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

Shengshi Li¹  Lin F. Yang²

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 $\hat{O}(\frac{SA}{\epsilon^2})$ episodes interacting with such an environment to achieve an $\epsilon$-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 $\hat{O}(1)$ 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 determinicity 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 $\epsilon$-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

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

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 approaches extend to two-player zero-sum games – our assumption.

**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 \( \text{poly}(H) \) 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 \( \epsilon \)-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 \( S^{3/4}A^1 \) 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.

### 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.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Online PAC</th>
<th>Generative PAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Li et al., 2022)</td>
<td>( \tilde{O}(\text{poly}(S, A, H^{1/2} \epsilon^{-3})) )</td>
<td>( O(\frac{S^3 A^3}{\epsilon^2}) )</td>
</tr>
<tr>
<td>(Zhang et al., 2022)</td>
<td>( \tilde{O}(\frac{S^3 A^3}{\epsilon^2}) )</td>
<td>-</td>
</tr>
<tr>
<td><strong>Our work</strong></td>
<td>( \tilde{O}(\frac{S^3 A^3}{\epsilon^2} + \frac{S^3 A^3}{\epsilon^2}) )</td>
<td>( \tilde{O}(\frac{S^3 A^3}{\epsilon^2}) )</td>
</tr>
</tbody>
</table>

Table 1. Comparison to existing horizon-independent results.
2. Preliminaries

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

Markov Decision Process (MDP). In this paper we consider the finite-horizon time-homogeneous MDP \(M = (S, A, P, R, H, s_0)\), where \(S\) is the finite state space, \(A\) is the finite action space, \(P : S \times A \rightarrow \Delta(S)\) is the unknown (but fixed) transition operator which takes a state-action pair and returns a distribution over states, \(R : S \times A \rightarrow \Delta([-1, 1])\) is the reward function, \(H \in \mathbb{Z}^+\) 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 \(\mu : S \rightarrow \Delta(A)\) stands for the policy of the max-player and \(\nu : S \rightarrow \Delta(B)\) stands for the policy of the min-player. For a given strategy, the corresponding state-value and action-value functions are defined as follows.

\[
V_h^\pi(s_h) := \mathbb{E} \left[ \sum_{t=0}^H r(s_t, a_t) | \pi, s_h \right],
\]

\[
Q_h^\pi(s_h, a_h) := \mathbb{E} \left[ \sum_{t=0}^H r(s_t, a_t) | \pi, s_h, a_h \right].
\]

Our goal is to find an optimal policy \(\pi^*\) that maximizes the value, i.e. \(\max_{\pi} \mathbb{E} \left[ \sum_{t=0}^H r(s_t, a_t) \right]\) by only interacting with the environment. We use \(V_h^\pi\) and \(Q_h^\pi\) to denote the value function of \(\pi^*\), respectively. We call a policy \(\epsilon\)-optimal if \(V_1^\pi(s_1) - V_1^\pi(s_1) \leq \epsilon\).

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

\[
V_h^\pi(s_h) := \mathbb{E} \left[ \sum_{t=0}^H r(s_t, a_t, b_t) | \pi, s_h \right],
\]

\[
Q_h^\pi(s_h, a_h, b_h) := \mathbb{E} \left[ \sum_{t=0}^H r(s_t, a_t, b_t) | \pi, s_h, a_h, b_h \right].
\]

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 \(\nu\) is fixed, the MG degenerates to an MDP, and the optimal policy in this MDP is the best response strategy \(br_1(\nu)\). Similarly, we can define the best-response strategy, \(br_2(\mu)\), for the min-player. The subscript in \(br_1\) and \(br_2\) will be ignored in the clear context. For all \(h \in [H], s_h \in S\), we define

\[
V_h^{\pi,\nu}(s_h) := V_h^{br_1(\nu),\nu}(s_h) = \max_\mu V_h^{\mu,\nu}(s_h),
\]

\[
V_h^{\mu,\nu}(s_h) := V_h^{\mu,br_2(\mu)}(s_h) = \min_\nu V_h^{\mu,\nu}(s_h).
\]

There exists Nash equilibrium (NE) policy pair \((\mu^*, \nu^*)\) that \(\mu^*\) and \(\nu^*\) are the best responses to each other. We define \(V_h^*(s_h) = V_h^{\mu^*,\nu^*}(s_h)\) for all \(s_h \in S, h \in [H]\).

The following weak duality property holds for all policy pairs \((\mu, \nu)\) in MG:

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

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

\[
\text{Gap}(\pi) = V_1^{\pi,\nu}(s_1) - V_1^{\pi^*,s_1}(s_1).
\]

We call a policy pair \(\epsilon\)-approximate NE if \(\text{Gap}(\pi) \leq \epsilon\).

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

\[
\text{Regret}(K) = \sum_{k=1}^K V_1^{\pi^k}(s_1^k) - V_1^{\pi^*,s_1}(s_1^k),
\]

while the regret in MG setting is defined as

\[
\text{Regret}(K) = \sum_{k=1}^K V_1^{\pi^k,\nu}(s_1^k) - V_1^{\pi^*,\nu}(s_1^k).
\]

The measurement we will use is PAC-RL sample complexity, which counts the total number of episodes to find an
Proof. To prove the optimal policy exists, we can construct an optimal policy with the Bellman optimality equation. By setting up the problem as a MDP, we can use dynamic programming to find the optimal policy. The existence of an optimal policy is guaranteed because the MDP is finite and the reward function is bounded.


c-approximate NE policy in MDP or an c-optimal policy 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 τ0 to denote all the trajectories in RFKSP and τh to denote the trajectory in the h-th episode afterward. Moreover, we denote Γk = (τ0, τ1, . . . , τh) as the trajectories before the k + 1-th episode interacting with the aux MDP/MG. We use Nk(s, a) to denote the visit count of (s, a) in Γk−1 and Nk(s, a) denote the visit count of (s, a) in RFKSP. Similarly we define Nk(s, a, s′) and set all Nk(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 A by A × B.

Assumption 3.1 (Reward Uniformity). For all (s, a) ∈ S × A, we have 0 ≤ r(s, a) ≤ 1/H. 3

Assumption 3.2 (Total Boundedness). 0 ≤ r(s, a) ≤ 1 and ∑h=1 H r(sh, ah) ≤ 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 ∈ [H], h-reachable state sh and policy π, |r(sh, a)| ≤ 1 and

Eπ [∑h=h H |r(sh, ah)|] ≤ 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) + c(s, a), where c(s, a) is the observation noise with zero-expectation. We can assume the noise is small enough so 0 ≤ r(s, a) ≤ 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|s, c) = 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.
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 ground-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 $M$, $\epsilon, \delta$.
2: Set $\epsilon_{ksp}, \epsilon_{ucb}, \epsilon_{eval} = O(\epsilon)$, $\delta_{ksp} = \frac{\delta}{2}$, $\delta_{eval} = \frac{\delta}{2T}$.
3: Run $T = \log \left( \frac{2}{\epsilon} \right)$ independently.
4: for $t = 1, 2, \ldots, T$ do
5: $M_t \leftarrow$ RFKSP($M$, $\epsilon_{ksp}, \delta_{ksp}$).
6: $\pi_t \leftarrow$ MDP-RBUCBI($M_t$, $\epsilon_{ucb}$).
7: $V^{\pi_t}_1(s_0) \leftarrow$ MDP-Evaluation($M, \pi_t, \epsilon_{eval}, \delta_{eval}$).
8: end for
9: $i = \arg \max_{i \in [T]} V^{\pi_i}_1(s_0)$.
10: Output: $\pi_i$ or $V^{\pi_i}_1(s_0)$ 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 $O(\log(\frac{1}{\epsilon}))$ 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

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

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

$$
\begin{align*}
\tilde{P}_{s,a} &= P_{s,a}, & \tilde{R}(s, a) &= R(s, a), & \forall(s, a) \in \mathcal{O}^c, \\
\tilde{P}_{z,a} &= 1_z, & \tilde{R}(s, a) &= 0, & \forall(s, a) \in \mathcal{O}, \\
\tilde{P}_{z,a} &= 1_z, & \tilde{R}(z, a) &= 0, & \forall a \in A.
\end{align*}
$$
If the max visiting probability to $O$ is small enough, $V^*_1(s_0)$ is close to $V^*_2(s_0)$ for any policy $\pi$. ($V$ is the value function in the aux MDP/MG). If we further have sufficient initial samples for the “important” state-action pairs in $O^C$, we call such aux MDP/MG to be good conditioned.

**Definition 5.2.** We call an auxiliary Markovian environment to be $\epsilon$-good conditioned if it satisfies that

1. $\max_{\pi} \mathbb{P}_x [3h \in [H], (s_h, a_h) \in O] \leq \epsilon$,

2. $N^1(s, a) \geq \frac{U(s, a)}{\text{poly}(3, \delta)}, \forall (s, a) \in O^C$.

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 $K_1 = \text{poly} \left( S, A, \frac{1}{\epsilon} \right)$ episodes, the auxiliary Markovian environment it returned is $\epsilon$-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 reliable reward-free key-state preserving algorithm with $K_1 = \tilde{O}(S^2A^3/\epsilon)$. This algorithm is given in Appendix C and is used in our current result. The core idea is that it maintains an omitted set $O \subset S \times A$. Each episode is divided into two phases, where the algorithm plans optimistically to reach $O$ in the first phase and collects samples in the second phase. The collected target is the first reached state-action pair in $O$ when the estimated transition probability is accurate enough. Once enough samples for a state-action pair are collected, it is removed from $O$ to $O^C$. If not enough samples for some $(s, a)$ are collected during the entire phase, it remains in the omitted set $O$. 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 $H \geq U(s, a)$. 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-RBUCB is formally presented in Algorithm 2. In each episode $k$ and state $s$, we set $\hat{\nu}^k_{H+1} = 0$, and calculate $\hat{Q}^k_{H}, \hat{\nu}^k_H, \ldots, \hat{Q}^k_1, \hat{\nu}^k_1$ iteratively as follows. For $(s, a) \in O$, $\hat{Q}^k_{h}(s, a) = 0$. For $(s, a) \in O^C$,

$$\hat{Q}^k_{h}(s, a) = \min \left( \pi^k(s, a), \max_{\pi \in \mathcal{P}_{s, a}^k} h \nu^k_{h+1}, 1 \right).$$

Furthermore, $\pi^k(s, a) = \arg \max_a \hat{Q}^k_{h}(s, a)$ and $\nu^k_{h} = \mathbb{E}_{a \sim \pi^k_h(s, |a)\hat{Q}^k_{h}(s, a)}$.

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

**Algorithm 2** MDP-RBUCB (Reward-Based UCB with Initialization)

1: **Input:** $\mathcal{M}$ $(\epsilon_{\text{kb}}, \delta_{\text{kb}} = \frac{1}{4})$ (aux MDP), $\epsilon_{\text{ucb}}$.
2: **Initialization:** $\hat{V}^k_{h+1}(s) = 0, \forall k, s \in S$.  
3: Use $K = \tilde{O}(S^2A^3)$ episodes.
4: for episode $k = 1, 2, \ldots, K$ do
5: for step $h = H, H - 1, H - 2, \ldots, 1$ do
6: Compute $\hat{Q}^k_{h}(s, a)$ as in equation 1.
7: for $s \in S$ do
8: $\pi^k_h(s) \leftarrow \arg \max_a \hat{Q}^k_{h}(s, a)$.
9: Compute $\nu^k_{h}(s) = \mathbb{E}_{a \sim \pi^k_h(s, |a)\hat{Q}^k_{h}(s, a)}$.
10: end for
11: end for
12: Receive initial state $s^k_1 = s_0$, play policy $\pi^k$, collect trajectory $\tau_k$. Calculate $\mathcal{P}^{k+1}, \mathcal{R}^{k+1}$ based on $\Gamma_k$.
13: end for
14: **Output:** Randomly select one policy $\pi^k$.

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

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

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

$$\rho \in \Delta(S) : \left| \rho_s - \hat{P}^k_{s, a, s'} \right| \leq \sqrt{\frac{2\hat{P}^k_{s, a, s'} + \delta_{\text{conf}}}{3N^k(s, a)}} \right\}. \tag{2}$$

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 $r^k(s, a), i \in [N^k(s, a)]$. We build the confidence set $\mathcal{R}^k_{s, a}$ based on the sample mean $\hat{r}^k(s, a)$ as

$$\left\{ r : |r - \hat{r}^k(s, a)| \leq \sqrt{\frac{4\hat{V}^k_{s, a} + 10\delta_{\text{conf}}}{N^k(s, a)}} \right\},$$
where $\hat{V}^k$ is the sample variance. It can be shown that with probability at least $1 - \delta$, $\mathbb{E} [R(s, a)] \in \mathcal{R}_{s,a}^k$ holds for any $(s, a, k)$. We set the overestimation $\tau^k(s, a)$ to be $\min \{\max_{r \in \mathcal{R}_{s,a}^k} r, 1\}$, and the underestimation $\bar{\tau}^k(s, a, b)$, $\bar{\tau}^{k+1}(s, a, b)$ used in MG to be $\max \{\min_{r \in \mathcal{R}_{s,a}^k} r, -1\}$.

5.3. Algorithm for MG

The whole algorithm MG-Full (Algorithm 8) is given in Appendix F, which resembles MDP-Full. However, we fix the gap for a policy $\pi$ in MG as $V_\pi^k(s_0) - V_\pi^k(s_0)$, and the gap in MG is $V_\pi^k(s_0) - V_\pi^k(s_0)$. We used CCE in MG to be $\min \{\max_{r \in \mathcal{R}_{s,a}^k} r, 1\}$, and the underestimation $\bar{\tau}^k(s, a, b)$, $\bar{\tau}^{k+1}(s, a, b)$ used in MG to be $\max \{\min_{r \in \mathcal{R}_{s,a}^k} r, -1\}$.

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 $(\phi, \psi) \in \Delta(A) \times \Delta(B)$ for general sum game, which satisfies

$$\phi^T P \psi \geq \max_{a \in A} \mathbb{E}_{P} \phi_a, \quad \phi^T Q \psi \leq \min_{b \in B} \phi^T Q b.$$

When extending the algorithm from MDP to MG, the estimation function are modified as follows. For $(s, a, b) \in \mathcal{O}$,

$$\hat{Q}^k(s, a, b) = \mathcal{Q}^k(s, a, b) = 0.$$

For $(s, a, b) \in \mathcal{O}^C$,

$$\mathcal{Q}^k(s, a, b) = \min \left\{ \tau^k(s, a, b) + \max_{r \in \mathcal{R}_{s,a}^k} \mathbb{E}_{P} \mathbb{E}_{s', a'} \hat{V}_{k+1}^{s', a'}, 1 \right\}, \quad (3)$$

$$\mathcal{Q}^k(s, a, b) = \max \left\{ \tau^k(s, a, b) + \min_{r \in \mathcal{R}_{s,a}^k} \mathbb{E}_{P} \mathbb{E}_{s', a'} \hat{V}_{k+1}^{s', a'}, -1 \right\}.$$

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 $(\mu, \nu)$. We multiply rewards by $-1$ when estimating $V_\mu^k(s_0)$ 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 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 $\epsilon$-optimal policy by sampling at most $K_{Reward} + K_{RFKSP}$ episodes, where $K_{Reward} = O \left( S^2 A / \epsilon^2 \right), K_{RFKSP} = O \left( S^9 A^3 / \epsilon \right)$.

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

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 $\epsilon$-approximate NE policy pair by sampling at most $K_{Reward} + K_{RFKSP}$ episodes, where $K_{Reward} = O \left( S^2 A / \epsilon^2 \right), K_{RFKSP} = O \left( S^9 A^3 B^3 / \epsilon \right)$.

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 $O( S^9 A^3 / \epsilon^2 )$ PAC result in the generative setting. The formal guarantee is presented as Theorem 6.1 in Appendix G.

Below, we provide a proof sketch for Theorem 6.1. The proofs for Theorem 6.3 and Theorem 6.1 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
\[ i = \arg \max_{t \in [T]} \hat{V}^\pi_t(s_0), \quad j = \arg \max_{t \in [T]} V^\pi_t(s_0). \]

MDP-Full outputs policy \( \pi, \) whose gap \( V^\pi_1(s_0) - V^\pi_1(s_0) \) can be decomposed as
\[
\leq |V^\pi_1(s_0) - V^\pi_1(s_0)| + |V^\pi_1(s_0) - \hat{V}^\pi_1(s_0)| + |\hat{V}^\pi_1(s_0) - V^\pi_1(s_0)|.
\]

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, \( \pi \) 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 \( \hat{V}^\pi_1(s_0) \) returned by MDP-Evaluation satisfies that with probability \( 1 - \delta_{\text{eval}} \),
\[ |V_T^\pi(s_0) - V^\pi_1(s_0)| \leq O(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}). \]

**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 \( \Gamma_k \) satisfies:

1. The auxiliary Markovian environment built on \( \tau_0 \) is \( \epsilon_{\text{ksp}} \cdot \text{good conditioned for } \epsilon_{\text{ksp}} \) given in RFKSP.
2. For any \((s, a)\) and \( t \in [k + 1] \), \( \hat{P}_{s, a}^k \) satisfies equation 2 and \( E[R(s, a)] \in R_{s, a}^k \).

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 \( V_n^k(s_h) \geq \hat{V}_n^k(s_h) \). By applying Markov inequality twice to the expectation of regret and setting \( K, \delta_{\text{conf}} \) and \( \delta_{\text{ksp}} \) appropriately, we conclude that both \( G_K \) and \( V_T^\pi(s_0) - V^\pi_1(s_0) \leq O(\epsilon_{\text{ucb}}) \) hold with probability \( 1/2 \). We denote the value of the optimal policy for the true MDP in the aux MDP as \( \hat{V}^{**} \). By definition, \( \hat{V}^\pi_1(s_0) \geq \hat{V}^{**}_1(s_0) \), and we have that
\[
\hat{V}^\pi_1(s_0) - V^\pi_1(s_0) \leq \hat{V}^{**}_1(s_0) - \hat{V}^\pi_1(s_0) + 2\epsilon_{\text{ksp}} \leq \hat{V}^{**}_1(s_0) + 2\epsilon_{\text{ksp}} \leq O(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}).
\]

**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 \( O \) is less than \( \epsilon \), i.e.\( \max_{s} P \left[ h \in [H], (s_h, a_h) \in O^{(c)} \right] \leq \epsilon \), then for any fixed policy \( \pi \),
\[ |V^\pi_1(s_0) - \hat{V}^\pi_1(s_0)| \leq \epsilon. \]

We denote
\[ L = \max_{(s, a) \in O^{(c)}} \sum_{k=1}^K \min \left( \log_2 \left( \frac{N^{k+1}(s, a)}{N^k(s, a)} \right), 1 \right). \]

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 \( 1/N_k(s, a) \) can be generally bounded by \( \text{SAL} \). Since \( N^{k+1}(s, a) - N^k(s, a) \leq KU(s, a)/\epsilon \) holds with probability \( 1 - \delta \), \( L \) is bounded when there is sufficient initial samples.

**Lemma 6.7.** \( E_{\Gamma_K} L 1_{G_0} \leq O(\text{polylog}(S, A, K)). \)

**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
\[ E_{\Gamma_K} \left\{ \sum_{k=1}^K \left[ V^\pi_1(s_0) - V^\pi_k(s_0) \right] 1_{G_k} \right\} \leq \bar{O}(S\sqrt{AK}). \]

**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 \( \hat{V}^\pi_1 \) and \( \hat{V}^{**}_1 \), which are both expected sum over the trajectory space following \( \pi^k \). We expand their difference into the expectation form as
\[ T_1 = \sum_{k=1}^K \sum_{h=1}^H E_{\Gamma_k} \left( \max_{p \in \pi^k, s_h} p - \hat{P}_{s_h}^k \right) \left( \hat{V}^{**}_h 1_{G_{k-1}} \right). \]

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
We propose a relaxed reward-boundedness assumption for studying the horizon-dependence problem. Under our relaxed assumption, we propose a generic algorithmic framework that 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


Horizon-free Learning for MDP and Games: Stochastically Bounded Rewards and Improved Bounds


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 \([0, \ell]\). Define \( F_k = \sigma(X_1, X_2, \ldots, X_{k-1}) \) and \( Y_k = \mathbb{E}[X_k | Y_k] \) for \( k \geq 1 \). For any \( \delta > 0 \), we have that

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

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

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 \( \forall Z = \mathbb{E}[(Z - \mathbb{E}[Z])^2] \). Then we have

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

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 \( S_n = \sum_{i=1}^{n} X_i \), then for all \( \epsilon, k \geq 0 \),

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

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 \( Z = \frac{1}{n} \sum_{i=1}^{n} Z_i \) and \( V_n = \frac{1}{n} \sum_{i=1}^{n} (Z_i - Z)^2 \). Then we have

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

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 \( \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{N^k(s_h^k, a_h^k)} \) where \((s_h^k, a_h^k)\) 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

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{N^k(s_h^k, a_h^k)} = O \left( \sum_{s,a} \sum_{k=1}^{K} \log \left( \frac{N^{k+1}(s, a)}{N^k(s, a)} \right) \right)
\]

\[
\leq O(SA \log(KH)).
\]

(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 \( U(s, a) = \max_{\pi} \mathbb{E}_{\pi} \left[ \sum_{h=1}^{H} 1_{(s_h, a_h) = (s, a)} \right] \) 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 \( N^1(s, a) \geq U(s, a)/\exp(\text{poly}(S, A)) \),

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

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

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

3. \( X^\pi_d(O, p, \mu_1) \): the probability of reaching \( O \) in \( d \) steps under transition probability \( p \), policy \( \pi \), initial distribution \( \mu_1 \).

4. \( X^\pi_s(O, p, \mu_1) := \sum_{i \geq 1} \gamma^{i-1} P[(s_i, a_i, s_{i+1}) \in O, (s'_i, a'_i, s'_{i+1}) \notin O, \forall i \leq i' \leq i - 1 | s_1 \sim \mu_1] \).

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

1: Input: MDP \( M \), \( \epsilon \), \( \delta \).
2: Initialization: \( N(s, a, s') \leftarrow 0, \forall s, a, s' \), \( \tilde{N}(s, a) \leftarrow 0, \forall (s, a), d \leftarrow \frac{(S+1)H}{s+2} \). \( O^1 \leftarrow S \times A \). \( n_1 \leftarrow C_2 S^7 A^3 t \).
3: for \( k = 1, 2, \ldots, K_1 \) do
4: \( \mathcal{P}^k \leftarrow \text{Build confidence set } \mathcal{P}_{s,a}^k \text{ based on } \{N(s, a, s')\}_{s,a,s'} \).
5: \( (\pi^k, \tilde{P}^k) \leftarrow \max_{\pi, P \in \mathcal{P}^k} X^\pi_d(O^k, p, \mu_1) \).
6: for \( h = 1, 2, \ldots, d \) do
7: Observe \( s^k_h \), takes action \( \pi^k_h(s^k_h) \), receives \( r^k_h \) and transits to \( s^k_{h+1} \).
8: \( N(s^k_h, a^k_h, s^k_{h+1}) \leftarrow N(s^k_h, a^k_h, s^k_{h+1}) + 1 \).
9: if \( \exists a, (s^k_{h+1}, a) \in \mathcal{O}^k \) then
10: \( (s^*_1, a^*_1) \leftarrow (s^k_{h+1}, a) \).
11: \( \tilde{N}(s, a, s') \leftarrow \{n(s, a, s')\}_{s,a,s'} \).
12: \( \mathcal{K}^k \leftarrow \{(s, a, s') : n(s, a, s') \geq N_0\}, \mathcal{K}^k(s, a) \leftarrow \{s' : (s, a, s') \in \mathcal{K}^k\} \).
13: \( n(s, a) \leftarrow \max_{(s', a', s') \in \mathcal{K}^k} n(s, a, s'), 1 \) \( \forall (s, a) \).
14: \( \mathcal{P}_{s,a,s'} \leftarrow n(s,a,s') / n(s,a), \mathcal{P}_{s,a,z} \leftarrow 0, \forall (s, a, s') \in \mathcal{K}^k \).
15: \( \mathcal{P}_{s,a,s'} \leftarrow 0, \mathcal{P}_{s,a,z} = 1, \forall (s, a, s') \text{ such that } \mathcal{K}^k(s, a) = 0 \).
16: (Trigger, \( \{n(s, a, s')\}_{s,a,s'} \) \( \leftarrow \) Algorithm 5 with inputs \((s^*_1, a^*_1), \mathcal{P}_{s,a,s'} \).
17: if Trigger = FALSE then
18: \( \{n(s, a, s')\}_{s,a,s'} \) \( \leftarrow \) Algorithm 6 with inputs \((s^*_1, a^*_1), \mathcal{P}_{s,a,s'}, d' \).
19: \( m(s^*_1, a^*_1) \leftarrow m(s^*_1, a^*_1) + 1 \).
20: if \( m(s^*_1, a^*_1) \geq 400 \log(1/\delta) \) then
21: \( \mathcal{O}^{k+1} \leftarrow \mathcal{O}^k / (s^*_1, a^*_1) \).
22: end if
23: end if
24: \{n(s, a, s')\}_{s,a,s'} \leftarrow \{N(s, a, s')\}_{s,a,s'} \).
25: break.
26: end if
27: end for
28: If there are remaining steps, run a random policy and update \( \{N(s, a, s')\}_{s,a,s'} \).
29: end for
30: Return: an auxiliary Markovian environment \( \tilde{M} \) which is built based on \( \mathcal{O}^{K_1+1} \).
Proof. Combining Lemma 6 and Lemma 24 in (Zhang et al., 2022), we have that with probability $1 - O \left( \frac{K^2}{S^8A^8\log K} \delta_0 \right)$.
1. \( \max_{\pi} \mathbb{P}_{\pi} [ \exists h \in [H], (s_h, a_h) \in \mathcal{O}] \leq O\left( \frac{S^9A^3\epsilon_{0} + S^3A\epsilon^2}{K_1} \right) \).

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 \( O\left( \frac{S^9A^3\epsilon_{0} + S^3A\epsilon^2}{K_1} \right) \leq \epsilon \) and \( O\left( \frac{K_1^2}{S^9A^3\epsilon_{0} \delta_0} \right) \leq \delta \). The first equation can be satisfied by setting \( K_1 = \Omega\left( \frac{S^9A^3\epsilon_{0} + S^3A\epsilon^2}{\epsilon} \right) \). Substituting it into the second equation and noting that \( \delta_0^{-0.5} \geq O\left( \delta_0^3 \right) \), we can meet both of our requirements by setting \( \delta_0 = \frac{S^9A^3}{S^9A\epsilon} \).

Wrapping up all these results, we have that if we set \( K_1 = O\left( \frac{S^9A^3\epsilon_{0} + S^3A\epsilon^2}{\epsilon} \text{polylog}(S, A, \frac{1}{\epsilon}) \right) \), with probability \( 1 - \delta \),

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

2. \( N^1(s, a) \geq \frac{U(s, a)}{\text{poly}(S)} \) 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 $A \times B$ in MG as the action space $A$ in MDP.

D.1. Proofs for Auxiliary Markovian environment

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

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

then for any fixed policy $\pi$, $|V^{\pi}_i(s_0) - \tilde{V}^{\pi}_i(s_0)| \leq \epsilon$.

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 $\Gamma_1$ denotes the trajectory before the first visit to $O$ and $\Gamma_2$ denotes the left trajectory. For example, if $(s_2, a_2)$ is the first time the trajectory visits to $O$, $\Gamma_1 = (s_1, a_1, s_2)$ and $\Gamma_2 = (s_2, a_2, \ldots, s_{H+1})$. $\Gamma_2$ can be empty in the extreme case where the trajectory $\Gamma$ never visits to $O$.

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 $\Gamma_1$, we denote the set of suitable $\Gamma_2$ as $S(\Gamma_1) := \{ \Gamma_2 : \exists \Gamma = \Gamma_1 \cup \Gamma_2 \}$. With these notations, we have that

$$V^{\pi}_i(s_0) = \sum_{\Gamma_1} \sum_{\Gamma_2 \in S(\Gamma_1)} (r(\Gamma_1) + r(\Gamma_2)) \cdot P(\Gamma_1) \cdot P(\Gamma_2),$$

$$\tilde{V}^{\pi}_i(s_0) = \sum_{\Gamma_1} \sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_1) \cdot P(\Gamma_1) \cdot P(\Gamma_2).$$

So the difference can be calculated as

$$\left| V^{\pi}_i(s_0) - \tilde{V}^{\pi}_i(s_0) \right| = \left| \sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_2) \cdot P(\Gamma_1) \cdot P(\Gamma_2) \right| \leq \sum_{\Gamma_2 \in S(\Gamma_1)} r(\Gamma_2) \cdot P(\Gamma_2).$$

When $\Gamma_2$ is an empty set, $r(\Gamma_2) = 0$. From the requirement that the max visiting probability to $O$ is $\epsilon$, we know that the probability of $\Gamma_1 \neq \Gamma$ is less than $\epsilon$. If $\Gamma_1 \neq \Gamma$, we assume the last term in $\Gamma_1$ is $(s_h, a_h)$. We set $\pi' = \pi$ except $\pi'_h(s_h) = a$.

Then from our reward assumption, we have that

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

Thus we conclude our proof by noticing that

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

\[ \square \]

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 $N^k(s, a)$ to be the visit count before the $h$-th step in the $k$-th episode, $J = \{ (k, h) : \exists (s, a) \in O^C, s.t. N^k(s, a) \geq 2N^k(s, a) \}$. $I^k_h = 1$ if $(k, h) \notin J$ else 0. By the definition, $I^k_1 = 1$ and $I^k_h$ do not depend on the action taken at the $h$-th step in the $k$-th episode. We define $L = \max_{(s, a) \in O^C} \sum_{k=1}^{K} \min \left( \log_2 \left( \frac{N^{k+1}(s, a)}{N^k(s, a)} \right), 1 \right)$.
Lemma D.2.

\[
\sum_{k=1}^{K} I_{k}^{h} = 0 \leq SAL \quad \text{and} \quad \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{I_{k}^{h}}{N^{k}(s_{h}^{k}, a_{h}^{k})} \mathbf{1}_{(s_{h}^{k}, a_{h}^{k})} \in \mathcal{O^{C}} \leq SAL.
\] (4)

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

\[
\sum_{k=1}^{K} I_{\exists h, I_{k}^{h} = 0} \leq \sum_{k=1}^{K} \sum_{(s, a) \in \mathcal{O}^{C}} \min \left\{ \log_{2} \left( \frac{N^{k+1}(s, a)}{N^{k}(s, a)} \right), 1 \right\} = SAL.
\]

Meanwhile, due to the existence of \( I_{k}^{h} \), we can derive that

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{I_{k}^{h}}{N^{k}(s_{h}^{k}, a_{h}^{k})} \mathbf{1}_{(s_{h}^{k}, a_{h}^{k})} \in \mathcal{O}^{C} \leq \sum_{(s, a) \in \mathcal{O}^{C}} \sum_{k=1}^{K} \min \left\{ \frac{N^{k+1}(s, a) - N^{k}(s, a)}{N^{k}(s, a)}, 1 \right\} = SAL.
\]

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

\[
\mathbb{E}_{G_{k}} L_{1} \leq O(\text{polylog}(S, A, K)).
\] (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 \( B(s, a) = \{ k \in [K] : N^{k+1}(s, a) - N^{k}(s, a) \geq K^{2}U(s, a) \} \). 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 \( X_{k} = 1_{k \in B(s, a)} \). \( Y_{k} = \mathbb{E}[X_{k} | F_{k}] = P(N^{k+1}(s, a) - N^{k}(s, a) \geq K^{2}U(s, a) | F_{k}) \leq 1/K^{2} \). Therefore with probability \( 1 - \delta \), \( |B(s, a)| = \sum_{k=1}^{K} X_{k} \leq 3 \sum_{k=1}^{K} Y_{k} + \epsilon \leq 3/K + \epsilon \). Taking union bound, we have that with probability \( 1 - SA\delta \), \( |B(s, a)| \leq \frac{3}{K} + \epsilon \) hold for \((s, a) \). Under such event, for any \((s, a) \in \mathcal{O}^{C} \), suppose...
Let \( B(s,a) = \{ k_1,k_2,\ldots \} \) and let \( k_0 = 0 \), we have that

\[
\sum_{k \notin B(s,a)} \min \left\{ \log_2 \left( \frac{N^{k+1}(s,a)}{N^k(s,a)} \right), 1 \right\} \leq \sum_{i \geq 0} \sum_{k = k_i}^{k_{i+1}-1} \log_2 \left( \frac{N^{k+1}(s,a)}{N^k(s,a)} \right)
\]

\[
\leq \sum_{i \geq 0} \log_2 \left( \frac{K^3U(s,a) + N^{k_i}(s,a)}{N^k(s,a)} \right)
\]

\[
\leq |B(s,a)| \log_2 \left( \frac{K^3U(s,a) + N^1(s,a)}{N^1(s,a)} \right)
\]

\[
\leq \left( \frac{3}{K} + \epsilon \right) \log_2 \left( \frac{K^3U(s,a) + N^1(s,a)}{N^1(s,a)} \right).
\]

When \( G_0 \) holds, \( N_1(s,a) \geq \frac{U(s,a)}{\text{poly}(S)} \) 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

\[
\mathbb{E}_{\Gamma_K} L1G_0 \leq \left( \frac{3}{K} + \epsilon \right) \log_2 \left( \frac{K^3U(s,a) + N^1(s,a)}{N^1(s,a)} \right) + \left( \frac{3}{K} + \epsilon \right) + SA \delta K \leq O(\text{polylog}(S,A,K)).
\]

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 \( \hat{P} \) and \( \hat{R} \).

Lemma D.4. For any \( \delta_{\text{conf}} > 0 \), with probability at least \( 1 - S^2 A K \delta_{\text{conf}} \),

\[
|P_{s,a,s'} - \hat{P}_{s,a,s'}^k| \leq 5 \sqrt{\frac{\hat{P}_{s,a,s'}^{k_{\text{conf}}} N^k(s,a)}{N^{k+1}(s,a)}} + \frac{5 \delta_{\text{conf}}}{N^k(s,a)}
\]

holds for any \((s,a,s')\) and \( k \). With probability at least \( 1 - S A K \delta_{\text{conf}} \),

\[
|\mathbb{E}R(s,a) - \hat{r}^k(s,a)| \leq 4 \sqrt{\frac{\hat{r}_{\text{conf}}^{k} N^k(s,a)}{N^{k+1}(s,a)}} + \frac{10 \delta_{\text{conf}}}{N^k(s,a)}
\]

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 \( X_1, X_2, \ldots, X_{N^k(s,a)} \), in which \( \mathbb{E}[X_1, X_2, \ldots, X_{i-1}] = P_{s,a,s'}(1 - P_{s,a,s'}) \leq P_{s,a,s'} \). By further setting \( k = N^{k}(s,a)P_{s,a,s'} \), we can derive from lemma A.3 that with probability \( 1 - \delta_{\text{conf}} \),

\[
|P_{s,a,s'} - \hat{P}_{s,a,s'}^k| \leq \sqrt{2 \frac{P_{s,a,s'}^{k_{\text{conf}}} N^k(s,a)}{N^{k+1}(s,a)}} + \frac{\delta_{\text{conf}}}{3N^k(s,a)}.
\]
When the above line holds, we have

\[
5 \sqrt{\frac{p_{s,a,s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + \frac{5 \epsilon_{\text{conf}}}{N^{k}(s,a)} \geq 5 \sqrt{\frac{p_{s,a,s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + \frac{5 \epsilon_{\text{conf}}}{N^{k}(s,a)}
\]

\[
\geq 5 \sqrt{\frac{p_{s,a,s',s^{'}}^{k} N^{k}(s,a)}{N^{k}(s,a)}} - 5 \sqrt{2 \frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} \cdot \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)} + \left( 5 - \frac{5}{\sqrt{3}} \right) \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)}
\]

\[
\geq (5 - \frac{5 \sqrt{2}}{2}) \sqrt{\frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + \left( 5 - \frac{5}{\sqrt{3}} \right) \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)}
\]

\[
\geq 2 \frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)} + \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)}
\]

We need to mention that when \( p_{s,a,s',s'}^{k} N^{k}(s,a) \) 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 \( V_{a}^{k}, \hat{a}^{k} \) as the sample variance and the sample mean of \( \{a^{i}\} \). By definition \( V_{a}^{k} = 4 V_{a}^{k} \). Again we apply lemma A.4 to \( \{a^{i}\} \), with probability \( 1 - \delta_{\text{conf}} \).

\[
|\mathbb{E}R(s, a) - \hat{r}(s, a)| = 2 |\mathbb{E}a - \hat{a}^{k}| \leq 2 \left| 4 \frac{\hat{V}_{a}^{k} \epsilon_{\text{conf}}}{N^{k}(s,a)} + 2 \frac{5 \epsilon_{\text{conf}}}{N^{k}(s,a)} \right|
\]

\[
\leq 2 \left| \frac{\hat{V}_{a}^{k} \epsilon_{\text{conf}}}{N^{k}(s,a)} + \frac{10 \epsilon_{\text{conf}}}{N^{k}(s,a)} \right|
\]

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

\[\Box\]

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

\[
|P'(s') - P''(s')| \leq C \left( \sqrt{\frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)} \right)
\]

hold for any \((s, a, s', s'') \in S \times A \times S \).

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

**proof of Lemma D.5.** Since \( P' \) and \( P'' \in \{P_{s,a}^{k}\} \), using equation 2 leads to

\[
|P'(s') - P''(s')| \leq 10 \sqrt{\frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + 10 \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)}
\]

\[
\leq 10 \left( \sqrt{\frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)} \right) + 10 \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)}
\]

\[
\leq 10 \left( \sqrt{\frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)} \right) + 16 \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)}
\]

\[
\leq 25 \left( \sqrt{\frac{p_{s,a,s',s'}^{k} N^{k}(s,a)}{N^{k}(s,a)}} + \frac{\epsilon_{\text{conf}}}{N^{k}(s,a)} \right)
\]

Thus taking \( C = 25 \) conclude this proof. \[\Box\]
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 $\epsilon$-optimal policy by sampling at most $K = K_{\text{Reward}} + K_{\text{RFKSP}}$ episodes, where

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

**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 $1/2$,

\[
0 \leq V^*_t(s_0) - V^t(s_0) \leq O\left(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}\right).
\]

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

\[
\left| \hat{V}^*_t(s_0) - V^*_t(s_0) \right| \leq O\left(\epsilon_{\text{eval}}\right).
\]

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 $i = \arg\max_{t \in [T]} \hat{V}^*_t(s_0)$. Taking union bound, we have that the following equation holds with probability $1 - \delta$.

\[
V^*(s_0) - V^i(s_0) \leq V^*(s_0) - \hat{V}^i(s_0) + O(\epsilon_{\text{eval}}) \\
\leq V^*(s_0) - \hat{V}^i(s_0) + O(\epsilon_{\text{eval}}) \\
\leq V^*(s_0) - V^i(s_0) + O(\epsilon_{\text{eval}}) \\
\leq O(\epsilon_{\text{eval}} + \epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}).
\]

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 $\epsilon$-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

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

episodes. Each time we run MDP-RBUCBI, we use $K = O\left(\frac{S^2 A}{\epsilon_{\text{ucb}}} \text{polylog} \left( S, A, \frac{1}{\epsilon_{\text{ucb}}} \right) \right)$ episodes. Each time we run MDP-Evaluation, we use

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

episodes(See discussion under MDP-Evaluation(Algorithm 7)). We run $\iota_0$ times RFKSP, MDP-RBUCBI, and MDP-Evaluation in MDP-Full. Summing up, we use

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

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 ∀k ∈ [K], if \( \tilde{P}_{s,a} \in \mathcal{P}_{s,a}^k \) and \( \mathbb{E} [\tilde{R}(s,a)] \in \mathcal{R}_{s,a}^k \) holds for any \((s,a)\), for ∀h ∈ [H] and ∀h-reaching state \( s_h \), \( \tilde{V}_h^k(s_h) \geq \tilde{V}_h^k(s_h) \).

**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 \( h = H + 1 \),

\[
\nabla_{H+1}^k(s_{H+1}) = \tilde{V}_{H+1}^k(s_{H+1}) = 0.
\]

Suppose the target equation holds for \( h + 1 \), then for ∀a,

\[
\tilde{Q}_h^k(s_h, a) = \min \left( \tilde{r}^k(s_h, a) + \max_{p \in \tilde{P}_{s_h,a}^k} \tilde{p} \tilde{V}_{h+1}^k, 1 \right)
\geq \min \left( \mathbb{E} [\tilde{R}(s_h, a)] + \tilde{P}_{s_h,a}^k \tilde{V}_{h+1}^k, 1 \right),
\]

\[
\geq \min \left( \mathbb{E} [\tilde{R}(s_h, a)] + \tilde{P}_{s_h,a}^k \tilde{V}_{h+1}^k, 1 \right),
\]

\[
= \tilde{Q}_h^*(s_h, a).
\]  

(7)

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

\[
\nabla_h^k(s_h) = \max_a \tilde{Q}_h^k(s_h, a) \geq \mathbb{E}_{a \sim \pi'_h(s_h)} \tilde{Q}_h^k(s_h, a) \geq \mathbb{E}_{a \sim \pi'_h(s_h)} \tilde{Q}_h^*(s_h, a) = \tilde{V}_h^*(s_h).
\]

(9)

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

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

**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

\[
\mathbb{E}_{\Gamma_K} \left\{ \sum_{k=1}^{K} \left[ \tilde{V}_{1}^k(s_0) - \tilde{V}_{1}^*k(s_0) \right] \mathbf{1}_{G_K} \right\} \leq \mathbb{E}_{\Gamma_K} \left\{ \sum_{k=1}^{K} \left[ \nabla_1^k(s_0) - \tilde{V}_{1}^*k(s_0) \right] \mathbf{1}_{G_{k-1}} \right\}
\]

\[
= \sum_{k=1}^{K} \mathbb{E}_{\Gamma_{k-1}} \left\{ \left[ \nabla_1^k(s_0) - \tilde{V}_{1}^*k(s_0) \right] \mathbf{1}_{G_{k-1}} \right\}.
\]

(10)

For a single episode \( k \), using the definition of \( \nabla \) and \( \tilde{V} \), we can turn the difference of \( \tilde{V}_1^k(s_0) \) and \( \tilde{V}_1^*k(s_0) \) into the expectation form of some term on the trajectory of the \( k \)-th episode. Here we introduce the cut-off indicator \( I_K^k \) 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 $h' > h$.

\[
\begin{align*}
&\nabla_1^k (s_1^k) - \nabla_1^k (s_{\pi_2}^k) = \nabla_1^k (s_1^k) I_1^k - \nabla_1^k (s_{\pi_2}^k) I_1^k \\
&\leq \mathbb{E}_{a_1^k \sim \pi_2^k} \left( \left( \max_{p \in \mathcal{P}_{s_1^k,a_1^k}^k} p - \hat{P}_{s_1^k,a_1^k}^k \right) \nabla_2^k I_1^k + \mathbb{E}_{a_2^k \sim \hat{P}_{s_2^k,a_2^k}^k} \left( \nabla_2^k (s_2^k) - \nabla_2^k (s_{\pi_2}^k) \right) I_1^k \right) + \mathbb{E}_{a_1^k \sim \pi_2^k} \left( \left( \max_{p \in \mathcal{P}_{s_1^k,a_1^k}^k} p - \hat{P}_{s_1^k,a_1^k}^k \right) \nabla_2^k I_1^k + \mathbb{E}_{a_2^k \sim \hat{P}_{s_2^k,a_2^k}^k} \left( \nabla_2^k (s_2^k) - \nabla_2^k (s_{\pi_2}^k) \right) I_1^k \right) + 2\mathbb{E}_{a_1^k \sim \pi_1^k} (I_1^k - I_2^k) \\
&\leq \mathbb{E}_{a_1^k \sim \pi_2^k} \left( \left( \max_{p \in \mathcal{P}_{s_1^k,a_1^k}^k} p - \hat{P}_{s_1^k,a_1^k}^k \right) \nabla_2^k I_1^k + \mathbb{E}_{a_2^k \sim \hat{P}_{s_2^k,a_2^k}^k} \left( \nabla_2^k (s_2^k) - \nabla_2^k (s_{\pi_2}^k) \right) I_1^k \right) + 2\mathbb{E}_{a_1^k \sim \pi_1^k} (I_1^k - I_2^k) \\
&\leq \mathbb{E}_{a_1^k \sim \pi_2^k} \left( \left( \max_{p \in \mathcal{P}_{s_1^k,a_1^k}^k} p - \hat{P}_{s_1^k,a_1^k}^k \right) \nabla_2^k I_1^k + \mathbb{E}_{a_2^k \sim \hat{P}_{s_2^k,a_2^k}^k} \left( \nabla_2^k (s_2^k) - \nabla_2^k (s_{\pi_2}^k) \right) I_1^k \right) + 2\mathbb{E}_{a_1^k \sim \pi_1^k} (I_1^k - I_2^k) \\
&\leq \mathbb{E}_{a_1^k \sim \pi_2^k} \left( \left( \max_{p \in \mathcal{P}_{s_1^k,a_1^k}^k} p - \hat{P}_{s_1^k,a_1^k}^k \right) \nabla_2^k I_1^k + \mathbb{E}_{a_2^k \sim \hat{P}_{s_2^k,a_2^k}^k} \left( \nabla_2^k (s_2^k) - \nabla_2^k (s_{\pi_2}^k) \right) I_1^k \right) + 2\mathbb{E}_{a_1^k \sim \pi_1^k} (I_1^k - I_2^k)
\end{align*}
\]

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

\[
\begin{align*}
&\mathbb{E}_{\gamma_K} \left\{ \sum_{k=1}^{K} \left[ \nabla_1^k (s_0) - \nabla_1^k (s_{\pi_2}^k) \right] 1_{G_K} \right\} \leq \sum_{k=1}^{K} \mathbb{E}_{\gamma_K} \left\{ \left[ \sum_{h=1}^{H} \left( \max_{p \in \mathcal{P}_{s_1^h,a_1^h}^k} p - \hat{P}_{s_1^h,a_1^h}^k \right) \nabla_2^h I_1^h \right] 1_{G_{k-1}} \right\} + \sum_{k=1}^{K} \mathbb{E}_{\gamma_K} \left\{ \left[ \sum_{h=1}^{H} \left( \nabla_2^h (s_2^h) - \nabla_2^h (s_{\pi_2}^h) \right) I_1^h \right] 1_{G_{k-1}} \right\} \\
&\quad + 2\mathbb{E}_{\gamma_K} \left[ \left( \sum_{k=1}^{K} 1_{I_{h+1}^k = 0} \right) 1_{G_0} \right] \\
\end{align*}
\]

For simplicity, we use the following notations. We denote

\[
\begin{align*}
M_1 &= \sum_{k=1}^{K} \mathbb{E}_{\gamma_K} \left\{ \left[ \sum_{h=1}^{H} \left( \max_{p \in \mathcal{P}_{s_1^h,a_1^h}^k} p - \hat{P}_{s_1^h,a_1^h}^k \right) \nabla_2^h I_1^h \right] 1_{G_{k-1}} \right\} , \\
M_2 &= \sum_{k=1}^{K} \mathbb{E}_{\gamma_K} \left\{ \left[ \sum_{h=1}^{H} \left( \nabla_2^h (s_2^h) - \nabla_2^h (s_{\pi_2}^h) \right) I_1^h \right] 1_{G_{k-1}} \right\} \\
M_3 &= \mathbb{E}_{\gamma_K} \left[ \left( \sum_{k=1}^{K} 1_{I_{h+1}^k = 0} \right) 1_{G_0} \right] .
\end{align*}
\]

Thus our target equation turns into

\[
\begin{align*}
\mathbb{E}_{\gamma_K} \left\{ \sum_{k=1}^{K} \left[ \nabla_1^k (s_0) - \nabla_1^k (s_{\pi_2}^k) \right] 1_{G_K} \right\} \leq M_1 + M_2 + 2M_3 .
\end{align*}
\]

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

\[
\begin{align*}
\mathbb{E}_{\gamma_K} \left\{ \sum_{k=1}^{K} \left[ \nabla_1^k (s_0) - \nabla_1^k (s_{\pi_2}^k) \right] 1_{G_K} \right\} \leq O(S \sqrt{AK \log(S, A, K)^2}) ,
\end{align*}
\]

holds if $K \geq \Omega(S^2 A)$. 

\[\square\]
Lemma E.4. In Lemma 6.8, 
\[ M_1 \leq O(\sqrt{K}\sqrt{\text{polylog}(S, A, K)\frac{I_{\text{conf}}}{2}}), \]
\[ M_2 \leq O(\sqrt{SA}\sqrt{\text{polylog}(S, A, K)\frac{I_{\text{conf}}}{2}}), \]
\[ M_3 \leq O(S\text{polylog}(S, A, K)\frac{I_{\text{conf}}}{2}). \]

We further denote 
\[ M_4 = \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] N^k(s_h^k, a_h^k) \right] \mathbb{1}_{G_k-1}. \]

And we have that 
\[ M_4 \leq O(S\text{polylog}(S, A, K)\frac{I_{\text{conf}}}{2}). \]

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

For \( M_2 \),
\[ M_2 = \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) - \mathbb{E}R(s_h^k, a_h^k) I_h^k \right] \mathbb{1}_{G_k-1} \]
\[ \leq C \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) + \tilde{r}_k I_h^k \mathbb{1}_{G_k-1} \right] \]
\[ \leq C \mathbb{E}_{\Gamma_k} \sqrt{M_4\frac{I_{\text{conf}}}{2}} \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) + \tilde{r}_k I_h^k \mathbb{1}_{G_k-1} \right]. \]

Here \( \mathbb{E}V_k(s, a) = \mathbb{E}\frac{1}{N_k(s, a)} \sum_{i=1}^{N_k(s, a)} (r_i(s, a) - \tilde{r}_k(s, a))^2 \leq \mathbb{E}r(s, a)^2 \leq \mathbb{E}|r(s, a)|. \) Thus if \( K \geq \Omega(SA) \),
\[ M_2 \leq C \mathbb{E}_{\Gamma_k} \sqrt{M_4\frac{I_{\text{conf}}}{2}} \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] r(s_h^k, a_h^k) + \tilde{r}_k I_h^k \mathbb{1}_{G_k-1} \right] \]
\[ \leq C \mathbb{E}_{\Gamma_k} \sqrt{M_4\frac{I_{\text{conf}}}{2}} \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) + \tilde{r}_k I_h^k \mathbb{1}_{G_k-1} \right]. \]

For \( M_1 \),
\[ M_1 = \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) \mathbb{1}_{G_k-1} \right] \]
\[ \leq C \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) \mathbb{1}_{G_k-1} \right] \]
\[ \leq C \mathbb{E}_{\Gamma_k} \sqrt{M_4\frac{I_{\text{conf}}}{2}} \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) \mathbb{1}_{G_k-1} \right]. \]

We further denote
\[ M_5 = \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \mathbb{E}_{s, a} \left[ s, a \right] \tilde{r}_k(s_h^k, a_h^k) \mathbb{1}_{G_k-1} \right]. \]
By the above notations, we have $M_1 \leq C \sqrt{SM_4 \text{conf}} \cdot \sqrt{M_5} + 2C \cdot M_4 \text{St}_{\text{conf}}$. Next, we try to bound $M_5$ with $M_1$ and thus construct a recursion structure.

$$M_5 = \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ V \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right) I_h^k \right] 1_{G_{k-1}} \right\}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k - \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right)^2 \right] I_h^k 1_{G_{k-1}} \right\}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ \mathcal{V}_{h+1}^k (s_{h+1}^k)^2 - \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right)^2 \right] I_h^k 1_{G_{k-1}} \right\} + \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left[ \left( I_{H+1}^k = 1 \right) 1_{G_{k-1}} \right]$$

$$\leq \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ \mathcal{V}_{h+1}^k (s_{h+1}^k)^2 - \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right)^2 \right] I_h^k 1_{G_{k-1}} \right\} + M_3.$$  

In line 12, we use the fact that

$$\sum_{h=1}^{H} \mathcal{V}_{h+1}^k (s_{h+1}^k)^2 I_h^k \leq \sum_{h=1}^{H} \mathcal{V}_{h+1}^k (s_{h+1}^k)^2 I_h^k + 1_{H+1} = 0.$$  

(14)

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

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

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

$$\sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \sum_{h=1}^{H} \left[ \mathcal{V}_h^k (s_h^k)^2 I_h^k \right] 1_{G_{k-1}} \leq \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \sum_{h=1}^{H} \left[ \mathcal{V}_h^k (s_h^k)^2 + \max_{p \in P_{s_h^k, a_h}^k} p \mathcal{V}_{h+1}^k \right] I_h^k 1_{G_{k-1}}.$$  

(15)

Here the equation 15 holds since $a_h^k$ is fixed given $s_h^k$ and policy $\pi^k$. 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.

$$M_5 \leq \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ \mathcal{V}_h^k (s_h^k)^2 - \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right)^2 \right] I_h^k 1_{G_{k-1}} \right\} + M_3$$

$$\leq \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ \mathcal{V}_h^k (s_h^k)^2 + \max_{p \in P_{s_h^k, a_h}^k} p \mathcal{V}_{h+1}^k \right] - \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right)^2 \right\} I_h^k 1_{G_{k-1}} \right\} + M_3$$

$$+ \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ \mathcal{V}_h^k (s_h^k)^2 + \max_{p \in P_{s_h^k, a_h}^k} p \mathcal{V}_{h+1}^k \right] - \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right)^2 \right\} I_h^k 1_{G_{k-1}} \right\} + M_3$$

$$\leq 4M_2 + 3 \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \left\{ \sum_{h=1}^{H} \left[ \mathcal{V}_h^k (s_h^k)^2 + \max_{p \in P_{s_h^k, a_h}^k} p \mathcal{V}_{h+1}^k \right] - \left( \tilde{P}_{s_k^h, a_h}^k \mathcal{V}_{h+1}^k \right)^2 \right\} I_h^k 1_{G_{k-1}} \right\} + M_3$$

(17)

$$\leq 4M_2 + 3K + 3M_1 + M_3.$$  

(18)
To be more specific, we derive line 17 as follows. Since $I_k^*$ and $\Gamma_{k-1}$ are independent of $a_k^*$, we can focus on the difference of squares. For the second difference of squares,

$$
\left(\mathbb{E}\left(s_k^*, a_k^*\right) + \max_{p \in P_k^*} \mathbb{E}\left(s_k^*, a_k^* \mathbb{E}\left|h_{k+1}\right| \right)^2 \right) - \left(\mathbb{E}\left(s_k^*, a_k^* \mathbb{E}\left|h_{k+1}\right| \right)^2 \right)
$$

$$
\leq \left(\mathbb{E}\left(s_k^*, a_k^*\right) + \max_{p \in P_k^*} \mathbb{E}\left(s_k^*, a_k^* \mathbb{E}\left|h_{k+1}\right| \right)^2 \right) \cdot \left(\mathbb{E}\left(s_k^*, a_k^*\right) + \max_{p \in P_k^*} \mathbb{E}\left(s_k^*, a_k^* \mathbb{E}\left|h_{k+1}\right| \right)^2 \right)
$$

Combining line 18 and $M_1 \leq C \sqrt{SM_4 \epsilon_{\text{conf}}} \cdot \sqrt{M_5} + 2C \cdot M_4 \epsilon_{\text{conf}}$, we can solve $M_1$ satisfies that

$$
M_1 \leq O((S \sqrt{A} \sqrt{K} + S^2 A) \text{polylog}(S, A, K)^2_{\epsilon_{\text{conf}}})
$$

$$
\leq O(S \sqrt{A} \sqrt{K} \text{polylog}(S, A, K)^2) \quad \text{If } K \geq \Omega(S^2 A).
$$

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

$$
V_1^*(s_0) - V_1^*(s_0) \leq O(\epsilon_{\text{ucb}} + \epsilon_{\text{ksp}}).
$$

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

$$
\mathbb{E}_{\Gamma_k} \left\{ \left[ \hat{V}_1^*(s_0) - \hat{V}_1^{\pi_k}(s_0) \right] \mathbb{1}_{G_K} \right\} \leq 8 \left[ O\left(\frac{S \sqrt{A} \sqrt{K} \text{polylog}(S, A, K)^2_{\epsilon_{\text{conf}}}}{\sqrt{K}}\right)\right].
$$

Since $\left[ \hat{V}_1^*(s_0) - \hat{V}_1^{\pi_{k_1}}(s_0) \right] \mathbb{1}_{G_K} \geq 0$ also always hold, we apply Markov inequality and derive that with probability at least $\frac{1}{16}$,

$$
\left[ \hat{V}_1^*(s_0) - \hat{V}_1^{\pi_{k_1}}(s_0) \right] \mathbb{1}_{G_K} \leq 8 \mathbb{E}_{\Gamma_k} \left\{ \left[ \hat{V}_1^*(s_0) - \hat{V}_1^{\pi_{k_1}}(s_0) \right] \mathbb{1}_{G_K} \right\}.
$$

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

$$
\hat{V}_1^*(s_0) - \hat{V}_1^{\pi_{k_1}}(s_0) = \left[ \hat{V}_1^*(s_0) - \hat{V}_1^{\pi_{k_1}}(s_0) \right] \mathbb{1}_{G_K}
$$

$$
\leq O\left(\frac{S \sqrt{A} \sqrt{K} \text{polylog}(S, A, K)^2_{\epsilon_{\text{conf}}}}{\sqrt{K}}\right).
$$

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

$$V_1^*(s_0) - \hat{V}_1^{\pi_k}(s_0) \leq O(\epsilon_{\text{uch}}).$$

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

$$V_1^*(s_0) - \hat{V}_1^{\pi_k}(s_0) \leq \hat{V}_1^{**}(s_0) - \hat{V}_1^{\pi_k}(s_0) + 2\epsilon_{\text{ksp}}$$
$$\leq \hat{V}_1^{**}(s_0) - \hat{V}_1^{\pi_k}(s_0) + 2\epsilon_{\text{ksp}}$$
$$\leq O(\epsilon_{\text{uch}} + \epsilon_{\text{ksp}}).$$

\[\Box\]

### E.2. MDP-Evaluation

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

**Algorithm 7** MDP-Evaluation

1. **Input**: MDP $M$, Policy $\pi$, $\epsilon_{\text{eval}}$, $\delta_{\text{eval}}$.
2. **Initialization**: $V_{H+1}^k(s) = 0$, $V_k^H(s) = 0, \forall k, s,$
3. Set $\epsilon_{\text{ksp}} \gets \epsilon_{\text{eval}}$, $\delta_{\text{ksp}} \gets \frac{\delta_{\text{eval}}}{2T}$.
4. Run $T = \log\left(\frac{2}{\delta_{\text{eval}}}\right)$ times independently.
5. for $t = 1, 2, \ldots, T$ do
6. $M_t \leftarrow \text{RFKSP}(M, \epsilon_{\text{ksp}}, \delta_{\text{ksp}})$.
7. Use $K = \tilde{O}\left(\frac{S^2A}{\epsilon_{\text{eval}}}\right)$ episodes.
8. for episode $k = 1, 2, \ldots, K$ do
9. for step $h = H, H - 1, H - 2, \ldots, 1$ do
10. Compute $Q_h^k(s, a)$ as in equation 1.
11. Compute $\nabla_h^k(s) = \mathbb{E}_{a \sim \pi_k(.|s)}[Q_h^k(s, a)]$.
12. end for
13. Play policy $\pi$, collect trajectory $\tau_k$.
14. Calculate $P_{k+1}, R_{k+1}$ based on $\Gamma_k$.
15. end for
16. Randomly select $\nabla_1^k(s_0)$, denote as $\nabla^1(s_0)$.
17. end for
18. **Output**: $\hat{V}_1^*(s_0) = \min\{\nabla_1(s_0), \ldots, \nabla_T(s_0)\}$.

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

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

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

$$K = O\left(\frac{S^9A^3\epsilon_{\text{eval}}}{\epsilon_{\text{eval}}}\text{polylog}\left(S, A, \frac{1}{\epsilon_{\text{eval}}}\right)\right) + O\left(\frac{S^2A\epsilon_{\text{eval}}}{\epsilon_{\text{eval}}}\text{polylog}\left(S, A, \frac{1}{\epsilon_{\text{eval}}}\right)\right).$$
Lemma E.6. In MDP-Evaluation (Algorithm 7), for all $k \in [K]$, 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 all $h \in [H]$ and all $h$-reachable state $s_h$, $\nabla^k_h(s_h) \geq \hat{V}^\pi_h(s_h)$.

Proof. For fixed $k$, we do induction on $h = H + 1, H, \ldots, 1$. When $h = H + 1$,$$
\nabla^k_{H+1}(s_{H+1}) = \hat{V}^\pi_{H+1}(s_{H+1}) = 0.
$$
Suppose the equation holds for $h + 1$, then for all $a$,
$$\nabla^k_h(s_h, a) = \min \left( \tau^k(s_h, a) + \max_{\mathcal{P}^k_{s_h,a}} \mathbb{P}^k \mathcal{V}^k_{h+1}, 1 \right) 
\geq \min \left( \mathbb{E}\tilde{R}(s_h, a) + \tilde{P}_{s_h,a} \mathcal{V}^k_{h+1}, 1 \right) 
\geq \min \left( \mathbb{E}\tilde{R}(s_h, a) + \tilde{P}_{s_h,a} \hat{V}^\pi_{h+1}, 1 \right) 
= \hat{Q}^\pi_h(s_h, a).
$$
Here line 23 holds since we assume $\tilde{P}_{s_h,a} \in \mathcal{P}^k_{s,a}$. 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, $\mathcal{V}^k_{h+1}(s_{h+1}) \geq \hat{V}^\pi_{h+1}(s_{h+1})$. Thus line 24 also holds. As for line 25, by definition we have $\hat{Q}^\pi_h(s_h, a) = \mathbb{E}\tilde{R}(s_h, a) + \tilde{P}_{s_h,a} \hat{V}^\pi_{h+1}$.
We are left to prove that $\hat{Q}^\pi_h(s_h, a) \leq 1$. We can take $\pi'$ where $\pi' = \pi$ except $\pi'_h(s_h) = a$. Hence by our reward assumption, $\hat{Q}^\pi_h(s_h, a) = \hat{V}^\pi_h(s_h) \leq 1$. We further conclude our induction by
$$\nabla^k_h(s_h) = \mathbb{E}_{a \sim \pi_h(s_h)} \hat{Q}^k_h(s_h, a) \geq \mathbb{E}_{a \sim \pi_h(s_h)} \hat{Q}^\pi_h(s_h, a) = \hat{V}^\pi_h(s_h).
$$

Lemma E.7. In each independent running time $t \in [T]$ in MDP-Evaluation (Algorithm 7),
$$\mathbb{E}_{\Gamma_K} \left\{ \sum_{k=1}^{K} \left[ \mathcal{V}^k_1(s_0) - \hat{V}^\pi_1(s_0) \right] 1_{G_K} \right\} \leq O(S \sqrt{AK} \text{polylog}(S, A, K) t_{\text{conf}}^2).
$$

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 $\mathcal{V}_t(s_0)$ satisfies that with probability $\frac{1}{2}$,
$$0 \leq \mathcal{V}_t(s_0) - \hat{V}^\pi_1(s_0) \leq O(\epsilon_{\text{eval}}).
$$

Proof. We focus on a fixed independent running time $t \in [T]$. From Lemma E.7 we have that
$$\mathbb{E}_{\Gamma_K} \left\{ \sum_{k=1}^{K} \left[ \mathcal{V}^k_1(s_0) - \hat{V}^\pi_1(s_0) \right] 1_{G_K} \right\} \leq O(S \sqrt{AK} \text{polylog}(S, A, K) t_{\text{conf}}^2).
$$
By Lemma E.6, we know that $\left[ \mathcal{V}^k_1(s_0) - \hat{V}^\pi_1(s_0) \right] 1_{G_K} \geq 0$. And therefore $\mathbb{E}_{\Gamma_K} \left\{ \left[ \mathcal{V}^k_1(s_0) - \hat{V}^\pi_1(s_0) \right] 1_{G_K} \right\}$ is positive for any $k \in [K]$. We randomly choose episode $k_1 \in [K]$ and denote $\mathcal{V}^{k_1}_1(s_0)$ as $\mathcal{V}_t(s_0)$. Using Markov inequality twice and taking union bound, we have that with probability at least $\frac{1}{2}$,
$$\mathbb{E}_{\Gamma_K} \left\{ \left[ \mathcal{V}^{k_1}_1(s_0) - \hat{V}^\pi_1(s_0) \right] 1_{G_K} \right\} \leq O(S \sqrt{AK} \text{polylog}(S, A, K) t_{\text{conf}}^2),
$$
$$\mathbb{E}_{\Gamma_K} \left\{ \left[ \mathcal{V}^{k_1}_1(s_0) - \hat{V}^\pi_1(s_0) \right] 1_{G_K} \right\} \leq 16 \mathbb{E}_{\Gamma_K} \left\{ \left[ \mathcal{V}^{k_1}_1(s_0) - \hat{V}^\pi_1(s_0) \right] 1_{G_K} \right\}. 
$$
Since $\delta_{\text{exp}} = \frac{1}{4}$, by further setting $\delta_{\text{conf}} = \frac{1}{16S^2AK}$, we have that with probability at least $\frac{5}{8} = 1 - 2S^2AK\delta_{\text{conf}} - \delta_{\text{exp}}$, $G_K$ holds. By taking union bound with equation 26 and equation 27, we have that with probability at least $\frac{1}{2}$, 

\[
\begin{align*}
\left[ \hat{V}_1^{k_1}(s_0) - \hat{V}_1^{\pi}(s_0) \right] &= \left[ \hat{V}_1^{k_1}(s_0) - \hat{V}_1^{\pi}(s_0) \right] 1_{G_K} \\
&\leq O \left( \frac{S\sqrt{A}}{\sqrt{K}} \text{polylog}(S, A, K) \delta_{\text{conf}} \right) \\
&= O \left( \frac{S\sqrt{A}}{\sqrt{K}} \text{polylog}(S, A, K) \right) .
\end{align*}
\]  

(28)

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

\[
0 \leq \left[ \hat{V}_1^{k_1}(s_0) - \hat{V}_1^{\pi}(s_0) \right] \leq O(\epsilon_{\text{eval}}).
\]

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

\[
|\hat{V}_1^{\pi}(s_0) - V_1^{\pi}(s_0)| \leq O(\epsilon_{\text{eval}}).
\]

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

\[
0 \leq \left[ \hat{V}_t(s_0) - \hat{V}_1^{\pi}(s_0) \right] \leq O(\epsilon_{\text{eval}}).
\]

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

\[
0 \leq V_i(s_0) - \hat{V}_1^{\pi}(s_0) \leq O(\epsilon_{\text{eval}}).
\]

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

\[
0 \leq \min_{t \in [T]} V_t(s_0) - \hat{V}_1^{\pi}(s_0) \\
\leq \min_{t \in [T]} \hat{V}_t(s_0) - V_1^{\pi}(s_0) + \epsilon_{\text{eval}} \\
\leq \hat{V}_i - V_1^{\pi}(s_0) + \epsilon_{\text{eval}} \\
\leq \hat{V}_i - \hat{V}_1^{\pi}(s_0) + 2\epsilon_{\text{eval}} \\
\leq O(\epsilon_{\text{eval}}).
\]

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

\[
|\min_{t \in [T]} V_t(s_0) - V^{\pi}(s_0)| \leq O(\epsilon_{\text{eval}}).
\]

Here $\min_{t \in [T]} V_t(s_0)$ 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 $\mathcal{G} (S, A, B, P, R, H, \mu_0), \epsilon, \delta$
2: Set $\epsilon_{\text{ksp}}, \epsilon_{\text{ucb}}, \epsilon_{\text{eval}} = O(\epsilon), \delta_{\text{ksp}} = \frac{1}{16}, \delta_{\text{eval}} = \frac{\delta}{16}$.
3: Run $T = \log \left( \frac{16}{\epsilon^2} \right)$ times independently.
4: for $t = 1, 2, \ldots, T$
5: $\tilde{G}_t \leftarrow \text{RFKSP} (G, \epsilon_{\text{ksp}}, \delta_{\text{ksp}})$.
6: $\pi^t = (\mu^t, \nu^t) \leftarrow \text{MG-RBUCB}(\tilde{G}_t, \epsilon_{\text{ucb}})$.
7: $\hat{V}^*_t, \hat{V}^t (s_0), \hat{V}^{\mu^t, \nu^t} (s_0) \leftarrow \text{MG-Evaluation}(G, \epsilon_{\text{eval}}, \delta_{\text{eval}}, \pi^t)$.
8: end for
9: $i \leftarrow \arg \min_{t \in [T]} \left[ \hat{V}^{*, \nu^t} (s_0) - \hat{V}^{\mu^t, \nu^t} (s_0) \right]$.
10: **Output:** $\pi^i$.

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

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

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

**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 $1/2$,

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

As we run the subroutine $T = \log \left( \frac{16}{\epsilon^2} \right)$ times independently, with probability $1 - \frac{\delta}{2}$, there exists $j \in [T]$ that the above equation holds. By Theorem F.5, the estimation $\hat{V}^{*, \nu^j} (s_0)$ and $\hat{V}^{\mu^j, \nu^j} (s_0)$ returned by MDP-Evaluation satisfy that with probability $1 - 2\delta_{\text{eval}},$

$$|\hat{V}^{*, \nu^j} (s_0) - \hat{V}^{*} (s_0)| \leq O(\epsilon_{\text{eval}}).$$

$$|\hat{V}^{\mu^j, \nu^j} (s_0) - \hat{V}^{\mu} (s_0)| \leq O(\epsilon_{\text{eval}}).$$

Since we set $\delta_{\text{eval}} = \frac{\delta}{16}$ in MG-Full, with probability $1 - \frac{\delta}{2}$, the above equation hold for $\forall t \in [T]$. Suppose we denote $i = \arg \min_{t \in [T]} \left[ \hat{V}^{*, \nu^t} (s_0) - \hat{V}^{\mu^t, \nu^t} (s_0) \right]$. Taking union bound, we have that the following equation holds with probability $1 - \delta$.

$$V^{*, \nu^i} (s_0) - V^{\mu^i, \nu^i} (s_0) \leq V^{*, \nu^j} (s_0) - V^{\mu^j, \nu^j} (s_0) + O(\epsilon_{\text{eval}}) \leq V^{*, \nu^i} (s_0) - \hat{V}^{\mu^j, \nu^j} (s_0) + O(\epsilon_{\text{eval}}) \leq V^{*, \nu^i} (s_0) - \hat{V}^{\mu^i, \nu^i} (s_0) + O(\epsilon_{\text{eval}}) \leq O \left( \epsilon_{\text{eval}} + \epsilon_{\text{ucb}} + \epsilon_{\text{ksp}} \right).$$

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 $\epsilon$-approximate NE policy pair. Each time we run RFKSP with $\delta_{\text{ksp}} = \frac{1}{16}$ and $\epsilon_{\text{ksp}} = O(\epsilon)$, we use

$$K = O \left( \frac{S^9 A^3 B^3 \epsilon_{\text{ksp}}}{\epsilon} \text{polylog} \left( S, A, B, \frac{1}{\epsilon_{\text{ksp}}} \right) \right) = O \left( \frac{S^9 A^3 B^3}{\epsilon} \text{polylog} \left( S, A, B, \frac{1}{\epsilon} \right) \right).$$
episodes. Each time we run MG-RBUCBI, we use \( K = \tilde{O}\left(\frac{S^2AB}{\epsilon_{eval}}\right) \) episodes. Each time we run MG-Evaluation, we use
\[
K = O\left(\frac{S^0A^3B^3\epsilon^2_{eval}}{\epsilon_{eval}}\log\left(S, A, \frac{1}{\epsilon_{eval}}\right)\right) + O\left(\frac{S^2AB^2\epsilon^2_{eval}}{\epsilon^2}\log\left(S, A, B, \frac{1}{\epsilon}\right)\right)
\]
episodes (See discussion under MDP-Evaluation (Algorithm 7)). Summing up, we have that we use
\[
K = O\left(\frac{S^0A^3B^3\epsilon^2_{eval}}{\epsilon_{eval}}\log\left(S, A, B, \frac{1}{\epsilon}\right)\right)
\]
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 \( \hat{P}_{s,a,b} \in \mathcal{P}_{s,a,b}^k \) and \( \mathbb{E}\left[\hat{R}(s, a, b)\right] \in \mathcal{R}_{s,a,b}^k \) holds for any \( (s, a, b) \), for \( \forall h \in [H] \) and \( \forall h \)-reachable state \( s_h \),
\[
\hat{V}_h^{k}(s_h) \geq \hat{V}_h^{\pi^k}(s_h) \geq \hat{V}_h^{\pi^k}(s_h) \geq \hat{V}_h^{k}(s_h).
\]

Proof. In the above equation, \( \hat{V}_h^{\pi^k}(s_h) \geq \hat{V}_h^{k}(s_h) \geq \hat{V}_h^{k}(s_h) \) 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 \( h = H + 1 \), \( \hat{V}_h^{k}(s_{H+1}) = \hat{V}_h^{k}(s_{H+1}) = 0 \). Suppose the equation holds for \( h + 1 \), then for \( \forall (a, b) \in \mathcal{A} \times \mathcal{B} \),
\[
\hat{Q}_h^{k}(s_h, a, b) = \min\left(\pi^k(s_h, a, b) + \max_{P_{s_h,a,b} \in \mathcal{P}_{s_h,a,b}} P_{s_h,a,b} \hat{V}_h^{k}, 1\right)
\]
\[
\geq \min\left(\mathbb{E}R(s_h, a, b) + \hat{P}_{s_h,a,b} \hat{V}_h^{k+1}, 1\right) \tag{30}
\]
\[
\geq \min\left(\mathbb{E}R(s_h, a, b) + \hat{P}_{s_h,a,b} \hat{V}_h^{k+1}, 1\right) \tag{31}
\]
\[
= \hat{Q}_h^{k}(s_h, a, b). \tag{32}
\]

Here (30) holds since we assume \( \hat{P}_{s_h,a,b} \in \mathcal{P}_{s_h,a,b}^k \) and \( \mathbb{E}\left[\hat{R}(s, a, b)\right] \in \mathcal{R}_{s,a,b}^k \). And if \( \hat{P}_{s_h,a,b,s_{h+1}} \neq 0 \), \( s_{h+1} \) is \( h + 1 \)-reachable. So by induction (31) also holds. As for (32), we can take \( \pi^k \) where \((\mu^k, \nu^k) = (\mu^*, \nu^*)\) except \( \mu^k(s_h) = a, \nu^k(s_h) = b \). Hence by our reward assumption, \( \hat{Q}_h^{*}(s_h, a, b) = \hat{V}_h^{*,\nu^*}(s_h) \leq 1 \). We further conclude our induction by
\[
\hat{V}_h^{k}(s_h) = \mathbb{E}_{a \sim \mu^k(s_h), b \sim \nu^k(s_h)} \hat{Q}_h^{k}(s_h, a, b)
\]
\[
\geq \mathbb{E}_{a \sim s, b \sim \nu^k(s_h)} \hat{Q}_h^{k}(s_h, a, b) \tag{33}
\]
\[
\geq \mathbb{E}_{a \sim s, b \sim \nu^k(s_h)} \hat{Q}_h^{k}(s_h, a, b)
\]
\[
= \hat{V}_h^{*,\nu^k}(s_h).
\]

Here (33) holds by the property of CCE since there exists a deterministic policy to be the best response of \( \nu \).

Lemma F.3. In MG-RBUCBI (Algorithm 3), the expectation of regret concerning the auxiliary Markovian environment can be bounded by
\[
\mathbb{E}_{\Gamma_K} \left\{ \sum_{k=1}^{K} \left[ \hat{V}_1^{*,\nu^k}(s_0) - \hat{V}_1^{\mu^k,*}(s_0) \right] 1_{G_K} \right\} \leq O(S\sqrt{ABK}\log(S, A, B, K)^2_{conf}).
\]
Proof. By Lemma F.2, we have the following equation.

\[ \mathbb{E}_{\Gamma_K} \sum_{k=1}^{K} \left\{ \left[ \hat{V}^*_k(s_0) - \hat{V}^k(s_0) \right] 1_{G_K} \right\} \leq \mathbb{E}_{\Gamma_K} \sum_{k=1}^{K} \left\{ \left[ \bar{V}^k(s_0) - \bar{V}^k(s_0) \right] 1_{G_K} \right\} \]

\[ = \mathbb{E}_{\Gamma_K} \sum_{k=1}^{K} \left\{ \left[ \bar{V}^k(s_0) - \bar{V}^k(s_0) \right] 1_{G_K} \right\} + \mathbb{E}_{\Gamma_K} \sum_{k=1}^{K} \left\{ \left[ \bar{V}^k(s_0) - \bar{V}^k(s_0) \right] 1_{G_K} \right\}. \]

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.

\[ \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \sum_{h=1}^{H} \bar{V}^k_h(s_h)^2 I^k_h 1_{G_K - 1} \leq \sum_{k=1}^{K} \mathbb{E}_{\Gamma_k} \sum_{h=1}^{H} \bar{V}^k_h(s_h)^2 + \max_{\pi \in P_{s_h}^{a_h,b_h}} p^V_{h+1} \left( I^k_h 1_{G_K - 1} \right)^2. \]

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

\[ \mathbb{E}_{\Gamma_k} \left[ \bar{V}^k_h(s_h)^2 I^k_h 1_{G_K - 1} \right] = \mathbb{E}_{\Gamma_k} \left\{ \left[ \mathbb{E}_{\alpha \sim \mu^*_h(s_h), \beta \sim \nu^*_h(s_h)} \left( \bar{V}^k_h(s_h, a, b) + \max_{\pi \in P_{s_h}^{a_h,b_h}} p^V_{h+1} \right) I^k_h 1_{G_K - 1} \right]^2 \right\} \]

\[ \leq \mathbb{E}_{\Gamma_k} \left\{ \left[ \mathbb{E}_{\alpha \sim \mu^*_h(s_h), \beta \sim \nu^*_h(s_h)} \left( \bar{V}^k_h(s_h, a, b) + \max_{\pi \in P_{s_h}^{a_h,b_h}} p^V_{h+1} \right)^2 \right] I^k_h 1_{G_K - 1} \right\} \]

Here line 34 holds since \( \mathbb{E} [X]^2 \leq \mathbb{E} [X^2] \). Since \( \gamma_k \) is the trajectory following \( \pi^k \) 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 \( \frac{1}{2} \),

\[ V^*_1(s_0) - V^*_{\hat{\mu}_1}(s_0) \leq O \left( \epsilon_{ucb} + \epsilon_{esp} \right). \]

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

\[ \mathbb{E}_{\Gamma_K} \left[ \hat{V}^*_k(s_0) - \hat{V}^*_k(s_0) \right] 1_{G_K} \geq 16 \left[ O \left( \frac{S \sqrt{AB}}{\sqrt{K}} \text{polylog}(S, A, B, K), \epsilon_{conf} \right) \right]. \]

Since \( \left[ \hat{V}^*_k(s_0) - \hat{V}^*_k(s_0) \right] 1_{G_K} \geq 0 \) hold for \( \forall k \in [K] \), by Markov inequality, the following equation holds with probability at least \( \frac{15}{16} \).

\[ \left[ \hat{V}^*_k(s_0) - \hat{V}^*_k(s_0) \right] 1_{G_K} \geq 16 \mathbb{E}_{\Gamma_K} \left[ \hat{V}^*_k(s_0) - \hat{V}^*_k(s_0) \right] 1_{G_K}. \]
We set $K$ which satisfies that $O \left( \frac{S\sqrt{AB}}{\sqrt{K}} \right) = \epsilon_{\text{ucb}}$. Since $G_0$ holds, by Lemma 6.6 we have the following equation.

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

**F.2. MG-Evaluation**

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

**Algorithm 9 MG-Evaluation**

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

**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 $\hat{V}_1^{*}, \nu'(s_0)$ and $\hat{V}_1^{\mu}, \nu'(s_0)$ satisfy that

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

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

**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 $\mu$ 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 $\epsilon$-optimal policy by sampling at most $K$ episodes, where

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

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

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

**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.