Reinforcement Learning in Factored MDPs: Oracle-Efficient Algorithms and Tighter Regret Bounds for the Non-Episodic Setting

Ziping Xu
Department of Statistics
University of Michigan
zipingxu@umich.edu

Ambuj Tewari
Department of Statistics
University of Michigan
tewaria@umich.edu

Abstract

We study reinforcement learning in non-episodic factored Markov decision processes (FMDPs). We propose two near-optimal and oracle-efficient algorithms for FMDPs. Assuming oracle access to an FMDP planner, they enjoy a Bayesian and a frequentist regret bound respectively, both of which reduce to the near-optimal bound $\widetilde{O}(DS\sqrt{AT})$ for standard non-factored MDPs. We propose a tighter connectivity measure, factored span, for FMDPs and prove a lower bound that depends on the factored span rather than the diameter D. In order to decrease the gap between lower and upper bounds, we propose an adaptation of the REGAL.C algorithm whose regret bound depends on the factored span. Our oracle-efficient algorithms outperform previously proposed near-optimal algorithms on computer network administration simulations.

1   Introduction

Designing computationally and statistically efficient algorithms is a core problem in Reinforcement Learning (RL). There is a rich line of works that achieve a strong sample efficiency guarantee with regret analysis in tabular MDPs, where state and action spaces are finite and small (Jaksch et al., 2010; Osband et al., 2013; Dann and Brunskill, 2015; Kearns and Singh, 2002). A current challenge in RL is dealing with large state and action spaces where even polynomial dependence of regret on state and action spaces size is unacceptable. One idea to meet this challenge is to consider MDPs with compact representations. For example, factored MDPs (FMDPs) (Boutilier et al., 2000) represent transition functions of MDPs using a compact Dynamic Bayesian network (DBN) (Ghahramani, 1997). FMDPs have a variety of applications in important real-world problems, e.g. multi-agent RL, and they also serve as important case studies in theoretical RL research (Guestrin et al., 2002a,c; Tavakol and Brefeld, 2014; Sun et al., 2019).

There is no FMDP planner that is both computationally efficient and accurate (Goldsmith et al., 1997; Littman, 1997). Guestrin et al. (2003) proposed approximate algorithms with prespecified basis functions and bounded approximation errors. For the even harder online learning setting, we study *oracle-efficient algorithms*, which learn an unknown FMDP efficiently when given access to a planning oracle. In this paper, our goal is to design efficient online algorithms that only make a polynomial number of calls to the oracle planning oracle. Side-stepping the computational intractability of the offline problem by assuming oracle access to a solver has yielded insights into simpler decision making problems. For example, oracle-based efficient algorithms have been proposed for the contextual bandit problem (Syrgkanis et al., 2016; Luo et al., 2018).

Online learning in *episodic* FMDP has been studied by Osband and Van Roy (2014). They proposed two algorithms, PSRL (Posterior Sampling RL) and UCRL-factored with near-optimal Bayesian and

Algorithms	Works	Regret bounds	Oracle
F-RMAX	Strehl et al. (2007)	(mixing rate) $T^{3/4}$	Planning on FMDP
UCRL-Factored	Osband and Van Roy (2014)	$DT^{1/2}$	Planning on Bounded FMDP
PSRL	This work	$DT^{1/2}$ (Bayesian)	Planning on FMDP
DORL	This work	$DT^{1/2}$	Planning on FMDP
FSRL	This work	$QT^{1/2}$	Optimizing average reward with bounded factored span

Table 1: Improvements of our methods over previous results on non-episodic FMDP in terms of regret bounds. There are three types of oracle: 1) Planning on a known FMDP; 2) Planning on a known bounded FMDP; 3) Optimizing average rewards over a confidence set with bounded factored span. As far as we known, the first oracle has multiple efficient approximate solutions, while the last two do not. Of all the algorithms, FSRL gives the tightest bound.

frequentist regret bounds, respectively. Their UCRL-factored algorithm relies on solving a Bounded FMDP (Givan et al., 2000), which is an even stronger assumption than the access to a planning oracle. Recently, Tian et al. (2020) improved the bound using a Bernstein-style confidence set.

This work studies FMDPs in the more challenging *non-episodic* setting. Previous studies in non-episodic FMDPs either have some high order terms in their analysis (Strehl, 2007) or depend on some strong connectivity assumptions, e.g. mixing time (Kearns and Koller, 1999). There is no near-optimal regret analysis in this setting yet.

Regret analysis in the non-episodic setting relies on the connectivity assumptions. Previous available connectivity assumptions include *mixing time* (Lewis and Puterman, 2001), *diameter* (Jaksch et al., 2010) and *span of bias vector* (Bartlett and Tewari, 2009). Mixing time is the strongest assumption and span of bias vector gives the tightest regret bound among the three. However, we show that even upper bound using span can be loose if the factor structure is not taken into account.

This paper makes three main contributions:

1. We provide two oracle-efficient algorithms, DORL (Discrete Optimism RL) and PSRL (Posterior Sampling RL), with near-optimal frequentist regret bound and Bayesian regret bound respectively. Both upper bounds depend on the *diameter* of the unknown FMDP. The algorithms call the FMDP planner only a polynomial number of times. The upper bound of DORL, when specialized to the standard non-factored MDP setting, matches that of UCRL2 (Jaksch et al., 2010). The same applies to the upper bound of PSRL in the non-factored setting (Ouyang et al., 2017).
2. We propose a tighter connectivity measure especially designed for FMDPs, called *factored span* and prove a regret lower bound that depends on the *factored span* of the unknown FMDP rather than its *diameter*.
3. Our last algorithm FSRL is not oracle efficient but its regret scales with factored span, and using it, we are able to *reduce the gap between upper and lower bounds on regret* in terms of both the dependence on diameter and on m, the number of factors.

2 Preliminaries

We first introduce necessary definitions and notation for non-episodic MDPs and FMDPs.

2.1 Non-episodic MDP

We consider a setting where a learning agent interacts without resets or episodes with a Markov decision process (MDP), represented by $M = \{\mathcal{S}, \mathcal{A}, P, R\}$, with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, the transition probability $P \in \mathcal{P}_{\mathcal{S} \times \mathcal{A}, \mathcal{S}}$ and reward distribution $R : \mathcal{P}_{\mathcal{S} \times \mathcal{A}, [0,1]}$. Here $\Delta(\mathcal{X})$ denotes a distribution over the space $\mathcal{X}$. Let $\mathcal{G}(\mathcal{X})$ be the space of all possible distributions over $\mathcal{X}$ and $\mathcal{P}_{\mathcal{X}_1, \mathcal{X}_2}$ is the class of all the mappings from space $\mathcal{X}_1$ to $\mathcal{G}(\mathcal{X}_2)$. Let $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$.

An MDP M and a learning algorithm $\mathcal{L}$ operating on M with an arbitrary initial state $s_1 \in \mathcal{S}$ constitute a stochastic process described by the state s_t visited at time step t, the action a_t chosen by

$\mathcal{L}$ at step t, the reward $r_t \sim R(s_t, a_t)$ and the next state $s_{t+1} \sim P(s_t, a_t)$ obtained for $t = 1, \ldots, T$. Let $H_t = \{s_1, a_1, r_1, \ldots, s_{t-1}, a_{t-1}, r_{t-1}\}$ be the trajectory up to time t.

Below we will define our regret measure in terms of undiscounted sum of rewards. To derive non-trivial upper bounds, we need some connectivity constraint. There are several subclasses of MDPs corresponding to different types of connectivity constraints (e.g., see the discussion in Bartlett and Tewari (2009)). We first focus on the class of *communicating* MDPs, i.e., the diameter of the MDP, which is defined below, is upper bounded by some $D < \infty$.

Definition 1 (Diameter). *Consider the stochastic process defined by a stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ operating on an MDP M with initial state s. Let $T(s' \mid M, \pi, s)$ be the random variable for the first time step in which state s' is reached in this process. Then the diameter of M is defined as*

$$D(M) := \max_{s \neq s' \in \mathcal{S}} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}\left[T\left(s' | M, \pi, s\right)\right].$$

A stationary policy π on an MDP M is a mapping $\mathcal{S} \mapsto \mathcal{A}$. An average reward (also called gain) of a policy π on M with an initial distribution s_1 is defined as

$$\lambda(M, \pi, s_1) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}[\sum_{t=1}^{T} r(s_t, \pi(s_t))],$$

where the expectation is over trajectories H_T. We restrict the choice of policies within the set Π of all policies whose average reward is independent of the starting state s_1. It can be shown that for a communicating MDP the optimal policies with the highest average reward are in the set and neither the optimal policy nor the optimal reward depends on the initial state. Let $\pi(M) = \arg\max_{\pi \in \Pi} \lambda(M, \pi, s_1)$ denote the optimal policy for MDP M and $\lambda^*(M)$ denote the optimal average reward or optimal gain. For any optimal policy $\pi(M)$, following the definition in Puterman (2014), we define

$$h(M, s) = \mathbb{E}[\sum_{t=1}^{\infty}(r_t - \lambda^*(M)) \mid s_1 = s], \text{ for } s = 1, \ldots S,$$

where the expectation is taken over the trajectory generated by policy $\pi(M)$. And the bias vector of MDP M is $\boldsymbol{h}(M) := (h(M, 1), \ldots, h(M, S))^T$. Let the span of a vector $\boldsymbol{h}$ be $sp(\boldsymbol{h}) := \max_{s_1, s_2} \boldsymbol{h}(s_1) - \boldsymbol{h}(s_2)$. Note that if there are multiple optimal policies, we consider the policy with the largest span for its bias vector.

We define the regret of a reinforcement learning algorithm $\mathcal{L}$ operating on MDP M up to time T as

$$R_T := \sum_{t=1}^{T}\left(\lambda^*(M) - r_t\right),$$

and Bayesian regret w.r.t. a prior distribution ϕ on a set of MDPs as $\mathbb{E}_{M \sim \phi} R_T$.

Optimality equation for average reward criterion. We let $\boldsymbol{R}(M, \pi)$ denote the S-dimensional vector with each element representing $\mathbb{E}_{r \sim R(s, \pi(s))}[r]$ and $P(M, \pi)$ denote the $S \times S$ matrix with each row as $P(s, \pi(s))$. For any communicating MDP M, the bias vector $\boldsymbol{h}(M)$ satisfies the following equation (Puterman, 2014):

$$\mathbf{1}\lambda^*(M) + \boldsymbol{h}(M) = \boldsymbol{R}(M, \pi^*) + P(M, \pi^*)\boldsymbol{h}(M). \tag{1}$$

2.2 Factored MDP

Factored MDP is modeled with a DBN (Dynamic Bayesian Network) (Dean and Kanazawa, 1989), where transition dynamics and rewards are factored and each factor only depends on a finite scope of state and action spaces. We use the definition in Osband and Van Roy (2014). We call $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ factored set if it can be factored by $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$. Note this formulation generalizes those in Strehl (2007); Kearns and Koller (1999) to allow the factorization of the action space as well. A visual illustration is shown in Figure 2.2.

Definition 2 (Scope operation for factored sets). *For any subset of indices $Z \subseteq \{1, 2, .., n\}$, let us define the scope set $\mathcal{X}[Z] := \bigotimes_{i \in Z} \mathcal{X}_i$. Further, for any $x \in \mathcal{X}$ define the scope variable $x[Z] \in \mathcal{X}[Z]$ to be the value of the variables $x_i \in \mathcal{X}_i$ with indices $i \in Z$. For singleton sets $\{i\}$, we write $x[i]$ for $x[\{i\}]$ in the natural way.*
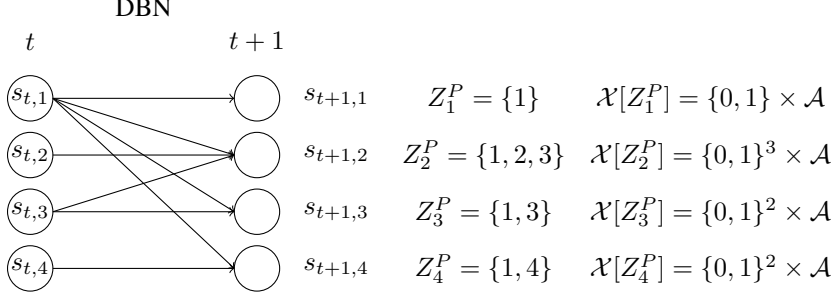
3

DBN

t $t+1$

$s_{t,1}$ $s_{t+1,1}$ $Z_1^P = \{1\}$ $\mathcal{X}[Z_1^P] = \{0,1\} \times \mathcal{A}$

$s_{t,2}$ $s_{t+1,2}$ $Z_2^P = \{1,2,3\}$ $\mathcal{X}[Z_2^P] = \{0,1\}^3 \times \mathcal{A}$

$s_{t,3}$ $s_{t+1,3}$ $Z_3^P = \{1,3\}$ $\mathcal{X}[Z_3^P] = \{0,1\}^2 \times \mathcal{A}$

$s_{t,4}$ $s_{t+1,4}$ $Z_4^P = \{1,4\}$ $\mathcal{X}[Z_4^P] = \{0,1\}^2 \times \mathcal{A}$

Figure 1: An example FMDP with 4 factors. Each node represents a dimension on a four dimensional vector. Each dimension is from $\{0,1\}$. Arrows represents the dependence between two state vectors at step t and $t+1$. In this example, all four dimensions depend on dimension 1 and themselves.

Definition 3 (Factored reward distribution). *A reward distribution R is factored over $\mathcal{X}$ with scopes $Z_1^R, \ldots, Z_l^R$ if and only if, for all $x \in \mathcal{X}$, there exist distributions $\left\{ R_i \in \mathcal{P}_{\mathcal{X}[Z_i^P], [0,1]} \right\}_{i=1}^l$ such that any $r \sim R(x)$ can be decomposed as $\sum_{i=1}^l r_i$, with each $r_i \sim R_i(x[Z_i^R])$ individually observable. Throughout the paper, we also let $R(x)$ denote reward function of the distribution $R(x)$, which is the expectation $\mathbb{E}_{r \sim R(x)}[r]$.*

Definition 4 (Factored transition probability). *A transition function P is factored over $\mathcal{S} \times \mathcal{A} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ and $\mathcal{S} = \mathcal{S}_1 \times \ldots \mathcal{S}_m$ with scopes $Z_1^P, \ldots, Z_m^P$ if and only if, for all $x \in \mathcal{X}, s \in \mathcal{S}$ there exists some $\left\{ P_i \in \mathcal{P}_{\mathcal{X}[Z_i], \mathcal{S}_i} \right\}_{i=1}^m$ such that,*

$$P(s|x) = \prod_{i=1}^m P_i \left(s[i] \mid x\left[Z_i \right] \right).$$

For simplicity, let $P(x)$ also denote the vector for the probability of each next state from current pair x. We define $P_i(x)$ in the same way.

Assumptions on FMDP. To ensure a finite number of parameters, we assume that $|\mathcal{X}[Z_i^R]| \leq L$ for $i \in [n]$, $|\mathcal{X}[Z_i^P]| \leq L$ for $i \in [m]$ and $|\mathcal{S}_i| \leq W$ for all $i \in [m]$ for some finite L and W. Furthermore, we assume that $r \sim R$ is in $[0,1]$ with probability 1.

Empirical estimates. We first define number of visits for each factored set. Let $N_{R_i}^t(x) \coloneqq \sum_{\tau=1}^{t-1} \mathbb{1}\{x_\tau[Z_i^R] = x\}$ be the number of visits to $x \in \mathcal{X}[Z_i^R]$ until t, $N_{P_i}^t(x)$ be the number of visits to $x \in \mathcal{X}[Z_i^P]$ until t and $N_{P_i}^t(s, x)$ be the number of visits to $x \in \mathcal{X}[Z_i^P], s \in \mathcal{S}_i$ until t. The empirical estimate for $R_i(x)$ is $\hat{R}_i^t(x) = \sum_\tau^{t-1} r_\tau \mathbb{1}\{x_\tau[Z_i^R] = x\} / \max\{1, N_{R_i}^t(x)\}$ for $i \in [l]$. Estimate for transition probability is $\hat{P}_i^t(s \mid x) = \frac{N_{P_i}^t(s,x)}{\max\{1, N_{P_i}^t(x)\}}$ for $i \in [m]$. Algorithms for non-episodic MDP typically proceed in internal episodes, where no reset takes place at the end of an episode. We let $N_{R_i}^k, \hat{R}_i^k$ and $\hat{P}_i^k$ be $N_{R_i}^{t_k}, \hat{R}_i^{t_k}$ and $\hat{P}_i^{t_k}$ with t_k being the first step of episode k.

3 Oracle-efficient Algorithms

We use PSRL (Posterior Sampling RL) and a modified version of UCRL-factored, called DORL (Discrete Optimism RL). The main idea of DORL is the simple fact that keeping a discrete set of MDPs reflecting the extreme points in the confidence set is sufficient for optimism. *By keeping a discrete set, one can avoid the hard computational oracle of planning on a bounded FMDP.*

Both PSRL and DORL use a fixed policy within an episode. For PSRL, we apply optimal policy for an MDP sampled from the posterior distribution of the true MDP. For DORL, instead of optimizing over a bounded MDP, we construct a new extended MDP, which is also factored with the number of parameters polynomial in that of the true MDP. Then the optimal policy of the extended FMDP is mapped to the policy space of the true FMDP. Instead of using dynamic episodes, we show that a

simple fixed episode scheme can also give us near-optimal regret bounds. The fixed episode scheme is proposed to ease the technical difficulty on the PSRL. Ouyang et al. (2017) combines doubling trick (Jaksch et al., 2010) and constraining that $T_{k+1} \leq T_k + 1$, which gives the same regret bound. However, this scheme changes episodes slower, which could be more efficient in practice. Algorithm details are shown in Algorithm 1.

3.1 Extended FMDP

Previous two constructions. Previous near-optimal algorithms on regular MDP depend on constructing an extended MDP with a high probability of being optimistic. Jaksch et al. (2010) constructs the extended MDP with a continuous action space to allow choosing any transition probability in a confidence set. This construction generates a bounded-parameter MDP. Agrawal and Jia (2017) instead sample MDPs only from the extreme points of a similar confidence set and combine them by adding an extra dimension on action space to decide which MDP to use.

Solving the bounded-parameter MDP with the first construction, which requires storing and ordering the S-dimensional bias vector, is not feasible for FMDPs. There is no direct adaptation that mitigates this computation issue. We show that the second construction using only a discrete set of MDPs can be solved with a much lower complexity in the FMDP setting.

We formally describe the construction. For simplicity, we ignore the notations for k in this section. First define the error bounds as an input. For every $x \in \mathcal{X}[Z_i^P]$, $s \in \mathcal{S}$, we have an error bound $E_{P_i}(s \mid x)$ for transition probability $\hat{P}_i(s \mid x)$. For every $x \in \mathcal{X}[Z_i^R]$, we have an error bound $E_{R_i}(x)$ for $\hat{R}_i(x)$. At the start of episode k the construction takes the inputs of $\hat{M}_k$ and the error bounds, and outputs the extended MDP M_k.

Extreme transition dynamic. We first define the extreme transition probability mentioned above in factored setting. Let $P_i(x)^{s+}$ be the transition probability that encourages visiting $s \in \mathcal{S}_i$, denoted by

$$P_i(x)^{s+} = P_i(x) - E_{P_i}(\cdot \mid x) + \mathbb{1}_s \sum_j E_{P_i}(j \mid x),$$

where $\mathbb{1}_j$ is the vector with all zeros except for a one on the j-th element. By this definition, $P_i(x)^{s+}$ is a new transition probability that puts all the uncertainty onto the direction s. An example is shown in Figure 2. Our construction assigns an action for each extreme transition dynamic.

$$\underbrace{\begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \end{pmatrix}}_{\text{Estimated dynamic}} - \underbrace{\begin{pmatrix} 0.1 \\ 0.05 \\ 0.05 \end{pmatrix}}_{\text{Uncertainty}} + \underbrace{\begin{pmatrix} 0 \\ 0.2 \\ 0 \end{pmatrix}}_{\text{Encourage visiting } s_2} = \underbrace{\begin{pmatrix} 0.4 \\ 0.45 \\ 0.15 \end{pmatrix}}_{\text{Extreme dynamic}}$$

Figure 2: An extreme transition dynamic that encourages visiting the second state out of three states.

Construction of extended FMDP. Our new factored MDP is denoted by $M_k = \{\mathcal{S}, \tilde{\mathcal{A}}, \tilde{P}, \tilde{R}\}$, where $\tilde{\mathcal{A}} = \mathcal{A} \times \mathcal{S}$ and the new scopes $\{\tilde{Z}_i^R\}_{i=1}^l$ and $\{\tilde{Z}_i^P\}_{i=1}^m$ are the same as those for the original MDP.

Let $\tilde{\mathcal{X}} = \mathcal{X} \times \mathcal{S}$. The new transition probability is factored over $\tilde{\mathcal{X}} = \bigotimes_{i \in [m]} (\mathcal{X}[Z_i^P] \times \mathcal{S}_i)$ and $\mathcal{S} = \bigotimes_{i \in [m]} \mathcal{S}_i$ with the factored transition probability to be

$$\tilde{P}_i(x, s[i]) := \hat{P}_i(x)^{s[i]+}, \text{ for any } x \in \mathcal{X}[Z_i^P], s \in \mathcal{S}.$$

The new reward function is factored over $\tilde{\mathcal{X}} = \bigotimes_{i \in [l]} (\mathcal{X}[Z_i^P] \times \mathcal{S}_i)$, with reward functions to be

$$\tilde{R}_i(x, s[i]) = \hat{R}_i(x) + E_{R_i}(x),$$

for any $x \in \mathcal{X}[Z_i^R], s \in \mathcal{S}$.

Claim 1. *The factored set $\tilde{\mathcal{X}} = \mathcal{S} \times \tilde{\mathcal{A}}$ of the extended MDP M_k satisfies each $|\tilde{\mathcal{X}}[Z_i^P]| \leq LW$ for any $i \in [m]$ and each $|\tilde{\mathcal{X}}[Z_i^R]| \leq LW$ for any $i \in [l]$.*

By Claim 1, any planner that efficiently solves the original MDP, can also solve the extended MDP. We find the best policy $\tilde{\pi}_k$ for M_k using the planner. To run a policy π_k on original action space, we choose π_k such that $(s, \pi_k(s)) = f(s, \tilde{\pi}_k(s))$ for every $s \in \mathcal{S}$, where $f : \tilde{\mathcal{X}} \mapsto \mathcal{X}$ maps any new state-action pair to the pair it is extended from, i.e. $f((x, s)) = x$ for any $(x, s) \in \tilde{\mathcal{X}}$.

Algorithm 1 PSRL and DORL

Input: $\mathcal{S}, \mathcal{A}$, accuracy ρ for DORL and prior distribution for PSRL, T, encoding $\mathcal{G}$ and L, upper bound on the size of each factor set.

$k \leftarrow 1; t \leftarrow 1; t_k \leftarrow 1; T_k \leftarrow 1; \mathcal{H} = \{\}$.

repeat

    For DORL:

        Construct the extended MDP M_k using error bounds:

$$E_{P_i}^k(s \mid x) = \min\{\sqrt{\frac{18\hat{P}_i(s|x)\log(c_{i,k})}{\max\{N_{P_i}^k(x), 1\}}} + \frac{18\log(c_{i,k})}{\max\{N_{P_i}^k(x), 1\}}, \hat{P}_i^k(s|x)\}, \tag{2}$$

        for $c_{i,k} = 6mS_i|\mathcal{X}[Z_i^P]|t_k/\rho$ and

$$E_{R_i}^k = \sqrt{\frac{12\log(6l|\mathcal{X}[Z_i^R]|t_k/\rho)}{\max\{N_{R_i}(x), 1\}}}. \tag{3}$$

        Compute $\tilde{\pi}_k = \pi(M_k)$ and find corresponding π_k in original action space.

    For PSRL:

        Sample M_k from $\phi(M|\mathcal{H})$.

        Compute $\pi_k = \pi(M_k)$.

    for $t = t_k$ **to** $t_k + T_k - 1$ **do**

        Apply action $a_t = \pi_k(s_t)$

        Observe new state s_{t+1}.

        Observe new rewards $r_{t+1} = (r_{t+1,1}, \ldots r_{t+1,l})$.

        $\mathcal{H} = \mathcal{H} \cup \{(s_t, a_t, r_{t+1}, s_{t+1})\}$.

    end for

    $k \leftarrow k + 1$.

    $T_k \leftarrow \lceil k/L \rceil; t_k \leftarrow t + 1$.

until $t_k > T$

---

4 Upper bounds for PSRL and DORL

We achieve the near-optimal Bayesian regret bound by PSRL and frequentist regret bound by DORL, respectively. Let $\tilde{O}$ denote the order ignoring the logarithmic term and the universal constant.

Theorem 1 (Regret of PSRL). *Let M be the factored MDP with graph structure $\mathcal{G} = \left(\{\mathcal{S}_i\}_{i=1}^m; \{\mathcal{X}_i\}_{i=1}^n; \{Z_i^R\}_{i=1}^l; \{Z_i^P\}_{i=1}^m\right)$, all $|\mathcal{X}[Z_i^R]|$ and $|\mathcal{X}[Z_j^P]| \leq L$, $|\mathcal{S}_i| \leq W$ and diameter upper bounded by D. Then if ϕ is the true prior distribution supported on the set of MDPs with diameter $\leq D$, then we bound Bayesian regret of PSRL:*

$$\mathbb{E}[R_T] = \tilde{O}(D(l + m\sqrt{W})\sqrt{TL}).$$

Theorem 2 (Regret of DORL). *Let M be the factored MDP with graph structure $\mathcal{G} = \left(\{\mathcal{S}_i\}_{i=1}^m; \{\mathcal{X}_i\}_{i=1}^n; \{Z_i^R\}_{i=1}^l; \{Z_i^P\}_{i=1}^m\right)$, all $|\mathcal{X}[Z_i^R]|$ and $|\mathcal{X}[Z_j^P]| \leq L$, $|\mathcal{S}_i| \leq W$ and diameter upper bounded by D. Then, with high probability, regret of DORL is upper bounded by:*

$$R_T = \tilde{O}(D(l + m\sqrt{W})\sqrt{TL}).$$

The two bounds match the frequentist regret bound in Jaksch et al. (2010) and Bayesian regret bound in Ouyang et al. (2017) for non-factored communicating MDP. We also give a condition of designing the speed of changing policies.

Remark. *Replacing the episode length in Algorithm 1 with any $\{T_k\}_{k=1}^K$ that satisfies $K = O(\sqrt{LT})$ and $T_k = O(\sqrt{T/L})$ for all $k \in [K]$, the frequentist bound in Theorem 2 still holds. Furthermore, if $\{T_k\}_{k=1}^K$ is fixed the Bayesian bound in Theorem 2 also holds.*

5 Lower Bound and Factored Span

Any regret bound depends on a difficulty measure determining the connectivity of the MDP. The upper bounds of DORL and PSRL use diameter. A tighter alternative is the span of bias vector (Bartlett and Tewari, 2009), defined as $sp(\boldsymbol{h}^*)$, where $\boldsymbol{h}^*$ is the bias vector of the optimal policy. However, none of those connectivity measures address the complexity of the graph structure. Indeed, some graph structure allows a tighter regret bound. In this section, we first show a lower bound with a Cartesian product structure. We further propose a new connectivity measure that can scale with the complexity of the graph structure.

Large diameter case. We consider a simple FMDP with infinite diameter but bounded span. The FMDP is a Cartesian product of two identical MDPs, M_1 and M_2 with $\mathcal{S} = \{0, 1, 2, 3\}$, $\mathcal{A} = \{1, 2\}$. The transition probability is chosen such that from any state and action pair, the next state will either move forward or move backward with probability one (state 0 is connected with state 3 as a circle).

We can achieve a low regret easily by learning each MDP independently. However, since the sum of the two states always keeps the same parity, vector state $(0, 1)$ can never be reached from $(0, 0)$. Thus, the FMDP has an infinite diameter. The span of bias vector, on the other hand, is upper bounded by $D(M_1) + D(M_2)$, which is tight in this case.

Lower bound with only dependency on span. Let us formally state the lower bound. Our lower bound casts some restrictions on the scope of transition probability, i.e. the scope contains itself, which we believe is a natural assumption. We provide Theorem 3 here.

Theorem 3 (Lower bound). *For any algorithm, any graph structure satisfying $\mathcal{G} = \left(\{\mathcal{S}_i\}_{i=1}^n; \{\mathcal{S}_i \times \mathcal{A}_i\}_{i=1}^n; \{Z_i^R\}_{i=1}^n; \{Z_i^P\}_{i=1}^n\right)$ with $|\mathcal{S}_i| \leq W$, $|\mathcal{X}[Z_i^R]| \leq L$, $|\mathcal{X}[Z_i^P]| \leq L$ and $i \in Z_i^P$ for $i \in [n]$, there exists an FMDP with an optimal bias vector $\boldsymbol{h}$, such that for any initial state $s \in \mathcal{S}$, the expected regret of the algorithm after T steps is $\Omega(\sqrt{sp(\boldsymbol{h})LT})$.*

The proof is given in Appendix B. Note that this bound of $\Omega(\sqrt{sp(\boldsymbol{h})LT})$ seems weaker than the lower bound $\sqrt{DSAT}$ when reduced to non-factored MDP case. However, the two bounds are exactly the same as in the construction from Jaksch et al. (2010), the diameter scales with span of bias vector by a constant factor. As we can see, the upper bound in Theorem 1 is larger than the lower bound by a factor of $\frac{D}{\sqrt{sp(\boldsymbol{h})}}$, m, l and $\sqrt{W}$. We now discuss how to reduce the first three excesses.

5.1 Tighter connectivity measure

The mismatch in the dependence on m is due to not taking the factor structure into account properly in the definition of the span. A tighter bound should be able to detect special structure, e.g. product of independent MDPs. For that purpose, recall that regret analysis involves in bounding a deviation term with the form

$$\sum_{s \in \mathcal{S}} \left(\tilde{P}^k(s) - P^k(s)\right) h_k(s) \leq span(\boldsymbol{h}_k) \max_s |\tilde{P}_k(s) - P_k(s)|_1,$$

where $\tilde{P}^k$ and $\tilde{P}^k$ are the transition probability of M_k and M using the policy $\tilde{\pi}_k$ and π_k. Here h_k is the optimal bias vector of M_k. This inequality could be extremely loose as $\boldsymbol{h}_k$ has exponentially many dimensions. We use the following example to illustrate how to get a tighter inequality by looking at each factor on the bias vector.

Reducing span by looking at each factor. Let us write a bias vector of an FMDP with two factors as a matrix in the following table, where columns and rows represent the first and second dimension of the state vector respectively. The state set is given by $\{(x, y) : x \in \{x_1, x_2, x_3, x_4\}, y \in \{y_1, y_2, y_3\}\}$.

factor 2 & factor 1	x_1	x_2	x_3	x_4	span of rows
y_1	1	3	3	4	3
y_2	4	3	6	7	4
y_3	7	8	9	10	3
span of columns	6	5	6	6	9

While the total span of the vector is **9**, the span for each column and each row is at most **6** and **4**. We also notice that in the lower bound construction with two independent MDPs, the value that restricts the connectivity is the span of each column and each row. Even for the non-independent case, using the span of the whole vector will still lead a quite loose bound.

This example encourages us to construct a tighter connectivity measure *Factored Span* in Definition 5: for each factor, we write the bias vector into a matrix with the current factor as columns and the rest of factors as rows. Factored span is the maximum span of all the columns.

Definition 5 (Factored span). *For an FMDP M with an bias vector $\boldsymbol{h}$ of its optimal policy and a factorization of state space $\mathcal{S} = \bigotimes_{i=1}^{m} \mathcal{S}_i$, we define factored span $sp_1, \ldots, sp_m$ as:*

$$sp_i := \max_{s_{-i}} sp(\boldsymbol{h}(\cdot, s_{-i})) \text{ and let } Q(\boldsymbol{h}) := \sum_{i=1}^{m} sp_i,$$

where $s_{-i} := (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_m)$ and $sp(h(\cdot, s_{-i})) := (h(s, s_{-i}))_{s \in \mathcal{S}_1}$.

Proposition 1. *For any bias vector $\boldsymbol{h}$, $sp(\boldsymbol{h}) \leq Q(\boldsymbol{h}) \leq m \, sp(\boldsymbol{h})$. The first equality holds when the FMDP has the construction of Cartesian product of m independent MDPs. The lower bound can also be written as $\Omega(\sqrt{Q(\boldsymbol{h})LT})$.*

5.2 Tighter upper bound

We now provide another algorithm called FSRL (Factored-span RL) with a tighter regret bound of $\tilde{O}(Q(\boldsymbol{h})\sqrt{WLT})$ as shown in Theorem 4. The bound reduces the gap on m, l and replaces D with the sum of factored span Q. Proposition 1 guarantees that $Q(\boldsymbol{h}) \leq msp(\boldsymbol{h}) \leq mD$ such that the upper bound is at least as good as the upper bound in Theorem 1.

FSRL (full description in Appendix D, Algorithm 2), mimics REGAL.C by solving the following optimization,

$$M = \arg\max_{M \in \mathcal{M}_k} \lambda^*(M) \quad \text{subject to} \quad Q(\boldsymbol{h}(M)) \leq Q \text{ for some prespecified } Q > 0,$$

where $\mathcal{M}_k$ is the confidence set defined in (2) and (3). FSRL relies on the computational oracle of optimizing average rewards over the confidence set with the sum of factored span bounded by a prespecified value. Therefore, FSRL cannot be run by just calling an FMDP planning oracle.

Theorem 4 (Regret of FSRL (Factored-span RL)). *Let M be the factored MDP with graph structure $\mathcal{G} = \left(\{\mathcal{S}_i\}_{i=1}^{m}; \{\mathcal{X}_i\}_{i=1}^{n}; \{Z_i^R\}_{i=1}^{l}; \{Z_i^P\}_{i=1}^{m}\right)$, all $|\mathcal{X}[Z_i^R]|$ and $|\mathcal{X}[Z_j^P]| \leq L$, $|\mathcal{S}_i| \leq W$, bias vector of optimal policy $\boldsymbol{h}$ and its sum of factored spans $Q(\boldsymbol{h})$. Then, with high probability, regret of FSRL is upper bounded by: $R_T = \tilde{O}(Q(\boldsymbol{h})\sqrt{WLT})$.*

Main idea is on bounding the deviation of transition probabilities between the true MDP and M_k in episode k with factored span. The details are shown in Appendix C.

6 Simulation

There are two previously available sample efficient and implementable algorithms for FMDPs: factored E^3 and factored Rmax (f-Rmax). F-Rmax was shown to have better empirical performance (Guestrin et al., 2002b). Thus, we compare PSRL, DORL and f-Rmax. For PSRL, at the start of each episode, we simply sample each factored transition probability and reward functions from a Dirichlet distribution and a Gaussian distribution, i.e. $P_i^k(x) \sim \text{Dirichlet}(N_{P_i}^t(\cdot, x)/c)$ and $R_i^k(x) \sim N(\hat{R}_i^k(x), c/N_{P_i}^t(x))$, where c is searched over $(0.05, 0.1, 0.3, 0.75, 1, 5, 20)$. The total number of samplings for PSRL in each round is upper bounded by the number of parameters of the FMDP. For DORL, we replace the coefficients 18 and 12 in (2) and (3) with a hyper-parameter c searched over
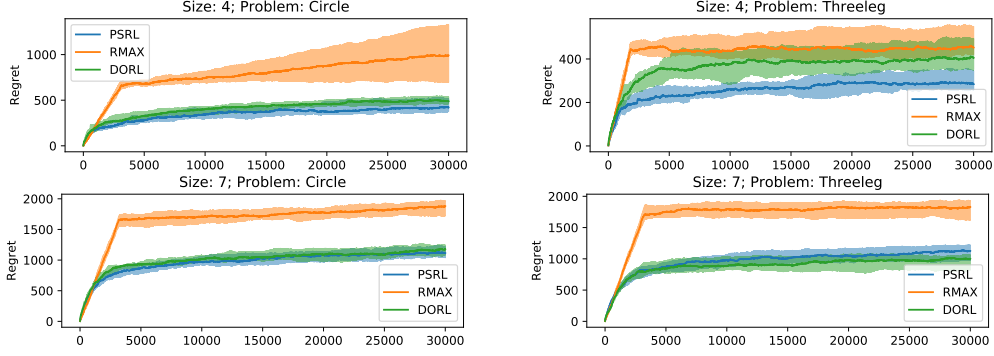
Figure 3: Regrets of PSRL, f-Rmax and DORL on circle and three-leg MDP with a size 4, 7. For PSRL, $c = 0.75$. For f-Rmax, $m = 300, 500, 500, 500$ and for DORL, $c = 0.03$ in Circle 4, Circle 7, Three-leg 4, Three-leg 7, respectively.

the set $\{0.05, 0.1, 0.3, 0.75, 1, 5, 20\}$. The choice of hyper-parameters is based on the fact that our regret analysis bounds the worst-case scenario, which lead to over-exploration, while in practice the level of exploration should be lower. For f-Rmax, m, the number of visits needed to be known are chosen from 100, 300, 500, 700 and the best choice is selected for each experiment.

For the approximate planner used by our algorithm, we implement approximate linear programming (Guestrin et al., 2003) with the basis $h_i(s) = s_i$ for $i \in [m]$. For regret evaluation, we use an accurate planner to find the true optimal average reward.

We compare three algorithms on computer network administrator domain with a circle and a three-leg structure (Guestrin et al., 2001; Schuurmans and Patrascu, 2002). To avoid the extreme case in our lower bound, both the MDPs are set to have limited diameters. The details on the environment are in Appendix E.

Figure 3 shows the regret of the two algorithms on circle and three-leg structure with a size 4, 7, respectively. Each experiment is run 20 times, with which median, 75% and 25% quantiles are computed. Our DORL and PSRL have very similar performance in all the environment except for Three-leg with a size 4. Optimal hyper-parameter for PSRL and DORL is stable in the way that c around 0.75 and 0.03 are the optimal parameter for PSRL and DORL respectively for all the experiments. Note that we use the exact, not approximate, optimal reward in regret evaluation. So we see that DORL and PSRL was always able to find a near-optimal optimal policy despite the use of an approximate planner.

7 Discussion

FSRL relies on a harder computational oracle that is not efficiently solvable yet. Fruit et al. (2018) achieved a regret bound depending on span using an implementable Bias-Span-Constrained value iteration on non-factored MDP. It remains unknown whether FSRL could be approximately solved using an efficient implementation.

In non-factored MDP, Zhang and Ji (2019) achieved the lower bound of $\sqrt{DSAT}$. On the lower bound of FMDP, it remains an open problem to close the remaining gap involving $\sqrt{W}$ and $\sqrt{Q}$. Recently, Sadegh Talebi et al. (2020) improved the exploration using a Berstein-style confidence set and was able to reduce the extra term on $\sqrt{W}$. However, their algorithm depends on the harder planning oracle on bounded FMDP. UCRL3 (Bourel et al., 2020) improves the dependence on actions and states by restricting the set of successor states of the state-action pair on non-factored MDP. We believe this construction is quite independent from the FMDP construction and the similar adaptation can easily be applied. Our algorithms require the full knowledge of the graph structure of the FMDP, which can be impractical. The structural learning scenario has been studied by Strehl et al. (2007); Chakraborty and Stone (2011); Hallak et al. (2015). Their algorithms either rely on an admissible structure learner or do not have a regret or sample complexity guarantee. Recently, Rosenberg and Mansour (2020) proposed an oracle-efficient structure learner with regret guarantee.

9

8 Broader Impacts

As a theoretical paper, we can not foresee any direct societal consequences in the near future. Factored MDP, the main problem we study in this paper, may be used in multi-agent Reinforcement Learning scenario.

References

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194.

Bartlett, P. L. and Tewari, A. (2009). Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press.

Bourel, H., Maillard, O.-A., and Talebi, M. S. (2020). Tightening exploration in upper confidence reinforcement learning. *arXiv preprint arXiv:2004.09656*.

Boutilier, C., Dearden, R., and Goldszmidt, M. (2000). Stochastic dynamic programming with factored representations. *Artificial intelligence*, 121(1-2):49–107.

Chakraborty, D. and Stone, P. (2011). Structure learning in ergodic factored mdps without knowledge of the transition function's in-degree. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 737–744. Citeseer.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational intelligence*, 5(2):142–150.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586.

Ghahramani, Z. (1997). Learning dynamic bayesian networks. In *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pages 168–197. Springer.

Givan, R., Leach, S., and Dean, T. (2000). Bounded-parameter markov decision processes. *Artificial Intelligence*, 122(1-2):71–109.

Goldsmith, J., Littman, M. L., and Mundhenk, M. (1997). The complexity of plan existence and evaluation in robabilistic domains. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 182–189.

Guestrin, C., Koller, D., and Parr, R. (2001). Max-norm projections for factored mdps. In *IJCAI*, volume 1, pages 673–682.

Guestrin, C., Koller, D., and Parr, R. (2002a). Multiagent planning with factored mdps. In *Advances in neural information processing systems*, pages 1523–1530.

Guestrin, C., Koller, D., Parr, R., and Venkataraman, S. (2003). Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468.

Guestrin, C., Patrascu, R., and Schuurmans, D. (2002b). Algorithm-directed exploration for model-based reinforcement learning in factored mdps. In *ICML*, pages 235–242. Citeseer.

Guestrin, C., Venkataraman, S., and Koller, D. (2002c). Context-specific multiagent coordination and planning with factored mdps. In *AAAI/IAAI*, pages 253–259.

Hallak, A., Schnitzler, F., Mann, T., and Mannor, S. (2015). Off-policy model-based learning under unknown factored dynamics. In *International Conference on Machine Learning*, pages 711–719.

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, M. and Koller, D. (1999). Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747.

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

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM.

Lewis, M. E. and Puterman, M. L. (2001). A probabilistic analysis of bias optimality in unichain markov decision processes. *IEEE Transactions on Automatic Control*, 46(1):96–100.

Littman, M. L. (1997). Probabilistic propositional planning: representations and complexity. In *Proceedings of the fourteenth national conference on artificial intelligence*, pages 748–754.

Luo, H., Wei, C.-Y., Agarwal, A., and Langford, J. (2018). Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory*, pages 1739–1776.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

Osband, I. and Van Roy, B. (2014). Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pages 604–612.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017). Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342.

Puterman, M. L. (2014). *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rosenberg, A. and Mansour, Y. (2020). Oracle-efficient reinforcement learning in factored mdps with unknown structure. *arXiv preprint arXiv:2009.05986*.

Sadegh Talebi, M., Jonsson, A., and Maillard, O.-A. (2020). Improved exploration in factored average-reward mdps. *arXiv e-prints*, pages arXiv–2009.

Schuurmans, D. and Patrascu, R. (2002). Direct value-approximation for factored mdps. In *Advances in Neural Information Processing Systems*, pages 1579–1586.

Strehl, A. L. (2007). Model-based reinforcement learning in factored-state mdps. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 103–110. IEEE.

Strehl, A. L., Diuk, C., and Littman, M. L. (2007). Efficient structure learning in factored-state mdps. In *AAAI*, volume 7, pages 645–650.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933.

Syrgkanis, V., Luo, H., Krishnamurthy, A., and Schapire, R. E. (2016). Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems*, pages 3135–3143.

Tavakol, M. and Brefeld, U. (2014). Factored mdps for detecting topics of user sessions. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 33–40.

Tian, Y., Qian, J., and Sra, S. (2020). Towards minimax optimal reinforcement learning in factored markov decision processes. *arXiv preprint arXiv:2006.13405*.

Zhang, Z. and Ji, X. (2019). Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pages 2823–2832.

A Proof of Theorem 1 and 2

A standard regret analysis consists of proving the optimism, bounding the deviations and bounding the probability of failing the confidence set. Our analysis follows the standard procedure while adapting them to a FMDP setting.

Some notations. For simplicity, we let π^* denote the optimal policy of the true MDP, $\pi(M)$. Let t_k be the starting time of episode k and K be the total number of episodes. Since $\tilde{R}^k(x, s)$ for any $(x, s) \in \tilde{\mathcal{X}}$ does not depend on s, we also let $\tilde{R}^k(x)$ denote $\tilde{R}^k(x, s)$ for any s. Let λ^* and λ_k denote the optimal average reward for M and M_k.

Confidence set. Before proving the theorems, we first introduce the confidence set for both transition probability and reward functions. Let $\mathcal{M}_k$ be the confidence set of FMDPs at the start of episode k with the same factorization, such that for and each $i \in [l]$,

$$|R_i(x) - \hat{R}_i^k(x)| \le E_{R_i}^k(x), \forall x \in \mathcal{X}[Z_i^R],$$

where $E_{R_i}^k(x) := \sqrt{\frac{12 \log(6l|\mathcal{X}[Z_i^R]|t_k/\rho)}{\max\{N_{R_i}^k(x),1\}}}$ as defined in (3);
and for each $j \in [m]$

$$|P_j(s|x) - \hat{P}_j^k(s|x)| \le E_{P_j}^k(s|x), \forall x \in \mathcal{X}[Z_j^P], s \in \mathcal{S}_j,$$

where $W_{P_j}^k(s|x)$ is defined in (2). It can be shown that

$$|P_j(x) - \hat{P}_j^k(x)|_1 \le 2\sqrt{\frac{18|\mathcal{S}_i| \log(6S_i m|\mathcal{X}[Z_i^P]|t_k/\rho)}{\max\{N_{P_i}^k(x),1\}}},$$

where $\bar{E}_{P_i}^k(x) := 2\sqrt{\frac{18|\mathcal{S}_i| \log(6S_i m|\mathcal{X}[Z_i^P]|t_k/\rho)}{\max\{N_{P_i}^k(x),1\}}}$.

In the following analysis, we all assume that the true MDP M for both PSRL and DORL are in $\mathcal{M}_k$ and M_k by PSRL are in $\mathcal{M}_k$ for all $k \in [K]$. In the end, we will bound the regret caused by the failure of confidence set.

Regret decomposition. We follow the standard regret analysis framework by Jaksch et al. (2010). We first decompose the total regret into three parts in each episode:

$$R_T = \sum_{t=1}^{T} (\lambda^* - r_t)$$

$$= \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} (\lambda^* - \lambda_k) \tag{4}$$

$$+ \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} (\lambda_k - R(s_t, a_t)) \tag{5}$$

$$+ \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} (R(s_t, a_t) - r_t). \tag{6}$$

Using Hoeffding's inequality, the regret caused by (6) can be upper bounded by $\sqrt{\frac{5}{2} T \log\left(\frac{8}{\rho}\right)}$, with probability at least $\frac{\rho}{12}$.

A.1 Bounding term (4)

We bound the regret caused by (4).

PSRL. For PSRL, since we use fixed episodes, we follow the techniques from Osband et al. (2013) and show that the expectation of (4) equals to zero.

Lemma 1 (Lemma 1 in Osband et al. (2013)). *If ϕ is the distribution of M, then, for any $\sigma(H_{t_k}) - measurable$ function g,*

$$\mathbb{E}\left[g\left(M\right)|H_{t_k}\right] = \mathbb{E}\left[g\left(M_k\right)|H_{t_k}\right].$$

We let $g = \lambda(M, \pi(M))$. As g is a $\sigma(H_{t_k}) - measurable$ function. Since t_k, K are fixed value for each k, we have $(4) = \mathbb{E}[\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}(\lambda^* - \lambda_k)] = 0$.

DORL. For DORL, we need to prove optimism, i.e, $\lambda(M_k, \tilde{\pi}_k) \geq \lambda^*$ with high probability. We follow the proof in Agrawal and Jia (2017). In the case of FMDP, we show that for any policy π for the true FMDP, there exists a policy $\tilde{\pi}$ for M_k such that $(P(M_k, \tilde{\pi}) - P(M, \pi))h \geq 0$ for any $h \in \mathbb{R}^S$. This is proved in Lemma 2.

Lemma 2. *For any policy π for M and any vector $h \in \mathbb{R}^S$, let $\tilde{\pi}$ be the policy for M_k satisfying $\tilde{\pi}(s) = (\pi(s), s^*)$, where $s^* = \arg\max_s h(s)$. Then, given $M \in \mathcal{M}_k$, $(P(M_k, \tilde{\pi}) - P(M, \pi))h \geq 0$.*

Proof. We fix some $s \in \mathcal{S}$ and let $x = (s, \pi(s)) \in \mathcal{X}$. Recall that for any $s_i \in \mathcal{S}_i$, $\Delta_i^k(s_i|x) =$

$$\min\left\{\sqrt{\frac{18\hat{P}_i^k(s_i|x)\log(c_{i,k})}{\max\left\{N_{P_i}^k(x), 1\right\}}} + \frac{18\log(c_{i,k})}{\max\left\{N_{P_i}^k(x), 1\right\}}, \hat{P}_i^k(s_i|x)\right\}.$$

and define $P_i^-(\cdot|x) = \hat{P}_i^k(\cdot|x) - \Delta_i^k(\cdot|x)$. Slightly abusing the notations, let $\tilde{P} = P(M_k, \tilde{\pi})_{s,\cdot}$, $P = P(M, \pi)_{s,\cdot}$. Define two S-dimensional vectors $\hat{P}$ and P^- with $\hat{P}(\bar{s}) = \prod_i \hat{P}_i(\bar{s}[Z_i^P]|x)$ and $P^-(\bar{s}) = \Pi_i P_i^-(\bar{s}[Z_i^P]|x)$ for $\bar{s} \in \mathcal{S}$.

As $M \in \mathcal{M}_k$, $P^- \leq P$. Define $\alpha := \hat{P} - P \leq \hat{P} - P^- =: \Delta$. Without loss of generality, we let $\max_s h(s) = D$.

$$
\begin{aligned}
\sum_i \tilde{P}(i)h(i) &= \sum_i P(i)^- h(i) + D\left(1 - \sum_j P(j)^-\right) \\
&= \sum_i P(i)^- h(i) + D\sum_j \Delta(j) \\
&= \sum_i \left(\hat{P}(i) - \Delta(i)\right)h(i) + D\Delta(i) \\
&= \sum_i \hat{P}(i)h(i) + (D - h(i))\Delta(i) \\
&\geq \sum_i \hat{P}(i)h(i) + (D - h(i))\alpha(i) \\
&= \sum_i \left(\hat{P}(i) - \alpha(i)\right)h(i) + D\alpha(i) \\
&= \sum_i P(i)h(i) + D\sum_i \alpha(i) = \sum_i P(i)h(i)
\end{aligned}
$$

$\square$

Corollary 1. *Let $\tilde{\pi}^*$ be the policy that satisfies $\tilde{\pi}^*(s) = (\pi^*(s), s^*)$, where $s^* = \arg\max_s h(M)_s$ and π^* is the true optimal policy for M. Then $\lambda(M_k, \tilde{\pi}^*, s_1) \geq \lambda^*$ for any starting state s_1.*

Proof. Let $\boldsymbol{d}(s_1) := \boldsymbol{d}(M_k, \tilde{\pi}^*, s_1) \in \mathbb{R}^{1 \times S}$ be the row vector of stationary distribution starting from some $s_1 \in \mathcal{S}$. By optimal equation,

$$
\begin{aligned}
&\lambda(M_k, \tilde{\pi}^*, s_1) - \lambda^* \\
&= \boldsymbol{d}(s_1)\boldsymbol{R}(M_k, \tilde{\pi}^*) - \lambda^*(\boldsymbol{d}(s_1)\mathbf{1}) \\
&= \boldsymbol{d}(s_1)(\boldsymbol{R}(M_k, \tilde{\pi}^*) - \lambda^*\mathbf{1}) \\
&= \boldsymbol{d}(s_1)(\boldsymbol{R}(M_k, \tilde{\pi}^*) - \boldsymbol{R}(M, \pi^*)) \\
&\quad + \boldsymbol{d}(s_1)(I - P(M, \pi^*))\boldsymbol{h}(M) \\
&\geq \boldsymbol{d}(s_1)(\boldsymbol{R}(M_k, \tilde{\pi}^*) - \boldsymbol{R}(M, \pi^*)) \\
&\quad + \boldsymbol{d}(s_1)(P(M_k, \tilde{\pi}^*) - P(M, \pi^*))\boldsymbol{h}(M) \\
&\geq 0,
\end{aligned}
$$

where the last inequality is by Lemma 2 and Corollary 1 follows. $\square$

Thereon, $\lambda(M_k, \tilde{\pi}_k) \geq \lambda(M_k, \tilde{\pi}^*, s_1) \geq \lambda^*$. The total regret of (4) ≤ 0.

A.2 Regret caused by deviation (5)

We further bound regret caused by (5), which can be decomposed into the deviation between our brief M_k and the true MDP. We first show that the diameter of M_k can be upper bounded by D.

Bounded diameter. We need diameter of extended MDP to be upper bounded to give a sublinear regret. For PSRL, since prior distribution has no mass on MDP with diameter greater than D, the diameter of MDP from posterior is upper bounded by D almost surely. For DORL, we have the following Lemma 3.

Lemma 3. *When M is in the confidence set $\mathcal{M}_k$, the diameter of the extended MDP $D(M_k) \leq D$.*

Proof. Fix a $s_1 \neq s_2$, there exist a policy π for M such that the expected time to reach s_2 from s_1 is at most D, without loss of generality we assume s_2 is the last state. Let E be the $(S - 1) \times 1$ vector with each element to be the expected time to reach s_2 except for itself. We find $\tilde{\pi}$ for M_k such that the expected time to reach s_2 from s_1 can be bounded by D. We choose the $\tilde{\pi}$ that satisfies $\tilde{\pi}(s) = (\pi(s), s_2)$.

Let Q be the transition matrix under $\tilde{\pi}$ for M_k. Let Q^- be the matrix removing s_2-th row and column and P^- defined in the same way for M. We immediately have $P^{-1}E \geq Q^{-1}E$, given $M \in \mathcal{M}_k$. Let $\tilde{E}$ be the expected time to reach s_2 from every other states except for itself under $\tilde{\pi}$ for M_k.

We have $\tilde{E} = \mathbf{1} + Q^-\tilde{E}$. The equation for E gives us $E = \mathbf{1} + P^-E \geq \mathbf{1} + Q^-E$. Therefore,

$$
\tilde{E} = (1 - Q^-)^{-1}\mathbf{1} \leq E,
$$

and $\tilde{E}_{s_1} \leq E_{s_1} \leq D$. Thus, $D(M_k) \leq D$. $\square$

Deviation bound. Now we formally bound (5). In this section, the regrets for PSRL and DORL can be bounded in the same way. Let $\nu_k(s, a)$ be the number of visits on s, a in episode k and $\boldsymbol{\nu}_k$ be the row vector of $\nu_k(\cdot, \pi_k(\cdot))$. Let $\Delta_k = \sum_{s,a} \nu_k(s, a)(\lambda(M_k, \tilde{\pi}_k) - R(s, a))$. Using optimal equation,

$$
\begin{aligned}
\Delta_k &= \sum_{s,a} \nu_k(s, a)\left[\lambda(M_k, \tilde{\pi}_k) - \tilde{R}^k(s, a)\right] \\
&\quad + \sum_{s,a} \nu_k(s, a)\left[\tilde{R}^k(s, a) - R(s, a)\right] \\
&= \boldsymbol{\nu}_k(\tilde{P}^k - I)\boldsymbol{h}_k + \boldsymbol{\nu}_k(\tilde{\boldsymbol{R}}^k - \boldsymbol{R}^k) \\
&= \underbrace{\boldsymbol{\nu}_k(P^k - I)\boldsymbol{h}_k}_{①} + \underbrace{\boldsymbol{\nu}_k(\tilde{P}^k - P^k)\boldsymbol{h}_k}_{②} + \underbrace{\boldsymbol{\nu}_k(\tilde{\boldsymbol{R}}^k - \boldsymbol{R}^k)}_{③},
\end{aligned}
$$

14

where $\tilde{P}^k := P(M_k, \tilde{\pi}_k), P^k := P(M, \pi_k), \boldsymbol{h}_k := \boldsymbol{h}^*(M_k)$, and $\tilde{\boldsymbol{R}}^k := \boldsymbol{R}(M_k, \tilde{\pi}_k), \boldsymbol{R}^k := \boldsymbol{R}(M, \pi_k)$.

Using Azuma-Hoeffding inequality and the same analysis in Jaksch et al. (2010), we bound ① with probability at least $1 - \frac{\rho}{12}$,

$$\sum_k ① = \sum_k \boldsymbol{\nu}_k \left(P^k - I \right) \boldsymbol{h}_k \leq D\sqrt{\frac{5}{2}T \log\left(\frac{8}{\rho}\right)} + KD. \tag{7}$$

To bound ② and ③, we analyze the deviation in transition and reward function between M and M_k. For DORL, the deviation in transition probability is upper bounded by

$$\max_{s'} |\tilde{P}_i^k(x, s') - \hat{P}_i^k(x)|_1$$

$$\leq \min\{2\sum_{s \in \mathcal{S}_i} E_{P_i}^k(s \mid x), 1\}$$

$$\leq \min\{2\bar{E}_{P_i}^k(x), 1\} \leq 2\bar{E}_{P_i}^k(x),$$

The deviation in reward function $|\tilde{R}_i^k - \hat{R}_i^k|(x) \leq E_{R_i}^k(x)$.

For PSRL, since $M_k \in \mathcal{M}_k$, $|\tilde{P}_i^k - \hat{P}_i^k|(x) \leq \bar{E}_{P_i}^k(x)$ and $|\tilde{R}_i^k - \hat{R}_i^k|(x) \leq E_{R_i}^k(x)$.

Decomposing the bound for each scope provided by $M \in \mathcal{M}_k$ and M_k for PSRL $\in \mathcal{M}_k$, it holds for both PSRL and DORL that:

$$\sum_k ② \leq 3\sum_k D \sum_{i=1}^{m} \sum_{x \in \mathcal{X}[Z_i^P]} \nu_k(x) \bar{E}_{P_i}^k(x), \tag{8}$$

$$\sum_k ③ \leq 2\sum_k \sum_{i=1}^{l} \sum_{x \in \mathcal{X}[Z_i^R]} \nu_k(x) E_{R_i}^k(x); \tag{9}$$

where with some abuse of notations, define $\nu_k(x) = \sum_{x' \in \mathcal{X}: x'[Z_i] = x} \nu_k(x')$ for $x \in \mathcal{X}[Z_i]$. The second inequality is from the fact that $|\tilde{P}^k(\cdot|x) - P^k(\cdot|x)|_1 \leq \sum_1^m |\tilde{P}_i^k(\cdot|x[Z_i^R]) - P_i^k(\cdot|x[Z_i^R])|_1$ (Osband and Van Roy, 2014).

A.3 Bound (7), (8) and (9) by balancing episode length and episode number

We give a general criterion for bounding (7), (8) and (9), which we believe, is a new technique. We first introduce Lemma 4 which implies that bounding (7), (8) and (9) is to balance total number of episodes and the length of the longest episode. The proof, relies on defining the last episode k_0, such that $N_{k_0}(x) \leq \nu_{k_0}(x)$.

Lemma 4. *For any fixed episodes $\{T_k\}_{k=1}^K$, if there exists an upper bound $\bar{T}$, such that $T_k \leq \bar{T}$ for all $k \in [K]$, we have the bound*

$$\sum_{x \in \mathcal{X}[Z]} \sum_k \nu_k(x) / \sqrt{\max\{1, N_k(x)\}} \leq L\bar{T} + \sqrt{LT},$$

where Z is any scope with $|\mathcal{X}[Z]| \leq L$, and $\nu_k(x)$ and $N_k(x)$ are the number of visits to x in and before episode k. Furthermore, total regret of (7), (8) and (9) can be bounded by $\tilde{O}\big((\sqrt{W}Dm + l)(L\bar{T} + \sqrt{LT}) + KD\big)$

Proof. We bound the random variable $\sum_{k=1}^K \frac{\nu_k(x)}{\sqrt{\max\{N_k(x), 1\}}}$ for every $x \in \mathcal{X}[Z]$, where $\nu_k(x) = \sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}(x_t = x)$ and $N_k(x) = \sum_{i=1}^{k-1} \nu_i(x)$.

Let $k_0(x)$ be the largest k such that $N_k(x) \leq \nu_k(x)$. Thus $\forall k \geq k_0(x), N_k(x) > \nu_k(x)$, which gives $N_t(x) := N_k(x) + \sum_{\tau=t_k}^t \mathbb{1}(x_\tau = x) < 2N_k(x)$ for $t_k \leq t < t_{k+1}$.

15

Conditioning on $k_0(x)$, we have

$$\sum_{k=1}^{K} \frac{\nu_k(x)}{\sqrt{\max\{N_k(x),1\}}}$$

$$\leq N_{k_0(x)}(x) + \nu_{k_0(x)}(x) + \sum_{k>k_0(x)} \frac{\nu_k(x)}{\sqrt{\max\{N_k(x),1\}}}$$

$$\leq 2\nu_{k_0(x)}(x) + \sum_{k>k_0(x)} \frac{\nu_k(x)}{\sqrt{\max\{N_k(x),1\}}}$$

$$\leq 2\bar{T} + \sum_{k>k_0(x)} \frac{\nu_k(x)}{\sqrt{\max\{N_k(x),1\}}},$$

where the first inequality uses $\max\{N_k(x),1\} \geq 1$ for $k = 1, \ldots k_0(x)$, the second inequality is by the fact that $N_{k_0(x)}(x) \leq \nu_{k_0(x)}(x)$ and the third one is by $\nu_{k_0}(x) \leq T_{k_0(x)} \leq T_K$.

And letting $k_1(x) = k_0(x) + 1$ and $N(x) := N_K(x) + \nu_K(x)$, we have

$$\sum_{k>k_0(x)} \frac{\nu_k(x)}{\sqrt{\max\{N_k(x),1\}}}$$

$$\leq \sum_{t=t_{k_1(x)}}^{T} 2\frac{\mathbb{1}(x_t = x)}{\sqrt{\max\{N_t(x),1\}}}$$

$$\leq \sum_{t=t_{k_1(x)}}^{T} 2\frac{\mathbb{1}(x_t = x)}{\sqrt{\max\{N_t(x) - N_{k_1(x)},1\}}}$$

$$\leq 2\int_{1}^{N(x)-N_{k_1(x)}} \frac{1}{\sqrt{x}} dx$$

$$\leq (2 + \sqrt{2})\sqrt{N(x)}.$$

Given any $k_0(x)$, we can bound the term with a fixed value $2\bar{T} + (2+\sqrt{2})\sqrt{N(x)}$. Thus, the random variable $\sum_{k=1}^{K} \frac{\nu_k(x)}{\sqrt{\max\{N_k(x),1\}}}$ is upper bounded by $2\bar{T} + (2 + \sqrt{2})\sqrt{N(x)}$ almost surely. Finally, $\sum_x \sum_{k=1}^{K} \frac{\nu_k(x)}{\sqrt{\max\{N_k(x),1\}}} \leq L\bar{T} + (2 + \sqrt{2})\sqrt{LT}$. The regret by (8) is

$$\sum_k 3D \sum_{i\in[m]} \sum_{x\in\mathcal{X}[Z_i^P]} \nu_k(x)\bar{W}_{P_i}^k(x)$$

$$= \tilde{O}(\sqrt{W}Dm(L\bar{T} + \sqrt{LT}) + KD).$$

The regret by (9) is

$$\sum_k 2 \sum_{i\in[l]} \sum_{x\in\mathcal{X}[Z_i^R]} \nu_k(x)\bar{W}_{R_i}^k(x) = \tilde{O}(l(L\bar{T} + \sqrt{LT}) + KD).$$

The last statement is completed by directly summing (7), (8) and (9). $\qquad\square$

Instead of using the doubling trick that was used in Jaksch et al. (2010). We use an arithmetic progression: $T_k = \lceil k/L \rceil$ for $k \geq 1$. As in our algorithm, $T \geq \sum_{k=1}^{K-1} T_k \geq \sum_{k=1}^{K-1} k/L = \frac{(K-1)K}{2L}$, we have $K \leq \sqrt{3LT}$ and $T_k \leq T_K \leq K/L \leq \sqrt{3T/L}$ for all $k \in [K]$. Thus, by Lemma 4, putting (6), (7), (9), (8) together, the total regret for $M \in \mathcal{M}_k$ is upper bounded by

$$\tilde{O}\big((\sqrt{W}Dm + l)\sqrt{LT}\big), \tag{10}$$

with a probability at least $1 - \frac{\rho}{6}$.

A.4 Failure of the confidence set

For the failure of confidence set, we prove the following Lemma.

Lemma 5. *For all $k \in [K]$, with probability greater than $1 - \frac{3\rho}{8}$, $M \in \mathcal{M}_k$ holds.*

Proof. We first deal with the probabilities, with which in each round a reward function of the true MDP M is not in the confidence set. Using Hoeffding's inequality, we have for any t, i and $x \in \mathcal{X}[Z_i^R]$,

$$\mathbb{P}\left\{|\hat{R}_i^t(x) - R_i(x)| \geq \sqrt{\frac{12\log(6l|\mathcal{X}[Z_i^R]|t/\rho)}{\max\{1, N_{R_i}^t(x)\}}}\right\}$$

$$\leq \frac{\rho}{3l|\mathcal{X}[Z_i^R]|t^6}, \text{ with a summation} \leq \frac{3}{12}\rho.$$

Thus, with probability at least $1 - \frac{3\rho}{12}$, the true reward function is in the confidence set for every $t \leq T$.

For the transition probability, we use a different concentration inequality.

Lemma 6 (Multiplicative Chernoff Bound (Kleinberg et al., 2008) Lemma 4.9)**.** *Consider n, i.i.d random variables $X_1, \ldots, X_n$ on $[0, 1]$. Let μ be their mean and let X be their average. Then with probability $1 - \rho$,*

$$|X - \mu| \leq \sqrt{\frac{3\log(2/\rho)X}{n}} + \frac{3\log(2/\rho)}{n}.$$

Using Lemma 6, for each x, i, k, it holds that with probability $1 - \rho/(6m\left|\mathcal{X}\left[Z_i^P\right]\right|t_k^6)$,

$$|\hat{P}_i(\cdot|x) - P_i(\cdot|x)|_1 \leq \sqrt{\frac{18S_i\log(c_{i,k})}{\max\{N_{P_i}^k(x), 1\}}} + \frac{18\log(c_{i,k})}{\max\{N_{P_i}^k(x), 1\}}.$$

Then with a probability $1 - \frac{3\rho}{24}$, it holds for all x, i, k. Therefore, with a probability $1 - \frac{3\rho}{8}$, the true MDP is in the confidence set for each k. $\square$

Combined with (10), with probability at least $1 - \frac{2\rho}{3}$ the regret bound in Theorem 2 holds.

For PSRL, M_k and M has the same posterior distribution. The expectation of the regret caused by $M \notin \mathcal{M}_k$ and $M_k \notin \mathcal{M}_k$ are the same. Choosing sufficiently small $\rho \leq \sqrt{1/T}$, Theorem 1 follows.

B Proof of the lower bound

Our lower bound construction is a Cartesian product of n independent MDPs. We start by discussing the bias vector of such FMDP in Lemma 7.

Lemma 7. *Let M^+ be the Cartesian product of n independent MDPs $\{M_i\}_{i=1}^n$, each with a span of bias vector $sp(h_i)$. The optimal policy for M^+ has a span $sp(h^+) = \sum_i sp(h_i)$.*

Proof. Let λ_i^* for $i \in [n]$ be the optimal gain of each MDP. Optimal gain of M^+ is direct $\lambda^* = \sum_{i \in [n]} \lambda_i^*$. As noted in Puterman (2014) (8.2.3), by the definition of bias vector we have

$$h_i(s) = \mathbb{E}[\sum_{t=1}^{\infty}(r_t^i - \lambda_i^*) \mid s_1^i = s], \quad \forall s \in \mathcal{S}_i,$$

where r_t^i is the reward of the i-th MDP at time t and $s_t^i := s_t[i]$.

The lemma is directly by

$$h^+(s) = \mathbb{E}[\sum_{t=1}^{\infty}(r_t - \lambda^*) \mid s_1 = s]$$

$$= \mathbb{E}[\sum_{t=1}^{\infty}(\sum_{i \in [n]}(r_t^i - \lambda_i^*)) \mid s_1 = s]$$

$$= \sum_{i \in [n]} \mathbb{E}[\sum_{t=1}^{\infty}(r_t^i - \lambda_i^*) \mid s_1^i = s[i]]$$

$$= \sum_{i \in [n]} h_i(s[i]).$$

We immediately have $sp(h^+) = \sum_i sp(h_i)$. $\qquad\square$

Recall Theorem 3 states *for any algorithm, any graph structure satisfying $\mathcal{G} = (\{\mathcal{S}_i\}_{i=1}^n ; \{\mathcal{S}_i \times \mathcal{A}_i\}_{i=1}^n ; \{Z_i^R\}_{i=1}^n ; \{Z_i^P\}_{i=1}^n)$ with $|\mathcal{S}_i| \leq W$, $|\mathcal{X}[Z_i^R]| \leq L$, $|\mathcal{X}[Z_i^P]| \leq L$ and $i \in Z_i^P$ for $i \in [n]$, there exists an FMDP with an optimal bias vector h^+, such that for any initial state $s \in \mathcal{S}$, the expected regret of the algorithm after T step is*

$$\Omega(\sqrt{sp(\boldsymbol{h}^+)LT}). \tag{11}$$

Proof. Let $l = |\cup_i^n Z_i^R|$. As $i \in Z_i^P$, a special case is the FMDP with graph structure $\mathcal{G} = \left(\{\mathcal{S}_i\}_{i=1}^n ; \{\mathcal{S}_i \times \mathcal{A}_i\}_{i=1}^n ; \{\{i\}\}_{i=1}^l \text{ and } \{\emptyset\}_{i=l+1}^n ; \{\{i\}\}_{i=1}^n\right)$, which can be decomposed into n independent MDPs as in the previous example. Among the n MDPs, the last $n - l$ MDPs are trivial. By simply setting the rest l MDPs to be the construction used by Jaksch et al. (2010), which we refer to as "JAO MDP", the regret for each MDP with the span $sp(\boldsymbol{h})$, is $\Omega(\sqrt{sp(\boldsymbol{h})WT})$ for $i \in [l]$. The total regret is $\Omega(l\sqrt{sp(\boldsymbol{h})WT})$.

Using Lemma 7, $sp(\boldsymbol{h}^+) = l\, sp(\boldsymbol{h})$ and the total expected regret is $\Omega(\sqrt{l\, sp(\boldsymbol{h}^+)WT})$. Normalizing the reward function to be in $[0,1]$, the expected regret of the FMDP is $\Omega(\sqrt{sp(\boldsymbol{h}^+)WT})$, which completes the proof. $\qquad\square$

C   Proof of Theorem 4

The only difference between the proof of Theorem 4 and 2 lies in the bound of term ②.

Proof. Starting from ②, for each $s \in \mathcal{S}$, we bound $(\tilde{P}^k(\cdot \mid s) - P^k(\cdot \mid s))\boldsymbol{h}_k$. For simplicity, we remove the subscriptions of s and use $\tilde{P}^k$ and P^k to denote the vector for s-th row of the two matrix.

$$\sum_{s \in \mathcal{S}} (\tilde{P}^k(s) - P^k(s))h_k(s)$$

$$= \sum_{s_1 \in \mathcal{S}_1} \sum_{s_{-1} \in \mathcal{S}^{-1}} (P_1(s_1)P_{-1}(s_{-1}) - \tilde{P}_1(s_1)\tilde{P}_{-1}(s_{-1}))h_k(s_1, s_{-1})$$

$$= \sum_{s_1} \left[(P_1(s_1) - \tilde{P}_1(s_1)) \sum_{s_{-1}} \tilde{P}_{-1}(s_{-1})h_k(s_1, s_{-1}) \right] +$$

$$\sum_{s_{-1}} \left[(P_{-1}(s_{-1}) - \tilde{P}_{-1}(s_{-1})) \sum_{s} P_1(s_1)h_k(s_1, s_{-1}) \right]$$

$$= \sum_{s_1} (P_1(s_1) - \tilde{P}_1(s_1))h_{1k}(s_1) + \sum_{s_{-1}} (P_{-1}(s_{-1}) - \tilde{P}_{-1}(s_{-1}))h_{-1k}(s_{-1}),$$

where $h_{1k}(s_1) := \sum_{s_{-1}} \tilde{P}_{-1}(s_{-1})h_k(s_1, s_{-1})$ and $h_{-1k}(s_{-1}) := \sum_{s_1} P_1(s_1)h_k(s_1, s_{-1})$. As $span(h_{1k}) \leq sp_1(M_k)$,

$$\sum_{s \in \mathcal{S}} (\tilde{P}^k(s) - P^k(s))h_k(s) \leq |P_1 - \tilde{P}_1|_1 sp_1(M_k) + \sum_{s_{-1}} (P_{-1}(s_{-1}) - \tilde{P}_{-1}(s_{-1}))h_{-1k}(s_{-1}). \quad (12)$$

By applying (12) recurrently, we have

$$\sum_{s \in \mathcal{S}} (\tilde{P}^k(s) - P^k(s))h_k(s) \leq \sum_{i=1}^{m} |P_i - \tilde{P}_i|_1 sp_i(M_k).$$

Note that $sp_i(M_k)$ is generally smaller than $span(h_k)$. In our lower bound case each $sp_i = \frac{1}{m} span(h_k)$, which improves our upper bound by a scale of $1/m$.

The reduction of l can be achieved by bounding each factored reward to be in $[1, 1/l]$. The following proof remains the same. $\quad\square$

D FSRL algorithm

Here we provide a complete description of the FSRL algorithm that was omitted in the main paper due to space considerations.

Algorithm 2 FSRL

Input: $\mathcal{S}, \mathcal{A}, T$, encoding $\mathcal{G}$ and upper bound on sum of factored span Q.
$k \leftarrow 1; t \leftarrow 1; t_k \leftarrow 1; T_k = 1; \mathcal{H} \leftarrow \{\}$
repeat
    Choose $M_k \in \mathcal{M}_k$ by solving the following optimization over $M \in \mathcal{M}_k$,

$$\max \lambda^*(M) \quad \text{subject to} \quad Q(h) \leq Q \text{ for } h \text{ being the bias vector of } M.$$

    Compute $\tilde{\pi}_k = \pi(M_k)$.
    for $t = t_k$ **to** $t_k + T_k - 1$ **do**
        Apply action $a_t = \pi_k(s_t)$
        Observe new state s_{t+1}
        Observe new rewards $r_{t+1} = (r_{t+1,1}, \ldots r_{t+1,l})$
        $\mathcal{H} = \mathcal{H} \cup \{(s_t, a_t, r_{t+1}, s_{t+1})\}$
        $t \leftarrow t + 1$
    end for
    $k \leftarrow k + 1.$
    $T_k \leftarrow \lceil k/L \rceil; t_k \leftarrow t + 1.$
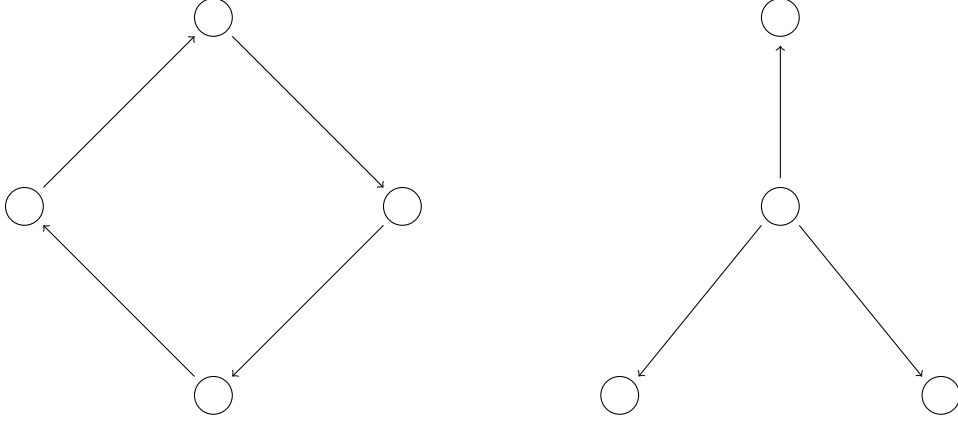**until** $t_k > T$

Figure 4: Circle and three-leg structure with a size 4. State space is a 4-dimensional vector with each dimension as $\{0, 1\}$ representing whether the computer is working or not. Arrows represent the scopes of dimension. Each node has an arrow to itself, which we ignored in the figure.

E Experiment Setups

Circle and Three-leg structures.   Our computer network administrator domain with a circle and a three-leg structure Guestrin et al. (2001); Schuurmans and Patrascu (2002) are shown in Figure 4. Each computer gives a 1 reward when it is work and a 0 reward otherwise. The factored transition matrix for network with size m is

$$P(s[i] = 0 \mid s[i] = 1, s) = \min\{1, \alpha_1|\epsilon_i^1| + \sum_{j \in Z_i^P} \alpha_2|\eta_{ij}^1| \mathbb{1}(s[j] = 0)\}, \ \forall i \in [m],$$

$$P(s[i] = 0 \mid s[i] = 0, s) = \min\{\max\{|\epsilon_i^0|, 0.5\} + \sum_{j \in Z_i^P} \alpha_2|\eta_{ij}^0| \mathbb{1}(s[j] = 0)\}, \ \forall i \in [m],$$

where $\alpha_1, \alpha_2 = 0.1$ are constant and $\epsilon_i^1, \epsilon_i^0, \eta_{ij}^1, \eta_{ij}^0$ are all white noise. To avoid the extreme cases in our lower bound, both the MDPs are set to have limited diameters.