# Reinforcement Learning in Rich-Observation MDPs using Spectral Methods

**Kamyar Azizzadenesheli** [1] **Alessandro Lazaric** [2] **Animashree Anandkumar** [1]

## Abstract

We focus on the problem of learning in rich-observation Markov decision processes (ROMDP), where a low-dimensional MDP with $X$ hidden states is observable through a possibly large number of observations. In ROMDPs, hidden states are mapped to observations through an injective mapping, so that an observation $y$ can be generated by only one hidden state $x$. Exploiting this specific structure, we devise a spectral learning method guaranteed to correctly reconstruct the mapping between hidden states and observations. We then integrate this method into UCRL to obtain a reinforcement learning algorithm able to achieve a regret of order $\widetilde{O}(X\sqrt{AN})$ which matches the regret of UCRL running directly on the hidden MDP.

## 1. Introduction

One of the main challenges in reinforcement learning (RL) (Sutton & Barto, 1998) is how to properly trade off the *exploration* of an unknown environment and the *exploitation* of an estimate of the optimal policy to collect as much reward as possible over time. The exploration-exploitation trade-off has been vastly investigated under the PAC-MDP framework (Strehl et al., 2009) and the regret minimization model (Jaksch et al., 2010), for which nearly-optimal algorithms are available. In particular, the UCRL (Jaksch et al., 2010) achieves a regret over $n$ steps of order $\widetilde{O}(D_{\mathcal{Y}}Y\sqrt{AN})$ in environments with $Y$ observable states, $A$ actions, and diameter $D_{\mathcal{Y}}$ (i.e., the longest shortest path between any two states). Despite being nearly-optimal, in practice the performance rapidly degrades as the number of states increases. Nonetheless, in many domains (e.g., robotics) the high-dimensional states observed by the agent are actually generated from an underlying low-dimensional latent space with $X \ll Y$ states that effectively summarize the MDP dynamics and rewards. If the mapping between the latent states and the observations is known, then we could directly learn on the low-dimensional latent space and dramatically reduce the regret to $\widetilde{O}(D_{\mathcal{X}}X\sqrt{AN})$, where $D_{\mathcal{X}}$ is the diameter of the latent MDP. However, this mapping is typically unknown and needs to be learned by the agent. In this case, the main challenge is to efficiently learn the mapping so that the regret can scale with the number of latent states $X$ and not the number of observations $Y$.

**Summary of the results.** In this paper, we show that for the class of rich-observation MDPs (ROMDP), this is indeed possible. In ROMDPs the mapping between hidden states to observations is injective, so that an observation can be generated by only one hidden state and both rewards and dynamics are direct functions of the hidden states. In other words, the hidden states form a non-overlapping clustering of observations and ROMDPs can be seen as a specific class of latent variable models (LVM). A number of tensor spectral decomposition methods have been developed to solve different LVMs, such as Gaussian mixture models, latent Dirichlet allocation, hidden Markov models, with strong sample-complexity guarantees (see e.g., Anandkumar et al. (2014), Anandkumar et al. (2012), Song et al. (2013)).

In this paper, we build on this approach and we derive a novel spectral learning method based on the decomposition of the third moment of the observation process that allows us to correctly recover the mapping between hidden states and observations. In particular, we show that under the assumption of ergodicity and a technical assumption on the stochasticity of the ROMDP, the clustering returned by our method is correct with high probability. Furthermore, as the number of samples used to estimate the third moment increases the number of clusters reduces to a number between $X$ (i.e., actual number of hidden state times) and $XA$. We then integrate this spectral method into the UCRL algorithm and we obtain an algorithm that trades off exploration and exploitation as the mapping from observations to hidden states is learned. We prove a regret bound of order $\widetilde{O}(D_{\mathcal{X}}X\sqrt{AN})$, thus matching the regret that can be obtained when the ROMDP structure is known, while the number of observations $Y$ only appear in lower-order terms.

**Related work.** The existence of a latent space is an assumption that is commonly used in many online learning settings to reduce the learning complexity. In multi-armed bandit, Gheshlaghi-Azar et al. (2013) and Maillard & Mannor (2014) assume that a bandit problem is generated from

[1]University of California, Irvine [2]Institut National de Recherche en Informatique et en Automatique, (Inria).

*Appearing at RLDM2017, Ann Arbor, Michigan, USA*

an unknown (latent) finite set of problems and show how the regret can be significantly reduced by identifying this set. Gentile et al. (2014) consider the more general scenario of latent contextual bandits, where the contexts belong to a few number of underlying hidden classes. Uniform exploration strategy over the contexts, combined with an online clustering algorithm is shown to achieve a regret scaling only with the number of hidden clusters. An extension to recommender systems is considered in (Gopalan et al., 2016) where the contexts type for both the users and items are unknown a priori. Again, uniform exploration is used, and the spectral algorithm of Anandkumar et al. (2014) is deployed to learn the latent classes. The ROMDP model considered in this paper is a generalization of the latent contextual bandits, where actions influence the contexts (i.e., the states) and the objective is to maximize the long-term reward. ROMDPs have been studied by Krishnamurthy et al. (2016), where a PAC analysis is given in the episodic framework, in which the agent tries to learn the best $Q$-function by searching in a given function space. They investigated the setting of deterministic MDPs and extended their results to the general case of contextual decision processes in (Jiang et al., 2016). While the resulting algorithms are proved to achieve a PAC-complexity scaling with the number of hidden states/factors $X$, they are computationally intractable. Learning in ROMDPs can be also seen as a state-aggregation problem, where observations are aggregated to form a small latent MDP. While the literature on state-aggregation in RL is vast, most of the results have been derived for the batch setting (see e.g., Li et al. (2006)). In online RL, Ortner (2013) proposes a method to integrate state aggregation with UCRL. While the resulting algorithm may significantly reduce the computational complexity of extended value iteration, the analysis fails at showing any improvement in the regret. Finally, we notice that ROMDPs are a special class of partially observable Markov decision processes (POMDP). Azizzadenesheli et al. (2016b) recently proposed an algorithm that leverages spectral learning methods to recover the mapping from states to observations, the reward, and the dynamics. While we follow a similar approach, learning in POMDPs is considerably more difficult than in ROMDPs and their final regret still scales with the number of observations, while the computation of the optimal memoryless policy relies on an optimization oracle, which in general is NP-hard (Azizzadenesheli et al., 2016a; Littman, 1994). On the other hand, computing the optimal policy in ROMDPs amounts to solving a standard MDP.

## 2. Preliminaries

A rich-observation MDP (ROMDP) is a tuple $M = \langle \mathcal{X}, \mathcal{Y}, \mathcal{A}, R, f_T, f_O \rangle$, where $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{A}$ are the sets of hidden states, observations, and actions. We denote by $X$, $Y$, and $A$ the cardinality of the sets, while their elements are

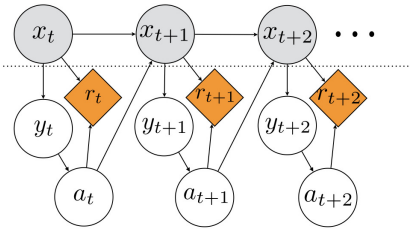


*Figure 1.* Graphical model of a ROMDP.

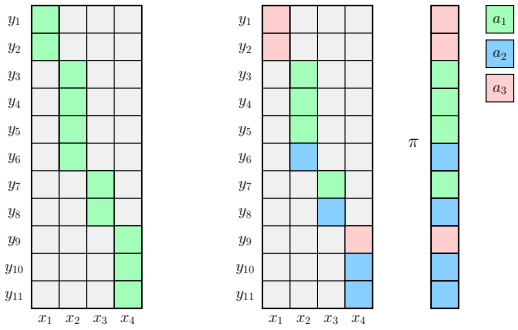


*Figure 2. (left)* Example of an observation matrix $O$. Since state and observation labelling is arbitrary, we arranged the non-zero values so as to display the block diagonal structure of the matrix. *(right)* Example of clustering that can be achieved by policy $\pi$. We have, e.g., that $\mathcal{X}_\pi^{(a_1)} = \{x_2, x_3\}$. Using each action we can recover *partial* clusterings corresponding to 7 hidden states $\mathcal{S} = \{s_1, \dots, s_7\}$ with clusters, e.g., $\mathcal{Y}_{s_1} = \{y_1, y_2\}$ $\mathcal{Y}_{s_2} = \{y_3, y_4, y_5\}$, and $\mathcal{Y}_{s_8} = \{y_{10}, y_{11}\}$, while the remaining elements are the singletons $y_6$, $y_7$, $y_8$, and $y_9$. Clusters coming from different actions cannot be merged together because of different labelling of the hidden state, where, e.g., $x_2$ may be labelled differently depending on whether action $a_1$ or $a_2$ is used.

enumerated by $i \in [X] = \{1..X\}$, $j \in [Y] = \{1..Y\}$, and $l \in [A] = \{1..A\}$. We assume that the number of hidden states is much smaller than the observations, i.e., $X \ll Y$. We assume the rewards to be bounded in $[0, 1]$ and to depend only on hidden states and actions, where the reward matrix $R \in \mathbb{R}^{A \times X}$ is such that $[R]_{i,l} = \mathbb{E}[r(x = i, a = l)]$. The dynamics of the MDP is defined directly on the hidden states as $T_{i',i,l} := f_T(i'|i, l) = \mathbb{P}(x' = i'|x = i, a = l)$, where we introduce the transition tensor $T \in \mathbb{R}^{X \times X \times A}$. Finally, the observations are generated as $[O]_{j,i} = f_O(j|i) = \mathbb{P}(y = j|x = i)$, where the observation matrix $O \in \mathbb{R}^{Y \times X}$ has minimum *non-zero* entry $O_{\min}$. The graphical model associated to a ROMDP is illustrated in Fig. 1.

While this model matches the general partially-observable MDP model, the observations of a ROMDP have a specific structure, where each observation $y$ can be generated by one and only one hidden state. In other words, in ROMDPs the observation matrix $O$ has a block structure where there is exactly one non-zero element in each row (see Fig. 2-*left*). This structure corresponds to assuming the existence of a non-overlapping clustering of the observations into $X$ clusters/hidden states. In particular, we denote by $\mathcal{Y}_x =$

$\mathcal{Y}_i = \{y = j \in \mathcal{Y} : [O]_{j,i} > 0\}$ the set of elements (i.e., observations) in cluster $x$ (i.e., hidden state), while $x_y = x_j$ is the cluster observation $y = j$ belongs to.[1] This structure implies the existence of an MDP $M_{\mathcal{Y}} = \langle \mathcal{Y}, \mathcal{A}, R', f'_T \rangle$ defined directly on the observations, where $R' = R$ since the reward of an observation-action pair $(y, a)$ is the same as the reward in the hidden state-action pair $(x_y, a)$, and the dynamics can be obtained directly from the hidden dynamics and the observation model, such that $f'_T(j'|j,a) = \mathbb{P}(y'=j'|y=j, a=l) = \mathbb{P}(y'=j'|x'=x_{j'})\mathbb{P}(x'=x_{j'}|x=x_j, a=l) = [O]_{j',x_{j'}}[T]_{x_{j'},x_j,l}$. We measure the performance of an observation-based policy $\pi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{A}$ starting from a hidden state $x$ by its asymptotic average reward

$$\rho(x; \pi_{\mathcal{Y}}) = \lim_{N\to\infty} \mathbb{E}\left[\frac{\sum_{t=1}^{N} r_t}{n}\middle| x_0 = x, \pi_{\mathcal{Y}}\right].$$

While the set of observation-based policies is a strict superset of the set of hidden state-based policies $\pi_{\mathcal{X}} : \mathcal{X} \to \mathcal{A}$, it is easy to see that the specific structure of a ROMDP is such that the optimal policy $\pi^*_{\mathcal{Y}}(y)$ for MDP $M_{\mathcal{Y}}$ selects the same action in all observations belonging to the same hidden state (i.e., $\pi^*_{\mathcal{Y}}(y) = a$ for some $a \in \mathcal{A}$ for all observations $y' \in \mathcal{Y}_{x_y}$) and then it can be directly translated into an optimal policy for $M$ by taking $\pi^*_{\mathcal{X}}(x) = \pi^*_{\mathcal{Y}}(y)$ for any $y \in \mathcal{Y}_x$. This equivalence suggests $\pi^*_{\mathcal{X}}$ can be learned by ignoring the hidden structure and directly solving $M_{\mathcal{Y}}$ by applying, e.g., UCRL. Nonetheless, the corresponding high-probability regret over $N$ rounds would be

$$R_N = N\rho^* - \left[\sum_{t=1}^{N} r_t\right] \le \widetilde{O}\big(D_{\mathcal{Y}} Y \sqrt{AN}\big),$$

where $\rho^* = \rho(\pi^*_{\mathcal{X}})$ and $D_{\mathcal{Y}}$ is the diameter of the observation MDP defined as

$$D_{\mathcal{Y}} = \max_{y,y'\in\mathcal{Y}} \min_{\pi:\mathcal{Y}\to\mathcal{A}} \mathbb{E}\big[\tau_\pi(y, y')\big],$$

with $\tau_\pi(y, y')$ being the (random) number of steps from $y$ to $y'$ by following the observation-based policy $\pi$. Unfortunately, this performance may not be satisfactory in practice since the number of observations $Y$ is greater than $X$ (i.e., the low-dimensional structure of the ROMDP is completely ignored) and the diameter of $M_{\mathcal{Y}}$ may be very large. In fact, a policy $\pi_{\mathcal{Y}}$ cannot directly influence the realization of an observation $y' = j$ and it can only try to reach the corresponding hidden state $x_{y'} = i$ as fast as possible from an observation $y$. Nonetheless, if the probability of observing that specific observation is small (i.e., $O_{j,i}$ is small), the number of steps to go from $y$ to $y'$ may be very large. On the other hand, if the hidden representation of a ROMDP is known, the hidden MDP can be solved with an overall regret $R_N = \widetilde{O}\big(D_{\mathcal{X}} X \sqrt{AN}\big)$, where $D_{\mathcal{X}}$ is the diameter of the hidden-state dynamics when executing hidden-state based policies. Our objective in the next section is to devise an algorithm whose regret tends to the regret attainable when the ROMDP structure is known in advance.

[1] Notice that throughout the paper we use the indices $i$, $j$, and $l$ and the "symbolic" values $x$, $y$, and $a$ interchangeably.

## 3. Stochastic latent MDP

In this section we introduce an algorithm combining UCRL with spectral matrix/tensor methods to recover the ROMDP representation. We first describe the methods that can be used to learn the structure of the observation matrix $O$, then we report our UCRL variant, and finally we derive its regret performance. In particular, we show that in order to improve the learning performance, we do not need to estimate the matrix $O$ exactly, but only reconstruct the clusters $\{\mathcal{Y}_x\}_{x\in\mathcal{X}}$ correctly, which can be achieved by identifying the non-zero entries of $O$ and not their exact value.

### 3.1. ROMDP Recovery Through Spectral Methods

**Assumptions.** We focus on a specific subset of ROMDPs satisfying the two following assumptions.

**Assumption 1** (Ergodicity)**.** *For any policy $\pi_{\mathcal{Y}}$, the Markov chain induced on the hidden MDP $M$ is ergodic.*

Relying on this assumption, we define the stationary distribution over hidden states induced by a policy $\pi$ as $\omega_\pi(i)$ for any $i \in [X]$. We also introduce the stationary distribution on $\mathcal{X}$ conditioned on a specific action $a$, that is $\omega^{(l)}_\pi(i) = \mathbb{P}_\pi(x = i|a = l)$. Let $\mathcal{X}^{(l)}_\pi = \{i \in [X] : \omega^{(l)}_\pi(i) > 0\}$ be the hidden states where action $l$ could be taken according to policy $\pi$. In other words, if $\mathcal{Y}^{(l)}_\pi = \{j \in [Y] : \pi(j) = l\}$ is the set of observations in which policy $\pi$ takes action $l$, then $\mathcal{X}^{(l)}_\pi$ is the set of hidden states $\{x_y\}$ with $y \in \mathcal{Y}^{(l)}_\pi$ (see Fig. 2-*right*). We also define the set of all hidden states that can be reached starting from states in $\mathcal{X}^{(l)}_\pi$ and taking action $l$, that is

$$\overline{\mathcal{X}}^{(l)}_\pi = \bigcup_{i\in\mathcal{X}^{(l)}_\pi} \Big\{i' \in [X] : \mathbb{P}\big(x' = i'|x = i, a = l\big) > 0\Big\}.$$

Similarly we $\underline{\mathcal{X}}^{(l)}_\pi$ is the set of latent states that are mapped to $\mathcal{X}^{(l)}_\pi$ by policy $\pi$. We need the following assumption.

**Assumption 2** (Full-Rank)**.** *For any policy $\pi_{\mathcal{Y}}$, the transition matrix is full-rank.*

The Asm. 2 results that for any action $l$, $M$ is expansive, i.e., $\big|\mathcal{X}^{(l)}_\pi\big| \le \big|\overline{\mathcal{X}}^{(l)}_\pi\big|$.

This assumption simply requires that the number of hidden states where policy $\pi$ can take an action $a$ (i.e., $\mathcal{X}^{(l)}_\pi$) is smaller than the number of states that can be reached when executing action $a$ itself (i.e., $\overline{\mathcal{X}}^{(l)}_\pi$).

**The multi-view model and exact recovery.** We are now ready to introduce the multi-view model (Anandkumar et al.,

2014) that allows us to reconstruct the clustering structure of the ROMDP. We consider the trajectory of observations and actions generated by an arbitrary policy $\pi$ and we focus on three consecutive observations $y_{t-1}, y_t, y_{t+1}$ at any step $t$. As customary in multi-view models, we *vectorize* the observations into three one-hot view vectors $\vec{v}_1, \vec{v}_2, \vec{v}_3$ in $\{0,1\}^Y$ such that $\vec{v}_1 = \vec{e}_j$ corresponds to saying that the observation in the first view is $j \in [Y]$ and where we remap time indices $t-1, t, t+1$ onto 1, 2, and 3. We notice that these views are indeed independent random variables when conditioning on the state $x_2$ (i.e., the hidden state at time $t$) and the action $a_2$ (i.e., the action at time $t$), thus defining a multi-view model for the hidden state process. Let $k_1 = \left|\underline{\mathcal{X}}_\pi^{(l)}\right|$, $k_2 = \left|\mathcal{X}_\pi^{(l)}\right|$ and $k_3 = \left|\overline{X}_\pi^{(l)}\right|$, then we define the factor matrices $V_1^{(l)} \in \mathbb{R}^{Y \times k_1}, V_2^{(l)} \in \mathbb{R}^{Y \times k_2}, V_3^{(l)} \in \mathbb{R}^{Y \times k_3}$ as follows

$$[V_1^{(l)}]_{j,i} = \mathbb{P}(\vec{v}_1 = \vec{e}_j | x_2 = i, a_2 = l),\ i \in \underline{\mathcal{X}}_\pi^{(l)}$$
$$[V_2^{(l)}]_{j,i} = \mathbb{P}(\vec{v}_2 = \vec{e}_j | x_2 = i, a_2 = l),\ i \in \mathcal{X}_\pi^{(l)}$$
$$[V_3^{(l)}]_{j,i} = \mathbb{P}(\vec{v}_3 = \vec{e}_j | x_2 = i, a_2 = l),\ i \in \overline{\mathcal{X}}_\pi^{(l)}.$$

We are interested in estimating $V_2^{(l)}$ since it directly relates to the observation matrix as

$$[V_2^{(l)}]_{j,i} = \frac{\mathbb{P}(a_2 = l | y_2 = j)\mathbb{P}(y_2 = j | x_2 = i)}{\mathbb{P}(a_2 = l | x_2 = i)} = \frac{\mathbb{I}\{\pi(j) = l\} f_O(j|i)}{\mathbb{P}(a_2 = l | x_2 = i)}, \quad (1)$$

where $\mathbb{I}$ is the indicator function. As it can be noticed, $V_2^{(l)}$ borrows the same structure as the observation matrix $O$ and since we want to recover only the clustering structure of $M$ (i.e., $\{\mathcal{Y}_i\}_{i \in [X]}$), it is sufficient to compute the columns of $V_2^{(l)}$ up to any multiplicative constant. In fact, any non-zero entry of $V_2^{(l)}$ corresponds to a non-zero element in the original observation matrix (i.e., $[V_2^{(l)}]_{j,i} > 0 \Rightarrow [O]_{j,i} > 0$) and for any hidden state $i$, we can construct a cluster $\mathcal{Y}_i^{(l)} = \{j \in [Y] : [V_2^{(l)}]_{j,i} > 0\}$, which is accurate up to a re-labelling of the states. More formally, there exists a mapping function $\sigma^{(l)} : \mathcal{X} \to \mathcal{X}$ such that any pair of observations $j, j' \in \mathcal{Y}_i^{(l)}$ is such that $j, j' \in \mathcal{Y}_{\sigma(i)}$. Nonetheless, as illustrated in Fig. 2-*right*, the clustering may not be minimal. In fact, we have $[O]_{j,i} > 0 \not\Rightarrow [V_2^{(l)}]_{j,i} > 0$ since $[V_2^{(l)}]_{j,i}$ may be zero because of policy $\pi$ even if $[O]_{j,i} > 0$. Since the (unknown) mapping function $\sigma^{(l)}$ changes with actions, we are unable to correctly "align" the clusters and we may obtain more clusters than hidden states. We define $\mathcal{S}$ as the auxiliary state space obtained by the partial aggregation of observations into clusters $\{\mathcal{Y}_i^{(l)}\}_{i,l}$ together with observations which are not included in any cluster $\mathcal{Y}_i^{(l)}$. We can prove the following.

**Lemma 1.** *Given a policy $\pi$, for any action $l$ and any hidden state $i \in \mathcal{X}_\pi^{(l)}$, let $\mathcal{Y}_i^{(l)}$ be the observations that can be clustered together according to $V_2^{(l)}$ and $\mathcal{Y}^{\mathsf{c}} = \mathcal{Y} \setminus \bigcup_{i,l} \mathcal{Y}_i^{(l)}$ be the observations not clustered, then the auxiliary state space $\mathcal{S}$ contains all the clusters $\{\bigcup_{i,l} \mathcal{Y}_i^{(l)}\}$ and the singletons in $\mathcal{Y}^{\mathsf{c}}$ for a total number of elements $S = |\mathcal{S}| \le AX$.*

In the next section we discuss how the size of $\mathcal{S}$ can be further reduced (to $X$) by exploiting clusterings returned by different policies over time.

We now show how to recover the factor matrix $V_2^{(l)}$. We introduce mixed second and third order moments as

$$K_{p,q}^{(l)} = \mathbb{E}[\vec{v}_p \otimes \vec{v}_q], \quad K_{p,q,r}^{(l)} = \mathbb{E}[\vec{v}_p \otimes \vec{v}_q \otimes \vec{v}_r]$$

where $p, q, r$ is any permutation of $\{1, 2, 3\}$. Exploiting the conditional independence of the views, the second moments (similarly for the third moment) can be written as

$$K_{p,q}^{(l)} = \sum_{i \in \mathcal{X}_\pi^{(l)}} \mathbb{P}(x_2 = i | a_2 = l) \times \mathbb{E}\big[\vec{v}_p | x_2 = i, a_2 = l\big] \otimes \mathbb{E}\big[\vec{v}_q | x_2 = i, a_2 = l\big] = \sum_{i \in \mathcal{X}_\pi^l} \omega_\pi^{(l)}(i) [V_p^{(l)}]_{:,i} \otimes [V_q^{(l)}]_{:,i},$$

where $[V_p^{(l)}]_{:,i}$ denotes the $i$-th column of matrix $V_p^{(l)}$. In general the second moment matrices are rank deficient, with rank $X_\pi^{(l)}$. We now proceed with defining a symmetric second moment by introducing the symmetrized views

$$\widetilde{v}_1 = K_{2,3}^{(l)} (K_{1,3}^{(l)})^\dagger \vec{v}_1, \qquad \widetilde{v}_3 = K_{2,1}^{(l)} (K_{3,1}^{(l)})^\dagger \vec{v}_3, \quad (2)$$

where $K^\dagger$ denotes Moore–Penrose pseudoinverse. Then we can construct the second moment $M_2^{(l)} \in \mathbb{R}^{Y \times Y}$ as

$$M_2^{(l)} = \mathbb{E}\big[\widetilde{v}_1 \otimes \widetilde{v}_3\big] = \sum_{i \in \mathcal{X}_\pi^{(l)}} \omega_\pi^{(l)}(i) [V_2^{(l)}]_{:,i} \otimes [V_2^{(l)}]_{:,i}. \quad (3)$$

We notice that the columns of $V_2^{(l)}$ are all orthogonal (but not orthonormal) since the clusters are non-overlapping and an observation $j$ that can be obtained from a state $i$ cannot be generated by any other state $i'$ (i.e., for any $i \neq i'$, $[V_2^{(l)}]_{:,i}^\mathsf{T} [V_2^{(l)}]_{:,i'} = 0$). As a result, Eq. 3 can be seen as an eigendecomposition of $M_2^{(l)}$, where the columns $[V_2^{(l)}]_{:,i}$ are the eigenvectors and $\omega_\pi^{(l)}(i)$ are the eigenvalues. More formally, let $M_2^{(l)} = U \Sigma U^\mathsf{T}$ be the eigendecomposition of $M_2^{(l)}$, if all eigenvalues are distinct, the eigenvectors in $U$ can be used to recover $V_2^{(l)}$ up to a mapping function and multiplicative factors.

**Lemma 2.** *For any action $l \in [A]$, let $M_2^{(l)}$ be the second moment constructed on the symmetrized views as in Eq. 3 and $M_2^{(l)} = U \Sigma U^\mathsf{T}$ be its eigendecomposition. If all eigenvalues of $M_2^{(l)}$ have multiplicity 1, there exists a mapping $\sigma^{(l)} : X \to X$ and multiplicative constants $\{C_i^{(l)}\}_{i \in [X]}$, such that for any $i \in \mathcal{X}_\pi^{(l)}$ and $j \in [Y]$,*

$[V_2^{(l)}]_{j,\sigma^{(l)}(i)} = C_i^{(l)}[U]_{j,i}$. *As a result, for any hidden state* $i \in \mathcal{X}_\pi^{(l)}$ *we define the cluster* $\widetilde{\mathcal{Y}}_i^{(l)}$ *as*

$$\widetilde{\mathcal{Y}}_i^{(l)} = \{j \in [Y] : [U]_{j,i} > 0\} \tag{4}$$

*and we have that if* $j, j' \in \widetilde{\mathcal{Y}}_i^{(l)}$ *then* $j, j' \in \mathcal{Y}_{\sigma(i)}^{(l)}$ *(i.e., observations that are clustered together in* $\widetilde{\mathcal{Y}}_i^{(l)}$ *are clustered in the original ROMDP).*

This provides us with a first spectral decomposition technique to correctly reconstruct part of the hidden structure of $M$. Nonetheless, the previous lemma relies on the assumption that all eigenvalues of $M_2$ has single multiplicity, which is not true in general. In this case, the columns of $V_2^{(l)}$ associated to eigenvalues with multiplicity cannot be recovered correctly and the corresponding clustering may be wrong since observations generated by distinct states (and thus with different rewards and dynamics) may be aggregated together. In this case, we have to move to the third order statistics to disambiguate between observations and cluster them properly. We introduce the symmetric tensor

$$\begin{aligned} M_3^{(l)} &= \mathbb{E}\big[\widetilde{v}_1 \otimes \widetilde{v}_3 \otimes \vec{v}_2\big] \\ &= \sum_{i \in \mathcal{X}_\pi^l} \omega_\pi^{(l)}(i) [V_2^{(l)}]_{:,i} \otimes [V_2^{(l)}]_{:,i} \otimes [V_2^{(l)}]_{:,i}. \end{aligned} \tag{5}$$

We can now employ the standard machinery of tensor decomposition methods to orthogonalize the tensor $M_3^{(l)}$ using $M_2^{(l)}$ and recover $V_2^{(l)}$ (we refer to (Anandkumar et al., 2014) for further details) and a suitable clustering.

**Lemma 3.** *For any action* $l \in [A]$*, let* $M_3^{(l)}$ *be the third moment constructed on the symmetrized views as in Eq. 5, then we can orthogonalize it using the second moment* $M_2^{(l)}$ *and obtain a unique spectral decomposition from which we compute the exact factor matrix* $[V_2^{(l)}]_{j,i}$*. As a result, for any hidden state* $i \in \mathcal{X}_\pi^{(l)}$ *we define the cluster* $\widetilde{\mathcal{Y}}_i^{(l)}$ *as*

$$\widetilde{\mathcal{Y}}_i^{(l)} = \{j \in [Y] : [V_2^{(l)}]_{j,i} > 0\} \tag{6}$$

*and there exists a mapping* $\sigma^{(l)} : X \to X$ *such that if* $j, j' \in \widetilde{\mathcal{Y}}_i^{(l)}$ *then* $j, j' \in \mathcal{Y}_{\sigma^{(l)}(i)}$ *(i.e., observations that are clustered together in* $\widetilde{\mathcal{Y}}_i^{(l)}$ *are clustered in the original ROMDP).*

Unlike Lem. 2, performing a tensor decomposition on $M_3^{(l)}$ guarantees the correct recovery of $V_2^{(l)}$ with no restriction on $M_2^{(l)}$ (i.e., even when it has eigenvalues with multiplicity).

**Spectral learning algorithms.**[2] Lem. 3 provides guarantees for the *exact* recovery of $V_2^{(l)}$ when $M_3^{(l)}$ is known,

[2]We only report the spectral learning algorithm for the tensor decomposition but a very similar algorithm and guarantees can be derived for the matrix decomposition approach when the eigenvalues of $\widehat{M}_2^{(l)}$ have multiplicity 1.

**Algorithm 1** Spectral learning algorithm.

**Input:** Trajectory $(y_1, a_1, \dots, y_N)$
**for** Action $l \in [A]$ **do**
  Estimate second moments $\widehat{K}_{2,3}^{(l)}$, $\widehat{K}_{1,3}^{(l)}$, $\widehat{K}_{2,1}^{(l)}$, and $\widehat{K}_{3,1}^{(l)}$
  Estimate the rank of matrix $\widehat{K}_{2,3}^{(l)}$ (see App. A.2)
  Compute symmetrized views $\widetilde{v}_{1,t}$ and $\widetilde{v}_{3,t}$, for $t = 2..N-2$
  Compute second and third moments $\widehat{M}_2^{(l)}$ and $\widehat{M}_3^{(l)}$
  Compute $\widehat{V}_2^{(l)}$ from the tensor decomposition of (an orthogonalized version of) $\widehat{M}_3^{(l)}$ and return clusters

$$\widehat{\mathcal{Y}}_i^{(l)} = \{j \in [Y] : [\widehat{V}_2^{(l)}]_{j,i} > 0\}$$

**end for**

while in practice we can only estimates $M_3^{(l)}$ through samples. Let $N$ be the length of the trajectory generated by policy $\pi$, then we can construct $N-2$ triples $\{y_{t-1}, y_t, y_{t+1}\}$ that can be used to construct the corresponding views $\vec{v}_{1,t}, \vec{v}_{2,t}, \vec{v}_{3,t}$ and to estimate second mixed moments as

$$\widehat{K}_{p,q}^{(l)} = \frac{1}{N(l)} \sum_{t=1}^{N-1} \mathbb{I}(a_t = l)\, \vec{v}_{p,t} \otimes \vec{v}_{q,t},$$

with $p, q \in \{1, 2, 3\}$ and $N(l) := \max\{1, \sum_t^{N-1} \mathbb{I}(a_t = l)\}$ . Furthermore, Lem. 3 requires a knowledge of exact $|\mathcal{X}_\pi^{(l)}|$ which is not known apriori. Under Asm. 1 and 2, for any action $l$, the rank of $K_{2,3}^{(l)}$ is indeed $|\mathcal{X}_\pi^{(l)}|$ and thus $\widehat{K}_{2,3}^{(l)}$ can be used to recover the rank. The actual way to calculate the efficient rank of $\widehat{K}_{2,3}^{(l)}$ is quite intricate and we postpone the details to App. A.2. From $\widehat{K}_{p,q}^{(l)}$ we can construct the symmetric views $\widetilde{v}_{1,t}$ and $\widetilde{v}_{3,t}$ as in Eq. 2 and compute the estimate second and third moments as

$$\widehat{M}_2^{(l)} = \frac{1}{N(l)} \sum_{t=1}^{N-1} \mathbb{I}(a_t = l) \widetilde{v}_{1,t} \otimes \widetilde{v}_{3,t},$$

$$\widehat{M}_3^{(l)} = \frac{1}{N(l)} \sum_{t=1}^{N-1} \mathbb{I}(a_t = l) \widetilde{v}_{1,t} \otimes \widetilde{v}_{3,t} \otimes \vec{v}_{2,t}.$$

Following the same procedure as in the exact case, we are then able to recover estimates of the factor matrix $\widehat{V}_2^{(l)}$, which enjoys the following error guarantee.

**Lemma 4.** *Under Asm. 1 and 2, let* $\widehat{V}_2^{(l)}$ *denotes the empirical estimate of* $V_2^{(l)}$*. Given a trajectory of* $N$ *samples obtained by executing a policy* $\pi$*, then there exists* $N_0$ *such that for any* $l \in \mathcal{A}$ *and* $i \in \mathcal{X}_\pi^{(l)}$ *if* $N(l) > N_0$

$$\|[V_2^{(l)}]_{\cdot,i} - [\widehat{V}_2^{(l)}]_{\cdot,i}\|_2 \le C_2 \sqrt{\frac{\log(2Y^{3/2}/\delta)}{N(l)}} := \mathcal{B}_O^{(l)} \tag{7}$$

*with probability at least* $1 - \delta$*,* $C_2$ *is a problem-dependent constant independent from the number of observations* $Y$*.*

While this estimate could be directly used to construct a clustering of observations, the noise in the empirical esti-

mates might lead to $[\widehat{V}_2^{(l)}]_{j,i} > 0$ for any $(j,i)$ pair, which prevents us from generating any meaningful clustering. On the other hand, we can use the guarantee in Lem. 4 to single-out the entries of $\widehat{V}_2^{(l)}$ that are non-zero w.h.p. In particular, we derive the binary matrix $\widetilde{V}_2^{(l)} \in \{0,1\}^{Y \times X}$ as

$$[\widetilde{V}_2^{(l)}]_{j,i} = \begin{cases} 1 & \text{if } [\widehat{V}_2^{(l)}]_{j,i} \geq \mathcal{B}_O^{(l)} \\ 0 & \text{otherwise} \end{cases},$$

which relies on the fact that if $[\widehat{V}_2^{(l)}]_{j,i} - \mathcal{B}_O^{(l)} > 0$ then $[V_2^{(l)}]_{j,i} > 0$. At this point, for any $l$ and any $i \in \mathcal{X}_\pi^{(l)}$, we can generate the cluster

$$\widehat{\mathcal{Y}}_i^{(l)} = \{j \in [Y] : [\widetilde{V}_2^{(l)}]_{j,i} > 0\}, \tag{8}$$

which is guaranteed to aggregate observations correctly in high-probability. We denote be $\widehat{\mathcal{Y}}^{\mathsf{c}} = \mathcal{Y} \setminus \bigcup_{i,l} \widehat{\mathcal{Y}}_i^{(l)}$ the set of observations which are not clustered through this process. Then we define the auxiliary state space $\widehat{\mathcal{S}}$ obtained by enumerating all the elements of non-clustered observations together with clusters $\{\widehat{\mathcal{Y}}_i^{(l)}\}_{i,l}$, for which we have the following guarantee.

**Corollary 1.** *Let $\widehat{\mathcal{S}}$ be the auxiliary states composed of clusters $\{\widehat{\mathcal{Y}}_i^{(l)}\}$ and singletons in $\mathcal{Y}^{\mathsf{c}}$ obtained by clustering observations according to $\widetilde{V}_2^{(l)}$, then for any pair of observations $j, j'$ clustered together in $\widehat{\mathcal{S}}$, there exists a hidden state $i$ such that $j, j' \in \mathcal{Y}_i$. Finally, $\widehat{S} = |\widehat{\mathcal{S}}| \geq S$ and $\widehat{\mathcal{S}} \to \mathcal{S}$ as $N$ tends to infinity.*

The computation complexity of Alg. 1 is mainly determined by the iterative robust power method used to decompose the third order tensor. This complexity has been studied by Song et al. (2013) and needs $poly(k)\log(1/\delta)$ initialization ($k$ being the rank), $\log(X)$ iteration for each initialization and each iteration has at most $\mathcal{O}(X^3)$ computation due to tensor multiplication. It is worth noting that in high dimensional environments ($Y \gg 1$), the moments are usually very sparse and the multiplications are done using sparse methods and the pseudo inverses can be computed using efficient randomized methods.

### 3.2. Spectral Learning UCRL.

**The algorithm.** We now describe the spectral Learning UCRL (SL-UC) obtained by integrating the spectral methods above with the UCRL strategy to optimize the exploration-exploitation trade-off and achieve small regret (Alg. 2). The learning process is split into epochs of increasing length. At the beginning of each epoch $k$, we first use the trajectory $(s_1, a_1, \dots, s_{N^{(k-1)}}), s \in \widehat{\mathcal{S}}^{(k)}$ from the previous epoch to construct the auxiliary state space $\widehat{\mathcal{S}}$ using Alg. 1[3].

[3]The input to the Alg. 1 is a sequence of $(s_1, a_1, \dots, s_{N^{(k-1)}}), s \in \widehat{\mathcal{S}}^{(k)}$ instead of $(y_1, a_1, \dots, y_{N^{(k-1)}})$, i.e. the algorithm considers $s$ as an observation and computes the decomposition on the space of size $\widehat{S}^{(k-1)}$ instead of $Y$.(it significantly reduces the computation complexity)

**Algorithm 2** Spectral-Learning UCRL(SL-UC).

**Input:** Confidence $\delta'$
**Initialize:** $t = 1$, initial state $x_1$, $k = 1$, $\delta/N^6$
**while** $t < N$ **do**
  Run Alg. 1 with samples from epoch $k-1$ and obtain $\widehat{\mathcal{S}}$
  Compute the auxiliary space $\widehat{\mathcal{S}}^{(k)}$ by merging $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}^{(k-1)}$
  Compute the estimate reward $r^{(k)}$ and dynamics $p^{(k)}$
  Construct plausible set of AuxMDP $\mathcal{M}^{(k)}$ out of set $\widehat{\mathcal{S}}^k$
  Compute the optimistic policy

$$\widetilde{\pi}^{(k)} = \arg\max_{\pi} \max_{M \in \mathcal{M}^{(k)}} \rho(\pi; M) \tag{9}$$

  Set $v^{(k)}(s,l) = 0$ for all actions $l \in \mathcal{A}$, $s \in \widehat{\mathcal{S}}^{(k)}$
  **while** $\forall l, \forall s, v^{(k)}(s,l) < \max\{1, N^{(k)}(s,l)\}$ **do**
    Execute $a_t = \widetilde{\pi}^{(k)}(s_t)$
    Observe reward $r_t$ and observation $y_t$
  **end while**
  Set $k = k + 1$
**end while**

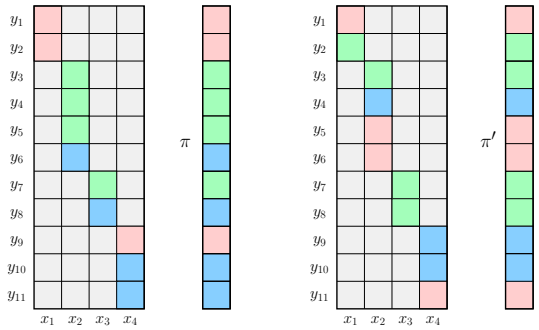


*Figure 3.* Example of clusterings obtained from two policies that can be effectively merged.

As discussed in the previous section, the limited number of samples and the specific policy executed at epoch $k-1$ may prevent from clustering many observations together, which means that despite $\widehat{\mathcal{S}}$ being a *correct* clustering (see Cor. 1), its size may still be large. While clusterings obtained at different epochs cannot be "aligned" because of different labelling, we can still effectively merge together any two clusterings $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}'$ generated by two different policies $\pi$ and $\pi'$.

In order to illustrate this procedure we consider the scenario in Fig. 3. Observations $y_3$, $y_4$, and $y_5$ are clustered together in the auxiliary space generated by $\pi$, while $y_5$ and $y_6$ are clustered together using $\pi'$. While the labelling of the auxiliary states is arbitrary, observations preserve their labels across epochs and thus we can safely conclude that observations $y_3$, $y_4$, $y_5$, and $y_6$ belong to the same hidden state. Similarly, we can construct a new cluster with $y_9$, $y_{10}$, and $y_{11}$, which, in this case, returns the ex-

act hidden space $\mathcal{X}$. Following this procedure we generate $\widehat{\mathcal{S}}^{(k)}$ as the clustering obtain by merging $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{S}}^{(k-1)}$ (where $\widehat{\mathcal{S}}^1 = \mathcal{Y}$). At this point we can directly estimate the reward and transition model of the auxiliary MDP constructed on $\widehat{\mathcal{S}}^{(k)}$ by using standard empirical estimators. Let $N^{(k)}(y, a, y')$ be the number of times a transition from observation $y$ to $y'$ through action $a$ has been observed up to the beginning of epoch $k$. Then we can easily construct $N^{(k)}(s, a, s') = \sum_{y \in s} \sum_{y' \in s'} N^{(k)}(y, a, y')$, where with an abuse of notation we write $y \in s$ to denote the fact that observation $y$ has been clustered into an auxiliary state $s$ at the epoch $k$. Similarly we can compute the number of visits to any auxiliary state-action pair $N^{(k)}(s, a)$ and the reward cumulated over time $R^{(k)}(s, a)$. Then we return the estimates[4]

$$\widehat{r}^{(k)}(s,a) = \frac{R^{(k)}(s,a)}{\max\{1, N^{(k)}(s,a)\}}; \widehat{p}^{(k)}(s'|s,a) = \frac{N^{(k)}(s,a,s')}{\max\{N^{(k)}(s,a)\}}$$

As a result, Chernoff-Hoeffding confidence bounds can be derived as in UCRL such that for any $s \in \widehat{\mathcal{S}}^{(k)}$ and $a \in \mathcal{A}$

$$\|p(\cdot|s,a) - \widehat{p}^{(k)}(\cdot|s,a)\|_1 \le d_p(s,a) = \sqrt{\frac{7S^{(k)} \log(\frac{2AN^{(k)}}{\delta})}{\max\{1, N^{(k)}(s,a)\}}}$$

$$|\bar{r}(s,a) - \widehat{r}^{(k)}(s,a)| \le d_r(s,a) = \sqrt{\frac{7 \log(\frac{2S^{(k)}AN^{(k)}}{\delta})}{2\max\{1, N^{(k)}(s,a)\}}}$$

with probability at least $1 - \delta$, where $p(\cdot|s, a)$ and $\bar{r}$ are the transition probabilities and reward of the auxiliary MDP $M_{\widehat{\mathcal{S}}^{(k)}}$. At this point we can simply apply the same steps as in standard UCRL, where an optimistic auxiliary MDP $\widetilde{M}^{(k)} := \langle \widehat{\mathcal{S}}^{(k)}, \mathcal{A}, \widetilde{R}, \widetilde{p}(\cdot|\cdot,\cdot) \rangle$ is constructed using the confidence intervals above and the extended value iteration algorithm (see (Jaksch et al., 2010) for more details). The resulting optimal optimistic policy $\widetilde{\pi}^{(k)}$ is then executed until the number samples at least for one pair of auxiliary state and action is doubled.

The optimization at Eq. 9 has be shown to have computation complexity of finding an optimal policy for an MDP of size $\widehat{S}^k$(Jaksch et al., 2010). This optimization can be done via dynamic programming with $\mathcal{O}(\widehat{S}^{(k)})$ iterations and $\mathcal{O}((\widehat{S}^{(k)})^2 A)$ computations per iteration. It reduces to $\mathcal{O}(X)$ iterations and $\mathcal{O}((X)^2 A)$ per iteration when the true clustering is learnt. Compare to UCRL, with $\mathcal{O}(Y)$ iterations and $\mathcal{O}((Y)^2 A)$ computations per iteration, it is a significant improvement.

[4]Since the clustering $\widehat{\mathcal{S}}^{(k)}$ is *monotonic* (i.e., observations clustered together at epoch $k$ they stay clustered at any other epoch $k' > k$), $\widehat{r}^{(k)}(s, a)$ and $\widehat{p}^{(k)}(s'|s, a)$ can be computed incrementally without storing the statistics $N^{(k)}(y, a, y')$, $N^{(k)}(y, a)$, and $R^{(k)}(y, a)$ at observation level, but simply updating the auxiliary state statistics from epoch to epoch, thus significantly reducing the space complexity of the algorithm.

**Regret guarantees.** Corollary 1 guarantees that the number of auxiliary states in $\widehat{\mathcal{S}}$ reduces as the epochs get longer up to $XA$ (i.e., the size of $\mathcal{S}$ in the worst case). Furthermore, as more clusterings are merged together into $\mathcal{S}^{(k)}$, its size may shrink well below the upper bound of $XA$. Nonetheless, it is not possible to prove that $\widehat{\mathcal{S}}^{(k)}$ converges to $\mathcal{X}$ and even if the number of clusters is nearly-minimal, the MDP constructed on the auxiliary state space may have a large diameter. In fact, it is enough that an observation $j$ with very low probability $O_{j,i}$ is not clustered (it is a singleton in $\mathcal{S}^{(k)}$) to have a diameter that scales as $1/O_{\min}$, which could be at least as large as $Y$. While its *actual* impact on the regret may be negligible (i.e., if $j$ is not visited by the current policy), in general the advantage obtained by clustering reduces the dependency on the number of states from $Y$ to $XA$ but it may not be effective in reducing the dependency on the diameter from $D_{\mathcal{Y}}$ to $D_{\mathcal{X}}$. In order to deal with this problem, we can integrate Alg. 2 with a clustering technique similar to the one used by Gentile et al. (2014) and Ortner (2013). At any epoch $k$, we proceed by merging together all the auxiliary states in $\widehat{\mathcal{S}}^{(k)}$ whose reward and transition confidence intervals overlap (i.e., $s$ and $s'$ are merged if the confidence interval $[\widehat{r}(s, a) \pm d_r(s, a)]$ overlaps with $[\widehat{r}(s', a) \pm d_r(s', a)]$ and $[\widehat{p}(\cdot|s, a) \pm d_p(s, a)]$[5] overlaps with $[\widehat{p}(\cdot|s', a) \pm d_p(s', a)]$). Then we check if the number new clusters is equal to $X$ we claim we leaned the true clustering, if it is less than $X$ we forget this further clustering and proceed to the next epoch. While in general merging different auxiliary states leads to a "biased" clustering (i.e., different observations may be clustered together), we can prove the following.

**Lemma 5.** *Exploiting the the structure of reward function and transition probability on the top of spectral learning speeds up the general clustering, results in early stopping and converges to latent MDP of size $X$.*

At this point, we have a method to refine the auxiliary state space $\mathcal{S}^{(k)}$ down to the minimal clustering defined by $\mathcal{X}$, thus recovering the same number of hidden state $X$ and the same diameter $D_{\mathcal{X}}$.

**Theorem 1** (Regret Bound of Stochastic Environment)**.** *Consider a ROMDP $M = \langle \mathcal{X}, \mathcal{Y}, \mathcal{A}, R, f_T, f_O \rangle$ characterized by a diameter $D_{\mathcal{X}}$ and $\tau_M = \max_\pi \max_x \mathbb{E}_\pi[\tau(x \to x)]$. If* SL-UC *is run over $N$ time steps, under Asm. 1 and 2, it suffers the total regret of*

$$Reg_N \le 34 D X \sqrt{A(N-\tau)\log(N/\delta)} + \sum_{k=1}^{k^*} \left( D_{\widehat{S}^{(k)}} \sqrt{14 \widehat{S}^{(k)} \log(N^{(k)}/\delta)} \sum_{s \in \widehat{\mathcal{S}}^{(k)}, a} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}(s,a)}} \right)$$

*with probability at least $1 - \delta$ where the second additive term is $N$-independent and loosely bounded by*

[5]Deviation $d_p(s, a)$ on a $\widehat{S}$ dimensional simplex

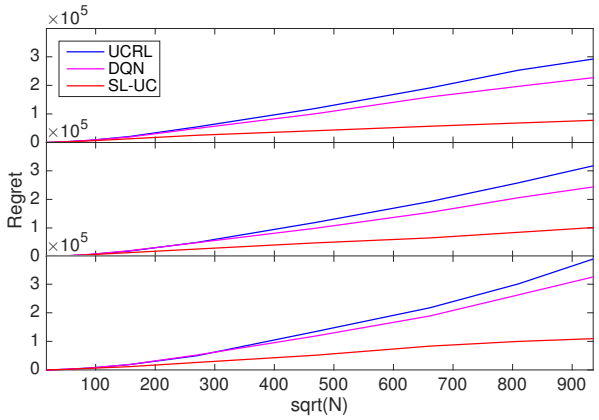


*Figure 4.* Regret comparison for ROMDPs with $X = 5, A = 4$ and from top to bottom $Y = 10, 20, 30$.

$\min\{\tau, 34 D_Y Y \sqrt{A\tau \log(N/\delta)}\}$. *The constant number of steps $\tau$ is the time that the algorithm takes to identify the whole mapping from observations to hidden states*

$$\tau := C_1 16 A Y \tau_M (4 C_2 \frac{\log(Y/\delta)}{O_{\min}^2} + \log(XA/\delta))$$

*where $C_1$ is a numerical constant which represents the similarity of observations in policy and $k^*$ is order of $log_2(\tau)$.*

The bound reported above show that SL-UC achieves a regret roughly of the same order as if the hidden MDP was known from the beginning. On the other hand, it suffers from a constant term which depends on the observation MDP through $Y$ and $D_Y$. The time to get to the right cluster $\tau$ also depends on the smallest probability of observation $O_{\min}$, since rare observations may take a long time before being correctly clustered. Notice that we left the additive term in the regret relatively implicit to stress the fact that the performance of SL-UC progressively gets better as more and more observations are clustered together since the number of aggregated $N^{(k)}(s, a)$ increases, the number of auxiliary states $\widehat{\mathcal{S}}^{(k)}$ decreases together with the diameter.

We can compare this result with the regret bound of Azizzadenesheli et al. (2016b) for POMDPs, which are a more general class than ROMDPs. Recalling that POMDPs are characterized by a diameter $D_{pomdp}$ defined as

$$D_{pomdp} := \max_{x,x',a,a'} \min_{\pi \in \mathcal{P}} \mathbb{E}[\tau\left((x, a) \to (x', a')\right)],$$

the regret derived by Azizzadenesheli et al. (2016b) scales as $\widetilde{\mathcal{O}}(D_{pomdp} X^{3/2} \sqrt{AYN})$. The regret suffers from additional term $\sqrt{Y}$ because the RL algorithm in POMDP put much effort on accurate estimation of entries of $O$ matrix and does not exploit its specific structure. Moreover, there is an additional factor $X$ in regret bound due to learning of transition tensor through spectral methods.

## 4. Experiments

We validate our theoretical results by comparing the performance of SL-UC, UCRL2 (model based, exact solution) and DQN (model free, function approximation) (Mnih et al., 2013), two well known RL algorithms. Since SL-UC is proposed specifically for infinite horizon environment, we are not able to compare it with episodic methods, e.g., PSRL (Osband et al., 2013). For DQN we implemented a three hidden-layers feed forward network (with no CNN block), equipped with RMSprop and reply buffer. We tuned the hyper parameters of the network and reported the best performance (achieved by network of size $30 \times 30 \times 30$)). For SL-UC, we need to compute the confidence bound $\mathcal{B}_O$ for the entries in matrix $V_2$. Since this bound contains many problem-dependent quantities which are usually unknown at run time, we bootstrap over samples and construct the corresponding confidence intervals.

We consider three randomly generated ROMDPs with $X = 5$, $A = 4$ and observation spaces of sizes $Y = 10, 20, 30$. Fig. 4 reports the regret on a $\sqrt{N}$ scale. We see that in each instance the regret of UCRL and DQN grows much faster than SL-UC's, which leverages on the low-dimensional latent space to cluster observations and reduce the regret. Furthermore, while all regrets tend to be linear (i.e., growing as $\sqrt{N}$), we clearly see that the performance of UCRL and DQN is negatively affected by the increasing number of observations, while the regret of SL-UC stays almost constant, confirming that the hidden space $\mathcal{X}$ is learned very rapidly. We report additional experiments in App. C.

## 5. Conclusion

We introduced SL-UC, a novel RL algorithm to learn in ROMDPs combining a spectral method for recovering the clustering structure of the problem and UCRL to effectively trade off exploration and exploitation. We proved theoretical guarantees showing that SL-UC progressively refines the clustering so that its regret tends to the regret that could be achieved when the hidden structure is known in advance. Despite this result almost matches the regret obtained by running UCRL directly on the latent MDP, the regret analysis requires ergodicity of the MDP. One of the main open questions is whether the spectral clustering method could still provide "useful" clusterings when the state space is not fully visited (i.e., in case of non-ergodic MDP), so that observations are properly clustered where it is actually needed to learn the optimal policy. We can provide a partial answer in the case of deterministic ROMDPs. In fact, despite not being ergodic, in this case we can first uniformly explore the latent space and collect sufficient number of sample to find the exact clustering and reduce the large MDP to the latent MDP and then apply UCRL on the latent MDP. These two-phase algorithm would suffer a constant regret of pure exploration at the beginning and regret of $\widetilde{\mathcal{O}}(D_{\mathcal{X}} X \sqrt{AN})$ due to UCRL in the second phase.

## References


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

Anandkumar, Animashree, Hsu, Daniel J, and Kakade, Sham M. A method of moments for mixture models and hidden markov models. In *COLT*, volume 1, pp. 4, 2012.

Azizzadenesheli, Kamyar, Lazaric, Alessandro, and Anandkumar, Animashree. Open problem: Approximate planning of pomdps in the class of memoryless policies. *arXiv preprint arXiv:1608.04996*, 2016a.

Azizzadenesheli, Kamyar, Lazaric, Alessandro, and Anandkumar, Animashree. Reinforcement learning of pomdps using spectral methods. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, 2016b.

Cesa-Bianchi, Nicolo, Gentile, Claudio, and Zappella, Giovanni. A gang of bandits. In *Advances in Neural Information Processing Systems*, pp. 737–745, 2013.

Gentile, Claudio, Li, Shuai, and Zappella, Giovanni. Online clustering of bandits. In *ICML*, pp. 757–765, 2014.

Gheshlaghi-Azar, Mohammad, Lazaric, Alessandro, and Brunskill, Emma. Sequential transfer in multi-armed bandit with finite set of models. In *Advances in Neural Information Processing Systems 26*, pp. 2220–2228. 2013.

Gopalan, Aditya, Maillard, Odalric-Ambrym, and Zaki, Mohammadi. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.

Jaksch, Thomas, Ortner, Ronald, and Auer, Peter. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Jiang, Nan, Krishnamurthy, Akshay, Agarwal, Alekh, Langford, John, and Schapire, Robert E. Contextual decision processes with low bellman rank are pac-learnable. *arXiv preprint arXiv:1610.09512*, 2016.

Krishnamurthy, Akshay, Agarwal, Alekh, and Langford, John. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pp. 1840–1848, 2016.

Li, Lihong, Walsh, Thomas J., and Littman, Michael L. Towards a unified theory of state abstraction for mdps. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics (ISAIM-06)*, 2006.

Littman, Michael L. 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, pp. 238–245, Cambridge, MA, USA, 1994. MIT Press. ISBN 0-262-53122-4.

Maillard, Odalric-Ambrym and Mannor, Shie. Latent bandits. In *Proceedings of the Thirty-First International Conference on Machine Learning (ICML'14)*, 2014.

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

Ortner, Ronald. Adaptive aggregation for reinforcement learning in average reward markov decision processes. *Annals of Operations Research*, 208(1):321–336, 2013.

Osband, Ian, Russo, Dan, and Van Roy, Benjamin. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

Puterman, Martin L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

Strehl, Alexander L., Li, Lihong, and Littman, Michael L. Reinforcement learning in finite mdps: Pac analysis. *J. Mach. Learn. Res.*, 10:2413–2444, 2009.

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

# A. Proofs of Section 3

## A.1. Proof of Lemma 4

At each epoch $k$, we estimate the factor matrix $V_2^{(l)}$ (for all $l \in [A]$) using all the samples collected during the previous epoch according to policy $\widetilde{\pi}^{(k)}$. In order to simplify the notation, in the following we remove the dependency on $k$, even if all the quantities should be intended as specifically computed at the beginning of epoch $k$.

In order to bound the empirical error of the moment estimators, we need to consider the properties of the Markov chain generated by policy $\widetilde{\pi}$ and the fact that a single continuous trajectory is observed. In particular, we have to carefully consider the mixing time of the underlying Markov chain (the amount of time it takes that the underlying Markov chain converges to its stationary distribution) and exploit the martingale property of the trajectory. This problem has been previously studied by Azizzadenesheli et al. (2016b) for the general case of partially observable MDPs. Since ROMDPs are a special case of POMDPs, we can directly rely on the following general concentration inequality.

**Proposition 1** (Theorem 11 (Azizzadenesheli et al., 2016b) POMDP concentration bound)**.** *Consider a sequence of $\nu$ observations $\{y_1, \dots, y_\nu\}$ obtained by executing a policy $\widetilde{\pi}$ in a POMDP starting from an arbitrary initial hidden state. For any action $l \in [A]$, $\nu^{(l)}$-length sequence $b^{(l)} = \{(y_{t-1}, y_t, y_{t+1}); a_t = l\}$, and any c-Lipschitz[6] matrix valued function $\Phi(\cdot) : b^{(l)} \to \mathbb{R}^{Y \times Y}$, we have*

$$\left\| \Phi(b^{(l)}) - \mathbb{E}[\Phi(b^{(l)})] \right\|_2 \leq \frac{G(\widetilde{\pi})}{1 - \theta(\widetilde{\pi})} \left(1 + \frac{1}{\sqrt{2} c (\nu^{(l)})^{\frac{3}{2}}}\right) \sqrt{8 c^2 \nu^{(l)} \log\left(\frac{2Y}{\delta}\right)}$$

*with probability at least $1 - \delta$, where $G(\widetilde{\pi})$ and $\theta(\widetilde{\pi})$ are, respectively, the geometric ergodicity and the contraction coefficient of the underlying Markov chain on the hidden states (they define how fast the underlying Markov chain converges to its stationary distribution), and the expectation is with respect to the distribution of initial state equals to the stationary distribution.*

The parameters $1 \leq G(\pi^{(k)}) < \infty$ and $0 \leq \theta(\pi^{(k)}) < 1$ are well defined for Markov chain and shows the state distribution of the Markov chain convergence to its stationary distribution (if such distribution exists) with rate of $G(\pi)\theta(\pi)^t$. (the lower $G(\pi)$ and $\theta(\pi)$ give the lower mixing for corresponding Markov chain.)

Given the ergodicity assumption (Asm. 1) under any policy, we can apply Proposition 1 to bound the errors for both second and third order moments. For any $\{p, q, r\}$ a permutation of set $\{1, 2, 3\}$

$$\|\widehat{K}_{p,q}^{(l)} - K_{p,q}^{(l)}\|_2 \leq G(\pi) \frac{1 + \frac{1}{\sqrt{2} c (\nu^{(k)}(l))^{\frac{3}{2}}}}{1 - \theta} \sqrt{8 c^2 \nu^{(k)}(l) log(\frac{2Y}{\delta})} \tag{10}$$

$$\|\widehat{M}_{p,q,r}^{(l)} - M_{p,q,r}^{(l)}\|_2 \leq G(\pi) \frac{1 + \frac{1}{\sqrt{2} c (\nu^{(k)}(l))^{\frac{3}{2}}}}{1 - \theta} \sqrt{8 c^2 \nu^{(k)}(l) log(\frac{2Y^{1.5}}{\delta})} \tag{11}$$

with probability at least $1 - \delta$. At this point we can proceed with applying the robust tensor power method proposed in (Anandkumar et al., 2012) to recover $V_2^{(l)}$ and obtain the guarantees of Lemma 5 of Azizzadenesheli et al. (2016b) through Proposition 6, where $c = \frac{1}{\nu^{(k)}(l)}$. We report a more detailed version of the statement of Lemma 4.

**Lemma 6** (Concentration Bounds)**.** *The robust power method of Anandkumar et al. (2012) applied to tensor $\widehat{M}_3^{(l)}$ returns the $X^{(l)}$ columns of matrix $V_2^{(l)}$ with the following confidence bounds*

$$\left\| [V_2^{(l)}]_{(\cdot|i)} - [\widehat{V}_2^{(l)}]_{(\cdot|i)} \right\|_2 \leq \epsilon_3^{(l)} = C_O^l \sqrt{\frac{\log(2Y^{3/2}/\delta)}{\nu^{(l)}}} := \mathcal{B}_O^{(l)} \tag{12}$$

[6] under the Hamming metric

*if*

$$\nu^{(l)} \geq \overline{N} := \max_{\pi} \left( \frac{4}{\omega^{(l)}_{\widetilde{\pi}_{\min}} \min_{m\in\{1,2,3\}} \{\sigma^2_{\min}(V_m^{(l)})\}} \right)^2 \log(2\frac{(Y^{1.5})}{\delta})\Theta^{(l)} \tag{13}$$

$$\Theta^{(l)} := \max \left\{ \frac{16(X^{(l)})^{\frac{1}{3}}}{C_1^{\frac{2}{3}}(\omega^{(l)}_{\widetilde{\pi}_{\min}})^{\frac{1}{3}}}, 4, \frac{2\sqrt{2}X^{(l)}}{C_1^2\omega^{(l)}_{\min} \min_{m\in\{1,2,3\}} \{\sigma^2_{\min}(V_m^{(l)})\}} \right\}, \tag{14}$$

*with probability at least $1-\delta$, where $C_1$ is a problem-independent constants and $\omega^{(l)}_{\widetilde{\pi}_{\min}} := \min_{i\in\mathcal{X}^{(l)}} \mathbb{P}_{\widetilde{\pi}}(x=i|a=l)$ where the minimization is over non-zero probabilities. In addition, the $\sigma_{\min}(\cdot)$ operator returns the smallest non-zero singular value of its input matrix. The values of the error $\mathcal{B}_O^{(l)}$ under policy $\widetilde{\pi}$ is defined (see Eq.29 of Azizzadenesheli et al. (2016b))*

$$\mathcal{B}_O^{(l)} := G(\widetilde{\pi}) \frac{4\sqrt{2}+2}{(\omega^{(l)}_{\widetilde{\pi}_{\min}})^{\frac{1}{2}}(1-\theta(\widetilde{\pi}))} \sqrt{\frac{\log(2\frac{(2Y)}{\delta})}{\nu^{(l)}}} + \frac{8\widetilde{\epsilon}^{(l)}}{\omega^{(l)}_{\widetilde{\pi}_{\min}}}, \tag{15}$$

*where*

$$\widetilde{\epsilon}^{(l)} \leq \frac{2\sqrt{2}G(\widetilde{\pi})\frac{2\sqrt{2}+1}{1-\theta(\widetilde{\pi})}\sqrt{\frac{\log(\frac{2(Y^{\frac{3}{2}})}{\delta})}{\nu^{(l)}}}}{((\omega^{(l)}_{\widetilde{\pi}_{\min}})^{\frac{1}{2}} \min_{m\in\{1,2,3\}} \{\sigma_{\min}(V_m^{(l)})\})^3} + \frac{\left(64G(\widetilde{\pi})\frac{2\sqrt{2}+1}{1-\theta(\widetilde{\pi})}\right)}{\min_{m\in\{1,2,3\}} \{\sigma^2_{\min}(V_m^{(l)})\}(\omega^{(l)}_{\widetilde{\pi}_{\min}})^{1.5}} \sqrt{\frac{\log\left(2\frac{Y^{\frac{3}{2}}}{\delta}\right)}{\nu^{(l)}}},$$

### A.2. Rank recovery

Lemma 6 holds when the rank of matrix $V_2^{(l)}$ is known in advance. While this is not the case in practice, here we show how one can estimate the rank $r = \left|\mathcal{X}^{(l)}_{\pi^{(k)}}\right|$ of $V_2^{(l)}$. Given the expansiveness of latent MDP (Asm. 2), we have that for any policy $\pi$ and any action $l$, $\left|\mathcal{X}^{(l)}_{\pi}\right| \leq \left|\overline{\mathcal{X}}^{(l)}_{\pi}\right|$. The rank of the second moment matrix $K^{(l)}_{2,3}$ is then $\min\{\left|\mathcal{X}^{(l)}_{\pi^{(k)}}\right|, \left|\overline{\mathcal{X}}^{(l)}_{\pi^{(k)}}\right|\} = r$, which also corresponds to the number of non-zero columns in matrix $V_2^{(l)}$. We can then try to estimate $r$ through the estimate second moment $\widehat{K}^{(l)}_{2,3}$, which according to Eq. 10, estimates $K^{(l)}_{2,3}$ up to an additive error $\epsilon^{(l)}_{2,3}$ that decreases as $O(\sqrt{\frac{1}{\nu^{(l)}}})$. This means that the highest perturbation over its singular values is also at most $O(\sqrt{\frac{1}{\nu^{(l)}}})$. We introduce a threshold function $g^{\epsilon}(\nu^{(l)})$ that satisfies the condition

$$\epsilon^{(l)}_{2,3} \leq g^{\epsilon}(\nu^{(l)}) \leq 0.5\sigma_r, \tag{16}$$

where $\sigma_r$ is the smallest non-zero singular value of $K^{(l)}_{2,3}$. We then perform a SVD of $\widehat{K}^{(l)}_{2,3}$ and discard all singular values with value below the threshold $g^{\epsilon}(\nu^{(l)})$. Therefore, with probability at least $1-\delta$, the number of remaining singular values is equal to the true rank $r$. We are left with finding a suitable definition for the threshold function $g^{\epsilon}$. From the condition on Eq. 16, we notice that we need $g^{\epsilon}$ to be smaller than a fixed value (RHS) and, at the same time, greater than a decreasing function of order $\mathcal{O}(\sqrt{\frac{1}{\nu^{(l)}}})$ (LHS). Then it is natural to define

$$g^{\epsilon}(\nu^{(k)}(l)) = \frac{g}{\nu^{(k)}(l)^{0.5-\epsilon}}$$

for a suitable $g > 0$ and with $0 < \epsilon < 0.5$. Therefore there is a number $N_0^{(l)}$ such that for all $\nu^{(l)} \geq N_0^{(l)}$ the condition on Eq. 16 is satisfied and Lemma 6 holds. Therefore we restate the sample complexity in Lemma 6 by adding the extra term to

$$\overline{N} \leftarrow \overline{N} + N_0(l).$$

Let $\overline{N}_{\max}$ denotes the maximum of this threshold for any action and policy.

### A.3. Proof of Lemma 1

Under policy $\pi$, Fig. 2-*right* shows the structure of $V_2^{(l)}$. Given action $l$, the matrix $V_2^{(l)}$ contains $X^l_{\pi}$ columns and each column corresponds to a column in emission matrix (up to permutation). We showed that the knowledge about a column of

$V_2^{(l)}$ reveals part of the corresponding column in emission matrix, the entries with non-zero $\pi(y|l)$. The policy, in general, partitions the observation space to at most $A$ partitions, $\mathcal{Y}_l \forall l \in \mathcal{A}$ and maps each partition to an action. It means that when we condition on an action, e.g., $l$, we restrict ourselves to the part of observation space $\mathcal{Y}_l$ and the input to the spectral learning algorithm is set $\mathcal{Y}_l$. Therefore, the algorithm is able to partition this set to $X_\pi^{(l)}$ partition. Because of the unknown permutation over columns of $V_2^{(l)}$ for different actions, we are not able to combine the resulting clustering give different action. If we enumerate over actions, we end up with $A$ partition $\mathcal{Y}_l$ and then we partition each set $\mathcal{Y}_l$ to at most $X$(upper bound on $X_\pi^{(l)}$), as a consequence, we might end up with at most $XA$ disjoint clusters.

### A.4. Proof of Lemma 2

In Eq. 3 we show that matrix $M_2^{(l)}$ is a symmetric matrix and has the following representation;

$$M_2^{(l)} \sum_{i \in \mathcal{X}_\pi^{(l)}} \omega_\pi^{(l)}(i) [V_2^{(l)}]_{:,i} \otimes [V_2^{(l)}]_{:,i}.$$

As long as $V_2^{(l)}]_{:,i}$ for $i \in \mathcal{X}_\pi^{(l)}$ are orthogonal vectors, this matrix has rank of $X_\pi^{(l)}$ with the following eigendecomposition;

$$M_2^{(l)} = U \Sigma U^\mathsf{T}$$

where the matrix $U$, up to permutation, is the orthonormal version of $V_2^{(l)}$, and $\Sigma$ is a diagonal matrix of rank $X_\pi^{(l)}$ with diagonal entries equal to $\omega_\pi^{(l)}$ multiplied by the normalization factors. Let's consider the $i$'th and $j$'th nonzero diagonal entries of matrix $\Sigma$, $\sigma_i$ and $\sigma_j$, with eigenvectors of $U_i, U_j$, i.e., $M_2^{(l)} U_i = \sigma_i U_i$ , $M_2^{(l)} U_j = \sigma_j U_j$. In the case of no eigengap, i.e., $\sigma_i = \sigma_j$, for any $0 \leq \lambda \leq 1$ we have $M_2^{(l)}(\lambda U_i + (1-\lambda) U_j) = \sigma_i(\lambda U_i + (1-\lambda) U_j) = \sigma_j(\lambda U_i + (1-\lambda) U_j)$. Therefore, any direction in the span of $span(U_i, U_j)$ is an eigenvector and the matrix decomposition is not unique, and we can not learn the true $V_2^{(l)}$. We relax this issue by deploying tensor decomposition of higher order moments.

### A.5. Proof of Lemma 3

This lemma has been deeply studied in (Anandkumar et al., 2014) and we refer to it.

## B. Proof of Theorem 1

We begin with decomposing the regret its components

$$\textit{Reg}_N = N\eta^* - \sum_{t=1}^{N} r_t$$

Recall $r_t$ is random variable reward that the agent receives at time $t$. At time $N$, $N(x,a)$ denotes the total number of time that pair of hidden state $x$ and action $a$ coincide together so far. Therefore, given set of $N(x,a)$'s and Hoeffding's inequality

$$\sum_{t=1}^{N} r_t \geq \sum_{x \in \mathcal{X}, a \in \mathcal{A}} N(x,a) \bar{r}(x,a) - \sqrt{N \log(\frac{1}{\delta})}$$

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

$$\textit{Reg}_N \leq N\eta^* - \sum_{x \in \mathcal{X}, a \in \mathcal{A}} N(x,a) \bar{r}(x,a) + \sqrt{N \log(\frac{1}{\delta})}$$

Without loss of generality, assume the time N is the time that the agent starts $(K+1)$'th epoch then

$$\textit{Reg}_N \leq \sum_{k=1}^{K} \Delta_k + \sqrt{N \log(\frac{1}{\delta})}$$

$\Delta_k$ is the regret of $k$'th epoch.

### B.1. Failing Confidence

Let's first consider the regret due to failing the confidence interval of estimated Aux-MDP parameters and failing the spectral method. At the beginning of each epoch, the spectral method estimates columns of $V_2^{(l)}(k)$ matrices and clusters the observations to construct the auxiliary set $\widehat{\mathcal{S}}^{(k)}$. The clustering procedure, with high probability, does not make a mistake and does not cluster two observation which belong to two different hidden states.

We first notice that if we redefine the confidence intervals in lemma 4 by substituting the term $(1/\delta)$ with $(At^6/\delta)$, we are able to show that at any time instants $t$, the clustering procedure does not make any mistake with probability at least$1-24\delta/t^6$. On the other hand, given set $\widehat{\mathcal{S}}^{(k)}$, we are going to show that the confidence intervals of estimated Aux-MDP parameters hold with probability at least $1-\delta/15t^6$.

The whole confidence bounds in the regret analysis holds when both of these confidence bound do not break down and hold. In this case $\mathbb{P}(M \notin \mathcal{M}^t) \leq 24\delta/t^6 + \delta/15t^6 \leq 25\delta/t^6$.

$$Reg_N^{fail} = \sum_{k=1}^{K} \left( \sum_{t=N^{(k)}}^{t=N^{(k+1)}} (\eta^* - \bar{r}_t)\, \mathbb{I}(M \notin \mathcal{M}^{(k)}) \right)$$

where $\mathcal{M}^{(t)}$ denotes the set of admissible Aux-MDPs according to the samples available at time $N^{(k)}$ (the end of epoch $k-1$). From sample complexity analysis in lemma 4, for large $N$ when $\bar{N}_{SM} \leq N^{1/4}$ we have

$$Reg^{fail} \leq \sum_{t=1}^{\lfloor N^{1/4} \rfloor} t\mathbb{1}(M \notin \mathcal{M}^{(t)}) + \sum_{\lfloor N^{1/4} \rfloor + 1}^{N} t\mathbb{1}(M \notin \mathcal{M}^{(t)}) \leq \sqrt{N} + \sum_{\lfloor N^{1/4} \rfloor + 1}^{N} t\mathbb{1}(M \notin \mathcal{M}^{(t)})$$

As it mentioned before $\mathbb{P}(M \notin \mathcal{M}^t) \leq 25\delta/t^6$. Therefore we are left with bounding the last term.

$$\sum_{t=\lfloor N^{1/4} \rfloor + 1}^{N} \frac{25}{t^6} \leq \frac{25}{N^{6/4}} + \int_{\lfloor N^{1/4} \rfloor}^{\infty} \frac{25}{t^6} dt = \frac{25}{N^{6/4}} + \frac{25}{5N^{5/4}} \leq \frac{150A}{N^{5/4}} = \frac{30A}{N^{5/4}},$$

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

### B.2. Per Epoch regret

For the rest of the proof, let assume $M \in \mathcal{M}^{(t)}$ holds. Moreover, assume that for the optimization Eq. 9 we use the Alg. 3.

**Algorithm 3** *Finding optimistic policy*

**Input:** Set $\mathcal{S}$, $\mathcal{A}$ and estimated $\widehat{p}(\cdot|s,a)$, $\widehat{R}(s,a)$
Reorder set $\widehat{\mathcal{S}} = \{s'_1, s'_2, \ldots, s'_{\widehat{S}}\}$ such that $u(s'_1) \geq u(s'_2) \geq \ldots \geq u(s'_S)$
**for all** $s$ **and** $a$
  **Set:**
    $p(s'_1) := \min\{1, \widehat{T}(s'_1|\widetilde{s}, a) + \frac{d(s,a)}{2}\}$
    $p(s'_j) := \widehat{T}(s'_j|s,a)\ \forall j > 1$
  set $\ell := \widehat{S}^{(k)}$
  **While** $\sum_j p(s'_j) > 1$ **do**
    Reset $p(s'_\ell) := \max\{0, 1 - \sum_{j \neq \ell} p(s'_j)\}$
    Set $\ell = \ell - 1$

Optimization in Eq. 9 is equal to finding an optimal policy for an augmented MDP. To solve this optimization, assign estimated rewards to their highest admissible values $\widetilde{r}(\overline{s}, a) = \widehat{r}(s,a) + d'(s,a)$. Define value function $u(s)$ for all $s$. Then

apply value iteration in Alg. 3. It reduces the optimization problem to: $\forall s \in \mathcal{S}^{(k)}$

$$u_0(s) = 0,$$
$$u_{t+1}(s) = \max_a \{\widetilde{r}(s,a) + \max_{p(\cdot)\in\mathcal{P}^k(s,a)} \{\sum_{s'} p(s') \cdot u_i(s')\}\} \tag{17}$$

where the set of vectors $\mathcal{P}^k(s,a)$ is set of admissible vectors $p(\cdot|s,a)$ in the confidence interval. The iterative algorithm in Eq. 17 is another view of Poisson Equation to find optimal policy. As it is shown in Eq. 17, this iterative procedure solves a simple optimization problem at each iteration, and updates the value vectors. The algorithm in Alg. 3 provides an efficient way to handle the optimization in 9 with computation cost of $\mathcal{O}((\widehat{\mathcal{S}}^{(k)})^3 A)$ per epoch. The computation cost for baseline algorihtm, UCRL2, is order of $\mathcal{O}((Y)^3 A)$ per epoch and it does not decrease while in the SL-UC when $\mathcal{S} = X$ the computation complexity decreases to $\mathcal{O}((X)^3 A)$ which is order of magnutude smaller than the UCRL2. For more detailed analysis see (Puterman, 2014), (Jaksch et al., 2010).

Assign the stopping criterion of $\frac{1}{\sqrt{N^{(k)}}}$ to iterative update Eq. 17 which is the maximum improvement in the Q-functions then we have

$$\Delta^{(k)} = \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \nu^{(k)}(s,a)(\eta^* - \bar{r}(s,a)) \leq \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \nu^{(k)}(s,a)(\widetilde{\eta} - \bar{r}(s,a)) + \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}$$

$$= \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \nu^{(k)}(s,a)(\widetilde{\eta} - \widetilde{r}(s,a)) + \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \nu(s,a)(\widetilde{r}(s,a) - \bar{r}(s,a)) + \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}$$

$$\leq \vec{\nu}^k(\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)} + \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \nu(s,a)(\widetilde{r}(s,a) - \bar{r}(s,a)) + 2\sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}$$

$$\leq \vec{\nu}^k(\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)} + \sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} 2\nu(s,a)(\widetilde{r}(s,a) - \widehat{r}(s,a)) + 2\sum_{s\in\mathcal{S}^{(k)},a\in\mathcal{A}} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}$$

Therefore

$$\Delta^{(k)} \leq \vec{\nu}^k(\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)} + \sum_{s,a} 2\nu(s,a)\sqrt{\frac{7\log(2\widehat{S}^{(k)}AN^{(k)}/\delta)}{2\max\{1,N^{(k)}(s,a)\}}} + 2\sum_{s,a}\frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}$$

$$\leq \underbrace{\vec{\nu}^k(\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)}}_{(a)} + \underbrace{2\sum_{s,a}\frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}}_{(b)} + \underbrace{\left(\sqrt{14\log(2\widehat{S}^{(k)}AN^{(k)}/\delta)}\right)\sum_{s,a}\frac{\nu^{(k)}(s,a)}{\sqrt{2\max\{1,N^{(k)}(s,a)\}}}}_{(b')}$$

Because the sum over rows of transition matrix $\widetilde{P}^{(k)}$ is one we can substitute vector $\vec{u}^{(k)}$ with vector $\bar{u}^{(k)}$ where for $s \in \mathcal{S}^{(k)}:\ \bar{u}^{(k)}(s) = \vec{u}^{(k)}(s) - \frac{\min_{s'}\vec{u}^{(k)}(s') - \max_{s'}\vec{u}^{(k)}(s')}{2}$.
Now $(a)$ can be decomposed as follows:

$$(a) := \vec{\nu}^{(k)}\left(\widetilde{P}^{(k)} - P^{(k)} + P^{(k)} - I\right)\bar{u}^{(k)} = \underbrace{\vec{\nu}^{(k)}\left(\widetilde{P}^{(k)} - P^{(k)}\right)\bar{u}^{(k)}}_{(c)} + \underbrace{\vec{\nu}^{(k)}\left(P^{(k)} - I\right)\bar{u}^{(k)}}_{(d)}$$

and for $(c)$

$$
\begin{aligned}
(c) &:= \sum_{s\in\mathcal{S}^{(k)}}\sum_{s'\in\mathcal{S}^{(k)}} \nu^{(k)}(s,\widetilde{\pi}^{(k)})\left(\widetilde{f}_T(s'|s,\widetilde{\pi}^{(k)}(s)) - f_T(s'|s,\widetilde{\pi}^{(k)}(s))\right)\bar{u}^{(k)}(s') \\
&\leq \sum_{s\in\mathcal{S}^{(k)}} \nu^{(k)}(s,\widetilde{\pi}^{(k)})\|\widetilde{f}_T(\cdot|s,\widetilde{\pi}^{(k)}(s)) - f_T(\cdot|s,\widetilde{\pi}^{(k)}(s))\|_1\|\bar{u}^{(k)}\|_\infty \\
&\leq \sum_{s\in\mathcal{S}^{(k)}} \nu^{(k)}(s,\widetilde{\pi}^{(k)})2\sqrt{\frac{14\widehat{S}^{(k)}\log(2AN^{(k)}/\delta)}{2\max\{1,N(s,a)\}}}.\frac{D_{\widehat{\mathcal{S}}^{(k)}}}{2} \\
&\leq D_{\widehat{\mathcal{S}}^{(k)}}\sqrt{14\widehat{S}^{(k)}\log(2AN^{(k)}/\delta)}\sum_{s\in\mathcal{S}^{(k)},a}\frac{\nu^{(k)}(s,a)}{\sqrt{\max\{1,N^{(k)}(s,a)\}}}
\end{aligned}
$$

### B.3. Grouping Rate

Currently, we studied the regret of each epoch separately. In the designed RL algorithm, in the exploration phase, the agent chooses the optimal policy with respect to the optimistic policy. In theory, the behavior of this policy on the true environment is not well known but we can argue that, in practice, for such models, e.g. Rich Observation MDP, where the number of observations is an order of magnitude larger than the number hidden states, the policy maps a significant portion of observation of a hidden state to an action. It means that it is not even the case that the policy maps just small number of observations of a hidden state to a specific action while it maps other observation to other actions. In other world, if the policy maps one observation to an action, then most likely it maps considerable portion of other observations from the same state to the same action. Let $\alpha_p$ denotes the smallest probability of this portion given their hidden state, $\alpha_p := \min_{k\in[K],x\in\mathcal{X},a\in\mathcal{A}}\sum_y \mathbb{P}(y|x)\mathbb{1}(a=\pi^{(k)}(\cdot|y))$ . Potentially this value can get as low as $O_{\min}$ but in practice, it is considerably large.

Now we need need to know how fast the set $\mathcal{S}^{(k)}$ converges to set $\mathcal{X}$, in other work how fast is the clustering process. Given Eq. [7], to cluster any observation $y$, make sure its value in $V_2$ is non zero, sufficient number of sample is required. It means that the number of samples of the corresponding action, in a given epoch $k$, $\nu^{(k)}(l)$, should satisfy:

$$
f_O(y|x(y)) \geq 2C_O^{(k)}(l)\sqrt{\frac{\log(1/\delta)}{\nu^k(l)}}
$$

which means the required number of samples for corresponding action is $\nu(l) \geq 4C_O^{(k)}(l)\frac{\log(1/\delta)}{(f_O(y|x(y)))^2}$. Let's call this number for a particular observation as $\overline{\nu}^{(k)}(y)$. Let's define $\tau_M = \max_\pi \tau_{M,\pi} = \max_\pi \mathbb{E}[\tau(x\to x)]$, where $\tau(x\to x)$ is a random variable and represents the passing time of starting from a state $x$ and ending to the same state according to policy $\pi$. By Markov inequality, the probability that it takes more than $2\tau_M$ time step to from first visit of state $x$ to its second visit is at most $1/2$. Given the definition of $\alpha_p$, it is clear that if the action $l$ is taken in state $x$ then, this action will be taken at state $x$ for $\alpha_p$ portion of the time. If we divide the episode of length $\nu$ into $\nu\alpha_p/2\tau_M$ intervals of length $2\tau_M/\alpha_p$, we have that within each interval we have a probability of $1/2$ to observe a sample from state $x$ and take a particular action. Therefore, the lower bound on the average number of time that the agent takes any feasible action (action with nonzero probability) is $\nu\alpha_p/4\tau_M$ samples. Thus from Chernoff-Hoeffding, we obtain that the number of samples of any feasible action in the epoch with length $\nu$ is as follows;

$$
\forall x\in\mathcal{X},\forall l\in range\{\widetilde{\pi}(\cdot|x)\};\ \nu(l) \geq \frac{\nu\alpha_p}{4\tau_M} - \sqrt{\frac{\nu\alpha_p\log(XA/\delta)}{2\tau_M}}
$$

with probability at least $1-\delta$.

At this point, we can derive a lower bound on the length of the episode that guarantee the desired number of samples to reveal the identity of any observation is reached. For observation $y$, we solve

$$
\frac{\nu\alpha_p}{4\tau_M} - \sqrt{\frac{\nu\alpha_p\log(XA/\delta)}{2\tau_M}} \geq \overline{N}(y)
$$

and we obtain the condition

$$\sqrt{\nu} \geq \sqrt{\frac{2\tau_M}{\alpha_p}\log(XA/\delta)} + \sqrt{\frac{2\tau_M}{\alpha_p}\log(XA/\delta) + \frac{4\tau_M}{\alpha_p}\overline{N}(y)},$$

which can be simplified to

$$\nu \geq \overline{\nu}(y) := \frac{4\tau_M}{\alpha_p}\left(\overline{N}(y) + \log(XA/\delta)\right). \tag{18}$$

With the same argument in Appendix [D] in (Azizzadenesheli et al., 2016b) the number of required epochs to reveal observation $y$ is $\widetilde{K}(y) \leq AY\log_2(\overline{\nu}(y)) + 1$. It means that the agent before epoch $k_1 := \min_y \widetilde{K}(y)$ encounters the problem with state dimensionality of $Y$. The amount of time step required to reach $k_1$ is, $4AY\min_y \overline{\nu}(y)$, and after epoch $k_2 := \max_y \widetilde{K}(y)$ it encounters problem with dimensionality of $X$ which takes $4AY\max_y \overline{\nu}(y)$ time steps. For the epochs between $k_1$ and $k_2$, the agent deals with the problem of size $\widehat{S}^{(k)}$.

**Regret due to sample complexity**

To deploy the spectral method, we need sufficient amount of samples to make use of spectral methods. In lemma 4 we show that a minimum number of $\bar{N}_{SM}$ samples is required to start the spectral method. With same analysis in the previous section, there is an epoch $k_{SM}$ with high probability the sample complexity is satisfied. By substituting the value of $k_1 \leftarrow \max\{k_1, k_{SM}\}$ and $k_2 \leftarrow \max\{k_2, k_{SM}\}$ the analyses remain same.

### B.4. Overall Regret

In this section we sum up all mentioned regret sources. For $(b)$ by applying Lemma [19] in (Jaksch et al., 2010)

$$\sum_k \sum_{s,a} 2\frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}} \leq \sum_k \sum_a 2\frac{\nu^{(k)}(a)}{\sqrt{N^{(k)}}} \leq 2(\sqrt{2}+1)\sqrt{N}$$

for $(c)$ :

$$\begin{aligned}
&\sum_k \left( D_{\widehat{\mathcal{S}}^{(k)}} \sqrt{14\widehat{S}^{(k)}\log(2AN^{(k)}/\delta)} \sum_{s\in\mathcal{S}^{(k)},a} \frac{\nu^{(k)}(s,a)}{\sqrt{\max\{1,N^{(k)}(s,a)\}}} \right) \\
&\leq \sum_{k=1}^{k_1} \left( D_{\mathcal{Y}} \sqrt{14Y\log(2AN^{(k)}/\delta)} \sum_{y,a} \frac{\nu^{(k)}(y,a)}{\sqrt{\max\{1,N^{(k)}(y,a)\}}} \right) \\
&+ \sum_{k=k_1+1}^{k_2} \left( D_{\widehat{\mathcal{S}}^{(k)}} \sqrt{14\widehat{S}^{(k)}\log(2AN^{(k)}/\delta)} \sum_{s\in\mathcal{S}^{(k)},a} \frac{\nu^{(k)}(s,a)}{\sqrt{\max\{1,N^{(k)}(s,a)\}}} \right) \\
&+ \sum_{k=k_2+1} \left( D_{\mathcal{X}} \sqrt{14X\log(2AN^{(k)}/\delta)} \sum_{x,a} \frac{\nu^{(k)}(x,a)}{\sqrt{\max\{1,N^{(k)}(x,a)\}}} \right)
\end{aligned}$$

where $D_{\widehat{\mathcal{S}}^{(k)}}$ is the diameter of MDP under $\widehat{\mathcal{S}}^{(k)}$ configuration. This part of regret can be simplified a bit more and can be shown that is loosely upper-bounded as

$$\begin{aligned}
&D_{\mathcal{Y}} Y\sqrt{14AN^{(k_2)}\log(2AN^{(k_2)}/\delta)} \\
&+ D_X X\sqrt{14AN\log(2A(N-N^{(k_2)})/\delta)}
\end{aligned}$$

where the first part is constant number.

And for $(d)$

$$\sum_k \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} (f_T(\cdot|s_t,a_t) - \vec{e}_{s_t})\,\bar{u}^{(k)} =$$
$$\sum_k \left( \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} f_T(\cdot|s_t,a_t) - \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \vec{e}_{s_{t+1}} + \vec{e}_{s_{t_{(k+1)}}} - \vec{e}_{s_{N^{(k)}}} \right) \bar{u}^{(k)}$$

Let's define $\zeta_t := (f_T(\cdot|s_t,a_t) - \vec{e}_{s_{t+1}})\bar{u}^{(k)}$ then we have

$$\sum_{k=1}^{k_2} \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + \bar{u}^{(k)}(s_{N^{(k+1)}}) - \bar{u}^{(k)}(s_{N^{(k)}}) + \sum_{k=k_2+1} \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + \bar{u}^{(k)}(s_{N^{(k+1)}}) - \bar{u}^{(k)}(s_{N^{(k)}})$$
$$\leq \sum_{k=1}^{k_2} \left( \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + D_{\mathcal{Y}} \right) + \sum_{k=k_2+1} \left( \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + D_{\mathcal{X}} \right)$$

As long as $|\zeta_t|$ is bounded by $D_{\widehat{\mathcal{S}}^{(k)}}$ and $\mathbb{E}[\zeta_t | s_{N^{(k)}}, a_{N^{(k)}}, \ldots, s_t, a_t] = 0$ this random variable is bounded martingale. Therefore Lemma [10] in (Jaksch et al., 2010) gives us

$$\sum_{t=1}^{t=N^{(k)}} \zeta_t = \sum_{t=1}^{t=N^{(k_2)}} \zeta_t + \sum_{t=N^{(k_2+1)}}^{t=N} \zeta_t \leq D_{\mathcal{Y}} \sqrt{2N^{(k_2)} \frac{5}{4} \log(\frac{8N}{\delta})} + D_{\mathcal{X}} \sqrt{2(N - N^{(k_2)}) \frac{5}{4} \log(\frac{8N}{\delta})}$$

with probability at least $1 - \frac{\delta}{12N^{5/4}}$.

Then the regret due to part $(d)$ is bounded by

$$\sum_k \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + D_{\widehat{\mathcal{S}}^{(k)}}$$
$$\leq D_{\mathcal{Y}} \sqrt{2N^{(k_2)} \frac{5}{4} \log(\frac{8N}{\delta})} + D_{\mathcal{Y}} A \log_2(8N^{(k_2)}/A) + D\mathcal{X} \sqrt{2(N - N^{(k_2)}) \frac{5}{4} \log(\frac{8N}{\delta})} + D_{\mathcal{X}} A \log_2(8N/A)$$

As we can see, at the beginning the agent suffers from huge regret due to the large MDP over space of observation. After collecting some samples, the agent starts to build new unbiased models which have smaller dimensionality and smaller diameter compared to model on $\mathcal{Y}$, i.e. $X \leq \widehat{S}^k \leq Y$ and $D_{\mathcal{X}} \leq D_{\widehat{\mathcal{S}}^k} \leq D_{\mathcal{Y}}$. At most at epoch $k_1 + 1$ the model that the agent deal with has lower dimension and the regret rate start to reduce even more. At epoch at most $k_2$ the agent totally identifies the surjective mapping $y \to x$ and then deal with the smaller model and then suffer from the regret of $\mathcal{O}(D_{\mathcal{X}} X \sqrt{AN})$.

**Remark 1.** *The values $D_{\mathcal{Y}}$ and $D_{\widehat{\mathcal{S}}^{(k)}}$ are the diameters of true model and Aux-MDPs in epoch $k$. These terms appear in the regret as an upper bound for $\max_s u^{(k)}(s) - \min_s u^{(k)}(s) := D^{(k)}$. Actually, in practice, because of the loose confidence bound and building the optimistic models on the top of that, we roughly have $D_{\mathcal{X}}{}^{(k)} << D_{\widehat{\mathcal{S}}^{(k)}} << D_{\mathcal{Y}}$*

### B.5. Cluster Aggregation

The spectral learning algorithm has been shown to efficiently cluster the observation set to an auxiliary state space of size $X \leq S \leq XA$. As long as different clusters are merged across epochs, we expect $\mathcal{S}^{(k)}$ to tends to $\mathcal{X}$, yet there is a change that it converges to a number of auxiliary states $S \neq X$. To make sure that the algorithm eventually converges to the hidden space $\mathcal{X}$, we include a further clustering technique. We adapt the idea of Gentile et al. (2014), Cesa-Bianchi et al. (2013) and the state aggregation analysis of Ortner (2013) and perform an additional step of *Reward and Transition Clustering*. In order to simplify the notation, in the following we remove the dependency on $k$, even if all the quantities should be intended as specifically computed at the beginning of epoch $k$.

We first recall that given any hidden state $x$ and any action $a$ we have $r(y,a) = r(x,a)$ (*reward similarity*) and $p(\cdot|y,a) = p(\cdot|x,a)$ for all observations $y \in \mathcal{Y}_x$ (*transition similarity*). The same similarity measures work for auxiliary states via

replacing observations with auxiliary states in the above definitions, i.e., given any hidden state $x$ and any action $a$ we have $r(s,a) = r(s,a)$ and $p(\cdot|s,a) = p(\cdot|x,a)$ for all auxiliary states $s \in \mathcal{S}$ that belong to hidden state $x$.[7] We also recall that high-probability confidence intervals can be computed for any $s \in \widehat{\mathcal{S}}$ any $a \in \mathcal{A}$ as

$$\begin{aligned} \|p(\cdot|s,a) - \widehat{p}(\cdot|s,a)\|_1 &\leq d(s,a) := \sqrt{\frac{14\widehat{S}\log(2AN/\delta)}{2\max\{1, N(s,a)\}}} \\ |\bar{r}(s,a) - \widehat{r}(s,a)| &\leq d'(s,a) := \sqrt{\frac{7\log(2\widehat{S}AN/\delta)}{2\max\{1, N(s,a)\}}}. \end{aligned} \tag{19}$$

At any epoch, we proceed by merging together all the auxiliary states in $\widehat{S}$ whose reward and transition confidence intervals overlap (i.e., $s$ and $s'$ are merged if the confidence interval $[\widehat{r}(s,a) \pm d_r(s,a)]$ overlaps with $[\widehat{r}(s',a) \pm d_r(s',a)]$ and $[\widehat{p}(\cdot|s,a) \pm d_p(s,a)]$ overlaps with $[\widehat{p}(\cdot|s',a) \pm d_p(s',a)]$) and construct a new set $\widetilde{\mathcal{S}}$. In practice, the set $\widetilde{S}$ is constructed by building a fully connected graph on $s \in \widehat{S}$ where each state $s$ as a node. The algorithm deletes the edges between the nodes when $|\widehat{r}(s,a) - \widehat{r}(s',a)| > d_r(s,a) + d_r(s',a)$ or $|\widehat{p}(\cdot|s,a) - \widehat{p}(\cdot|s',a)|_1 > d_p(s,a) + d_p(s',a)$. The algorithm temporarily aggregates the connected components of the graph and consider each disjoint component as a cluster. If the number of disjoint components is equal to $X$ then it returns $\widehat{\mathcal{S}}$ as the final hidden state $\mathcal{X}$, otherwise the original auxiliary state space $\widehat{\mathcal{S}}$ is preserved and the next epoch is started. Notice that if $s$ and $s'$ belong to the same hidden state $x$ then w.h.p. their confidence reward and transition intervals in Eq. 19 overlap at any epoch. Thus in general $\widetilde{S} \leq X$.

Let's define the reward gaps as follows (similar for the transitions) $\forall s, s' \in \widehat{\mathcal{S}}, \forall a \in \mathcal{A}$ and the corresponding $x, x'$

$$\begin{aligned} \gamma_r^a(s,s') = \gamma_r^a(x,x') &:= \left|\bar{r}(x,a) - \bar{r}(x',a)\right| = \left|\bar{r}(s,a) - \bar{r}(s',a)\right|, \\ \gamma_p^a(s,s') = \gamma_p^a(x,x') &:= \|p(\cdot|x,a) - p(\cdot|x',a)\|_1 = \|p(\cdot|s,a) - p(\cdot|s',a)\|_1. \end{aligned}$$

where $p(\cdot|x,a), p(\cdot|s,a) \in \widetilde{\Delta}_{\widehat{S}-1}$, where $\widetilde{\Delta}_{\widehat{S}-1}$ is $(\widehat{S}-1)$ dimensional simplex. To delete an edge between two states $s, s'$ belonging to two different hidden states, one of the followings needs to be satisfied for at least for one action

$$|\widehat{r}(s,a) - \widehat{r}(s',a)| > d_r(s,a) + d_r(s',a) \Rightarrow \gamma_r^a(s,s') > \sqrt{\frac{7\log(2\widehat{S}AN/\delta)}{2\max\{1, N(s,a)\}}} + \sqrt{\frac{7\log(2\widehat{S}AN/\delta)}{2\max\{1, N(s',a)\}}} \tag{20}$$

$$|\widehat{p}(\cdot|s,a) - \widehat{p}(\cdot|s',a)|_1 > d_p(s,a) + d_p(s',a) \Rightarrow \gamma_p^a(s,s') \sqrt{\frac{14\widehat{S}\log(2AN/\delta)}{2\max\{1, N(s,a)\}}} + \sqrt{\frac{14\widehat{S}\log(2AN/\delta)}{2\max\{1, N(s',a)\}}} \tag{21}$$

For simplicity, we proceed the analysis with respect to reward, the same analysis holds for transition probabilities. The Eq. 21 can be rewritten as follows;

$$\left(\frac{1}{\sqrt{\max\{1, N(s,a)\}}} + \frac{1}{\sqrt{\max\{1, N(s',a)\}}}\right)^{-1} \geq \sqrt{\frac{28\log(2\widehat{S}A^2N/\delta)}{2\gamma_r^a(s,s')^2}}$$

which hold when

$$\min\{N(s,a), N(s',a)\} > \frac{56\log(2\widehat{S}AN/\delta)}{\gamma_r^a(s,s')^2}. \tag{22}$$

This implies that after enough visits to the auxiliary states $s$ and $s'$, the two states would be split whenever belonging to different hidden states. We notice that as $\mathcal{S}$ becomes smaller, more and more samples from raw observations are clustered into the auxiliary states, thus making $N(s,a)$ larger and larger. Furthermore, we can expect that the transition gaps may become bigger and bigger as observations are clustered together.

The way that spectral method clusters the observation is effected by separability of observations' probability. But the clustering due to reward analysis (or transition or both) is influenced by the separability in reward function (or transition function or both) and depends on gaps. These two methods look at the clustering problem from different point of view, as a consequence, their combination speeds up the clustering task.

For simplicity we just again look at the reward function, same analysis applies to transition function as well.

[7] Notice that this holds since $\mathcal{S}$ is a "valid" clustering in high probability.

**Regret due to slowness of Reward Clustering (Transition Clustering)** Let $N_a^r(s,s')$ denote the required number of sample for each of $s$ and $s'$ to disjoint them.

$$N_a^r(s,s') := \frac{56\log(2\widehat{S}AN/\delta)}{\gamma_a(s,s')^2}$$

While the underlying Markov chain is ergodic, with high probability we can say at time step $N(s,s')$, at least for one action $\min\{N(s,a),N(s',a)\} \le \overline{N}_a(s,s')$ where

$$N_r(s,s') = \min_a\{\frac{4\tau_M}{\alpha_p}\left(N_a^r(s,s') + \log(YA/\delta)\right)\}$$

In the worse case analysis we might need to have $N_r = \max_{s,s'} N_r(s,s')$ samples, which corresponds to at most $AY\log(N_r)$ episode. At this time, the reward clustering procedure can output the exact mapping. This bound can be enhance even further by considering the reward function together with the transition process. With the same procedure we can define $N_p = \max_{s,s'} N_p(s,s')$ where

$$N_p(s,s') = \min_a\{\frac{4\tau_M}{\alpha_p}\left(N_a^p(s,s') + \log(YA/\delta)\right)\} \tag{23}$$

with

$$N_a^p(s,s') := \frac{112\widehat{S}\log(2\widehat{S}AN/\delta)}{\gamma_a(s,s')^2}$$

Again, in the worse case analysis we might need to have $N_p = \max_{s,s'} N_p(s,s')$ samples, which corresponds to at most $AY\log(N_r)$ episode. Therefore the number of required episode for the agent to declare the true mapping w.h.p., is $AY\log(\min\{N_r,N_p\})$.

## C. Additional Experiments

While the results reported in the main text are obtained on actual ROMDPs, here we test SL-UC on random MDPs with no explicit hidden space. The objective is to verify whether SL-UC can be used to identify (approximate) clusters. Since SL-UC in high probability only clusters observations that *actually* belong to the same hidden state, in this case SL-UC would reduce to run simple UCRL, as there is no two observations that can be *exactly* clustered. In order to encourage clustering, we half the (exact) confidence intervals in the attempt of trading off a small bias with a significant reduction in the variance. We compare the regret on three random MDPs with increasing number of states. As it is shown in Fig. 5, SL-UC is effective even in this scenario compared to UCRL and DQN. In fact, we see from Fig. 5-right that SL-UC is able to find clusters without compromising the overall regret. While the number of states now directly affects the performance of SL-UC, we see that it is more robust than the other algorithms and its regret is not severely affected by an increasing number of observations.

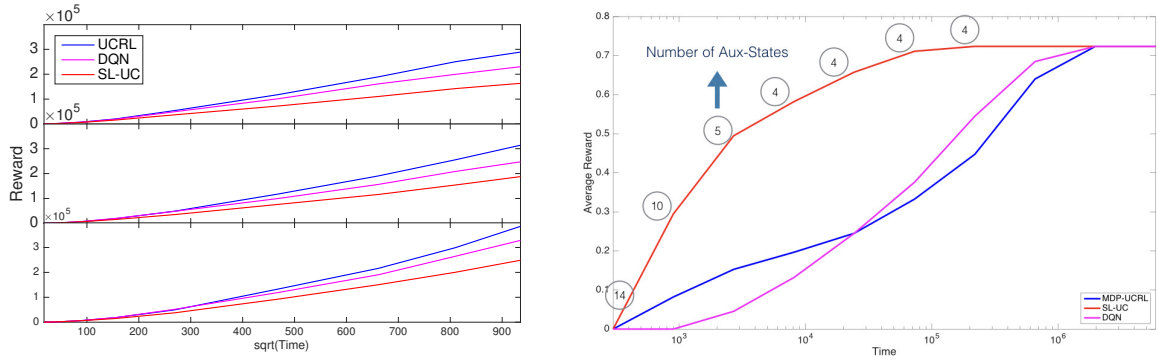


*Figure 5. (left)*The regret comparison, $A = 4$, from top to bottom, $Y = 10, 20, 30$. The scale is $\sqrt{T}$. *(right)* Learning rate of SL-UC compared to UCRL and DQN. After first few rounds, it learns the the true mapping matrix. The numbers in the bulbs are the cardinality of Aux-MDP.