by working on the second and third views only, as illustrated in Sect. 3.

Step 1 (bound on f_R). Using the empirical version of Eq. 5, the reward model in state i for action l is computed as

$$\widehat{f}_R(\boldsymbol{e}_{m'}|i,l) = \sum_{n'=1}^{Y} [\widehat{V}_2^{(l)}]_{(n',m'),i}.$$

Then the ℓ_1-norm of the error can be bounded as

$$\begin{aligned}
\|\widehat{f}_R(.|i,l) - f_R(.|i,l)\|_1 &= \sum_{m'=1}^{R} |\widehat{f}_R(\boldsymbol{e}_{m'}|i,l) - f_R(\boldsymbol{e}_{m'}|i,l)| \\
&\le \sum_{m'=1}^{R} \Big| \sum_{n'=1}^{Y} [\widehat{V}_2^{(l)}]_{(n',m'),i} - \sum_{n'=1}^{Y} [V_2^{(l)}]_{(n',m'),i} \Big| \\
&\le \sum_{m'=1}^{R} \sum_{n'=1}^{Y} \Big| [\widehat{V}_2^{(l)}]_{(n',m'),i} - [V_2^{(l)}]_{(n',m'),i} \Big| \\
&\le \sqrt{YR} \Big(\sum_{m'=1}^{R} \sum_{n'=1}^{Y} \Big([\widehat{V}_2^{(l)}]_{(n',m'),i} - [V_2^{(l)}]_{(n',m'),i} \Big)^2 \Big)^{1/2} \\
&= \sqrt{YR} \big\| [\widehat{V}_2^{(l)}]_{:,i} - [V_2^{(l)}]_{:,i} \big\|_2,
\end{aligned}$$

where we use $\|v\|_1 \le \sqrt{YR}\|v\|_2$ for any vector $v \in \mathbb{R}^{Y \cdot R}$. Applying Lemma 8 we obtain

$$\|\widehat{f}_R(.|i,l) - f_R(.|i,l)\|_1 \le \mathcal{B}_R := \frac{C_R}{\sigma_{3,1}^{(l)} (\omega_{\min}^{(l)})^{\frac{3}{2}} \min_{\nu \in \{1,2,3\}} \{\sigma_{\min}^3(V_\nu^{(l)})\}} \sqrt{\frac{YR \log(2\frac{Y^2 + YAR}{\delta})}{N(l)}},$$

where C_R is a numerical constant.

Step 2 (bound on $\rho(i,l)$). We proceed by bounding the error of the estimate the term $\rho(i,l) = 1/\mathbb{P}(a_2 = l | x_2 = i)$ which is computed as

$$\widehat{\rho}(i,l) = \sum_{m'=1}^{R} \sum_{n'=1}^{Y} \frac{[\widehat{V}_2^{(l)}]_{(n',m'),i}}{f_\pi(l|\boldsymbol{e}_{n'})},$$

and it is used to estimate the observation model. Similarly to the bound for f_R we have

$$\begin{aligned}
|\rho(i,l) - \widehat{\rho}(i,l)| &\le \sum_{m'=1}^{R} \sum_{n'=1}^{Y} \frac{|[V_2^{(l)}]_{(n',m'),i} - [\widehat{V}_2^{(l)}]_{(n',m'),i}|}{f_\pi(l|\boldsymbol{e}_{n'})} \le \frac{1}{\pi_{\min}^{(l)}} \big\| [V_2^{(l)}]_{:,i} - [\widehat{V}_2^{(l)}]_{:,i} \big\|_1 \\
&\le \frac{\sqrt{YR}}{\pi_{\min}^{(l)}} \big\| [V_2^{(l)}]_{:,i} - [\widehat{V}_2^{(l)}]_{:,i} \big\|_2 \le 21 \frac{\sqrt{YR}}{\sigma_{3,1}^{(l)} \pi_{\min}^{(l)}} \epsilon_3(i) =: \epsilon_\rho(i,l),
\end{aligned} \tag{32}$$

where $\pi_{\min}^{(l)} = \min_{\boldsymbol{y}\in\mathcal{Y}} f_\pi(l|\boldsymbol{y})$ is the smallest non-zero probability of taking an action according to policy π.

Step 3 (bound on f_O). The observation model in state i for action l can be recovered by plugging the estimates into Eq. 5 and obtain

$$\widehat{f}_O^{(l)}(\boldsymbol{e}_{n'}|i) = \sum_{m'=1}^{R} \frac{[\widehat{V}_2^{(l)}]_{(n',m'),i}}{f_\pi(l|\boldsymbol{e}_{n'})\widehat{\rho}(i,l)},$$

where the dependency on l is due do the fact that we use the view computed for action l. As a result, the ℓ_1-norm of the estimation error is bounded as follows

$$\begin{aligned}
\sum_{n'=1}^{Y} |\widehat{f}_O^{(l)}(\boldsymbol{e}_{n'}|i) - f_O(\boldsymbol{e}_{n'}|i)| &\le \sum_{n'=1}^{Y}\sum_{m'=1}^{R} \left| \frac{1}{f_\pi(l|\boldsymbol{e}_{n'})} \left(\frac{[\widehat{V}_2^{(l)}]_{(n',m'),i}}{\widehat{\rho}(i,l)} - \frac{[V_2^{(l)}]_{(n',m'),i}}{\rho(i,l)} \right) \right| \\
&\le \frac{1}{\pi_{\min}^{(l)}} \sum_{n'=1}^{Y}\sum_{m'=1}^{R} \left| \frac{\rho(i,l)\big([\widehat{V}_2^{(l)}]_{(n',m'),i} - [V_2^{(l)}]_{(n',m'),i}\big) + [V_2^{(l)}]_{(n',m'),i}\big(\rho(i,l) - \widehat{\rho}(i,l)\big)}{\widehat{\rho}(i,l)\rho(i,l)} \right| \\
&\le \frac{1}{\pi_{\min}^{(l)}} \left(\sum_{n'=1}^{Y}\sum_{m'=1}^{R} \frac{\big|[\widehat{V}_2^{(l)}]_{(n',m'),i} - [V_2^{(l)}]_{(n',m'),i}\big|}{\widehat{\rho}(i,l)} + \frac{\big|\rho(i,l) - \widehat{\rho}(i,l)\big|}{\widehat{\rho}(i,l)\rho(i,l)} \Big(\sum_{n'=1}^{Y}\sum_{m'=1}^{R} [V_2^{(l)}]_{(n',m'),i} \Big) \right) \\
&\overset{(a)}{\le} \frac{1}{\pi_{\min}^{(l)}} \left(\frac{\sqrt{YR}}{\widehat{\rho}(i,l)} \big\| [\widehat{V}_2^{(l)}]_{:,i} - [V_2^{(l)}]_{:,i} \big\|_2 + \frac{\big|\rho(i,l) - \widehat{\rho}(i,l)\big|}{\widehat{\rho}(i,l)\rho(i,l)} \Big(\sum_{m'=1}^{R} [V_2^{(l)}]_{(n',m'),i} \Big) \right) \\
&\overset{(b)}{\le} \frac{1}{\pi_{\min}^{(l)}} \left(\frac{\sqrt{YR}}{\widehat{\rho}(i,l)} \epsilon_2(i) + \frac{\epsilon_\rho(i,l)}{\widehat{\rho}(i,l)\rho(i,l)} \right) \\
&\overset{(c)}{\le} \frac{1}{\pi_{\min}^{(l)}} \left(21\sqrt{YR} \frac{\epsilon_3(i)}{\sigma_{3,1}^{(l)}} + \epsilon_\rho(i,l) \right),
\end{aligned}$$

where in (a) we used the fact that we are only summing over R elements (instead of the whole YR dimensionality of the vector $[V_2^{(l)}]_{:,i}$), in (b) we use Lemmas 5, 8, and in (c) the fact that $1/\rho(i,l) = \mathbb{P}[a_2 = l|x_2 = i] \le 1$ (similar for $1/\widehat{\rho}(i,l)$). Recalling the definition of $\epsilon_\rho(i,l)$ and Lemma 5 and Lemma 8 we obtain

$$\begin{aligned}
\|\widehat{f}_O^{(l)}(\cdot|i) - f_O(\cdot|i)\|_1 &\le \frac{62}{(\pi_{\min}^{(l)})^2}\sqrt{YR}\epsilon_3(l) \\
&\le \mathcal{B}_O^{(l)} := \frac{C_O}{(\pi_{\min}^{(l)})^2 \sigma_{1,3}^{(l)} (\omega_{\min}^{(l)})^{\frac{3}{2}} \min_{\nu\in\{1,2,3\}} \{\sigma_{\min}^3(V_\nu^{(l)})\}} \sqrt{\frac{YR\log(2\frac{Y^2+YAR}{\delta})}{N(l)}},
\end{aligned}$$

where C_O is a numerical constant. As mentioned in Sect. 3, since we obtain one estimate per action, in the end we define $\widehat{f}_O$ as the estimate with the smallest confidence interval,

that is

$$\widehat{f}_O = \widehat{f}_O^{(l^*)}, \quad l^* = \arg\min_{\{\widehat{f}_O^{(l)}\}} \mathcal{B}_O^{(l)},$$

whose corresponding error bound is

$$\|\widehat{f}_O(\boldsymbol{e}_{n'}|i) - f_O(\boldsymbol{e}_n|i)\| \le \mathcal{B}_O := \min_{l=1,\dots,A} \frac{C_O}{(\pi_{\min}^{(l)})^2 \sigma_{1,3}^{(l)} (\omega_{\min}^{(l)})^{\frac{3}{2}} \min_{\nu\in\{1,2,3\}} \{\sigma^3_{\min}(V_\nu^{(l)})\}} \sqrt{\frac{YR\log(2\frac{Y^2+YAR}{\delta})}{N(l)}}.$$

The columns of estimated $O^{(l)}$ matrices are up to different permutations over states, i.e. these matrices have different columns ordering. Let's assume that the number of samples for each action is such a way that satisfies $\mathcal{B}_O^{(l)} \le \frac{d_O}{4}$, $\forall l \in [A]$. Then, one can exactly match each matrix $O^{(l)}$ with $O^{(l^*)}$ and then propagate these orders to matrices $V_2^{(l)}$ and $V_3^{(l)}$, $\forall l \in [A]$. The condition $\mathcal{B}_O^{(l)} \le \frac{d_O}{4}$, $\forall l \in [A]$ can be represented as follow

$$N(l) \ge \frac{16 C_O^2 YR}{{\lambda^{(l)}}^2 d_O^2}, \quad \forall l \in [A]$$

Step 4 (bound on f_T). The derivation of the bound for $\widehat{f}_T$ is more complex since each distribution $\widehat{f}_T(\cdot|x,a)$ is obtained as the solution of the linear system of equations in Eq. 26, that is for any state i and action l we compute

$$[\widehat{T}]_{i,:,l} = \widehat{O}^\dagger [\widehat{V}_3^{(l)}]_{:,i}, \tag{33}$$

where $\widehat{O}$ is obtained plugging in the estimate $\widehat{f}_O$.[10] We first recall the following general result for the pseudo-inverse of a matrix and we instance it in our case. Let W and $\widehat{W}$ be any pair of matrix such that $\widehat{W} = W + E$ for a suitable error matrix E, then we have Meng and Zheng (2010)

$$\|W^\dagger - \widehat{W}^\dagger\|_2 \le \frac{1+\sqrt{5}}{2} \max\left\{\|W^\dagger\|_2, \|\widehat{W}^\dagger\|_2\right\} \|E\|_2, \tag{34}$$

where $\|\cdot\|_2$ is the spectral norm. Since Lemma 8 provides a bound on the error for each column of $V_2^{(l)}$ for each action and a bound on the error of $\rho(i,l)$ is already developed in Step 2, we can bound the ℓ_2 norm of the estimation error for each column of O and $\widehat{O}$ as

$$\|\widehat{O} - O\|_2 \le \|\widehat{O} - O\|_F \le \sqrt{X} \min_{l\in[A]} \mathcal{B}_O^{(l)}. \tag{35}$$

We now focus on the maximum in Eq. 34, for which we need to bound the spectral norm of the pseudo-inverse of the estimated W. We have $\|\widehat{O}^\dagger\|_2 \le (\sigma_X(\widehat{O}))^{-1}$ where $\sigma_X(\widehat{O})$ is the X-th singular value of matrix $\widehat{O}$ whose perturbation is bounded by $\|\widehat{O} - O\|_2$. Since matrix

10. We recall that $\widehat{f}_O$ corresponds to the estimate $\widehat{f}_O^{(l)}$ with the tightest bound $B_O^{(l)}$.

O has rank X from Asm. 2 then

$$\|\widehat{O}^\dagger\|_2 \leq (\sigma_X(\widehat{O}))^{-1} \leq \frac{1}{\sigma_X(O)}\left(1+\frac{\|\widehat{O}-O\|_2}{\sigma_X(O)}\right) \leq \frac{1}{\sigma_X(O)}\left(1+\frac{\|\widehat{O}-O\|_F}{\sigma_X(O)}\right).$$

We are now ready to bound the estimation error of the transition tensor. From the definition of Eq. 33 we have that for any state $i=1,\dots,X$ the error is bounded as

$$\|T_{i,:,l}-\widehat{T}_{i,:,l}\|_2 \leq \|T_{:,:,l}-\widehat{T}_{:,:,l}\|_2 \leq \|\widehat{O}^\dagger - O^\dagger\|_2 \|V_3^{(l)}\|_2 + \|\widehat{V}_3^{(l)} - V_3^{(l)}\|_2 \|\widehat{O}^\dagger\|_2.$$

In Lemma 5 we have a bound on the ℓ_2-norm of the error for each column of $V_3^{(l)}$, thus we have $\|\widehat{V}_3^{(l)} - V_3^{(l)}\|_2 \leq \|\widehat{V}_3^{(l)} - V_3^{(l)}\|_F \leq 18\sqrt{X}\epsilon_3(l)$. Using the bound on Eq. 34 and denoting $\|V_3^{(l)}\|_2 = \sigma_{\max}(V_3^{(l)})$ we obtain

$$\begin{aligned}
&\|T_{i,:,l}-\widehat{T}_{i,:,l}\|_2 \\
&\quad\leq \frac{1+\sqrt{5}}{2}\frac{\|\widehat{O}-O\|_F}{\sigma_X(O)}\left(1+\frac{\|\widehat{O}-O\|_F}{\sigma_X(O)}\right)\sigma_{\max}(V_3^{(l)}) + 18\sqrt{X}\epsilon_3(l)\frac{1}{\sigma_X(O)}\left(1+\frac{\|\widehat{O}-O\|_F}{\sigma_X(O)}\right) \\
&\quad\leq \frac{2}{\sigma_X(O)}\left(1+\frac{\|\widehat{O}-O\|_F}{\sigma_X(O)}\right)\left(\sigma_{\max}(V_3^{(l)})\|\widehat{O}-O\|_F + 18\sqrt{X}\epsilon_3(l)\right).
\end{aligned}$$

Finally, using the bound in Eq. 35 and bounding $\sigma_{\max}(V_3^{(l)}) \leq \sqrt{X}$,[11]

$$\begin{aligned}
\|T_{i,:,l}-\widehat{T}_{i,:,l}\|_2 &\leq \frac{4}{\sigma_X(O)}\left(X\min_{l\in[A]}\mathcal{B}_O^{(l)} + 18\sqrt{X}\epsilon_3(l)\right) \\
&\leq \frac{C_T}{\sigma_X(O)(\pi_{\min}^{(l)})^2\sigma_{1,3}^{(l)}(\omega_{\min}^{(l)})^{\frac{3}{2}}\min_{\nu\in\{1,2,3\}}\{\sigma_{\min}^3(V_\nu^{(l)})\}}\sqrt{\frac{X^2YR\log(8/\delta)}{N(l)}},
\end{aligned}$$

thus leading to the final statement. Since we require all these bounds to hold simultaneously for all actions, the probability of the final statement is $1-3A\delta$. Notice that for the sake of readability in the final expression reported in the theorem we use the denominator of the error of the transition model to bound all the errors and we report the statement with probability $1-24A\delta$ are change the logarithmic term in the bounds accordingly.

■

Appendix D. Proof of Theorem 4

Proof [Proof of Theorem 4] While the proof is similar to UCRL Jaksch et al. (2010), each step has to be carefully adapted to the specific case of POMDPs and the estimated models obtained from the spectral method.

11. This is obtained by $\|V_3^{(l)}\|_2 \leq \sqrt{X}\|V_3^{(l)}\|_1 = \sqrt{X}$, since the sum of each column of $V_3^{(l)}$ is one.

Step 1 (regret decomposition). We first rewrite the regret making it explicit the regret accumulated over episodes, where we remove the burn-in phase

$$\begin{aligned}\mathrm{Reg}_N &\leq \sum_{k=1}^{K}\Bigg(\sum_{t=t^{(k)}}^{t^{(k+1)}-1}\Big(\eta^+ - r_t(x_t,\widetilde{\pi}_k(\boldsymbol{y}_t))\Big)\Bigg)\\ &= \sum_{k=1}^{K}\sum_{t=t^{(k)}}^{t^{(k+1)}-1}\Big(\eta^+ - r_t(x_t,\widetilde{\pi}_k(\boldsymbol{y}_t))\Big),\end{aligned}$$

where $r_t(x_t,\widetilde{\pi}_k(\boldsymbol{y}_t))$ is the random reward observed when taking the action prescribed by the optimistic policy $\widetilde{\pi}_k$ depending on the observation triggered by state x_t. We introduce the time steps $\mathcal{T}^{(k)} = \big\{t : t^{(k)} \leq t < t^{(k+1)}\big\}$, $\mathcal{T}^{(k)}(l) = \big\{t \in \mathcal{T}^{(k)} : l_t = l\big\}$, $\mathcal{T}^{(k)}(x,l) = \big\{t \in \mathcal{T}^{(k)} : x_t = x, a_t = l\big\}$ and the counters $v^{(k)} = |\mathcal{T}^{(k)}|$, $v^{(k)}(l) = |\mathcal{T}^{(k)}(l)|$, $v^{(k)}(x,l) = |\mathcal{T}^{(k)}(x,l)|$, while we recall that $N^{(k)}(l)$ denotes the number of samples of action l available at the beginning of episodes k used to compute the optimistic policy $\widetilde{\pi}_k$. We first remove the randomness in the observed reward by Hoeffding's inequality as

$$\mathbb{P}\left[\sum_{t=t^{(k)}}^{t^{(k)}+v^{(k)}-1} r_t(x_t,\widetilde{\pi}_k(\boldsymbol{y}_t)) \leq \sum_{x,l} v^{(k)}(x,l)\bar{r}(x,l) - r_{\max}\sqrt{\frac{v^{(k)}\log\frac{1}{\delta}}{2}} \;\Big|\; \{N^{(k)}(l)\}_l\right] \leq \delta,$$

where the probability is taken w.r.t. the reward model $f_R(\cdot|x,a)$ and observation model $f_O(\cdot|x)$, $\overline{r}(x,l)$ is the expected reward for the state-action pair x,l. Recalling the definition of the optimistic POMDP $\widetilde{M}^{(k)} = \arg\max_{M\in\mathcal{M}^{(k)}}\max_{\pi\in\mathcal{P}}\eta(\pi;M)$, we have that $\eta^+ \leq \eta(\widetilde{M}^{(k)};\widetilde{\pi}^{(k)}) = \widetilde{\eta}^{(k)}$, then applying the previous bound in the regret definition we obtain

$$\mathrm{Reg}_N \leq \underbrace{\sum_{k=1}^{K}\sum_{x=1}^{X}\sum_{l=1}^{A} v^{(k)}(x,l)\Big(\widetilde{\eta}^{(k)} - \bar{r}(x,l)\Big)}_{\Delta^{(k)}} + r_{\max}\sqrt{N\log 1/\delta},$$

with high probability, where the last term follows from Jensen's inequality and the fact that $\sum_k v^{(k)} = N$.

Step 2 (condition on $N(l)$). As reported in Thm. 3, the confidence intervals are valid only if for each action $l = 1,\dots,A$ enough samples are available. As a result, we need to compute after how many episodes the condition in Eq. 9 is satisfied (with high probability). We first roughly simplify the condition by introducing $\overline{\omega}^{(l)}_{\min} = \min_{\pi\in\mathcal{P}}\min_{x\in\mathcal{X}}\omega^{(l)}_{\pi}(x)$ and

$$\overline{N} := \max_{l\in[A]}\max\left\{\frac{4}{(\sigma^{(l)}_{3,1})^2}, \frac{16C_O^2 YR}{{\lambda^{(l)}}^2 d_O^2}, \frac{C_2^2}{(\overline{\omega}^{(l)}_{\min})^2\min\limits_{\nu\in\{1,2,3\}}\{\sigma^4_{\min}(V^{(l)}_{\nu})\}}\overline{\Theta}^{(l)}\right\}\log(2\frac{Y^2+YAR}{\delta}).$$

$$\overline{\Theta}^{(l)} = \max\left\{ \frac{16X^{\frac{1}{3}}}{C_1^{\frac{2}{3}}(\overline{\omega}_{\min}^{(l)})^{\frac{1}{3}}}, 4, \frac{2\sqrt{2}X}{C_1^2\omega_{\min}^{(l)} \min_{\nu\in\{1,2,3\}}\{\sigma_{\min}^2(V_\nu^{(l)})\}} \right\}, \tag{36}$$

We recall that at the beginning of each episode k, the POMDP is estimated using $N^{(k)}(l)$ which is the largest number of samples collected for action l in any episode prior to k, i.e., $N^{(k)}(l) = \max_{k'<k} v^{(k')}(l)$. Thus we first study how many samples are likely to be collected for any action l in any episode of length v. Let $\tau_{M,\pi}^{(l)}$ is the mean passage time between two steps where action l is chosen according to policy $\pi \in \mathcal{P}$ then we define $\tau_M^{(l)} = \max_{\pi\in\mathcal{P}} \tau_{M,\pi}^{(l)} = \max_{\pi\in\mathcal{P}} \mathbb{E}[\mathcal{T}(l,l)]$, where $\mathcal{T}(l,l)$ is random variable and represent the passing time between two steps where action l is chosen according to policy $\pi \in \mathcal{P}$. By Markov inequality, the probability that it takes more than $2\tau_M^{(l)}$ to take the same action l is at most $1/2$. If we divide the episode of length v into $v/2\tau_M^{(l)}$ intervals of length $2\tau_M^{(l)}$, we have that within each interval we have a probability of $1/2$ to observe a sample from action l, and thus on average we can have a total of $v/4\tau_M^{(l)}$ samples. Thus from Chernoff-Hoeffding, we obtain that the number of samples of action l is such that

$$\mathbb{P}\left\{ \exists l \in [A] :\ v(l) \geq \frac{v}{4\tau_M^{(l)}} - \sqrt{\frac{v\log(A/\delta)}{2\tau_M^{(l)}}} \right\} \geq 1-\delta.$$

At this point we can derive a lower bound on the length of the episode that guarantee that the desired number of samples is collected. We solve

$$\frac{v}{4\tau_M^{(l)}} - \sqrt{\frac{v\log(A/\delta)}{2\tau_M^{(l)}}} \geq \overline{N},$$

and we obtain the condition

$$\sqrt{v} \geq \sqrt{2\tau_M^{(l)}\log(A/\delta)} + \sqrt{2\tau_M^{(l)}\log(A/\delta) + 16\tau_M^{(l)}\overline{N}},$$

which can be simplified to

$$v \geq \overline{v} := 24\tau_M^{(l)}\overline{N}\log(A/\delta). \tag{37}$$

Thus we need to find a suitable number of episodes $\widetilde{K}$ such that there exists an episode $k' < \widetilde{K}$ such that $v^{(k')}$ satisfies the condition in Eq. 37. Since an episode is terminated when an action l ($v^{(k)}(l)$) is selected twice the number of samples available at the beginning of the episode ($N^{(k)}(l)$), we have that at episode k there was an episode in the past ($k' < k$) with at least 2^c steps with $c = \max\{n \in \mathbb{N} : An \leq k\}$, where A is the number of actions (i.e., after Ac episodes there was at least one episode in which an action reached 2^c samples, which forced the episode to be at least that long). From condition in Eq. 37, we need $2^c \geq \overline{v}$, which in turn gives $\widetilde{K} \geq A\log_2(\overline{v})$, which finally implies that $\widetilde{K} \leq A\log_2(\overline{v}) + 1$ is a sufficient condition on the number of episodes needed to guarantee that all the actions have been selected enough so that the condition of Thm. 3 is satisfied. We are just left with

measuring the regret accumulated over the first $\widetilde{K}$ episodes, that is

$$\sum_{k=1}^{\widetilde{K}+1}\sum_{t=t^{(k)}}^{t^{(k+1)}-1}\Big(\eta^+ - r_t(x_t,\widetilde{\pi}_k(\boldsymbol{y}_t))\Big) \le r_{\max}\sum_{k=1}^{\widetilde{K}+1} v^{(k)} \le r_{\max}\sum_{k=1}^{\widetilde{K}+1} A2^k \le 4r_{\max}A2^{\widetilde{K}} \le 4Ar_{\max}(\overline{v}+1), \tag{38}$$

where in the first step we maximize the per-step regret by $r_{\max}$ and then we use a rough upper bound for the length of each episode (as if the length is doubled at each episode) and finally we use the upper-bound on $\widetilde{K}$.

Step 3 (failing confidence intervals). Even after the first $\widetilde{K}$ episodes, the confidence intervals used to construct the set of POMDPs $\mathcal{M}^{(k)}$ may not be correct, which implies that the true POMDP M is not contained in $\mathcal{M}^{(k)}$. We now bound the regret in the case of failing confidence intervals from Thm. 3. We have

$$R^{\text{fail}} = \sum_{k=1}^{K}\left(\sum_{t=t^{(k)}}^{t^{(k+1)}-1}\Big(\eta^+ - r_t(x_t,\widetilde{\pi}_k(\boldsymbol{y}_t))\Big)\mathbb{1}_{M\notin\mathcal{M}^{(k)}}\right) \le r_{\max}\sum_{k=1}^{K} v^{(k)}\mathbb{1}_{M\notin\mathcal{M}^{(k)}} \le r_{\max}\sum_{t=1}^{N} t\mathbb{1}_{M\notin\mathcal{M}^{(t)}},$$

where $\mathcal{M}^{(t)}$ denotes the set of admissible POMDPs according to the samples available at time t. We recall from Step 2 that the number of steps needed for the statement of Thm. 3 to be valid is $\overline{t} = 4(\overline{v}+1)$. If N is large enough so that $\overline{t} \le N^{1/4}$, then we bound the regret as

$$R^{\text{fail}} \le \sum_{t=1}^{\lfloor N^{1/4}\rfloor} t\mathbb{1}_{M\notin\mathcal{M}^{(t)}} + \sum_{t=\lfloor N^{1/4}\rfloor+1}^{N} t\mathbb{1}_{M\notin\mathcal{M}^{(t)}} \le \sqrt{N} + \sum_{t=\lfloor N^{1/4}\rfloor+1}^{N} t\mathbb{1}_{M\notin\mathcal{M}^{(t)}}.$$

We are left with bounding the last term. We first notice that if we redefine the confidence intervals in Thm. 3 by substituting the term $\log(1/\delta)$ by $\log(t^6/\delta)$, we obtain that at any time instants t, the statement holds with probability $1 - 24A\delta/t^6$. Since

$$\sum_{t=\lfloor N^{1/4}\rfloor+1}^{N}\frac{24A}{t^6} \le \frac{24A}{N^{6/4}} + \int_{\lfloor N^{1/4}\rfloor}^{\infty}\frac{24A}{t^6}dt = \frac{24A}{N^{6/4}} + \frac{24A}{5N^{5/4}} \le \frac{144A}{5N^{5/4}} \le \frac{30A}{N^{5/4}},$$

then M is in the set of $\mathcal{M}^{(k)}$ at any time step $\lfloor N^{1/4}\rfloor \le t \le N$ with probability $1 - 30A\delta/N^{5/4}$. As a result, the regret due to failing confidence bound is bounded by $\sqrt{N}$ with probability $1 - 30A\delta/N^{5/4}$.

Step 4 (reward model). Now we focus on the per-episode regret $\Delta^{(k)}$ for $k > \overline{K}$ when M is contained in $\widetilde{\mathcal{M}}_k$ and we decompose it in two terms

$$\Delta^{(k)} \le \sum_{x=1}^{X}\sum_{l=1}^{A} v^{(k)}(x,l)\underbrace{\Big(\widetilde{\eta}^{(k)} - \widetilde{r}^{(k)}(x,l)\Big)}_{(a)} + \underbrace{\sum_{x=1}^{X}\sum_{l=1}^{A} v^{(k)}(x,l)\Big(\widetilde{r}^{(k)}(x,l) - \bar{r}(x,l)\Big)}_{(b)},$$

where $\widetilde{r}^{(k)}$ is the state-action expected reward used in the optimistic POMDP $\widetilde{\mathcal{M}}^{(k)}$. We start by bounding the second term, which only depends on the size of the confidence intervals in estimating the reward model of the POMDP. We have

$$\begin{aligned}(b) &\le \sum_{l=1}^{A}\sum_{x=1}^{X} v^{(k)}(x,l)\max_{x'\in\mathcal{X}}\left|\widetilde{r}^{(k)}(x',l)-\bar{r}(x',l)\right| \\ &= \sum_{l=1}^{A} v^{(k)}(l)\max_{x\in\mathcal{X}}\left|\widetilde{r}^{(k)}(x,l)-\bar{r}(x,l)\right| \\ &\le 2\sum_{l=1}^{A} v^{(k)}(l)\mathcal{B}_R^{(k,l)},\end{aligned}$$

where $\mathcal{B}_R^{(k,l)}$ corresponds to the term $\mathcal{B}_R^{(l)}$ in Thm. 3 computed using the $N^{(k)}(l)$ samples collected during episode $k^{(l)} = \arg\max_{k'<k} v^{(k')}(l)$.

Step 5 (transition and observation models). We now proceed with studying the first term (a), which compares the (optimal) average reward in the optimistic model $\widetilde{\mathcal{M}}^{(k)}$ and the (optimistic) rewards collected on the states traversed by policy $\widetilde{\pi}^{(k)}$ in the true POMDP. We first recall the Poisson equation. For any POMDP M and any policy π, the action value function $Q_{\pi,M}: \mathcal{X}\times\mathcal{A}\to\mathbb{R}$ satisfies

$$\begin{aligned}Q_{\pi,M}(x,a) &= \overline{r}(x,a)-\eta_\pi+\sum_{x'} f_T(x'|x,a)\Big(\sum_{a'} f_\pi(a'|x')Q_{\pi,M}(x',a')\Big) \qquad (39)\\ \Rightarrow \eta_\pi-\overline{r}(x,a) &= \sum_{x'} f_T(x'|x,a)\Big(\sum_{a'} f_\pi(a'|x')Q_{\pi,M}(x',a')\Big)-Q_{\pi,M}(x,a),\end{aligned}$$

where $f_\pi(a'|x') = \sum_y f_O(y|x')f_\pi(a'|y)$ and terms such as $\overline{r}$ and f_T depend on the specific POMDP. We define the function $\overline{Q}_{\pi,M}(x,l)$ as

$$\overline{Q}_{\pi,M}(x,l) = Q_\pi(x,l) - \frac{\min_{x,l} Q_{\pi,M}(x,l)-\max_{x,l} Q_{\pi,M}(x,l)}{2},$$

which is a centered version of $Q_{\pi,M}(x,l)$. In order to characterize $\overline{Q}$, we introduce a notion of diameter specific to POMDPs and the family of policies considered in the problem

$$D := \max_{x,x'\in[X],\ l,l'\in[A]}\ \min_{\pi\in\mathcal{P}} \mathbb{E}[T(x',l'|M,\pi,x,l)],$$

where $T(x',l'|M,\pi,x,l)$ is the (random) time that takes to move from state x by first taking action l and then following policy π before reaching state x' and performing action l'. An important feature of the diameter is that it can be used to upper bound the range of the function $Q_{\pi,M}$ computed using a policy derived from Eq 14 in an optimistic model. The proof of this fact is similar to the case of the diameter for MDPs. We first recall the

definition of the optimistic policy

$$\widetilde{\pi}^{(k)} = \arg\max_{\pi \in \mathcal{P}} \max_{M \in \mathcal{M}^{(k)}} \eta(\pi; M), \tag{40}$$

while M_k is the optimistic model. The joint choice of the policy and the model can be seen as if a POMDP $\widetilde{M}^+$ with augmented action space $\mathcal{A}'$ is considered. Taking an action $a' \in \mathcal{A}'$ in a state x corresponds to a basic action $a \in \mathcal{A}$ and a choice of transition, reward, and observation model from $\mathcal{M}^{(k)}$. We denote by $\mathcal{P}^{(\mathcal{M}^{(k)})}$ the corresponding augmented policy space using $\mathcal{P}$ and the set of admissible POMDPs $\mathcal{M}^{(k)}$. As a result, for any augmented policy $\widetilde{\pi}^+$ executed in $\widetilde{M}^+$ we obtain transitions, rewards, and observations that are equivalent to executing a (standard) policy $\widetilde{\pi}$ in a specific POMDP $\widetilde{M} \in \mathcal{M}^{(k)}$ and viceversa. As a result, computing $\widetilde{\pi}^{(k)}$ and the corresponding optimistic model $\widetilde{M}^{(k)}$ is equivalent to choosing the optimal policy in the POMDP $\widetilde{M}^+$. Since the true POMDP of diameter D is in $\mathcal{M}^{(k)}$ with high-probability, then the diameter of the augmented POMDP $\widetilde{M}^+$ is at most D. Furthermore, we can show that the optimal policy has a Q-value with range bounded by D. Let us assume that there exists state-action pairs $(x, a), (x', a')$ such that $Q_{\widetilde{\pi}^{(k)}, \widetilde{M}^{(k)}}(x, a) - Q_{\widetilde{\pi}^{(k)}, \widetilde{M}^{(k)}}(x', a') \geq r_{\max} D$. Then it is easy to construct a policy different from $\widetilde{\pi}^{(k)}$ which achieves a better Q-value. We already know by definition of diameter that there exists a policy moving from x to x' in D steps on average. If from x' the optimal policy is followed, then only $r_{\max} D$ reward could have been missed at most and thus the difference in action-value function between x and x' cannot be larger than $r_{\max} D + 1$, thus contradicting the assumption. As a result, we obtain

$$\max_{x,a} Q_{\widetilde{\pi}^{(k)}, \widetilde{M}^{(k)}}(x, a) - \min_{x,a} Q_{\widetilde{\pi}^{(k)}, \widetilde{M}^{(k)}}(x, a) \leq r_{\max} D \tag{41}$$

and thus

$$\overline{Q}_{\widetilde{\pi}^{(k)}, \widetilde{M}^{(k)}}(x, a)(x, l) \leq r_{\max} \frac{(D+1)}{2}.$$

By replacing Q with $\overline{Q}$ in the Poisson equation for the optimistic POMDP characterized by the transition model $\widetilde{f}_T^{(k)}$ and where the observation model is such that the policy $\widetilde{\pi}_k$ takes

actions according to the distribution $\widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(\cdot|x)$, we obtain

$$\begin{aligned}
(a) &= \sum_{x'} \widetilde{f}^{(k)}_T(x'|x,l)\Big(\sum_{l'} \widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(a'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',a')\Big) - \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x,l) \\
&= \sum_{x'} \widetilde{f}^{(k)}_T(x'|x,l)\Big(\sum_{l'} \widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(l'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')\Big) \\
&- \sum_{x'} f_T(x'|x,l)\Big(\sum_{l'} f_{\widetilde{\pi}^{(k)}}(l'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')\Big) \\
&\quad + \sum_{x'} f_T(x'|x,l)\Big(\sum_{l'} f_{\widetilde{\pi}^{(k)}}(l'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')\Big) - \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x,l) \\
&= \underbrace{\sum_{x'}\sum_{l'} \big(\widetilde{f}^{(k)}_T(x'|x,l)\widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(l'|x') - f_T(x'|x,l)f_{\widetilde{\pi}^{(k)}}(l'|x')\big)\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')}_{(c)} \\
&\quad \underbrace{+ \sum_{x'} f_T(x'|x,l)\Big(\sum_{l'} f_{\widetilde{\pi}^{(k)}}(l'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')\Big) - \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x,l)}_{\zeta^{(k)}(x,l)}.
\end{aligned}$$

The term (c) can be further expanded as

$$\begin{aligned}
(c) &= \sum_{x'}\sum_{l'} \Big(\big(\widetilde{f}^{(k)}_T(x'|x,l) - f_T(x'|x,l)\big)\widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(l'|x') - f_T(x'|x,l)\big(\widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(l'|x') - f_{\widetilde{\pi}^{(k)}}(l'|x')\big)\Big)\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l') \\
&\le \Big(\underbrace{\sum_{x'} |\widetilde{f}^{(k)}_T(x'|x,l) - f_T(x'|x,l)|}_{(d)} + \sum_{x'} f_T(x'|x,l) \underbrace{\sum_{l'} |\widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(l'|x') - f_{\widetilde{\pi}^{(k)}}(l'|x')|}_{(d')}\Big)\|\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}\|_\infty,
\end{aligned}$$

where we used the fact that $\sum_{l'} \widetilde{f}^{(k)}_{\widetilde{\pi}^{(k)}}(l'|x') = 1$. For the first term we can directly apply the bound from Thm. 3, Eq. 12 and obtain

$$(d) = \|\widetilde{f}^{(k)}_T(\cdot|x,l) - f_T(\cdot|x,l)\|_1 \le 2\sqrt{X}\mathcal{B}^{(k,l)}_T.$$

As for (d'), the error in estimating the observation model is such that

$$(d') = \sum_{l'}\sum_{y} |\widetilde{f}^{(k)}_O(y|x') - f_O(y|x')| f^{(k)}_{\widetilde{\pi}^{(k)}}(l'|y) = \sum_{y} |\widetilde{f}^{(k)}_O(y|x') - f_O(y|x')| \le 2\mathcal{B}^{(k)}_O.$$

Plugging back these two bounds into (c) together with the bound on $\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x,l)$, we obtain

$$(c) \le 2(\sqrt{X}\mathcal{B}^{(k,l)}_T + \mathcal{B}^{(k)}_O) r_{\max}\frac{(D+1)}{2}.$$

The term (a) in the per-episode regret is thus bounded as

$$(a) \leq 2\big(\sqrt{X}\mathcal{B}_T^{(k,l)} + \mathcal{B}_O^{(k)}\big) r_{\max} \frac{(D+1)}{2} + \zeta^{(k)}(x,l).$$

Step 6 (Residual error). We now bound the cumulative sum of the terms $\zeta^{(k)}(x,l)$. At each episode k we have

$$\sum_{x=1}^{X}\sum_{l=1}^{A} v^{(k)}(x,l)\zeta^{(k)}(x,l) = \sum_{t=t^{(k)}}^{t^{(k+1)}} \sum_{x'} f_T(x'|x_t,l_t)\Big(\sum_{l'} f_{\widetilde{\pi}^{(k)}}(l'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')\Big) - \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x_t,l_t),$$

we introduce the term $\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x_{t+1},l_{t+1})$ and we obtain two different terms

$$\begin{aligned}
&\sum_{x=1}^{X}\sum_{l=1}^{A} v^{(k)}(x,l)\zeta^{(k)}(x,l) \\
&= \sum_{t=t^{(k)}}^{t^{(k+1)}} \sum_{x'} f_T(x'|x_t,l_t)\Big(\sum_{l'} f_{\widetilde{\pi}^{(k)}}(l'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')\Big) - \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x_{t+1},l_{t+1}) \\
&\quad + \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x_{t+1},l_{t+1}) - \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x_t,l_t) \\
&\leq \sum_{t=t^{(k)}}^{t^{(k+1)}} \underbrace{\sum_{x'} f_T(x'|x_t,l_t)\Big(\sum_{l'} f_{\widetilde{\pi}^{(k)}}(l'|x')\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x',l')\Big) - \overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}(x_{t+1},l_{t+1})}_{Y_t} + r_{\max} D,
\end{aligned}$$

where we use the fact that the range of $\overline{Q}_{\widetilde{\pi}^{(k)},\widetilde{M}^{(k)}}$ is bounded by the diameter D. We notice that $\mathbb{E}[Y_t|x_1,a_1,y_1,\ldots,x_t,a_t,y_t] = 0$ and $|Y_t| \leq r_{\max} D$, thus Y_t is a martingale difference sequence and we can use Azuma's inequality to bound its cumulative sum. In fact, we have

$$\sum_{t=1}^{N} Y_t \leq D\sqrt{2N\log(N^{5/4}/\delta)}$$

with probability $1-\delta/N^{5/4}$. As a result we can now bound the total sum of the terms $\zeta^{(k)}$ as

$$\sum_{x=1}^{X}\sum_{l=1}^{A} v^{(k)}(x,l)\zeta^{(k)}(x,l) \leq \sum_{t=1}^{N} Y_t + r_{\max} KD \leq r_{\max} D\sqrt{2N\log(N^{5/4}/\delta)} + r_{\max} KD.$$

Step 7 (per-episode regret). We can now bound the per-episode regret as

$$\Delta^{(k)} \leq \sum_{l=1}^{A} v^{(k)}(l) 2\bigg(\mathcal{B}_R^{(k,l)} + \big(\sqrt{X}\mathcal{B}_T^{(k,l)} + \mathcal{B}_O^{(k)}\big) r_{\max}\frac{(D+1)}{2}\bigg).$$

Recalling the results from Thm. 3, we can bound the first term in the previous expression as

$$\Delta^{(k)} \le 3r_{\max}(D+1)\sqrt{d' \log(N^6/\delta)}\big(C_O + C_R + C_T X^{3/2}\big) \sum_{l=1}^{A} \frac{v^{(k)}(l)}{\lambda^{(k,l)}} \sqrt{\frac{1}{N^{(k)}(l)}}.$$

Since the number of samples $N^{(k)}(l)$ collected in the previous episode is at most doubled in the current episode k, we have that $N^{(k)}(l) \ge v^{(k)}(l)/2$, then we obtain

$$\Delta^{(k)} \le 9r_{\max}(D+1)\sqrt{v^{(k)} d' \log(N^6/\delta)}\big(C_O + C_R + C_T X^{3/2}\big) \max_{l=1,\dots,A} \frac{1}{\lambda^{(k,l)}}.$$

Step 8 (bringing all together). Now we have to recollect all the previous terms: the number of episodes needed to use Thm. 3 (Step 2), regret in case of failing confidence intervals (Step 3), and the per-episode regret (Step 7). The result is

$$\begin{aligned}&\mathrm{Reg}_N \\ &\le r_{\max}\Big(\underbrace{\sqrt{N \log(N^6/\delta)}}_{\text{Step 1}} + \underbrace{4Ar_{\max}(\overline{v}+1)}_{\text{Step 2}} + \underbrace{\sqrt{N}}_{\text{Step 3}} + \underbrace{D\sqrt{2N \log(N^{5/4}/\delta)} + r_{\max} K D}_{\text{Step 6}} \Big) + \sum_{k=\widetilde{K}+1}^{K} \Delta^{(k)}.\end{aligned}$$

The last term can be bounded as

$$\sum_{k=\widetilde{K}+1}^{K} \Delta^{(k)} \le \frac{9r_{\max}(D+1)}{\overline{\lambda}} \sqrt{N d' \log(N^6/\delta)}\big(C_O + C_R + C_T X^{3/2}\big).$$

where $\overline{\lambda} = \min_{k,l} \lambda^{(k,l)}$ and it is defined as in the statement of the theorem. Since K is a random number, we need to provide an upper-bound on it. We can use similar arguments as in Step 2. Given the stopping criterion of each episode, at most every A steps, then length of an episode is doubled. As a result, after K episodes, we have these inequalities

$$N = \sum_{k=1}^{K} v^{(k)} \ge \sum_{k'=1}^{K/A} 2^{k'} \ge 2^{K/A}.$$

As a result, we obtain the upper bound $K \le \overline{K}_N \le A \log_2 N$. Bringing all the bounds together we obtain the final statement with probability $1 - \delta/(4N^{5/4})$.

■

Appendix E. Proof of Remark 2 in Section 4

We first prove the bound on the transition tensor, which requires re-deriving step 4 in the proof of Thm. 3.

Proof

Step 4 (bound on f_T). The derivation of the bound for $\widehat{f}_T$ is more complex since each distribution $\widehat{f}_T(\cdot|x,a)$ is obtained as the solution of the linear system of equations like Eq. 7, that is for any state i and action l we compute

$$[T]_{i,:,l} = W^\dagger [V_3^{(l)}]_{:,i},$$

and derive transition tensor as follows

$$[\widehat{T}]_{i,:,l} = \widehat{W}^\dagger [\widehat{V}_3^{(l)}]_{:,i}, \tag{42}$$

where $\widehat{W}$ is obtained plugging in the estimates of $\widehat{f}_O$ and $\widehat{f}_R$ and the policy f_π. We first recall the following general result for the pseudo-inverse of a matrix and we instance it in our case. Let W and $\widehat{W}$ be any pair of matrix such that $\widehat{W} = W + E$ for a suitable error matrix E, then we have Meng and Zheng (2010)

$$\|W^\dagger - \widehat{W}^\dagger\|_2 \leq \frac{1+\sqrt{5}}{2} \max\left\{\|W^\dagger\|_2, \|\widehat{W}^\dagger\|_2\right\} \|E\|_2, \tag{43}$$

where $\|\cdot\|_2$ is the spectral norm. From the definition of W and $V_2^{(l)}$ we have

$$[W]_{s,j} = [W]_{(n,m,k),j} = f_\pi(k|\boldsymbol{e}_n) f_R(\boldsymbol{e}_m|j,k) f_O(\boldsymbol{e}_n|j),$$
$$[V_2^{(l)}]_{s,i} = [V_2^{(l)}]_{(n',m'),i} = \rho(i,l) f_\pi(l|\boldsymbol{e}_{n'}) f_O(\boldsymbol{e}_{n'}|i) f_R(\boldsymbol{e}_{m'}|i,l).$$

Then it is clear that any column j of W is the result of stacking the matrices $V_2^{(l)}$ over actions properly re-weighted by $\rho(i,l)$, that is

$$[W]_{:,j} = \left[\frac{[V_2^{(1)}]_{:,j}^\top}{\rho(j,1)} \;\cdots\; \frac{[V_2^{(l)}]_{:,j}^\top}{\rho(j,l)} \;\cdots\; \frac{[V_2^{(A)}]_{:,j}^\top}{\rho(j,A)}\right]^\top.$$

The same relationship holds for $\widehat{W}$ and $\widehat{V}_2^{(l)}$. Since Lemmas 8 and Eq. 32 provide a bound on the error for each column of $V_2^{(l)}$ for each action and a bound on the error of $\rho(i,l)$ is already developed in Step 2, we can bound the ℓ_2 norm of the estimation error for each column of W and $\widehat{W}$ as

$$\|[\widehat{W}]_{:,i} - [W]_{:,i}\|_2^2 = \sum_l^A \sum_{m',n'}^{R,Y} \left(\frac{[\widehat{V}_2^{(l)}]_{(n',m'),i}}{\widehat{\rho}(i,l)} - \frac{[V_2^{(l)}]_{(n',m'),i}}{\rho(i,l)}\right)^2.$$

Following similar steps as in Step 3, each summand can be bounded as

$$\left|\frac{[\widehat{V}_2^{(l)}]_{(n',m'),i}}{\widehat{\rho}(i,l)} - \frac{[V_2^{(l)}]_{(n',m'),i}}{\rho(i,l)}\right| \leq \left|[\widehat{V}_2^{(l)}]_{(n',m'),i} - [V_2^{(l)}]_{(n',m'),i}\right| + \left|\frac{1}{\rho(i,l)} - \frac{1}{\widehat{\rho}(i,l)}\right| [V_2^{(l)}]_{(n',m'),i}.$$

Then the ℓ_2-norm of the error is bounded as

$$\begin{aligned}\|[\widehat{W}]_{:,i} - [W]_{:,i}\|_2 &\le \sqrt{\sum_{l=1}^{A} \|[\widehat{V}_2^{(l)}]_{:,i} - [V_2^{(l)}]_{:,i}\|_2^2} + \sqrt{\sum_{l=1}^{A}\left(\frac{1}{\rho(i,l)} - \frac{1}{\widehat{\rho}(i,l)}\right)^2 \sum_{m',n'}^{R,Y} [V_2^{(l)}]^2_{(n',m'),i}} \\ &\le \sqrt{\sum_{l=1}^{A} \|[\widehat{V}_2^{(l)}]_{:,i} - [V_2^{(l)}]_{:,i}\|_2^2} + \sqrt{\sum_{l=1}^{A}\left(\frac{1}{\rho(i,l)} - \frac{1}{\widehat{\rho}(i,l)}\right)^2} \\ &\le 20\sqrt{\sum_{l=1}^{A} \frac{\epsilon_3(l)}{\sigma_{3,1}^{(l)}}^2} + \sqrt{\sum_{l=1}^{A} \epsilon_\rho^2(i,l)} \\ &\le \sum_{l=1}^{A}\left(20\frac{\epsilon_3(l)}{\sigma_{3,1}^{(l)}} + \epsilon_\rho(i,l)\right) \le 40\sqrt{YR}\sum_{l=1}^{A}\frac{\epsilon_3(l)}{\sigma_{3,1}^{(l)}\pi_{\min}^{(l)}}.\end{aligned}$$

Now we can bound the spectral norm of the error in estimating W as

$$\|\widehat{W} - W\|_2 \le \|\widehat{W} - W\|_F \le 40\sqrt{XYR}\sum_{l=1}^{A}\frac{\epsilon_3(l)}{\sigma_{3,1}^{(l)}\pi_{min}^{(l)}}. \tag{44}$$

We now focus on the maximum in Eq. 34, for which we need to bound the spectral norm of the pseudo-inverse of the estimated W. We have $\|\widehat{W}^\dagger\|_2 \le (\sigma_X(\widehat{W}))^{-1}$ where $\sigma_X(\widehat{W})$ is the X-th singular value of matrix $\widehat{W}$ whose perturbation is bounded by $\|\widehat{W} - W\|_2$. Since matrix W is a rank X matrix on Asm. 2 then

$$\|\widehat{W}^\dagger\|_2 \le (\sigma_X(\widehat{W}))^{-1} \le \frac{1}{\sigma_X(W)}\left(1 + \frac{\|\widehat{W} - W\|_2}{\sigma_X(W)}\right) \le \frac{1}{\sigma_X(W)}\left(1 + \frac{\|\widehat{W} - W\|_F}{\sigma_X(W)}\right).$$

We are now ready to bound the estimation error of the transition tensor. From the definition of Eq. 33 we have that for any state $i = 1, \dots, X$ the error is bounded as

$$\|T_{i,:,l} - \widehat{T}_{i,:,l}\|_2 \le \|T_{:,:,l} - \widehat{T}_{:,:,l}\|_2 \le \|\widehat{W}^\dagger - W^\dagger\|_2 \|V_3^{(l)}\|_2 + \|\widehat{V}_3^{(l)} - V_3^{(l)}\|_2 \|\widehat{W}^\dagger\|_2.$$

In Lemma 5 we have a bound on the ℓ_2-norm of the error for each column of $V_3^{(l)}$, thus we have $\|\widehat{V}_3^{(l)} - V_3^{(l)}\|_2 \le \|\widehat{V}_3^{(l)} - V_3^{(l)}\|_F \le 18\sqrt{X}\epsilon_3(l)$. Using the bound on Eq. 34 and denoting $\|V_3^{(l)}\|_2 = \sigma_{\max}(V_3^{(l)})$ we obtain

$$\begin{aligned}\|T_{i,:,l} - \widehat{T}_{i,:,l}\|_2 &\le \frac{1+\sqrt{5}}{2}\frac{\|\widehat{W} - W\|_F}{\sigma_X(W)}\left(1 + \frac{\|\widehat{W} - W\|_F}{\sigma_X(W)}\right)\sigma_{\max}(V_3^{(l)}) + 18\sqrt{X}\epsilon_3(l)\frac{1}{\sigma_X(W)}\left(1 + \frac{\|\widehat{W} - W\|_F}{\sigma_X(W)}\right) \\ &\le \frac{2}{\sigma_X(W)}\left(1 + \frac{\|\widehat{W} - W\|_F}{\sigma_X(W)}\right)\left(\sigma_{\max}(V_3^{(l)})\|\widehat{W} - W\|_F + 18\sqrt{X}\epsilon_3(l)\right)\end{aligned}$$

Using the bound in Eq. 44 and $\sigma_{\max}(V_3^{(l)}) \leq \sqrt{X}$ we obtain

$$\|T_{i,:,l} - \widehat{T}_{i,:,l}\|_2 \leq \frac{4}{\sigma_X(W)}\Big(40\sqrt{X^2YR}\sum_{l=1}^{A}\frac{\epsilon_3(l)}{\sigma_{3,1}^{(l)}\pi_{min}^{(l)}} + 18\sqrt{X}\epsilon_3(l)\Big)$$
$$\leq C_T\sqrt{AX^2YR}\max_{l'=1,\dots,A}\frac{1}{\lambda^{l'}}\sqrt{\frac{\log(8/\delta)}{N_{l'}}},$$

thus leading to the final statement for the bound over confidence of transition tensor. ■

We now move to analyzing how the new estimator for the transition tensor affects the regret of the algorithm. The proof is exactly the same as in Thm. 4 except for step 8.
Proof

per-episode regret: The per-episode regret is bounded as

$$\Delta^{(k)} \leq 2\sum_{l=1}^{A} v^{(k)}(l)\Big(\mathcal{B}_R^{(k,l)} + (\mathcal{B}_T^{(k,l)} + \mathcal{B}_O^{(k)})r_{\max}\frac{(D+1)}{2}\Big).$$

All the terms can be treated as before except for the cumulative regret due to the transition model estimation error. We define $\Delta_N = \sum_{k=\widetilde{K}+1}^{K}\sum_{l=1}^{A} v^{(k)}(l)(\mathcal{B}_T^{(k,l)})r_{\max}(D+1)$, which gives

$$\Delta_T = \sum_k r_{max}(D+1)\sum_{l=1}^{A} v^{(k)}(l)\sqrt{X}\max_{l'=1,\dots,A}\frac{C_T^{(k)}\sqrt{X^2YR}}{\lambda^{(k)(l')}}\sqrt{\frac{\log(N^6/\delta)}{v^{(k)}(l')}}.$$

Let $\tau_{M,\pi}^{(l)}$ the mean passage time between two steps where action l is chosen according to policy $\pi \in \mathcal{P}$ and restate a π-diameter ration D_{ratio}^{π}

$$D_{\text{ratio}}^{\pi} = \frac{\max_{l\in\mathcal{A}}\tau_{M,\pi}^{(l)}}{\min_{l\in\mathcal{A}}\tau_{M,\pi}^{(l)}}$$

and D_{ratio}

$$D_{\text{ratio}} = \max_{\pi\in\mathcal{P}} D_{\text{ratio}}^{\pi}.$$

We need the following lemma which directly follows from Chernoff-Heoffding inequality.

Lemma 9 *By Markovian inequality, the probability that during $2\tau_{M,\pi}^{(l)}$ the action l is not visited is less than $\frac{1}{2}$. Then during episode k*

$$\mathbb{P}\Big\{v^{(k)}(l) \leq \frac{1}{2}\frac{v^{(k)}}{2\tau_{M,\pi}^{(l)}} - \sqrt{v^{(k)}\log(\frac{1}{\delta})}\Big\} \leq \delta$$

On the other hand we have

$$\mathbb{P}\Big\{v^{(k)}(l) \geq \frac{v^{(k)}}{2\tau^{(l)}_{M,\pi}} + \sqrt{v^{(k)}\log(\frac{1}{\delta})}\Big\} \leq \delta$$

Let $C_T = \max_{l\in\mathcal{A},k\in\{k'|t_{k'}\leq N\}} \frac{C_T^{(k)}\sigma_{\max}(V_3^{(k)(l')})}{\lambda^{(k)(l')}}$ then

$$\Delta_T \leq X^{\frac{3}{2}}\sqrt{YR}r_{max}\sqrt{\log\frac{1}{\delta}}(D+1)\sum_k\sum_{l=1}^{A}\sqrt{v^{(k)}(l)}\sqrt{\underbrace{\frac{v^{(k)}(l)}{\min_{l'\in\mathcal{A}} v^{(k)}(l')}}_{(a')}}$$

From Lemma 9 we have that

$$(a') = \frac{v^{(k)}(l)}{\min_{l'\in\mathcal{A}} v^{(k)}(l')} \leq \frac{\frac{v^{(k)}}{2\min_{l\in\mathcal{A}}\tau^{(l)}_{M,\pi}} + \sqrt{v^{(k)}\log(\frac{1}{\delta})}}{\frac{1}{2}\frac{v^{(k)}}{2\max_{l\in\mathcal{A}}\tau^{(l)}_{M,\pi}} - \sqrt{v^{(k)}\log(\frac{1}{\delta})}}$$

$$\leq 2D^{\pi}_{\text{ratio}} + 8D^{\pi}_{\text{ratio}}\max_{l\in\mathcal{A}}\tau^{(l)}_{M,\pi}\sqrt{\frac{\log(\frac{1}{\delta})}{v^{(k)}}} + 4\max_{l\in\mathcal{A}}\tau^{(l)}_{M,\pi}\sqrt{\frac{\log(\frac{1}{\delta})}{v^{(k)}}} + 16(\max_{l\in\mathcal{A}}\tau^{(l)}_{M,\pi})^2\frac{\log(\frac{1}{\delta})}{v^{(k)}},$$

with probability $1-2\delta$. The first term dominates all the other terms and thus

$$(a') \leq 2D^{\pi}_{\text{ratio}} \leq 2D_{\text{ratio}}$$

with probability at least $1-\delta$. Thus we finally obtain

$$\Delta_T = \sum_{k=\widetilde{K}+1}^{K}\sum_{l=1}^{A} v^{(k)}(l)\big(\mathcal{B}_T^{(k,l)}\big)r_{\max}(D+1) \leq X^{\frac{3}{2}}\sqrt{YR}r_{max}\sqrt{\log\frac{1}{\delta}}(D+1)\sqrt{2D_{\text{ratio}}}\frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{AN}$$

with probability at least $1-2\overline{K}_N\delta$. Then with probability at least $1-\delta(8A+5\overline{K}_N)$ the regret is bounded by the final statement. ■

Appendix F. Experiments

In the following section, we show how SM-UCRL algorithm outperforms other well-known methods in both synthetic environment and simple computer game.

F.1. Synthetic Environment

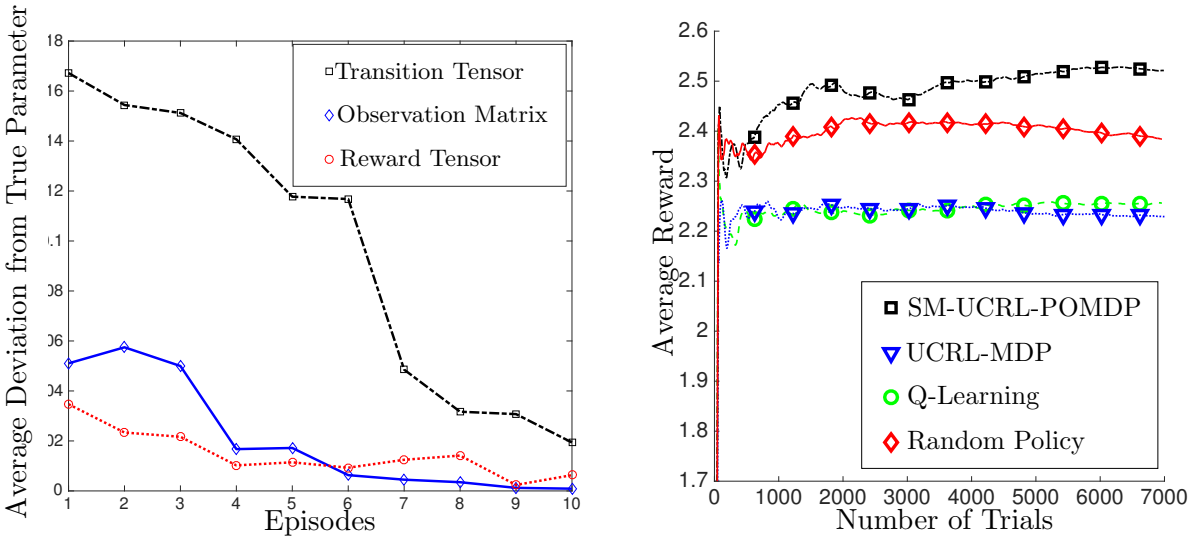
In this subsection, we illustrate the performance of our method on a simple synthetic environment which follows a POMDP structure with $X = 2$, $Y = 4$, $A = 2$, $R = 4$, and

$r_{max} = 4$. We find that spectral learning method converges quickly to the true model parameters, as seen in Fig. [F.1]. Estimation of the transition tensor T takes longer compared to estimation of observation matrix O and reward Tensor R. This is because the observation and reward matrices are first estimated through tensor decomposition, and the transition tensor is estimated subsequently through additional manipulations. Moreover, the transition tensor has more parameters since it is a tensor (involving observed, hidden and action states) while the observation and reward matrices involve fewer parameters.

For planning, given the POMDP model parameters, we find the memoryless policy using a simple alternating minimization heuristic, which alternates between updates of the policy and the stationary distribution. We find that in practice this converge to a good solution. The regret bounds are shown in Fig. [F.1]. We compare against the following policies: (1) baseline random policies which simply selects random actions without looking at the observed data, (2) UCRL-MDP Auer et al. (2009) which attempts to fit a MDP model to the observed data and runs the UCRL policy, and (3) Q-Learning Watkins and Dayan (1992) which is a model-free method that updates policy based on the Q-function. We find that our method converges much faster than the competing methods. Moreover, it converges to a much better policy. Note that the MDP-based policies UCRL-MDP and Q-Learning perform very poorly, and are even worse than the baseline are too far from SM-UCRL policy. This is because the MDP policies try to fit data in high dimensional observed space, and therefore, have poor convergence rates. On the other hand, our spectral method efficiently finds the correct low dimensional hidden space quickly and therefore, is able to converge to an efficient policy.

F.2. Simple Atari Game

In the following subsection, we illustrate the performance of our method on an environment of simple computer Game. In this game, as it is shown in Figs. [F.2,F.2], the environment is 10×10 grid world and this grid world environment contains 5 sweet *(green)* and 5 poisonous apples *(red)*. The environment uniformly spread out these apples in the grid world. In addition, each of these sweat and poisonous apples lasts either uniformly for $(1, 2)$ time steps or being eaten by the agent. In this game, there exist an agent who is in interaction with the corresponding environment and has either set of actions *(N,W,S,E)* or set of actions *(N,NW,W,SW,S,SE,E,NE)*. At each time step the agent chooses an action and based on the direction of that action, she deterministically moves to new location by one step toward that direction. If there is an sweat apple at new location which is not expired yet, the agent will score up by one, score down by one if it is poisonous one. The agent score nothing when the reached apple is expired by the time. This process adds more randomness to the rewarding process. To bring more detail about the game, when the action is one of *(N,W,S,E)* and results in new location which is wall, the agent remains at her previous location. If one of *(NW,SW,SE,NE)*, which are a combination of two directions ends up to the wall, the agent remains at her previous location in the coordinate orthogonal to the wall and goes one step further toward the other direction. In this game, at each time step, the agent just partially observes the environment, Fig. [F.2] just one single box above of the agent is visible to her, and Fig. [F.2] just three boxes above of her are observable. The randomness on rewarding process and partial observability bring the notion of hidden



a)Learning by Spectral Method

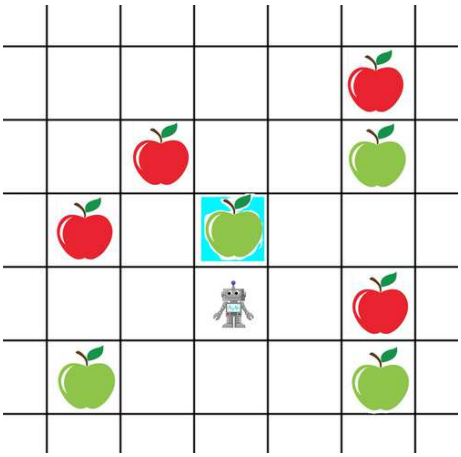
b) Regret Performance

Figure 3: (a)Accuracy of estimated model parameter through tensor decomposition. See Eqs. 11,10 and 12. (b)Comparison of SM-UCRL-POMDP is our method, UCRL-MDP which attempts to fit a MDP model under UCRL policy, $\epsilon - greedy$ Q-Learning, and a Random Policy.

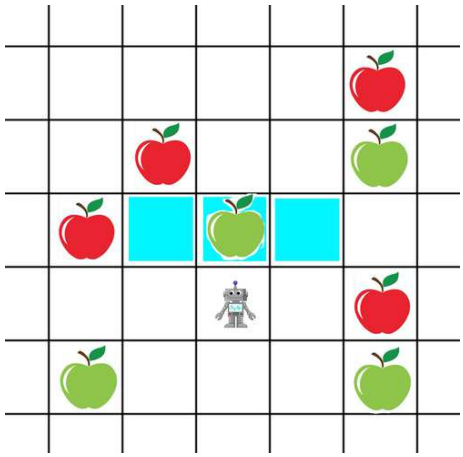
structure and push the environment more toward to POMPD models rather than MDP models.

For single box observable setting, the observation set has cardinality of 4, *(wall, sweat apple, poisonous apple, nothing)*. For this environment, let's model the environment behavior as POMDP model with number of states equal to 3 *(X=3)*. We apply our SM-UCRL method on this environment and, for planning, given the POMDP model parameters, same as before, we find the memoryless policy using a simple alternating minimization heuristic, except here we add some level of randomness to the policy matrix and reduce this randomness level when the number of episodes grows. In addition, we reimplement DNN (Deep Neural Network) algorithm proposed in Mnih et al. (2013) and make to interact with same environment. For this DNN, we consider three hidden layers DNN with 10 hidden units at each hidden layer with *hyperbolic tangent* activation function. For back propagation, we use *RMSProp* method which is shown to be robust and stable. Figs. [F.2,F.2] show the performance of both SM-UCRL and DNN when the action set is *(N,W,S,E)* and when it is *(N,NW,W,SW,S,SE,E,NE)*. We show that not only SM-UCRL captures the environment behavior faster than DNN but also reaches to the better long term average reward. We run DNN couple of times and represent the average performance as it is shown in Figs. [F.2,F.2]. DNN some times gets stuck in some local minima and it results in bad performance which deduces the its average performance.

In other setting, when three of boxes are observable Fig. [F.2], the observation set has cardinality of 64, four possible observation for each of three boxes (we know that some

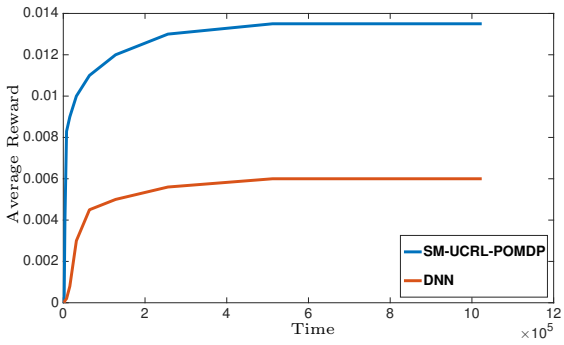


a)Single box observable environment

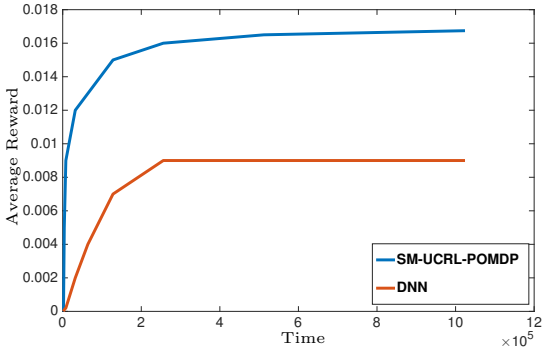


b) Triple boxes observable environment

Figure 4: (a)Accuracy of estimated model parameter through tensor decomposition. See Eqs. 11,10 and 12. (b)Comparison of SM-UCRL-POMDP is our method, UCRL-MDP which attempts to fit a MDP model under UCRL policy, $\epsilon - greedy$ Q-Learning, and a Random Policy.



a)Performances in Single box observable environment, Action set *(N,W,S,E)*



b) Performances in Single box observable environment, Action set *(N,NW,W,SW,S,SE,E,NE)*.

Figure 5: (a) Action set *(N,W,S,E)* (b)Action set *(N,NW,W,SW,S,SE,E,NE)*.

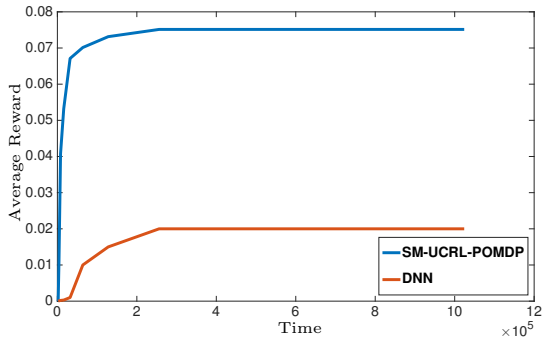


Figure 6: Performances in triple boxes observable environment, Action set *(N,NW,W,SW,S,SE,E,NE)*.

configurations are not possible and there are some constraint, e.g. when the middle box is wall the rests have to be wall as well, for simplicity we do not consider this reduction). We model the environment behavior as POMDP model with number of states equal to 8 *(X=8)*. We apply our SM-UCRL method on this environment and for planning we use same technique that we used in previous subsection. We show how SM-UCRL outperforms DNN with same structure except 30 hidden units at each hidden layer.

During the implementation, we observed that SM-UCRL does not need to estimate the model parameter very well to get to reasonable policy. It comes up with the stochastic and reasonably good policy even from the beginning. On the other hand, we observed that the policy makes balance between moving upward and downward and makes balance between moving rightward and leftward to keep the agent away from the walls. It helps the agent to gain more and wander around the area which has less limitation compared to the area around the walls.

Appendix G. Concentration Bound

Concentration of functions of a HMM We now provide concentration bounds for any matrix valued function $\phi(\cdot)$ over samples drawn from a HMM. This extends the result on scalar functions by Kontorovich et al. (2014).

For any ergodic Markov chain, lets consider ω as its stationary distribution and $f_{1\to t}(x_t|x_1)$ as a distribution over states at time t given initial state x_1. Lets define inverse mixing time $\rho_{mix}(t)$ as follows

$$\rho_{mix}(t) = \sup_{x_1} \|f_{1\to t}(\cdot|x_1) - \omega\|_{TV}$$

Kontorovich et al. (2014) show that this measure can be bounded by

$$\rho_{mix}(t) \leq G\theta^{t-1},$$

where $1 \leq G < \infty$ is *geometric ergodicity* and $0 \leq \theta < 1$ is contraction coefficient of Markov chain.

As before, let $\mathbf{y}^n := [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \mathcal{Y}^n$ denote the sequence of observations from HMM and let $x^n := [x_1, \ldots, x_n] \in \mathcal{X}^n$ denote the sequence of hidden states. We now consider matrix valued function $\Phi : \mathcal{Y}^n \to \mathbb{R}^{d_1 \times d_2}$. It is said to be c-Lipschitz with respect to spectral norm when

$$\sup_{\mathbf{y}^n, \mathbf{y}'^n \in \mathcal{Y}^n} \frac{\|\Phi(\mathbf{y}^n) - \Phi(\mathbf{y}'^n)\|_2}{\|\mathbf{y}^n - \mathbf{y}'^n\|_H} \leq c$$

where $\|\cdot\|_H$ is norm with respect to Hamming metric, and $\mathbf{y}^n, \mathbf{y}'^n$ are any two sequences of sample observations.

Theorem 10 (HMM Concentration Bound) *Consider Hidden Markov Model with finite sequence of n samples $\boldsymbol{y}_i$ as observations from finite observation set $\mathbf{y}^n$ and arbitrary initial state distribution. For any c-Lipschitz matrix valued function $\Phi(\cdot)$, we have*

$$\|\Phi(\mathbf{y}^n) - \mathbb{E}[\Phi(\mathbf{y}^n)]\|_2 \leq G\frac{1 + \frac{1}{\sqrt{2}cn^{\frac{3}{2}}}}{1-\theta}\sqrt{8c^2 n \log(\frac{(d_1+d_2)}{\delta})}$$

with probability at least $1-\delta$, where G is geometric ergodicity constant of corresponding Markov chain, and the $\mathbb{E}[\Phi(\mathbf{y}^n)]$ is expectation over samples of HMM when the initial distribution corresponds to the stationary distribution.

Proof In the Appendix. H ■

Theorem 11 (POMDP Concentration Bound) *Consider Partially Observable Markov Decision Process with finite sequence of $n(l)$ samples $\boldsymbol{y}_i^{(l)}$ for all $i \in \{1, 2, \ldots, n(l)\}$ $\forall l \in [A]$ as observations from finite observation sets $\mathbf{y}^{n(l)}$ and arbitrary initial state distribution. For*

any c-Lipschitz matrix valued function $\Phi^l(\cdot)$ function, we have

$$\left\|\Phi^l(\mathbf{y}^{n(l)}) - \mathbb{E}[\Phi^l(\mathbf{y}^{n(l)})]\right\|_2 \leq G\frac{1+\frac{1}{\sqrt{2}cn^{\frac{3}{2}}}}{1-\theta}\sqrt{8c^2 n\log(\frac{(d_1+d_2)}{\delta})}$$

with probability at least $1-\delta$, where G is geometric ergodicity constant of corresponding Markov chain, and the $\mathbb{E}[\Phi(\mathbf{y}^{n(l)})]$ is expectation over samples of POMDP with middle action l when the initial distribution corresponds to the stationary distribution.

Proof In the Appendix. H ■

Appendix H. Proof of Thms. 10 and 11

The proof is based on the results in Tropp (2012), Kontorovich et al. (2008), and Kontorovich et al. (2014) with minor modifications and applying the following inequality

$$\frac{G}{1-\theta}\sqrt{8c^2\frac{\log(\frac{(d_1+d_2)}{\delta})}{n}} + \frac{2G}{n(1-\theta)} \leq G\frac{1+\frac{1}{\sqrt{2}cn^{\frac{3}{2}}}}{1-\theta}\sqrt{8c^2 n\log(\frac{(d_1+d_2)}{\delta})}$$

here we just bring the sketch of the proof. In Thm. 12, we give the upper confidence bound over $\|\Phi - \mathbb{E}[\Phi]\|_2$ where the expectation is with same initial distribution as it used for Φ. The next step is finding upper bound for difference between $\mathbb{E}[\Phi]$ with arbitrary initial distribution and $\mathbb{E}_{stat}[\Phi]$ with initial distribution equal to stationary distribution. It is clear through Kontorovich et al. (2014) that this quantity is upper bounded by $\sum_i G\theta^{-(i-1)}$ which is upper bounded by $\frac{G}{(1-\theta)}$.

Appendix I. Concentration Bound

Theorem 12 (Matrix Azuma) *Consider Hidden Markov Model with finite sequence of n samples S_i as observations given arbitrary initial states distribution and $c-$ Lipschitz matrix valued function $\Phi : S_1^n \to \mathbb{R}^{d_1\times d_2}$ in dimension d_1 by d_2, then*

$$\|\Phi - \mathbb{E}[\Phi]\|_2 \leq \frac{1}{1-\theta}\sqrt{8c^2 n\log(\frac{(d_1+d_2)}{\delta})}$$

with probability at least $1-\delta$. The $\mathbb{E}[\Phi]$ is given same initial distribution of samples.

Proof The Thm. 7.1 Tropp (2012) presents the upper confidence bound over the summation of matrix random variables. Consider a finite sequence of d by d' matrix Ψ_i, then for variance parameter σ^2 which is upper bound for $\sum_i \left[\Psi_i - \mathbb{E}_{i-1}[\Psi_i]\right]$, $\forall i$

$$\left\|\sum_i \left[\Psi_i - \mathbb{E}_{i-1}[\Psi_i]\right]\right\|_2 \leq \sqrt{8\sigma^2\log(\frac{d+d'}{\delta})}$$

with probability at least $1-\delta$.
For the function Φ, lets define the martingale difference of function Φ as the input random variable with arbitrary initial distribution over states.

$$MD_i(\Phi; S_1^i) = \mathbb{E}[\Phi|S_1^i] - \mathbb{E}[\Phi|S_1^{i-1}]$$

where S_i^j is sub set of samples from $i'th$ position in sequence to $j'th$ one. then the summation over these set of random variable gives $\mathbb{E}[\Phi|S_1^n] - \mathbb{E}[\Phi]$ which is $\Phi(S_1^n) - \mathbb{E}[\Phi]$ and $\mathbb{E}[\Phi]$ is expectation with same initial state distribution . The remaining part is finding σ which is upper bound for $\left\|\sum_i MD_i(\Phi; S_1^i)\right\|_2$ for all possible sequence. Lets define $MD_i(\Phi) = \max_{S_1^i} MD_i(\Phi; S_1^i)$ and through Kontorovich et al. (2008) it is easy to show that $\|MD_i(\Phi)\|_2$ is $c-Lipchitz$ function and it is upper bounded by $cH_{i,n}$. In Kontorovich et al. (2014) it is shown that $H_{i,n}$ is upper bounded by $G\theta(n-i)$. ■

For the case when Φ is symmetric matrix, d_1+d_2 can be reduced to just d and constant 8 can be reduced to 2.
The result in Thm. 12 can be extended to the situation when distribution of next state depends on current state and current observation and even more complicated models like memory-less policy POMDP.

Theorem 13 (Concentration Bound) *Consider finite sequence of multiple views are drawn from memory less policy POMDP with common middle action and their corresponding covariance matrix* $\boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)}$ *for* $\nu, \nu' \in \{1,2,3\}$ *and* $\nu \neq \nu'$. *For simplicity, lets just consider one set of* ν, ν', *one specific middle action, and* n *samples are drawn. Define random variable* $\Phi_i := \frac{1}{N(l)}\left[\mathbb{E}\left[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \middle| S_1^i\right] - \mathbb{E}\left[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \middle| S_1^{i-1}\right]\right]$ *with dimensions* $d_\nu \times d_{\nu'}$ *where* d_ν *and* $d_{\nu'}$ *for* $\nu, \nu' \in \{1,2,3\}$ *are the dimension along the* ν *and* ν' *views.*

$$\left\|\sum_i \Phi_i\right\|_2 = \left\|\frac{1}{N(l)}\sum_t \left[\boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)}\right] - \frac{1}{N(l)}\mathbb{E}[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)}]\right\|_2 \leq \frac{G(\pi)}{1-\theta(\pi)}\sqrt{8\frac{\log\frac{(d_\nu+d_{\nu'})}{\delta}}{N(l)}}$$

with probability at least $1-\delta$.
For tensor case, $\frac{1}{N(l)}\left[\mathbb{E}\left[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \otimes \boldsymbol{v}_{\nu'',t}^{(l)} \middle| S_1^i\right] - \mathbb{E}\left[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \otimes \boldsymbol{v}_{\nu'',t}^{(l)} \middle| S_1^{i-1}\right]\right]$ *where* $[\nu, \nu', \nu'']$ *can be any permutation of set* $\{1,2,3\}$.

$$\left\|\frac{1}{N(l)}\sum_t \left[\boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \otimes \boldsymbol{v}_{\nu'',t}^{(l)}\right] - \frac{1}{N(l)}\mathbb{E}[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \otimes \boldsymbol{v}_{\nu'',t}^{(l)}]\right\|_2 \leq \frac{G(\pi)}{1-\theta(\pi)}\sqrt{8\frac{\log\frac{(d_\nu d_{\nu'}+d_{\nu''})}{\delta}}{N(l)}}$$

with probability at least $1-\delta$.

Proof

For simplicity lets just proof the first claim in Thm. 13 and the proof for the second claim would be followed by same procedure. To proof the Thm. 13 it is needed to bring together the results from Tropp (2012), Kontorovich et al. (2008), Thms. 10 , and 12 and then modify them. The Thm. 7.1 in Tropp (2012) presents following upper confidence

bounds

$$\left\| \frac{1}{N(l)} \sum_t \left[\boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \right] - \frac{1}{N(l)} \mathbb{E}[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)}] \right\|_2 \leq \sqrt{8(\widetilde{\sigma}_{Pairs}^{\nu,\nu'})^2 \log \frac{(d_\nu + d_{\nu'})}{\delta}}$$

with probability at least $1 - \delta$. And

$$\left\| \frac{1}{N(l)} \sum_t \left[\boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \otimes \boldsymbol{v}_{\nu'',t}^{(l)} \right] - \frac{1}{N(l)} \mathbb{E}[\sum_t \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \otimes \boldsymbol{v}_{\nu'',t}^{(l)}] \right\|_2 \leq \sqrt{8(\widetilde{\sigma}_{Triples}^{\nu,\nu',\nu''})^2 \log \frac{(d_\nu d_{\nu'} + d_{\nu''})}{\delta}}$$

with probability at least $1 - \delta$. It is needed to show that $(\widetilde{\sigma}_{Pairs}^{i,i'})^2 \leq \frac{G(\pi)^2}{n(1-\theta(\pi))^2}$ and $(\widetilde{\sigma}_{Triples}^{i,i',i''})^2 \leq \frac{G(\pi)^2}{n(1-\theta(\pi))^2}$.
■

For the function $\Phi : S_1^n \to R^{d_1 \times d_2}$, where S_1^n is a collection of all possible $\{S_1, S_2, \ldots, S_n\}$ with length n. Its martingale difference is defined as follows

$$MD_i(\Phi; S_1^i) = \mathbb{E}[\Phi | S_1^i] - \mathbb{E}[\Phi | S_1^{i-1}]$$

and then $MD_i(\Phi) = \max_{S_1^i} MD_i(\Phi; S_1^i)$.
The upper bound over $\widetilde{\sigma}_{Pairs}^{\nu,\nu'}$ is as follows

$$(\widetilde{\sigma}_{Pairs}^{\nu,\nu'})^2 \leq \left\| \sum_{t=1}^n U_t^2 \right\|_2$$

Where U_t is a fixed sequence of matrices which follows $MD_t \preceq U_t$ for all possible MD_t and $\forall t$. This bound over Triple tensor can be derived after matricizing the martingale difference. Next step is to upper bound the $\left\| \sum_t U_t^2 \right\|_2$.
Lets define new set of variables; given each action (middle action) $a_i = l$ there are the following set of variables; $B^{i|l}$ is collection of
$\boldsymbol{y}_{p(i,l)-1}, a_{p(i,l)-1}, r_{p(i,l)-1}, \boldsymbol{y}_{p(i,l)}, r_{p(i,l)}, \boldsymbol{y}_{p(i,l)+1}$,
where $B^{i|l}$ is $i'th$ triple with middle action equal to l and $p(i,l)$ is its corresponding position in the original sequence. Lets define variable $S^{i|l}$, which is consequence of four hidden states $x_{p(i,l)-1}, x_{p(i,l)}, x_{p(i,l)+1}, x_{p(i,l)+2}$. The variables $B_i^{j|l}$ and $S_i^{j|l}$, for $i \leq j$, are corresponding to set of consecutive $i \to j$ triple views and quadruple hidden states. Note that this is the time to define mixing coefficients.

$$\eta_{i,j}^{(l)}(b_1^{i-1|l}, \varrho, \varrho') := \left\| \mathbb{P}(B_j^{N(l)|l} | B_1^{i|l} = b_1^{i-1|l}, \varrho, l) - \mathbb{P}(B_j^{N(l)|l} | B_1^{i-1|l} = b_1^{i-1|l}, \varrho', l) \right\|_{TV}$$

where TV is total variation distance between distributions and

$$\bar{\eta}_{i,j}^{(l)} := \sup_{b_1^{i-1|l}, \varrho, \varrho'} \eta_{ij}^{(l)}(b_1^{i-1|l}, \varrho, \varrho')$$

where $\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho, l)$ and $\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho'.l)$ are nonzero for all possible input variables. Then for $\Delta_{N(l)}$

$$(\Delta_{N(l)})_{i,j} = \begin{cases} 1 & \text{if } i = j \\ \bar{\eta}_{i,j}^{(l)} & \text{if } i < j \\ 0 & otherwise. \end{cases}$$

and $H_{n(l),i} = 1 + \bar{\eta}_{i,i+1}^{(l)} + \ldots + \bar{\eta}_{i,n}^{(l)}$

Martingale Difference . To upper bound for $\widetilde{\sigma}_{Pairs}^{\nu,\nu'}$, it is enough to upper bound $\left\|\sum_{t=1}^n U_t^2\right\|_2$ or directly upper bound $\left\|\sum_{t=1}^n MD_t^2\right\|_2$ for all possible sequence of samples. The result in Kontorovich et al. (2008) shows that this is upper bounded by $\sum_{i=1}^{n(l)} \|MD_i(\Phi))\|_2^2$ and each $\|MD_i(\Phi))\|_2 \leq cH_{n,i}$ when the $\|\Phi\|_2$ is $c - Lipschitz$.

In addition, it is obvious that for the class of moment functions with elements in $[0, 1]$ the c is upper bounded by $\frac{1}{N(l)}$ for the purpose of this paper. The remaining shows the upper bound over $H_{n,i}$ and then $\sum_{i=1}^{n(l)} H_{n,i}^2$.

Lemma 14 *The function $H_{n,i}$ is upper bounded by $\frac{G(\pi)}{1-\theta(\pi)}$ and then $\sum_{i=1}^n (cH_{n,i})^2 \leq nc^2 \frac{G^2(\pi)}{(1-\theta(\pi))^2} \leq \frac{G^2(\pi)}{n(1-\theta(\pi))^2}$*

Proof As it mentioned, $H_{n(l),i} = 1 + \bar{\eta}_{i,i+1}^{(l)} + \ldots + \bar{\eta}_{i,n}^{(l)}$, and it is needed to find the upper bound over $1 + \bar{\eta}_{i,i+1}^{(l)} + \ldots + \bar{\eta}_{i,n}^{(l)}$

$$\begin{aligned} &\eta_{ij}^{(l)}(b_1^{i-1|l}, \varrho, \varrho') \\ &= \frac{1}{2} \sum_{b_j^{N(l)|l}} |\mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l} | B_1^{i|l} = b_1^{i-1|l}, \varrho, l) - \mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l} | B_1^{i-1|l} = b_1^{i-1|l}, \varrho', l)| \end{aligned}$$

For the first part

$$\begin{aligned} &\mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l} | B_1^{i|l} = b_1^{i-1|l}, \varrho, l) \\ &= \sum_{s_1^{i|l}, s_j^{N(l)|l}} \mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l}, S_1^{i|l} = s_1^{i|l}, S_j^{N(l)|l} = s_j^{N(l)|l} | B_1^{i|l} = b_1^{i-1|l}, \varrho, l) \end{aligned}$$

Lets assume, for simplicity, that the hidden states on $s^{i|l}$ do not have overlap with states on $s^{i-1|l}$ and $s^{i+1|l}$.

$$\begin{aligned}\mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l} | B_1^{i|l} = b_1^{i-1|l}, \varrho, l) = \\ \sum_{s_1^{i|l}, s_j^{N(l)|l}} \mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l}, B_1^{i-1|l} = b_1^{i-1|l}, \varrho, l | S_1^{i|l} = s_1^{i|l}, S_j^{N(l)|l} = s_j^{N(l)|l}) \\ \mathbb{P}(S_1^{i|l} = s_1^{i|l}, S_j^{N(l)|l} = s_j^{N(l)|l}) \frac{1}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho, l)} \\ = \sum_{s_1^{i|l}, s_j^{N(l)|l}} \mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l}, l | S_j^{N(l)|l} = s_j^{N(l)|l}) \mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho, l | S_1^{i|l} = s_1^{i|l}, S) \\ \mathbb{P}(S_1^{i|l} = s_1^{i|l}, S_j^{N(l)|l} = s_j^{N(l)|l}) \frac{1}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho, l)}\end{aligned}$$

with this representation

$$\begin{aligned}\eta_{ij}^{(l)}(b_1^{i-1|l}, \varrho, \varrho') = \frac{1}{2} \sum_{b_j^{N(l)|l}} | \sum_{s_1^{i|l}, s_j^{N(l)|l}} \mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l} | S_j^{N(l)|l} = s_j^{N(l),l|l}) \mathbb{P}(S_1^{i|l} = s_1^{i|l}, S_j^{N(l)|l} = s_j^{N(l)|l}) \\ \mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho', l | S_1^{i|l} = s_1^{i|l}) \big(\frac{\mathbb{P}(\varrho, l | S^{i|l} = s^{i|l})}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho, l)} - \frac{\mathbb{P}(\varrho', l | S^{i|l} = s^{i|l})}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho', l)} \big) |\end{aligned}$$

$$\begin{aligned}\eta_{ij}^{(l)}(b_1^{i-1|l}, \varrho, \varrho') \leq \frac{1}{2} \sum_{s^{j|l}} | \sum_{s_1^{i|l}} \mathbb{P}(B_j^{N(l)|l} = b_j^{N(l)|l}, l | S_j^{N(l)|l} = s_j^{N(l)|l}) \mathbb{P}(S_1^{i|l} = s_1^{i|l}, S_j^{N(l)|l} = s_j^{N(l)|l}) \\ \mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho', l | S_1^{i|l} = s_1^{i|l}) \big(\frac{\mathbb{P}(\varrho, l | S^{i|l} = s^{i|l})}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho, l)} - \frac{\mathbb{P}(\varrho', l | S^{i|l} = s^{i|l})}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho', l)} \big) |\end{aligned}$$

$$\eta_{ij}^{(l)}(b_1^{i-1|l}, \varrho, \varrho') \leq \frac{1}{2} \sum_{x^{p(j)-1|l}} | \sum_{s_1^{i|l}} \mathbb{P}(S_1^{i|l} = s_1^{i|l}) \mathbb{P}(x^{p(j)-1|l} | x^{p(i,l)+2|l}) \mathbb{P}(B_1^{i-1|l} = b_1^{i-1|l}, l | S_1^{i|l} = s_1^{i|l}) q(s^{i|l}) |$$

where

$$q(v, l) := \frac{\mathbb{P}(\varrho, l | S^{i|l} = v)}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho, l)} - \frac{\mathbb{P}(\varrho', l | S^{i|l} = v)}{\mathbb{P}(B_1^{i|l} = b_1^{i-1|l}, \varrho', l)}$$

then

$$\begin{aligned}\eta_{ij}^{(l)}(b_1^{i-1|l},\varrho,\varrho') &\le \frac{1}{2}\sum_{x^{p(j)-1|l}}|\sum_{s^{i|l}}\mathbb{P}(x^{p(j)-1|l}|x^{p(i,l)+2|l})h(s^{i|l},l)|\\ &\le \frac{1}{2}\sum_{x^{p(j)-1|l}}|\sum_{x^{p(i,l)+1|l}}\mathbb{P}(x^{p(j)-1|l}|x^{p(i,l)+2|l})\sum_{x^{p(i,l)+1|l},x^{p(i,l)|l},x^{p(i,l)-1|l}}h(s^{i|l},l)|\\ &\le \frac{1}{2}\sum_{x^{p(j)-1|l}}|\sum_{x^{p(i,l)+2|l}}\mathbb{P}(x^{p(j)-1|l}|x^{p(i,l)+2|l})\bar{h}(x^{p(i,l)+2|l},l)|\end{aligned}$$

where

$$h(v,l) := \sum_{s_1^{i-1|l}}\mathbb{P}(S_1^{i|l}=s_1^{i|l})\mathbb{P}(B_1^{i-1|l}=b_1^{i-1|l},l|S_1^{i|l}=s_1^{i|l})q(v,l)$$

$$\bar{h}(x^{p(i,l)+2|l},l) := \sum_{x^{p(i,l)+1|l},x^{p(i,l)|l},x^{p(i,l)-1|l}}h(s^{i|l},l)$$

as a consequence

$$\eta_{ij}^{(l)}(b_1^{i-1|l},\varrho,\varrho') \le \left\|\frac{1}{2}\bar{h}^\top P^{i,j}\right\|_1$$

where $P^{i,j}=\mathbb{P}(x^{p(j)-1|l}|x^{p(i,l)+2|l})$. Through Lemma A.2 in Kontorovich et al. (2008) and Kontorovich et al. (2014), when $\sum_x\bar{h}(x,l)=0$ and $\frac{1}{2}\left\|\bar{h}\right\|_1\le 1$, it is clear that $\eta_{ij}^{(l)}(b_1^{i-1|l},\varrho,\varrho')$, and also $\bar{\eta}_{ij}^{(l)}$ can be bounded by $\frac{1}{2}\left\|\bar{h}^\top\right\|_1 G(\pi)\theta(\pi)^{p(j,l)-p(i,l)-4}$. To verify $\sum_x\bar{h}=0$ and $\frac{1}{2}\left\|\bar{h}\right\|_1\le 1$

$$\begin{aligned}\sum_{x^{p(i,l)+2|l}}\bar{h}(x^{p(i,l)+2|l},l) &= \sum_{x^{p(i,l)+2|l}}\sum_{x^{p(i,l)+1|l},x^{p(i,l)|l},x^{p(i,l)-1|l}}h(s^{i|l},l)=\sum_{s^{i|l}}h(s^{i|l},l)\\ &=\sum_{s_1^{i|l}}\mathbb{P}(S_1^{i|l}=s_1^{i|l})\mathbb{P}(B_1^{i-1|l}=b_1^{i-1|l},l|S_1^{i|l}=s_1^{i|l})q(s^{i|l},l)\\ &=\sum_{s_1^{i|l}}\mathbb{P}(S_1^{i|l}=s_1^{i|l})\mathbb{P}(B_1^{i|l}=b_1^{i-1|l},l|S_1^{i|l}=s_1^{i|l})\left(\frac{\mathbb{P}(\varrho,l|S^{i|l}=s^{i|l})}{\mathbb{P}(B_1^{i|l}=b_1^{i-1|l},\varrho,l)}-\frac{\mathbb{P}(\varrho',l|S^{i|l}=s^{i|l})}{\mathbb{P}(B_1^{i-1|l}=b_1^{i-1|l},\varrho',l)}\right)\end{aligned}$$

For the first part of parenthesis

$$\sum_{s_1^{i|l}}\mathbb{P}(S_1^{i|l}=s_1^{i|l})\mathbb{P}(B_1^{i-1|l}=b_1^{i-1|l},l|S_1^{i|l}=s_1^{i|l})\left(\frac{\mathbb{P}(\varrho,l|S^{i|l}=s^{i|l})}{\mathbb{P}(B_1^{i|l}=b_1^{i-1|l},\varrho,l)}\right)=1$$

and same for the second one. This shows the conditions for the Lemma A.2 in Kontorovich et al. (2008), $\sum_x\bar{h}(x,l)=0$ and $\frac{1}{2}\left\|\bar{h}\right\|_1\le 1$, are met.
The presented proof is for the case of non-overlapped with states of $S^{i|l}$, for the other cases,

the overlapped situation, the proof is pretty much similar to the non-overlapped case. Now, it is the time to upper bound $\left\|\Delta_{N(l)}\right\|_\infty$.

$$H_{n,i} = 1 + \sum_{j=i}^{N(l)} \bar{\eta}_{ij}^{(l)} \leq 1 + \max_i \sum_{j>i} \bar{\eta}_{ij}^{(l)} \leq G(\pi) \sum_{i=0}^{n(l)} \theta(\pi)^{p(i,l)} \leq G(\pi) \sum_{i=0}^{n(l)} \theta(\pi)^i$$
$$= G(\pi) \frac{1-\theta(\pi)^{n(l)}}{1-\theta(\pi)} \leq \frac{G(\pi)}{1-\theta(\pi)}$$

■

Define $\mathbb{E}_{stat}$ as the expectation with initial distribution equals to stationary distribution. Generally, in tensor decomposition, we are interested in

$$\left\| \frac{1}{N(l)} \sum_t \left[\boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)} \right] - \frac{1}{N(l)} \mathbb{E}_{stat} [\sum_i \boldsymbol{v}_{\nu,t}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)}] \right\|_2$$
$$\left\| \frac{1}{N(l)} \sum_t \left[\boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)} \otimes \boldsymbol{v}_{\nu'',i}^{(l)} \right] - \frac{1}{N(l)} \mathbb{E}_{stat} [\sum_i \boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)} \otimes \boldsymbol{v}_{\nu'',i}^{(l)}] \right\|_2$$

instead of

$$\left\| \frac{1}{N(l)} \sum_t \left[\boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)} \right] - \frac{1}{N(l)} \mathbb{E} [\sum_i \boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)}] \right\|_2$$
$$\left\| \frac{1}{N(l)} \sum_t \left[\boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)} \otimes \boldsymbol{v}_{\nu'',i}^{(l)} \right] - \frac{1}{N(l)} \mathbb{E} [\sum_i \boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',t}^{(l)} \otimes \boldsymbol{v}_{\nu'',i}^{(l)}] \right\|_2$$

which are derived through Thm. 13. To come up with the upper confidence bound over the above mentioned interesting deviation, it is needed to derive the upper bound for deviation over expectation with arbitrary initial state distribution and expectation with stationary distribution over initial states, for simplicity, lets just derive the bound for second order moment.

$$\left\| \frac{1}{N(l)} \mathbb{E}_n [\sum_i \boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)}] - \frac{1}{N(l)} \mathbb{E}_{stat} [\sum_i \boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)}] \right\|_2$$

As the bound ϵ_i over deviation from stationary distribution of Markov chain follows $\epsilon = G(\pi)\theta(\pi)^{-i}$. It results in

$$\left\| \frac{1}{N(l)} \mathbb{E} [\sum_i \boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)}] - \frac{1}{N(l)} \mathbb{E}_{stat} [\sum_i \boldsymbol{v}_{\nu,i}^{(l)} \otimes \boldsymbol{v}_{\nu',i}^{(l)}] \right\|_2 \leq 2 \frac{G(\pi)}{N(l)(1-\theta(\pi))}$$

which is negligible compared to $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{n}})$

Corollary 15 *These result hold for pure HMM model. For tensor case, where* $[\nu, \nu', \nu'']$ *is any permutation of set* $\{1, 2, 3\}$.

$$\left\| \frac{1}{N(l)} \sum_t \left[\boldsymbol{v}_{\nu,t} \otimes \boldsymbol{v}_{\nu',t} \right] - \frac{1}{N} \mathbb{E}[\sum_t \boldsymbol{v}_{\nu,t} \otimes \boldsymbol{v}_{\nu',t}] \right\|_2 \leq \frac{G}{1-\theta} \sqrt{8 \frac{\log \frac{(d_\nu + d_{\nu'})}{\delta}}{N(l)}}$$

with probability at least $1 - \delta$ *and*

$$\left\| \frac{1}{N} \sum_t \left[\boldsymbol{v}_{\nu,t} \otimes \boldsymbol{v}_{\nu',t} \otimes \boldsymbol{v}_{\nu'',t} \right] - \frac{1}{N} \mathbb{E}[\sum_t \boldsymbol{v}_{\nu,t} \otimes \boldsymbol{v}_{\nu',t} \otimes \boldsymbol{v}_{\nu'',t}] \right\|_2 \leq \frac{G}{1-\theta} \sqrt{8 \frac{\log \frac{(d_\nu d_{\nu'} + d_{\nu''})}{\delta}}{N}}$$

with probability at least $1 - \delta$. *The deviation bound is as follows*

$$\left\| \frac{1}{N} \mathbb{E}[\sum_i \boldsymbol{v}_{\nu,i} \otimes \boldsymbol{v}_{\nu',i}] - \frac{1}{N} \mathbb{E}_{stat}[\sum_i \boldsymbol{v}_{\nu,i} \otimes \boldsymbol{v}_{\nu',i}] \right\|_2 \leq 2 \frac{G}{N(1-\theta)}$$

Proof Through Kontorovich et al. (2008) and Kontorovich et al. (2014) it is shown that for the HMM models, the value of $H_{n,i}$ is bounded by $\frac{G}{1-\theta}$ and then it means that the corresponding martingale difference is bounded by $\frac{cG}{1-\theta}$. In the consequence, the $\sigma^2_{HMM,\Phi}$ is bounded by $\frac{G^2}{n(1-\theta)^2}$. ■

Appendix J. Whitening and Symmetrization Bound

Theorem 16 (Whitening, Symmetrization and De-Whitening Bound) *Pick any* δ. *Then for HMM model with* k *hidden state and its multi-view representation with factor matrices* A_1, A_2, A_3, *and finite observation set with dimension* d_1, d_2, d_3 *corresponds to multi-view representation, when the number of samples with arbitrary initial state distribution satisfies*

$$n \geq \left(\frac{G^{\frac{2\sqrt{2}+1}{1-\theta}}}{\omega_{\min} \min_i \{\sigma_k^2(A_i)\}} \right)^2 \log(2 \frac{(d_1 d_2 + d_3)}{\delta}) \max \left\{ \frac{16 k^{\frac{1}{3}}}{C^{\frac{2}{3}} \omega_{\min}^{\frac{1}{3}}}, 4, \frac{2\sqrt{2}k}{C^2 \omega_{\min} \min_i \{\sigma_k^2(A_i)\}} \right\}$$

for some constant C . *After tensor symmetrizing and whitening, with low order polynomial computation complexity, the robust power method in Anandkumar et al. (2012) yield to whitened component of the views* $\mu_1, \ldots, \mu_k$, *such that with probability at least* $1 - \delta$, *we have*

$$\|\mu_j - (\widehat{\mu}_j))\|_2 \leq 18 \epsilon_M$$

for $j \in \{1, \dots, k\}$ *up to permutation and*

$$\epsilon_M \leq \frac{2\sqrt{2}G\frac{2\sqrt{2}+1}{1-\theta}\sqrt{\frac{\log(\frac{2(d_1d_2+d_3)}{\delta})}{n}}}{(\omega_{\min}^{\frac{1}{2}}\min_i\{\sigma_k(A_i))\}^3} + \frac{\left(4G\frac{2\sqrt{2}+1}{1-\theta}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}}\right)^3}{(\min_i\{\sigma_k(A_i)\})^6\omega_{\min}^{3.5}}$$

Therefore

$$\left\|(A_i)(:,j) - (\widehat{A}_i)_{:,j}\right\|_2 \leq \epsilon_3$$

for $i \in \{1,2,3\}$, $j \in \{1, \dots, k\}$ *up to permutation and*

$$\epsilon_3 := G\frac{4\sqrt{2}+4}{(\omega_{\min})^{\frac{1}{2}}(1-\theta)}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}} + \frac{8\epsilon_M}{\omega_{\min}}$$

Proof Appendix K. ■

Appendix K. Whitening and Symmetrization Bound Proof

Proof of Thm. 16

In Appendix I, the upper confidence bounds for deviation between empirical pairs matrices and tensor from their original ones are derived. As it is shown in Song et al. (2013) and Anandkumar et al. (2014) for multi-view models with factors $A_1 \in \mathbb{R}^{d_1 \times k}$, $A_2 \in \mathbb{R}^{d_2 \times k}$, $A_3 \in \mathbb{R}^{d_3 \times k}$ (three view model with k hidden states), to derive the factor matrices, applying tensor decomposition method is one of the most efficient way. They show that for tensor decomposition method, it is needed to first; symmetrize the initial raw empirical tensor and then whiten it to get orthogonal symmetric tensor. It is well known that orthogonal symmetric tensors have unique eigenvalues and eigenvectors and can be obtained thorough power method Anandkumar et al. (2014).

Without loss of generality, lets assume we are interested in A_3, the derivation can be done for other view by just permuting them. Assume tensor $M_3 = \mathbb{E}[\boldsymbol{v}_1 \otimes \boldsymbol{v}_2 \otimes \boldsymbol{v}_3]$ is triple raw cross correlation between views, and matrix R_2 and R_3 are rotation matrices for rotating second and third view to first view. It means that it results in symmetric tensor $M_3(R_1, R_2, I)$. Through Anandkumar et al. (2014) these rotation matrices are as follow

$$R_1 = \mathbb{E}[\boldsymbol{v}_3 \otimes \boldsymbol{v}_2]\mathbb{E}[\boldsymbol{v}_1 \otimes \boldsymbol{v}_2]^{-1}$$
$$R_2 = \mathbb{E}[\boldsymbol{v}_3 \otimes \boldsymbol{v}_1]\mathbb{E}[\boldsymbol{v}_2 \otimes \boldsymbol{v}_1]^{-1}$$

Define second order moment as $M_2 = \mathbb{E}[\boldsymbol{v}_1 \otimes \boldsymbol{v}_2]$ and its symmetrized version as $M_2(R_1, R_2)$. Lets $W \in \mathbb{R}^{d_1 \times k}$ be a linear transformation such that

$$M_2(R_1W, R_2W) = W^\top M_2(R_1, R_2)W = I$$

where I is $k \times k$ identity matrix. Then the matrix $W = U\Lambda^{-\frac{1}{2}}$ where $M_2(R_1, R_2) = U\Lambda V^\top$ is singular value decomposition of $M_2(R_1, R_2)$. It is well known result that tensor $M_3(W_1, W_2, W_3) = M_3(R_1W, R_2W, W)$ is symmetric orthogonal tensor and ready for power iteration to compute the unique $(A_3)_i \ \forall i \in [1 \dots k]$.

To come up with upper confidence bound over $\left\|(\widehat{A}_3)_i - (A_3)_i\right\|_2$ (columns of factor matrices), it is needed to aggregate the different source of error. This deviation is due to empirical average error which derived in I, symmetrizing error, and whitening error.

To obtain the upper bound over the aggregated error, lets apply the following proof technique. It is clear that for matrix $\widehat{M}_2$, we have $\widehat{W}^\top \widehat{R}_1^\top \widehat{M}_2 \widehat{R}_2 \widehat{W} = I$. lets assume, matrices B, D, B as a singular value decomposition of $\widehat{W}^\top \widehat{R}_1^\top M_2 \widehat{R}_2 \widehat{W} = BDB^\top$. Then it is easy to show that for $\widetilde{W}_1 = \widehat{W}_1 B D^{-\frac{1}{2}} B^\top$, $\widetilde{W}_2 = \widehat{W}_2 B D^{-\frac{1}{2}} B^\top$, and $\widetilde{W}_3 = \widehat{W} B D^{-\frac{1}{2}} B^\top$ then

$$\widetilde{W}_2^\top M_2 \widetilde{W}_1 = I$$

and then the ϵ_M

$$\begin{aligned}\epsilon_M &= \left\|M_3(\widetilde{W}_1, \widetilde{W}_2, \widetilde{W}_3) - \widehat{M}_3(\widehat{W}_1, \widehat{W}_2, \widehat{W}_3)\right\|_2 \\ &\leq \left\|(\widehat{M}_3 - M_3)(\widehat{W}_1, \widehat{W}_2, \widehat{W}_3)\right\|_2 + \left\|M_3(\widehat{W}_1 - \widetilde{W}_1, \widehat{W}_2 - \widetilde{W}_2, \widehat{W}_3 - \widetilde{W}_3)\right\|_2\end{aligned}$$

It means

$$\epsilon_M \leq \left\|M_3 - \widehat{M}_3\right\|_2 \left\|\widehat{W}_1\right\|_2 \left\|\widehat{W}_2\right\|_2 \left\|\widehat{W}_3\right\|_2 + \left\|M_3(\widehat{W}_1 - \widetilde{W}_1, \widehat{W}_2 - \widetilde{W}_2, \widehat{W}_3 - \widetilde{W}_3)\right\|_2$$

Lets assume $U_{1,2}\Lambda_{1,2}V_{1,2}^\top = M_2$ is singular value decomposition of matrix M_2. From $W^\top R_1^\top M_2 R_2 W = I$ and the fact that $W_1 U_{1,2}\Lambda^{\frac{1}{2}} = W_2 V_{1,2}\Lambda^{\frac{1}{2}}$ which are the square root of matrix M_2 and to be able to learn all factor matrices we can show that $\|W_i\|_2 \leq \frac{1}{\min_i \sigma_k(A_i Diag(\omega)^{\frac{1}{2}})} \leq \frac{1}{\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i)}$ for $i \in \{1, 2, 3\}$. Now, it is clear to say, when $\left\|\widehat{M}_2 - M_2\right\|_2 \leq 0.5\sigma_k(M_2)$ then $\left\|\widehat{W}_i\right\|_2 \leq \frac{\sqrt{2}}{\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i)}$ for $i \in \{1, 2, 3\}$ and

$$\left\|M_3 - \widehat{M}_3\right\|_2 \left\|\widehat{W}_1\right\|_2 \left\|\widehat{W}_2\right\|_2 \left\|\widehat{W}_3\right\|_2 \leq \frac{2\sqrt{2}\left\|\widehat{M}_3 - M_3\right\|_2}{(\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i))^3}$$

To bound the second term in ϵ_M

$$\left\|M_3(\widehat{W}_1 - \widetilde{W}_1, \widehat{W}_2 - \widetilde{W}_2, \widehat{W}_3 - \widetilde{W}_3)\right\|_2 \leq \frac{1}{\sqrt{\omega_{\min}}} \prod_{i=1}^{3} \left\|Diag(\omega)^{\frac{1}{2}} A_i^\top (\widehat{W}_i - \widetilde{W}_i)\right\|_2$$

then

$$\left\|Diag(\omega)^{\frac{1}{2}} A_i (\widehat{W}_i - \widetilde{W}_i)\right\|_2 = \left\|Diag(\omega)^{\frac{1}{2}} A_i^\top \widetilde{W}_i (BD^{\frac{1}{2}} B^\top - I)\right\|_2 \leq \left\|Diag(\omega)^{\frac{1}{2}} A_i^\top \widetilde{W}_i\right\|_2 \left\|(D^{\frac{1}{2}} - I)\right\|_2$$

We have that $\left\|Diag(\omega)^{\frac{1}{2}}A_i^\top \widetilde{W}_i\right\|_2 = 1$. Now we control $\left\|(D^{\frac{1}{2}} - I)\right\|_2$. Let $\widetilde{E} := M_2 - F_k$ where $F = \widehat{M}_2$, and F_k is its restriction to top-k singular values. Then, we have $\left\|\widetilde{E}\right\|_2 \leq \left\|\widehat{M}_2 - M_2\right\|_2 + \sigma_{k+1}(F) \leq 2\left\|\widehat{M}_2 - M_2\right\|_2$. We now have

$$\left\|(D^{\frac{1}{2}} - I)\right\|_2 \leq \left\|(D^{\frac{1}{2}} - I)(D^{\frac{1}{2}} + I)\right\|_2 \leq \|(D - I)\|_2 = \left\|(BDB^\top - I)\right\|_2 = \left\|(\widehat{W}_1^\top M_2 \widehat{W}_2 - I)\right\|_2 \tag{45}$$

$$= \left\|(\widehat{W}_1^\top \widetilde{E} \widehat{W}_2)\right\|_2 \leq \left\|\widehat{W}_1\right\|_2 \left\|\widehat{W}_2\right\|_2 2\left\|\widehat{M}_2 - M_2\right\|_2 \leq \frac{4\left\|\widehat{M}_2 - M_2\right\|_2}{(\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i))^2} \tag{46}$$

As a conclusion it is shown that

$$\epsilon_M \leq \frac{2\sqrt{2}\left\|\widehat{M}_3 - M_3\right\|_2}{(\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i))^3} + \frac{\left(\frac{4\left\|\widehat{M}_2 - M_2\right\|_2}{(\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i))^2}\right)^3}{\sqrt{\omega_{\min}}} \tag{47}$$

when $\left\|\widehat{M}_2 - M_2\right\|_2 \leq 0.5\sigma_k(M_2)$.
Through Appendix I, the followings hold

$$\left\|M_2 - \widehat{M}_2\right\|_2 \leq G\frac{2\sqrt{2}+1}{1-\theta}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}}$$

$$\left\|M_3 - \widehat{M}_3\right\|_2 \leq G\frac{1+\frac{1}{\sqrt{8}n^{\frac{1}{2}}}}{1-\theta}\sqrt{8\frac{\log(\frac{2(d_1d_2+d_3)}{\delta})}{n}}$$

with probability at least $1-\delta$. It is followed by

$$\epsilon_M \leq \frac{2\sqrt{2}G\frac{2\sqrt{2}+1}{1-\theta}\sqrt{\frac{\log(\frac{2(d_1d_2+d_3)}{\delta})}{n}}}{(\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i))^3} + \frac{\left(4G\frac{2\sqrt{2}+1}{1-\theta}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}}\right)^3}{(\min_i \sigma_k(A_i))^6 \omega_{\min}^{3.5}}$$

with probability at least $1-\delta$. To this result holds, it is required $\left\|\widehat{M}_2 - M_2\right\|_2 \leq 0.5\sigma_k(M_2)$ and from Anandkumar et al. (2012) that $\epsilon_M \leq \frac{C_1}{\sqrt{k}}$. Then for the first requirement

$$n \geq \left(\frac{G\frac{2\sqrt{2}+1}{1-\theta}}{0.5(\omega_{\min}^{\frac{1}{2}} \min_i \sigma_k(A_i))^2}\right)^2 \log(2\frac{(d_1+d_2)}{\delta})$$

and for the second requirement $\epsilon_M \leq \frac{C_1}{\sqrt{k}}$ to be hold it is enough that each term in Eq 47 is upper bounded by $\frac{C}{\sqrt{k}}$ for some constant C.

$$\frac{C}{\sqrt{k}} \geq \frac{2\sqrt{2}G^{\frac{2\sqrt{2}+1}{1-\theta}}\sqrt{\frac{\log(\frac{2(d_1d_2+d_3)}{\delta})}{n}}}{(\omega_{\min}^{\frac{1}{2}}\min_i \sigma_k(A_i))^3}$$

then

$$n \geq \left(\frac{2\sqrt{2}G^{\frac{2\sqrt{2}+1}{1-\theta}}}{C(\omega_{\min}^{\frac{1}{2}}\min_i \sigma_k(A_i))^3}\right)^2 k\log(\frac{2(d_1d_2+d_3)}{\delta})$$

and for the second part

$$\frac{C}{\sqrt{k}} \geq \frac{\left(4G^{\frac{2\sqrt{2}+1}{1-\theta}}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}}\right)^3}{(\min_i \sigma_k(A_i))^6\omega_{\min}^{3.5}}$$

$$n \geq \left(\frac{4k^{\frac{1}{6}}G^{\frac{2\sqrt{2}+1}{1-\theta}}}{C^{\frac{1}{3}}(\min_i \sigma_k(A_i))^2\omega_{\min}^{\frac{3.5}{3}}}\right)^2 \log(2\frac{(d_1+d_2)}{\delta})$$

It means it is enough that

$$n \geq \left(\frac{G^{\frac{2\sqrt{2}+1}{1-\theta}}}{\omega_{\min}\min_i \sigma_k^2(A_i)}\right)^2 \max\left\{\log(2\frac{(d_1+d_2)}{\delta})\max\left\{\frac{16k^{\frac{1}{3}}}{C^{\frac{2}{3}}\omega_{\min}^{\frac{1}{3}}}, 4\right\}, \log(2\frac{(d_1d_2+d_3)}{\delta})\frac{2\sqrt{2}k}{C^2\omega_{\min}\min_i \sigma_k^2(A_i)}\right\}$$

which can be reduced to

$$n \geq \left(\frac{G^{\frac{2\sqrt{2}+1}{1-\theta}}}{\omega_{\min}\min_i \sigma_k^2(A_i)}\right)^2 \log(2\frac{(d_1d_2+d_3)}{\delta})\max\left\{\frac{16k^{\frac{1}{3}}}{C^{\frac{2}{3}}\omega_{\min}^{\frac{1}{3}}}, 4, \frac{2\sqrt{2}k}{C^2\omega_{\min}\min_i \sigma_k^2(A_i)}\right\}$$

In Anandkumar et al. (2014) it is shown that when $\epsilon_M = \left\|M_3(W_1,W_2,W_3) - \widehat{M}_3(\widehat{W}_1,\widehat{W}_2,\widehat{W}_3)\right\|$ then the robust power method in Anandkumar et al. (2012) decomposes the tensor and comes up with set $\widehat{\lambda}_i$ and orthogonal $\widehat{\mu}_i$ where

$$\left\|M_3(W_1,W_2,W_3) - \sum_i^k \widehat{\lambda}_i\widehat{\mu}_i^{\otimes^3}\right\|_2 \leq 55\epsilon_M$$

$$\left\|\omega_i^{-1/2}\mu_i - \widehat{\lambda}_i\widehat{\mu}_i\right\|_2 \leq 8\epsilon_M\omega_i^{-1/2}$$

and

$$\left|\omega_i^{-1/2} - \widehat{\lambda}_i\right| \leq 5\epsilon_M \tag{48}$$

It can be verified that

$$\|\mu_i - \widehat{\mu}_i\|_2 \leq 18\epsilon_M.$$

Proof In order to simplify the notation, in the following we use $\mu = \mu_i$, $\omega = \omega_i$, and $\zeta = \omega_i^{-1/2}$, similar terms for the estimated quantities. From above mentioned bound, we have

$$\left\|\zeta\mu - \widehat{\zeta}\widehat{\mu}\right\|_2 = \left\|\zeta(\mu - \widehat{\mu}) - (\widehat{\zeta} - \zeta)\widehat{\mu}\right\|_2 \leq 8\zeta\epsilon_3(l).$$

We take the square of the left hand side and we obtain

$$\begin{aligned}\left\|\zeta(\mu - \widehat{\mu}) - (\widehat{\zeta} - \zeta)\widehat{\mu}\right\|_2^2 &= \zeta^2\|\mu - \widehat{\mu}\|_2^2 + (\widehat{\zeta} - \zeta)^2\|\widehat{\mu}\|_2^2 - 2\zeta(\zeta - \widehat{\zeta})\sum_{s=1}^{d}[\widehat{\mu}]_s([\mu]_s - [\widehat{\mu}]_s) \\ &\geq \zeta^2\|\mu - \widehat{\mu}\|_2^2 - 2\zeta|\zeta - \widehat{\zeta}|\left|\sum_{s=1}^{d}[\widehat{\mu}]_s([\mu]_s - [\widehat{\mu}]_s)\right| \\ &\geq \zeta^2\|\mu - \widehat{\mu}\|_2^2 - 2\zeta|\zeta - \widehat{\zeta}|\|\widehat{\mu}\|_2\|\mu - \widehat{\mu}\|_2,\end{aligned}$$

where in the last step we used the Cauchy-Schwarz inequality. Thus we obtain the second-order equation

$$\zeta\|\mu - \widehat{\mu}\|_2^2 - 2(\zeta - \widehat{\zeta})\|\widehat{\mu}\|_2\|\mu - \widehat{\mu}\|_2 \leq 64\zeta\epsilon_3(l)^2.$$

Solving for $\|\mu - \widehat{\mu}\|_2$ we obtain

$$\|\mu - \widehat{\mu}\|_2 \leq \frac{|\zeta - \widehat{\zeta}|\|\widehat{\mu}\|_2 + \sqrt{(\zeta - \widehat{\zeta})^2\|\widehat{\mu}\|_2^2 + 64\zeta^2\epsilon_M^2}}{\zeta}.$$

Now we can use the bound in Eq. 48 and the fact that $\|\widehat{\mu}\|_2 \leq \|\widehat{\mu}\|_1 \leq 1$ since $\widehat{\mu}$ is a probability distribution and obtain

$$\|\mu - \widehat{\mu}\|_2 \leq \frac{5\epsilon_M + \sqrt{25\epsilon_M^2 + 64\zeta^2\epsilon_M^2}}{\zeta} = \frac{\epsilon_M}{\zeta}\big(5 + \sqrt{25 + 64\zeta^2}\big) \leq \frac{\epsilon_M}{\zeta}\big(10 + 8\zeta\big).$$

Plugging the original notation into the previous expression, we obtain the final statement. Finally, since $\zeta = \omega_\pi^{(l)}(i)^{-1/2}$ and $\omega_\pi^{(l)}$ is a probability, we have $1/\zeta \leq 1$ and thus

$$\|\mu - \widehat{\mu}\|_2 \leq 18\epsilon_M.$$

which all results are up to permutation. ■

Lemma 17 (De-Whitening) *The upper bound over the de-whitened μ_i is as follow*

$$\epsilon_3 := \left\|(A_3)_i - (\widehat{A_3})_i\right\|_2 \leq G\frac{4\sqrt{2}+2}{(\omega_{\min})^{\frac{1}{2}}(1-\theta)}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}} + \frac{8\epsilon_M}{\omega_{\min}} \tag{49}$$

Proof As it is shown in Anandkumar et al. (2012), to reconstruct the columns of views A_1, A_2, A_3, de-whitening process is needed. It is shown that the columns can be recovered by $(A_1)_i = \widetilde{W}_1^\dagger \lambda_i \mu_i$, $(A_2)_i = \widetilde{W}_2^\dagger \lambda_i \mu_i$, and $(A_3)_i = \widetilde{W}_3^\dagger \lambda_i \mu_i$. For simplicity, let just investigate the third view, the process for other two views is same as third view.

$$\left\|(A_3)_i - (\widehat{A_3})_i\right\|_2 \leq \left\|\widetilde{W}_3^\dagger - \widehat{W}_3^\dagger\right\|_2 \|\lambda_i\mu_i\|_2 + \left\|\widehat{W}_3^\dagger\right\|_2 \left\|\lambda_i\mu_i - \widehat{\lambda}_i\widehat{\mu}_i\right\|_2$$

it is clear that $\left\|\lambda_i\mu_i - \widehat{\lambda}_i\widehat{\mu}_i\right\|_2 \leq \frac{8\epsilon_M}{(\omega_{\min})^{\frac{1}{2}}}$, $\left\|\widehat{W}_3^\dagger\right\| \leq 1$, and $\|\lambda_i\mu_i\|_2 \leq \frac{1}{\omega_{\min}}$.

$$\left\|\widetilde{W}_3^\dagger - \widehat{W}_3^\dagger\right\|_2 = \left\|(BD^{\frac{1}{2}}B^\top - I)\widehat{W}_3^\dagger\right\|_2 \leq 2\left\|M_2 - \widehat{M_2}\right\|_2$$

where the last inequality is inspired by Eq 45. Then

$$\left\|(A_3)_i - (\widehat{A_3})_i\right\|_2 \leq \frac{2}{\omega_{\min}}\left\|M_2 - \widehat{M_2}\right\|_2 + \frac{8\epsilon_M}{(\omega_{\min})^{\frac{1}{2}}}$$

Therefore

$$\epsilon_3 := \left\|(A_3)_i - (\widehat{A_3})_i\right\|_2 \leq G\frac{4\sqrt{2}+2}{(\omega_{\min})^{\frac{1}{2}}(1-\theta)}\sqrt{\frac{\log(2\frac{(d_1+d_2)}{\delta})}{n}} + \frac{8\epsilon_M}{\omega_{\min}} \tag{50}$$

■