Reinforcement Learning of Contextual MDPs using Spectral Methods

Kamyar Azizzadenesheli
University of California, Irvine

Alessandro Lazaric
INRIA, France

Animashree Anandkumar
University of California, Irvine

Abstract

We propose a new reinforcement learning (RL) algorithm for contextual Markov decision processes (CMDP) using spectral methods. CMDPs are structured MDPs where the dynamics and rewards depend on a smaller number of hidden states or contexts. If the mapping between the hidden and observed states is known a priori, then standard RL algorithms such as UCRL are guaranteed to attain low regret. Is it possible to achieve regret of the same order even when the mapping is unknown? We provide an affirmative answer in this paper. We exploit spectral methods to learn the mapping from hidden to observed states with guaranteed confidence bounds, and incorporate it into the UCRL-based framework to obtain order-optimal regret.

Keywords: Contextual Markov decision process, spectral method, reinforcement learning, confidence bounds.

1 Introduction

Reinforcement learning (RL) is the study of agent-environment interaction, where the agent learns how to interact with the environment so as to maximize a certain objective function [Bertsekas and Tsitsiklis, 1996, Sutton and Barto, 1998]. At the beginning of the interaction, the agent is uncertain about the environment dynamics and requires exploration to obtain a better understanding. Once the agent is fairly certain, knowledge about the environment can be exploited to design a good planning policy, a mapping from context to action, that attains the maximum reward.

Designing algorithms which achieve an optimal tradeoff between exploration and exploitation is the primary goal of reinforcement learning. The tradeoff is commonly measured in terms of *cumulative regret*, which is the difference between the rewards accumulated by the optimal policy (which requires exact knowledge of the environment) and those by the given RL algorithm respectively.

In practice, we are faced with environments with large observation spaces, e.g. robotics. Designing algorithms with a low regret which has a weak dependence on the size of the observed space is extremely challenging. Typically, there is an underlying low dimensional latent space which can effectively summarize the large observation space. Such scenarios arise in a number of practical settings, e.g. robot navigation.

If the mapping between the latent states and the observations is known a priori, then it is trivial to exploit it and design a policy that operates on the latent space of a lower dimension rather than the large observation space. However, this mapping is typically unknown and needs to be learnt by the agent. It is then crucial to have efficient learning algorithms that can quickly provide an estimate of the mapping between hidden and observation states. What is the best possible performance by such an algorithm? In the ideal scenario, there would be no additional regret, and the performance would be *as if* the mapping is known a priori. We show that surprisingly, this indeed is possible (up to lower order terms) for the class of contextual Markov decision processes.

2 Summary of Results

We propose a RL algorithm referred to as Spectral Method CMDP or SM-CMDP. We utilize the spectral method for learning the mapping between observed and hidden states, incorporate it within the framework of upper confidence bound reinforcement learning (UCRL). The algorithm proceeds in epochs, and in each epoch, the estimate is updated, and an optimistic planning policy based on the confidence bounds is employed to collect new samples in the next epoch.

We assume an unknown CMDP with observed space $\mathcal{Y}$, action space $\mathcal{A}$, and hidden state space $\mathcal{X}$, with cardinalities Y, A, and $X \leq Y$ respectively. Let η^*

denote the optimal average reward for a given CMDP.

$$\eta^* := \max_{\pi} \mathbb{E}[\sum_t r_t] = \sum_i \mathbb{P}_{\pi^*}(x)\bar{r}(x, \pi^*(x))$$

where π^* is optimal policy and r_t is reward at time t. Let $\tau_\pi(x \to x')$ denotes mean passing time from state x to state x' given policy π, then the diameter D is defined as

$$D := \max_{x,x'} \min_{\pi} \mathbb{E}[\tau(x \to x')]$$

which is the smallest average time required to go from one hidden state to another hidden state for all pairs of hidden states.

Theorem (Informal Results on Regret Bound)**.** *If* SM-CMDP *in algorithms 1, 3 is run for N time steps, the regret is upper bounded w.h.p. as* [1]

$$Reg_N := N\eta^* - \sum_{t=1}^{N} r_t \leq \widetilde{\mathcal{O}}(DX\sqrt{AN})$$

Remark 1. *If the mapping between the hidden and observed spaces is known a priori, then the UCRL2 algorithm [Jaksch et al., 2010] can be run on the hidden space, and it results in the same regret (up to lower order terms). Thus, asymptotically, there is no loss in performance arising due to the spectral learning algorithm.*

Remark 2. *If the CMDP is treated as a MDP with no additional structure and the UCRL2 algorithm is run [Jaksch et al., 2010] on the observed space, then the regret is bounded by $\widetilde{\mathcal{O}}(D_Y Y\sqrt{AN})$ where $D_Y = \min_{\pi} \max_{y,y'} \mathbb{E}(\tau[y \to y'])$ is the smallest average time required to go from one observed state to another, for any pair of states. It is clear for the case $Y >> X$ we have $D_Y >> D$ and this algorithm suffers from a huge amount of regret since it does not exploit the structure in CMDP.*

3 Related Literature

Researchers have studied the class of latent contextual bandits [Gentile et al., 2014], where the contexts belong to unknown groups. Uniform exploration strategy, combined with an online clustering algorithm is shown to achieve order-optimal regret. An extension is considered in [Gopalan et al., 2016] for recommender systems where the groupings of contexts 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 employed to learn the latent classes. The CMDP considered in this paper is a generalization of the latent contextual bandits, where there is a temporal correlation. Simple uniform exploration no longer suffices for CMDPs, and we need to incorporate spectral learning algorithm in the framework of UCRL2 in order to derive efficient regret bounds.

CMDPs have been earlier studied by [Krishnamurthy et al., 2016], where a PAC analysis is given in the episodic case. They assume that the learner is given a set of functions, and that the true Q-function belongs to this set. The proposed algorithm progressively tries to discard the functions which are not close to the true Q-function, which it tries to learn. The bound derived does not explicitly depend on the cardinality of the context space, it only depends on the cardinality of the function set. However, the underlying Markovian process is required to be deterministic for the analysis to be valid, which is limiting. Moreover, the analysis for the PAC framework is significantly different from the cumulative regret framework in this paper.

CMDP is one approach to deal with MDPs with large state space and there have been other approaches in literature. [Kocsis and Szepesvári, 2006] introduce the MDP Monte-Carlo planning tree to find a near-optimal policy. They consider roll-out based Monte-Carlo planning and show that the probability of choosing optimal action asymptotically tends to 1. Their result, in the case of low chance of visiting states, breaks down to non-selective Monte-Carlo planning and there no theoretical guarantee for the regret analysis. [Kearns et al., 2002] studies episodic learning framework and provides regret analysis with a weak dependency on the cardinality of context space. However, it has an exponential dependence on the length of episodes.

[Hauskrecht et al., 1998] investigates the model with a large number of states and actions, proposes the notion of macro-action to tackle the curse of dimensionality in large MDPs. In [Hallak et al., 2015], the given MDP consists of multiple smaller MDPs and the algorithm is faced with one of these MDPs during each episode. At the beginning of each episode, the environment randomly chooses one small MDP out of the set of models without revealing the model identity and follows the dynamic w.r.t that MDP. In this paper, the authors provide an exploration and exploitation strategy which recognizes the identity of the MDP and analyze the upper bound on regret which grows linearly with a number of episodes.

CMDPs are a special class of partially observable Markov decision processes (POMDP). [Azizzadenesheli et al., 2016] studies

[1] Notation $\widetilde{\mathcal{O}}$ is upper bound notation with logarithmic factors ignored. w.h.p means with high probability.

reinforcement learning of partially observable MDPs (POMDP). They incorporate a spectral learning algorithm within the framework of UCRL2, and provide regret analysis for the class of memoryless planning policies. However, their method does not exploit special structure of CMDPs and suffers from regret which depends on the cardinality of the context space, Remark 3. The relationships between hidden and observed states are more constrained in CMDPs, compared to POMDPs: the transition from observed state to hidden is deterministic in a CMDP. We exploit this here to obtain better regret bounds for the class of CMDPs, and remove the dependence on the cardinality of the observed space in the dominant term of the regret bound. Moreover, finding the optimal memoryless policy for a POMDP is NP-hard [Littman, 1994], and the authors assume an oracle which can compute the optimistic memoryless policy over the class of POMDPs under consideration. In contrast, for CMDPs, the optimistic policy can be computed efficiently. Moreover, [Azizzadenesheli et al., 2016] can only handle stochastic policies which are suboptimal for CMDPs. We remove this requirement in this paper.

CMDPs are a class of latent variable models (LVM) which are popular in a number of domains. Traditional methods such as Expectation-Maximization (EM) [Dempster et al., 1977] have been used previously to learn LVMs, but have no consistency guarantees, are computationally expensive, and may converge to poor local optima. To overcome these drawbacks, spectral methods have been used and guarantee consistent estimation for a wide class of LVMs [Anandkumar et al., 2012, Anandkumar et al., 2014, Song et al., 2013], such as Gaussian mixture models, latent Dirichlet allocation, hidden Markov models, and so on. In this paper, we use spectral methods to learn the hidden structure of the CMDP and utilize the confidence bounds to obtain efficient regret bounds.

4 Preliminaries

A contextual MDP M is a tuple of $M = \langle \mathcal{X}, \mathcal{Y}, \mathcal{A}, R, f_T, f_O \rangle$, where $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{A}$ are set of hidden states, context states, and set of actions, moreover, X, Y, and A are the cardinality of these sets with elements in $[X]$, $[Y]$, and $[A]$ respectively. We assume the reward at any time step t is bounded $0 \leq r_t \leq 1$ and generated by hidden states and actions where matrix $R \in \mathbb{R}^{A \times X}$ is mean reward matrix, $R_{i,l} = \bar{r}(i,l) = \mathbb{E}_r[r(x=i, a=l)]$, $\forall i \in [X], l \in [A]$. Moreover, the transition process is characterized by $T_{i',i,l} := f_T(i'|i,l) = \mathbb{P}(x' = i'|x = i, a = l)$, where transition tensor $T \in \mathbb{R}^{X \times X \times A}$. At each time step, the environment of large MDP is at state y and hidden MDP at its corresponding hidden state. The agent chooses an action a which takes the environment to new state y' and the corresponding hidden state x'. When the environment is at hidden state $x = i$, it emits signal $y = \vec{e}_j$ as a context with distribution $O_{j,i} = f_O(j|i) = \mathbb{P}(y = \vec{e}_j | x = i)$ where matrix $O \in \mathbb{R}^{Y \times X}$ is the emission matrix with minimum nonzero entry $O_{\min}$ Fig. 1 ($\vec{e}_j$ is a basis vector where all elements are equal to zero except j'th element is equal to one). In CMDPs the emission matrix has block structure and has exactly one nonzero element at each row. The matrix O can be represented

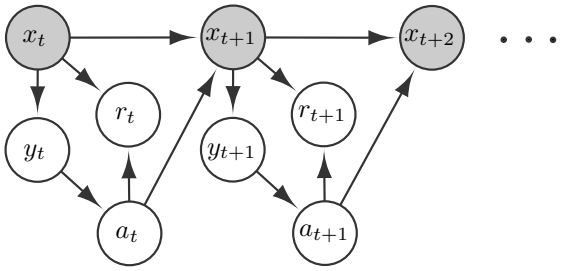

Figure 1: Graphical model of CMDP.

by a vector $\bar{o}$ where $\bar{o}_j$ is non-zero entry in j'th row of emission matrix. The main property of CMDP which distinguishes it from general POMDP is that there is a stochastic mapping from hidden state to context but surjective mapping $y \to x$. In the case of surjective mapping $y \to x$, the problem of POMDP reduces to low dimension MDP which is the underlying MDP on hidden states of CMDP.

5 Stochastic CMDP

In previous sections we stated the model structure of CMDP and provide the detailed notation for model dynamics. In this section, we turn our attention to the model where the small hidden MDP is ergodic and the transition process on these hidden states is stochastic. We propose an algorithm for efficient learning and planning under the stochastic setting. We provide order optimal regret analysis and show that, asymptotically, it does not dependent on Y. We introduce the novel spectral method to efficiently overcome the learning difficulty of the latent structure, and demonstrate how to learn the parameters of CMDP, f_T, surjective mapping $y \to x$, and R matrix. The block structure of emission matrix results in the surjective mapping from context to hidden state which is sufficient statistics for optimal policy and further effort on learning entries of the matrix O is redundant.

Assumption 1 (Stochastic CMDP)**.** *For any policy*

π, the Markov chain on hidden states of CMDPs is ergodic, and for any action a, slice of transition matrix on hidden states $T_{:,:,a}$ is invertible.

In the following, we deploy the multi-view analysis in [Anandkumar et al., 2014], establish the learning algorithm, and show how to combine the learning and planning process to design an efficient RL algorithm.

5.1 Spectral Learning Method

We learn the structure of emission matrix by utilizing multi-view model. Given N samples of under policy π, let $t \in [2, \dots, N]$ denote time steps at which $a_t = l$. We construct new variables $\vec{v}_1, \vec{v}_2$ and $\vec{v}_3$ with the realizations of y_{t-1}, y_t, and y_{t+1} respectively. Moreover, define the corresponding middle action and middle state $\bar{a}_2$ and $\bar{x}_2$ with the realizations of a_t and x_t. It is clear from Fig. 1 that these three views v_1, v_2 and v_3 are conditionally independent give $\bar{x}_2$ and $\bar{a}_2$ which reduces it to multi-view model, whose learning has been vastly studied in tensor methods literature. Define the factor matrices $V_1^{(l)}, V_2^{(l)}, V_3^{(l)} \in \mathbb{R}^{Y \times k_\pi^{(l)}}$ as follows

$$[V_1^{(l)}]_{j,i} = \mathbb{P}(\vec{v}_1 = \vec{e}_j | \bar{x}_2 = i, \bar{a}_2 = l)$$
$$[V_2^{(l)}]_{j,i} = \mathbb{P}(\vec{v}_2 = \vec{e}_j | \bar{x}_2 = i, \bar{a}_2 = l)$$
$$[V_3^{(l)}]_{j,i} = \mathbb{P}(\vec{v}_3 = \vec{e}_j | \bar{x}_2 = i, \bar{a}_2 = l)$$

We are interested to estimate factor matrices of middle view. Generally, the challenge in spectral method is to find the right moments to estimate the desired factors. In multi-view model, given set of samples $\vec{v}_1, \vec{v}_2, \vec{v}_3$, we form second order moments $K_{p,q}^{(l)}$ for $\forall p, q \in \{1, 2, 3\}$, as covariance matrices of pairs (p, q).

$$K_{p,q}^{(l)} = \mathbb{E}[v_p \otimes v_q]$$

with empirical averages $\widehat{K}_{p,q}^{(l)} = \frac{1}{N(l)} \sum_t y_{t+p-2} \otimes y_{t+q-2}$ where $N(l)$ is the number of triples with middle action equals to l. Moreover, for third order moment $K_{1,3,2}^{(l)} = \mathbb{E}[y_{t-1} \otimes y_{t+1} \otimes y_t]$ with empirical average $\widehat{K}_{1,3,2}^{(l)}$.

Let $\mathcal{I}(l)$ denote a set of hidden state which are mapped to action l under the policy pi. Under Asm. 1, we estimate the factor matrices $\widehat{V}_2^{(l)}$ for all actions and their corresponding set of hidden states $\mathcal{I}(l)$. Due to the estimation error in moment estimation, spectral learning estimates the factor matrices up to the confidence intervals which shrink to zero as N increases. Define $\mathcal{B}_O^{(l)}$ as upper bound on the estimation error, given samples with middle action equals to l

$$\|[V_2^{(l)}]_{\cdot,i} - [\widehat{V}_2^{(l)}]_{\cdot,i}\|_2 \leq \mathcal{B}_O^{(l)}$$

Given Asm. 1, $rank(V_2^{(l)}) \leq rank(V_3^{(l)})$, therefore the learning is consistent with factor matrix of second view. As it is mentioned before, CMDP can be seen as a large MDP on context space, therefore the optimal policy is a deterministic mapping from context to action. Furthermore, the columns of factor matrices $V_2^{(l)}$ for different actions contain many zeros due to the deterministic policy. As a consequence, we are not able to recover the elements of the O matrix, but we are easily able to recover its nonzero elements. As long as the underlying Markov chain is recurrent, the agent is able to go over all the hidden states and identify the non-zero elements of the O matrix. By increasing the number samples, the spectral methods provide tighter precision in parameter estimation. It means that for some elements of factor matrices, e.g. for context j, and one pair of hidden state i and action l, we have $[\widehat{V}^{(l)}]_{j,i} > 2\mathcal{B}_O^{(l)}$. It means that the context j revealed the hidden state to which it belongs.

At time N there are many contexts which have revealed their identities and some others whose identities are still vague to the learner and require tighter confidence bound, consequently more samples. The learner, for each action l and column i, identify the entries in matrix $[\widehat{V}_2^{(l)}]$ that have values above the threshold $2\mathcal{B}_O^{(l)}$. It means that those entries belong to the same column and as a result belong to same hidden state. The learner construct an auxiliary set $\mathcal{S}$ and assign each cluster an auxiliary state. The algorithm goes over all actions and hidden states, clusters the contexts whose hidden state is same. Up to permutation of the spectral method, the algorithm can not combine the clustering results of different actions and create larger clusters. But, if there are columns of different factor matrices which agree in at least one context, then these columns correspond to same hidden state and the algorithm combines the clustering results of these columns by creating one cluster out of union of those clusters.

The contexts, which are not assigned to any cluster, are the context with unknown identity up to current estimate. The learner assigned each of them an auxiliary state as well. It means that the elements in set $\mathcal{S}$ are either a set of contexts whose identities are revealed and agree on same hidden state or single context which has not been clustered yet.

The CMDP is a large MDP at context level, therefore, if the learner constructs the set of auxiliary states $\mathcal{S}$, the process on the level of the auxiliary states is still MDP but with fewer states. Therefore, given the set of auxiliary states $\mathcal{S}$, the agent constructs new MDP on set. At the beginning of learning process, the set $\mathcal{S} = \mathcal{Y}$, after collecting samples $S \leq Y$. For sufficient

amount of sample when $\mathcal{B}_O^{(l)}$ is small enough to identify all contexts, i.e. $[\widehat{V}^{(l)}]_{j,i} > 2\mathcal{B}_O^{(l)}$ is satisfied for all the non-zero elements, $S = X$ and $\mathcal{S} = \mathcal{X}$.

5.2 Spectral UCRL

In the previous section, we showed how to deploy spectral methods to quickly transfer large MDP on context space to smaller MDP on auxiliary set $\mathcal{S}$. In this section, we utilize the results from spectral methods, construct MDP on space $\mathcal{S}$ and provide planning strategy on the top of this auxiliary MDP. We mimic the idea of exploration and exploitation from UCRL2 and apply it on small auxiliary MDP on $\mathcal{S}$.

The algorithm illustrated in Alg. 1 is the result of the integration of the spectral method into UCRL which is designed to optimize the exploration-exploitation trade-off. The learning process is split into epochs of increasing length. At the starting point of each epoch k, the agent uses samples from past epochs to construct the set $\mathcal{S}^{(k)}$. Furthermore, the agent needs to estimate the model parameters of MDP with state space $\mathcal{S}$, i.e. transition tensor and reward process. By modified Chernoff-Hoeffding inequality for the set $\mathcal{S}^k$: $\forall s \in \mathcal{S}^{(k)}, \forall a \in \mathcal{A}$

$$d(s,a) := \|p(\cdot|s,a) - \widehat{p}(\cdot|s,a)\|_1 \leq \sqrt{\frac{7S^{(k)}\log(\frac{2AN^{(k)}}{\delta})}{\max\{1, N^{(k)}(s,a)\}}}$$
$$d'(s,a) := |\bar{r}(s,a) - \widehat{r}(s,a)| \leq \sqrt{\frac{7\log(\frac{2S^{(k)}AN^{(k)}}{\delta})}{2\max\{1, N^{(k)}(s,a)\}}} \quad (1)$$

with probability at least $1-\delta$. At the beginning of epoch k, $N^{(k)}(s,a)$ is the number of time that the pair of auxiliary state s and action a has happened, and $N^{(k)}(a)$ is the number time that action a has been chosen. Furthermore, define variables $\nu^{(k)}(s,a)$ and $\nu^{(k)}(a)$ as a corresponding counts during epoch k, i.e. $\nu^{(k)}(s,a) := N^{(k+1)}(s,a) - N^{(k)}(s,a)$ and $\nu^{(k)}(a) := N^{(k+1)}(a) - N^{(k)}(a)$, $\forall a \in \mathcal{A},\ s \in \mathcal{S}^{(\|)}$.

Among the ranges of possible parameters the SM-CMDP Alg. 1 constructs a set of admissible MDPs $\mathcal{M}^{(k)}$ on state space $\mathcal{S}$ whose parameters belong to the confidence interval 1. Thereafter the agent applies the popular principle of Optimism-in-Face-of-Uncertainty (OFU) to come up with optimal policy with respect to the most optimistic model in the class of plausible MDPs $\mathcal{M}^{(k)}$. Then, the agent applies the optimistic policy on next epoch and stop the epoch when the number samples at least for one pair of auxiliary state and action is doubled.

$$\widetilde{\pi}^{(k)} = \arg\max_{\pi} \max_{M' \in \mathcal{M}^{(k)}} \eta(\pi, M') \quad (2)$$

Algorithm 1 Stochastic *SM-CMDP* algorithm

Input: Confidence δ'
Variables:
Number of samples $N^{(k)}(l)$
Initialize: $t = 1$, initial state x_1, $k = 1$, δ/N^6
while $t < N$ **do**
Compute the estimated $V_2^{(l)}$ matrices $\forall l \in \mathcal{A}$ and construct set $\mathcal{S}^k$
Compute the estimate of reward and transition tensor in set $\mathcal{S}^k$ Eq. 1
Construct plausible set of MDP $\mathcal{M}^{(k)}$ out of set S^k
Compute the optimistic policy
$\widetilde{\pi}^{(k)} = \arg\max_{\pi} \max_{M \in \mathcal{M}^{(k)}} \eta(\pi; M)$
Set $v^{(k)}(s,l) = 0$ for all actions $l \in \mathcal{A}, s \in \mathcal{S}^{(k)}$
while $\forall l \in \mathcal{A}, \forall s \in \mathcal{S}^{(k)}$,
$v^{(k)}(s,l) < \max\{1, N^{(k)}(s,l)\}$ **do**
Execute $a_t \sim \widetilde{\pi}$
Obtain reward $\vec{r}_t$, observe next context $\vec{y}_{t+1}$, and set $t = t+1$
end while
Set $k = k+1$
end while

Let $\widetilde{M}^{(k)} := \langle \mathcal{S}^{(k)}, \mathcal{A}, \widetilde{R}, \widetilde{p}(\cdot|\cdot,\cdot)\rangle$ denote the optimistic MDP and $\widetilde{\eta}^{(k)}$ its average reward. This optimistic choice provides a smooth combination of the exploration, encouraged by the confidence intervals, and the exploitation of the current estimation of CMDP.

Algorithm 2 *Finding optimistic policy*

Input: Set $\mathcal{S}$, $\mathcal{A}$ and estimated $\widehat{p}(\cdot|s,a)$, $\widehat{R}(s,a)$
Reorder set $\mathcal{S} = \{s'_1, s'_2, \ldots, s'_S\}$ 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 $\ell := S^{(k)}$
While $\sum_j p(s'_j) > 1$ **do**
Reset $p(s'_\ell) := \max\{0, 1 - \sum_{j \neq \ell} p(s'_j)\}$
Set $\ell = \ell - 1$

To optimize over models and policies in Eq. 2 , 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. 2, it reduces the optimization problem to: $\forall s \in \mathcal{S}^{(k)}$

$$u_0(s) = 0,$$
$$u_{t+1}(s) = \max_a \{\widetilde{r}(s,a) + \max_{p(\cdot) \in \mathcal{P}^k(s,a)} \{\sum_{s'} p(s') \cdot u_i(s')\}\} \quad (3)$$

where the set of vectors $\mathcal{P}^k(s,a)$ is set of admissible vectors of $p(\cdot|s,a)$ in the confidence interval. The it-

erative algorithm in Eq. 3 is another view of Poisson Equation to find optimal policy. It can be shown with a suitable transformation that the value u_i is biased view of expected advantage value in Poisson Equation. As it is shown in Eq. 3, this iterative procedure solves a simple optimization problem at each iteration, and updates the value vectors. The algorithm in Alg. 2 provides an efficient way to handle this optimization with computation cost of $\mathcal{O}((S^{(k)})^2 A)$. The detailed analysis of finding optimistic policy is studied in [Puterman, 2014], [Jaksch et al., 2010].

5.3 Guarantee on Regret

In algorithm Alg. 2 we show how efficiently compute the optimistic policy. Generally, at the beginning of each epoch, the agent constructs the auxiliary space states and computes the optimal policy w.r.t optimistic model in $\mathcal{M}^{(k)}$. Then, the agent applies this policy to next epoch and stop the epoch when the number samples at least for one pair of auxiliary state and action is doubled.

Lemma 1. *Under Asm. 1, let $\widehat{V}_2^{(l)}$ denote the factor matrix of second view in multi-view representation of CMDP. Then for policy π which maps set of hidden states $\mathcal{I}(l)$ to action l, for $i \in \mathcal{I}(l)$*

$$\|[V_2^{(l)}]_{\cdot,i} - [\widehat{V}_2^{(l)}]_{\cdot,i}\|_2 \leq C_2 \sqrt{\frac{\log(Y/\delta)}{N(l)}} := \mathcal{B}_O^{(l)} \quad (4)$$

with probability at least $1-\delta$ where the constant C_2 is a model dependent constant.

Let's also define D_Y as diameter of corresponding large MDP over context space of CMDP,

$$D_Y := \max_{y,y'} \min_{\pi} \mathbb{E}[\tau(y \to y')]$$

which is the smallest average time required to go from one state to another state for all pairs of states in large MDP. Define $\tau_M = \max_\pi \max_x \mathbb{E}_\pi[\tau(x \to x)]$.[2]

Theorem 1 (Regret Bound of Stochastic Environment)**.** *Consider CMDP $M = \langle \mathcal{X}, \mathcal{Y}, \mathcal{A}, R, f_T, f_O \rangle$ characterized by a diameter D. If* SM-CMDP *is run over N time steps, under Asm. 1, it suffers the total regret of*

$$\begin{aligned} Reg_N \leq & 34DX\sqrt{A(N-\tau)\log(N/\delta)} \\ & + \min\{\tau, 34D_Y Y \sqrt{A\tau \log(N/\delta)}\} \end{aligned}$$

with probability at least $1-\delta$. The constant time τ is the time that the algorithm identifies the whole mapping from contexts to hidden states.

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

where C_1 is a numerical constant which represents the similarity of contexts in policy.

Remark 3. *[Regret on memory-less POMDP] Let M be a POMDP with X states, Y observations, A actions, and R rewards, with a diameter D_{pomdp} defined as*

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

where $\tau((x,a) \to (x',a'))$ denotes mean passing time from pair of (x,a) to (x',a') and $\mathcal{P}$ denotes set of memory-less policies with nonzero level of stochasticity. At time step N, the order optimal regret analysis of POMDP suffers regret w.r.t optimal policy in set $\mathcal{P}$

$$Reg_N \leq \widetilde{\mathcal{O}}(D_{pomdp} X^{3/2} \sqrt{AYN})$$

The regret of POMDP-based algorithm in comparison with SM-CMDP has a worse dependency on the model parameters. The regret suffers from additional term $\sqrt{Y}$ because the RL algorithm in POMDP put much effort on accurate estimation of entries in the 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. It worth to note that the notion of diameter in POMDP is in order of $A^2 D$.

6 Deterministic CMDPs

In the previous section, we proposed an RL algorithm for stochastic CMDP and elaborated the regret analysis of SM-CMDP on stochastic domain. Given ergodicity condition of the environment, the learner explores the behavior of all hidden states during epochs. Moreover, we showed that the learning is feasible through estimating factor matrix of middle view $V_2^{(l)}$.

In this section, we investigate deterministic environment and propose new method for efficient exploration and exploitation task. We provide a new spectral learning method and combine it with our planning strategy. In deterministic models, the deterministic policy followed by periodic Markov chain which might traverse just through a subset of hidden states. In general, there is no guarantee that the agent travels through all hidden states for sufficient amount of time. As stated before, for an efficient understanding of the hidden structure, the agent needs to explore all hidden states frequently. Therefore, we need to encourage our agent to explore the hidden states to which the optimistic policy does not plan to go.

Assumption 2 (Deterministic CMDP)**.** *The underlying MDP over hidden states is connected under some sequence of actions* [3] *and transition process is deterministic.*

[2] subscript π means expectation under policy π

[3] Bounded diameter D

6.1 Spectral Learning Method

Generally, the dynamics in deterministic environment is totally different from dynamics in stochastic environments, as a consequence the new learning procedure is required. Same as the previous section, at time N let $t \in [2, \dots, N]$ denote time steps at which $a_t = l$ and construct views $\vec{v}_1, \vec{v}_2$ and $\vec{v}_3$. Then define the corresponding middle action and middle state $\bar{a}_2$ and $\bar{x}_2$ and the factor matrices.

In the multi-view representation of deterministic dynamic, $rank(V_3^{(l)})$ is equal to $rank(K_{2,3}^{(l)})$ but it is less than or equal to $rank(V_1^{(l)})$ and $rank(V_2^{(l)})$. Consequently, the learning is feasible through matrix $V_3^{(l)}$ which represents the distribution of third view given middle hidden state and action. To learn the third view factor matrix we need to construct new third order moment $K_{1,2,3}^{(l)} = \mathbb{E}[y_{t-1} \otimes y_t \otimes y_{t+1}]$ with empirical average $\widehat{K}_{1,2,3}^{(l)}$. Let $i'(i,l)$ denote the next state of environment when the agent takes action l at hidden state i,

$$[V_3^{(l)}]_{j,i} = \sum_{i'} f_O(j|i') f_T(i'|i,l) = f_O(j|i'(i,l)) \quad (5)$$

Let set $\mathcal{I}'(l)$ denote the set of hidden state of the environment which are followed by taking action l at set $\mathcal{I}(l)$. Under Asm. 2, spectral methods estimate the columns of O matrix given samples from all the actions. Due to the estimation error in moment estimation, spectral learning estimates the these columns up to the confidence intervals which shrink to zero as N increases. Define $\mathcal{B}_{O_3}^{(l)}$ as upper bound on the estimation error given samples of action l

$$\|[O]_{\cdot,i} - [\widetilde{O}]_{\cdot,i}\|_2 \leq \mathcal{B}_{O_3}^{(l)}$$

At time step N, we mimic the clustering procedure from previous section and construct the auxiliary set $\mathcal{S}(N)$. We cluster the contexts which release their identity, $[V_3^{(l)}]_{j,i} > 2\mathcal{B}_{O_3}^{(l)}$, and agree in same hidden state, then treat them as one auxiliary state. We unify the set of clusters and remaining non-clustered context then construct auxiliary set $\mathcal{S}(N)$.

As it is mentioned before, for fixed deterministic policy the environment might result in underlying periodic Markov chain which might not traverse over all hidden states. Generally, learning under the samples of deterministic policy results in partial recovery of emission matrix which is not fulfillment situation for a consistent learner. We need to design a policy which explores all the hidden states and makes sufficiently tight confidence bound Eq. 6 to reveal the contexts identity.

To efficiently recover the nonzero elements in emission matrix, we apply uniform stochastic policy π_{exp} which chooses actions equally likely on the environment. Due to the connectivity condition of underlying hidden MDP, stochastic policy π_{exp} results in ergodic underlying Markov chain and therefore provides sufficient exploration of hidden states. At time N, let's define set $\mathcal{S}(N)$ the result of clustering of N samples. It is clear that for sufficiently large N the set $\mathcal{S}(N)$ tends to set $\mathcal{X}$.

6.2 Spectral UCRL

In the previous section, we showed how to deploy spectral methods under the stochastic policy and learn the non-zero entries in emission matrix. In this section, we propose a RL algorithm for the deterministic domain, which consists of two phases, (i) learning the true mapping from contexts to hidden states, (ii) exploration and exploration on hidden MDP of size X. To learn the hidden structure in phase i, the agent sets the policy to uniform policy π_{exp} and applies it to the environment. As long as the policy π_{exp} is not an optimal policy, the agent suffers from linear regret during phase (i).

Algorithm 3 Deterministic *SM-CMDP* algorithm

Input: Confidence δ'
Initialize: $t = 1$, $k = 1$, $k' = 1$, initial state x_1, δ/N^6, π_{exp} and $\mathcal{S} = \mathcal{Y}$,
while $t < N$ **do**
  Phase (i) :
  while $\mathcal{S} \neq \mathcal{X}$ **do**
    Set $\nu^{(k')}(l) = 0$ for all action $l \in \mathcal{A}$
    while $\forall l \in \mathcal{A}, \nu^{(k')}(l) < N^{(k')}(l)$ **do**
      Execute $a_t \sim \pi_{exp}$
      Obtain reward r_t and observe context y_t,
      Set $t = t + 1$
    end while
    Estimate $V_3^{(l)}$ matrices $\forall l \in \mathcal{A}$ and construct set $\mathcal{S}$
    Set $k' = k' + 1$ and update $N^{(k')}(l),\ \forall l \in \mathcal{A}$
  end while
  Phase (ii) :
  Compute the estimate of reward and transition tensor in set $\mathcal{X}$ Eq. 1
  Construct plausible set of MDP $\mathcal{M}^{(k)}$ out of set $\mathcal{X}$
  Compute the optimistic policy
    $$\widetilde{\pi}^{(k)} = \arg\max_{\pi} \max_{M \in \mathcal{M}^{(k)}} \eta(\pi; M)$$
  Set $v^{(k)}(x,l) = 0$ for all actions $l \in \mathcal{A}, x \in \mathcal{X}^{(k)}$
  while $\forall l \in \mathcal{A}, \forall x \in \mathcal{X}$,
    $v^{(k)}(x,l) < \max\{1, N^{(k)}(x,l)\}$ **do**
    Execute $a_t \sim \widetilde{\pi}$
    Obtain reward $\vec{r}_t$, observe next context $\vec{y}_{t+1}$, and set $t = t + 1$
  end while
  Set $k = k + 1$
end while

The agent stops this phase after collecting sufficient amount of samples for all the contexts to release their identities, i.e. $[V_3^{(l)}]_{j,i} > 2\mathcal{B}_{O_3^{(l)}}$ for all i, j and different actions. In other words, the agent applies the pure explorative policy π_{exp} until the set $\mathcal{S}(N)$ converges to set $\mathcal{X}$. In the learning process, there is no need to estimate $V_3^{(l)}$ at each time step. The learner goes through spectral method to estimate the factor matrix $V_3^{(l)}$ when the number of collected samples at least for one action is doubled.

At this time, the agent has learned the surjective assignment from contexts to hidden states and configured the required information to start the second phase. For phase ii, the agent does not deal with large MDP any more because of the full knowledge about $y \to x$ mapping. It means that the agent faces small MDP on hidden states and applies tradition UCRL algorithm.

Given samples from phase (i), the agent constructs an admissible set of MDPs $\mathcal{M}^{(k)}$ (where $k = 1$ for first epoch of phase ii) on space $\mathcal{X}$ Eq. 1, finds the optimistic model and computes its corresponding optimal policy through Eq. 3. Then the agent applies the optimistic policy on the environment until the number of samples at least for one pair of (x, a) is doubled. At this time, the agent shrinks the confidence intervals even more and constructs the tighter admissible set of MDPs to fine the optimistic policy which is closer to optimal policy Alg. 2, and applies it on next epoch Alg. 3.

6.3 Guarantee on Regret

The spectral algorithm in Alg. 3, suffers constant regret of pure exploration in phase i, but after learning the sufficient statistic information of hidden structure, it suffers from order optimal regret. Let's define $\tau'_{exp} := \max_x \mathbb{E}_{\pi_{exp}}[\tau(x \to x)]$.

Lemma 2. *Under Asm. 2, let $\widehat{V}_3^{(l)}$ denote the factor matrix of third view in multi-view representation of CMDP. Then, through spectral method, after $N(l)$ samples, for $i \in \mathcal{I}(l)$ we have*

$$\|[O]_{\cdot,i} - [\widetilde{O}]_{\cdot,i}\|_2 \leq C_3 \sqrt{\frac{\log(Y/\delta)}{N(l)}} := \mathcal{B}_{O_3^{(l)}} \quad (6)$$

with probability at least $1 - \delta$ where the constant C is a model dependent constant.

Remark 4. *$\widehat{V}_3^{(l)}$ is a matrix of right singular vectors of covariance matrix $K_{2,3}^{(l)}$. In the case of sufficiently large singular gap $\min_{i,i' \in \mathcal{I}(l)} |\mathbb{P}(\bar{x}_2 = i|\bar{a}_2 = l) - \mathbb{P}(\bar{x}_2 = i'|\bar{a}_2 = l)|$, singular value decomposition is preferred over third order tensor decomposition.*

Theorem 2 (Regret Bound of Deterministic Environment)**.** *Consider CMDP $M = \langle \mathcal{X}, \mathcal{Y}, \mathcal{A}, R, f_T, f_O \rangle$ characterized by a diameter D. If* SM-CMDP *is run over N time steps, under Asm. 2, it suffers the total regret of*

$$Reg_N \leq 34DX\sqrt{A(N - \tau')\log(N/\delta)} + \tau'$$

with probability at least $1 - \delta$. The constant time τ' is the time that the algorithm finishes the clustering task.

$$\tau' := 16A^2Y\tau'_{exp}(4C_3\frac{\log(Y/\delta)}{O_{\min}^2} + \log(XA/\delta))$$

7 Conclusion

We introduced a novel RL algorithm for CMDPs under different hidden process dynamics. We utilized the spectral method for learning the latent mapping from context to hidden states, argued that the knowledge of this mapping is sufficient for efficient planning, and finally incorporated this results within the framework of upper confidence bound reinforcement learning (UCRL). We demonstrated how efficiently reduce the problem of large MDP to the corresponding small latent MDP and provided regret analysis among this reduction process.

We showed that the regret of SM-CMDP has same rate as regret of the optimal RL algorithm on hidden MDP and stated that our agent does not suffer from an exhaustive exploration of large MDP.

In RL problems, the principle of Optimism-in-Face-of-Uncertainty contributes in designing a policy that locally improves the model uncertainty and average reward which has been shown to be an optimal strategy. It is an open question to analyze and modify this principle for the models with clustering where global improvement of the information in model uncertainty is required. While the SM-CMDP for deterministic models reaches order optimal regret, it is not still clear how to modify the exploration to enhance the constant regret of the pure exploration phase.

Acknowledgements

K. Azizzadenesheli is supported in part by NSF Career Award CCF- 1254106. A. Lazaric is supported in part by the French Ministry of Higher Education and Research 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, Google faculty award, Adobe grant, NSF Career Award CCF- 1254106, ONR Award N00014-14-1-0665, and AFOSRYIP FA9550-15-1-0221.

References

[Anandkumar et al., 2014] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.

[Anandkumar et al., 2012] Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*.

[Azizzadenesheli et al., 2016] Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. (2016). Reinforcement learning of pomdps using spectral methods. *arXiv preprint arXiv:1602.07764*.

[Bertsekas and Tsitsiklis, 1996] Bertsekas, D. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[Gentile et al., 2014] Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *ICML*, pages 757–765.

[Gopalan et al., 2016] Gopalan, A., Maillard, O.-A., and Zaki, M. (2016). Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*.

[Hallak et al., 2015] Hallak, A., Di Castro, D., and Mannor, S. (2015). Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*.

[Hauskrecht et al., 1998] Hauskrecht, M., Meuleau, N., Kaelbling, L. P., Dean, T., and Boutilier, C. (1998). Hierarchical solution of markov decision processes using macro-actions. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 220–229. Morgan Kaufmann Publishers Inc.

[Jaksch et al., 2010] Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

[Kearns et al., 2002] Kearns, M., Mansour, Y., and Ng, A. Y. (2002). A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, 49(2-3):193–208.

[Kocsis and Szepesvári, 2006] Kocsis, L. and Szepesvári, C. (2006). Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*, pages 282–293. Springer.

[Krishnamurthy et al., 2016] Krishnamurthy, A., Agarwal, A., and Langford, J. (2016). Contextual-mdps for pac-reinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722*.

[Littman, 1994] Littman, M. L. (1994). 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. MIT Press.

[Puterman, 2014] Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[Song et al., 2013] Song, L., Anandkumar, A., Dai, B., and Xie, B. (2013). Nonparametric estimation of multi-view latent variable models. *arXiv preprint arXiv:1311.3287*.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT Press.

[Tropp et al., 2011] Tropp, J. A. et al. (2011). Freedman's inequality for matrix martingales. *Electron. Commun. Probab*, 16:262–270.

Supplementary

Lemma 3 (Moments Confidence Bound)**.** *Estimation errors of second and third order moments given action l decay by $\mathcal{O}(\sqrt{\frac{1}{N(l)}})$ and the rank of these moments are identifiable.*

Proof. Given the policy π and time N let $t \in [2, \dots, N]$ denotes time steps at which $a_t = l$ with minor loss in notation $[t] = [2, \dots, N]$ and $|t|$ denotes the number of element in $[t]$ with values less than or equal to t. We construct new variables $\vec{v}_1, \vec{v}_2$ and $\vec{v}_3$ with the realizations of y_{t-1}, y_t, and y_{t+1} respectively. Moreover, define the corresponding middle action and middle state $\bar{a}_2$ and $\bar{x}_2$ with the realizations of a_t and x_t. It is clear from Fig. 1 that these three views v_1, v_2 and v_3 are conditionally independent give $\bar{x}_2$ and $\bar{a}_2$.

Given the results in Lemma. 6, and $t' \in [t]$; $t' \leq t$ define the following variable

$$\Phi_t := \sum_{t'}^{t} \left(y_{t'} \otimes y_{t'+1} - \mathbb{E}[y_{t'} \otimes y_{t'+1}] \right)$$

and its martingale difference

$$\Psi_t := \Phi_t - \Phi_{t-1} = y_t \otimes y_{t+1} - \mathbb{E}[y_t \otimes y_{t+1}]$$

as a consequence, for this case, $C_\Psi \leq 2$ and $W_j = \sum_i^j \mathbb{E}\left[\left[y_i \otimes y_{i+1} - \mathbb{E}(y_i \otimes y_{i+1}) \right]^2 | \mathcal{F}_{i-1} \right]$ where $\mathcal{F}_{i-1}$ is corresponding filter with $i, j \in [t]$.

$$W_j = \sum_i^j \mathbb{E}_{i-1}\left((y_i \otimes y_{i+1})^2 \right) - \left(\mathbb{E}_{i-1}\left(y_i \otimes y_{i+1} \right) \right)^2 = \sum_i^j \mathbb{E}_{i-1}\left(y_{i+1} \otimes y_{i+1} \right) - \left(\mathbb{E}_{i-1}\left(y_i \otimes y_{i+1} \right) \right)^2$$

then

$$\|W_j\|_2 \leq \sum_i^j \| \mathbb{E}_{i-1}\left(y_{i+1} \otimes y_{i+1} \right) \| + \| \left(\mathbb{E}_{i-1}\left(y_i \otimes y_{i+1} \right) \right) \|_2^2 \leq |j| + ||j = 2|j|$$

As a result

$$\widehat{\epsilon}_{2,3}^t := \frac{\Phi_t}{|t|} \leq \frac{4}{3|t|} + \sqrt{\frac{4 \log(2d/\delta)}{|t|}}$$

with same procedure, the variable $\sum_{t'}^t \mathbb{E}[y_{t'} \otimes y_{t'+1}] - \lim_{\tau' \to \infty} \frac{\sum_{\tau''}^{\tau'} \mathbb{E}[y_{\tau''} \otimes y_{\tau''+1}]}{\tau'}$ with $\tau' \in [t]$ is bounded by $\frac{4}{3} + \sqrt{4|t| \log(2d/\delta)}$ therefore

$$\epsilon_{2,3}^t := \frac{1}{|t|} \sum_{t'}^{t} \left(y_{t'} \otimes y_{t'+1} - \lim_{\tau' \to \infty} \frac{\sum_{\tau''}^{\tau'} \mathbb{E}[y_{\tau''} \otimes y_{\tau''+1}]}{\tau'} \right) \leq \frac{8}{3|t|} + \sqrt{\frac{16 \log(2d/\delta)}{|t|}} \tag{7}$$

It is clear that same inequality holds for other covariance matrices $p, q \in \{1, 2, 3\}$

$$\epsilon_{p,q}^t := \frac{1}{|t|} \sum_{t'}^{t} \left(y_{t'+p-2} \otimes y_{t'+q-2} - \lim_{\tau' \to \infty} \frac{\sum_{\tau''}^{\tau'} \mathbb{E}[y_{\tau''+p-2} \otimes y_{\tau''+q-2}]}{\tau'} \right) \tag{8}$$

Same argument works for third order moment

$$\epsilon_{p,q,z}^t := \frac{1}{|t|} \sum_{t'}^{t} \left(y_{t'+p-2} \otimes y_{t'+q-2} \otimes y_{t'+z-2} - \lim_{\tau' \to \infty} \frac{\sum_{\tau''}^{\tau'} \mathbb{E}[y_{\tau''+p-2} \otimes y_{\tau''+q-2} \otimes y_{\tau''+z-2}]}{\tau'} \right) \leq \frac{8}{3|t|} + \sqrt{\frac{16 \log(2\sqrt{d^3}/\delta)}{|t|}} \tag{9}$$

To recover $V_2^{(l)}$, we are interested in rank of $K_{2,3}^l$. If at time t the following condition holds

$$\epsilon_{2,3}^{N(l)}(l) \leq g^\epsilon(N(l)) \leq 0.5\sigma_{k_\pi^{(l)}(1,3)} \tag{10}$$

where $\sigma_{k_\pi^{(l)}(2,3)}$ is smallest non-zero singular value of matrix $M_2(2,3)$. Compute a set of singular values of $K_{2,3}^{(l)}$ and ignore the singular values with value less than $g^\epsilon(N(l)) = \frac{g}{N(l)^{0.5-\epsilon}}$ then the number of remaining singular values gives true rank of $K_{2,3}^{(l)}$ with probability at least $1-\delta$. Actually, with perturbation the additional singular values can have values at most ϵ_{pairs}. Actually, for any positive value g and any $0 < \epsilon < 0.5$, condition Eq. 10 holds after few number of samples, due to the fact that this function approaches zero by power of $-(0.5-\epsilon)$, moreover $0.5\sigma_{k_\pi^{(l)}}$ is fixed number and $\epsilon_{2,3}$ approaches zero in faster rate.

□

Lemma 4. *Let's recall rank-$k_\pi^{(l)}(2,3)$ covariance matrix $K_{2,3}^{(l)} = \mathbb{E}[y_2 \otimes y_3 | a_2 = l, \pi]$ with empirical average $\widehat{K}_{2,3}$ where $\otimes$ is tensor product. Assume that at time N, $\|K_{2,3}^{(l)} - \widehat{K}_{2,3}^{(l)}\|_2 \leq \epsilon_{2,3}^{(l)}$ with probability δ. Define function $g^\epsilon(N) = \frac{g}{N^{0.5-\epsilon}}$ where $g > 0$ and $0 < \epsilon < 0.5$. If $0.5\sigma_{k_\pi^{(l)}(2,3)} \geq g^\epsilon(N(l)) \geq \epsilon_{2,3}^{(N(l))}$ then with probability at least $(1-\delta)$, the rank $k_\pi^{(l)}$ is identifiable.*

Proof. If $\|K_{2,3}^l - \widetilde{K}_{2,3}^{(l)}\|_2 \leq \epsilon_{2,3}$ then the perturbation over singular values are bounded by value of $\epsilon_{2,3}$. If the rank of matrix $K_{2,3}^l$ is $k_\pi^{(l)}(2,3)$ then $k_\pi^{(l)}(2,3)$'th singular value of matrix $\widehat{K}_{2,3}^{(l)}$ can not get lower than $\sigma_{k_\pi^{(l)}(2,3)} - \epsilon_{2,3}^{(N(l))}$. In this case if $0.5\sigma_{k_\pi^{(l)}(2,3)} \geq g^\epsilon(N(l)) \geq \epsilon_{2,3}^{(N(l))}$ then one can decompose $\widehat{K}_{2,3}^{(l)}$ and find $k_\pi^{(l)}(2,3)$ by thresholding the singular values at level $g^\epsilon(N(l))$.

□

Proof : Lemma 1

Define the factor matrices $V_1^{(l)}, V_2^{(l)}, V_3^{(l)} \in \mathbb{R}^{Y \times k_\pi^{(l)}}$ as follows

$$\begin{aligned}
[V_1^{(l)}]_{j,i} &= \mathbb{P}(\vec{v}_1 = \vec{e}_j | \bar{x}_2 = i, \bar{a}_2 = l) \\
[V_2^{(l)}]_{j,i} &= \mathbb{P}(\vec{v}_2 = \vec{e}_j | \bar{x}_2 = i, \bar{a}_2 = l) \\
[V_3^{(l)}]_{j,i} &= \mathbb{P}(\vec{v}_3 = \vec{e}_j | \bar{x}_2 = i, \bar{a}_2 = l)
\end{aligned}$$

where second view and third view can be translated into

$$[V_2^{(l)}]_{j,i} = \frac{f_\pi(l|j) f_O(j|i)}{\mathbb{P}(\bar{a}_2 = l | \bar{x}_2 = i)} \ , \ [V_3^{(l)}]_{j,i} = \sum_{i'} f_O(j|i') f_T(i'|i,l)$$

The moments are restated as follows

$$\begin{aligned}
M_2^{(l)} &:= K_{1,3}^{(l)} = \mathbb{E}[v_1 \otimes v_3] \\
&= \sum_i^{k_\pi^{(l)}} \mathbb{P}(\bar{x}_2 = i | \bar{a}_2 = l) [V_1]_{(\cdot,i)} \otimes [V_3]_{(\cdot,i)}, \\
M_3^{(l)} &:= K_{1,3,2}^{(l)} = \mathbb{E}[v_1 \otimes v_3 \otimes v_2] \\
&= \sum_i^{k_\pi^{(l)}} \mathbb{P}(\bar{x}_2 = i | \bar{a}_2 = l) [V_1]_{(\cdot,i)} \otimes [V_3]_{(\cdot,i)} \otimes [V_2]_{(\cdot,i)}
\end{aligned}$$

where $k_\pi^{(l)}$ is cardinality of set $\mathcal{I}(l)$.

Lemma 5 (Concentration Bounds)**.** *By proposed robust power method in [Song et al., 2013], one can recover the columns of matrix V_2 with the following confidence bounds*

$$\|V_2(\cdot|i) - \widehat{V}_2(\cdot|i)\|_2 \leq C_O \sqrt{\frac{\log(1/\delta)}{N(l)}} := \mathcal{B}_O^{(l)} \tag{11}$$

if

$$N(l) > \overline{N} := \max_l \max_\pi \left(\frac{4}{\omega_{\min}^{(l)} \min_{\nu \in \{1,2,3\}} \{\sigma_{\min}^2(V_\nu^{(l)})\}} \right)^2 \log(2\frac{(Y^{1.5})}{\delta})\Theta^{(l)} \tag{12}$$

$$\Theta^{(l)} = \max \left\{ \frac{16X^{\frac{1}{3}}}{C_1^{\frac{2}{3}}(\omega_{\min}^{(l)})^{\frac{1}{3}}}, 4, \frac{2\sqrt{2}X}{C_1^2 \omega_{\min}^{(l)} \min_{\nu \in \{1,2,3\}} \{\sigma_{\min}^2(V_\nu^{(l)})\}} \right\}, \tag{13}$$

with probability at least $1-\delta$. Where C_1 is problem independent global constants and $\omega_{\min}^l := \min_\pi \min_i \mathbb{P}(x = i|a = l)$.

Proof. With simple modification of results in [Song et al., 2013], [Azizzadenesheli et al., 2016], and applying lemma 6 the results are followed.

At first we need to satisfy

$$\epsilon_{1,3}^t \leq 0.5\sigma_{k_\pi^{(l)}}(M_2^{(l)})$$

and then

lets define ϵ_M as follows

$$\epsilon_M \leq \frac{2\sqrt{2}\left\|\widehat{M}_3^{(l)} - M_3^{(l)}\right\|_2}{((\omega_{\min}^{(l)})^{\frac{1}{2}} \min_i \sigma_{k_\pi^{(l)}}(V_i^{(l)}))^3} + \frac{\left(\frac{4\left\|\widehat{M}_2^{(l)} - M_2^{(l)}\right\|_2}{((\omega_{\min}^{(l)})^{\frac{1}{2}} \min_i \sigma_{k_\pi^{(l)}}(V_i^{(l)}))^2}\right)^3}{\sqrt{(\omega^{(l)})_{\min}}} \tag{14}$$

Where $\sigma_{k_\pi^{(l)}}$ is $k_\pi^{(l)}$'th singular value of an input matrix, $\left(\omega^{(l)} = \min_i \left(\mathbb{P}\left(\bar{x}_2 = i|\bar{a}_2 = l\right)\right)\right)$ given applied policy. To have result in Eq. 11 we need to satisfy another sample complexity condition which is $\epsilon_M \leq \frac{C_1}{\sqrt{X}}$ where C_1 is global and problem independent constant which ends up to Eq. 12.

□

Proof. Theorem2

We begin with decomposing the regret equation

$$Reg_N = N\eta^* - \sum_{t=1}^N r_t$$

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

$$\sum_i^N r_t \geq \sum_{x,a} N(x,a)\bar{r}(x,a) - \sqrt{N\log(\frac{1}{\delta})}$$

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

$$Reg_N \leq N\eta^* - \sum_{x,a} N(x,a)\bar{r}(x,a) + \sqrt{N\log(\frac{1}{\delta})}$$

Let say at time N, the agent start new epoch $K+1$ then

$$Reg_N \leq \sum_{k}^{K} \Delta_k + \sqrt{N \log(\frac{1}{\delta})}$$

Δ_k is the average regret that the agent suffers at epoch k.

7.1 Failing Confidence

Let first consider the regret due to failing the confidence interval and failing spectral method. At the beginning of each epoch, the spectral method estimates columns of $V_2^{(l)}$ matrices and clusters the context set to construct the auxiliary set $S^{(k)}$. With high probability this clustering has no contradiction with the real model (there is no pair of contexts which are assigned to same state, when the do not belong to that state in true model). We first notice that if we redefine the confidence intervals in lemma 1 by substituting the term $\log(1/\delta)$ by $\log(At^6/\delta)$, we obtain, at any time instants t, the statement holds with probability $1-24\delta/t^6$. On the other hand, given set $S^{(k)}$, the confidence intervals in Eq. 1 hold by probability at least $1-\delta/15t^6$. The whole confidence bounds hold when both of these conditions hold. It means $\mathbb{P}(M \notin \mathcal{M}^t) \leq 24\delta/t^6 + \delta/15t^6 \leq 25\delta/t^6$

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

where $\mathcal{M}^{(t)}$ denotes the set of admissible MDPs according to the samples available at time and $N^{(k)}$ time at the beginning of epoch k. From sample complexity analysis in lemma 1, 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 M is in the set of $\mathcal{M}^{(k)}$ at any time step $\lfloor N^{1/4} \rfloor \leq t \leq N$ with probability $1 - 30\delta/N^{5/4}$. As a result, the regret due to failing confidence bound is bounded by $\sqrt{N}$ with probability $1 - 30\delta/N^{5/4}$.

7.2 Per Epoch regret

For the rest of the proof, let assume $M \in \mathcal{M}^{(t)}$ holds. Moreover, assume that the stopping criterion in iterative update Eq. 3 is when the accuracy is $\frac{1}{\sqrt{N^{(k)}}}$ then we have

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

where the state s belongs to $S^{(k)}$. Let's in epoch k, $\nu^{(k)}(l)$ denotes the number of time the action l has been chosen and $\nu^{(k)}(s,l)$ the number of time that the pair of state (s,l) has been happened.

$$\begin{aligned}
\sum_{s,a} \nu^{(k)}(s,a)(\widetilde{\eta} - \bar{r}(s,a)) + \sum_{s,a} \frac{\nu(s,a)}{\sqrt{N^{(k)}}} &= \sum_{s,a} \nu^{(k)}(s,a)(\widetilde{\eta} - \widetilde{r}(s,a)) + \sum_{s,a} \nu(s,a)(\widetilde{r}(s,a) - \bar{r}(s,a)) + \sum_{s,a} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}} \\
&\leq \vec{\nu}^k (\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)} + \sum_{s,a} \nu(s,a)(\widetilde{r}(s,a) - \bar{r}(s,a)) + 2\sum_{s,a} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}} \\
&\leq \vec{\nu}^k (\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)} + \sum_{s,a} 2\nu(s,a)(\widetilde{r}(s,a) - \widehat{r}(s,a)) + 2\sum_{s,a} \frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}
\end{aligned}$$

Therefore

$$\Delta^{(k)} \le \vec{\nu}^k(\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)} + \sum_{s,a} 2\nu(s,a)\sqrt{\frac{7\log(2S^{(k)}AN^{(k)}/\delta)}{2\max\{1,N^{(k)}(s,a)\}}} + 2\sum_{s,a}\frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}$$
$$\le \underbrace{\vec{\nu}^k(\widetilde{P}^{(k)} - I)\vec{u}_i^{(k)}}_{(a)} + \underbrace{2\sum_{s,a}\frac{\nu^{(k)}(s,a)}{\sqrt{N^{(k)}}}}_{(b)} + \underbrace{\left(\sqrt{14\log(2S^{(k)}AN^{(k)}/\delta)}\right)\sum_{s,a}\frac{\nu^{(k)}(s,a)}{\sqrt{2\max\{1,N^{(k)}(s,a)\}}}}_{(b')}$$

Because 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)

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

7.3 Clustering Rate

Currently, we studied the regret of each epoch separately. In theory, the behavior of the optimal policy of the optimistic mode on the real environment is not well known. In practice, in such models, e.g. $CMDP$, where the number of contexts is an order of magnitude larger than the number hidden states, it is not even the case that the optimistic policy maps just small number of contexts of a hidden state to one specific action. It means, if the policy maps one context to an action, then most likely it maps considerable portion of other contexts from the same state to the same action. Let α_p denotes the smallest probability of this portion of contexts given the hidden state. Potentially this value can get as low as $O_{\min}$ but in practice, it is considerably large.

We need to know how fast the set $\mathcal{S}^{(k)}$ reduce to set $\mathcal{X}$. Given Eq. [4], to cluster any context y, sufficient number of sample is needed to overcome the confidence bound. It means that the number of samples of the corresponding state and action, $\nu(l)$, should satisfy:

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

which means the required number of samples for corresponding action is $\nu(l) \ge 4C_O\frac{\log(1/\delta)}{(f_O(y|x(y)))^2}$ which called $\overline{N}(y)$. 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 represent the mean passing time between two steps of state x according to policy π. By Markov inequality, the probability that it takes more than $2\tau_M$ to start from state x and return back to it is at most $1/2$. Given the definition of α_p, it is clear that if the action l happens in state x the, in average this action will happen at state x, α_p portion of the time. If we divide the episode of length ν 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 plausible action, thus on average

we can have a total of $\nu\alpha_p/4\tau_M$ samples. Thus from Chernoff-Hoeffding, we obtain that the number of samples of state x and action$l$ is such that

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

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 context is reached. 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

$$\nu \geq \overline{\nu}(y) := \frac{4\tau_M}{\alpha_p}\left(\overline{N}(y) + \log(XA/\delta)\right). \tag{15}$$

With the same argument in Appendix [D] in [Azizzadenesheli et al., 2016] the number of required epochs to reveal context 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 step. Among this transition, at each epoch the agent face the problem with dimensionality of $S^{(k)}$.

Regret due to sample complexity

To deploy the spectral method, we need sufficient amount of samples to make use of spectral methods. In lemma 1 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.

7.4 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) :

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

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

$$D_Y Y\sqrt{14AN^{(k_2)}\log(2AN^{(k_2)}/\delta)} + DX\sqrt{14AN\log(2A(N-N^{(k_2)})/\delta)}$$

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

$$\sum_{k=1}^{k_2} \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + \bar{u}^{(k)}(s_{N^{(k+1)}}) - \bar{u}^{(k)}(s_{N^{(k)}}) + \sum_{k=k_2+1} \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + \bar{u}^{(k)}(s_{N^{(k+1)}}) - \bar{u}^{(k)}(s_{N^{(k)}})$$
$$\leq \sum_{k=1}^{k_2} \left(\sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + D_Y \right) + \sum_{k=k_2+1} \left(\sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + D_X \right)$$

As long as $|\zeta_t|$ is bounded by $D^{(k)}$ 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

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

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

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

$$\sum_k \sum_{t=N^{(k)}}^{t=N^{(k+1)}-1} \zeta_t + D^{(k)}$$
$$\leq D_Y\sqrt{2N^{(k_2)}\frac{5}{4}\log(\frac{8N}{\delta})} + D_Y A\log_2(8N^{(k_2)}/A) + D\sqrt{2(N-N^{(k_2)})\frac{5}{4}\log(\frac{8N}{\delta})} + DA\log_2(8N/A)$$

As we can see, at the beginning the agent suffers from huge regret due to the large MDP over space of context. 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. $X \leq S^k \leq Y$ and $D \leq D_{S^k} \leq D_Y$. 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 $\mathcal{O}(DX\sqrt{AN})$.

□

Remark 5. *The values D_Y and $D^{(k)}$ in the final bound are Diameter of true models under specific configuration in epoch k. But these term appears to upper bound the deviation values in optimistic models $\max_s u^{(k)}(s) - \min_s u^{(k)}(s) := \bar{D}^{(k)}$. Because of loose confidence bound before getting to epoch k_2, we get $\bar{D}^{(k)} << D^{(k)} << D_Y$*

Lemma 6 (Matrix Freedman [Tropp et al., 2011])**.** *Consider general d-dimension martingale matrices* $\{\chi_j : j = 0 \dots\}$ *and let* $\{\Psi_j : j = 1 \dots\}$ *be the difference martingale sequence. Assume the difference sequence is uniformly bounded*

$$\lambda_{\max}(\Psi_j) \leq C_\Psi \quad \textit{almost surely for all } j\textit{'s}$$

Define the predictable quadratic variation process of the martingale:

$$W_j := \sum_i^j \mathbb{E}_{i-1}[\Psi_i^2]. \textit{ for j=1,2,...}$$

Then for all $\epsilon \geq 0$ *and* $\sigma_\Psi^2 > 0$

$$\mathbb{P}[\exists j \geq 0 : \lambda_{\max}(\Phi_j) \geq \epsilon \textit{ and } \|W_j\| \leq \sigma_\Psi^2] \leq 2d. \exp\{-\frac{-\epsilon^2}{2\sigma_\Psi^2 + 2C_\Psi t/3}\}$$

in other word

$$\Phi_j \leq \frac{2C_\Psi}{3} + \sqrt{2\sigma_\Psi^2 \log(d/\delta)}$$

with probabilty at least $1 - \delta$

Proof. [Tropp et al., 2011] □