Horizon-Free Learning for Markov Decision Processes and Games: Stochastically Bounded Rewards and Improved Bounds

Shengshi Li [1] **Lin F. Yang** [2]

Abstract

Horizon dependence is an important difference between reinforcement learning and other machine learning paradigms. Yet, existing results tackling the (exact) horizon dependence either assume that the reward is bounded per step, introducing unfair comparison, or assume strict total boundedness that requires the sum of rewards to be bounded *almost surely* – allowing only restricted noise on the reward observation. This paper addresses these limitations by introducing a new relaxation – *expected boundedness* on rewards, where we allow the reward to be stochastic with only boundedness on the *expected* sum – opening the door to study horizon-dependence with a much broader set of reward functions with noises. We establish a novel generic algorithm that achieves *no-horizon dependence* in terms of sample complexity for both Markov Decision Processes (MDP) and Games, via reduction to a good-conditioned *auxiliary Markovian environment*, in which only "important" state-action pairs are preserved. The algorithm takes only $\tilde{O}(\frac{S^2A}{\epsilon^2})$ episodes interacting with such an environment to achieve an ϵ-optimal policy/strategy (with high probability), improving (Zhang et al., 2022) (which only applies to MDPs with deterministic rewards). Here S is the number of states and A is the number of actions, and the bound is independent of the horizon H.

1. Introduction

One of the most prominent differences between reinforcement learning (RL) and other learning paradigms is its dependence on the decision horizon. For instance, one evaluates a policy based on its long-term performance, which sums up a sequence of rewards received after each decision. In stark contrast, bandit learning problems evaluate a policy based on its single-shot performance. However, does the horizon-dependence makes RL considerably more difficult? Jiang & Agarwal (2018) ask this question formally and proposes to study the problem under the so-called "total boundedness" assumption, where the rewards have a bounded sum almost surely for any trajectory collected by a policy – given a relatively fair comparison between, e.g., bandit problems and RL. Recently, a line of research (Wang et al., 2020; Zhang et al., 2021; Li et al., 2022; Zhang et al., 2022) settles this question by showing the existence of algorithms, which only take $\tilde{O}(1)$[1] trajectories to learn a good policy – eliminating the dependence on the horizon in the learning sample complexity under the *total boundedness* assumption with *deterministic* rewards.

While being profound, the above works leave a slackness in the understanding of the horizon effect – the total boundedness and deterministicity of the rewards can be infeasible in systems with noise, which, however, are the standard assumptions of multi-arm bandit systems. Moreover, it is also unclear whether horizon-free learning can be achieved in multi-agent systems, e.g., two-player zero-sum games. In particular, Zhang et al. (2022) rely on the almost-sure-boundedness of reward sums to establish a high probability bound for regret. Such an approach fails when there is stochastic noise; on the other hand, Li et al. (2022) require an ϵ-net on the reward space, requiring which to possess special properties provided by the total boundedness. Both works only apply to single player MDP and do not extend to Markov games, where the multi-player nature makes the problem more challenging.

In this paper, we address the limitations introduced by the total boundedness. In particular, we study the RL problem under a relaxed *expected boundedness* assumption, which only imposes boundedness on the expectation of the sum of rewards – a standard assumption that only requires the value function to be bounded. Our proposed algorithm consists of two phases, where in the first phase, it applies a reward-free

[1]School of Mathematics and Sciences, Peking University [2]Department of Electrical and Computer Engineering, University of California, Los Angeles. Correspondence to: Shengshi Li <shengshi_li@pku.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]We ignore the $|S||A|$-dependence, where S is the number of states and A is the number of actions.

key-state preserving exploration on the environment that covers the important states; in the second phase, it optimizes a policy/strategy for a given arbitrary reward function via a carefully designed upper confidence bound (UCB) exploration algorithm. In particular, our first phase can take a reward-free exploration algorithm (e.g., we can leverage the techniques in (Zhang et al., 2022)) to output an auxiliary Markov environment. This new environment allows easy exploration that does not depend on the horizon length. The second phase explores the auxiliary environment and takes advantage of the expected boundedness via a novel analysis on the model-based UCB algorithm to bound the variation in the empirical value function of a policy – which provides a tight concentration bound for the relaxed boundedness assumption.

As a by-product of the generality of the algorithm, our approaches extend to two-player zero-sum games – our algorithm outputs an approximated Nash equilibrium with the number of episodes of interactions independent of the horizon. We summarize our contributions as below.

1. We propose a new and more natural reward-boundedness assumption for studying the horizon-dependence problem in RL. Our assumption allows the study of a broader set of noisy rewards, extending the results in recent advances (Li et al., 2022; Zhang et al., 2022) and also complements the open problem proposed in (Jiang & Agarwal, 2018).

2. We propose a generic algorithmic framework that consists of reward-free exploration and reward-based UCB exploration. This algorithm achieves horizon-free learning for both MDPs and games.

3. At the core of our analysis is a technical innovation on the UCB-type analysis of model-based RL. This new technique enables a direct computation of expected regret bound over the entire collected dataset rather than relying on a martingale argument, which requires the rewards to be almost surely bounded.

4. Our work improves the existing horizon-independent PAC bounds in both the online setting and the generative setting for MDPs and two-player games with S states and A actions. See Table 1 below.

Table 1. Comparison to existing horizon-independent results

Paper	Online PAC	Generative PAC
(Li et al., 2022)	$\tilde{O}\left(\frac{\text{poly}(S,A)^{O(S)}}{\epsilon^5}\right)$	$O\left(\frac{S^6A^4}{\epsilon^3}\right)$
(Zhang et al., 2022)	$\tilde{O}\left(\frac{S^9A^3}{\epsilon^2}\right)$	-
Our work	$\tilde{O}\left(\frac{S^2A}{\epsilon^2}+\frac{S^9A^3}{\epsilon}\right)$	$\tilde{O}\left(\frac{S^2A}{\epsilon^2}\right)$

1.1. Related Work

Tabular RL. There is a long line of research on the sample complexity and regret bound for RL in the tabular setting. See e.g.,(Kearns & Singh, 2002; Brafman & Tennenholtz, 2003; Kakade, 2003; Strehl et al., 2006; Strehl & Littman, 2008; Kolter & Ng, 2009; Bartlett & Tewari, 2009; Jaksch et al., 2010; Szita & Szepesvári, 2010; Lattimore & Hutter, 2012; Osband et al., 2013; Dann & Brunskill, 2015; Azar et al., 2017; Dann et al., 2017; Osband & Van Roy, 2017; Jin et al., 2018; Fruit et al., 2018; Talebi & Maillard, 2018; Dann et al., 2019; Dong et al., 2019; Simchowitz & Jamieson, 2019; Russo, 2019; Zhang & Ji, 2019; Zhang et al., 2020c; Yang et al., 2021; Pacchiano et al., 2020; Neu & Pike-Burke, 2020; Wang et al., 2020; Zhang et al., 2020b; Menard et al., 2021; Zhang et al., 2021; Ren et al., 2021) and references therein. Most of the prior works used the *Reward Uniformity* assumption, in which the reward values satisfy $r_h \in [0, 1/H]$ for all h, up to a scaling factor.

Dependence on Horizon. Jiang & Agarwal (2018) point out that to have a fair comparison between long horizon and short horizon problems, one should only impose an upper bound on the summation of the reward values, i.e., $\sum_{h=1}^{H} r_h \leq 1$. We refer as the *Total Boundedness* assumption. Under this assumption, they conjectured that there would be a poly(H) regret lower bound. This conjecture was first partially refuted by (Zanette & Brunskill, 2019), who gave an algorithm whose regret scales logarithmically with H in the regime $K = \text{poly}(S, A, H)$. Later this conjecture was substantially refuted by Wang et al. (2020), in which they provide an algorithm that requires only $\text{poly}(S, A, \log H, 1/\epsilon)$ episodes to learn a ϵ-optimal policy. Surprisingly, Li et al. (2022) settled this question by giving a horizon-independent algorithm, but with exponential dependence on S and A. This exponential dependence was further improved to S^9A^3 in Zhang et al. (2022).

Two-player zero-sum Markov Game. Markov games have been widely studied since the seminal work (Shapley, 1953). Early works (Littman, 1994; Hu & Wellman, 2003; Hansen et al., 2013) focused on the setting where the transition matrix and reward function are assumed to be known or in the asymptotic setting where the number of data goes to infinity. When the transition kernel is unknown, a line of works (Sidford et al., 2020; Cui & Yang, 2021; Zhang et al., 2020a; Jia et al., 2019) considers the generative setting, making strong reachability assumption under which no sophisticated exploration algorithm is required. Another line of works (Bai et al., 2020; Xie et al., 2020; Bai & Jin, 2020; Liu et al., 2021) look for the non-asymptotic guarantees without reachability assumptions. Our work is the first to consider the horizon-dependence problem proposed by (Jiang & Agarwal, 2018) for Markov Games.

2. Preliminaries

Notations. Throughout our paper, we use $[N]$ to denote the set $\{1, 2, \ldots, N\}$ for $N \in \mathbb{Z}^+$. We use $\mathbf{1}_s$ to denote the one-hot vector whose only non-zero element is in the s-th coordinate. For an event A, we use $\mathbf{1}_A$ as the indicator function. For a space $\mathcal{S}$, $\Delta(\mathcal{S})$ stands for all the probability distribution supported on $\mathcal{S}$. For two n-dimensional vectors x and y, we use a covariance-like function $V(x, y) = \sum_i x_i y_i^2 - (\sum_i x_i y_i)^2$ to prove our concentration results. We denote $\iota = \log(2/\delta)$ as the logarithm of confidence parameter δ.

Markov Decision Process(MDP). In this paper we consider the finite-horizon time-homogeneous MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, H, s_0)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the unknown (but fixed) transition operator which takes a state-action pair and returns a distribution over states, $R : \mathcal{S} \times \mathcal{A} \to \Delta([-1, 1])$ is the reward function, $H \in \mathbb{Z}^+$ is the planning horizon, and s_0 is the initial state.[2] A solution to the MDP is a policy π, which chooses an action a at a state $s \in \mathcal{S}$ and time step $h \in [H]$, i.e., $\pi := \{\pi_h\}_{h=1}^H$ where for each $h \in [H]$, $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ maps a given state to a distribution on the action space. Each *roll-out* of the policy π generates a random trajectory $s_1, a_1, r_1, s_2, \ldots, s_H, a_H, r_H, s_{H+1}$ where $s_1 = s_0, a_1 \sim \pi_1(s_1), r_1 \sim R(s_1, a_1), s_2 \sim P(s_1, a_1), \ldots, a_H \sim \pi_H(s_H), r_H \sim R(s_H, a_H), s_{H+1} \sim P(s_H, a_H)$. The state value function and state-action value function for the policy are then defined as

$$V_h^\pi(s_h) := \mathbb{E}\left[\sum_{t=h}^H r(s_t, a_t)|\pi, s_h\right],$$
$$Q_h^\pi(s_h, a_h) := \mathbb{E}\left[\sum_{t=h}^H r(s_t, a_t)|\pi, s_h, a_h\right].$$

Our goal is to find an optimal policy π^* that maximizes the value, i.e. $\max_\pi \mathbb{E}\left[\sum_{h=1}^H r(s_h, a_h)\right]$ by only interacting with the environment. We use Q_h^* and V_h^* to denote the value function of π^*, respectively. We call a policy ϵ-optimal if $V_1^*(s_1) - V_1^\pi(s_1) \le \epsilon$.

Markov Game(MG). We further consider the two-player finite-horizon time-homogeneous Markov Game $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, R, H, s_0)$. Similar to the MDP setting, $\mathcal{S}$ is the finite state space, $\mathcal{A}$ and $\mathcal{B}$ are the finite action space for the two players respectively, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ is the unknown transition operator which takes a state-action pair and returns a distribution over states. $R : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta([-1, 1])$ is the reward function. $H \in Z^+$ is the planning horizon, and s_0 is the initial state. Unlike the MDP setting, the solution in MG is a strategy (or policy pair) $\pi = (\mu, \nu)$, where $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ stands for the policy of the max-player and $\nu : \mathcal{S} \to \Delta(\mathcal{B})$ stands for the policy of the min-player. For a given strategy, the corresponding state-value and action-value functions are defined as follows.

$$V_h^\pi(s_h) := \mathbb{E}\left[\sum_{t=h}^H r(s_t, a_t, b_t)|\pi, s_h\right],$$
$$Q_h^\pi(s_h, a_h, b_h) := \mathbb{E}\left[\sum_{t=h}^H r(s_t, a_t, b_t)|\pi, s_h, a_h, b_h\right].$$

The max-player aims to maximize the value function, while the min-player aims to minimize the value function. If the min-player's strategy ν is fixed, the MG degenerates to an MDP, and the optimal policy in this MDP is the best response strategy $\mathrm{br}_1(\nu)$. Similarly, we can define the best-response strategy, $\mathrm{br}_2(\mu)$, for the min-player. The subscript in br_1 and br_2 will be ignored in the clear context. For all $h \in [H], s_h \in \mathcal{S}$, we define

$$V_h^{*,\nu}(s_h) := V_h^{\mathrm{br}(\nu),\nu}(s_h) = \max_\mu V_h^{\mu,\nu}(s_h),$$
$$V_h^{\mu,*}(s_h) := V_h^{\mu,\mathrm{br}(\mu)}(s_h) = \min_\nu V_h^{\mu,\nu}(s_h).$$

There exists Nash equilibrium (NE) policy pair $\pi^* = (\mu^*, \nu^*)$ that μ^* and ν^* are the best responses to each other. We define $V_h^*(s_h) = V_h^{\mu^*,\nu^*}(s_h)$ for all $s_h \in \mathcal{S}, h \in [H]$. The following weak duality property holds for all policy pairs (μ, ν) in MG:

$$V_h^{\mu,*} \le V_h^* \le V_h^{*,\nu}, \forall h \in [H].$$

Our goal is to minimize the duality gap of a policy pair $\pi = (\mu, \nu)$, which is defined as

$$\mathrm{Gap}(\pi) = V_1^{*,\nu}(s_1) - V_1^{\mu,*}(s_1).$$

We call a policy pair ϵ-approximate NE if $\mathrm{Gap}(\pi) \le \epsilon$.

Regret and PAC Bound. The agent interacts with the environment for K episodes. It chooses a policy (pair) π^k at the k-th episode. The regret in MDP setting is defined as

$$\mathrm{Regret}(K) = \sum_{k=1}^K V_1^*\left(s_1^k\right) - V_1^{\pi^k}\left(s_1^k\right),$$

while the regret in MG setting is defined as

$$\mathrm{Regret}(K) = \sum_{k=1}^K V_1^{*,\nu^k}\left(s_1^k\right) - V_1^{\mu^k,*}\left(s_1^k\right).$$

The measurement we will use is PAC-RL sample complexity, which counts the total number of episodes to find an

[2]This is without loss of generality: for the case of an unknown initial distribution $\mu_0 \in \Delta(S)$ in the MDP, it can be reduced to a fixed dummy initial state s_0, whose transition is μ_0 for any action played (all with reward 0).

ϵ-optimal policy in MDP or an ϵ-approximate NE policy pair in MG. As our algorithm and refined analysis provide bound on the expectation of regret, we can either derive PAC-RL result for mixed policy (a policy that randomly chooses the history policy), or use the idea of standard reduction in (Jin et al., 2018). Here we use a new evaluation algorithm to choose a good policy. More details to follow in later sections.

Trajectory. Each time we run RFKSP(See Section 5.1), we construct a new aux MDP/MG from scratch. We use τ_0 to denote all the trajectories in RFKSP and τ_k to denote the trajectory in the k-th episode afterward. Moreover, we denote $\Gamma_k = (\tau_0, \tau_1, \dots \tau_k)$ as the trajectories before the $k+1$-th episode interacting with the aux MDP/MG. We use $N^k(s,a)$ to denote the visit count of (s,a) in Γ_{k-1} and $N^1(s,a)$ denote the visit count of (s,a) in RFKSP. Similarly we define $N^k(s,a,s')$ and set all $N^k(s,a)$ to be at least 1.

3. Reward Assumption

In this section we compare the classic reward assumptions and our new reward assumption. They are given in MDP and can be translated to MG by substituting $\mathcal{A}$ by $\mathcal{A} \times \mathcal{B}$.

Assumption 3.1 (Reward Uniformity). For all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $0 \leq r(s,a) \leq 1/H$. [3]

Assumption 3.2 (Total Boundedness). $0 \leq r(s,a) \leq 1$ and $\sum_{h=1}^{H} r(s_h, a_h) \leq 1$ holds almost surely for any trajectory induced by any policy.

Reward Uniformity is commonly used in the literature when the comparison between long and short horizon problems is not a concern. (Jiang & Agarwal, 2018) suggests considering Assumption 3.2, the total boundedness assumption, which is considerably weaker than the reward uniformity assumption and allows the study of sparse-reward setting. However, the *almost-sure-boundedness* is restrictive and induces some surprising (unwanted) properties in the MDP (e.g., any state-action pair with a reward $O(1)$ can be visited by at most once for any policy and trajectory), reducing the practicality of such an assumption. Moreover, it is hard to capture the noisy reward setting.

Next, we state the more natural boundedness assumption, which relies on the notion of h-reachable states - states that can be visited at the h-th step starting from the initial state with a policy. The assumption is formally defined below.

Assumption 3.3 (Expected Boundedness). For all $h \in [H]$, h-reachable state s_h and policy π, $|r(s_h, a)| \leq 1$ and

$$\mathbb{E}_\pi \left[\sum_{t=h}^{H} |r(s_h, a_h)| \right] \leq 1.$$

Remark 3.4. We broaden the range of r to $[-1, 1]$ to provide convenience for tackling Markov Game. The expectation is taken over the trajectory space following policy π.

Our new reward assumption is a strict relaxation to the total boundedness. It allows the total sum to exceed 1 and even scale up to H under non-zero probability, which makes it considerably more challenging to achieve horizon-independence. The wide range of our reward assumption makes it natural to incorporate observation noise. In what follows, we show examples that distinguish the expected boundedness from the total boundedness assumptions.

Example 1. In real-world, for a designed reward $R(s,a)$ satisfying total-boundedness, the collected reward $r(s,a)$ may follow $R(s,a) + \epsilon(s,a)$, where $\epsilon(s,a)$ is the observation noise with zero-expectation. We can assume the noise is small enough so $0 \leq r(s,a) \leq 1$ still holds. Such rewards violate the total-boundedness since the sum of the rewards can exceed 1 with a positive probability.

Example 2. Consider a game where the player tries to remain alive for H steps. Action a and b get him killed with probability $1/2$ and 1. This game can be formulated as a MDP with states s(initial state) and z(death state). $P(s|s,a) = P(z|s,a) = r(s,a) = 1/2, P(z|s,b) = P(z|z,\cdot) = 1$. Other rewards are zero. Sticking to action a, the sum of the rewards is greater than 1 with constant probability and can be up to $O(H)$, clearly beyond the scope of total-boundedness. Using total-bounded rewards $r(s,a) = 1/H$ for this problem leads in an extra H factor.

4. Technical Overview

At a high level, our algorithm takes two phases. The first phase is the initialization phase, which explores the environment in a reward-free fashion that attempts to reach every reachable state-action pair. We do not restrict the algorithm to be used in the phase, and the algorithm can be applied to both games and MDPs (as no reward is considered). In fact, we believe many of the existing reward-free exploration algorithms (e.g., (Jin et al., 2020) and stage 1 of the main algorithm in (Zhang et al., 2022)) can be adapted to this phase. We clearly define the requirement of the algorithm output of the first phase – an auxiliary MDP that filters both the state-action space and probability transition of the ground-truth. We will show that stage 1 of the main algorithm in (Zhang et al., 2022) indeed outputs such an MDP. This auxiliary MDP makes the further algorithmic design less challenging and is also more friendly to horizon-free analysis with expected boundedness.

[3] The literature usually bounds $r(s,a) \in [0,1]$; we scale it to $[0, 1/H]$ for convenience.

Our second phase is a model-based algorithm on the auxiliary MDP. The algorithm itself is similar to many existing works, including R-max in (Brafman & Tennenholtz, 2003), RMIS in (Zhang et al., 2022). In each step, the algorithm establishes an approximate model of the auxiliary MDP and a confidence set that contains the groud-truth. Then the algorithm takes an optimistic policy computed using the largest-in-value model from the confidence set. The confidence bound is carefully designed so that no H dependence would appear – if some confidence bound is large (i.e., of order H) it should be canceled by the samples collected in the first phase.

Our core innovation in the second phase appears in the analysis. In fact, the analysis in (Zhang et al., 2022) follows a standard approach that first decomposes the regret of each episode along the collected trajectory. The trajectory-based decomposition necessarily introduces a martingale difference (the difference between the collected rewards and the expected rewards) that adapts to the history. Yet, deriving concentration bound on this martingale requires almost sure boundedness of the sum of rewards. Our innovation lies in the establishment of a "total expectation" argument that carefully bounds the sum of regret per episode under the expectation of the *entire history*. Therefore, a martingale-type argument is no longer needed. Thus expected boundedness of the rewards is sufficient to bound the expected sum of regrets. To further apply pigeonhole-type arguments to bound the final regret, we apply a filtering argument that selects the trajectories when the probability matrix is well-approximated to obtain the final horizon-free bound.

One last remark of the total expectation argument is that it only produces good policy with constant probability. We boost the probability via a classic probability boosting approach – repeat the algorithm instance independently and pick the best outcome among the outputs.

Algorithm 1 MDP-Full

1: **Input:** MDP $\mathcal{M}$, ϵ, δ.
2: Set $\epsilon_{\text{ksp}}, \epsilon_{\text{ucb}}, \epsilon_{\text{eval}} = O(\epsilon), \delta_{\text{ksp}} = \frac{1}{4}, \delta_{\text{eval}} = \frac{\delta}{2T}$.
3: Run $T = \log(2/\delta)$ times independently.
4: **for** $t = 1, 2, \dots, T$ **do**
5: $\quad \tilde{\mathcal{M}}_t \leftarrow$RFKSP$(\mathcal{M}, \epsilon_{\text{ksp}}, \delta_{\text{ksp}})$.
6: $\quad \pi^t \leftarrow$ MDP-RBUCBI$(\tilde{\mathcal{M}}_t, \epsilon_{\text{ucb}})$.
7: $\quad \hat{V}_1^{\pi^t}(s_0) \leftarrow$MDP-Evaluation$(\mathcal{M}, \pi^t, \epsilon_{\text{eval}}, \delta_{\text{eval}})$.
8: **end for**
9: $i \leftarrow \arg\max_{t\in[T]} \hat{V}_1^{\pi^t}(s_0)$.
10: **Output:** π^i or $\hat{V}_1^{\pi^i}(s_0)$in MG-Full.

5. Algorithm

Overview We first illustrate our algorithms for MDP, which is formally present in Algorithm 1. Our algorithm proceeds with following high level steps:

1. build an aux MDP by a reward-free key state preserving algorithm. Such an aux MDP has sufficient initial samples to achieve horizon-free while its value function is close to the original MDP with high probability.
2. apply a UCB-type algorithm, which we term as MDP-RBUCBI, to explore on aux MDP and obtain a policy, which is near-optimal for the aux MDP with constant probability.
3. estimate the value function of the returned policy in the original MDP by an algorithm called MDP-Evaluation.
4. use the idea of boosting, independently repeat the previous steps for $O(\log(\frac{1}{\delta}))$ times. Return the policy with the highest estimated value function.

Here the aux MDP is simulated using the true environment. Once an action is taken in the aux MDP, the action is translated to the true MDP and the feedback from the true MDP is translated back to the feedback in the aux MDP. Hence any algorithm runs on the aux MDP is in fact interacting with the true MDP using the aux MDP as a proxy.

Below we demonstrate the details of our algorithm. The RFKSP algorithm we selected to use in this paper is given in Appendix C. MDP-Evaluation is given in Appendix E.2 since it resembles MDP-Full(Algorithm 1) except that it keeps running a given policy in the reward-based phase.

5.1. Reward-Free Key State Preserving

Auxiliary Markovian Environment. We denote

$$U(s,a) = \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \mathbf{1}_{(s_h,a_h)=(s,a)}\right]$$

to be the max expected visit counts to (s,a) in an episode. The regret induced by UCB-type algorithm can be bounded independently of H if there are sufficient initial samples for all state-action pairs, i.e. $N^1(s,a) \geq U(s,a)/\text{poly}(S,A), \forall(s,a)$. See Appendix B. While it is hard to collect sufficient samples for all state-action pairs since some of them can rarely be visited, we build the **auxiliary Markovian environment** (or **aux MDP/MG**) by redirecting all these rarely visited state-action pairs to an absorbing state z and setting the reward from z to be 0.

Definition 5.1. For a MDP $M = (\mathcal{S}, \mathcal{A}, P, R, H, s_0)$, suppose we can partition $\mathcal{S}\times\mathcal{A} = \mathcal{O}\cup\mathcal{O}^C$($\mathcal{O}$ stands for omitted), then an **auxiliary Markovian environment** is defined as $\tilde{\mathcal{M}} = (\mathcal{S}\cup\{z\}, \mathcal{A}, \tilde{P}, \tilde{R}, H, s_0)$, where

$$\begin{aligned}
&\tilde{P}_{s,a} = P_{s,a}, && \tilde{R}(s,a) = R(s,a), && \forall(s,a)\in\mathcal{O}^C,\\
&\tilde{P}_{s,a} = \mathbf{1}_z, && \tilde{R}(s,a) = 0, && \forall(s,a)\in\mathcal{O},\\
&\tilde{P}_{z,a} = \mathbf{1}_z, && \tilde{R}(z,a) = 0, && \forall a\in\mathcal{A}.
\end{aligned}$$

If the max visiting probability to $\mathcal{O}$ is small enough, $V_1^\pi(s_0)$ is close to $\tilde{V}_1^\pi(s_0)$ for any policy π. ($\tilde{V}$ is the value function in the aux MDP/MG). If we further have sufficient initial samples for the "important" state-action pairs in $\mathcal{O}^C$, we call such aux MDP/MG to be good conditioned.

Definition 5.2. We call an auxiliary Markovian environment to be ϵ-good conditioned if it satisfies that

1. $\max_\pi \mathbb{P}_\pi [\exists h \in [H], (s_h, a_h) \in \mathcal{O}] \le \epsilon$,

2. $N^1(s,a) \ge \frac{U(s,a)}{\text{poly}(S,A)}, \forall (s,a) \in \mathcal{O}^C$.

We formally defined the RFKSP algorithm as below.

Definition 5.3. A **reward-free key-state preserving** algorithm satisfies that for any given $\epsilon, \delta > 0$, after using $K_1 = \text{poly}\left(S, A, \iota, \frac{1}{\epsilon}\right)$ episodes, the auxiliary Markovian environment it returned is ϵ-good conditioned with probability at least $1-\delta$.

Remark 5.4. By utilizing the partition of state-action space in the stage 1 of the main algorithm in (Zhang et al., 2022) to build the aux MDP, this algorithm serves as a viable reward-free key-state preserving algorithm with $K_1 = \tilde{O}\left(S^9 A^3/\epsilon\right)$. This algorithm is given in Appendix C and is used in our current result. The core idea is that it maintains an omitted set $\mathcal{O} \subset \mathcal{S} \times \mathcal{A}$. Each episode is divided into two phases, where the algorithm plans optimistically to reach $\mathcal{O}$ in the first phase and collects samples in the second phase. The collected target is the first reached state-action pair in $\mathcal{O}$ when the estimated transition probability is accurate enough. Once enough samples for a state-action pair are collected, it is removed from $\mathcal{O}$ to $\mathcal{O}^C$. If not enough samples for some (s,a) are collected during the entire phase, it remains in the omitted set $\mathcal{O}$. The optimistic design of the algorithm makes sure that it searches the entire reachable set of states.

Remark 5.5. A generative model (first proposed by (Kearns & Singh, 1998) and has inspired a number of follow works see e.g.(Singh & Yee, 1994; Gheshlaghi Azar et al., 2013; Sidford et al., 2018a;b; Agarwal et al., 2020; Li et al., 2020; 2022)) serves as the most straightforward and powerful reward-free key-state preserving algorithm with $K_1 = SA$. In particular, by sampling one batch (H samples) for each state-action pair, we have sufficient initial samples for all state-action pairs since $H \ge U(s,a)$. Thus the original MDP can be transformed into a 0-good conditioned auxiliary Markovian environment with an empty omitted set.

5.2. Reward-Based UCB with Initialization

MDP-RBUCBI is formally presented in Algorithm 2. In each episode k and state s, we set $\overline{V}_{H+1}^k(s)$ as 0, and calculate $\overline{Q}_H^k, \overline{V}_H^k, \ldots, \overline{Q}_1^k, \overline{V}_1^k$ iteratively as follows. For $(s,a) \in \mathcal{O}$, $\overline{Q}_h^k(s,a) = 0$. For $(s,a) \in \mathcal{O}^C$,

$$\overline{Q}_h^k(s,a) = \min\left(\overline{r}^k(s,a) + \max_{\overline{p} \in \mathcal{P}_{s,a}^k} \overline{p}\,\overline{V}_{h+1}^k, 1\right). \tag{1}$$

Furthermore, $\pi_h^k(s) = \arg\max_a \overline{Q}_h^k(s,a)$ and $\overline{V}_h^k(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot|s)} \overline{Q}_h^k(s,a)$.

$\mathcal{P}_{s,a}^k$ is the confidence set of transition probability, and $\overline{r}^k$ is the overestimation of the reward.

Algorithm 2 MDP-RBUCBI (**R**eward-**B**ased **UCB** with **I**nitialization)

1: **Input:** $\tilde{\mathcal{M}}\left(\epsilon_{\text{ksp}}, \delta_{\text{ksp}} = \frac{1}{4}\right)$(aux MDP), ϵ_{ucb}.
2: **Initialization:** $\overline{V}_{H+1}^k(s) = 0, \forall k, s \in \mathcal{S}$.
3: Use $K = \tilde{O}\left(\frac{S^2 A}{\epsilon_{\text{ucb}}^2}\right)$ episodes.
4: **for** episode $k = 1, 2, \ldots, K$ **do**
5:   **for** step $h = H, H-1, H-2, \ldots, 1$ **do**
6:     Compute $\overline{Q}_h^k(s,a)$ as in equation 1.
7:     **for** $s \in \mathcal{S}$ **do**
8:       $\pi_h^k(s) \leftarrow \arg\max_a \overline{Q}_h^k(s,a)$.
9:       Compute $\overline{V}_h^k(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot|s)} \overline{Q}_h^k(s,a)$.
10:     **end for**
11:   **end for**
12:   Receive initial state $s_1^k = s_0$, play policy π^k, collect trajectory τ_k. Calculate $\mathcal{P}^{k+1}, \mathcal{R}^{k+1}$ based on Γ_k.
13: **end for**
14: **Output:** Randomly select one policy π^k.

Confidence Set. The straightforward estimated transition probability in the k-th episode is $\hat{P}_{s,a,s'}^k = \frac{N^k(s,a,s')}{N^k(s,a)}$. By Freedman's inequality, with probability $1 - S^2 AK\delta_{\text{conf}}$,

$$|P_{s,a,s'} - \hat{P}_{s,a,s'}^k| \le \sqrt{2\frac{P_{s,a,s'}\iota_{\text{conf}}}{N^k(s,a)}} + \frac{\iota_{\text{conf}}}{3N^k(s,a)}. \tag{2}$$

And thus $P_{s,a} \in \mathcal{P}_{s,a}^k$ holds for the confidence set $\mathcal{P}_{s,a}^k$ as

$$\left\{p \in \Delta(S) : \left|p'_s - \hat{P}_{s,a,s'}^k\right| \le 5\sqrt{\frac{\hat{P}_{s,a,s'}^k \iota_{\text{conf}}}{N^k(s,a)}} + \frac{5\iota_{\text{conf}}}{N^k(s,a)}\right\}.$$

Previous works, including (Zhang et al., 2022), constructed similar confidence sets for the transition probability. We further build the confidence set for the reward. We denote the collected rewards for (s,a) before the k-th episode as $r^i(s,a), i \in [N^k(s,a)]$. We build the confidence set $\mathcal{R}_{s,a}^k$ based on the sample mean $\hat{r}^k(s,a)$ as

$$\left\{r : \left|r - \hat{r}^k(s,a)\right| \le \sqrt{4\frac{\hat{V}^k \iota_{\text{conf}}}{N^k(s,a)}} + \frac{10\iota_{\text{conf}}}{N^k(s,a)}\right\},$$

where $\hat{V}^k$ is the sample variance. It can be shown that with probability at least $1-SAK\delta_{\text{conf}}$, $\mathbb{E}\left[R(s,a)\right] \in \mathcal{R}^k_{s,a}$ holds for any (s,a,k). We set the overestimation $\overline{r}^k(s,a)$ to be $\min\{\max_{r\in\mathcal{R}^k_{s,a}} r, 1\}$. and the underestimation $\underline{r}^k(s,a,b)$ used in MG to be $\max\{\min_{r\in\mathcal{R}^k_{s,a,b}} r, -1\}$.

5.3. Algorithm for MG

The whole algorithm MG-Full(Algorithm 8) is given in Appendix F since it resembles MDP-Full. While the gap for a policy π in MDP is $V_1^*(s_0) - V_1^\pi(s_0)$, the gap for a strategy (μ,ν) in MG is $V_1^{*,\nu}(s_0) - V_1^{\mu,*}(s_0)$. We need a new MG-RBUCBI algorithm, which is formally presented in Algorithm 3, to derive strategy with low gap. The new MG-Evaluation(Algorithm 9) is given in Appendix F.2.

Recall that in MDP, we construct an overestimation $\overline{V}_h$ for V_h^* by setting π_h as argmax $\overline{Q}_h$ recursively. The gap for this policy π is bounded by the difference of $\overline{V}_1(s_0)$ and $V_1^\pi(s_0)$, which are both expected sum over the trajectory space following π. Correspondingly, we want to derive a strategy (μ,ν) for time step $h = H, H-1, \ldots, 1$ recursively in MG that its gap can be bounded by the difference of the overestimation and the underestimation as below.

$$\overline{V}_h(s_h) \ge V_h^{*,\nu}(s_h) \ge V_h^{\mu,*}(s_h) \ge \underline{V}_h(s_h).$$

We leverage CCE in (Moulin & Vial, 1978; Aumann, 1987).

Coarse Correlated Equilibrium CCE is a subroutine that takes two metrics $P, Q \in \mathbb{R}^{A\times B}$ and returns $(\phi,\psi) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$ for general sum game, which satisfies

$$\phi^T P\psi \ge \max_{a\in\mathcal{A}} \mathbf{1}_a^T P\psi, \quad \phi^T Q\psi \le \min_{b\in\mathcal{B}} \phi^T Q\mathbf{1}_b.$$

When extending the algorithm from MDP to MG, the estimation function are modified as follows. For $(s,a,b) \in \mathcal{O}$, $\overline{Q}_h^k(s,a,b) = \underline{Q}_h^k(s,a,b) = 0$. For $(s,a,b) \in \mathcal{O}^C$,

$$\overline{Q}_h^k(s,a,b) = \min\left(\overline{r}^k(s,a,b) + \max_{\overline{p}\in\mathcal{P}^k_{s,a,b}} \overline{p}\,\overline{V}^k_{h+1}, 1\right), \quad (3)$$

$$\underline{Q}_h^k(s,a,b) = \max\left(\underline{r}^k(s,a,b) + \min_{\underline{p}\in\mathcal{P}^k_{s,a,b}} \underline{p}\,\underline{V}^k_{h+1}, -1\right).$$

MG-Evaluation (Algorithm 9) is implemented with MDP-Full, which can return the near-optimal value for the opponent when a player is fixed, to evaluate the gap of a given policy pair (μ,ν). We multiply rewards by -1 when estimating $V_1^{\mu,*}(s_0)$ since MDP-Full tries to maximize the value function while the min player aims to minimize it. The modified model still satisfies our reward assumption since we have broaden the range of r.

6. Theoretical Guarantee

In this section, we provide the theoretical guarantee for our algorithms. We use K_{RFKSP} to denote the episodes used

Algorithm 3 MG-RBUCBI

Input: aux MG $\tilde{\mathcal{G}}(\epsilon_{\text{ksp}}, \delta_{\text{ksp}} = \frac{1}{4})$, ϵ_{ucb}.
Initialization: $\overline{V}^k_{H+1}(s) = 0, \underline{V}^k_{H+1}(s) = 0, \forall k, s$.
Use $K = \tilde{O}\left(\frac{S^2AB}{\epsilon^2_{\text{ucb}}}\right)$ episodes.
for episode $k = 1, 2, \ldots, K$ **do**
  for step $h = H, H-1, H-2, \ldots, 1$ **do**
    Compute $\overline{Q}^k_h(s,a,b), \underline{Q}^k_h(s,a,b)$ as equation 3.
    for $s \in \mathcal{S}$ **do**
      $\mu^k_h(\cdot|s), \nu^k_h(\cdot|s) \leftarrow \text{CCE}\left(\overline{Q}^k_h(s,\cdot,\cdot), \underline{Q}^k_h(s,\cdot,\cdot)\right)$
      $\overline{V}^k_h(s) = \mathbb{E}_{a\sim\mu^k_h(\cdot|s), b\sim\nu^k_h(\cdot|s)} \overline{Q}^k_h(s,a,b)$
      $\underline{V}^k_h(s) = \mathbb{E}_{a\sim\mu^k_h(\cdot|s), b\sim\nu^k_h(\cdot|s)} \underline{Q}^k_h(s,a,b)$
    end for
  end for
  Play policy μ^k and ν^k, collect trajectory τ_k.
  Calculate $\mathcal{P}^{k+1}, \mathcal{R}^{k+1}$ based on Γ_k.
end for
Output: Randomly select one policy pair π^k.

by the selected RFKSP algorithm and K_{Reward} to denote the episodes used by the reward-based part. In the online setting, we modify the collecting initial samples stage in (Zhang et al., 2022) as RFKSP(Appendix C). In the generative setting, we use the algorithm in Remark 5.5 as RFKSP. We also outline the proof of Theorem 6.1 for demonstration.

Theorem 6.1. *For any $\epsilon, \delta > 0$, with probability $1-\delta$, MDP-Full(Algorithm 1) returns an ϵ-optimal policy by sampling at most $K_{\text{Reward}} + K_{\text{RFKSP}}$ episodes, where $K_{\text{Reward}} = \tilde{O}\left(S^2A/\epsilon^2\right), K_{\text{RFKSP}} = \tilde{O}\left(S^9A^3/\epsilon\right)$.*

Remark 6.2. Compared to the PAC bound $\tilde{O}\left(S^9A^3/\epsilon^2\right)$ in (Zhang et al., 2022), our bound is much better, and can be further reduced to $\tilde{O}\left(S^2A/\epsilon^2\right)$ when $\epsilon \le O(1/S^7A^2)$.

Similarly, we provide the results for MGs as below.

Theorem 6.3. *For any $\epsilon, \delta > 0$, with probability $1-\delta$, MG-Full(Appendix F) returns an ϵ-approximate NE policy pair by sampling at most $K_{\text{Reward}} + K_{\text{RFKSP}}$ episodes, where $K_{\text{Reward}} = \tilde{O}\left(S^2AB/\epsilon^2\right), K_{\text{RFKSP}} = \tilde{O}\left(S^9A^3B^3/\epsilon\right)$.*

Our lower-order terms in the sample complexity can be additionally improved if we apply the generative model to initialize our auxiliary MDP. We achieve $\tilde{O}(S^2A/\epsilon^2)$ PAC result in the generative setting. The formal guarantee is presented as Theorem G.1 in Appendix G.

Below, we provide a proof sketch for Theorem 6.1. The proofs for Theorem 6.3 and Theorem G.1 are similar. The formal proofs are presented in Appendix E, F, G respectively.

Proof Sketch of Theorem 6.1. Among T returned policies in MDP-Full, we denote

$$i = \arg\max_{t\in[T]} \hat{V}_1^{\pi^t}(s_0), j = \arg\max_{t\in[T]} V_1^{\pi^t}(s_0).$$

MDP-Full outputs policy π^i, whose gap $V_1^*(s_0) - V_1^{\pi^i}(s_0)$ can be decomposed as

$$\le \underbrace{V_1^*(s_0) - V_1^{\pi^j}(s_0)}_{\text{exploration error}} + \underbrace{\left|V_1^{\pi^j}(s_0) - \hat{V}_1^{\pi^j}(s_0)\right|}_{\text{evaluation error}}$$
$$+ \underbrace{\left|\hat{V}_1^{\pi^i}(s_0) - V_1^{\pi^i}(s_0)\right|}_{\text{evaluation error}} + \underbrace{\hat{V}_1^{\pi^j}(s_0) - \hat{V}_1^{\pi^i}(s_0)}_{\le 0 \text{ by definition}}.$$

The exploration error and the evaluation error can be tackled by the following two theorems respectively. Lemma 6.5 states that the returned policy by MDP-RBUCBI is near-optimal with constant probability. As we run MDP-RBUCBI for $T = O(\log(1/\delta))$ times independently, π^j is near-optimal with high probability. We can also suitably estimate all the given policies by Lemma 6.4. Taking union bound to combine these two parts conclude our proof.

Lemma 6.4. *The estimate $\hat{V}^\pi(s_0)$ returned by MDP-Evaluation satisfies that with probability $1 - \delta_{\text{eval}}$, $|\hat{V}_1^\pi(s_0) - V_1^\pi(s_0)| \le O(\epsilon_{\text{eval}})$.*

MDP-Evaluation(Appendix E.2) resembles MDP-Full except that it runs a given policy π and overestimates its value instead of the optimal value in the UCB phase. The evaluation gap between the overestimation and the true value can be tackled similarly as in MDP-RBUCBI since they are also expected sum over the trajectory space following π.

Lemma 6.5. *The policy π returned by MDP-RBUCBI (Algorithm 2) satisfies that with probability $\frac{1}{2}$, $V_1^*(s_0) - V_1^\pi(s_0) \le O\left(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right)$.*

Proof Sketch of Lemma 6.5. This lemma is derived by combining lemma 6.8, which bound the expectation of the regret with respect to the aux MDP, and lemma 6.6, which states that the value function in a good conditioned aux MDP is close to the true MDP. Specifically, for any $k \in \{0\} \cup [K-1]$, we define good event G_k holds if Γ_k satisfies:

1. The auxiliary Markovian environment built on τ_0 is ϵ_{ksp}-good conditioned for ϵ_{ksp} given in RFKSP.

2. For any (s, a) and $t \in [k+1]$, $\hat{P}_{s,a}^t$ satisfies equation 2 and $\mathbb{E}\left[R(s,a)\right] \in \mathcal{R}_{s,a}^t$.

Defining good event for every k will be of use in our refined analysis. We further set G_K as G_{K-1}, which holds with probability $1 - \delta_{\text{ksp}} - 2S^2AK\delta_{\text{conf}}$. Property 1 is related to lemmas in Section 6.1. Property 2 ensures that our overestimation works as $\overline{V}_h^k(s_h) \ge \tilde{V}_h^*(s_h)$. By applying Markov inequality twice to the expectation of regret and setting K, δ_{conf} and δ_{ksp} appropriately, we conclude that both G_K and $\tilde{V}_1^*(s_0) - \tilde{V}_1^\pi(s_0) \le O\left(\epsilon_{\text{ucb}}\right)$ hold with probability $1/2$. We denote the value of the optimal policy for the true MDP in the aux MDP as $\tilde{V}^{**}$. By definition, $\tilde{V}_1^*(s_0) \ge \tilde{V}_1^{**}(s_0)$, and we have that

$$V_1^*(s_0) - V_1^\pi(s_0) \le \tilde{V}_1^{**}(s_0) - \tilde{V}_1^\pi(s_0) + 2\epsilon_{\text{ksp}}$$
$$\le \tilde{V}_1^*(s_0) - \tilde{V}_1^\pi(s_0) + 2\epsilon_{\text{ksp}} \le O\left(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right).$$

6.1. Auxiliary Markovian environment

In this section we give lemmas related to auxiliary markovian environment. For a good conditioned aux MDP, the value function within is close to the true MDP since the reward on the majority of trajectories are the same.

Lemma 6.6. *Suppose the max visiting probability to $\mathcal{O}$ is less than ϵ, i.e.$\max_\pi \mathbb{P}_\pi\left[\exists h \in [H], (s_h, a_h) \in \mathcal{O}^C\right] \le \epsilon$, then for any fixed policy π, $\left|V_1^\pi(s_0) - \tilde{V}_1^\pi(s_0)\right| \le \epsilon$.*

We denote

$$L = \max_{(s,a)\in\mathcal{O}^C} \sum_{k=1}^K \min\left(\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right).$$

Under most cases, the visit counts of a state-action pair in a single episode do not exceed its all visit counts before this episode, and thus the summation of $1/N^k(s_h^k, a_h^k)$ can be generally bounded by SAL. Since $N^{K+1}(s,a) - N^k(s,a) \le KU(s,a)/\delta$ holds with probability $1-\delta$, L is bounded when there is sufficient initial samples.

Lemma 6.7. $\mathbb{E}_{\Gamma_K} L \mathbf{1}_{\mathrm{G}_0} \le O(\mathrm{polylog}(S, A, K))$.

6.2. Refined Analysis for RBUCBI

In this section we use our refined analysis to bound the expectation of regret with respect to the aux MDP.

Lemma 6.8. *In MDP-RBUCBI, the expectation of regret with respect to the auxiliary MDP can be bounded by $\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K \left[\tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^k}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \le \tilde{O}(S\sqrt{AK})$.*

Proof Sketch of Lemma 6.8 Here we assume the reward is deterministic for simplicity. Under G_K, the regret can be bounded by the difference of $\overline{V}_1^k$ and $\tilde{V}_1^{\pi^k}$, which are both expected sum over the trajectory space following π^k. We expand their difference into the expectation form as

$$T_1 = \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\Gamma_k}\left(\max_{\overline{p}\in\mathcal{P}^k_{s_h^k, a_h^k}} \overline{p} - \tilde{P}_{s_h^k, a_h^k}\right)\overline{V}_{h+1}^k \mathbf{1}_{\mathrm{G}_{k-1}}$$

We further bound the above term by the width of the confidence set, which is accurate enough under the good event, as $O(\sqrt{ST_2} \cdot \sqrt{T_3} + ST_2)$, where

- $T_2 = \mathbb{E}_{\Gamma_K} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{N^k(s_h^k, a_h^k)} \mathbf{1}_{G_0}$,
- $T_3 = \mathbb{E}_{\Gamma_K} \sum_{k=1}^{K} \sum_{h=1}^{H} V(\tilde{P}_{s_h^k, a_h^k}, \overline{V}^k_{h+1}) \mathbf{1}_{G_{k-1}}$.

Generally $T_2 \leq SAL$ and can be bounded by lemma 6.7. For T_3, our refined analysis bound it by $O(K + T_1)$.

$$\begin{aligned}
&\mathbb{E}_{\Gamma_K} \sum_{k,h} \left[\tilde{P}_{s_h^k, a_h^k} \left(\overline{V}^k_{h+1}\right)^2 - \left(\tilde{P}_{s_h^k, a_h^k} \overline{V}^k_{h+1}\right)^2\right] \mathbf{1}_{G_{k-1}} \\
&= \mathbb{E}_{\Gamma_K} \sum_{k,h} \left[\overline{V}^k_{h+1}(s^k_{h+1})^2 - \left(\tilde{P}_{s_h^k, a_h^k} \overline{V}^k_{h+1}\right)^2\right] \mathbf{1}_{G_{k-1}} \\
&\leq \mathbb{E}_{\Gamma_K} \sum_{k,h} \left[\overline{V}^k_{h}(s^k_{h})^2 - \left(\tilde{P}_{s_h^k, a_h^k} \overline{V}^k_{h+1}\right)^2\right] \mathbf{1}_{G_{k-1}} \\
&= \mathbb{E}_{\Gamma_K} \sum_{k,h} \left[\overline{Q}^k_{h}(s^k_{h}, a^k_h)^2 - \left(\tilde{P}_{s_h^k, a_h^k} \overline{V}^k_{h+1}\right)^2\right] \mathbf{1}_{G_{k-1}} \\
&\leq 3\mathbb{E}_{\Gamma_K} \sum_{k,h} \left| r(s_h^k, a_h^k) \right| + 3T_1
\end{aligned}$$

The last line is derived by the formula for the difference of square. Therefore we bound T_1 by its recursive structure.

While $\overline{V}^k_h(s_h^k) = \overline{Q}^k_h(s_h^k, a_h^k)$ holds directly in MDP since the policy there is deterministic, its counterpart in MG setting can only hold under expectation since the policy here is nondeterministic. We can proceed as follows.

$$\begin{aligned}
&\mathbb{E}_{\Gamma_k} \overline{V}^k_h(s_h^k)^2 = \mathbb{E}_{\Gamma_k} \left(\mathbb{E}_{a_h^k, b_h^k} \overline{Q}^k_h(s_h^k, a_h^k, b_h^k)\right)^2 \\
\leq &\mathbb{E}_{\Gamma_k} \mathbb{E}_{a_h^k, b_h^k} \overline{Q}^k_h(s_h^k, a_h^k, b_h^k)^2 = \mathbb{E}_{\Gamma_k} \overline{Q}^k_h(s_h^k, a_h^k, b_h^k)^2.
\end{aligned}$$

7. Conclusion

We propose a relaxed reward-boundedness assumption for studying the horizon-dependence problem. Under our relaxed assumption, we propose a generic algorithmic framework consists of reward-free phase and reward-based phase that achieves horizon-free learning for both MDPs and Games. Our work improves the existing horizon-independent PAC bounds in both the online setting and the generative setting.

8. Acknowledgement

This work was supported in part by DARPA under agreement DARPA-HR00112190130, NSF grant #2221871, and an Amazon Research Grant.

References

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Aumann, R. J. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pp. 1–18, 1987.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pp. 551–560. PMLR, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

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

Brafman, R. I. and Tennenholtz, M. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3(Oct):213–231, March 2003. ISSN 1532-4435.

Cui, Q. and Yang, L. F. Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pp. 1496–1504. PMLR, 2021.

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

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5717–5727, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1507–1516, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Dong, K., Wang, Y., Chen, X., and Wang, L. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating markov decision processes. In *Advances in Neural Information Processing Systems*, pp. 2994–3004, 2018.

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

Hansen, T. D., Miltersen, P. B., and Zwick, U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1, 2013.

Hu, J. and Wellman, M. P. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

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

Jia, Z., Yang, L. F., and Wang, M. Feature-based q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Jiang, N. and Agarwal, A. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pp. 3395–3398, 2018.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Kakade, S. M. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Kearns, M. and Singh, S. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

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

Kolter, J. Z. and Ng, A. Y. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pp. 513–520, 2009.

Lattimore, T. and Hutter, M. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pp. 320–334. Springer, 2012.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.

Li, Y., Wang, R., and Yang, L. F. Settling the horizon-dependence of sample complexity in reinforcement learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 965–976. IEEE, 2022.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on International Conference on Machine Learning*, ICML'94, pp. 157–163, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1-55860-335-2. URL `http://dl.acm.org/citation.cfm?id=3091574.3091594`.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Menard, P., Domingues, O. D., Shang, X., and Valko, M. Ucb momentum q-learning: Correcting the bias without forgetting. *arXiv preprint arXiv:2103.01312*, 2021.

Moulin, H. and Vial, J.-P. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3):201–221, 1978.

Neu, G. and Pike-Burke, C. A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891*, 2020.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2701–2710. JMLR. org, 2017.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. *arXiv preprint arXiv:1306.0940*, 2013.

Pacchiano, A., Ball, P., Parker-Holder, J., Choromanski, K., and Roberts, S. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

Peel, T., Anthoine, S., and Ralaivola, L. Empirical bernstein inequality for martingales: Application to online learning. 2013.

Ren, T., Li, J., Dai, B., Du, S. S., and Sanghavi, S. Nearly horizon-free offline reinforcement learning. *arXiv preprint arXiv:2103.14077*, 2021.

Russo, D. Worst-case regret bounds for exploration via randomized value functions. *arXiv preprint arXiv:1906.02870*, 2019.

Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018a.

Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 770–787. Society for Industrial and Applied Mathematics, 2018b.

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002. PMLR, 2020.

Simchowitz, M. and Jamieson, K. Non-asymptotic gap-dependent regret bounds for tabular mdps. *arXiv preprint arXiv:1905.03814*, 2019.

Singh, S. P. and Yee, R. C. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Strehl, A. L. and Littman, M. L. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 881–888. ACM, 2006.

Szita, I. and Szepesvári, C. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, 2010.

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. *arXiv preprint arXiv:1803.01626*, 2018.

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pp. 3674–3682. PMLR, 2020.

Yang, K., Yang, L., and Du, S. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312, 2019.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020a.

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

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020b.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020c.

Zhang, Z., Ji, X., and Du, S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021.

Zhang, Z., Ji, X., and Du, S. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pp. 3858–3904. PMLR, 2022.

A. Technical Lemmas

Lemma A.1 (**Lemma 10 in (Zhang et al., 2022)**). *Let $X_1, X_2, \ldots$ be a sequence of random variables taking value in $[0, l]$. Define $\mathcal{F}_k = \sigma(X_1, X_2, \ldots, X_{k-1})$ and $Y_k = \mathbb{E}[X_k|Y_k]$ for $k \geq 1$. For any $\delta > 0$, we have that*

$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} X_k \geq 3\sum_{k=1}^{n} Y_k + l\log(1/\delta)\right] \leq \delta,$$
$$\mathbb{P}\left[\exists n, \sum_{k=1}^{n} Y_k \geq 3\sum_{k=1}^{n} X_k + l\log(1/\delta)\right] \leq \delta.$$

Lemma A.2 (**Bernstein's Inequality**). *Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables with values in $[0, 1]$ and let $\delta > 0$. Define $\mathbb{V}Z = \mathbb{E}\left[(Z - \mathbb{E}Z)^2\right]$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\mathbb{V}Z\log(2/\delta)}{n}} + \frac{\log(2/\delta)}{3n}\right] \leq \delta.$$

Lemma A.3 (**Freedman's Inequality Lemma 1 in (Peel et al., 2013)**). *Suppose $X_1, \ldots, X_n$ is a sequence of random variables such that $0 \leq X_i \leq 1$. Define the martingale difference sequence $\{Y_n = \mathbb{E}[X_n|X_1, \ldots, X_{n-1}] - X_n\}$ and note K_n the sum of the conditional variances*

$$K_n = \sum_{t=1}^{n} \mathbb{V}[X_t|X_1, \ldots, X_{t-1}].$$

Let $S_n = \sum_{i=1} X_i$, then for all $\epsilon, k \geq 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n} \mathbb{E}[X_i|X_1, \ldots, X_{i-1}] - S_n \geq \epsilon, K_n \leq k\right] \leq \exp\left(-\frac{\epsilon^2}{2k + 2\epsilon/3}\right).$$

Lemma A.4 (**Theorem 4 in (Maurer & Pontil, 2009)**). *Let $Z, Z_1, \ldots, Z_n (n \geq 2)$ be i.i.d. random variables with values in $[0, 1]$ and let $\delta > 0$. Define $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ and $\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\hat{V}_n\ln(2/\delta)}{n-1}} + \frac{7\ln(2/\delta)}{3(n-1)}\right] \leq \delta.$$

B. Discussion of horizon-independence

This section discusses one of the key ideas in achieving horizon-independence. The idea comes from (Zhang et al., 2022). For the integrity of our paper, we follow part of their analysis and list it here.

In nearly all UCB-based algorithms, we need to bound the term like $\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{N^k\left(s_h^k, a_h^k\right)}$ where $\left(s_h^k, a_h^k\right)$ is the state-action pair of the h-th step in the k-th episode. If we assume $N^{k+1}(s, a) \leq 2N^k(s, a)$, which is natural when we have already collected many samples, the classic analysis by pigeonhole will further lead us to

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{N^k\left(s_h^k, a_h^k\right)} = O\left(\sum_{s,a}\sum_{k=1}^{K}\log\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right)\right)$$
$$\leq O(SA\log(KH)).$$

(Zhang et al., 2022) observes that we can avoid dependence on H once we have enough initial samples for every state-action pair. To be more specific, we define $U(s, a) = \max_\pi \mathbb{E}_\pi\left[\sum_{h=1}^{H} \mathbf{1}_{(s_h, a_h)=(s,a)}\right]$ to be the maximum expected visitation count of (s, a) in one episode. By Markov inequality, the total count of (s, a) in K episodes satisfies $N^{K+1}(s, a) - N^1(s, a) \leq KU(s, a)/\delta$ with probability $1 - \delta$. So if $N^1(s, a) \geq U(s, a)/\exp(\text{poly}(S, A))$,

$$\sum_{s,a}\log\left(\frac{N^{K+1}(s,a)}{N^1(s,a)}\right) = O\left(\text{poly}(S, A)\log\left(K/\delta\right)\right),$$

which is independent of H.

C. Reward-Free Key-State Preserving Algorithm

In this section we give a viable RFKSP algorithm, which is modified from the collecting initial sample state in (Zhang et al., 2022). The main algorithm is given in Algorithm 4 and two other supplementary algorithms are given in Algorithm 5 and Algorithm 6. Some extra notations are being used in the following algorithms. In particular,

1. $W_d^\pi(r, p, \mu_1) := \mathbb{E}\left[\sum_{h=1}^H r_h | s_1 \sim \mu_1\right]$: the general value function.

2. $W_\gamma^\pi(r, p, \mu_1) := \mathbb{E}\left[\sum_{i \geq 1} \gamma^{i-1} r_i | s_1 \sim \mu_1\right]$: the value function in the discounted MDP.

3. $X_d^\pi(\mathcal{O}, p, \mu_1)$: the probability of reaching $\mathcal{O}$ in d steps under transition probability p, policy π, initial distribution μ_1.

4. $X_\gamma^\pi(\mathcal{O}, p, \mu_1) := \sum_{i \geq 1} \gamma^{i-1} \mathbb{P}\left[(s_i, a_i, s_{i+1}) \in \mathcal{O}, (s_{i'}, a_{i'}, s_{i'+1}) \notin \mathcal{O}, \forall 1 \leq i' \leq i-1 | s_1 \sim \mu_1\right]$.

We further prove in Lemma C.1 that Algorithm 4 serves as a viable RFKSP algorithm as we defined. The proof is based on the lemmas provided in (Zhang et al., 2022).

Algorithm 4 Reward-Free Key-State Preserving

Input:MDP $\mathcal{M}$, ϵ, δ.
Initialization: $N(s, a, s') \leftarrow 0, \forall s, a, s'$, $\bar{N}(s, a) \leftarrow 0, \forall (s, a)$, $d \leftarrow \frac{(S+1)H}{S+2}$. $\mathcal{O}^1 \leftarrow \mathcal{S} \times \mathcal{A}$. $n_1 \leftarrow C_2 S^7 A^3 \iota$. $d' = H - d$. $m(s, a) \leftarrow 0$. $N_0 \leftarrow 256 S^2 \log(1/\delta)$,$K_1 = \tilde{O}(\frac{S^9 A^3}{\epsilon})$.
for $k = 1, 2, \ldots, K_1$ **do**
  $\mathcal{P}^k \leftarrow$Build confidence set $\mathcal{P}^k_{s,a}$ based on $(\{N(s, a, s')\}_{s,a,s'})$.
  $(\pi^k, \tilde{P}^k) \leftarrow \max_{\pi, p \in \mathcal{P}^k} X_d^\pi(\mathcal{O}^k, p, \mu_1)$
  for $h = 1, 2, \ldots, d$ **do**
    Observes s_h^k, takes action $\pi_h^k(s_h^k)$, receives r_h^k and transits to s_{h+1}^k.
    $N(s_h^k, a_h^k, s_{h+1}^k) \leftarrow N(s_h^k, a_h^k, s_{h+1}^k) + 1$.
    if $\exists a, (s_{h+1}^k, a) \in \mathcal{O}^k$ **then**
      $(s_1^*, a_1^*) \leftarrow (s_{h+1}^k, a)$.
      $\{N(s, a, s')\}_{s,a,s'} \leftarrow \{n(s, a, s')\}_{s,a,s'}$.
      $\mathcal{K}^k \leftarrow \{(s, a, s') : n(s, a, s') \geq N_0\}$, $\mathcal{K}^k(s, a) \leftarrow \{s' : (s, a, s') \in \mathcal{K}^k\}$.
      $n(s, a) \leftarrow \max\{\sum_{s' : (s,a,s') \in \mathcal{K}^k} n(s, a, s'), 1\} \; \forall (s, a)$.
      $P^{\text{ref}}_{s,a,s'} \leftarrow \frac{n(s,a,s')}{n(s,a)}$, $P^{\text{ref}}_{s,a,z} \leftarrow 0, \forall (s, a, s') \in \mathcal{K}^k$.
      $P^{\text{ref}}_{s,a,s'} \leftarrow 0$, $P^{\text{ref}}_{s,a,z} = 1, \forall (s, a, s')$ such that $\mathcal{K}^k(s, a) = \emptyset$.
      $(\text{Trigger}, \{n(s, a, s')\}_{s,a,s'}) \leftarrow \mathit{Algorithm\ 5}$ with inputs$((s_1^*, a_1^*), P^{\text{ref}}, \{n(s, a, s')\}_{(s,a,s')}, \mathcal{K}^k, d')$.
      if Trigger $=$ FALSE **then**
        $\{n(s, a, s')\}_{(s,a,s')} \leftarrow \mathit{Algorithm\ 6}$ with inputs $((s_1^*, a_1^*), P^{\text{ref}}, \{n(s, a, s')\}_{s,a,s'}, d')$
        $m(s_1^*, a_1^*) \leftarrow m(s_1^*, a_1^*) + 1$.
        if $m(s_1^*, a_1^*) \geq 400 \log(1/\delta)$ **then**
          $\mathcal{O}^{k+1} \leftarrow \mathcal{O}^k / (s_1^*, a_1^*)$.
        end if
      end if
      $\{n(s, a, s')\}_{s,a,s'} \leftarrow \{N(s, a, s')\}_{s,a,s'}$.
      break.
    end if
  end for
  If there are remaining steps, run a random policy and update $\{N(s, a, s')\}_{s,a,s'}$.
end for
Return: an auxiliary Markovian environment $\tilde{M}$ which is built based on $\mathcal{O}^{K_1+1}$.

Algorithm 5 Explicit Exploration

Input: starting state-action pair (s_1, a_1), reference model P^{ref}, sample count $\{n(s,a,s')\}_{s,a,s'}$, known set $\mathcal{K}$, horizons d', $d_2 = d'/(20S\log(S))$, $d_1 = d' - d_2$
Initialization: discounted factor $\gamma = 1 - 1/d_2$, $N_0 \leftarrow 256S^2\log(1/\delta)$;Trigger = FALSE;
for $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**
  if $\exists s' \in \mathcal{S}$ such that $(s,a,s') \notin \mathcal{K}$ **then**
    $\pi_1^k \leftarrow \arg\max_{\pi\in\Pi_{\text{sta}},\pi_1(s_1)=a_1} X_\gamma^\pi(\{s\}, P^{\text{ref}}, \mathbf{1}_{s_1})$;
    $u^k(s) \leftarrow X_\gamma^{\pi_1^k}(\{s\}, P^{\text{ref}}, \mathbf{1}_{s_1})$;
    $\pi_2^k \leftarrow \arg\max_{\pi\in\Pi_{\text{sta}}} W_\gamma^\pi(\mathbf{1}_{s,a}, P^{\text{ref}}, \mathbf{1}_s)$;
    $v^k(s,a) \leftarrow W_\gamma^{\pi_2^k}(\mathbf{1}_{s,a}, P^{\text{ref}}, \mathbf{1}_s)$;
    if $u^k(s) \geq \frac{1}{1200S}$ and $n(s,a) \leq 810SAN_0u^k(s)v^k(s,a)$ **then**
      $\text{Trigger} \leftarrow \text{TRUE}$;
      Run π_1^k for d_1 steps. Stop if (s,a) is reached or some *unknown* state-action-state tuple is visited;
      if (s,a) is reached **then**
        Play π_2^k for d_2 steps, then play random policies till the end;
      else
        Play random policies till the end;
      end if
      Let $\{s_i, a_i, s_{i+1}\}_{i=1}^{d'}$ denote the data collected in the length d'-trajectory;
      for $i = 1, 2, \ldots, d'$ **do**
        $n(s_i, a_i, s_{i+1}) \leftarrow n(s_1, a_i, s_{i+1}) + 1$;
      end for
      Break;
    end if
  end if
end for
Return: Trigger, $\{n(s,a,s')\}_{(s,a,s')}$;

Algorithm 6 Sample Collection with a Reference Model

Input: initial state-action pair (s_1, a_1), reference model P^{ref} , visit count $\{n(s,a,s')\}_{s,a,s'}$, horizon d'.
Initialization: discounted factor $\gamma = 1 - 1/d_2$ where $d_2 = d'/(20S\log(S))$.
$\pi \leftarrow \arg\max_{\pi\in\Pi_{\text{sta}}} W_\gamma^\pi(\mathbf{1}_{s_1,a_1}, P^{\text{ref}}, \mathbf{1}_{s_1})$
Run π and collect d' samples $\{s_i, a_i, s_{i+1}\}_{i=1}^{d'}$;
for $i = 1, 2, \ldots, d'$ **do**
  $n(s_i, a_i, s_{i+1}) \leftarrow n(s_i, a_i, s_{i+1}) + 1$;
end for
Return: $\{n(s,a,s')\}_{(s,a,s')}$;

Lemma C.1. *Algorithm 4 serves as a viable reward-free key-state preserving algorithm with*

$$K_1 = O\left(\frac{S^9A^3\iota^2}{\epsilon}\text{polylog}\left(S, A, \frac{1}{\epsilon}\right)\right).$$

Remark C.2. Recall the definition of RFKSP in Definition 5.3. The above lemma indicates the following facts. For any given $\epsilon, \delta > 0$, after using $K_1 = \text{poly}\left(S, A, \iota, \frac{1}{\epsilon}\right)$ episodes, the auxiliary Markovian environment returned by Algorithm 4 is ϵ-good conditioned with probability at least $1-\delta$. In MG setting, this lemma still holds by substituting $\mathcal{A}$ by $\mathcal{A}\times\mathcal{B}$. In particular, $K_1 = O\left(\frac{S^9A^3B^3\iota_0^2}{\epsilon}\text{polylog}\left(S, A, B, \frac{1}{\epsilon}\right)\right)$.

Proof. Combining Lemma 6 and Lemma 24 in (Zhang et al., 2022), we have that with probability $1 - O\left(\frac{K_1^2}{S^8A^2\iota_0}\delta_0\right)$

1. $\max_\pi \mathbb{P}_\pi\left[\exists h \in [H], (s_h, a_h) \in \mathcal{O}\right] \leq O\left(\frac{S^9 A^3 \iota_0 + S^3 A \iota_0^2}{K_1}\right)$.

2. $N^1(s,a) \geq O\left(\frac{U(s,a)}{S(S+1)\log(S)}\right)$ for all $(s,a) \in \mathcal{O}^C$.

To meet our requirements for RFKSP, we need $O\left(\frac{S^9 A^3 \iota_0 + S^3 A \iota_0^2}{K_1}\right) \leq \epsilon$ and $O\left(\frac{K_1^2}{S^8 A^2 \iota_0}\delta_0\right) \leq \delta$. The first equation can be satisfied by setting $K_1 = \Omega\left(\frac{S^9 A^3 \iota_0^2}{\epsilon}\right)$. Substituting it into the second equation and noting that $\delta_0^{-0.5} \geq O\left(\iota_0^3\right)$, we can meet both of our requirements by setting $\delta_0 = \frac{\delta^2 \epsilon^4}{S^{20} A^8}$.

Wrapping up all these results, we have that if we set $K_1 = O\left(\frac{S^9 A^3 \iota^2}{\epsilon} \mathrm{polylog}\left(S, A, \frac{1}{\epsilon}\right)\right)$, with probability $1-\delta$,

1. $\max_\pi \mathbb{P}_\pi\left[\exists h \in [H], (s_h, a_h) \in \mathcal{O}\right] \leq \epsilon$.

2. $N^1(s,a) \geq \frac{U(s,a)}{\mathrm{poly}(S)}$ for all $(s,a) \in \mathcal{O}^C$.

$\square$

D. Auxiliary Proofs

In this section, we illustrate the auxiliary lemmas that will be used in both MDP and MG settings. In particular, we give proof to the lemma concerning the auxiliary Markovian environment and the confidence set we built in our algorithm. All the lemmas in this section are given in the MDP setting. They can be translated into MG setting by viewing the product of two players' action space $\mathcal{A} \times \mathcal{B}$ in MG as the action space $\mathcal{A}$ in MDP.

D.1. Proofs for Auxiliary Markovian environment

Lemma D.1 (**Restatement of Lemma 6.6**). *Suppose the maximum visiting probability to $\mathcal{O}^C$ is ϵ, i.e.*

$$\max_{\pi} \mathbb{P}_{\pi}\left[\exists h \in [H], (s_h, a_h) \in \mathcal{O}^C\right] \leq \epsilon,$$

then for any fixed policy π, $\left|V_1^{\pi}(s_0) - \tilde{V}_1^{\pi}(s_0)\right| \leq \epsilon$.

Proof of Lemma 6.6. In this proof we denote a single trajectory as $\Gamma = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H, s_{H+1})$. We further divide $\Gamma = \Gamma_1 \cup \Gamma_2$ where Γ_1 denotes the trajectory before the first visit to $\mathcal{O}$ and Γ_2 denotes the left trajectory. For example, if (s_2, a_2) is the first time the trajectory visits to $\mathcal{O}$, $\Gamma_1 = (s_1, a_1, s_2)$ and $\Gamma_2 = (s_2, a_2, \ldots, s_{H+1})$. Γ_2 can be empty in the extreme case where the trajectory Γ never visits to $\mathcal{O}$.

In the original model, we use $r(\Gamma)$ and $P(\Gamma)$ to denote the expected reward and the probability of the trajectory respectively. For a given Γ_1, we denote the set of suitable Γ_2 as $S(\Gamma_1) := \{\Gamma_2 : \exists \Gamma = \Gamma_1 \cup \Gamma_2\}$. With these notations, we have that

$$\begin{aligned}
V_1^{\pi}(s_0) &= \sum_{\Gamma_1} \sum_{\Gamma_2 \in S(\Gamma_1)} (r(\Gamma_1) + r(\Gamma_2)) \cdot P(\Gamma_1) \cdot P(\Gamma_2), \\
\tilde{V}_1^{\pi}(s_0) &= \sum_{\Gamma_1} \sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_1) \cdot P(\Gamma_1) \cdot P(\Gamma_2).
\end{aligned}$$

So the difference can be calculated as

$$\begin{aligned}
\left|V_1^{\pi}(s_0) - \tilde{V}_1^{\pi}(s_0)\right| &= \left|\sum_{\Gamma_1} \sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_2) \cdot P(\Gamma_1) \cdot P(\Gamma_2)\right| \\
&\leq \sum_{\Gamma_1} P(\Gamma_1) \left|\sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_2) \cdot P(\Gamma_2)\right|.
\end{aligned}$$

When Γ_2 is an empty set, $r(\Gamma_2) = 0$. From the requirement that the max visiting probability to $\mathcal{O}$ is ϵ, we know that the probability of $\Gamma_1 \neq \Gamma$ is less than ϵ. If $\Gamma_1 \neq \Gamma$, we assume the last term in Γ_1 is (s_h, a_h). We set $\pi^{'} = \pi$ except $\pi_h^{'}(s_h) = a$. Then from our reward assumption, we have that

$$1 \geq \mathbb{E}_{\pi}\left[\sum_{t=h}^{H} |r(s_h, a_h)|\right] \geq \left|\mathbb{E}_{\pi}\left[\sum_{t=h}^{H} r(s_h, a_h)\right]\right| = \left|\sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_2) \cdot P(\Gamma_2)\right|.$$

Thus we conclude our proof by noticing that

$$\sum_{\Gamma_1} P(\Gamma_1) \left|\sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_2) \cdot P(\Gamma_2)\right| \leq \sum_{\Gamma_1 : \Gamma_1 \neq \Gamma} P(\Gamma_1) \leq \epsilon.$$

□

Cut off. Note that $t \leq \log_2(1+t)$ only holds when $0 < t < 1$, we need to cut off some term when a single (s,a) pair is visited too many times in a single episode. We define $N_h^k(s,a)$ to be the visit count before the h-th step in the k-th episode, $\mathcal{J} = \left\{(k,h) : \exists (s,a) \in \mathcal{O}^C, s.t. N_h^k(s,a) \geq 2N^k(s,a)\right\}$. $I_h^k = 1$ if $(k,h) \notin \mathcal{J}$ else 0. By the definition, $I_1^k = 1$ and I_h^k do not depend on the action taken at the h-th step in the k-th episode. We define $L = \max_{(s,a)\in\mathcal{O}^C} \sum_{k=1}^{K} \min\left(\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right)$.

Lemma D.2.

$$\sum_{k=1}^{K} \mathbf{1}_{I_{H+1}^k=0} \leq SAL \quad \text{and} \quad \sum_{k=1}^{K}\sum_{h=1}^{H} \frac{I_h^k}{N^k\left(s_h^k, a_h^k\right)} \mathbf{1}_{\left(s_h^k, a_h^k\right)\in\mathcal{O}^C} \leq SAL. \tag{4}$$

Proof of Lemma D.2. For fixed k, if $I\left[\exists h, I_h^k = 0\right] = 0$, $I\left[\exists h, I_h^k = 0\right] \leq \sum_{(s,a)\in\mathcal{O}^C} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\}$ holds naturally. If $I\left[\exists h, I_h^k = 0\right] = 1$, there exist $(s,a)\in\mathcal{O}^C$ such that $N^{k+1}(s,a) \geq 2N^k(s,a)$. In this case $I\left[\exists h, I_h^k = 0\right] \leq \sum_{(s,a)\in\mathcal{O}^C} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\}$, and we have that

$$\begin{aligned}
\sum_{k=1}^{K} I\left[\exists h, I_h^k = 0\right] &\leq \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{O}^C} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\} \\
&= \sum_{(s,a)\in\mathcal{O}^C}\sum_{k=1}^{K} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\} \\
&\leq SA \max_{(s,a)\in\mathcal{O}^C} \sum_{k=1}^{K} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\} = SAL.
\end{aligned}$$

Meanwhile, due to the existence of I_h^k, we can derive that

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{I_h^k}{N^k\left(s_h^k, a_h^k\right)} \mathbf{1}_{\left(s_h^k, a_h^k\right)\in\mathcal{O}^C} &\leq \sum_{(s,a)\in\mathcal{O}^C}\sum_{k=1}^{K} \min\left\{\frac{N^{k+1}(s,a) - N^k(s,a)}{N^k(s,a)}, 1\right\} \\
&= \sum_{(s,a)\in\mathcal{O}^C}\sum_{k=1}^{K} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\} \\
&\leq SA \max_{(s,a)\in\mathcal{O}^C} \left\{\sum_{k=1}^{K} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\}\right\} = SAL.
\end{aligned}$$

$\square$

Lemma D.3 (Restatement of lemma 6.7). *L can be bounded by*

$$\mathbb{E}_{\Gamma_K} L\mathbf{1}_{\mathrm{G}_0} \leq O(\mathrm{polylog}(S, A, K)). \tag{5}$$

Note: This lemma is similar to Lemma 26 in (Zhang et al., 2022). The difference is that here we do not need to deal with the case of $(s,a)\in\mathcal{O}$ due to the construction of our auxiliary Markovian environment.

Proof of Lemma 6.7. We define $B(s,a) = \left\{k\in[K] : N^{k+1}(s,a) - N^k(s,a) \geq K^2 U(s,a)\right\}$. For different k, we bound the corresponding term in L as follows.

$$\min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), 1\right\} \leq \begin{cases} \log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right), & k\in B(s,a); \\ 1, & k\notin B(s,a). \end{cases}$$

To apply Lemma A.1, we denote $X_k = \mathbf{1}_{k\in B(s,a)}$. $Y_k = \mathbb{E}[X_k|\mathcal{F}_k] = P(N^{k+1}(s,a) - N^k(s,a) \geq K^2U(s,a)|\mathcal{F}_k) \leq 1/K^2$. Therefore with probability $1-\delta$, $|B(s,a)| = \sum_{k=1}^K X_k \leq 3\sum_{k=1}^K Y_k + \iota \leq 3/K + \iota$. Taking union bound, we have that with probability $1 - SA\delta$, $|B(s,a)| \leq \frac{3}{K} + \iota$ hold for $\forall(s,a)$. Under such event, for any $(s,a)\in\mathcal{O}^C$, suppose

$B(s,a)=\{k_1,k_2,\dots\}$ and let $k_0=0$, we have that

$$
\begin{aligned}
\sum_{k\notin B(s,a)} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right),1\right\} &\le \sum_{i\ge 0}\sum_{k=k_i}^{k_{i+1}-1}\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right)\\
&\le \sum_{i\ge 0}\log_2\left(\frac{K^3U(s,a)+N^{k_i}(s,a)}{N^{k_i}(s,a)}\right)\\
&\le |B(s,a)|\log_2\left(\frac{K^3U(s,a)+N^1(s,a)}{N^1(s,a)}\right)\\
&\le \left(\frac{3}{K}+\iota\right)\log_2\left(\frac{K^3U(s,a)+N^1(s,a)}{N^1(s,a)}\right).\\
\sum_{k\in B(s,a)} \min\left\{\log_2\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right),1\right\} &\le |B(s,a)|\le \frac{3}{K}+\iota.
\end{aligned}
$$

When G_0 holds, $N_1(s,a)\ge\frac{U(s,a)}{\mathrm{poly}(S)}$ holds for any $(s,a)\in\mathcal{O}^C$ by definition. Thus by setting $\delta=\frac{1}{SAK^2}$ and adding up all the terms, we can conclude that

$$
\mathbb{E}_{\Gamma_K}L\mathbf{1}_{\mathrm{G}_0}\le\left(\frac{3}{K}+\iota\right)\log_2\left(\frac{K^3U(s,a)+N^1(s,a)}{N^1(s,a)}\right)+\left(\frac{3}{K}+\iota\right)+SA\delta K\le O(\mathrm{polylog}(S,A,K)).
$$

□

D.2. Proofs for Confidence Set

This section provides proof of the lemmas concerning the confidence set. Note that when interacting with the auxiliary Markovian environment, P and R in the following lemmas should be replaced by $\tilde{P}$ and $\tilde{R}$.

Lemma D.4. *For any $\delta_{\mathrm{conf}}>0$, with probability at least $1-S^2AK\delta_{\mathrm{conf}}$,*

$$
|P_{s,a,s'}-\hat{P}^k_{s,a,s'}|\le 5\sqrt{\frac{\hat{P}^k_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{5\iota_{\mathrm{conf}}}{N^k(s,a)},
$$

holds for any (s,a,s') and k. With probability at least $1-SAK\delta_{\mathrm{conf}}$,

$$
\left|\mathbb{E}R(s,a)-\hat{r}^k(s,a)\right|\le\sqrt{4\frac{\hat{V}^k\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{10\iota_{\mathrm{conf}}}{N^k(s,a)},
$$

holds for any (s,a) and k.

Proof of lemma D.4. For any fixed (s,a,s') and k, we have visited (s,a) for $N^k(s,a)$ times before the k-th episode. For $i\in[N^k(s,a)]$, if the state transits to s' after the i-th time we visited (s,a), we denote $X_i=1$. Otherwise, we denote $X_i=0$. We apply Freedman inequality (lemma A.3) to $X_1,X_2,\dots,X_{N^k(s,a)}$, in which $\mathbb{E}\left[X_i|X_1,\dots,X_{i-1}\right]=P_{s,a,s'}$ and $\mathbb{V}\left[X_i|X_1,\dots,X_{i-1}\right]=P_{s,a,s'}\left(1-P_{s,a,s'}\right)\le P_{s,a,s'}$. By further setting $k=N^k(s,a)P_{s,a,s'}$, we can derive from lemma A.3 that with probability $1-\delta_{\mathrm{conf}}$,

$$
|P_{s,a,s'}-\hat{P}^k_{s,a,s'}|\le\sqrt{2\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{\iota_{\mathrm{conf}}}{3N^k(s,a)}.
$$

.

When the above line holds, we have

$$
\begin{aligned}
5\sqrt{\frac{\hat{P}^k_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{5\iota_{\mathrm{conf}}}{N^k(s,a)} &\geq 5\sqrt{\frac{\left(P_{s,a,s'}-\sqrt{2\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}-\frac{\iota_{\mathrm{conf}}}{3N^k(s,a)}\right)\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{5\iota_{\mathrm{conf}}}{N^k(s,a)} \\
&\geq 5\sqrt{\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}-5\sqrt{\sqrt{2\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}\cdot\frac{\iota_{\mathrm{conf}}}{3N^k(s,a)}}+\left(5-\frac{5}{\sqrt{3}}\right)\frac{\iota_{\mathrm{conf}}}{N^k(s,a)} \\
&\geq \left(5-\frac{5\sqrt{2}}{2}\right)\sqrt{\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\left(5-\frac{5}{\sqrt{3}}-\frac{5}{6}\right)\frac{\iota_{\mathrm{conf}}}{N^k(s,a)} \\
&\geq \sqrt{2\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{\iota_{\mathrm{conf}}}{3N^k(s,a)}.
\end{aligned}
$$

We need to mention that when $P_{s,a,s'}-\sqrt{2\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}-\frac{\iota_{\mathrm{conf}}}{3N^k(s,a)}\leq 0$, we can skip the first line above and derive the second line directly. Taking union bound over (s,a,s') and k conclude our proof.

Note that in our new reward assumption, $r(s,a)$ is bounded in $[-1,1]$ instead of $[0,1]$. For fixed (s,a) and k, we denote $a^i(s,a)=(r^i(s,a)+1)/2, \forall i\in[N^k(s,a)]$. We further denote $\hat{V}^k_a$, $\hat{a}^k$ as the sample variance and the sample mean of $\{a^i\}$. By definition $\hat{V}^k=4\hat{V}^k_a$. Again we apply lemma A.4 to $\{a^i\}$, with probability $1-\delta_{\mathrm{conf}}$,

$$
\begin{aligned}
\left|\mathbb{E}R(s,a)-\hat{r}^k(s,a)\right| = 2\left|\mathbb{E}a-\hat{a}^k\right| &\leq 2\sqrt{4\frac{\hat{V}^k_a\iota_{\mathrm{conf}}}{N^k(s,a)}}+2\frac{5\iota_{\mathrm{conf}}}{N^k(s,a)} \\
&\leq 2\sqrt{\frac{\hat{V}^k\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{10\iota_{\mathrm{conf}}}{N^k(s,a)}.
\end{aligned}
$$

Taking union bound, we have that the above equation holds for any (s,a) and k with probability $1-SAK\delta_{\mathrm{conf}}$. ☐

Lemma D.5. *For given (s,a) and k, if equation 2 holds for any $s'\in\mathcal{S}$, for any P' and $P''\in\left\{\mathcal{P}^k_{s,a}\right\}$, we have that*

$$
|P'(s')-P''(s')|\leq C\left(\sqrt{\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{\iota_{\mathrm{conf}}}{N^k(s,a)}\right), \tag{6}
$$

hold for any $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$.

Note: Here the probability P is the true transition probability of the model we interact with. Moreover, substituting a by a,b and $\mathcal{A}$ by $\mathcal{A}\times\mathcal{B}$ can transform the above result into MG setting.

proof of Lemma D.5. Since P' and $P''\in\mathcal{P}^k_{s,a}$, using equation 2 leads to

$$
\begin{aligned}
|P'(s')-P''(s')| &\leq 10\sqrt{\frac{\hat{P}_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+10\frac{\iota_{\mathrm{conf}}}{N^k(s,a)} \\
&\leq 10\sqrt{\frac{\left(P_{s,a,s'}+\sqrt{2\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{\iota_{\mathrm{conf}}}{3N^k(s,a)}\right)\iota_{\mathrm{conf}}}{N^k(s,a)}}+10\frac{\iota_{\mathrm{conf}}}{N^k(s,a)} \\
&\leq 10\sqrt{\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+10\sqrt{\sqrt{2\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}\cdot\frac{\iota_{\mathrm{conf}}}{N^k(s,a)}}+16\frac{\iota_{\mathrm{conf}}}{N^k(s,a)} \\
&\leq 25\left(\sqrt{\frac{P_{s,a,s'}\iota_{\mathrm{conf}}}{N^k(s,a)}}+\frac{\iota_{\mathrm{conf}}}{N^k(s,a)}\right).
\end{aligned}
$$

Thus taking $C=25$ conclude this proof. ☐

E. Proofs for MDP(Theorem 6.1)

In this section, we give proofs and algorithms in MDP setting.

Theorem E.1 (**Restatement of Theorem 6.1**). *For any $\epsilon, \delta > 0$, with probability $1-\delta$, MDP-Full(Algorithm 1) returns an ϵ-optimal policy by sampling at most $K = K_{\text{Reward}} + K_{\text{RFKSP}}$ episodes, where*

$$K_{\text{Reward}} = O\left(\frac{S^2 A \iota^2}{\epsilon^2}\text{polylog}\left(S, A, \frac{1}{\epsilon}\right)\right),$$
$$K_{\text{RFKSP}} = O\left(\frac{S^9 A^3 \iota^2}{\epsilon}\text{polylog}\left(S, A, \frac{1}{\epsilon}\right)\right).$$

Proof of Theorem 6.1. Theorem 6.1 is mainly based on Lemma 6.5 and Lemma 6.4. Given these two theorems, we derive Theorem 6.1 as follows. In each running time $t \in [T]$, by Lemma 6.5 we have that with probability $1/2$,

$$0 \le V_1^*(s_0) - V_1^{\pi^t}(s_0) \le O\left(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right).$$

As we run the subroutine $T = \log(\frac{2}{\delta})$ times independently, with probability $1 - \frac{\delta}{2}$, there exists $j \in [T]$ that the above equation holds. By Lemma 6.4, the estimation $\hat{V}_1^{\pi^t}(s_0)$ returned by MDP-Evaluation satisfies that with probability $1 - \delta_{\text{eval}}$,

$$\left|\hat{V}_1^{\pi^t}(s_0) - V_1^{\pi^t}(s_0)\right| \le O\left(\epsilon_{\text{eval}}\right).$$

Since we set $\delta_{\text{eval}} = \frac{\delta}{2T}$ in MDP-Full, with probability $1 - \frac{\delta}{2}$, the above equation hold for $\forall t \in [T]$. Suppose we denote $i = \arg\max_{t\in[T]} \hat{V}^{\pi^t}(s_0)$. Taking union bound, we have that the following equation holds with probability $1-\delta$.

$$\begin{aligned}
V^*(s_0) - V^{\pi^i}(s_0) &\le V^*(s_0) - \hat{V}^{\pi^i}(s_0) + O(\epsilon_{\text{eval}}) \\
&\le V^*(s_0) - \hat{V}^{\pi^j}(s_0) + O(\epsilon_{\text{eval}}) \\
&\le V^*(s_0) - V^{\pi^j}(s_0) + O(\epsilon_{\text{eval}}) \\
&\le O\left(\epsilon_{\text{eval}} + \epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right).
\end{aligned}$$

Since we set $\epsilon_{\text{ksp}}, \epsilon_{\text{ucb}}, \epsilon_{\text{eval}} = O(\epsilon)$, we conclude that with probability $1-\delta$, MDP-Full returns an ϵ-optimal policy.

Next, we calculate the sum of episodes we used. Each time we run RFKSP in MDP-Full with $\delta_{\text{ksp}} = \frac{1}{4}$ and $\epsilon_{\text{ksp}} = O(\epsilon)$, we use

$$K = O\left(\frac{S^9 A^3 \iota_{\text{ksp}}^2}{\epsilon_{\text{ksp}}}\text{polylog}\left(S, A, \frac{1}{\epsilon_{\text{ksp}}}\right)\right) = O\left(\frac{S^9 A^3}{\epsilon}\text{polylog}\left(S, A, \frac{1}{\epsilon}\right)\right)$$

episodes. Each time we run MDP-RBUCBI, we use $K = \tilde{O}\left(\frac{S^2 A}{\epsilon_{\text{ucb}}^2}\right)$ episodes. Each time we run MDP-Evaluation, we use

$$K = O\left(\frac{S^9 A^3 \iota_{\text{eval}}}{\epsilon_{\text{eval}}}\text{polylog}\left(S, A, \frac{1}{\epsilon_{\text{eval}}}\right)\right) + O\left(\frac{S^2 A \iota_{\text{eval}}}{\epsilon_{\text{eval}}^2}\text{polylog}\left(S, A, \frac{1}{\epsilon_{\text{eval}}}\right)\right)$$

episodes(See discussion under MDP-Evaluation(Algorithm 7)). We run ι_0 times RFKSP, MDP-RBUCBI, and MDP-Evaluation in MDP-Full. Summing up, we use

$$K = O\left(\frac{S^9 A^3 \iota_0^2}{\epsilon}\text{polylog}\left(S, A, \frac{1}{\epsilon}\right)\right) + O\left(\frac{S^2 A \iota_0^2}{\epsilon^2}\text{polylog}\left(S, A, \frac{1}{\epsilon}\right)\right)$$

episodes in total in MDP-Full. □

E.1. MDP-RBUCBI

In this section, we give proof to Lemma 6.5. We first prove some auxiliary lemmas that will be of use.

Lemma E.2. *In MDP-RBUCBI (Algorithm 2), for $\forall k \in [K]$, if $\tilde{P}_{s,a} \in \mathcal{P}^k_{s,a}$ and $\mathbb{E}\left[\tilde{R}(s,a)\right] \in \mathcal{R}^k_{s,a}$ holds for any (s,a), for $\forall h \in [H]$ and $\forall h$-reachable state s_h, $\overline{V}^k_h(s_h) \geq \tilde{V}^*_h(s_h)$.*

Note: We want to mention that MDP-RBUCBI interacts with the auxiliary Markovian environment instead of the original MDP. Thus $\tilde{P}$ and $\tilde{R}$ are the true transition probability and the reward function that generate the collected trajectory.

Proof of Lemma E.2. For fixed k, we do induction on $h = H+1, H, \ldots, 1$. When $h = H+1$,

$$\overline{V}^k_{H+1}(s_{H+1}) = \tilde{V}^*_{H+1}(s_{H+1}) = 0.$$

Suppose the target equation holds for $h+1$, then for $\forall a$,

$$\begin{aligned}
\overline{Q}^k_h(s_h,a) &= \min\left(\overline{r}^k(s_h,a) + \max_{\overline{p}\in\mathcal{P}^k_{s_h,a}} \overline{p}\,\overline{V}^k_{h+1}, 1\right) \\
&\geq \min\left(\mathbb{E}\left[\tilde{R}(s_h,a)\right] + \tilde{P}_{s_h,a}\overline{V}^k_{h+1}, 1\right) && (7)\\
&\geq \min\left(\mathbb{E}\left[\tilde{R}(s_h,a)\right] + \tilde{P}_{s_h,a}\tilde{V}^*_{h+1}, 1\right) && (8)\\
&= \tilde{Q}^*_h(s_h,a). && (9)
\end{aligned}$$

Here line 7 holds since we assume $\tilde{P}_{s_h,a} \in \mathcal{P}^k_{s_h,a}$ and $\mathbb{E}\left[\tilde{R}(s_h,a)\right] \in \mathcal{R}^k_{s_h,a}$. If $\tilde{P}_{s_h,a,s_{h+1}} \neq 0$, s_{h+1} is $h+1$-reachable as long as s_h is h-reachable. So by our induction, $\overline{V}^k_{h+1}(s_{h+1}) \geq \tilde{V}^*_{h+1}(s_{h+1})$ and thus line 8 also holds. As for line 9, $\tilde{Q}^*_h(s_h,a) = \mathbb{E}\left[\tilde{R}(s_h,a)\right] + \tilde{P}_{s_h,a}\tilde{V}^*_{h+1}$ by definition. We are left to prove $\tilde{Q}^*_h(s_h,a) \leq 1$. We can take π' where $\pi' = \pi^*$ except $\pi'_h(s_h) = a$. Hence by our reward assumption, $\tilde{Q}^*_h(s_h,a) = \tilde{V}^{\pi'}_h(s_h) \leq 1$. We further conclude our induction by

$$\overline{V}^k_h(s_h) = \max_a \overline{Q}^k_h(s_h,a) \geq \mathbb{E}_{a\sim\pi^*_h(s_h)}\overline{Q}^k_h(s_h,a) \geq \mathbb{E}_{a\sim\pi^*_h(s_h)}\tilde{Q}^*_h(s_h,a) = \tilde{V}^*_h(s_h).$$

$\square$

The following lemma bound the expectation of regret while interacting with the auxiliary Markovian environment. It is the most critical lemma in our paper.

Lemma E.3 (Restatement of Lemma 6.8). *In MDP-RBUCBI (Algorithm 2), the expectation of regret concerning the auxiliary Markovian environment can be bounded by*

$$\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K\left[\tilde{V}^*_1(s_0) - \tilde{V}^{\pi^k}_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \leq O(S\sqrt{AK}\mathrm{polylog}(S,A,K)\iota^2_{\mathrm{conf}}).$$

Proof of Lemma 6.8. Recall the definition of G_K, the preconditions in Lemma E.2 hold once G_K holds. From Lemma E.2 we have that

$$\begin{aligned}
\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K\left[\tilde{V}^*_1(s_0) - \tilde{V}^{\pi^k}_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} &\leq \mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K\left[\overline{V}^k_1(s_0) - \tilde{V}^{\pi^k}_1(s_0)\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&= \sum_{k=1}^K \mathbb{E}_{\Gamma_{k-1}}\left\{\left[\overline{V}^k_1(s_0) - \tilde{V}^{\pi^k}_1(s_0)\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\}. && (10)
\end{aligned}$$

For a single episode k, using the definition of $\overline{V}$ and $\tilde{V}$, we can turn the difference of $\overline{V}^k_1(s_0)$ and $\tilde{V}^{\pi^k}_1(s_0)$ into the expectation form of some term on the trajectory of the k-th episode. Here we introduce the cut-off indicator I^k_h into the equation. We mention again that by definition, I_1^k is always 1. And if $I_h^k = 0$, $I_{h'}^k = 0$ for any $h' > h$.

$$
\begin{aligned}
&\overline{V}_1^k\left(s_1^k\right) - \tilde{V}_1^{\pi^k}\left(s_1^k\right) = \overline{V}_1^k\left(s_1^k\right) I_1^k - \tilde{V}_1^{\pi^k}\left(s_1^k\right) I_1^k \\
&\leq \mathbb{E}_{a_1^k\sim\pi_1^k}\left[\left(\overline{r}^k\left(s_1^k, a_1^k\right) + \max_{\overline{p}\in\mathcal{P}^k_{s_1^k,a_1^k}} \overline{p}\overline{V}_2^k\right) I_1^k\right] - \mathbb{E}_{a_1^k\sim\pi_1^k}\left[\mathbb{E}R\left(s_1^k, a_1^k\right) + \tilde{P}_{s_1^k,a_1^k}\tilde{V}_2^{\pi^k}\right] I_1^k \\
&= \mathbb{E}_{a_1^k\sim\pi_1^k}\left[\left(\max_{\overline{p}\in\mathcal{P}^k_{s_1^k,a_1^k}} \overline{p} - \tilde{P}_{s_1^k,a_1^k}\right)\overline{V}_2^k I_1^k + \mathbb{E}_{s_2^k\sim\tilde{P}_{s_1^k,a_1^k}}(\overline{V}_2^k(s_2^k) - \tilde{V}_2^{\pi^k}(s_2^k)) I_1^k\right] \\
&+ \mathbb{E}_{a_1^k\sim\pi_1^k}\left[\left(\overline{r}^k(s_1^k, a_1^k) - \mathbb{E}R(s_1^k, a_1^k)\right) I_1^k\right] \\
&\leq \mathbb{E}_{a_1^k\sim\pi_1^k}\left[\left(\max_{\overline{p}\in\mathcal{P}^k_{s_1^k,a_1^k}} \overline{p} - \tilde{P}_{s_1^k,a_1^k}\right)\overline{V}_2^k I_1^k + \mathbb{E}_{s_2^k\sim\tilde{P}_{s_1^k,a_1^k}}(\overline{V}_2^k(s_2^k) - \tilde{V}_2^{\pi^k}(s_2^k)) I_2^k\right] + 2\mathbb{E}_{a_1^k\sim\pi_1^k}(I_1^k - I_2^k) \\
&+ \mathbb{E}_{a_1^k\sim\pi_1^k}\left[\left(\overline{r}^k(s_1^k, a_1^k) - \mathbb{E}R(s_1^k, a_1^k)\right) I_1^k\right] \leq \dots \\
&\leq \mathbb{E}_{\gamma_k}\left[\sum_{h=1}^H\left(\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p} - \tilde{P}_{s_h^k,a_h^k}\right)\overline{V}_{h+1}^k I_h^k\right] + \mathbb{E}_{\gamma_k}\left[\sum_{h=1}^H\left(\overline{r}^k(s_h^k, a_h^k) - \mathbb{E}R(s_h^k, a_h^k)\right) I_h^k\right] + 2\mathbb{E}_{\gamma_k}\left[\mathbf{1}_{I_{H+1}^k=0}\right].
\end{aligned}
$$

Substituting the above term back to line 10, we can arrange our target equation into the following form.

$$
\begin{aligned}
\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K\left[\tilde{V}_1^*(s_0) - V_1^{\pi^k}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \leq &\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\left[\sum_{h=1}^H\left(\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p} - \tilde{P}_{s_h^k,a_h^k}\right)\overline{V}_{h+1}^k I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&+ \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\left[\sum_{h=1}^H\left(\overline{r}^k(s_h^k, a_h^k) - \mathbb{E}R(s_h^k, a_h^k)\right) I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&+ 2\mathbb{E}_{\Gamma_K}\left[\left(\sum_{k=1}^K \mathbf{1}_{I_{H+1}^k=0}\right)\mathbf{1}_{\mathrm{G}_0}\right].
\end{aligned}
$$

For simplicity, we use the following notations. We denote

$$
\begin{aligned}
M_1 &= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\left[\sum_{h=1}^H\left(\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p} - \tilde{P}_{s_h^k,a_h^k}\right)\overline{V}_{h+1}^k I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\}, \\
M_2 &= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\left[\sum_{h=1}^H\left(\overline{r}^k(s_h^k, a_h^k) - \mathbb{E}R(s_h^k, a_h^k)\right) I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\}, \\
M_3 &= \mathbb{E}_{\Gamma_K}\left[\left(\sum_{k=1}^K \mathbf{1}_{I_{H+1}^k=0}\right)\mathbf{1}_{\mathrm{G}_0}\right].
\end{aligned}
$$

Thus our target equation turns into

$$
\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K\left[\tilde{V}_1^*(s_0) - V_1^{\pi^k}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \leq M_1 + M_2 + 2M_3. \tag{11}
$$

With the help of Lemma E.4, we have that

$$
\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K\left[\tilde{V}_1^*(s_0) - V_1^{\pi^k}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \leq O(S\sqrt{AK}\mathrm{polylog}(S, A, K)\iota_{\mathrm{conf}}^2),
$$

holds if $K \geq \Omega(S^2 A)$. □

Lemma E.4. *In Lemma 6.8,*

$$M_1 \le O(S\sqrt{AK}\text{polylog}(S,A,K)\iota_{\text{conf}}^2),$$
$$M_2 \le O(\sqrt{SAK}\text{polylog}(S,A,K)\iota_{\text{conf}}^2),$$
$$M_3 \le O(SA\text{polylog}(S,A,K)\iota_{\text{conf}}).$$

We further denote $M_4 = \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left[\sum_{h=1}^H \frac{I_h^k}{N^k\left(s_h^k,a_h^k\right)}\mathbf{1}_{\mathrm{G}_0}\right]$. *And we have that*

$$M_4 \le O(SA\text{polylog}(S,A,K)\iota_{\text{conf}}).$$

Proof. To begin with, M_3 and M_4 can be directly bounded by Lemma D.2 and Lemma 6.7.

For M_2,

$$\begin{aligned}
M_2 &= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\left[\sum_{h=1}^H \left(\overline{r}^k(s_h^k,a_h^k) - \mathbb{E}R(s_h^k,a_h^k)\right) I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&\le C\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\left[\sum_{h=1}^H\left(\sqrt{\frac{\hat{V}^k(s_h^k,a_h^k)\iota_{\text{conf}}}{N^k(s_h^k,a_h^k)}} + \frac{I_h^k\iota_{\text{conf}}}{N^k(s_h^k,a_h^k)}\right)\mathbf{1}_{\mathrm{G}_{k-1}}\right]\right\} \\
&\le C\sqrt{M_4\iota_{\text{conf}}}\cdot\sqrt{\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\sum_{h=1}^H \hat{V}^k(s_h^k,a_h^k)} + CM_4\iota_{\text{conf}}.
\end{aligned}$$

Here $\mathbb{E}\hat{V}^k(s,a) = \mathbb{E}\frac{1}{N^k(s,a)}\sum_{i=1}^{N^k(s,a)}(r^i(s,a) - \hat{r}^k(s,a))^2 \le \mathbb{E}r(s,a)^2 \le \mathbb{E}\left|r(s,a)\right|$. Thus if $K \ge \Omega(SA)$,

$$M_2 \le C\sqrt{M_4\iota_{\text{conf}}}\cdot\sqrt{\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\sum_{h=1}^H \left|r(s_h^k,a_h^k)\right|} + CM_4\iota_{\text{conf}} \le O(\sqrt{SAK}\text{polylog}(S,A,K)\iota_{\text{conf}}^2).$$

For M_1,

$$\begin{aligned}
M_1 &= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left[\left[\sum_{h=1}^H\left(\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}}\overline{p} - \tilde{P}_{s_h^k,a_h^k}\right)\overline{V}_{h+1}^k I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right] \\
&= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left[\left[\sum_{h=1}^H\left(\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}}\overline{p} - \tilde{P}_{s_h^k,a_h^k}\right)\left(\overline{V}_{h+1}^k - \tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right) I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right] \\
&\le C\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left[\left[\sum_{h=1}^H\sum_{s'\in\mathcal{S}}\left(\sqrt{\frac{\tilde{P}_{s_h^k,a_h^k,s'}\iota_{\text{conf}}}{N^k\left(s_h^k,a_h^k\right)}} + \frac{\iota_{\text{conf}}}{N^k\left(s_h^k,a_h^k\right)}\right)\left|\overline{V}_{h+1}^k(s') - \tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right| I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right] \\
&\le C\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left[\left[\sum_{h=1}^H\sqrt{\frac{SI_h^k\iota_{\text{conf}}}{N^k\left(s_h^k,a_h^k\right)}}\cdot\sqrt{V\left(\tilde{P}_{s_h^k,a_h^k},\overline{V}_{h+1}^k\right)I_h^k} + \frac{2S\iota_{\text{conf}}I_h^k}{N^k\left(s_h^k,a_h^k\right)}\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right] \\
&\le C\sqrt{SM_4\iota_{\text{conf}}}\cdot\sqrt{\mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H\left[V\left(\tilde{P}_{s_h^k,a_h^k},\overline{V}_{h+1}^k\right)I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\}} + 2CSM_4\iota_{\text{conf}}.
\end{aligned}$$

We further denote

$$M_5 = \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H\left[V\left(\tilde{P}_{s_h^k,a_h^k},\overline{V}_{h+1}^k\right)I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\}.$$

By the above notations, we have $M_1 \leq C\sqrt{SM_4\iota_{\text{conf}}} \cdot \sqrt{M_5} + 2C \cdot M_4 S\iota_{\text{conf}}$. Next, we try to bound M_5 with M_1 and thus construct a recursion structure.

$$
\begin{aligned}
M_5 &= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[V\left(\tilde{P}_{s_h^k,a_h^k}, \overline{V}_{h+1}^k\right) I_h^k\right]\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[\tilde{P}_{s_h^k,a_h^k}\left(\overline{V}_{h+1}^k\right)^2 - \left(\tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right)^2\right] I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&= \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[\overline{V}_{h+1}^k\left(s_{h+1}^k\right)^2 - \left(\tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right)^2\right] I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&\leq \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[\overline{V}_h^k(s_h^k)^2 - \left(\tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right)^2\right] I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} + \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left[\left(\mathbf{1}_{I_{H+1}^k=0}\right)\mathbf{1}_{\mathrm{G}_{k-1}}\right] && (12) \\
&\leq \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[\overline{V}_h^k(s_h^k)^2 - \left(\tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right)^2\right] I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} + M_3. && (13)
\end{aligned}
$$

In line 12, we use the fact that

$$
\sum_{h=1}^H \overline{V}_{h+1}^k\left(s_{h+1}^k\right)^2 I_h^k \leq \sum_{h=1}^H \overline{V}_h^k\left(s_h^k\right)^2 I_h^k + \mathbf{1}_{I_{H+1}^k=0}. \tag{14}
$$

In particular, When $I_{H+1}^k = 1$, $I_h^k = 1$ holds for $\forall k \in [K]$. Note that $\overline{V}_{H+1}^k = 0$, equation 14 holds. When $I_{H+1}^k = 0$, there exists $h \in [H]$ such that $I_t^k = 1, \forall t \in [h-1]$ and $I_t^k = 0, \forall t \in [h, H]$. Hence equation 14 holds by

$$
\sum_{h=1}^H \overline{V}_{h+1}^k\left(s_{h+1}^k\right)^2 I_h^k - \sum_{h=1}^H \overline{V}_{h+1}^k\left(s_h^k\right)^2 I_h^k = \overline{V}_1^k\left(s_1^k\right)^2 - \overline{V}_h^k\left(s_h^k\right)^2 \leq 1.
$$

By the definition of $\overline{V}_h^k\left(s_h^k\right)$, we have

$$
\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\sum_{h=1}^H \overline{V}_h^k(s_h^k)^2 I_h^k\mathbf{1}_{\mathrm{G}_{k-1}} \leq \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\sum_{h=1}^H \left(\overline{r}^k\left(s_h^k,a_h^k\right) + \max_{\overline{p}\in\mathcal{P}_{s_h^k,a_h^k}^k} \overline{p}\,\overline{V}_{h+1}^k\right)^2 I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}. \tag{15}
$$

Here the equation 15 holds since a_h^k is fixed given s_h^k and policy π^k. This is different from the MG setting. We will mention it again in the proof for MG. (See Lemma F.3.)

Substituting equation 15 into line 13, we derive the recursive structure for M_1. Here we use the square difference formula in line 16. Applying our new reward assumption to line 17 leads to line 18.

$$
\begin{aligned}
M_5 &\leq \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left(\overline{V}_h^k(s_h^k)^2 - \left(\tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right)^2\right) I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} + M_3 \\
&\leq \sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[\left(\overline{r}^k\left(s_h^k,a_h^k\right) + \max_{\overline{p}\in\mathcal{P}_{s_h^k,a_h^k}^k}\overline{p}\,\overline{V}_{h+1}^k\right)^2 - \left(\mathbb{E}R\left(s_h^k,a_h^k\right) + \max_{\overline{p}\in\mathcal{P}_{s_h^k,a_h^k}^k}\overline{p}\,\overline{V}_{h+1}^k\right)^2\right] I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} \\
&\quad + \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[\left(\mathbb{E}R\left(s_h^k,a_h^k\right) + \max_{\overline{p}\in\mathcal{P}_{s_h^k,a_h^k}^k}\overline{p}\,\overline{V}_{h+1}^k\right)^2 - \left(\tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right)^2\right] I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} + M_3 && (16) \\
&\leq 4M_2 + 3\sum_{k=1}^K \mathbb{E}_{\Gamma_k}\left\{\sum_{h=1}^H \left[\left|\mathbb{E}R\left(s_h^k,a_h^k\right)\right| + \left(\max_{\overline{p}\in\mathcal{P}_{s_h^k,a_h^k}^k}\overline{p}\,\overline{V}_{h+1}^k - \tilde{P}_{s_h^k,a_h^k}\overline{V}_{h+1}^k\right)\right] I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\} + M_3 && (17) \\
&\leq 4M_2 + 3K + 3M_1 + M_3. && (18)
\end{aligned}
$$

To be more specific, we derive line 17 as follows. Since I_h^k and Γ_{k-1} are independent of a_h^k, we can focus on the difference of squares. For the second difference of squares,

$$\begin{aligned}
&\left(\mathbb{E}R\left(s_h^k, a_h^k\right) + \max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p}\overline{V}^k_{h+1}\right)^2 - \left(\tilde{P}_{s_h^k,a_h^k}\overline{V}^k_{h+1}\right)^2 \\
&= \left(\mathbb{E}R\left(s_h^k, a_h^k\right) + \max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p}\overline{V}^k_{h+1} + \tilde{P}_{s_h^k,a_h^k}\overline{V}^k_{h+1}\right)\cdot\left(\mathbb{E}R\left(s_h^k, a_h^k\right) + \max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p}\overline{V}^k_{h+1} - \tilde{P}_{s_h^k,a_h^k}\overline{V}^k_{h+1}\right) \\
&\le \left(\left|\mathbb{E}R\left(s_h^k, a_h^k\right)\right| + \left|\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p}\overline{V}^k_{h+1}\right| + \left|\tilde{P}_{s_h^k,a_h^k}\overline{V}^k_{h+1}\right|\right)\cdot\left(\left|\mathbb{E}R\left(s_h^k, a_h^k\right)\right| + \max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p}\overline{V}^k_{h+1} - \tilde{P}_{s_h^k,a_h^k}\overline{V}^k_{h+1}\right) \\
&\le 3\left[\left|\mathbb{E}R\left(s_h^k, a_h^k\right)\right| + \left(\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}} \overline{p}\overline{V}^k_{h+1} - \tilde{P}_{s_h^k,a_h^k}\overline{V}^k_{h+1}\right)\right].
\end{aligned}$$

Combining line 18 and $M_1 \le C\sqrt{SM_4\iota_{\mathrm{conf}}}\cdot\sqrt{M_5} + 2C\cdot M_4S\iota_{\mathrm{conf}}$, we can solve M_1 satisfies that

$$\begin{aligned}
M_1 &\le O((S\sqrt{A}\sqrt{K} + S^2A)\mathrm{polylog}(S,A,K)\iota^2_{\mathrm{conf}}) \\
&\le O(S\sqrt{A}\sqrt{K}\mathrm{polylog}(S,A,K)\iota^2_{\mathrm{conf}}) \quad \text{If } K \ge \Omega(S^2A).
\end{aligned}$$

□

Theorem E.5 (**Restatement of Lemma 6.5**). *The policy π returned by MDP-RBUCBI (Algorithm 2) satisfies that with probability $\frac{1}{2}$,*

$$V_1^*(s_0) - V_1^\pi(s_0) \le O\left(\epsilon_{\mathrm{ucb}} + \epsilon_{\mathrm{ksp}}\right).$$

Proof of Lemma 6.5. We randomly choose $k_1 \in [K]$. Since $\mathbb{E}_{\Gamma_K}\left\{\left[\tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^k}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \ge 0$ hold for $\forall k \in [K]$, by Markov inequality and Lemma 6.8, the following holds with probability at least $\frac{1}{16}$.

$$\mathbb{E}_{\Gamma_K}\left\{\left[\tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^k}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \le 8\left[O\left(\frac{S\sqrt{A}}{\sqrt{K}}\mathrm{polylog}(S,A,K)\iota^2_{\mathrm{conf}}\right)\right]. \tag{19}$$

Since $\left[\tilde{V}^*(s_0) - \tilde{V}^{\pi^{k_1}}(s_0)\right]\mathbf{1}_{\mathrm{G}_K} \ge 0$ also always hold, we apply Markov inequality and derive that with probability at least $\frac{1}{16}$,

$$\left[\tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^{k_1}}(s_0)\right]\mathbf{1}_{\mathrm{G}_K} \le 8\mathbb{E}_{\Gamma_K}\left\{\left[\tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^{k_1}}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\}. \tag{20}$$

From Lemma D.4 and the definition of G_K we know that G_K holds with probability at least $1 - \delta_{\mathrm{ksp}} - 2S^2AK\delta_{\mathrm{conf}}$. Since $\delta_{\mathrm{ksp}} = \frac{1}{4}$ in MDP-RBUCBI, by further setting $\delta_{\mathrm{conf}} = \frac{1}{16S^2AK}$, we have that with probability at least $\frac{5}{8}$, G_K holds. By taking union bound with equation 19 and equation 20, we have that with probability at least $\frac{1}{2}$,

$$\begin{aligned}
\tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^{k_1}}(s_0) &= \left[\tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^{k_1}}(s_0)\right]\mathbf{1}_{\mathrm{G}_K} \qquad (21)\\
&\le O\left(\frac{S\sqrt{A}}{\sqrt{K}}\mathrm{polylog}(S,A,K)\iota^2_{\mathrm{conf}}\right) \\
&= O\left(\frac{S\sqrt{A}}{\sqrt{K}}\mathrm{polylog}(S,A,K)\right). \qquad (22)
\end{aligned}$$

Here line 21 holds since G_K and line 22 holds since we have set $\delta_{\mathrm{conf}} = \frac{1}{16S^2AK}$. We can further set $K = \tilde{O}\left(\frac{S^2A}{\epsilon_{\mathrm{ucb}}^2}\right)$ which satisfies $O\left(\frac{S\sqrt{A}}{\sqrt{K}}\mathrm{polylog}(S, A, K)\right) = \epsilon_{\mathrm{ucb}}$. Therefore

$$\tilde{V}^*(s_0) - \tilde{V}^{\pi^{k_1}}(s_0) \leq O(\epsilon_{\mathrm{ucb}}).$$

Since G_0 holds, by Lemma 6.6 we have that $\left|V_1^\pi(s_0) - \tilde{V}_1^\pi(s_0)\right| \leq \epsilon_{\mathrm{ksp}}$ holds for any π. Here we use $\tilde{V}^{**}$ to denote the value function of the best policy for the original model in the auxiliary Markovian environment. Note that by definition, $\tilde{V}_1^*(s_0) \geq \tilde{V}_1^{**}(s_0)$.

$$\begin{aligned} V_1^*(s_0) - V_1^{\pi^{k_1}}(s_0) &\leq \tilde{V}_1^{**}(s_0) - \tilde{V}_1^{\pi^{k_1}}(s_0) + 2\epsilon_{\mathrm{ksp}} \\ &\leq \tilde{V}_1^*(s_0) - \tilde{V}_1^{\pi^{k_1}}(s_0) + 2\epsilon_{\mathrm{ksp}} \\ &\leq O\left(\epsilon_{\mathrm{ucb}} + \epsilon_{\mathrm{ksp}}\right). \end{aligned}$$

□

E.2. MDP-Evaluation

In this section, we present MDP-Evaluation(Algorithm 7) and prove Lemma 6.4.

Algorithm 7 MDP-Evaluation

```
Input: MDP $M$, Policy $\pi$, $\epsilon_{\mathrm{eval}}$, $\delta_{\mathrm{eval}}$.
Initialization: $\overline{V}_{H+1}^k(s) = 0, \underline{V}_{H+1}^k(s) = 0, \forall k, s$.
Set $\epsilon_{\mathrm{ksp}} \leftarrow \epsilon_{\mathrm{eval}}$, $\delta_{\mathrm{ksp}} \leftarrow \frac{\delta_{\mathrm{eval}}}{2T}$.
Run $T = \log\left(\frac{2}{\delta_{\mathrm{eval}}}\right)$ times independently.
for $t = 1, 2, \ldots, T$ do
  $\tilde{M}_t \leftarrow \mathrm{RFKSP}(M, \epsilon_{\mathrm{ksp}}, \delta_{\mathrm{ksp}})$.
  Use $K = \tilde{O}\left(\frac{S^2A}{\epsilon_{\mathrm{eval}}^2}\right)$ episodes.
  for episode $k = 1, 2, \ldots, K$ do
    for step $h = H, H-1, H-2, \ldots, 1$ do
      Compute $\overline{Q}_h^k(s,a)$ as in equation 1.
      Compute $\overline{V}_h^k(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\overline{Q}_h^k(s,a)$.
    end for
    Play policy $\pi$, collect trajectory $\tau_k$.
    Calculate $\mathcal{P}^{k+1}, \mathcal{R}^{k+1}$ based on $\Gamma_k$.
  end for
  Randomly select $\overline{V}_1^k(s_0)$, denote as $\overline{V}_t(s_0)$.
end for
Output: $\hat{V}_1^\pi(s_0) = \min\left\{\overline{V}_1(s_0), \ldots, \overline{V}_T(s_0)\right\}$.
```

Note: Each time we run MDP-Evaluation, we run the subroutine in it for $T = \log(\frac{2}{\delta_{\mathrm{eval}}})$ times independently. In particular, each time we run RFKSP with $\epsilon_{\mathrm{ksp}} = \epsilon_{\mathrm{eval}}$ and $\delta_{\mathrm{ksp}} = \frac{\delta_{\mathrm{eval}}}{2T}$, we use

$$K = O\left(\frac{S^9A^3\iota_{\mathrm{ksp}}}{\epsilon_{\mathrm{ksp}}}\mathrm{polylog}\left(S, A, \frac{1}{\epsilon_{\mathrm{ksp}}}\right)\right) = O\left(\frac{S^9A^3\iota_{\mathrm{eval}}}{\epsilon_{\mathrm{eval}}}\mathrm{polylog}\left(S, A, \frac{1}{\epsilon_{\mathrm{eval}}}\right)\right)$$

episodes. Here we use the fact that $\log\left(\frac{\iota_{\mathrm{eval}}}{\delta_{\mathrm{eval}}}\right) \leq O\left(\iota_{\mathrm{eval}}\right)$. The total number of episodes used in MDP-Evaluation is

$$K = O\left(\frac{S^9A^3\iota_{\mathrm{eval}}}{\epsilon_{\mathrm{eval}}}\mathrm{polylog}\left(S, A, \frac{1}{\epsilon_{\mathrm{eval}}}\right)\right) + O\left(\frac{S^2A\iota_{\mathrm{eval}}}{\epsilon_{\mathrm{eval}}^2}\mathrm{polylog}\left(S, A, \frac{1}{\epsilon_{\mathrm{eval}}}\right)\right).$$

Lemma E.6. *In MDP-Evaluation (Algorithm 7), for $\forall k \in [K]$, if $\tilde{P}_{s,a} \in \mathcal{P}^k_{s,a}$ and $\mathbb{E}\left[\tilde{R}(s,a)\right] \in \mathcal{R}^k_{s,a}$ holds for any (s,a), for $\forall h \in [H]$ and $\forall h$-reachable state s_h, $\overline{V}^k_h(s_h) \geq \tilde{V}^\pi_h(s_h)$.*

Proof. For fixed k, we do induction on $h = H+1, H, \ldots, 1$. When $h = H+1$,

$$\overline{V}^k_{H+1}(s_{H+1}) = \tilde{V}^\pi_{H+1}(s_{H+1}) = 0.$$

Suppose the equation holds for $h+1$, then for $\forall a$,

$$\begin{aligned}
\overline{Q}^k_h(s_h,a) &= \min\left(\overline{r}^k(s_h,a) + \max_{\overline{p} \in \mathcal{P}^k_{s_h,a}} \overline{p}\overline{V}^k_{h+1}, 1\right) \\
&\geq \min\left(\mathbb{E}\tilde{R}(s_h,a) + \tilde{P}_{s_h,a}\overline{V}^k_{h+1}, 1\right) &(23)\\
&\geq \min\left(\mathbb{E}\tilde{R}(s_h,a) + \tilde{P}_{s_h,a}\tilde{V}^\pi_{h+1}, 1\right) &(24)\\
&= \tilde{Q}^\pi_h(s_h,a). &(25)
\end{aligned}$$

Here line 23 holds since we assume $\tilde{P}_{s_h,a} \in \mathcal{P}^k_{s_h,a}$. And if $\tilde{P}_{s_h,a,s_{h+1}} \neq 0$, then s_{h+1} is $h+1$-th reachable. So by our induction, $\overline{V}^k_{h+1}(s_{h+1}) \geq \tilde{V}^\pi_{h+1}(s_{h+1})$. Thus line 24 also holds. As for line 25, by definition we have $\tilde{Q}^\pi_h(s_h,a) = \mathbb{E}\tilde{R}(s_h,a) + \tilde{P}_{s_h,a}\tilde{V}^\pi_{h+1}$. We are left to prove that $\tilde{Q}^\pi_h(s_h,a) \leq 1$. We can take π' where $\pi' = \pi$ except $\pi'_h(s_h) = a$. Hence by our reward assumption, $\tilde{Q}^*_h(s_h,a) = \tilde{V}^\pi_h(s_h) \leq 1$. We further conclude our induction by

$$\overline{V}^k_h(s_h) = \mathbb{E}_{a\sim\pi_h(s_h)}\overline{Q}^k_h(s_h,a) \geq \mathbb{E}_{a\sim\pi_h(s_h)}\tilde{Q}^\pi_h(s_h,a) = \tilde{V}^\pi_h(s_h).$$

□

Lemma E.7. *In each independent running time $t \in [T]$ in MDP-Evaluation (Algorithm 7),*

$$\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K \left[\overline{V}^k_1(s_0) - \tilde{V}^\pi_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \leq O(S\sqrt{AK}\mathrm{polylog}(S,A,K)\iota^2_{\mathrm{conf}}).$$

Proof. This proof is similar to the proof of Lemma 6.8. The only difference is that we run one policy throughout this procedure and overestimate it here. □

Lemma E.8. *In each independent running time $t \in [T]$ in MDP-Evaluation (Algorithm 7), the returned estimation $\overline{V}_t(s_0)$ satisfies that with probability $\frac{1}{2}$,*

$$0 \leq \left[\overline{V}_t(s_0) - \tilde{V}^\pi_1(s_0)\right] \leq O(\epsilon_{\mathrm{eval}}).$$

Proof. We focus on a fixed independent running time $t \in [T]$. From Lemma E.7 we have that

$$\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^K \left[\overline{V}^k_1(s_0) - \tilde{V}^\pi_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} \leq O(S\sqrt{AK}\mathrm{polylog}(S,A,K)\iota_{\mathrm{conf}}).$$

By Lemma E.6, we know that $\left[\overline{V}^k_1(s_0) - \tilde{V}^\pi_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K} \geq 0$. And therefore $\mathbb{E}_{\Gamma_K}\left\{\left[\overline{V}^k_1(s_0) - \tilde{V}^\pi_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\}$ is positive for any $k \in [K]$. We randomly choose episode $k_1 \in [K]$ and denote $\overline{V}^{k_1}_1(s_0)$ as $\overline{V}_t(s_0)$. Using Markov inequality twice and taking union bound, we have that with probability at least $\frac{7}{8}$,

$$\begin{aligned}
\mathbb{E}_{\Gamma_K}\left\{\left[\overline{V}^{k_1}_1(s_0) - \tilde{V}^\pi_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\} &\leq O(S\sqrt{AK}\mathrm{polylog}(S,A,K)\iota^2_{\mathrm{conf}}), &(26)\\
\left[\overline{V}^{k_1}_1(s_0) - \tilde{V}^\pi_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K} &\leq 16\mathbb{E}_{\Gamma_K}\left\{\left[\overline{V}^{k_1}_1(s_0) - \tilde{V}^\pi_1(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\}. &(27)
\end{aligned}$$

Since $\delta_{\rm ksp} = \frac{1}{4}$, by further setting $\delta_{\rm conf} = \frac{1}{16S^2AK}$, we have that with probability at least $\frac{5}{8} = 1 - 2S^2AK\delta_{\rm conf} - \delta_{\rm ksp}$, G_K holds. By taking union bound with equation 26 and equation 27, we have that with probability at least $\frac{1}{2}$,

$$\begin{aligned}\left[\overline{V}_1^{k_1}(s_0) - \tilde{V}_1^{\pi}(s_0)\right] &= \left[\overline{V}_1^{k_1}(s_0) - \tilde{V}_1^{\pi}(s_0)\right]\mathbf{1}_{\mathrm{G}_K} \qquad (28)\\ &\le O\left(\frac{S\sqrt{A}}{\sqrt{K}}\mathrm{polylog}(S,A,K)\iota_{\rm conf}^2\right)\\ &= O\left(\frac{S\sqrt{A}}{\sqrt{K}}\mathrm{polylog}(S,A,K)\right). \qquad (29)\end{aligned}$$

Here line 28 holds since G_K holds. Line 29 holds since we have set $\delta_{\rm conf} = \frac{1}{16S^2AK}$. We can further set $K = \tilde{O}\left(\frac{S^2A}{\epsilon_{\rm ucb}^2}\right)$ which satisfies $O\left(\frac{S\sqrt{A}}{\sqrt{K}}\mathrm{polylog}(S,A,K)\right) = \epsilon_{\rm eval}$. Therefore with probability at least $\frac{1}{2}$,

$$0 \le \left[\overline{V}_1^{k_1}(s_0) - \tilde{V}_1^{\pi}(s_0)\right] = \left[\overline{V}_1^{k_1}(s_0) - \tilde{V}_1^{\pi}(s_0)\right]\mathbf{1}_{\mathrm{G}_K} \le O(\epsilon_{\rm eval}).$$

□

Lemma E.9 (Restatement of Lemma 6.4). *The estimate $\hat{V}^{\pi}(s_0)$ returned by MDP-Evaluation satisfies that with probability* $1 - \delta_{\rm eval}$,

$$|\hat{V}_1^{\pi}(s_0) - V_1^{\pi}(s_0)| \le O(\epsilon_{\rm eval}).$$

Proof. By Lemma 6.4 we have that for any independent running time $t \in [T]$, with probability $\frac{1}{2}$,

$$0 \le \left[\overline{V}_t(s_0) - \tilde{V}_1^{\pi}(s_0)\right] \le O(\epsilon_{\rm eval}).$$

Since we run the algorithm $T = \log\left(\frac{2}{\delta_{\rm eval}}\right)$ times independently, with probability $1 - \frac{\delta_{\rm eval}}{2}$, there exists $i \in [T]$ that

$$0 \le \overline{V}_i(s_0) - \tilde{V}_1^{\pi}(s_0) \le O(\epsilon_{\rm eval}).$$

In MDP-Evaluation, we set $\delta_{\rm ksp} = \frac{\delta_{\rm eval}}{2T}$. Taking union bound, G_0 holds in any running time $t \in [T]$ with probability at least $1 - \frac{\delta_{\rm eval}}{2}$. Combining with Lemma 6.6, we have that with probability at least $1 - \delta_{\rm eval}$,

$$\begin{aligned}0 &\le \min_{t\in[T]} \overline{V}_t(s_0) - \tilde{V}_1^{\pi}(s_0)\\ &\le \min_{t\in[T]} \overline{V}_t(s_0) - V_1^{\pi}(s_0) + \epsilon_{\rm eval}\\ &\le \overline{V}_i - V_1^{\pi}(s_0) + \epsilon_{\rm eval}\\ &\le \overline{V}_i - \tilde{V}_1^{\pi}(s_0) + 2\epsilon_{\rm eval}\\ &\le O(\epsilon_{\rm eval}).\end{aligned}$$

We must mention that the auxiliary Markovian environment $\tilde{M}_t$ differs between different episodes. Here $\tilde{V}$ in the first line refers to the auxiliary Markovian environment in episode $\arg\min_{t\in[T]} \overline{V}_t(s_0)$ and the latter $\tilde{V}$ refers to the auxiliary Markovian environment in the i-th episode. Rearranging the above equation leads to

$$|\min_{t\in[T]} \overline{V}_t(s_0) - V^{\pi}(s_0)| \le O(\epsilon_{\rm eval}).$$

Here $\min_{t\in[T]} \overline{V}_t(s_0)$ is the returned value $\hat{V}^{\pi}(s_0)$ of MDP-Evaluation (Algorithm 7). □

F. Proofs for MG(Theorem 6.3)

In this section we give proofs and algorithms in MG setting. The main structure resembles the MDP.

Algorithm 8 MG-Full

1: **Input:** MG $\mathcal{G}\left(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, R, H, \mu_0\right)$, ϵ, δ.
2: Set $\epsilon_{\text{ksp}}, \epsilon_{\text{ucb}}, \epsilon_{\text{eval}} = O(\epsilon), \delta_{\text{ksp}} = \frac{1}{4}, \delta_{\text{eval}} = \frac{\delta}{4T}$.
3: Run $T = \log\left(\frac{2}{\delta}\right)$ times independently.
4: **for** $t = 1, 2, \ldots, T$ **do**
5: $\quad\tilde{\mathcal{G}}_t \leftarrow$ RFKSP $(\mathcal{G}, \epsilon_{\text{ksp}}, \delta_{\text{ksp}})$.
6: $\quad\pi^t = (\mu^t, \nu^t) \leftarrow$ MG-RBUCBI$(\tilde{\mathcal{G}}_t, \epsilon_{\text{ucb}})$.
7: $\quad\hat{V}_1^{*,\nu^t}(s_0), \hat{V}_1^{\mu^t,*}(s_0) \leftarrow$MG-Evaluation$(\mathcal{G}, \epsilon_{\text{eval}}, \delta_{\text{eval}}, \pi^t)$.
8: **end for**
9: $i \leftarrow \arg\min_{t\in[T]}\left[\hat{V}_1^{*,\nu^t}(s_0) - \hat{V}_1^{\mu^t,*}(s_0)\right]$.
10: **Output:** π^i.

Theorem F.1 (**Restatement of Theorem 6.3**). *For any $\epsilon, \delta > 0$, with probability $1-\delta$, MG-Full returns an ϵ-approximate NE policy pair by sampling at most $K = K_{\text{Reward}} + K_{\text{RFKSP}}$ episodes, where*

$$K_{\text{Reward}} = O\left(\frac{S^2 AB\iota^3}{\epsilon^2}\text{polylog}\left(S, A, B, \frac{1}{\epsilon}\right)\right),$$
$$K_{\text{RFKSP}} = O\left(\frac{S^9 A^3 B^3\iota^3}{\epsilon}\text{polylog}\left(S, A, B, \frac{1}{\epsilon}\right)\right).$$

Proof of Theorem 6.3. Theorem 6.3 is mainly based on Theorem F.4 and Theorem F.5. Given these two theorems, we derive Theorem 6.3 as follows. In each running time $t \in [T]$, by Theorem F.4 we have that with probability $1/2$,

$$0 \le V_1^{*,\hat{\nu}}(s_0) - V_1^{\hat{\mu},*}(s_0) \le O\left(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right).$$

As we run the subroutine $T = \log(\frac{2}{\delta})$ times independently, with probability $1 - \frac{\delta}{2}$, there exists $j \in [T]$ that the above equation holds. By Theorem F.5, the estimation $\hat{V}_1^{*,\nu^t}(s_0)$ and $\hat{V}_1^{\mu^t,*}(s_0)$ returned by MDP-Evaluation satisfy that with probability $1 - 2\delta_{\text{eval}}$,

$$\left|\hat{V}_1^{*,\nu^t}(s_0) - V_1^{*,\nu^t}(s_0)\right| \le O(\epsilon_{\text{eval}}).$$
$$\left|\hat{V}_1^{\mu^t,*}(s_0) - V_1^{\mu^t,*}(s_0)\right| \le O(\epsilon_{\text{eval}}).$$

Since we set $\delta_{\text{eval}} = \frac{\delta}{4T}$ in MG-Full, with probability $1 - \frac{\delta}{2}$, the above equation hold for $\forall t \in [T]$. Suppose we denote $i = \arg\min_{t\in[T]}\left[\hat{V}_1^{*,\nu^t}(s_0) - \hat{V}_1^{\mu^t,*}(s_0)\right]$. Taking union bound, we have that the following equation holds with probability $1 - \delta$.

$$\begin{aligned}
V_1^{*,\nu^i}(s_0) - V_1^{\mu^i,*}(s_0) &\le \hat{V}_1^{*,\nu^i}(s_0) - \hat{V}_1^{\mu^i,*}(s_0) + O(\epsilon_{\text{eval}}) \\
&\le \hat{V}_1^{*,\nu^j}(s_0) - \hat{V}_1^{\mu^j,*}(s_0) + O(\epsilon_{\text{eval}}) \\
&\le V_1^{*,\nu^j}(s_0) - V_1^{\mu^j,*}(s_0) + O(\epsilon_{\text{eval}}) \\
&\le O\left(\epsilon_{\text{eval}} + \epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right).
\end{aligned}$$

Since we set $\epsilon_{\text{ksp}}, \epsilon_{\text{ucb}}, \epsilon_{\text{eval}} = O(\epsilon)$, we conclude that with probability $1 - \delta$, MG-Full returns an ϵ-approximate NE policy pair. Each time we run RFKSP with $\delta_{\text{ksp}} = \frac{1}{4}$ and $\epsilon_{\text{ksp}} = O(\epsilon)$, we use

$$K = O\left(\frac{S^9 A^3 B^3 \iota_{\text{ksp}}^2}{\epsilon_{\text{ksp}}}\text{polylog}\left(S, A, B, \frac{1}{\epsilon_{\text{ksp}}}\right)\right) = O\left(\frac{S^9 A^3 B^3}{\epsilon}\text{polylog}\left(S, A, B, \frac{1}{\epsilon}\right)\right)$$

episodes. Each time we run MG-RBUCBI, we use $K = \tilde{O}\left(\frac{S^2AB}{\epsilon_{\text{ucb}}^2}\right)$ episodes. Each time we run MG-Evaluation, we use

$$K = O\left(\frac{S^9A^3B^3\iota_{\text{eval}}^2}{\epsilon_{\text{eval}}}\text{polylog}\left(S, A, \frac{1}{\epsilon_{\text{eval}}}\right)\right) + O\left(\frac{S^2AB\iota_{\text{eval}}^2}{\epsilon_{\text{eval}}^2}\text{polylog}\left(S, A, B, \frac{1}{\epsilon_{\text{eval}}}\right)\right)$$

episodes(See discussion under MDP-Evaluation(Algorithm 7)). Summing up, we have that we use

$$K = O\left(\frac{S^9A^3B^3\iota_0^3}{\epsilon}\text{polylog}\left(S, A, B, \frac{1}{\epsilon}\right)\right) + O\left(\frac{S^2AB\iota_0^3}{\epsilon^2}\text{polylog}\left(S, A, B, \frac{1}{\epsilon}\right)\right)$$

episodes in MG-Full. □

F.1. MG-RBUCBI

In this section, we prove the lemmas regarding MG-RBUCBI(Algorithm 3).

Lemma F.2. *In MG-RBUCBI (Algorithm 3), for $\forall k \in [K]$, if $\tilde{P}_{s,a,b} \in \mathcal{P}_{s,a,b}^k$ and $\mathbb{E}\left[\tilde{R}(s,a,b)\right] \in \mathcal{R}_{s,a,b}^k$ holds for any (s,a,b), for $\forall h \in [H]$ and $\forall h$-reachable state s_h,*

$$\overline{V}_h^k(s_h) \geq \tilde{V}_h^{*,\nu^k}(s_h) \geq \tilde{V}_h^{\pi^k}(s_h) \geq \tilde{V}_h^{\mu^k,*}(s_h) \geq \underline{V}_h^k(s_h).$$

Proof. In the above equation, $\tilde{V}_h^{*,\nu^k}(s_h) \geq \tilde{V}_h^{\pi^k}(s_h) \geq \tilde{V}_h^{\mu^k,*}(s_h)$ hold naturally by the definition. Here we only prove the overestimation while the underestimation is almost the same.

For fixed k, we do induction on $h = H+1, H, \ldots, 1$. When $h = H+1$, $\overline{V}_{H+1}^k(s_{H+1}) = \tilde{V}_{H+1}^{*,\nu^k}(s_{H+1}) = 0$. Suppose the equation holds for $h+1$, then for $\forall (a,b) \in \mathcal{A} \times \mathcal{B}$,

$$\begin{aligned}
\overline{Q}_h^k(s_h,a,b) &= \min\left(\overline{r}^k(s_h,a,b) + \max_{\overline{p} \in \mathcal{P}_{s_h,a,b}^k} \overline{p}\,\overline{V}_{h+1}^k, 1\right) \\
&\geq \min\left(\mathbb{E}R(s_h,a,b) + \tilde{P}_{s_h,a,b}\overline{V}_{h+1}^k, 1\right) && (30)\\
&\geq \min\left(\mathbb{E}R(s_h,a,b) + \tilde{P}_{s_h,a,b}\tilde{V}_{h+1}^{*,\nu^k}, 1\right) && (31)\\
&= Q_h^{*,\nu^k}(s_h,a,b). && (32)
\end{aligned}$$

Here 30 holds since we assume $\tilde{P}_{s_h,a,b} \in \mathcal{P}_{s_h,a,b}^k$ and $\mathbb{E}\left[\tilde{R}(s,a,b)\right] \in \mathcal{R}_{s,a,b}^k$. And if $\tilde{P}_{s_h,a,b,s_{h+1}} \neq 0$, s_{h+1} is $h+1$-th reachable. So by induction 31 also holds. As for 32, we can take π' where $(\mu',\nu') = (\mu^*,\nu^*)$ except $\mu_h'(s_h) = a, \nu_h'(s_h) = b$. Hence by our reward assumption, $Q_h^*(s_h,a,b) = \tilde{V}_h^{\mu',\nu'}(s_h) \leq 1$. We further conclude our induction by

$$\begin{aligned}
\overline{V}_h^k(s_h) &= \mathbb{E}_{a\sim\mu_h^k(\cdot|s_h), b\sim\nu_h^k(\cdot|s_h)}\overline{Q}_h^k(s_h,a,b) \\
&\geq \mathbb{E}_{a\sim *, b\sim\nu_h^k(\cdot|s_h)}\overline{Q}_h^k(s_h,a,b) && (33)\\
&\geq \mathbb{E}_{a\sim *, b\sim\nu_h^k(\cdot|s_h)}\tilde{Q}_h^{*,\nu^k}(s_h,a,b) \\
&= \tilde{V}_h^{*,\nu^k}(s_h).
\end{aligned}$$

Here 33 holds by the property of CCE since there exists a deterministic policy to be the best response of ν. □

Lemma F.3. *In MG-RBUCBI (Algorithm 3), the expectation of regret concerning the auxiliary Markovian environment can be bounded by*

$$\mathbb{E}_{\Gamma_K}\left\{\sum_{k=1}^{K}\left[\tilde{V}_1^{*,\nu^k}(s_0) - \tilde{V}_1^{\mu^k,*}(s_0)\right]\mathbf{1}_{G_K}\right\} \leq O(S\sqrt{ABK}\text{polylog}(S,A,B,K)\iota_{\text{conf}}^2).$$

Proof. By Lemma F.2, we have the following equation.

$$\mathbb{E}_{\Gamma_K}\sum_{k=1}^{K}\left\{\left[\tilde{V}_1^{*,\nu^k}(s_0)-\tilde{V}_1^{\mu^k,*}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\}\leq\mathbb{E}_{\Gamma_K}\sum_{k=1}^{K}\left\{\left[\overline{V}_1^k(s_0)-\underline{V}_1^k(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\}$$
$$=\mathbb{E}_{\Gamma_K}\sum_{k=1}^{K}\left\{\left[\overline{V}_1^k(s_0)-\tilde{V}_1^{\pi^k}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\}+\mathbb{E}_{\Gamma_K}\sum_{k=1}^{K}\left\{\left[\tilde{V}_1^k(s_0)-\underline{V}_1^k(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\right\}.$$

By our reward assumption and the construction of overestimation and underestimation, we can similarly bound the two terms above. Following the same analysis in Lemma 6.8, we can bound the first term and thus conclude the proof for this lemma.

The only difference in the proof is that since the policy is deterministic in MDP, we can derive the following equation directly by definition.

$$\sum_{k=1}^{K}\mathbb{E}_{\Gamma_k}\sum_{h=1}^{H}\overline{V}_h^k(s_h^k)^2 I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\leq\sum_{k=1}^{K}\mathbb{E}_{\Gamma_k}\sum_{h=1}^{H}\left(\overline{r}^k\left(s_h^k,a_h^k\right)+\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k}}\overline{p}\overline{V}_{h+1}^k\right)^2 I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}.$$

A similar equation also holds in the MG setting, but it requires more refined analysis since the policy in MG is nondeterministic.

$$\begin{aligned}\mathbb{E}_{\Gamma_k}\left[\overline{V}_h^k(s_h^k)^2 I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right]&=\mathbb{E}_{\Gamma_k}\left\{\left\{\mathbb{E}_{a\sim\mu_h^k(s_h^k),b\sim\nu_h^k(s_h^k)}\left[\overline{r}^k(s_h^k,a,b)+\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a,b}}\overline{p}\overline{V}_{h+1}^k(s_h^k)\right]\right\}^2 I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\}\\&\leq\mathbb{E}_{\Gamma_k}\left\{\left\{\mathbb{E}_{a\sim\mu_h^k(s_h^k),b\sim\nu_h^k(s_h^k)}\left[\overline{r}^k(s_h^k,a,b)+\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a,b}}\overline{p}\overline{V}_{h+1}^k(s_h^k)\right]^2\right\} I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\}\quad(34)\\&=\mathbb{E}_{\Gamma_k}\left\{\left[\overline{r}^k\left(s_h^k,a_h^k,b_h^k\right)+\max_{\overline{p}\in\mathcal{P}^k_{s_h^k,a_h^k,b_h^k}}\overline{p}\overline{V}_{h+1}^k(s_h^k)\right]^2 I_h^k\mathbf{1}_{\mathrm{G}_{k-1}}\right\}\quad(35)\end{aligned}$$

Here line 34 holds since $[\mathbb{E}[X]]^2\leq\mathbb{E}\left[X^2\right]$. Since γ_k is the trajectory following π^k and the term in the large bracket is independent of the other part, we can absorb the action's expectation into the trajectory's expectation (i.e., line 35). This step shows the power of taking expectations. □

Theorem F.4. *The policy pair $\hat{\pi}=(\hat{\mu},\hat{\nu})$ returned by MG-RBUCBI (Algorithm 3) satisfies that with probability $\frac{1}{2}$,*

$$V_1^{*,\hat{\nu}}(s_0)-V_1^{\hat{\mu},*}(s_0)\leq O\left(\epsilon_{\mathrm{ucb}}+\epsilon_{\mathrm{ksp}}\right).$$

Proof. Suppose we randomly choose $k_1\in[K]$. Since $\mathbb{E}_{\Gamma_K}\left[\tilde{V}^{*,\nu^k}(s_0)-\tilde{V}^{\mu^k,*}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\geq 0$ hold for $\forall k\in[K]$, by Markov inequality, the following equation holds with probability at least $\frac{15}{16}$,

$$\mathbb{E}_{\Gamma_K}\left[\tilde{V}^{*,\nu^{k_1}}(s_0)-\tilde{V}^{\mu^{k_1},*}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\geq 16\left[O\left(\frac{S\sqrt{AB}}{\sqrt{K}}\mathrm{polylog}(S,A,B,K)\iota_{\mathrm{conf}}^2\right)\right].\quad(36)$$

Since $\left[\tilde{V}^{*,\nu^k}(s_0)-\tilde{V}^{\mu^k,*}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\geq 0$ hold for $\forall k\in[K]$, by Markov inequality, the following equation holds with probability at least $\frac{15}{16}$.

$$\left[\tilde{V}^{*,\nu^{k_1}}(s_0)-\tilde{V}^{\mu^{k_1},*}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}\geq 16\mathbb{E}_{\Gamma_K}\left[\tilde{V}^{*,\nu^{k_1}}(s_0)-\tilde{V}^{\mu^{k_1},*}(s_0)\right]\mathbf{1}_{\mathrm{G}_K}.\quad(37)$$

Since $\delta_{\text{ksp}} = \frac{1}{4}$, by setting $\delta_{\text{conf}} = \frac{1}{16S^2ABK}$, G_K holds with probability at least $\frac{5}{8} = 1 - \delta_{\text{ksp}} - 2S^2ABK\delta_{\text{conf}}$. Taking union bound, we have that with probability at least $\frac{1}{2}$, G_K holds and both equation 36, equation 37 hold. Therefore

$$\tilde{V}^{*,\nu^{k_1}}(s_0) - \tilde{V}^{\mu^{k_1},*}(s_0) \leq O\left(\frac{S\sqrt{AB}}{\sqrt{K}}\text{polylog}(S,A,B,K)\right).$$

We set $K = \tilde{O}\left(\frac{S^2AB}{\epsilon_{\text{ucb}^2}}\right)$ which satisfies that $O\left(\frac{S\sqrt{AB}}{\sqrt{K}}\text{polylog}(S,A,B,K)\right) = \epsilon_{\text{ucb}}$. Since G_0 holds, by Lemma 6.6 we have the following equation.

$$V^{*,\nu^{k_1}}(s_0) - V^{\mu^{k_1},*}(s_0) \leq \tilde{V}^{*,\nu^{k_1}}(s_0) - \tilde{V}^{\mu^{k_1},*}(s_0) + 2\epsilon_{\text{ksp}} \leq O\left(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right).$$

□

F.2. MG-Evaluation

In this section, we illustrate our MG-Evaluation algorithm and give its proof.

Algorithm 9 MG-Evaluation

1: **Input:** MG $\mathcal{G}\left(\mathcal{S},\mathcal{A},\mathcal{B},P,r,H,\mu_0\right)$, ϵ_{eval}, δ_{eval}, $\pi = (\mu,\nu)$.
2: $\hat{V}_1^{*,\nu}(s_0) \leftarrow$ MDP-Full$(\mathcal{G}+\nu, \epsilon_{\text{eval}}, \delta_{\text{eval}})$.
3: $\mathcal{G}' \leftarrow (\mathcal{S},\mathcal{A},\mathcal{B},P,-1*r,H,\mu_0)$.
4: $\hat{V}_1^{\mu,*}(s_0) \leftarrow -1\cdot$MDP-Full$(\mathcal{G}'+\mu, \epsilon_{\text{eval}}, \delta_{\text{eval}})$.
5: **Output:**$\hat{V}_1^{*,\nu}(s_0), \hat{V}_1^{\mu,*}(s_0)$.

Theorem F.5. *In each running time $t \in [T]$ in MG-Full (Algorithm 8), with probability $1 - 2\delta_{\text{eval}}$, the returned estimated value $\hat{V}_1^{*,\nu^t}(s_0)$ and $\hat{V}_1^{\mu^t,*}(s_0)$ satisfy that*

$$\left|\hat{V}_1^{*,\nu^t}(s_0) - V_1^{*,\nu^t}(s_0)\right| \leq O(\epsilon_{\text{eval}}).$$
$$\left|\hat{V}_1^{\mu^t,*}(s_0) - V_1^{\mu^t,*}(s_0)\right| \leq O(\epsilon_{\text{eval}}).$$

Proof. This theorem is a direct extension of Theorem 6.1. For the given MG environment $\mathcal{G}$, if one of the players is fixed, the environment degenerates into MDP. Here $\mathcal{G}+\mu$ refers to the case in which the max player is fixed while $\mathcal{G}+\nu$ refers to the case in which the min player is fixed. Applying Theorem 6.1 to $\mathcal{G}+\nu$ and $\mathcal{G}'+\mu$ respectively, and taking union bound will lead to the result. Note that the second Markov game is slightly modified to turn the fixed player μ to be the min player to apply our theorem in MDP setting where the unfixed player aims to maximize the sum of rewards. □

G. Proofs for Generative Setting

In this section, we present our PAC results for generative setting formally.

Theorem G.1. *In the generative setting, for any $\epsilon,\delta > 0$, with probability $1-\delta$, MDP-Full(Algorithm 1) returns an ϵ-optimal policy by sampling at most K episodes, where*

$$K = O\left(\frac{S^2A\iota^2}{\epsilon^2}\text{polylog}\left(S,A,\frac{1}{\epsilon}\right)\right).$$

MG-Full returns an ϵ-approximate NE policy pair by sampling at most K episodes, where

$$K = O\left(\frac{S^2AB\iota^3}{\epsilon^2}\text{polylog}\left(S,A,B,\frac{1}{\epsilon}\right)\right).$$

Proof. This theorem is the direct extension of Theorem 6.1 and Theorem 6.3. The only difference is that the generative setting provides us with an RFKSP algorithm with $K_1 = O(SA)$. Substituting into the proof of Theorem 6.1 and Theorem 6.3 leads to the result. □