Horizon-Free Reinforcement Learning in Polynomial Time: the Power of Stationary Policies

Zihan Zhang ZIHAN-ZH17@MAILS.TSINGHUA.EDU.CN
Tsinghua University

Xiangyang Ji XYJI@TSINGHUA.EDU.CN
Tsinghua University

Simon S. Du SSDU@CS.WASHINGTON.EDU
University of Washington

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

This paper gives the first polynomial-time algorithm for tabular Markov Decision Processes (MDP) that enjoys a regret bound *independent on the planning horizon*. Specifically, we consider tabular MDP with S states, A actions, a planning horizon H, total reward bounded by 1, and the agent plays for K episodes. We design an algorithm that achieves an $O\left(\mathrm{poly}(S, A, \log K)\sqrt{K}\right)$ regret in contrast to existing bounds which either has an additional polylog(H) dependency (Zhang et al., 2021b) or has an exponential dependency on S (Li et al., 2021b). Our result relies on a sequence of new structural lemmas establishing the approximation power, stability, and concentration property of stationary policies, which can have applications in other problems related to Markov chains.

1. Introduction

Tabular Markov Decision Process (MDP) is one of the most fundamental models for reinforcement learning (RL). The first algorithm that enjoys polynomial time and sample complexity guarantee at least dates back to 1990s (Kearns and Singh, 1998). However, despite of nearly two and half decades of research, the sample complexity on this fundamental model remains open. We study the canonical episodic time-homogeneous MDP with S states, A actions, planning horizon H, and total reward upper bounded by 1.[1]

The main challenges that differentiate RL and its special case, contextual bandits, are the unknown state-dependent transition and the long planning horizon. In contextual bandits, the planning horizon is one and there is no state-dependent transition to learn. Till today, it is unclear whether RL requires more samples than contextual bandits in the minimax sense.[2] Specifically, the lower bound for the RL setting considered in this paper is $\Omega\left(\sqrt{SAK}\right)$, which is the same for contextual bandits.

Due to these two challenges in RL, Jiang and Agarwal (2018) conjectured a poly(H) lower bound. Recent work refuted this conjecture by providing algorithms whose regret scales only *logarithmically* with H (Wang et al., 2020; Zhang et al., 2021b). Specifically, when the dependency

1. The upper bounded total reward is without the loss of generality. If the total reward is upper bounded by some $V_{\max} > 0$, then our regret will scale with $V_{\max}$.
2. For gap-dependent bounds, it has been shown that there is a gap between tabular MDP and contextual bandits (Xu et al., 2021).

Paper	Regret	PAC Bound
Zhang et al. (2021b)	$O\left(\left(\sqrt{SAK}+S^2A\right)\text{polylog}\left(S,A,K,H\right)\right)$	$O\left(\left(\frac{SA}{\epsilon^2}+\frac{S^2A}{\epsilon}\right)\text{polylog}\left(S,A,\frac{1}{\epsilon},H\right)\right)$
Li et al. (2021b)	-	$\frac{(SA)^{O(S)}}{\epsilon^5}$
This work	$O\left(\left(\sqrt{S^9A^3K}\right)\text{polylog}\left(S,A,K\right)\right)$	$O\left(\left(\frac{S^9A^3}{\epsilon^2}\right)\text{polylog}\left(S,A,\frac{1}{\epsilon}\right)\right)$
Contextual bandits lower bound	$\Omega\left(\sqrt{SAK}\right)$	$\Omega\left(\frac{SA}{\epsilon^2}\right)$

Table 1: Comparisons of our result with prior arts. S: number of states, A: number of actions, H: planning horizon, K: number of episodes, ϵ: target error.

on H is allowed, state-of-the-art result shows one can have an $O\left(\sqrt{SAK}\cdot\text{polylog}(S,A,H,K)\right)$ regret (Zhang et al., 2021b). More recently, Li et al. (2021b) gave a surprising result showing the dependency on H is not necessary. However, their sample complexity has an exponential dependency on the number of states. Therefore, one natural and conceptually important open question is:

Is there an algorithm whose regret (1) scales polynomially with S and A, and (2) does not depend on H?

1.1. Our Result

Our paper answers this question positively.

Theorem 1 *Suppose the reward at each step is non-negative and the total reward of each episode is bounded by* 1. *Given a failure probability $0<\delta<1$, then with probability at least $1-\delta$, the regret of our algorithm is bounded by $O\left(\left(\sqrt{S^9A^3K}\right)\text{polylog}\left(S,A,\log K,\log 1/\delta\right)\right)$ where S is the number of episodes, A is the number of actions, and K is the total number of episodes.*

Using a standard reduction (Jin et al., 2018), this regret bound also implies a PAC sample complexity of $O\left(\frac{S^9A^3\text{polylog}(S,A,1/\epsilon)}{\epsilon^2}\right)$, where $0<\epsilon<1$ is the target error. In Table 1, we compare our results with prior arts.

Several comments are in sequel. First, this is the first polynomial algorithm for tabular MDP whose regret has no dependence on H. Therefore, we achieve an exponential improvement over Li et al. (2021b). Second, our dependency on K (or $1/\epsilon$) is optimal up to logarithimic factors. Third, the dependencies on S and A are not optimal. A fundamental open problem is to design an algorithm for tabular MDP whose regret bound exactly matches the lower bound of contextual bandits.

1.2. Related Work

We focus on papers that study episodic tabular MDP. Other closely related settings include infinite-horizon discounted MDP, learning with a generative model, etc. We believe our techniques can be applied to those settings and obtain improvements, which we leave as future work.

Tabular MDP. There is a long list of sample complexity guarantees for tabular MDP (Kearns and Singh, 2002; Brafman and Tennenholtz, 2003; Kakade, 2003; Strehl et al., 2006; Strehl and Littman, 2008; Kolter and Ng, 2009; Bartlett and Tewari, 2009; Jaksch et al., 2010; Szita and Szepesvári, 2010; Lattimore and Hutter, 2012; Osband et al., 2013; Dann and Brunskill, 2015; Azar et al., 2017; Dann et al., 2017; Osband and Van Roy, 2017; Agrawal and Jia, 2017; Jin et al., 2018; Fruit et al., 2018; Talebi and Maillard, 2018; Dann et al., 2019; Dong et al., 2019; Simchowitz and Jamieson, 2019; Russo, 2019; Zhang and Ji, 2019; Cai et al., 2019; Zhang et al., 2020; Yang et al., 2020; Pacchiano et al., 2020; Neu and Pike-Burke, 2020; Zhang et al., 2021b; Li et al., 2021b; Ménard et al., 2021; Xiong et al., 2021; Li et al., 2021a; Pacchiano et al., 2020). We note that some previous works consider the time-inhomogeneous MDP where the transition and the reward can vary on different time steps (Jin et al., 2018; Zhang et al., 2020; Li et al., 2021a; Ménard et al., 2021). The regret for time-inhomogeneous setting will have an $\sqrt{H}$ factor in the regret, which is necessary because the degree of freedom increases by H compared with the time-homogeneous setting. Transforming a regret bound tightly from the time-homogenous setting to that for time-inhomogeneous setting is often straightforward (with an additional $\sqrt{H}$ factor), but not vice-versa, because one of the main difficulties to obtain sharp bounds in time-homogeneous MDP is how to exploit the property that the transition and reward do not vary on different time steps.

In this paper, we assume the total reward from all steps are upper bounded (cf. Assumption 1). Many prior work used the assumption that the reward from each step is upper bounded $1/H$, a.k.a., the uniformly bounded assumption. The bounded total reward is strictly more general than the uniformly bounded assumption. From a practical point of view, the bounded total reward assumption can model environments with spiky rewards, which are often considered to be a challenging problem (Jiang and Agarwal, 2018).

Dependence on Horizon. Then main focus of this work is the dependence on the planning horizon H. This problem was throughly discussed in a COLT 2018 Open Problem (Jiang and Agarwal, 2018) where it was conjectured that there would be a poly(H) regret lower bound. Zanette and Brunskill (2019) partially refuted this conjecture by giving an algorithm whose regret only scales logarithmically with H in the regime where $K = \text{poly}(S, A, H)$. This conjecture was refuted by Wang et al. (2020) who built an ϵ-net for the policy set and used it to develop a computationally inefficient algorithm which only requires poly $(S, A, \log H, 1/\epsilon)$ to learn an ϵ-optimal policy. This result was substantially improved by Zhang et al. (2021b) who gave a computationally efficient algorithm which enjoys an $O\left(\left(\sqrt{SAK} + S^2A\right)\text{polylog}\,(S, A, K, H)\right)$ regret. Their technique was later adopted in several other setting to tighten the dependency on the horizon (Zhang et al., 2021a,c; Ren et al., 2021; Chen et al., 2021a; Tarbouriech et al., 2021; Chen et al., 2021b). In Section 3, We will discuss why their work has polylog$H$ dependency and how we design new techniques to remove it.

Comparison with Li et al. (2021b). The recent breakthrough by Li et al. (2021b) gave an $\frac{(SA)^{O(S)}}{\epsilon^5}$ sample complexity bound. Notably, this is the first result showing the sample complexity *can be completely independent of* H. They have two key ideas: (1) a refined perturbation analysis in the generative model setting,[3] and (2) bounding the reaching probability based on the analysis on the entire trajectory instead of dynamic programming which is typically used in the literature. An implication of the second idea is an approximation bound to non-stationary policies using stationary policies of discounted MDPs. The approximation to non-stationary policies incurs the exponential

3. In the generative model setting, the agent can query any state-action pair. In this setting, they have can have polynomial sample complexity.

dependency on S. At a high level, they first uses all stationary policies (which is exponential in size) to collect enough samples to reduce the problem to the generative model setting, and uses the refined bound for to prove the final result.

We adopt their idea on the analyzing the entire trajectory. We give a refined analysis in bounding the reaching probability with an exponentially improved multiplicative constant (see discussions below Lemma 2). Our algorithm framework is different from theirs: our algorithm follows a more conventional approach based on upper confidence bound (UCB), and thus we do not use their perturbation analysis for the generative model setting. Nevertheless, the goal of the stage 1 of our algorithm, initial sample collection, is the same as their algorithm. The main differences are (1) we divide each episode in stage into two phases, each with a different policy, whereas they used a single stationary policy; and (2) our sample collection is adaptive in that we change the exploration policy based on the collected samples whereas theirs is oblivious in the sense that they simply enumerate all stationary policies.

Stationary policy is a central object in both works. They established the approximation power of stationary policies to non-stationary policies. We give an exponentially improved bound of approximation power dedicated to the visitation count (Lemma 3), and new results on the concentration (Lemma 4) and stability (Lemma 5) of stationary policies.

Lastly, besides using discounted MDPs to establish approximation bounds, we also use discounted MDP to compute stationary policies to make our algorithm run in polynomial time. All these differences are crucial in obtaining our polynomial-time horizon-free algorithm. See Section 3 for more details.

2. Preliminaries

Notations. Throughout this paper, we use $[N]$ to denote the set $\{1, 2, \ldots, N\}$ for $N \in \mathbb{Z}_+$. We use $\mathbf{1}_s$ to denote the one-hot vector whose only non-zero element is in the s-th coordinate. For an event $\mathcal{E}$, we use $\mathbb{I}[\mathcal{E}]$ to denote the indicator function, i.e., $\mathbb{I}[\mathcal{E}] = 1$ if $\mathcal{E}$ holds and $\mathbb{I}[\mathcal{E}] = 0$ otherwise. For notational convenience, we set $\iota = \ln(2/\delta)$ throughout the paper. For two n-dimensional vectors x and y, we use xy to denote $x^\top y$, use $\mathbb{V}(x, y) = \sum_i x_i y_i^2 - (\sum_i x_i y_i)^2$. In particular, when x is a probability vector, i.e., $x_i \geq 0$ and $\sum_i x_i = 1$, $\mathbb{V}(x, y) = \sum_i x_i (y_i - (\sum_i x_i y_i))^2 = \min_{\lambda \in \mathbb{R}} \sum_i x_i (y_i - \lambda)^2$. We also use x^2 to denote the vector $[x_1^2, x_2^2, ..., x_n^2]^\top$ for $x = [x_1, x_2, ..., x_n]^\top$. For two vectors x, y, $x \geq y$ denotes $x_i \geq y_i$ for all $i \in [n]$ and $x \leq y$ denotes $x_i \leq y_i$ for all $i \in [n]$.

Episodic Tabular MDP. We consider finite-horizon time-homogeneous Markov Decision Process (MDP) which can be described by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, H, \mu_1)$. $\mathcal{S}$ is the finite state space with cardinality S. $\mathcal{A}$ is the finite action space with cardinality A. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the unknown transition operator which takes a state-action pair and returns a distribution over the states. $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. For simplicity, we assume the reward function is known because the main difficulty is in estimating the transition function. Prior work, e.g., Jin et al. (2018), also made this assumption. $H \in \mathbb{Z}_+$ is the planning horizon. $\mu_1 \in \Delta(\mathcal{S})$ is the initial state distribution.

For notational convenience, we use $P_{s,a}$ and $P_{s,a,s'}$ to denote $P(\cdot|s, a)$ and $P(s'|s, a)$ respectively.

A policy π chooses an action a based on the current state $s \in \mathcal{S}$ and the time step $h \in [H]$. Note even though transition operator and the reward distribution do not depend on the level $h \in [H]$, the policy can choose different actions for the same state at different level h. Formally, we define $\pi = \{\pi_h\}_{h=1}^H$ where for each $h \in [H]$, $\pi_h : \mathcal{S} \to \mathcal{A}$ maps a given state to an action. The policy π

induces a trajectory $\{s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_H, a_H, r_H\}$, where $s_1 \sim \mu_1$, $a_1 = \pi_1(s_1)$, $r_1 \sim R(s_1, a_1)$, $s_2 \sim P(\cdot|s_1, a_1)$, $a_2 = \pi_2(s_2)$, etc. Our goal is to find a policy π that maximizes the expected total reward, i.e., $\max_\pi \mathbb{E}\left[\sum_{h=1}^H r_h \mid \pi\right]$, where the expectation is over μ_1, P and R. We make the following normalization assumption about the reward.

Assumption 1 (Bounded Total Reward) *The reward satisfies that $r_h \geq 0$ for all $h \in [H]$. Besides, for all policy π, $\sum_{h=1}^H r_h \leq 1$ almost surely.*

Q-function and V-function. Given a policy π and a level $h \in [H]$ the Q-function is defined as: $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to R$, $Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'} \mid s_h = s, a_h = a, \pi\right]$. Similarly, given a policy π, a level $h \in [H]$, the value function is defined as: $V_h^\pi : \mathcal{S} \to R$, $V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'} \mid s_h = s, \pi\right]$. Then Bellman equation states the following identities for policy π and $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$: $Q_h^\pi(s,a) = r(s,a) + P_{s,a}^\top V_{h+1}^\pi$ and $V_h^\pi(s) = Q_h^\pi(s, \pi_h(a))$. Throughout the paper, we let $V_{H+1}(s) = 0$ and $Q_{H+1}(s,a) = 0$ for simplicity. We use Q_h^* and V_h^* to denote the optimal Q-function and V-function, which satisfies for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, $Q_h^*(s,a) = \max_\pi Q_h^\pi(s,a)$ and $V_h^*(s) = \max_\pi V_h^\pi(s)$.

Regret and PAC Bound. The agent interacts with the environment for K episodes, and it chooses a policy π^k at the k-th episode. The total regret is defined as

$$\text{Regret}(K) = \sum_{k=1}^K V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k).$$

PAC-RL sample complexity is another measure which counts the total number of episodes to find an ϵ-optimal policy π, i.e., $\mathbb{E}_{s_1 \sim \mu_1}\left[V_1^*(s_1) - V^\pi(s_1)\right] \leq \epsilon$.

A regret bound can be transformed into a PAC bound (Jin et al., 2018). Specifically, if an algorithm achieves a $CK^{1-\alpha}$ regret for some $\alpha \in (0,1)$ and some C independent of K, by randomly selecting from policy π^k used in K episodes, π will satisfy $\mathbb{E}_{s_1 \sim \mu_1}\left[V_1^*(s_1) - V^\pi(s_1)\right] = O\left(CK^{-\alpha}\right)$. Setting $CK^{-\alpha} = \epsilon$, we can obtain a PAC-RL bound, which we also use.

Additional Notations. Let Π denote the set of all policies and Π_{sta} denote the set of all stationary policies (a policy π is stationary if $\pi_1 = \pi_2 \cdots = \pi_H$). We use $\mathbb{E}_{\pi,p}[\cdot]$ and $P_{\pi,p}[\cdot]$ to denote the expectation and probability following a policy π under a transition p. We let $W_d^\pi(r', p, \mu) := \mathbb{E}_{\pi,p}[\sum_{h=1}^d r'(s_h, a_h) | s_1 \sim \mu]$ be the value function for a reward function r' and a transition model p with horizon length d and initial distribution μ. With a slight abuse of notation, we also define $W_d^\pi(r', p, \mu) := \mathbb{E}_{\pi,p}[\sum_{h=1}^d r'(s_h, a_h) | (s_1, a_1) \sim \mu]$ for μ as a distribution over state-action space. We also use $\mathbf{1}_s$ and $\mathbf{1}_{s,a}$ to denote the reward function r' such that $r'(s', a') = \mathbb{I}[s' = s]$ and $r'(s', a') = \mathbb{I}[(s', a') = (s, a)]$, respectively. Sometimes we also abuse the notation to use $\mathbf{1}_s$ and $\mathbf{1}_{s,a}$ to denote a distribution with $\Pr(s) = 1$ and $\Pr(s, a) = 1$ respectively, With these notations, $W_d^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_s)$ denotes the expected number of visits to (s, a) under the policy π in a transition p with a planing horizon d and the agent starts from the fixed state s. This is a crucial function which we will use in our proof.

3. Technical Overview

Our algorithm follows the conventional UCB-based framework. Different from existing work, to avoid the dependency on H, we also design a new stage to *explicit explore* each state-action pair.

To illustrate why we introduce this new stage, along with our other technical ideas, we discuss each major source that incurs a $\log H$ dependency in Zhang et al. (2021b), and then describe our techniques to remove the $\log H$.

Source 1: Higher Order Expansion. One key idea in Zhang et al. (2021b) in bounding the regret is to use a recursive structure to relate the estimated variance to the higher moments. They expanded for $O(\log H)$ times, which incurred a $\log H$ in their regret bound.

This source is relatively simple to remove. We use an observation in Chen et al. (2021a) (cf. Lemma 9) that bounds the variance of the product two random variables, together with several probability bounds (cf. Lemma 10, 11, 12, 14, (40), and (53)), to avoid the use of recursion. We note that this analysis technique directly applies to the algorithm in Zhang et al. (2021b), and therefore can simplify their proof.

Source 2: Counting in Pigeonhole. A standard proof step in nearly all UCB-based algorithms is bounding $\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\max\{N^k(s_h^k,a_h^k),1\}}$ where $N^k(s,a)$ is the number of visits to state-action pair (s,a) before the k-th episode, and (s_h^k, a_h^k) is the state-action pair of the h-step in the k-th episode. By the pigeonhole principle, we can bound

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\max\{N^k(s_h^k,a_h^k),1\}} = O\left(\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{k=1}^{K}\min\left(\log\left(\frac{N^{k+1}(s,a)}{N^k(s,a)}\right),1\right)\right)$$

Since in total we have KH state-action pair visitations, we can have a straightforward bound by $\sum_{k=1}^{K}\min\left(\frac{\log N^{k+1}(s,a)}{\log N^k(s,a)},1\right) = O(\log(KH))$, which is used in all prior work. However, in the regime that H is large, e.g., $H = 2^K$, this naive bound gives linear regret. This source is much more difficult to remove. To remote it, we start with the following observation.

The Benefit of Initial Samples. Our first observation is that if we have enough initial samples, i.e., $N^1(s,a)$ is above a certain threshold, then we can avoid using the naive $\log(KH)$ bound. Formally, define $U(s,a) := \max_\pi W_H^\pi(\mathbf{1}_{s,a}, P, \mu_1)$ be the maximum expected visitation count of (s,a) in one episode. Then by Markov's inequality, with probability $1-\delta$, the total count of (s,a) in K episodes satisfies $N^K(s,a) \le KU(s,a)/\delta$. Conditioned on this event, if we make $N^1(s,a)$ comparable to $U(s,a)$, for example, $N^1(s,a) \ge U(s,a)/\exp(\mathrm{poly}(S))$, then we have that

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

which is independent of H. Now, the problem reduces to collect enough initial samples to make $N^1(s,a)$ comparable to $U(s,a)$. This problem of collecting initial samples is highly non-trivial and we devote the following subsection to describe our technical ideas.

3.1. Collecting Initial Samples

Now we focus on collecting the initial samples. For a fixed (s^*,a^*), our goal is to collect samples of (s^*,a^*). For the ease of discussion, here we assume that $N^1(s,a) \ge \frac{U(s,a)}{\exp(\mathrm{poly}(S))}$ for all (s,a) except for (s^*,a^*).

For this task, we divide one epoch into two phases. In the first phase, we aim to reach the target state s^*. In the second phase, we aim to collect as many samples of (s^*,a^*) as possible with the agent

starting from s^*. We note that this two-phase procedure uses *two* stationary policies, in contrast to Li et al. (2021b) who used a single stationary policy to collect samples. This difference is one of the key ingredients in obtaining the polynomial bound.

3.1.1. PHASE 1: REACHING THE TARGET STATE s^*

We first decide the length for each phase, which relies on the following lemma. The formal statement requires more notations and defer to appendix.

Lemma 2 *(Informal) Let $\mathcal{O} \subset \mathcal{S} \times \mathcal{A}$. Let $X^{\pi}_{d'}(\mathcal{O}, p, \mu_1)$ denote the probability of reaching $\mathcal{O}$ in d' steps following π under the transition p. We have the following bound: for any $\tilde{d} \in \mathbb{Z}_+$, $\max_\pi X^{\pi}_{(S+2)\tilde{d}}(\mathcal{O}, p, \mu_1) \le S^2 \max_\pi X^{\pi}_{(S+1)\tilde{d}}(\mathcal{O}, p, \mu_1)$.*

Lemma 2 establishes a bound of two reaching probabilities induced by the same policy, transition, initial distribution but slightly different planning horizons ($(S+1)\tilde{d}$ v.s. $S\tilde{d}$). We believe this lemma will have applications in other problems. To prove this lemma, we count the probability of all possible trajectories under two horizons and construct a mapping between the trajectories.

We note this lemma is similar in spirit to Lemma 4.2, 4.3 and 4.4 of Li et al. (2021a), which bound the reaching probability of a longer horizon Markov chain by that of a shorter horizon Markov chain and a multiplicative factor. The main difference is that we are using the reaching probability of a horizon-$S\tilde{d}$ Markov chain to approximate that of a horizon-$(S+1)\tilde{d}$ Markov chain whereas they used the reaching probability of a horizon-$S\tilde{d}$ Markov chain to approximate that of a horizon-$4S\tilde{d}$ Markov chain. This difference ($S\tilde{d}$ and $(S+1)\tilde{d}$ versus $S\tilde{d}$ and $4S\tilde{d}$) results in an exponential improvement in the multiplicative factor: from S^{4S} in Li et al. (2021b) to S^2 in Lemma 2. The proof for both results are based on counting arguments although the details are substantially different.

To use this lemma, we view $H = (S+1)\tilde{d}$, and from the bound, it is natural to use the first $\frac{HS}{S+1}$ steps in one episode to reach s^* and use the remaining steps to collect (s^*, a^*). [4]

To find a policy that reaches s^*, we can use $\mathbf{1}_{s^*}$ as the reward function, and perform a regret minimization algorithm. Since we have assumed $N^1(s', a') \ge U(s', a')/\exp(\mathrm{poly}(S))$ for any $(s, a) \ne (s^*, a^*)$, running the regret minimization problem for K_1 episodes gives a *first-order regret bound* of $O(\mathrm{poly}(SA)\mathrm{polylog}(K_1)\sqrt{K_1 v^* \iota})$, where v^* is the optimal value, i.e., the maximal probability of reaching s. Now we have two cases: (1) $v^* \ge \frac{f(SA)\mathrm{polylog}(K_1)}{K_1}$ for some polynomial f, then the cumulative reward, i.e., the number of times of reaching s^* is large enough; (2) $v^* < \frac{f(SA)\mathrm{polylog}(K_1)}{K_1}$, then (s^*, a^*) could be ignored with most $O\left(\frac{Kf(SA)\mathrm{polylog}(K_1)}{K_1}\right) = O(\mathrm{poly}(SA)\mathrm{polylog}(K)\sqrt{K\iota})$ regret by choosing $K_1 = O(\sqrt{K\iota})$. We note that the actual algorithm simultaneously explores all under-explored states by setting reward to be 1 for all under-explored states. See Algorithm 1 for details.

3.1.2. PHASE 2: COLLECTING SAMPLES OF (s^*, a^*) STARTING FROM s^*

In this phase, we start from the state s^*, and we would like to collect as many samples of (s^*, a^*) as possible. Inspired by recent work (Li et al., 2021b), we also consider using *stationary* policies to collect samples. Below we will give three key lemmas (Lemma 3, 4, 5) to characterize the approximation power, the concentration property, and the stability of stationary policies. We believe these lemmas will have applications in other problems.

4. With loss of generality, we assume $\frac{H}{S+1}$ is an integer and $H \gg S$ because we are interested in the regime H is large.

Recall that $W_d^\pi(\mathbf{1}_{s,a}, P, \mathbf{1}_s)$ denotes the expected number of visits to (s,a), starting from s, following π in a transition P with planning horizon d. The following lemma establishes that the power of stationary policies in collecting samples is not much worse than that of non-stationary policies. We prove this lemma using the discounted approximation by noting that there is an optimal stationary policy for the discounted planning.

This lemma can be compared to Corollary 4.7 of Li et al. (2021b). Their lemma is more general because it applies general reward and arbitrary initial distribution but ours only applies to reward of the form $\mathbf{1}_{s,a}$ with the starting distribution being $\mathbf{1}_s$. On the other hand, our multiplicative factor is exponentially smaller than theirs (roughly speaking, $O(S)$ vs. $S^{O(S)}$) and this improvement is crucial in obtaining our polynomial-time algorithm.

Lemma 3 *[Approximation Power of Stationary Policies] Let k and d be positive integers. We have that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\max_{\pi \in \Pi} W_{kd}^\pi(\boldsymbol{1}_{s,a}, P, \boldsymbol{1}_s) \le 6k \max_{\pi \in \Pi_{\text{sta}}} W_d^\pi(\boldsymbol{1}_{s,a}, P, \boldsymbol{1}_s).$$

The following lemma is a concentration bound for stationary policy, which shows the number of samples we collect empirically is close to the expectation. The proof is by regarding the recurrent time as i.i.d. random variables and constructing a stopping time.

Lemma 4 *[Concentration Property of Stationary Policies] For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\pi \in \Pi_{\text{sta}}$ such that $\pi(s) = a$, we have that $\Pr\left[N \ge \frac{1}{4} W_d^\pi(P, \boldsymbol{1}_{s,a}, \boldsymbol{1}_s)\right] \ge \frac{1}{2}$ for any horizon d, where N is the visit count of (s,a) following π under P in d steps with the initial distribution as $\boldsymbol{1}_s$.*

Therefore, if we successfully find a stationary policy that maximizes $W_{H/(S+2)}^\pi(\mathbf{1}_{s,a}, P, \mathbf{1}_s)$, then by Lemma 4, we can collect $\Omega(\max_{\pi \in \Pi_{\text{sta}}} W_{H/(S+2)}^\pi(\mathbf{1}_{s,a}, P, \mathbf{1}_s))$ samples, which, by Lemma 3, is larger than $\Omega\left(\frac{1}{S+1} \max_{\pi \in \Pi} W_H^\pi(\mathbf{1}_{s,a}, P, \mathbf{1}_s)\right) = \Omega(U(s,a)/S)$.

To learn a stationary policy with large enough visitation count to (s,a), we consider to learn a reference model P^{ref} close to P to help plan. The next lemma can be viewed as a *multiplicative performance difference lemma*. This lemma establishes the stability of stationary policies in the relative sense. Importantly, the multiplciative factor is completely independent of H. The proof is based on a local perturbation analysis. In each time, we perturb one (s,a) and aggregate the perturbation error in the end.

Lemma 5 (Multiplicative Performance Difference Lemma for Stationary Policies) *Let the initial distribution μ_1 be fixed. For two transition model P' and P'' such that $e^{-\epsilon} P''_{s,a,s'} \le P'_{s,a,s'} \le e^{\epsilon} P''_{s,a,s'}$, we have that*

$$e^{-4S\epsilon} W_d^\pi(r, P', \mu_1) \le W_d^\pi(r, P'', \mu_1) \le e^{4S\epsilon} W_d^\pi(r, P', \mu_1) \tag{2}$$

for any stationary policy π, horizon $d \ge 1$ and non-negative reward r.

Viewing $W_d^\pi(P', r, \mu_1)$ as a function of P', Lemma 5 shows that $\log(W(P', r, \mu_1))$ is $O(S)$-Lipschtiz continuous in $\log(P')$. It is crucial that the Lipschtiz constant is independent of H, which allows us to choose $\epsilon = O(1/S)$ in Lemma 5.

Now our goal is to find a transition model P^{ref} such that $e^{-\epsilon} P^{\text{ref}}_{s,a,s'} \le P_{s,a,s'} \le e^{\epsilon} P^{\text{ref}}_{s,a,s'}$ for any (s,a,s'). By concentration inequalities for the multinomial distribution, to learn such a transition model, we need to sample from (s,a) until (s,a,s') is visited more than $\frac{C_l}{\epsilon^2}$ times for each (s,a,s').

Clipped MDP and Explicit Exploration. The main difficulty is to deal with the case that $P_{s,a,s'}$ is small. For example if $P_{s,a,s'} \le \frac{1}{KH}$, we can hardly collect enough samples of (s,a,s'). To address this problem, we simply ignore such (s,a,s') tuples since the probability of visiting them is also very small. More precisely, we maintain a set $(\mathcal{K})^C$ for such tuples and construct a *clipped MDP*, where we redirect all $(s,a,s') \in (\mathcal{K})^C$ tuples to the virtual ending state, denoted as z. We note that we will update $(\mathcal{K})^C$ throughout the training process because after we collect new samples, we can assert that certain $P_{s,a,s'}$ is large and we can move (s,a,s') out of $(\mathcal{K})^C$.

In addition, we conduct *explicit exploration*. Roughly speaking, for each (s,a), if we have the chance to visit (s,a), then we design a policy to visit (s,a) as much as possible to judge whether z can be reached by (s,a). More precisely, in the beginning of the first sub-phase, we test that if there exists some (s,a) such that the maximal possible expected count of (s,a) under the clipped transition model exceeds the current visitation count of (s,a) by a $\frac{1}{\mathrm{poly}(SA)}$ ratio. Then we have two cases: (1) There exists such a (s,a). In this case we conduct exploration to collect samples of (s,a). Our target is the maximal possible expected count under the clipped transition model, which is smaller than that under the original transition model. Therefore, this task is easier and could be completed by naive planning. (2) Otherwise, we can ensure that the probability of visiting z is bounded by an universal constant. Then we can plan to visit the target state-action pair (s^*,a^*) by ignoring z.

Using Discounted MDP for Efficient Planning with Stationary Policies. Our final major technical idea is for the computational purpose. Given a finite-horizon MDP, finding the best stationary policy that maximizes the reward may not be computationally efficient. Recall that all we need is a multiplicative approximation. Therefore, we use discounted MDP to approximate the finite-horizon MDP. See Lemma 18 for the guarantees. We note that the idea of using discounted MDP was also used in Li et al. (2021b), although they did not use it for computational reasons.

4. Main Algorithm

Now we present our main algorithm. There are two stages in Algorithm 1. In the first stage, for each episode, we let the agent explore in its first $\frac{HS}{S+1}$ steps to reach new state-action pairs, and collect the initial samples using the remaining $\frac{H}{S+1}$ steps. The number of this stage is bounded by $O(\mathrm{poly}(S,A,\log(K))\sqrt{K})$, and incurring at most $O(\mathrm{poly}(S,A,\log K)\sqrt{K})$ regret. In the second stage, we play optimistic value iteration to learning the MDP with initial samples. Below we give two important notions used in Algorithm 1.

In stage 1, the algorithm maintains an omitted set denoted as $\mathcal{O}^k \subset \mathcal{S}\times\mathcal{A}$ for the k-th episode. If a state-action pair (s,a) is *not* in $\mathcal{O}^k$, we know we have have collected enough samples for (s,a). We note that in this end we may not have $\mathcal{O}^k = \emptyset$ because there can be states that are hard to reach using any policy and we cam simply ignore them. To explore, we plan *optimistically* according to a confidence set of the transition matrix, constructed by the collected samples.

Confidence set. Given $\{N(s,a,s')\}_{s,a,s'}$, we define $N(s,a) = \max\{\sum_{s'} N(s,a,s'),1\}$, and $\mathcal{P} = \texttt{ConfidenceSet}(\{N(s,a,s')\}_{s,a,s'})$ by setting $\mathcal{P} = \otimes_{h,s,a}\mathcal{P}_{h,s,a}$ where

$$\mathcal{P}_{h,s,a} = \left\{ p \in \Delta^S : |p_{s'} - \frac{N(s,a,s')}{N(s,a)}| \le \sqrt{4\frac{N(s,a,s')\iota}{N^2(s,a)}} + \frac{5\iota}{N(s,a)} \right\}.$$

We note that $\mathcal{P}_{h,s,a}$ does not depend on h. We add h in the subscript only for the writing purpose when we use $\mathcal{P}_{h,s,a}$.

Algorithm 1 Main Algorithm

Input: state space S, action space A, reward r, horizon H, confidence parameter δ;
Initialization: $N(s,a,s') \leftarrow 0, \forall s,a,s'$, $\bar{N}(s,a) \leftarrow 0, \forall (s,a)$, $d \leftarrow \frac{(S+1)H}{S+2}$; $O^1 \leftarrow \mathcal{S} \times \mathcal{A}$; $\mathcal{P}^1 \leftarrow (\Delta^S)^{SA}$; $K_1 \leftarrow C_1\sqrt{S^9A^3K\iota}$, $n_1 \leftarrow C_2S^7A^3\iota$; $d' = H - d$; $m(s,a) \leftarrow 0$; $N_0 \leftarrow 256S^2\log(1/\delta)$;
// Stage 1: Collecting initial samples
for $k = 1, 2, \ldots, K_1$ **do**
  $\mathcal{P}^k \leftarrow$ `ConfidenceSet`$(\{N(s,a,s')\}_{s,a,s'})$;
  $(\pi^k, \tilde{P}^k) \leftarrow \max_{\pi, p \in \mathcal{P}^k} X_d^{\pi}(O^k, p, \mu_1)$
  for $h = 1, 2, \ldots, d$ **do**
    Observes s_h^k, takes action $\pi_h^k(s_h^k)$, receives r_h^k and transits to s_{h+1}^k;
    $N(s_h^k, a_h^k, s_{h+1}^k) \leftarrow N(s_h^k, a_h^k, s_{h+1}^k) + 1$;
    if $\exists a, (s_{h+1}^k, a) \in O^k$ **then**
      $(s_1^*, a_1^*) \leftarrow (s_{h+1}^k, a)$;
      $\{N(s,a,s')\}_{s,a,s'} \leftarrow \{n(s,a,s')\}_{s,a,s'}$;
      $\mathcal{K}^k \leftarrow \{(s,a,s') : n(s,a,s') \geq N_0\}$, $\mathcal{K}^k(s,a) \leftarrow \{s' : (s,a,s') \in \mathcal{K}^k\}$;
      $n(s,a) \leftarrow \max\{\sum_{s':(s,a,s')\in\mathcal{K}^k} n(s,a,s'), 1\}\ \forall (s,a)$;
      $P^{\text{ref}}_{s,a,s'} \leftarrow \frac{n(s,a,s')}{n(s,a)}$, $P^{\text{ref}}_{s,a,z} \leftarrow 0, \forall (s,a,s') \in \mathcal{K}^k$;
      $P^{\text{ref}}_{s,a,s'} \leftarrow 0$, $P^{\text{ref}}_{s,a,z} = 1, \forall (s,a,s')$ such that $\mathcal{K}^k(s,a) = \emptyset$;
      (Trigger, $\{n(s,a,s')\}_{s,a,s'}$) $\leftarrow$ *Algorithm 2* with inputs $((s_1^*, a_1^*), P^{\text{ref}}, \{n(s,a,s')\}_{(s,a,s')}, \mathcal{K}^k, d')$;
      if Trigger = FALSE **then**
        $\{n(s,a,s')\}_{(s,a,s')} \leftarrow$ *Algorithm 3* with inputs $((s_1^*, a_1^*), P^{\text{ref}}, \{n(s,a,s')\}_{s,a,s'}, d')$
        $m(s_1^*, a_1^*) \leftarrow m(s_1^*, a_1^*) + 1$;
        if $m(s_1^*, a_1^*) \geq 400\log(1/\delta)$ **then**
          $O^{k+1} \leftarrow O^k/(s_1^*, a_1^*)$;
        end if
      end if
      $\{n(s,a,s')\}_{s,a,s'} \leftarrow \{N(s,a,s')\}_{s,a,s'}$;
      break;
    end if
  end for
  If there are remaining steps, run a random policy and update $\{N(s,a,s')\}_{s,a,s'}$;
end for
// Stage 2: Regret Minimization with Initial Samples
Run Algorithm 4 with inputs $\{N_{s,a,s'}\}_{s,a,s'}$.

For each $k \in [K]$, we use $N^k(s,a,s')$ to denote the value of $N(s,a,s')$ before the k-th episode. Define $N^k(s,a) = \max\{\sum_{s'} N^k(s,a,s'), 1\}$ and $\hat{P}^k_{s,a,s'} = \frac{N^k(s,a,s')}{N^k(s,a)}$. Define $\mathcal{G}$ be the event where

$$|P_{s,a,s'} - \hat{P}^k_{s,a,s'}| \leq \min\left\{\sqrt{2\frac{P_{s,a,s'}\iota}{N^k(s,a)}} + \frac{\iota}{3N(s,a)}, \sqrt{4\frac{\hat{P}^k_{s,a,s'}\iota}{N^k(s,a)}} + \frac{5\iota}{N(s,a)}\right\} \tag{3}$$

Algorithm 2 `Explicit Exploration`

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

holds for any k, s, a, s'. By Bennets's inequality and Bernstein's inequality, we have that $\mathbb{P}[\mathcal{G}] \geq 1 - 2S^2AK\delta$. In the analysis below, we assume $\mathcal{G}$ holds.

Now we describe Stage 1. We divide each episode into two phases. The first phase has length $d = \frac{SH}{H+1}$ and the second phase $d' = H - d$. In Line 5, we plan and try to arrive at a state-action pairs that we have not collected enough samples, a.k.a., maximize the reaching probability of $\mathcal{O}^k$. In the episode k and during phase 1, $h = 1, \ldots, d$, whenever we meet a state s_{h+1}^k such that there exists a that $(s_{h+1}^k, a) \in \mathcal{O}^k$, we stop phase 1 because we have reached one state-action pair that we have not collected enough samples of.

In phase 2, we denote $(s_1^*, a_1^*) = (s_{h+1}^k, a)$ and try to collect as many (s_1^*, a_1^*) as possible. Instead of using the confidence set of the transition matrix to do planning optimistically, we split state-action-state triples as known set ($\mathcal{K}^k$) and unknown set $\left(\mathcal{K}^k\right)^c$ (cf. Line 13), and then we compute a clipped reference transition model to plan defined below (also see Line 14 - Line 16 in Algorithm 1).

Algorithm 3 `Sample Collection with a Reference Model`

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

Clipped Reference Transition Model. Given $\mathcal{K}^C \subset \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and a transition model p, we define $p' := \text{clip}(p, \mathcal{K}^C)$ be the transition model such that $p'_{s,a,s'} = p_{s,a,s'}, \forall (s, a, s') \notin \mathcal{K}^C$, $p'_{s,a,s'} = 0, \forall (s, a, s') \in \mathcal{K}^C$, $p'_{s,a,z} = \sum_{s':(s,a,s')\in\mathcal{K}^C} p_{s,a,s'}$, $p'_{z,a} = \mathbf{1}_{z'}, \forall a$ and $p'_{z',a} = \mathbf{1}_{z'}, \forall a$. In words, we redirect the (s, a, s') triples in $\mathcal{K}^C$ to a virtual state z, which transits to a virtual absorbed state z' with probability 1. The reason why we need an additional z' instead of just z is make the total reward bounded by 1. As a result, we have the following identity by definition:

$$X_{\tilde{d}}^\pi(\mathcal{K}^C, p, \mu_1) = W_{\tilde{d}}^\pi(\mathbf{1}_z, \text{Clip}(p, \mathcal{K}^C), \mu_1), \quad \forall \tilde{d} \in \mathbb{Z}_+. \tag{4}$$

In a similar way, we define $\text{clip}(p, \mathcal{K}^C)$ for $\mathcal{K}^C \subset \mathcal{S}\times\mathcal{A}$ and $\mathcal{K}^C \subset \mathcal{S}$.

In our context, $P^{\text{ref}} = \text{Clip}(\hat{P}, \left(\mathcal{K}^k\right)^C)$ where $\hat{P}$ is the empirical model. This clipping operation is crucial to enable us to use Lemma 27.

Explicit Exploration. Given the a starting state-action pair (s_1^*, a_1^*), a reference model and the known set, we apply Algorithm 2. In Algorithm 2, we first try to explicit explore the unknown set in order to make our reference model estimation more accurate. To do so, for every state-action-state triple (s, a, s') not in the known set, we compute two *stationary policies*, π_1 and π_2 where π_1 tries to reach s from (s_1^*, a_1^*) and π_2 tries to collect as many (s, a) as possible *starting from* s. The stationary policies are computed by using a discounted MDP to approximate a finite-horizon MDP. The purpose is that we can compute the stationary policies in polynomial time.

Besides the policies, we also obtain estimates $u^k(s)$ and $v^k(s, a)$ on how many samples we can expect to collect. In Line 10 of Algorithm 2, we check whether our estimation is large, and we have not collect enough samples. If this is the case, we execute π_1 and π_2. Otherwise, we either have collected enough samples or s is hard to reach.

We iterate all state-action pairs, and if for all pairs we have either collected enough samples or identified that this triple is hard to reach (which we an ignore), we are confident the reference model is good enough for our purpose (Trigger = FALSE in this case). In this case, we use the reference model to collect as many (s_1^*, a_1^*) as possible (cf. Algorithm 3). Again, for computational efficiency purpose, we use a stationary policy computed from a discounted MDP that approximates the finite-horizon MDP.

Stage 2: Regret Minimization with Initial Samples After collecting initial samples, in Stage 2, we perform standard optimistic model-based planning using dynamic programming. See Algorithm 4 for details.

Algorithm 4 Regret Minimization with Initial Samples (`RMIS`)

Input: $\{N(s,a,s')\}_{(s,a,s')}$;
$N(s,a) \leftarrow \max\{\sum_{s'} N(s,a,s'), 1\}$;
for $k = 1, 2, \ldots, K - K_1$ **do**
  $\mathcal{P}^k \leftarrow$ `ConfidenceSet`$(\{N(s,a,s')\}_{s,a,s'})$
  $V^k_{H+1}(s) \leftarrow 0, \forall s$;
  $V^k_h(s) \leftarrow \min\{\max_{a,p\in\mathcal{P}^k_{s,a}}(r(s,a) + pV^k_{h+1}), 1\}, \forall (h,s) \in [H] \times \mathcal{S}$;
  $\pi^k_h(s) \leftarrow \arg\max_a \max_{p\in\mathcal{P}^k_{s,a}}(r(s,a) + pV^k_{h+1}), \forall (h,s) \in [H] \times \mathcal{S}$;
  for $h = 1, 2, \ldots, H$ **do**
    Observes s^k_h, takes action $\pi^k_h(s^k_h)$, receives reward r^k_h and transits to s^k_{h+1};
    $N(s^k_h, a^k_h, s^k_{h+1}) \leftarrow N(s^k_h, a^k_h, s^k_{h+1}) + 1$;
  end for
end for

5. Regret Analysis

Setting $K_1 = C_1\sqrt{S^9A^3K\iota}$ with some constant C_1, we have two key lemmas below.

Lemma 6 *Let $\mathcal{O}^{K_1+1}$ be defined in Algorithm 1. With probability $1 - 10SAK\delta$, we have that*

$$\max_\pi \mathbb{P}_\pi[\exists h \in [H], (s_h, a_h) \in \mathcal{O}^{K_1}] \le O\left(\frac{S^9A^3\iota + S^3A\iota^2}{K_1}\right). \tag{5}$$

Lemma 6 states that, we can collect enough initial samples for most state-action pairs. And the probability of visiting the remaining state-action pairs (those in the omitted set) is comparably small. See Appendix D for details.

Lemma 7 *With probability $1 - 10S^3A^2K\delta$, it holds that*

$$N^{\tilde{k}+1}(\tilde{s},\tilde{a}) \ge 2\max_{\pi\in\Pi_{\mathrm{sta}}} W^\pi_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}}, P, \mathbf{1}_{\tilde{s}})\log(1/\delta)$$

for any $(\tilde{s},\tilde{a}) \in \mathcal{O}^{\tilde{k}+1}/\mathcal{O}^{\tilde{k}}$ and any $1 \le \tilde{k} \le K_1$.

Lemma 7 states that the initial number of state-action pairs not in $\mathcal{O}^{K_1+1}$ is large enough. The proof of Lemma 7 is given in Appendix F.1.1

Lemma 8 *With probability $1 - 10SAK\delta$, the regret in the second stage is bounded by*

$$O\left(\mathrm{polylog}(SAK)\left(\sqrt{S^2AK\iota^2} + \frac{S^9A^3K\iota + S^3A\iota^2}{K_1}\right)\right).$$

Lemma 8 is based on classical regret analysis for finite horizon-MDP. The second term comes from the error from stage 1. In the proof we also need refined analysis to remove the extra $\log(H)$ factors. See Appendix E for details.

By Lemma 8, and noting that the regret in the first stage is bounded by K_1, we have that the total regret is upper bounded by $O(\mathrm{polylog}(SAK)\sqrt{(S^9A^3 + S^3A\iota)K\iota})$, and we finish the proof.

6. Conclusion

In this paper, we presented the first polynomial-time algorithm for tabular MDP whose regret is completely independent of the horizon. Our result crucially relies a series of structural lemmas of stationary policies, which we believe will be useful in other setting.

A fundamental open problem is whether we can design an algorithm with $O\left(\sqrt{SAK}\right)$ regret. A positive answer would have a surprising implication that tabular MDP is as easy as contextual bandits in the minimax sense.

Acknowledgements

The authors thank Ruosong Wang for insightful discussions. Zihan Zhang and Xiangyang Ji are supported by Beijing Municipal Science and Technology Commission grant Z201100005820005. Simon S. Du acknowledges funding from NSF Award's IIS-2110170 and DMS- 2134106.

References

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

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272, 2017.

Peter L Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, 2009.

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

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021a.

Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear mdp. *arXiv preprint arXiv:2112.09859*, 2021b.

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

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference*

on Neural Information Processing Systems, NIPS'17, page 5717–5727, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1507–1516, 2019.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, 3(1):100–118, 1975.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2994–3004, 2018.

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

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

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

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

Michael J Kearns and Satinder P Singh. Near-optimal reinforcement learning in polynominal time. In *Proceedings of the Fifteenth International Conference on Machine Learning*, page 260–268, 1998.

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

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

Yuanzhi Li, Ruosong Wang, and Lin F Yang. Settling the horizon-dependence of sample complexity in reinforcement learning. In *IEEE Symposium on Foundations of Computer Science*, 2021b.

Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618. PMLR, 2021.

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

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

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

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14433–14443, 2019.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, pages 1153–1162, 2019.

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

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888. ACM, 2006.

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

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. *arXiv preprint arXiv:1803.01626*, 2018.

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in Neural Information Processing Systems*, 34, 2021.

Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? In *Advances in Neural Information Processing Systems*, 2020.

Zhihan Xiong, Ruoqi Shen, and Simon S Du. Randomized exploration is near-optimal for tabular mdp. *arXiv preprint arXiv:2102.09703*, 2021.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pages 4438–4472. PMLR, 2021.

Kunhe Yang, Lin F Yang, and Simon S Du. Q-learning with logarithmic regret. *arXiv preprint arXiv:2006.09118*, 2020.

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

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

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, 2020.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *International Conference on Machine Learning*, 2021a.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021b.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. In *Advances in Neural Information Processing Systems*, 2021c.

Appendix A. Technical Lemmas

Lemma 9 (Lemma 30 in Chen et al. (2021a)) $\mathrm{Var}(XY) \le 2\mathbb{E}^2[Y]\mathrm{Var}(X) + 2\sup X^2\mathrm{Var}(Y)$.

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

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

Proof Let $t \in [0, 1/l]$ be fixed. Consider to bound $Z_k := \mathbb{E}[\exp(t\sum_{k'=1}^{k}(X_{k'} - 3Y_{k'}))]$. By definition, we have that

$$\begin{aligned}
\mathbb{E}[Z_k|\mathcal{F}_k] &= \exp(t\sum_{k'=1}^{k-1}(X_{k'} - 3Y_{k'}))\mathbb{E}\left[\exp(X_k - 3Y_k) \mid \mathcal{F}_k\right] \\
&\le \exp(t\sum_{k'=1}^{k-1}(X_{k'} - 3Y_{k'}))\exp(-3Y_k)\cdot\mathbb{E}[1 + tX_k + 2t^2X_k^2 \mid \mathcal{F}_k] \\
&\le \exp(t\sum_{k'=1}^{k-1}(X_{k'} - 3Y_{k'}))\exp(-3Y_k)\cdot\mathbb{E}[1 + 3tX_k \mid \mathcal{F}_k] \\
&= \exp(t\sum_{k'=1}^{k-1}(X_{k'} - 3Y_{k'}))\exp(-3Y_k)\cdot(1 + 3tY_k) \\
&\le \exp(t\sum_{k'=1}^{k-1}(X_{k'} - 3Y_{k'})) \\
&= Z_{k-1},
\end{aligned}$$

where the second line is by the fact that $e^x \le 1 + x + 2x^2$ for $x \in [0, 1]$. Define $Z_0 = 1$ Then $\{Z_k\}_{k\ge0}$ is a super-martingale with respect to $\{\mathcal{F}_k\}_{k\ge1}$. Let τ be the smallest n such that $\sum_{k=1}^{n} X_k - 3\sum_{k=1}^{n} Y_k > l\log(1/\delta)$. It is easy to verify that $Z_{\min\{\tau,n\}} \le \exp(tl\log(1/\delta) + tl) < \infty$. Choose $t = 1/l$. By the optimal stopping time theorem, we have that for any N:

$$\begin{aligned}
&\mathbb{P}\left[\exists n \le N, \sum_{k=1}^{n} X_k \ge 3\sum_{k=1}^{n} Y_k + l\log(1/\delta)\right] \\
&= \mathbb{P}\left[\tau \le N\right] \\
&\le \mathbb{P}\left[Z_{\min\{\tau,N\}} \ge \exp(tl\log(1/\delta))\right] \\
&\le \frac{\mathbb{E}[Z_{\min\{\tau,N\}}]}{\exp(tl\log(1/\delta))} \\
&\le \frac{Z_0}{\exp(tl\log(1/\delta))} \\
&\le \delta.
\end{aligned}$$

Letting $N \to \infty$, we have that

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

Considering $W_k = \mathbb{E}[\exp(t \sum_{k'=1}^{k}(Y_k/3 - X_k))]$, using similar arguments and choosing $t = 1/(3l)$, we have that

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

The proof is completed. ■

Lemma 11 *Let $v \in [0,1]^S$ be fixed vectir. Let $X_1, X_2, \ldots, X_t$ be i.i.d. multinomial distribution with parameter $p \in \Delta^S$. Define $\widehat{\mathrm{Var}} = \frac{1}{t}\left(\sum_{i=1}^{t} v_{X_i}^2 - \frac{1}{t}\left(\sum_{i=1}^{t} v_{X_i}\right)^2\right)$ be the empirical variance of $\{v_{X_i}\}_{i=1}^{t}$ and $\mathrm{Var} = \mathbb{V}(p, v)$ be the true variance of v_{X_1}. With probability $1 - 3\delta$, it holds that*

$$\frac{1}{3}\mathrm{Var} - \frac{7\log(1/\delta)}{3t} \leq \widehat{\mathrm{Var}} \leq 3\mathrm{Var} + \frac{\log(1/\delta)}{t}. \tag{6}$$

Proof Without loss of generality, we assume $\mathbb{E}[v_{X_1}] = 0$. By Lemma 10, with probability $1 - \delta$, it holds that

$$t\widehat{\mathrm{Var}} \leq \sum_{i=1}^{t} v_{X_i}^2 \leq 3t\mathbb{E}[v_{X_i}^2] + \log(1/\delta). \tag{7}$$

Dividing both side with t, we prove the right hand side of (6). For the other side, by Hoeffding's inequality, we have that $|\sum_{i=1}^{t} v_{X_i}| \leq \sqrt{2t\log(1/\delta)}$ holds with probability $1 - \delta$. Using Lemma 10 again, with probability $1 - \delta$ it holds that

$$\sum_{i=1}^{t} v_{X_i}^2 \geq \frac{t}{3}\mathbb{E}[v_{X_i}^2] - \frac{1}{3}\log(1/\delta).$$

As a result, with probability $1 - 2\delta$ it holds that (by the definition of $\widehat{Var}$)

$$\begin{aligned} t\widehat{Var} &\geq \frac{t}{3}\mathbb{E}[v_{X_i}^2] - \frac{1}{3}\log(1/\delta) - \frac{1}{t}\cdot 2t\log(1/\delta) \\ &= \frac{t}{3}\mathbb{E}[v_{X_i}^2] - \frac{7}{3}\log(1/\delta). \end{aligned}$$

The proof is completed by dividing both side by t.

■

Lemma 12 (Freedman's Inequality, Theorem 1.6 of Freedman (1975)) *Let $(M_n)_{n\geq 0}$ be a martingale such that $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$. Let $\mathrm{Var}_n = \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}]$ for $n \geq 0$, where $\mathcal{F}_k = \sigma(M_0, M_1, M_2, \ldots, M_k)$. Then, for any positive x and for any positive y,*

$$\mathbb{P}\left[\exists n : M_n \geq x \text{ and } \mathrm{Var}_n \leq y\right] \leq \exp\left(-\frac{x^2}{2(y + cx)}\right). \tag{8}$$

Lemma 13 (Bennet's Inequality) *Let $Z, Z_1, ..., Z_n$ be i.i.d. random variables with values in $[0,1]$ and let $\delta > 0$. Define $\mathbb{V}Z = \mathbb{E}\left[(Z - \mathbb{E}Z)^2\right]$. Then we have*

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

Lemma 14 *Let $(M_n)_{n\geq 0}$ be a martingale such that $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$ for some $c > 0$ and any $n \geq 1$. Let $\mathrm{Var}_n = \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2 | \mathcal{F}_{k-1}]$ for $n \geq 0$, where $\mathcal{F}_k = \sigma(M_1, M_2, ..., M_k)$. Let $C > 0$ be a constant. For any $p > 0$, we have that*

$$\mathbb{P}\left[\exists n, i, |M_n| \geq 10 \cdot 2^i ci\log(1/p), |\mathrm{Var}_n| \leq 2^{2i} ic^2 \log(1/p)\right] \leq p. \tag{9}$$

Proof By using Lemma 12 with $x = 10 \cdot 2^i ci\log(1/p)$, $y = 2^{2i} ic^2 \log(1/p)$, we have that

$$\mathbb{P}\left[\exists n, |M_n| \geq 10 \cdot 2^i ci\log(1/p), |\mathrm{Var}_n| \leq 2^{2i} ic^2 \log(1/p)\right] \leq \frac{p}{e^i} \leq \frac{p}{2^i}. \tag{10}$$

Taking sum over i we finish the proof. ■

Lemma 15 *Let $X_1, X_2, \ldots$ be i.i.d. random variables with multinomial distribution. For each $y \in \mathrm{supp}(X_1)$, we define $N_t(y) = \sum_{i=1}^{t} \mathbb{I}[X_i = y]$. Let $t(y)$ be the first time $N_t(y) = N := \frac{\log(2/\delta)}{\epsilon^2}$ (assume that $\frac{\log(2/\delta)}{\epsilon^2}$ is an integer) . It then holds that*

$$\mathbb{P}\left[e^{-8\epsilon}\frac{N}{t(y)} \leq \mathbb{P}[X_1 = y] \leq e^{8\epsilon}\frac{N}{t(y)}\right] \geq 1 - \delta.$$

Proof Let $p = \mathbb{P}[X_1 = y]$. Note that $N_{t(y)}(y) = N$, it suffices to prove that

$$\mathbb{P}[e^{-8\epsilon}N \leq t(y)p \leq e^{8\epsilon}N] \geq 1 - \delta.$$

Let $X_i' = \mathbb{I}[X_i = y]$. Note that

$$\Pr\left[\exists n \geq 1, \sum_{i=1}^{n} X_i' \geq e^{\epsilon} n\mathbb{E}[X_i'] + \frac{4}{\epsilon^2}\log(2/\delta)\right] \leq \delta/2;$$
$$\Pr\left[\exists n \geq 1, \sum_{i=1}^{n} X_i' \leq e^{-\epsilon} n\mathbb{E}[X_i'] - \frac{4}{\epsilon^2}\log(2/\delta)\right] \leq \delta/2,$$

we have that

$$\Pr\left[\forall n \geq 1, e^{-\epsilon} n\mathbb{E}[X_i'] - \frac{4}{\epsilon}\log(2/\delta) \leq \sum_{i=1}^{n} X_i' \leq e^{\epsilon} n\mathbb{E}[X_i'] + \frac{4}{\epsilon}\log(2/\delta)\right] \geq 1 - \delta. \tag{11}$$

Then with probability $1 - \delta$, it holds that

$$e^{-\epsilon}t(y)p - \frac{4}{\epsilon}\log(2/\delta) \leq N \leq e^{\epsilon}t_y p + \frac{4}{\epsilon}\log(2/\delta). \tag{12}$$

The proof is finished by noting that $N + \frac{4}{\epsilon}\log(2/\delta) \leq N(1 + 4\epsilon) \leq e^{4\epsilon}N$ and $N - \frac{4}{\epsilon}\log(2/\delta) = N(1 - 4\epsilon) \geq e^{-5\epsilon}$. ■

Appendix B. Collection of Notations

- $N^k(s,a,s')$: the value of $N(s,a,s')$ before in the k-th episode;
- $N^k(s,a) = \max\{\sum_{s'} N^k(s,a,s'), 1\}$;
- $\mathcal{K}^k := \{(s,a,s') : N^k(s,a,s') \geq N_0 := 256S^2\log(1/\delta)\}$: *known* state-action-state triples at the beginning of the k-th pair;
- $\mathcal{K}^k(s,a) := \{s' : (s,a,s') \in \mathcal{K}^k\}$;
- $\mathcal{U}^k := \{(s,a) : \mathcal{K}^k(s,a) = \emptyset\}$: the *unknown* state-action pairs;
- z: an additional state, which transits to z' with probability 1 for any action;
- z': an absorbed state, i.e., $P_{z,a} = \mathbf{1}_z, \forall a$;
- P: the true transition model;
- $\bar{P}^k$: the clipped transition model with respect to $(\mathcal{K}^k)^C$, i.e., the set of *unknown* state-action-state triples

$$\begin{aligned}
\bar{P}^k_{s,a,s'} &= P_{s,a,s'}, \forall (s,a,s') \in \mathcal{K}^k;\\
\bar{P}^k_{s,a,s'} &= 0, \forall (s,a,s') \notin \mathcal{K}^k;\\
\bar{P}^k_{s,a,z} &= \sum_{s' \notin \mathcal{K}^k(s,a)} P_{s,a,s'};
\end{aligned}$$

- $\bar{P}^{\text{cut},k}$: the transition model which ignore the probabilities transiting to z;

$$\begin{aligned}
\bar{P}^{\text{cut},k}_{s,a,s'} &= \frac{P_{s,a,s'}}{\sum_{s'' \in \mathcal{K}^k(s,a)} P_{s,a,s''}}, \forall (s,a,s') \in \mathcal{K}^k;\\
\bar{P}^{\text{cut},k}_{s,a,s'} &= 0, \forall (s,a,s') \notin \mathcal{K}^k;\\
\bar{P}^{\text{cut},k}_{s,a,z} &= 1, \forall (s,a) \in \mathcal{U}^k;
\end{aligned}$$

- $P^{\text{ref},k}$: the cut-off reference model

$$\begin{aligned}
P^{\text{ref},k}_{s,a,s'} &= \frac{N^k(s,a,s')}{\sum_{s'' \in \mathcal{K}^k(s,a)} N^k_{s,a,s''}}, \forall (s,a,s') \in \mathcal{K}^k;\\
P^{\text{ref},k}_{s,a,s'} &= 0, \forall (s,a,s') \notin \mathcal{K}^k;\\
P^{\text{ref},k}_{s,a,z} &= 1, \forall (s,a) \in \mathcal{U}^k;
\end{aligned}$$

- $\mathbb{E}_{p,\pi}[\cdot]$: the expectation(probability) following π under transition p;
- $\mathbb{P}_{p,\pi}[\cdot]$: the probability following π under transition p;
- $W^\pi_d(r,p,\mu_1) := \mathbb{E}_{p,\pi}\left[\sum_{h=1}^H r_h | s_1 \sim \mu_1\right]$: the general value function;
- $W^\pi_\gamma(r,p,\mu_1) := \mathbb{E}_{p,\pi}[\sum_{i\geq 1} \gamma^{i-1} r_i | s_1 \sim \mu_1]$;

- $X_d^\pi(O, p, \mu_1)$: the probability of reaching O in d steps with (p, π) as transition-policy pair and μ_1 as initial distribution;
- $X_\gamma^\pi(O, p, \mu_1) := \sum_{i\geq 1} \gamma^{i-1}\mathbb{P}_{p,\pi}[(s_i, a_i, s_{i+1}) \in O, (s_{i'}, a_{i'}, s_{i'+1}) \notin O, \forall 1 \leq i' \leq i-1 | s_1 \sim \mu_1]$,
- $\text{cut}(p)$: the cutting function for a transition model p. $p' = \text{Cut}(p)$ is defined by

$$p'_{s,a,s'} = \frac{p_{s,a,s'}}{\sum_{s''\neq z} p_{s,a,s''}} \forall (s,a) \quad \text{s.t.} p_{s,a,z} < 1;$$
$$p'_{s,a,z} = 0, \forall (s,a) \quad \text{s.t.} p_{s,a,z} < 1;$$
$$p'_{s,a,z} = 1, \forall (s,a) \quad \text{s.t.} p_{s,a,z} = 1.$$

  Note that $\bar{P}^{\text{cut},k} = \text{cut}(\bar{P}^k)$.
- $\mathbf{1}_s$: the vector which is 1 at s and 0 otherwise.
- $\mathbf{1}_{s,a}$: the vector which is 1 at (s, a) and 0 otherwise.
- $\mathbf{1}_{s,a,s'}$: the vector which is 1 at (s, a, s') and 0 otherwise.
- $\mathcal{J} := \{k \in [K_1] | \exists h \in [d], a, (s_{h+1}^k, a) \in O^k\}$;
- $(s_1^{*,k}, a_1^{*,k})$: the value of (s_1^*, a_1^*) in the k-th episode for $k \in \mathcal{J}$;
- Π_{sta}: the set of stationary policies;
- Π : the set of all possible policies.

Appendix C. Structural Lemmas for Stationary Policies

Lemma 16 (Restatement of Lemma 3) *Let k and d be positive integers. We have that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$\max_{\pi\in\Pi} W_{kd}^\pi(\boldsymbol{1}_{s,a}, P, \boldsymbol{1}_s) \leq 6k \max_{\pi\in\Pi_{\text{sta}}} W_d^\pi(\boldsymbol{1}_{s,a}, P, \boldsymbol{1}_s).$$

Proof [Proof of Lemma 3] Let π be fixed. Let $\tilde{\mu}_i$ be the distribution of s_{di+1} following π. We have that

$$W_{kd}^\pi(\mathbf{1}_{s,a}, P, \mathbf{1}_s) = \sum_{i=0}^{k-1} W_d^{\pi^{(i)}}(\mathbf{1}_{s,a}, P, \tilde{\mu}_i) \leq \sum_{i=0}^{k-1} W_d^{\pi^{(i)}}(\mathbf{1}_{s,a}, P, \mathbf{1}_s) \leq 6k \max_{\pi\in\Pi_{\text{sta}}} W_d^\pi(\mathbf{1}_{s,a}, P, \mathbf{1}_s),$$

where $\pi^{(i)}$ is defined as $\pi_{h'}^{(i)}(a|s) = \pi_{id+h'}(a|s)$. The first inequality uses the fact that to get a reward $\mathbf{1}_{s,a}$, the best initial state is s. The second inequality uses Lemma 17. The proof is finished by taking maximization over π. ■

Lemma 17 *For any horizon d and state-action pair (s, a), it holds that*

$$\max_{\pi\in\Pi_{\text{sta}}} W_d^\pi(\boldsymbol{1}_{s,a}, P, \boldsymbol{1}_s) \geq \frac{1}{6} \max_{\pi} W_d^\pi(\boldsymbol{1}_{s,a}, P, \boldsymbol{1}_s).$$

Proof Let π^* be the optimal stationary policy with respect to reward $\mathbf{1}_{s,a}$ under transition P and discounted factor $\gamma = 1 - \frac{1}{d}$. Let $\tilde{\mu}_i$ denote the distribution of s_{id+1} under P following π with initial distribution $\mathbf{1}_s$.

Then we have that for any policy π',

$$\begin{aligned}
W_d^{\pi^*}(\mathbf{1}_{s,a}, P, \mathbf{1}_s) &\geq \frac{1}{2}\sum_{i=0}^{\infty}\gamma^{di} W_d^{\pi^*}(\mathbf{1}_{s,a}, P, \mathbf{1}_s) \\
&\geq \frac{1}{2}\sum_{i=0}^{\infty}\gamma^{di} W_d^{\pi^*}(\mathbf{1}_{s,a}, P, \tilde{\mu}_i) \\
&\geq \frac{1}{2}\sum_{i=1}^{\infty}\gamma^{i-1}\mathbb{P}_{\pi^*}[(s_i, a_i) = (s, a)] \\
&\geq \frac{1}{2}\sum_{i=1}^{\infty}\gamma^{i-1}\mathbb{P}_{\pi'}[(s_i, a_i) = (s, a)] \\
&\geq \frac{1}{6}\sum_{i=1}^{d}\mathbb{P}_{\pi'}[(s_i, a_i) = (s, a)] \\
&= \frac{1}{6} W_d^{\pi'}(\mathbf{1}_{s,a}, P, \mathbf{1}_s).
\end{aligned}$$

The proof is completed by taking maximization over π'. ■

Recall that $W_\gamma^\pi(r, P, \mu_1)$ denotes the discounted accumulative reward with reward r, transition P, policy π, initial distribution μ_1 and discounted factor γ.

Lemma 18 *Let d_1, d_2 be positive integers such that $d_1 \geq 10S\log(S)d_2$. Let $\gamma = 1 - 1/d_2$. Let (s, a), a non-negative reward r and a stationary policy π be fixed. Then we have that*

$$\frac{1}{3} W_{d_2}^\pi(\boldsymbol{1}_{s,a}, p, \boldsymbol{1}_s) \leq W_\gamma^\pi(\boldsymbol{1}_{s,a}, p, \boldsymbol{1}_s) \leq 3 W_{d_2}^\pi(\boldsymbol{1}_{s,a}, p, \boldsymbol{1}_s) \tag{13}$$

$$W_\gamma^\pi(r, p, \boldsymbol{1}_s) \geq \frac{1}{3} W_{d_2}^\pi(r, p, \boldsymbol{1}_s) \tag{14}$$

$$W_\gamma^\pi(r, p, \boldsymbol{1}_s) \leq 10 W_{d_1}^\pi(r, p, \boldsymbol{1}_s). \tag{15}$$

Proof The left side of (13) holds because $\gamma^{d_2} = (1 - 1/d_2)^{d_2} \geq 1/3$. As for the right side, letting μ_i denote the distribution of s_{id_2+1} following π starting from s, we have that

$$W_\gamma^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_s) \leq \sum_{i=0}^{\infty}\gamma^{d_2 i} W_{d_2}^\pi(\mathbf{1}_{s,a}, p, \mu_i) \leq W_{d_2}^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_s)\sum_{i=0}^{\infty}(1 - d_2)^{d_2 i} \leq 3 W_{d_2}^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_s). \tag{16}$$

The first and second inequalities are by ignoring the terms with index larger than d_2. (14) holds by the fact $\gamma^{d_2} = (1 - 1/d_2)^{d_2} \geq 1/3$.

To prove (15), letting μ_i denote the distribution of $s_{d_1 i+1}$ following π starting from s, we have that

$$W_\gamma^\pi(r, p, \mathbf{1}_s) \leq \sum_{i=0}^{\infty}\gamma^{d_1 i} W_{d_1}^\pi(r, p, \mu_i) \leq \sum_{i=0}^{\infty} e^{-10S\log(S)i} W_{d_1}^\pi(r, p, \mu_i).$$

Next, we have

$$\begin{aligned}\sum_{i=2^k}^{2^{k+1}-1} W_{d_1}^{\pi}(r,p,\mu_i) &\le \sum_{i=0}^{2^{k+1}-1} W_{d_1}^{\pi}(r,p,\mu_i)\\ &= W_{2^{k+1}d_1}^{\pi}(r,p,\mathbf{1}_s)\\ &\le \exp(5(k+1)S\log(S))W_{d_1}^{\pi}(r,p,\mathbf{1}_s).\end{aligned}$$

The second inequality we used Lemma 19 for $(k+1)$ times. Therefore, we obtain

$$\begin{aligned}&\sum_{i=0}^{\infty} e^{-10S\log(S)i} W_{d_1}^{\pi}(r,p,\mu_i) \le \sum_{k=0}^{\infty} e^{-10S\log(S)2^k} \sum_{i=2^k}^{2^{k+1}-1} W_{d_1}^{\pi}(r,p,\mu_i)\\ &\le \sum_{k=0}^{\infty} e^{-S\log(S)(10\cdot 2^k-5(k+1))} W_{\gamma}^{\pi}(r,p,\mathbf{1}_s) \le 10 W_{\gamma}^{\pi}(r,p,\mathbf{1}_s). \qquad (17)\end{aligned}$$

The proof is completed. ■

Lemma 19 *[Lemma 4.6 in Li et al. (2021b)] Suppose $d \ge S \ge 5$. Then*

$$W_{2d}^{\pi}(r,p,\mu) \le 4S^{4S} W_d^{\pi}(r,p,\mu) \le \exp(5S\log(S)) W_d^{\pi}(r,p,\mu)$$

for any proper r, p, μ and stationary policy π.

Lemma 20 *Let $X_1, X_2, \ldots, X_n$ be i.i.d. positive random variables. Define $\tau_H := \min\{i | \sum_{j=1}^{i} X_j \ge H\}$. Then we have that*

$$\Pr\left[\tau_H \ge \frac{1}{2}\mathbb{E}[\tau_H] - 1\right] \ge \frac{1}{2}. \qquad (18)$$

Proof The proof comes from the analysis in Corollary 4.9 in Li et al. (2021a). Clearly $\mathbb{E}[\tau_H] \ge 1$. Let $\tau' = \lceil \tau_H \rceil - 1$, it suffices to prove that

$$\Pr\left[\sum_{j=1}^{\tau'/2} X_j < H\right] \ge \frac{1}{2}. \qquad (19)$$

Define $X_1' = \min\{X_1, H\}$.

By the stopping time theorem, we have that $\tau'\mathbb{E}[X_1'] \le H$, it then holds that $\frac{\tau'}{2}\mathbb{E}[X_1'] \le \frac{H}{2}$. By Markov's inequality we have that

$$\Pr\left[\sum_{j=1}^{\tau'/2} X_j' < H\right] \ge \frac{1}{2}. \qquad (20)$$

Noting that $\sum_{j=1}^{\tau'/2} X_j' < H$ implies $\sum_{j=1}^{\tau'/2} X_j < H$, we finish the proof.

■

By Lemma 20, we further have that

Lemma 21 (Restatement of Lemma 4) *For any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\pi \in \Pi_{\text{sta}}$ such that $\pi(s) = a$, we have that $\Pr\left[N \ge \frac{1}{4} W_d^{\pi}(P, \mathbf{1}_{s,a}, \mathbf{1}_s)\right] \ge \frac{1}{2}$ for any horizon d, where N is the visit count of (s,a) following π under P in d steps with the initial distribution as $\mathbf{1}_s$.*

Appendix D. Proof of Lemma 6

Proof [Proof of Lemma 6] Recall the definition of π^k, $\tilde{P}^k$ and O^k in Algorithm 1. Recall that $d = \frac{SH}{S+1}$. We use the following two lemmas below.

Lemma 22 *With probability* $1-\delta$, *we have that*

$$\max_{\pi} W_d^{\pi}(\boldsymbol{1}_z, \mathrm{Clip}(P, O^{K_1+1}), \mu_1) \le O\left(\frac{S^7A^3\iota}{K_1}\mathrm{polylog}(SAK)\right).$$

Lemma 23 (Formal statement of Lemma 2) *For any* $O \subset \mathcal{S}\times\mathcal{A}$ *and* $\tilde{d} \ge 1$, *we have that*

$$\max_{\pi} X^{\pi}_{(S+2)\tilde{d}}(O, P, \mu_1) \le S^2 \max_{\pi} X^{\pi}_{(S+1)\tilde{d}}(O, P, \mu_1).$$

Given these two lemmas and setting $\tilde{d} = \frac{H}{S+2}$, we can have that

$$\begin{aligned}
\max_{\pi} W_H^{\pi}(\mathbf{1}_z, \mathrm{Clip}(P, O^{K_1+1}), \mu_1) &= \max_{\pi} \mathbb{P}_{\pi}\left[\exists h \in [(S+1)\tilde{d}], (s_h, a_h) \in O^{K_1+1}\right] \\
&= \max_{\pi} X^{\pi}_{(S+1)\tilde{d}}(O^{K_1+1}, P, \mu_1) \\
&\le S^2 \max_{\pi} X^{\pi}_{S\tilde{d}}(O^{K_1+1}, P, \mu_1) \\
&= S^2 \max_{\pi} X^{\pi}_{d}(O^{K_1+1}, P, \mu_1) \\
&\le O\left(\frac{S^9A^3\iota}{K_1}\mathrm{polylog}(SAK)\right)
\end{aligned}$$

This completes the proof of Lemma 6. ■

Below we prove these two lemmas.

Proof [Proof of Lemma 23] Inspired by the analysis in Li et al. (2021b), we regard each $\tilde{d}$ steps as one *big step*, which reducing the problem to a special case where $\tilde{d} = 1$. Then we construct a mapping from the set of trajectories of length $(S+1)$ with final state as z to the set of trajectories of length S with final state as z, which bounds the probability of the former trajectories using the probability of latter trajectories.

Define $\overline{P}_{s,a} = P_{s,a}$ for any $(s,a) \notin O$, $\overline{P}_{s,a} = \mathbf{1}_z$ for $(s,a) \in O$ and $\overline{P}_{z,a} = \mathbf{1}_z$ for any a. In words, $\overline{P}$ is a copy of P, except for redirecting $(s,a) \in O$ to a absorbed state z.

Let $\mathcal{P}(1) = \{p | p_s \in \mathrm{Conv}(\{\overline{P}_{s,a}\}_{a\in\mathcal{A}})\}$ be the set of all possible $1-$step transition probability under $\overline{P}$, where $\mathrm{Conv}(\mathcal{X})$ denote the convex hull of a set $\mathcal{X}$. Let $\mathcal{P}(\tilde{d}) = \{\Pi_{i=1}^{\tilde{d}} p_i | p_i \in \mathcal{P}(1), \forall i\}$, which is the set of l-th step transition probability with respect to $\overline{P}$.

By definition, we have that

$$\max_{\pi} X^{\pi}_{(S+2)\tilde{d}}(O, P, \mu_1) = \max_{\{p^{(i)}\in\mathcal{P}(l)\}_{i=1}^{S+1}} \mu_1^{\top}\Pi_{i=1}^{S+1} p^{(i)}\mathbf{1}_z \tag{21}$$

$$\max_{\pi} X^{\pi}_{(S+1)\tilde{d}}(O, P, \mu_1) = \max_{\{p^{(i)}\in\mathcal{P}(l)\}_{i=1}^{S+1}} \mu_1^{\top}\Pi_{i=1}^{S} p^{(i)}\mathbf{1}_z. \tag{22}$$

Let $\mathcal{T} = \{\{\tilde{s}_i \in \mathcal{S} \cup \{z\}\}_{i=1}^{S+1}\}$ be the set of all possible trajectories with length $S+1$. Then we have that for any $\{p^{(i)} \in \mathcal{P}(l)\}_{i=1}^{S+1}$

$$\mu_1^\top \Pi_{i=1}^{S+1} p^{(i)} \mathbf{1}_z = \sum_{\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T}} \mu_1(\tilde{s}_1) \Pi_{i=1}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z}. \tag{23}$$

For any trajectory $\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T}$, by the pigeon hole principle, it either holds that $\tilde{s}_{S+1} = z$ or $\exists i_1, i_2$ such that $\tilde{s}_{i_1} = \tilde{s}_{i_2}$. In the first case, we define that $\mathcal{T}' = \{\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T}, : \tilde{s}_{S+1} = z\}$. Then we have that for any $\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T}'$,

$$\mu_1(\tilde{s}_1) \Pi_{i=1}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} = \mu_1(\tilde{s}_1) \Pi_{i=1}^{S-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S)}_{\tilde{s}_S, z}. \tag{24}$$

Taking sum, we have that

$$\sum_{\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T}'} \mu_1(\tilde{s}_1) \Pi_{i=1}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} = \sum_{\{\tilde{s}_i\}_{i=1}^{S} \in \mathcal{T}} \mu_1(\tilde{s}_1) \Pi_{i=1}^{S-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S)}_{\tilde{s}_S, z} = \mu_1^\top \Pi_{i=1}^{S} p^{(i)} \mathbf{1}_z. \tag{25}$$

In the second case, for a fixed (i_1, i_2), we define $\mathcal{T}(i_1, i_2) = \{\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T} : \tilde{s}_{S+1} \neq z, \tilde{s}_{i_1} = \tilde{s}_{i_2}\}$ for $1 \leq i_1 < i_2 \leq S+1$.

Then we have that

$$\begin{aligned}
&\sum_{\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T}(i_1, i_2)} \mu_1(\tilde{s}_1) \Pi_{i=1}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} \\
&= \sum_{\{\tilde{s}_i\}_{i=1}^{S+1} \in \mathcal{T}(i_1, i_2)} \mu_1(\tilde{s}_1) \Pi_{i=1}^{i_1-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot \Pi_{i=i_1}^{i_2-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot \Pi_{i=i_2}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z}. \\
&\leq \sum_{\{\tilde{s}_i\}_{i=1}^{i_1}, \{\tilde{s}\}_{i=i_2+1}^{S+1}} \mu_1(\tilde{s}_1) \Pi_{i=1}^{i_1-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot \Pi_{i=i_2}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} \cdot \sum_{\{\tilde{s}_i\}_{i=i_1+1}^{i_2-1}} \Pi_{i=i_1}^{i_2-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \\
&\leq \sum_{\{\tilde{s}_i\}_{i=1}^{i_1}, \{\tilde{s}\}_{i=i_2+1}^{S+1}} \mu_1(\tilde{s}_1) \Pi_{i=1}^{i_1-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot \Pi_{i=i_2}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} \qquad (26) \\
&= \mu_1^\top \Pi_{i=1}^{i_1-1} p^{(i)} \cdot \Pi_{i=i_2}^{S+1} p^{(i)} \mathbf{1}_z. \qquad (27)
\end{aligned}$$

Here (26) holds by the fact that $\sum_{\{\tilde{s}_i\}_{i=i_1+1}^{i_2-1}} \Pi_{i=i_1}^{i_2-1} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}}$ is the probability of transiting to $\tilde{s}_{i_1}$ from $\tilde{s}_{i_1}$ using $i_2 - i_1$ steps under transition $\{p^{(i)}\}_{i=i_1}^{i_2-1}$, which is bounded by 1.

By (27) and (25), we obtain that

$$\begin{aligned}
&\mu_1^\top \Pi_{i=1}^{S+1} p^{(i)} \mathbf{1}_z \\
&= \sum_{\{\tilde{s}\}_{i=1}^{S+1} \in \mathcal{T}} \mu_1(\tilde{s}_1) \Pi_{i=1}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} \\
&\leq \sum_{\{\tilde{s}\}_{i=1}^{S+1} \in \mathcal{T}'} \mu_1(\tilde{s}_1) \Pi_{i=1}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} + \sum_{1 \leq i_1 < i_2 \leq S+1} \sum_{\{\tilde{s}\}_{i=1}^{S+1} \in \mathcal{T}(i_1, i_2)} \mu_1(\tilde{s}_1) \Pi_{i=1}^{S} p^{(i)}_{\tilde{s}_i, \tilde{s}_{i+1}} \cdot p^{(S+1)}_{\tilde{s}_{S+1}, z} \\
&\leq \mu_1^\top \Pi_{i=1}^{S} p^{(i)} \mathbf{1}_z + \sum_{1 \leq i_1 < i_2 \leq S+1} \mu_1^\top \Pi_{i=1}^{i_1-1} p^{(i)} \cdot \Pi_{i=i_2}^{S+1} p^{(i)} \mathbf{1}_z \\
&\leq S^2 \max_{\{p^{(i)} \in \mathcal{P}(l)\}_{i=1}^{S}} \mu_1^\top \Pi_{i=1}^{S} p^{(i)} \mathbf{1}_z. \qquad (28)
\end{aligned}$$

Noting that (28) holds for any $\{p^{(i)} \in \mathcal{P}(l)\}_{i=1}^{S+1}$, we conclude that

$$\max_{\{p^{(i)}\in\mathcal{P}(l)\}_{i=1}^{S+1}} \mu_1^\top \Pi_{i=1}^{S+1} p^{(i)} \mathbf{1}_z \le S^2 \max_{\{p^{(i)}\in\mathcal{P}(l)\}_{i=1}^{S}} \mu_1^\top \Pi_{i=1}^{S} p^{(i)} \mathbf{1}_z.$$

The proof is completed by (21) and (22). ■

D.1. Proof of Lemma 22

The following lemma guarantees that all state-action pairs in the known set (i.e., $\notin O^k$), we have collected enough data.

Lemma 24 *Recall the definition of $U(s,a) = \max_\pi W_H^\pi(\mathbf{1}_{s,a}, P, \mu_1)$. With probability $1 - 10SAK\delta$, for each $1 \le k \le K_1$ and each $(s,a) \notin O^k$, we have that*

$$N^k(s,a) \ge \frac{C}{71S(S+1)\log(S)} U(s,a), \tag{29}$$

where C is an universal constant.

Proof [Proof of Lemma 24] By Lemma 7, with probability $1 - SAK\delta$, it holds that $N^k(s,a) \ge C \max_{\pi\in\Pi_{\text{sta}}} W^\pi_{\frac{H}{2S(S+1)\log(S)}}(\mathbf{1}_{s,a}, P, \mathbf{1}_s)$ for some constant C.

Now we can lower bound

$$\begin{aligned} N^k(s,a) &\ge C \max_{\pi\in\Pi_{\text{sta}}} W^\pi_{\frac{H}{2S(S+1)\log(S)}}(\mathbf{1}_{s,a}, P, \mathbf{1}_s) \\ &\ge \frac{C}{12S(S+1)\log(S)} \max_\pi W_H^\pi(\mathbf{1}_{s,a}, P, \mathbf{1}_s) \\ &\ge \frac{C}{12S(S+1)\log(S)} \max_\pi W_H^\pi(\mathbf{1}_{s,a}, P, \mu_1) \\ &= \frac{C}{12S(S+1)\log(S)} U(s,a) \end{aligned}$$

where the first inequality we used Lemma 3, and the second inequality we used that $\mathbf{1}_s$ is the optimal initial distribution for reward $\mathbf{1}_{s,a}$.

■

Proof [Proof of Lemma 22] Now we proceed to prove Lemma 22. We first make some definitions.

Define $\bar{P}^k = \text{Clip}(P, O^k)$. Note that O^k never appears under $\bar{P}^k$. We then define the state-action space of $\bar{P}^k$ as Φ^k. Below we continue the analysis for the k-th episode in the first stage under the context of $\bar{P}^k$. Let $\bar{r}^k = \mathbf{1}_z$. Note that O^k varies in k, then the definition of z varies in different episodes. As a result, $\bar{P}^k$ and $\bar{r}^k$ also vary in k.

Define $\tilde{R} = \sum_{k=1}^{K_1} \max_\pi W^\pi(\mathbf{1}_z, \bar{P}^k, \mu_1) - \sum_{k=1}^{K_1}\sum_{h=1}^{H} \bar{r}^k(s_h^k, a_h^k)$, which can be viewed as the regret. Note this is different from the regret in standard MDP because the reward is not fixed (z depends on k).

Recall that $(\pi^k, \tilde{P}^k) = \arg\max_{\pi, p\in\mathcal{P}^k} X_d^\pi(O^k, p, \mu_1)$. Let $\tilde{p}^k = \text{Clip}(\tilde{P}^k, O^k)$, $\hat{p}^k = \text{Clip}(\hat{P}^k, O^k)$ and $\{\tilde{V}_h^k(s)\}_{(h,s)\in[H]\times\mathcal{S}^k}$ be the value function with reward $\mathbf{1}_z$ and transition $\tilde{p}^k$.

Define $\check{\mathcal{J}} := \{(k,h) : k \in [K_1], \exists h' \le h, (s,a) \in \mathcal{S}\times\mathcal{A}, N^k_{h'}(s,a) > 2N^k(s,a)+1\}$ and $i^k_h = \mathbb{I}[(k,h) \notin \check{\mathcal{J}}]$.

Following the regret analysis in Zhang et al. (2021b), by the optimality of $\tilde{P}^k$, we have that

$$
\begin{aligned}
&\sum_{k=1}^{K_1} \max_{\pi} W^{\pi}_d(\mathbf{1}_z, \bar{P}^k, \mu_1) - \sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s^k_h, a^k_h) \\
&\le \sum_{k=1}^{K_1} \tilde{V}^k_1(s^k_1) i^k_1 - \sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s^k_h, a^k_h) \\
&= \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\tilde{V}^k_h(s^k_h) i^k_h - \tilde{V}^k_{h+1}(s^k_{h+1}) i^k_{h+1}\right) - \sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s^k_h, a^k_h) \\
&= \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\tilde{p}^k_{s^k_h, a^k_h} \tilde{V}^k_{h+1} i^k_h + \bar{r}^k(s^k_h, a^k_h) i^k_h - \mathbf{1}_{s^k_{h+1}} \tilde{V}^k_{h+1} i^k_{h+1}\right) - \sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s^k_h, a^k_h) \\
&\le \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\tilde{p}^k_{s^k_h, a^k_h} \tilde{V}^k_{h+1} - \mathbf{1}_{s^k_{h+1}} \tilde{V}^k_{h+1}\right) i^k_{h+1} + \sum_{k=1}^{K_1}\sum_{h=1}^{d} \tilde{p}^k_{s^k_h, a^k_h} \tilde{V}^k_{h+1}(i^k_h - i^k_{h+1}) \\
&= \sum_{k=1}^{K_1} \left(\sum_{h=1}^{d} (\tilde{p}^k_{s^k_h, a^k_h} - \bar{P}^k_{s^k_h, a^k_h}) \tilde{V}^k_{h+1} i^k_{h+1} + (\bar{P}^k_{s^k_h, a^k_h} - \mathbf{1}_{s^k_{h+1}}) \tilde{V}^k_{h+1} i^k_{h+1} \right) + \sum_{k=1}^{K_1}\sum_{h=1}^{d} \tilde{p}^k_{s^k_h, a^k_h} \tilde{V}^k_{h+1}(i^k_h - i^k_{h+1}) \\
&= \sum_{k=1}^{K_1} \left(\sum_{h=1}^{d} \left((\tilde{p}^k_{s^k_h, a^k_h} - \bar{P}^k_{s^k_h, a^k_h})(\tilde{V}^k_{h+1} - \bar{P}^k_{s^k_h, a^k_h} \tilde{V}^k_{h+1} \cdot \mathbf{1}) i^k_{h+1} + (\bar{P}^k_{s^k_h, a^k_h} - \mathbf{1}_{s^k_{h+1}}) \tilde{V}^k_{h+1} i^k_{h+1} \right) \right) \\
&\qquad\qquad + \sum_{k=1}^{K_1}\sum_{h=1}^{d} \tilde{p}^k_{s^k_h, a^k_h} \tilde{V}^k_{h+1}(i^k_h - i^k_{h+1}) \qquad (30) \\
&\le 20 \sum_{k=1}^{K_1}\sum_{h=1}^{d}\sum_{s'\neq z} \left(\sqrt{\frac{\bar{P}^k_{s^k_h, a^k_h, s'} \iota}{N^k(s^k_h, a^k_h)}} + \frac{\iota}{N^k(s^k_h, a^k_h)} \right) |\tilde{V}^k_{h+1}(s') - \bar{P}^k_{s^k_h, a^k_h} \tilde{V}^k_{h+1}| i^k_{h+1} \\
&\qquad\qquad + \sum_{k=1}^{K_1}\sum_{h=1}^{d} (\bar{P}^k_{s^k_h, a^k_h} - \mathbf{1}_{s^k_{h+1}}) \tilde{V}^k_{h+1} i^k_{h+1} + \sum_{k=1}^{K_1} \mathbb{I}[\exists h \in [d], (k,h) \in \bar{\mathcal{J}}] \qquad (31) \\
&\le 20 \sqrt{\sum_{k=1}^{K_1}\sum_{h=1}^{d} \frac{S\, i^k_{h+1} \iota}{N^k(s^k_h, a^k_h)}} \cdot \sqrt{\sum_{k=1}^{K_1}\sum_{h=1}^{d} \mathbb{V}(\bar{P}^k_{s^k_h, a^k_h}, \tilde{V}^k_{h+1}) i^k_{h+1}} + 20 \sum_{k=1}^{K_1}\sum_{h=1}^{d} \frac{S\, i^k_{h+1} \iota}{N^k(s^k_h, a^k_h)} \\
&\qquad\qquad + \sum_{k=1}^{K_1}\sum_{h=1}^{d} (\bar{P}^k_{s^k_h, a^k_h} - \mathbf{1}_{s^k_{h+1}}) \tilde{V}^k_{h+1} i^k_{h+1} + \sum_{k=1}^{K_1} \mathbb{I}[\exists h \in [d], (k,h) \in \bar{\mathcal{J}}]. \qquad (32)
\end{aligned}
$$

In the first equality we used the fact that $i_1^k = 1$. In (30), we used the fact that $\bar{P}^k_{s_h^k,a_h^k}\tilde{V}^k_{h+1}\cdot\mathbf{1}$ is a constant factor and $\|\tilde{p}^k_{s_h^k,a_h^k}\|_1 = \|\bar{P}^k_{s_h^k,a_h^k}\|_1 = 1$. In (31), we used the fact that

$$|\tilde{p}^k_{s_h^k,a_h^k,s'} - \bar{P}^k_{s_h^k,a_h^k,s'}| \le \left(\sqrt{\frac{4\hat{P}^k_{s_h^k,a_h^k,s'}\iota}{N^k(s_h^k,a_h^k)}} + \frac{5\iota}{N^k(s_h^k,a_h^k)}\right) \le 20\left(\sqrt{\frac{\bar{P}^k_{s_h^k,a_h^k,s'}\iota}{N^k(s_h^k,a_h^k)}} + \frac{\iota}{N^k(s_h^k,a_h^k)}\right)$$

for $s' \neq z$. Lastly, (32) holds by Cauchy's inequality.

Define

$$\begin{aligned}
T_1 &= \sum_{k=1}^{K_1}\sum_{h=1}^{d}\frac{i^k_{h+1}\iota}{N^k(s_h^k,a_h^k)}\\
T_2 &= \sum_{k=1}^{K_1}\sum_{h=1}^{d}\mathbb{V}(\bar{P}^k_{s_h^k,a_h^k},\tilde{V}^k_{h+1})i^k_{h+1}\\
T_3 &= \sum_{k=1}^{K_1}\sum_{h=1}^{d}(\bar{P}^k_{s_h^k,a_h^k} - \mathbf{1}_{s^k_{h+1}})\tilde{V}^k_{h+1}i^k_{h+1}\\
T_4 &= \sum_{k=1}^{K_1}\mathbb{I}[\exists h\in[d],(k,h)\in\bar{\mathcal{J}}].
\end{aligned}$$

The following lemma bounds these four terms.

Lemma 25 *With probability* $1-4\delta$, $T_1, T_4 \le SAB$ *with* $B = O(\iota\text{polylog}(SAK))$,

$$\begin{aligned}
T_2 &\le O\left(\text{polylog}(SAK)\left(\sum_{k=1}^{K_1}\sum_{h=1}^{d}\bar{r}^k(s_h^k,a_h^k) + S^2AB\iota\right)\right)\\
&\le O\left(\text{polylog}(SAK)(S^7A^3\iota + S^2A\iota^2)\right),
\end{aligned}$$

and $T_3 \le O\left(\sqrt{S^7A^3\iota^2 + S^2A\iota^3}\text{polylog}(SAK)\right)$.

By Lemma 25, with probability $1-4\delta$, we have that

$$\begin{aligned}
&\sum_{k=1}^{K_1}\max_{\pi} W^{\pi}(\mathbf{1}_z,\bar{P}^k,\mu_1) - \sum_{k=1}^{K_1}\sum_{h=1}^{d}\bar{r}^k(s_h^k,a_h^k)\\
&\le O\left(\text{polylog}(SAK)\left(\sqrt{S^2AB\iota\sum_{k=1}^{K_1}\sum_{h=1}^{d}\bar{r}^k(s_h^k,a_h^k)} + S^2AB\iota\right)\right)\\
&\le O\left(\text{polylog}(SAK)\left(\sum_{k=1}^{K_1}\sum_{h=1}^{d}\bar{r}^k(s_h^k,a_h^k) + S^2AB\iota\right)\right),
\end{aligned}$$

where it follows that

$$\sum_{k=1}^{K_1}\max_{\pi} W^{\pi}(\mathbf{1}_z,\bar{P}^k,\mu_1) \le O\left(\text{polylog}(SAK)\left(\sum_{k=1}^{K_1}\sum_{h=1}^{d}\bar{r}^k(s_h^k,a_h^k) + S^2AB\iota\right)\right). \tag{33}$$

Let $u = \max_\pi W_d^\pi(\mathbf{1}_z, \bar{P}^{K_1+1}, \mu_1)$ be the maximal possible probability of visiting O^{k+1}. Noting that O^k is non-increasing in k, the probability of visiting O^k is also non-increasing in k. By (33) we have that

$$K_1 u \leq O\left(\text{polylog}(SAK)\left(\sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s_h^k, a_h^k) + S^2AB\iota\right)\right) \tag{34}$$

$$= O((S^7A^3\iota + S^2A\iota^2)\text{polylog}(SAK)), \tag{35}$$

which implies that $u = O\left(\frac{S^7A^3\iota+S^2A\iota}{K_1}\text{polylog}(SAK)\right)$.

The proof is completed.

■

It remains to prove Lemma 25.

Proof [Proof of Lemma 25] We start with bounding T_1 and T_4. Define $\mathcal{L}(s,a) = \{k \in [K_1] : N^{k+1}(s,a) - N^k(s,a) \geq K^2U(s,a)\}$. By Lemma 10, $|\mathcal{L}(s,a)| \leq O(1/K^2 + \iota)$ with probability $1-\delta$. By Lemma 24, $N^k(s,a) \geq \frac{C}{12S(S+1)\log(S)}U(s,a)$ for any $(s,a) \notin O^k$. By definition

$$T_1 \leq 2\sum_{k=1}^{K_1}\sum_{(s,a)\notin O^k} \min\left\{\log(\frac{N^{k+1}(s,a)}{N^k(s,a)}), 1\right\}.$$

Fix $(s,a) \notin O^{K_1+1}$. Noting that O^k is non-increasing in k, there exists some k', such that $(s,a) \notin O^k$ for $k \geq k'$ and $(s,a) \in O^{k'-1}$.

Suppose $\mathcal{L}_2(s,a) \cap \{k : k' \leq k \leq K_1\} = \{k_1, k_2, ..., \}$. Let $k_0 = k' - 1$. We have that

$$\begin{aligned}
&\sum_{k=k':k\notin\mathcal{L}(s,a)}^{K_1} \min\{\log(N^{k+1}(s,a)/N^k(s,a)), 1\} \\
&= \sum_{i\geq 0}\sum_{k=k_i+1}^{k_{i+1}-1} \log(N^{k+1}(s,a)/N^k(s,a)) \\
&\leq \sum_{i\geq 0} \log\left(\frac{K^3U(s,a) + N^{k_i+1}}{N^{k_i+1}}\right) \\
&\leq |\mathcal{L}(s,a)|\log(K^3U(s,a)/N^{k'}(s,a)) \\
&\leq O(\text{polylog}(SAK)\iota).
\end{aligned}$$

It then holds that

$$\begin{aligned}
&\sum_{k=k'}^{K_1} \min\left\{\log(\frac{N^{k+1}(s,a)}{N^k(s,a)}), 1\right\} \\
&\leq |\mathcal{L}(s,a)| + |\mathcal{L}(s,a)|\log(K^3U(s,a)/N^{k'}(s,a)) \\
&\leq O(\text{polylog}(SAK)\iota).
\end{aligned} \tag{36}$$

Taking sum over (s,a), we obtain that $T_1 \leq O(\text{polylog}(SAK)SA\iota)$

In a similar way we have that

$$T_4 \le \sum_{k=1}^{K_1} \sum_{(s,a)\notin O^k} \min\left\{\log(\frac{N^{k+1}(s,a)}{N^k(s,a)}), 1\right\} \le O(\text{polylog}(SAK)SA\iota). \tag{37}$$

To bound T_2, following regret analysis in Zhang et al. (2021b), we have that

$$\begin{aligned}
T_2 &= \sum_{k=1}^{K_1}\sum_{h=1}^{d} \mathbb{V}(\bar{P}^k_{s_h^k,a_h^k}, \tilde{V}^k_{h+1}) i^k_{h+1} \\
&= \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\bar{P}^k_{s_h^k,a_h^k}(\tilde{V}^k_{h+1})^2 - (\bar{P}^k_{s_h^k,a_h^k}\tilde{V}^k_{h+1})^2\right) i^k_{h+1} \\
&= \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\bar{P}^k_{s_h^k,a_h^k}(\tilde{V}^k_{h+1})^2 i^k_{h+1} - (\tilde{V}^k_h(s^k_h))^2 i^k_h\right) + \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left((\tilde{V}^k_h(s^k_h))^2 i^k_h - (\bar{P}^k_{s_h^k,a_h^k}\tilde{V}^k_{h+1})^2 i^k_{h+1}\right) \\
&\le \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\bar{P}^k_{s_h^k,a_h^k}(\tilde{V}^k_{h+1})^2 i^k_{h+1} - (\tilde{V}^k_h(s^k_h))^2 i^k_h\right) + 2\sum_{k=1}^{K_1}\sum_{h=1}^{d} \min\{\tilde{V}^k_h(s^k_h) i^k_h - \bar{P}^k_{s_h^k,a_h^k}\tilde{V}^k_{h+1} i^k_{h+1}, 0\} \\
&\le \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\bar{P}^k_{s_h^k,a_h^k}(\tilde{V}^k_{h+1})^2 i^k_{h+1} - (\tilde{V}^k_{h+1}(s^k_{h+1}))^2 i^k_{h+1}\right) + 2\sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s^k_h, a^k_h) + 8\sqrt{S^2ABtT_2} + 2T_4 \\
&\le \sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\bar{P}^k_{s_h^k,a_h^k}(\tilde{V}^k_{h+1})^2 i^k_{h+1} - (\tilde{V}^k_{h+1}(s^k_{h+1}))^2 i^k_{h+1}\right) + 2\sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s^k_h, a^k_h) + \frac{1}{2}T_2 + 8S^2ABt + 2B \\
&\le 2\sum_{k=1}^{K_1}\sum_{h=1}^{d} \left(\bar{P}^k_{s_h^k,a_h^k}(\tilde{V}^k_{h+1})^2 i^k_{h+1} - (\tilde{V}^k_{h+1}(s^k_{h+1}))^2 i^k_{h+1}\right) + 4\sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s^k_h, a^k_h) + 20S^2ABt.
\end{aligned} \tag{38}$$

where the last step is by solving T_2.

Define

$$T_5 = \sum_{k=1}^{K_1}\sum_{h=1}^{d} (\bar{P}^k_{s_h^k,a_h^k}(\tilde{V}^k_{h+1})^2 i^k_{h+1} - \tilde{V}^k_{h+1}(s^k_{h+1})^2 i^k_{h+1})$$

and

$$T_6 = \sum_{k=1}^{K_1}\sum_{h=1}^{d} \mathbb{V}(\bar{P}^k_{s_h^k,a_h^k}, (\tilde{V}^k_{h+1})^2 i^k_{h+1}).$$

By Lemma 9, we have that

$$T_6 \le 4\sum_{k,h} \mathbb{V}(\bar{P}^k_{s_h^k,a_h^k}, \tilde{V}^k_{h+1} i^k_{h+1}) = 4T_2. \tag{39}$$

We note that here we cannot directly use the Freedman's inequality because we do not know a tight upper bound of variance and using a naive upper bound will lead to a dependency on H. Instead, we resort to Lemma 14.

By Lemma 14, we have that

$$\mathbb{P}\left[\exists i, T_5 \ge 10\cdot 2^i \iota, T_6 \le 2^{2i}\iota\right] \le \delta, \tag{40}$$

which implies that

$$\mathbb{P}\left[\exists i, T_5 \geq 10 \cdot 2^i \iota, T_2 \leq 2^{2i-2} i\iota\right] \leq \delta. \tag{41}$$

Therefore, with probability $1-\delta$, for any $i \geq 1$, it either holds $T_5 < 10 \cdot 2^i \iota$ or $T_2 > 2^{2i-2} i\iota$. Then we have that

$$T_2 \leq T_5 + 4K + 20\sqrt{S^2AKL\iota} + 60S^2AB\iota \leq T_5 + 8K + 60S^2AL\iota.$$

Suppose $T_5 \geq C \geq 8K + 60S^2AB\iota$, then we have that

$$T_2 \geq \frac{T_5^2}{800\iota \log_2(C)} \geq \frac{C^2}{800\iota \log_2(C)} \geq 3C. \tag{42}$$

Then we have that $T_5 \geq T_2 - (8K + 60S^2AB\iota) \geq T_2 - C \geq 2C$. In this case, T_5 is infinite, which leads to a contradiction. Therefore, with probability $1-\delta$, $T_5 < 8K + 60S^2AB\iota$, and it follows that $T_2 \leq 16K + 480S^2AB\iota$. As a result, $T_6 \leq 64K + 480S^2AB\iota$ and $T_5 \leq \sqrt{800\iota(\log_2(K)+10)T_6} = O(\sqrt{\iota T_2})\text{polylog}(S, A, K, 1/\delta)$. Recall that $\mathcal{J} == \{k \in [K_1] | \exists h \in [d], a, (s_{h+1}^k, a) \in \mathcal{O}^k\}$. By Lemma 28, we have that $|\mathcal{J}| \leq O(S^7A^3\iota\text{polylog}(SAK))$. Therefore, we have

$$\begin{aligned}
T_2 &\leq 4\sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s_h^k, a_h^k) + 20S^2AB\iota + O(\sqrt{\iota T_2}\text{polylog}(S, A, K, 1/\delta)) \\
&\leq O\left(\text{polylog}(SAK)\left(\sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s_h^k, a_h^k) + S^2AB\iota\right)\right) \\
&\leq O\left(\text{polylog}(SAK)\left(S^7A^3\iota + S^2A\iota^2\right)\right).
\end{aligned} \tag{43}$$

Here (43) is by the fact that $\sum_{k=1}^{K_1}\sum_{h=1}^{d} \bar{r}^k(s_h^k, a_h^k) \leq |\mathcal{J}| \leq O(S^7A^3\iota\text{polylog}(SAK))$.

Using Lemma 14 again, and noting that with probability $1-\delta$, $T_2 \leq O(\text{polylog}(SAK)(S^7A^3\iota + S^2A\iota^2))$, we learn that with probability $1-\delta$

$$T_3 \leq O(\sqrt{T_2\iota} + \iota) \leq O\left(\sqrt{S^7A^3\iota^2 + S^2A\iota^3}\text{polylog}(SAK)\right).$$

The proof is finished.

■

Appendix E. Proof of Lemma 8

Notations Since the proof is independent of our main proof, we will re-use some notations for simplicity. We re-define $N^k(s,a,s')$ be the count of (s,a,s') before the k-th episode in the second stage. Let $N_h^k(s,a,s')$ be the count of (s,a,s') before the h-th step in the k-th episode in the second stage. We also define $N^k(s,a) = \max\{\sum_{s'} N^k(s,a,s'), 1\}$ and $N_h^k(s,a) = \max\{\sum_{s'} N_h^k(s,a,s'), 1\}$.

Define $(k,h) \leq (k',h')$ when $k' > k$ or $k' = k, h' \geq h$. Similarly we define $(k,h) < (k',h')$ when $k' > k$ or $k' = k, h' > h$. Let $\mathcal{F}_h^k = \sigma(\{s_{h'}^{k'}\}_{(k',h')<(k,h)})$

Now we bound the regret. Conditioned on $\mathcal{G}$, we have that $P \in \mathcal{P}^k$ for any $1 \leq k \leq K$. By induction on h, we have that $\max_\pi W_H^\pi(r, P, \mathbf{1}_{s_1^k}) \leq V_1^k(s_1^k)$ for any k.

Define $\mathcal{J} := \{(k,h) : \exists h' \le h, (s,a) \in \mathcal{S}\times\mathcal{A}, N^k_{h'}(s,a) > 2N^k(s,a)+1\}$ and $I^k_h = \mathbb{I}[(k,h)\notin\mathcal{J}]$. Let $\check{V}^k_h = V^k_h \cdot I^k_h$. Let h, s be fixed and $a = \pi^k_h(s)$. Using a similar argument in the proof of Lemma 22, . we have that

$$\begin{aligned}
&\check{V}^k_h(s) - P_{s,a}\check{V}^k_{h+1} \\
&\le r^k(s,a)I^k_h + \max_{p\in\mathcal{P}^k_{s,a}} pV^k_{h+1}\cdot(I^k_h - I^k_{h+1}) + \max_{p\in\mathcal{P}^k_{s,a}}(p-P_{s,a})\check{V}^k_{h+1} \qquad (44)\\
&\le r^k(s,a)I^k_h + (I^k_h - I^k_{h+1}) + \sum_{s'}\left(\sqrt{\frac{2P_{s,a,s'}\iota}{N^k(s,a)}} + \frac{\iota}{3N^k(s,a)}\right)|\check{V}^k_{h+1}(s') - P_{s,a}\check{V}^k_{h+1}|. \qquad (45)
\end{aligned}$$

By definition of π^k, the regret is bounded by

$$\begin{aligned}
&\sum_{k=1}^K\left(\max_\pi W^\pi_H(r,P,\mu_1) - \sum_{h=1}^H r(s^k_h,a^k_h)\right)\\
&\le \sum_{k=1}^K\left(\check{V}^k_1(s^k_1) - \sum_{h=1}^H r(s^k_h,a^k_h)\right)\\
&= \sum_{k=1}^K\left(\sum_{h=1}^H\left(r^k(s^k_h,a^k_h)I^k_h - r(s^k_h,a^k_h) + (\max_{p\in\mathcal{P}^k_{s^k_h,a^k_h}} p - P_{s^k_h,a^k_h})\check{V}^k_{h+1}\right) + (P_{s^k_h,a^k_h} - \mathbf{1}_{s^k_{h+1}})\check{V}^k_{h+1}\right)\\
&\qquad + \sum_{k=1}^K\sum_{h=1}^H \max_{p\in\mathcal{P}^k_{s,a}} pV^k_{h+1}(I^k_h - I^k_{h+1})\\
&\le 10\sum_{k=1}^K\left(\sum_{h=1}^H\left(\sum_{s'}\left(\sqrt{\frac{P_{s^k_h,a^k_h,s'}\iota}{N^k(s^k_h,a^k_h)}} + \frac{\iota}{N^k(s^k_h,a^k_h)}\right)|\check{V}^k_{h+1}(s') - P_{s^k_h,a^k_h}\check{V}^k_{h+1}|\right)\right) +\\
&\qquad + \sum_{k=1}^K\sum_{h=1}^H(P_{s^k_h,a^k_h} - \mathbf{1}_{s^k_{h+1}})\check{V}^k_{h+1} + \sum_{k=1}^K\mathbb{I}[\exists h,(k,h)\in\mathcal{J}]. \qquad (46)
\end{aligned}$$

Here (46) is by the definition of I^k_h and Lemma 11.

Let

$$\begin{aligned}
M_1 &= 10\sum_{k=1}^K\sum_{h=1}^H\left(\sum_{s'}\left(\sqrt{\frac{P_{s^k_h,a^k_h,s'}\iota}{N^k(s^k_h,a^k_h)}} + \frac{\iota}{N^k(s^k_h,a^k_h)}\right)|\check{V}^k_{h+1}(s') - P_{s^k_h,a^k_h}\check{V}^k_{h+1}|\right)\\
M_2 &= \sum_{k=1}^K\sum_{h=1}^H(P_{s^k_h,a^k_h} - \mathbf{1}_{s^k_{h+1}})\check{V}^k_{h+1}\\
M_3 &= \sum_{k=1}^K\mathbb{I}[\exists h,(k,h)\in\mathcal{J}].
\end{aligned}$$

The following lemma is crucial in bounding M_1 and M_3 and it shows the usefulness of stage 1.

Lemma 26 *Define* $L = \max_{(s,a)\notin O^{k+1}}\sum_k \min\{\log(N^{K_1+1}(s,a)/N^k(s,a)),1\}$. *With probability* $1-2SA\delta$, $M_3 \le U := SAL + O(S^8A^3K\iota/K_1) \le O\left(\frac{S^8A^3K\iota}{K_1}\text{polylog}(SAK)\right)$ *and* $L \le O(\iota\text{polylog}(SAK))$.

Proof [Proof of Lemma 26] Define $\mathcal{B}_1 = \{k \in [K] : \exists h, (s_h^k, a_h^k) \in O^{K_1+1}\}$ and $\mathcal{B}_2(s,a) = \{k \in [K] : (s,a) : N^{k+1}(s,a) - N^k(s,a) \geq K^2 U(s,a)\}$. By Lemma 6 and 10, with probability $1-\delta$, we have that $|\mathcal{B}_1| \leq O\left(\frac{S^8A^3K\iota}{K_1} + \iota\right)$. By definition of $U(s,a)$ and Lemma 10, with probability $1 - SA\delta$, $|\mathcal{B}_2(s,a)| \leq (1/K^2 + \iota)$ for any (s,a). By definition, we have that

$$\begin{aligned} M_3 &= \sum_{k=1}^{K} \mathbb{I}[\exists h, (k,h) \in \mathcal{J}] \\ &\leq |\mathcal{B}_1| + \sum_{k=1}^{K} \max_{(s,a)\notin O^{K_1+1}} \min\{\log(N^{k+1}(s,a)/N^k(s,a)), 1\}. \end{aligned} \tag{47}$$

Let $(s,a) \notin O^{K_1+1}$ be fixed. Suppose $\mathcal{B}_2(s,a) = \{k_1, k_2, ..., \}$. Let $k_0 = 0$. Then

$$\begin{aligned} &\sum_{k\notin\mathcal{B}_2(s,a)} \min\{\log(N^{k+1}(s,a)/N^k(s,a)), 1\} \\ &= \sum_{i\geq 0} \sum_{k=k_i+1}^{k_{i+1}-1} \log(N^{k+1}(s,a)/N^k(s,a)) \\ &\leq \sum_{i\geq 0} \log\left(\frac{K^3U(s,a) + N^{k_i+1}}{N^{k_i+1}}\right) \\ &\leq |\mathcal{B}_2(s,a)| \log(K^3U(s,a)/N^1(s,a)) \\ &\leq O(\text{polylog}(SAK)\iota). \end{aligned}$$

The proof is completed by taking sum over SA. ■

Now we use this lemma to bound M_1. We have that

$$\begin{aligned} M_1 &\leq 10 \sum_{k=1}^{K} \sum_{h=1}^{H} \left(\sqrt{\frac{S \sum_{s'} P_{s_h^k,a_h^k,s'} |\check{V}_{h+1}^k(s') - P_{s_h^k,a_h^k} \check{V}_{h+1}^k|^2 \iota}{N^k(s_h^k, a_h^k)}} + \frac{S I_{h+1}^k \iota}{N^k(s_h^k, a_h^k)} \right) \\ &\leq 10 \sqrt{\sum_{k,h} \frac{I_{h+1}^k \iota}{N^k(s_h^k, a_h^k)}} \cdot \sqrt{S \sum_{k,h} \sum_{s'} P_{s_h^k,a_h^k,s'} |\check{V}_{h+1}^k(s') - P_{s_h^k,a_h^k} \check{V}_{h+1}^k|^2} + 10S \sum_{k,h} \frac{I_{h+1}^k \iota}{N^k(s_h^k, a_h^k)} \\ &\leq 10\sqrt{S^2AL\iota} \cdot \sqrt{\sum_{k,h} \sum_{s'} P_{s_h^k,a_h^k,s'} |\check{V}_{h+1}^k(s') - P_{s_h^k,a_h^k} \check{V}_{h+1}^k|^2} + 40S^2AL\iota, \end{aligned} \tag{48}$$

where the last line is by the fact that $I_{h+1}^k \leq I_h^k$ and $\sum_{k,h} \frac{1}{N^k(s_h^k,a_h^k)} \leq SAL$.

Let

$$M_4 = \sum_{k,h} \sum_{s'} P_{s_h^k,a_h^k,s'} |\check{V}_{h+1}^k(s') - P_{s_h^k,a_h^k} \check{V}_{h+1}^k|^2. \tag{49}$$

By (48) we have that

$$M_1 \leq 10\sqrt{S^2ALM_4\iota} + 40S^2AL\iota \leq \frac{1}{4}M_4 + 140S^2AL\iota. \tag{50}$$

We continue with bounding M_4.

$$
\begin{aligned}
&M_4 \\
&= \sum_{k,h} \sum_{s'} P_{s_h^k,a_h^k,s'} |\check{V}_{h+1}^k(s') - P_{s_h^k,a_h^k} \check{V}_{h+1}^k|^2 \\
&= \sum_{k,h} \left(P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - (P_{s_h^k,a_h^k} \check{V}_{h+1}^k)^2 \right) \\
&= \sum_{k,h} (P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - \check{V}_h^k(s_h^k)^2) + \sum_{k,h} \left((\check{V}_h^k(s_h^k))^2 - (P_{s_h^k,a_h^k} \check{V}_{h+1}^k)^2 \right) \\
&\le \sum_{k,h} (P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - \check{V}_h^k(s_h^k)^2) + 2 \sum_{k,h} \max\{\check{V}_h^k(s_h^k) - P_{s_h^k,a_h^k} \check{V}_{h+1}^k, 0\} \\
&\le \sum_{k,h} (P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - \check{V}_{h+1}^k(s_{h+1}^k)^2) \\
&\quad + 2 \sum_{k,h} \left(r^k(s_h^k, a_h^k) I_h^k + (I_h^k - I_{h+1}^k) + 10 \sum_{s'} \left(\sqrt{\frac{P_{s_h^k,a_h^k,s'} \iota}{N^k(s_h^k, a_h^k)}} + \frac{\iota}{N^k(s_h^k, a_h^k)} \right) |\check{V}_{h+1}^k(s') - P_{s_h^k,a_h^k} \check{V}_{h+1}^k| \right) \\
&\le \sum_{k,h} (P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - \check{V}_{h+1}^k(s_{h+1}^k)^2) + 2 \sum_{k,h} (I_h^k - I_{h+1}^k) + 2 \sum_{k,h} r(s_h^k, a_h^k) + 2M_1 \\
&\le \sum_{k,h} (P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - \check{V}_{h+1}^k(s_{h+1}^k)^2) + 2 \sum_{k,h} r(s_h^k, a_h^k) + 2(M_1 + M_3) \\
&\le 2 \sum_{k,h} (P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - \check{V}_{h+1}^k(s_{h+1}^k)^2) + 4K + 20\sqrt{S^2AKL\iota} + 60S^2AL\iota + 4U, \qquad (51)
\end{aligned}
$$

where $U = O\left(\frac{S^8A^3K\iota}{K_1} \text{polylog}(SAK)\right)$ is an upper bound of M_3. Here (51) is by (50) and rearrangement.

Let $M_5 = \sum_{k,h} (P_{s_h^k,a_h^k} (\check{V}_{h+1}^k)^2 - \check{V}_{h+1}^k(s_{h+1}^k)^2)$ and $M_6 = \sum_{k,h} \mathbb{V}(P_{s_h^k,a_h^k}, (\check{V}_{h+1}^k)^2)$. By Lemma 9, we have that

$$
M_6 \le 4 \sum_{k,h} \mathbb{V}(P_{s_h^k,a_h^k}, \check{V}_{h+1}^k) = 4M_4. \qquad (52)
$$

By Lemma 14, we have that

$$
\mathbb{P}\left[\exists i, M_5 \ge 10 \cdot 2^i \iota, M_6 \le 2^{2i} \iota\right] \le \delta, \qquad (53)
$$

which implies that

$$
\mathbb{P}\left[\exists i, M_5 \ge 10 \cdot 2^i \iota, M_4 \le 2^{2i-2} \iota\right] \le \delta. \qquad (54)
$$

Therefore, with probability $1 - \delta$, for any $i \ge 1$, it either holds $M_5 < 10 \cdot 2^i \iota$ or $M_4 > 2^{2i-2} \iota$. By (51), we have that

$$
M_4 \le M_5 + 4K + 40\sqrt{S^2AKL\iota} + 60S^2AL\iota + 4U \le M_5 + 8K + 100S^2AL\iota + 4U. \qquad (55)
$$

Suppose $M_5 \geq C \geq 8K + 100S^2AL\iota + 4U$, then we have that

$$M_4 \geq \frac{M_6^2}{800\iota \log_2(C)} \geq \frac{C^2}{800\iota \log_2(C)} \geq 3C. \tag{56}$$

By (55), we have that $M_5 \geq M_4 - (8K + 100S^2AL\iota + 4U) \geq M_4 - C \geq 2C$. In this way, M_5 is infinite, which leads to contradiction. Therefore, with probability $1-\delta$, $M_5 < 8K + 100S^2AL\iota + 4U$, and it follows that

$$M_4 \leq 16K + 100S^2AL\iota + 4U = O\left(K + \frac{S^8A^3K\iota}{K_1}\text{polylog}(SAK)\right). \tag{57}$$

Next, we bound M_2. Using Lemma 12, we have that

$$\begin{aligned}
&\mathbb{P}\left[M_2 \geq 10\sqrt{16K + 200S^2AL\iota + 4U},\right] \\
&\leq \mathbb{P}\left[M_2 \geq 10\sqrt{16K + 200S^2AL\iota + 4U}, M_4 \leq 16K + 200S^2AL\iota + 4U\right] + \delta \\
&\leq 2\delta.
\end{aligned}$$

Finally, putting all together, with probability $1 - 6SA\delta$,

$$\sum_{k=1}^{K}\left(\max_{\pi} W_H^{\pi}(r, P, \mu_1) - \sum_{h=1}^{H} r(s_h^k, a_h^k)\right) \leq O\left(\text{polylog}(SAK)\left(\sqrt{S^2A\iota^2} + \frac{S^8A^3K\iota}{K_1}\right)\right).$$

The proof is completed.

Appendix F. Missing Proofs about Collecting Initial Samples

F.1. Approximated Reference Model

Define $\iota = \log(1/\delta)$. Define $\bar{[}\mathcal{S}\} = \mathcal{S} \cup \{z, z'\}$. Define $\mathcal{G}_1$ be the event where

$$|P_{s,a,s'} - \frac{N^k(s,a,s')}{N^k(s,a)}| \leq \sqrt{\frac{2P_{s,a,s'}\iota}{N^k(s,a)}} + \frac{\iota}{3N^k(s,a)} \tag{58}$$

holds for any proper $k \in [K], (s,a) \in \mathcal{S} \times \mathcal{A}$. By Bennet's inequality (see Lemma 13), we have that $\mathbb{P}[\mathcal{G}_1] \geq 1 - S^2AK\delta$. We continue the analysis assuming $\mathcal{G}_1$ holds.

Lemma 27 *Let k be fixed. With probability $1 - S^2A\delta$, $e^{-1/S}P^{\text{ref},k}_{s,a,s'} \leq \bar{P}^{\text{cut},k}_{s,a,s'} \leq e^{1/S}P^{\text{ref},k}_{s,a,s'}$ for any $(s,a,s') \in \bar{\mathcal{S}} \times \mathcal{A} \times \bar{\mathcal{S}}$.*

Proof For each $(s,a) \in \mathcal{U}^k$, $P^{\text{ref},k}_{s,a,z} = \bar{P}^{\text{cut},k}_{s,a,z} = 1$. For any $(s,a,s') \notin \mathcal{K}^k$, $P^{\text{ref},k}_{s,a,s'} = \bar{P}^{\text{cut},k}_{s,a,s'} = 0$.

For $s = z, z'$, we have that $P^{\text{ref},k}_{z,a} = \bar{P}^{\text{cut},k}_{z,a} =^{\text{ref},k}_{z',a} = \bar{P}^{\text{cut},k}_{z',a} = \mathbf{1}_{z'}$ for any a.

For $(s,a,s') \in \mathcal{K}^k$, by the definition of $\mathcal{G}$, and noting that $n^k(s,a,s') \geq 256S^2\iota$, $1/S \leq 1/2$, we have that

$$\begin{aligned}
&|N^k(s,a)P_{s,a,s'} - N^k(s,a,s')| \\
&\leq \sqrt{2P_{s,a,s'}N^k(s,a)\iota} + \frac{\iota}{3} \\
&\leq \frac{1}{4S}N^k(s,a)P_{s,a,s'} + 8S\iota + \frac{\iota}{3} \\
&\leq \frac{1}{4S}N^k(s,a)P_{s,a,s'} + \frac{1}{4S}N^k(s,a,s'),
\end{aligned}$$

which implies that

$$N^k(s,a)e^{-\frac{1}{2S}}P_{s,a,s'} \leq N^k(s,a,s') \leq N^k(s,a)e^{\frac{1}{2S}}P_{s,a,s'}.$$

Taking sum over s' such that $(s,a,s') \in \mathcal{K}$, we have that

$$N^k(s,a)e^{-\frac{1}{2S}}\sum_{s':(s,a,s')\in\mathcal{K}} P_{s,a,s'} \leq \sum_{(s,a,s')\in\mathcal{K}} n^k(s,a,s') \leq N^k(s,a)e^{\frac{1}{2S}}\sum_{(s,a,s')\in\mathcal{K}} P_{s,a,s'}.$$

Therefore, it holds that

$$\frac{N^k(s,a)e^{-\frac{1}{2S}}P_{s,a,s'}}{N^k(s,a)e^{\frac{1}{2S}}\sum_{(s,a,s')\in\mathcal{K}}P_{s,a,s'}} \leq \frac{N^k(s,a,s')}{\sum_{(s,a,s')\in\mathcal{K}}N^k(s,a,s')} \leq \frac{N^k(s,a)e^{\frac{1}{2S}}P_{s,a,s'}}{N^k(s,a)e^{-\frac{1}{2S}}\sum_{(s,a,s')\in\mathcal{K}}P_{s,a,s'}}.$$

The proof is completed. ■

F.1.1. PROOF OF LEMMA 7

lemma[Restatement of Lemma 7] *With probability* $1 - 10S^3A^2K\delta$, *it holds that*

$$N^{\tilde{k}+1}(\tilde{s},\tilde{a}) \geq 2\max_{\pi\in\Pi_{\mathrm{sta}}} W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}}, P, \mathbf{1}_{\tilde{s}})\log(1/\delta) \tag{59}$$

for any $(\tilde{s},\tilde{a}) \in \mathcal{O}^{\tilde{k}+1}/\mathcal{O}^{\tilde{k}}$ *and any* $1 \leq \tilde{k} \leq K_1$.

Proof

Let $\mathrm{Trigger}^k$ denote the value of Trigger in the end of the k-th round for $k \in [K_1]$. Fix $\tilde{k}$ and $(\tilde{s},\tilde{a}) \in \mathcal{O}^{\tilde{k}+1}/\mathcal{O}^{\tilde{k}}$.

Recall that $\mathcal{J} := \{k \in [K_1] | \exists (h,a), (s^k_{h+1}, a) \in \mathcal{O}^k\}$. Define $\mathcal{C} := \{k \in \mathcal{J}, k \leq \tilde{k} | \mathrm{Trigger}^k = \mathrm{FALSE}, (s^{*,k}_1, a^{*,k}_1) = (\tilde{s},\tilde{a})\}$.

By Algorithm 1, we have that $|\mathcal{C}| \geq 400\log(1/\delta)$. Let $k \in \mathcal{C}$ be fixed. We first show that

$$\max_{\pi\in\Pi_{\mathrm{sta}},\pi(\tilde{s})=\tilde{a}} W^{\pi}_{d_2}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_{\tilde{s}}) \leq \frac{1}{10}. \tag{60}$$

Recall $\gamma = 1 - \frac{1}{d_2}$. By Lemma 18

$$u^k(s) = \max_{\pi\in\Pi_{\mathrm{sta}}} X^{\pi}_{\gamma}(\{s\}, P^{\mathrm{ref},k}, \mathbf{1}_{\tilde{s}}) \geq \frac{1}{3}\max_{\pi\in\Pi_{\mathrm{sta}},\pi(\tilde{a})=\tilde{a}} X^{\pi}_{d_2}(\{s\}, P^{\mathrm{ref},k}, \mathbf{1}_{\tilde{s}}) \tag{61}$$

$$v^k(s,a) = \max_{\pi\in\Pi_{\mathrm{sta}}} W^{\pi}_{\gamma}(\mathbf{1}_{s,a}, P^{\mathrm{ref},k}, \mathbf{1}_s) \geq \frac{1}{3}\max_{\pi\in\Pi_{\mathrm{sta}},\pi(\tilde{s})=\tilde{a}} W^{\pi}_{d_2}(\mathbf{1}_{s,a}, P^{\mathrm{ref},k}, \mathbf{1}_s). \tag{62}$$

For each $s \in \mathcal{S}$, if there exists a and s' such that $(s,a,s') \notin \mathcal{K}^k$, then we either have $u^k(s) < \frac{1}{1200S}$ or $N^k(s,a) > 100S^2AN_0u^k(s)v^k(s,a)$. Denote $\mathcal{S}_1^k := \{s : u^k(s) \le \frac{1}{1200S}\}$. In the first case, we have that

$$\begin{aligned}
\max_{\pi\in\Pi_{\text{sta}},\pi(\tilde{s})=\tilde{a}} X^\pi_{d_2}(\{s\},\bar{P}^k,\mathbf{1}_{\tilde{s}}) &\le \max_{\pi\in\Pi_{\text{sta}},\pi(\tilde{s})=\tilde{a}} X^\pi_{d_2}(\{s\},\bar{P}^{\text{cut},k},\mathbf{1}_{\tilde{s}}) && (63)\\
&\le 3\max_{\pi\in\Pi_{\text{sta}},\pi(\tilde{s})=\tilde{a}} X^\pi_{d_2}(\{s\},P^{\text{ref},k},\mathbf{1}_{\tilde{s}}) && (64)\\
&\le 12u^k(s) && (65)\\
&\le \frac{1}{100S}.
\end{aligned}$$

Here (63) holds by Lemma 31 and the fact that $\bar{P}^{\text{cut},k} = \text{cut}(\bar{P}^k)$, (64) holds by Lemma 27 and Lemma 5.

In the second case, by definition of $\mathcal{G}$, we have that

$$\bar{P}^k_{s,a,z} = \sum_{s':(s,a,s')\notin\mathcal{K}} P_{s,a,s'} \le \frac{2SN_0}{N^k(s,a)} \le \frac{1}{810SAu^k(s,a)v^k(s,a)}. \tag{66}$$

For any stationary policy π such that $\pi(\tilde{s}) = \tilde{a}$, noting that z is a transient state, the probability of reaching z from some state s is bounded by the probability of reaching s. That is, for any state s, it holds that $\sum_a W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_{\tilde{s}})\bar{P}^k_{s,a,z} \le X^\pi_{d_2}(\{s\},\bar{P}^k,\mathbf{1}_{\tilde{s}})$. As a result, we have that

$$\begin{aligned}
W^\pi_{d_2}(\mathbf{1}_z,\bar{P}^k,\mathbf{1}_{\tilde{s}}) &= \sum_{s,a} W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_{\tilde{s}})\bar{P}^k_{s,a,z}\\
&= \sum_{s\in\mathcal{S}_1^k}\sum_a W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_{\tilde{s}})\bar{P}^k_{s,a,z} + \sum_{s\notin\mathcal{S}_1^k}\sum_a W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_{\tilde{s}})\bar{P}^k_{s,a,z}\\
&\le \sum_{s\in\mathcal{S}_1^k} X^\pi_{d_2}(\{s\},\bar{P}^k,\mathbf{1}_{\tilde{s}}) + \sum_{s\notin\mathcal{S}_1^k}\sum_a W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_{\tilde{s}})\bar{P}^k_{s,a,z}. && (67)
\end{aligned}$$

Continuing the computation, we obtain that

$$\begin{aligned}
&W^\pi_{d_2}(\mathbf{1}_z,\bar{P}^k,\mathbf{1}_{\tilde{s}})\\
&\le 3\sum_{s\in\mathcal{S}_1^k} u^k(s) + \sum_{s\notin\mathcal{S}_1^k}\sum_a W^\pi_{d_2}(\mathbf{1}_{s,a},\check{P}^k,\mathbf{1}_{\tilde{s}})\bar{P}^k_{s,a,z} && (68)\\
&\le \frac{1}{400} + \sum_{s\notin\mathcal{S}_1^k}\sum_a W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_{\tilde{s}})\bar{P}^k_{s,a,z}\\
&\le \frac{1}{400} + \sum_{s,a} 9u^k(s)\cdot v^k(s,a)\cdot\frac{1}{810SAu^k(s)v^k(s,a)} && (69)\\
&\le \frac{1}{10}.
\end{aligned}$$

Here (68) is by (67) and (62), and (69) holds because (66) and the fact below

$$\begin{aligned}
&W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_{\tilde{s}})\\
&\le \max_{\pi\in\Pi_{\text{sta}},,\pi(\tilde{s})=\tilde{a}} X^\pi_{d_2}(\{s\},\bar{P}^k,\mathbf{1}_{\tilde{s}})\cdot\max_{\pi\in\Pi_{\text{sta}},\pi(\tilde{s})=\tilde{a}} W^\pi_{d_2}(\mathbf{1}_{s,a},\bar{P}^k,\mathbf{1}_s)\\
&\le 9u^k(s)\cdot 9v^k(s,a).
\end{aligned}$$

Define E_1^k to be the event z is visited under $\bar{P}^k$. So E_1^k is corresponding to visiting $(\mathcal{K}^k)^C$ under P. Let be X^k be the count of $(\tilde{s},\tilde{a})$ in the first d_2 steps, i.e., $X^k = \sum_{i=1}^{d} \mathbb{I}[(s_i,a_i)=(\tilde{s},\tilde{a})]$. Then we have that for any stationary policy π such that $\pi(\tilde{s})=\tilde{a}$,

$$W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^k,\mathbf{1}_{\tilde{s}}) = \mathbb{E}_{\bar{P}^k,\pi}[X^k|\mu_1=\mathbf{1}_{\tilde{s}}] \geq \mathbb{E}_{\bar{P}^k,\pi}[X^k\mathbb{I}[(E_1^k)^C]|\mu_1=\mathbf{1}_{\tilde{s}}]. \tag{70}$$

By definition of $\bar{P}^k$ and E_1^k, we obtain that

$$\mathbb{E}_{\bar{P}^k,\pi}[X^k\mathbb{I}[(E_1^k)^C]|\mu_1=\mathbf{1}_{\tilde{s}}] = \mathbb{E}_{P,\pi}[X^k\mathbb{I}[(E_1^k)^C]|\mu_1=\mathbf{1}_{\tilde{s}}], \tag{71}$$

which implies that $W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^k,\mathbf{1}_{\tilde{s}}) \geq \mathbb{E}_{P,\pi}[X^k\mathbb{I}[(E_1^k)^C]|\mu_1=\mathbf{1}_{\tilde{s}}]$.

On the other hand, by Lemma 20, we have that

$$\Pr\left[X^k \geq \frac{1}{2}\mathbb{E}_{P,\pi}[X^k|\mu_1=\mathbf{1}_{\tilde{s}}] \text{ and } (E_1^k)^C\right] \geq \frac{1}{2}-\frac{1}{10}. \tag{72}$$

As a result, it holds that

$$\begin{aligned}
&W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^k,\mathbf{1}_{\tilde{s}})\\
\geq&\mathbb{E}_{\pi,P}[X^k\mathbb{I}[(E_1^k)^C]|\mu_1=\mathbf{1}_{\tilde{s}}]\\
\geq&\frac{1}{2}(1-\frac{1}{10})\mathbb{E}_{\pi,P}[X^k|\mu_1=\mathbf{1}_{\tilde{s}}]\\
=&\frac{9}{20}W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},P,\mathbf{1}_{\tilde{s}})
\end{aligned} \tag{73}$$

By (60) and Lemma 30, we have that for any stationary policy π such that $\pi(\tilde{s})=\tilde{a}$,

$$W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^k,\mathbf{1}_{\tilde{s}}) \leq W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^{\mathrm{cut},k},\mathbf{1}_{\tilde{s}}) \leq 2W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^k,\mathbf{1}_{\tilde{s}}). \tag{74}$$

By Lemma 27 and 5, we further have that

$$W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^k,\mathbf{1}_{\tilde{s}}) \leq 3W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},P^{\mathrm{ref},k},\mathbf{1}_{\tilde{s}}) \leq 18W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},\bar{P}^k,\mathbf{1}_{\tilde{s}}). \tag{75}$$

Combining (73) with (75), we learn that

$$W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},P,\mathbf{1}_{\tilde{s}}) \leq 9W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},P^{\mathrm{ref},k},\mathbf{1}_{\tilde{s}}) \leq 54W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},P,\mathbf{1}_{\tilde{s}}) \tag{76}$$

for any stationary policy π such that $\pi(\tilde{s})=\tilde{a}$.

Recall that $\gamma = 1-\frac{1}{d_2}$. By running the policy

$$\pi_2^k := \arg\max_{\pi\in\Pi_{\mathrm{sta}},\pi(\tilde{s})=\tilde{a}} W^{\pi}_{\gamma}(\mathbf{1}_{\tilde{s},\tilde{a}},P^{\mathrm{ref},k},\mathbf{1}_{\tilde{s}}),$$

using (72) and (76), with probability $1/2$,

$$\begin{aligned}
X^k \geq& \frac{1}{4}W^{\pi_2^k}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},P^{\mathrm{ref},k},\mathbf{1}_{\tilde{s}})\\
\geq& \frac{1}{12}W^{\pi_2}_{\gamma}(\mathbf{1}_{\tilde{s},\tilde{a}},P^{\mathrm{ref},k},\mathbf{1}_{\tilde{s}}) &(77)\\
=& \frac{1}{12}\max_{\pi\in\Pi_{\mathrm{sta}},\pi(\tilde{s})=\tilde{a}}\frac{1}{12}W^{\pi}_{\gamma}(\mathbf{1}_{\tilde{s},\tilde{a}},P^{\mathrm{ref},k},\mathbf{1}_{\tilde{s}})\\
\geq& \frac{1}{36}\max_{\pi\in\Pi_{\mathrm{sta}},\pi(\tilde{s})=\tilde{a}}\frac{1}{12}W^{\pi}_{\gamma}(\mathbf{1}_{\tilde{s},\tilde{a}},P^{\mathrm{ref},k},\mathbf{1}_{\tilde{s}}) &(78)\\
\geq& \frac{1}{108}\max_{\pi\in\Pi_{\mathrm{sta}},\pi(\tilde{s})=\tilde{a}}W^{\pi}_{d_2}(\mathbf{1}_{\tilde{s},\tilde{a}},P,\mathbf{1}_{\tilde{s}}) &(79)
\end{aligned}$$

samples of $(\tilde{s}, \tilde{a})$ in the k-th episode. Here (77) and (78) are by Lemma 18, and (79) holds by Lemma 27 and Lemma 5.

By Lemma 10, with probability $1-\delta$ it holds that

$$\begin{aligned} N^{\tilde{k}+1}(\tilde{s},\tilde{a}) &\geq \sum_{k\in C} X^k \\ &\geq 2 \max_{\pi\in\Pi_{\text{sta}},\pi(s)=a} W^{\pi}_{d_2}(\mathbf{1}_{s,a}, P, \mathbf{1}_s)\log(1/\delta) \\ &= 2 \max_{\pi\in\Pi_{\text{sta}}} W^{\pi}_{d_2}(\mathbf{1}_{s,a}, P, \mathbf{1}_s)\log(1/\delta). \end{aligned} \tag{80}$$

The proof is completed. ■

F.1.2. STATEMENT AND PROOF OF LEMMA 28

Lemma 28 *With probability* $1-3S^3A^2K\delta$, *it holds that* $|\mathcal{J}| \leq O(\text{polylog}(SAK)S^7A^3\log(1/\delta))$.

Proof For fixed (s,a), we define

$$\text{True}(s,a) = \{k \in \mathcal{J} | \text{ Trigger}^k \text{ is set to be True with respect to } (s,a)\}.$$

Now we analyze the size of $\text{True}(s,a)$. Let $\mathcal{I} = \{k : \mathcal{K}^k \neq \mathcal{K}^{k-1}\}$. Since $\mathcal{K}^1 \subset \mathcal{K}^2 \subset \ldots \mathcal{K}^k \subset \ldots$, and the $|\mathcal{K}^k| \leq S^2A$ for any k, we have that $|\mathcal{I}| \leq S^2A$. Suppose $\mathcal{I} = \{k_1, k_2, \ldots, k_{|\mathcal{I}|}\}$.

Then we have that

$$|\text{True}(s,a)| = \sum_{i=1}^{|\mathcal{I}|} |\text{True}(s,a) \cap [k_{i-1}, k_i - 1]|, \tag{81}$$

where k_0 is defined as 1.

Then we have the lemma to bound $|\text{True}(s,a)|$.

Lemma 29 *For any* $1 \leq i \leq |\mathcal{I}|$, *with probability* $1-3S^3A^2\delta$, *it holds that*

$$\sum_{(s,a)} \sum_{1\leq i\leq|\mathcal{I}|} |\text{True}(s,a) \cap [k_{i-1}, k_i - 1]| \leq 480S(9600S^4A^3N_0 + 5S^3A^2\log(1/\delta) + 2S^2AN_0).$$

By Lemma 29, we obtain that

$$\begin{aligned} |\mathcal{J}| &\leq \sum_{1\leq i\leq|\mathcal{I}|} \sum_{s,a} |\text{True}(s,a) \cap [k_{i-1}, k_i - 1]| + \sum_{k\in[\mathcal{J}]} \mathbb{I}\left[\text{Trigger}^k = \text{FALSE}\right] \\ &= O(\text{polylog}(SAK)S^7A^3\log(1/\delta)). \end{aligned}$$

The proof is completed. ■

Proof [Proof of Lemma 29] Let i and (s,a) be fixed. Let $\Lambda_{1,i}(s,a) = \text{True}(s,a) \cap [k_{i-1}, k_i - 1]$. Let $\mathcal{K} = \mathcal{K}^k$ for some $k \in \Lambda_{1,i}(s,a)$. The definition is proper since $\mathcal{K}^k$ is the same for any $k \in \Lambda_{1,i}(s,a)$.

Recall that $\gamma = 1 - \frac{1}{d_2}$ and

$$u^k(s) = \max_{\pi \in \Pi_{\text{sta}}} X_\gamma^\pi(\{s\}, P^{\text{ref},k}, \mathbf{1}_{\tilde{s}}) \geq \frac{1}{3} \max_{\pi \in \Pi_{\text{sta},\pi(s_1^{*,k})=a_1^{*,k}}} X_{d_2}^\pi(\{s\}, P^{\text{ref},k}, \mathbf{1}_{s_1^{*,k}}) \tag{82}$$

$$v^k(s,a) = \max_{\pi \in \Pi_{\text{sta}}} W_\gamma^\pi(\mathbf{1}_{s,a}, P^{\text{ref},k}, \mathbf{1}_s) \geq \frac{1}{3} \max_{\pi \in \Pi_{\text{sta},\pi(s_1^{*,k})=a_1^{*,k}}} W_{d_2}^\pi(\mathbf{1}_{s,a}, P^{\text{ref},k}, \mathbf{1}_s). \tag{83}$$

Recall that $\pi_1^k, \pi_2^k \in \Pi_{\text{sta}}$ are such that $\pi_1^k(s_1^{*,k}) = \pi_2^k(s_1^{*,k}) = a_1^{*,k}$ and

$$\begin{aligned} X_\gamma^{\pi_1^k}(\{s\}, P^{\text{ref},k}, \mathbf{1}_{s_1^{*,k}}) &= u^k(s) \\ W_\gamma^{\pi_2^k}(\mathbf{1}_{s,a}, P^{\text{ref},k}, \mathbf{1}_s) &= v^k(s,a). \end{aligned} \tag{84}$$

By Lemma 18, and recalling $d_1 = d - d_2 \geq 10S\log(S)d_2$, we also have that

$$X_{d_1}^{\pi_1^k}(\{s\}, P^{\text{ref},k}, \mathbf{1}_{s_1^{*,k}}) \geq \frac{1}{10} X_\gamma^{\pi_1^k}(\{s\}, P^{\text{ref},k}, \mathbf{1}_{s_1^{*,k}}) = \frac{1}{10} u^k(s) \tag{85}$$

$$W_{d_2}^{\pi_2^k}(\mathbf{1}_{s,a}, P^{\text{ref},k}, \mathbf{1}_s) \geq \frac{1}{3} v^k(s,a). \tag{86}$$

By definition, for any $k \in \Lambda_{1,i}(s,a)$, we have that $N^k(s,a) \leq 300SAN_0 u^k(s)v^k(s,a)$ and $u^k(s) \geq \frac{1}{1200S}$. By Lemma 5 we have that

$$X_{d_1}^{\pi_1^k}(\{s\}, \bar{P}^{\text{cut},k}, \mathbf{1}_{s_1^{*,k}}) \geq \frac{1}{3} X_{d_1}^{\pi_1^k}(\{s\}, P^{\text{ref},k}, \mathbf{1}_{s_1^{*,k}}) \geq \frac{1}{30} u^k(s) \tag{87}$$

By Lemma 31, Lemma 27 and Lemma 5, we further have that

$$\begin{aligned} &X_{d_1}^{\pi_1^k}(\{s\}, \bar{P}^k, \mathbf{1}_{s_1^{*,k}}) \\ &\geq X_{d_1}^{\pi_1^k}(\{s\}, \bar{P}^{\text{cut},k}, \mathbf{1}_{s_1^{*,k}}) - W_{d_1}^{\pi_1^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_{s_1^{*,k}}) \\ &\geq \frac{1}{30} u^k(s) - W_{d_1}^{\pi_1^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_{s_1^{*,k}}). \end{aligned}$$

By rearranging the inequality, we have that

$$X_{d_1}^{\pi_1^k}(\{s\}, \bar{P}^k, \mathbf{1}_{s_1^{*,k}}) + W_{d_1}^{\pi_1^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_{s_1^{*,k}}) \geq X_{d_1}^{\pi_1^k}(\{s\}, \bar{P}^{\text{cut},k}, \mathbf{1}_{s_1^{*,k}}) \geq \frac{1}{30} u^k(s) \geq \frac{1}{300S}. \tag{88}$$

Let E_2^k be reaching s without visiting $\mathcal{K}^C$ in the k-th episode and E_3^k be reaching $\mathcal{K}^C$ in the first d steps in the k-th episode. Then we have that

$$\Pr_{P,\pi_1^k}[E_2^k \cup E_3^k] = X_{d-1}^{\pi_1^k}(\{s\}, \bar{P}^k, \mathbf{1}_{s_1^{*,k}}) + W_{d_1}^{\pi_1^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_{s_1^{*,k}}) \tag{89}$$

$$\geq X_{d_1}^{\pi_1^k}(\{s\}, \bar{P}^{\text{cut},k}, \mathbf{1}_{s_1^{*,k}}) \geq \frac{1}{30} u^k(s) \geq \frac{1}{3600S}. \tag{90}$$

Therefore, if $|\Lambda_{1,i}(s,a)| \geq 28800S\log(1/\delta)$, by Lemma 10, with probability $1-\delta$, it holds that

$$\sum_{k \in \Lambda_{1,i}(s,a)} \mathbb{I}[E_2^k] + \mathbb{I}[E_3^k] \geq \frac{1}{14400S} |\Lambda_{1,i}(s,a)| - \log(1/\delta). \tag{91}$$

Define $\Lambda_{2,i}(s,a) = \{k \in \Lambda_{1,i}(s,a) : \mathbb{I}[E_2^k] = 1\}$. Let $k \in \Lambda_{2,i}(s,a)$ be fixed. Also recall that $\pi_2^k(s) = a$ and $W_\gamma^{\pi_2^k}(\mathbf{1}_{s,a}, P^{\text{ref},k}, \mathbf{1}_s) = v^k(s,a)$. By Lemma 5 and (86), we have that

$$W_{d_2}^{\pi_2^k}(\mathbf{1}_{s,a}, \bar{P}^{\text{cut},k}, \mathbf{1}_s) \geq \frac{1}{9} v^k(s,a).$$

By Lemma 30, we have that

$$\begin{aligned} W_{d_2}^{\pi_2^k}(\mathbf{1}_{s,a}, \bar{P}^k, \mathbf{1}_s) \geq & (1 - W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_s)) \cdot W_{d_2}^{\pi_2^k}(\mathbf{1}_{s,a}, \bar{P}^{\text{cut},k}, , \mathbf{1}_s) & (92) \\ \geq & \frac{1}{3}(1 - W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_s)) \cdot v^k(s,a). & (93) \end{aligned}$$

Let Z^k be the number of samples of (s,a) collected in the d_2 steps following π_2^k. Noting that $\mathbb{E}[Z_k] = W_{d_2}^{\pi_2^k}(\mathbf{1}_{s,a}, P, \mathbf{1}_s) \geq W_d^{\pi_2^k}(\mathbf{1}_{s,a}, \bar{P}^k, \mathbf{1}_s)$, by Lemma 20, we have that

$$\Pr\left[Z^k \geq \frac{1}{12}(1 - W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_s)) \cdot v^k(s,a)\right] \geq \frac{1}{2}. \tag{94}$$

Note that for any $k \in \Lambda_{1,i}(s,a)$, $\bar{P}^{\text{cut},k}$ does not vary in k. By Lemma 27 and Lemma 5, we learn that $\min_{k \in \Lambda_{1,i}(s,a)} v^k(s,a) \geq \frac{1}{9} \max_{k \in \Lambda_{1,i}(s,a)} v^k(s,a)$. As a result, by (94), we have that

$$\Pr\left[Z^k \geq \frac{1}{96}(1 - W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_s)) \cdot \max_{k' \in \Lambda_{1,i}(s,a)} v^{k'}(s,a)\right] \geq \frac{1}{2}. \tag{95}$$

Let $k_{\max} = \max_{k' \in \Lambda_{2,i}(s,a)} k'$. By Lemma 10, with probability $1 - \delta$ it holds that

$$\sum_{k \in \Lambda_{2,i}(s,a), k < k_{\max}} \mathbb{I}\left[Z^k \geq \frac{1}{96}(1 - W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}^k, \mathbf{1}_s)) \cdot \max_{k' \in \Lambda_{1,i}(s,a)} v^{k'}(s,a)\right] \geq \frac{|\Lambda_{2,i}(s,a) - 1|}{4} - \log(1/\delta),$$

where it follows that

$$\sum_{k \in \Lambda_{2,i}(s,a), k < k_{\max}} Z^k \geq \left(\frac{|\Lambda_{2,i}(s,a) - 1|}{392} - \frac{1}{96}\log(1/\delta) - \frac{1}{96} \sum_{k \in \Lambda_{2,i}(s,a)} W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}_k, \mathbf{1}_s)\right) \max_{k \in \Lambda_{1,i}(s,a)} v^k(s,a). \tag{96}$$

Let E_4^k be the event E_2^k occurs, and then the agent reaches $\mathcal{K}^C$ in the following d steps under π_2^k. Note that $W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}_k, \mathbf{1}_s) = \mathbb{E}[E_4^k]$. By Lemma 10, with probability $1 - \delta$ it holds that

$$\sum_{k \in \Lambda_{2,i}(s,a), k < k_{\max}} W_{d_2}^{\pi_2^k}(\mathbf{1}_z, \bar{P}_k, \mathbf{1}_s) \leq 2 \sum_{k \in \Lambda_{2,i}(s,a), k < k_{\max}} \mathbb{I}[E_4^k] + 4\log(1/\delta). \tag{97}$$

By (91), (96) and (97), with probability $1-3\delta$, it holds that

$$\begin{aligned}
&\sum_{k\in\Lambda_{2,i}(s,a),k<k_{\max}} Z^k \\
&\geq \frac{1}{96}\left(\frac{|\Lambda_{2,i}(s,a)-1|}{4} - 5\log(1/\delta) - 2\sum_{k\in\Lambda_{2,i}(s,a)}\mathbb{I}[E_4^k]\right)\max_{k\in\Lambda_{1,i}(s,a)} v^k(s,a) \\
&\geq \frac{1}{96}\left(\frac{1}{4}\left(\frac{1}{14400S}|\Lambda_{1,i}(s,a)| - \log(1/\delta) - \sum_{k\in\Lambda_{1,i}(s,a)}\mathbb{I}[E_3^k]\right) - 6\log(1/\delta) - 2\sum_{k\in\Lambda_{2,i}(s,a)}\mathbb{I}[E_4^k]\right) \\
&\quad\cdot \max_{k\in\Lambda_{1,i}(s,a)} v^k(s,a) \\
&\geq \frac{1}{96}\left(\frac{1}{57600S}|\Lambda_{1,i}(s,a)| - 6\log(1/\delta) - 2\sum_{k\in\Lambda_{1,i}(s,a)}\mathbb{I}[E_3^k\cup E_4^k]\right)\max_{k\in\Lambda_{1,i}(s,a)} v^k(s,a).
\end{aligned}$$

Also note that by definition,

$$\sum_{k'\in\Lambda_{2,i}(s,a),k'<k_{\max}} Z^{k'} \leq N^{k_{\max}}(s,a) \leq 810SAN_0u^k(s)v^k(s,a) \leq 810SAN_0\max_{k\in\Lambda_{1,i}(s,a)} v^k(s,a),$$

we have that

$$\left(\frac{1}{57600S}|\Lambda_{1,i}(s,a)| - 6\log(1/\delta) - 2\sum_{k\in\Lambda_{1,i}(s,a)}\mathbb{I}[E_3^k\cup E_4^k]\right) \leq 810000SAN_0. \tag{98}$$

Taking sum over i, and noting that $\sum_{s,a}\sum_{i,k\in\Lambda_{1,i}(s,a)}\mathbb{I}[E_3^k\cup E_4^k] \leq S^2AN_0$, we learn that

$$\sum_{s,a}\sum_{1\leq i\leq|\mathcal{I}|}|\Lambda_{1,i}(s,a)| \leq 57600S\left(810000S^4A^3N_0 + 6S^3A^2\log(1/\delta) + 2S^2AN_0\right) \tag{99}$$

The proof is completed. ■

F.1.3. PUTTING ALL TOGETHER

Combining Lemma 7 and 29, the proof is completed.

F.2. Other Missing Proofs

Lemma 5 (restated) *Let the initial distribution μ_1 be fixed. For two transition model P' and P'' such that P' is ϵ-closed to P'', it holds that*

$$e^{-4S\epsilon}W_d^\pi(r,P',\mu_1) \leq W_d^\pi(r,P'',\mu_1) \leq e^{4S\epsilon}W_d^\pi(r,P'',\mu_1) \tag{100}$$

for any stationary policy π, horizon $d\geq 1$ and any non-negative reward r.

Proof Let the policy π be fixed. First, we assume P' only differs with P'' at (s^*,a^*), where $a^* = \pi(s^*)$[5]. Moreover, we assume that there are only two possible next states of (s^*,a^*), which are denoted as s_l and s_r.

5. Here we deal with a deterministic policy π. The proof also works for non-deterministic policies.

Let $p'_l = P'_{s^*,a^*,s_l}$, $p'_r = 1 - p'_l$, $p''_l = P''_{s^*,a^*,s_l}$ and $p''_r = 1 - p''_l$. We assume that the agent starts at (s^*, a^*), since the transitions of the two models are exactly the same before visiting (s^*, a^*). For each (s, a, s'), we define

$$w'_d(s,a,s') = \mathbb{E}_{\pi,P'}\left[\sum_{i=1}^{d} \mathbb{I}[(s_i,a_i,s_{i+1}) = (s,a,s')] \mid (s_1,a_1) = (s^*,a^*)\right];$$
$$w''_d(s,a,s') = \mathbb{E}_{\pi,P''}\left[\sum_{i=1}^{d} \mathbb{I}[(s_i,a_i,s_{i+1}) = (s,a,s')] \mid (s_1,a_1) = (s^*,a^*)\right].$$

For fixed h, we define $\kappa'_h(s,a,s') := \min_{1\le d\le h} \frac{w'_d(s,a,s')}{w''_d(s,a,s')}$.

Assuming $p'_r \ge p''_r$, by policy difference lemma (Lemma 32) we then have that for any d, $w'_d(s^*,a^*,s_r) \ge w''_d(s^*,a^*,s_r)$ and $w'_d(s^*,a^*,s_l) \le w''_d(s^*,a^*,s_l)$.

By definition it follows that

$$\kappa'_h(s^*,a^*,s_r) \ge 1$$
$$\kappa'_h(s^*,a^*,s_l) \le 1$$

Let (s,a,s') be fixed. Let $x_{l,d_1,d_2}(x_{r,d_1,d_2})$ be the probability of visiting (s,a,s') at the d_2-th step starting from $s_l(s_r)$ at the d_1-th step without visiting (s^*,a^*) between the d_1-th and d_2-th step. By definition, x_{l,d_1,d_2} only depends on $(d_2 - d_1)$ and we can rewrite x_{l,d_1,d_2} as $x_l(d_2 - d_1)$. Similarly we define $x_r(d_2 - d_1) = x_{r,d_1,d_2}$. Note that $x_l(d')(x_r(d'))$ do not depends on p'_r. Since the initial state-action pair is (s^*,a^*), for any $h' \in [h]$ we have that

$$\begin{aligned}
w'_{h'}(s,a,s') &= \sum_{d_1=1}^{h'}\sum_{d_2=d_1}^{h'} (\mathbb{P}_{P'}[s_{d_1} = s_l]x_{l,d_1,d_2} + \mathbb{P}_{P'}[s_{d_1} = s_r]x_{r,d_1,d_2}) \\
&= \sum_{d_1=1}^{h'}\left(\mathbb{P}_{P'}[s_{d_1} = s_l]\sum_{d_2=d_1}^{h'} x_{l,d_1,d_2} + \mathbb{P}_{P'}[s_{d_1} = s_r]\sum_{d_2=d_1}^{h'} x_{r,d_1,d_2}\right) \\
&= \sum_{d_1=1}^{h'}\left(\mathbb{P}_{P'}[s_{d_1} = s_l]\sum_{d_2=1}^{h'-d_1+1} x_l(d_2) + \mathbb{P}_{P'}[s_{d_1} = s_r]\sum_{d_2=1}^{h'-d_1+1} x_r(d_2)\right) \\
&= \sum_{d_2=1}^{h'} x_{l,1,d_2}\left(\sum_{d_1=1}^{h'-d_2+1}\mathbb{P}_{P'}[s_{d_1} = s_l]\right) + \sum_{d_2=1}^{h'} x_r(d_2)\left(\sum_{d_1=1}^{h'-d_2+1}\mathbb{P}_{P'}[s_{d_1} = s_r]\right) \\
&= \sum_{d_2=1}^{h'} x_l(d_2)w'_{h'-d_2+1}(s^*,a^*,s_l) + \sum_{d_2=1}^{h'} x_r(d_2)w'_{h'-d_2+1}(s^*,a^*,s_r) \\
&\ge \kappa_h(s^*,a^*,s_l)\sum_{d_2=1}^{h'} x_l(d_2)w''_{h'-d_2+1}(s^*,a^*,s_l) + \sum_{d_2=1}^{h'} x_r(d_2)w''_{h'-d_2+1}(s^*,a^*,s_r) \\
&= \kappa_{h'}(s^*,a^*,s_l)w''_{h'}(s,a,s').
\end{aligned}$$

By definition, we then have that

$$\kappa'_h(s^*,a^*,s_l) \le \kappa'_h(s,a,s') \tag{101}$$

Note that for any $d \geq 1$,

$$w'_d(s^*, a^*, s_r)/p'_r = w'_d(s^*, a^*, s_l)/p'_l;$$
$$w''_d(s^*, a^*, s_r)/p''_r = w''_d(s^*, a^*, s_l)/p''_l.$$

We then have that

$$1 \leq \kappa'_h(s^*, a^*, s_r) = \min_{1\leq d\leq h} \frac{w'_d(s^*, a^*, s_r)}{w''_d(s^*, a^*, s_r)} = \frac{p'_r p''_l}{p'_l p''_r} \min_{1\leq d\leq h} \frac{w'_d(s^*, a^*, s_l)}{w''_d(s^*, a^*, s_l)} \leq e^{2\epsilon}\kappa'_h(s^*, a^*, s_l) \leq e^{2\epsilon}. \quad (102)$$

By (101) and (102), we have that $\kappa'_h(s, a, s') \geq \kappa'_h(s^*, a^*, s_r) \geq e^{-2\epsilon}$ for any $h \geq 1$ and any (s, a, s'), which implies that

$$w'_d(s, a, s') \geq e^{-2\epsilon} w''_d(s, a, s')$$

for any $d \geq 1$ and any (s, a, s'). By reversing s_l and s_r, we can obtain that

$$w''_d(s, a, s') \geq e^{-2\epsilon} w'_d(s, a, s')$$

for any $d \geq 1$ and any (s, a, s'). The proof is completed by noting that any reward r is a positive linear combination of $\{\mathbf{1}_{s,a,s'}\}_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$.

As for the general case, we also the case where P' only differs with P'' at (s^*, a^*). Let $p' = P'_{s^*,a^*}$ and $p'' = P''_{s^*,a^*}$. We claim there exists $p_0 = p', p_1, p_2, \ldots, p_S = p''$ satisfying that: $\exists \mathcal{S}^i_1, \mathcal{S}^i_2, \mathcal{S}^i_3$ be a partition of $\mathcal{S}$ such that

$$p_{i,s} = p_{i+1,s}, \forall s \in \mathcal{S}^i_1, \quad p_{i,s} = e^{\epsilon'_i} p_{i+1,s}, \forall s \in \mathcal{S}^i_2, \quad p_{i,s} = e^{-\epsilon''_i} p_{i+1,s}, \forall s \in \mathcal{S}^i_3 \quad (103)$$

for $1 \leq i \leq S-1$ with $\epsilon'_i, \epsilon''_i \geq 0$ and $\sum_{i=1}^{S-1} \max\{\epsilon'_i, \epsilon''_i\} \leq 2\epsilon$. If this claim holds, then the conclusion holds by iteratively using the proof for the case where there are only two possible next states. Now we construct $\{p_i\}$. Give $p_1 = p'$, we define $\mathcal{S}' = \{s : p'_s > p''_s\}$. For $\lambda \in [0, 1]$, define $p_{1,\lambda}$ by setting $p_{1,\lambda,s} = \lambda p_{1,s}$ for $s \in \mathcal{S}'$ and $p_{1,\lambda,s} = \lambda' p_{1,s}$ for $s \in \mathcal{S}'$, where λ' is the unique real such that $\sum_{s'} p_{1,\lambda,s'} = 1$. Let λ_1 be the largest real in $[0, 1]$ such that $\exists s', p_{1,\lambda_1,s'} = p''_{s'}$. We then choose $p_2 = p_{1,\lambda_1}$. For $i \geq 2$, we define $p_{i,\lambda}$ by setting $p_{i,\lambda,s} = p_{i,s}, \forall s$ such that $p_{i,s} = p''_s$, $p_{i,\lambda,s} = \lambda p_{i,s}, \forall s$ such that $p_{i,s} > p''_s$ and $p_{i,\lambda,s} = \lambda' p_{i,s}, \forall s$ such that $p_{i,s} < p''_s$, where λ' is the unique real such that $\sum_{s'} p_{i,\lambda,s'} = 1$. Let λ_i be the largest real such that $\exists, s'$ such that $p_{i,s'} \neq p''_{s'}$ and $p_{i,\lambda,s'} = p''_{s'}$ and λ'_i be the corresponding λ'. Then we set $p_{i+1} = p_{i,\lambda_i}$. It is easy to note that for $s' = \arg\max_s \frac{p'_s}{p''_s}$, $p_{i,s'}$ is increasing in i, therefore, we have that $\Pi_{i=1}^{S-1} \frac{1}{\lambda_i} \leq e^{\epsilon}$. In a similar way, $\Pi_{i=1}^{S-1} \lambda'_i \leq e^{\epsilon}$. The proof is completed. ■

Lemma 30 *For any stationary policy π and any transition model p and any $(s, a) \in \mathcal{S} \times \mathcal{A}$ such that $\pi(s) = a$ and $p_{z,a} = \mathbf{1}_z, \forall a$, it holds that*

$$(1 - W^\pi_d(\mathbf{1}_z, p, \mathbf{1}_s))W^\pi_d(\mathbf{1}_{s,a}, \text{cut}(p), \mathbf{1}_s) \leq W^\pi_d(\mathbf{1}_{s,a}, p, \mathbf{1}_s) \leq W^\pi_d(\mathbf{1}_{s,a}, \text{cut}(p), \mathbf{1}_s).$$

Proof Let $p' = \mathrm{cut}(p)$. By policy difference lemma (Lemma 32), we have that

$$\begin{aligned}
&W_d^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_s) - W_d^\pi(\mathbf{1}_{s,a}, p', \mathbf{1}_s) \\
&= \mathbb{E}_{p,\pi}\left[\sum_{h=1}^{d-1}\sum_{s'}(p_{s_h,a_h,s'} - p'_{s_h,a_h,s'})W_{d-h}^\pi(\mathbf{1}_{s,a}, p', \mathbf{1}_{s'})\right] \\
&\geq -\mathbb{E}_{p,\pi}\left[\sum_{h=1}^{d-1} p_{s_h,a_h,z}\mathbb{I}[s_h \neq z]\right] W_d^\pi(\mathbf{1}_{s,a}, p', \mathbf{1}_s) \qquad (104) \\
&= -W_d^\pi(\mathbf{1}_z, p, \mathbf{1}_s)\cdot W_d^\pi(\mathbf{1}_{s,a}, p', \mathbf{1}_s). \qquad (105)
\end{aligned}$$

Here (104) uses the fact that

$$W_{d-h}^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_{s'}) \leq W_d^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_{s'}) \leq W_d^\pi(\mathbf{1}_{s,a}, p, \mathbf{1}_s).$$

The left part is proven by rearranging (105). As for the right side, it suffices to note that for any (s,a) such that $p_{s,a,z} < 1$,

$$(p_{s_h,a_h} - p'_{s_h,a_h})W_{d-h}^\pi(\mathbf{1}_{s,a}, p', \mathbf{1}_{s_h,a_h}) = -p_{s_h,a_h,z}W_{d-h}^\pi(\mathbf{1}_{s,a}, p', \mathbf{1}_{s_h,a_h}) \leq 0,$$

and for (s,a) such that $p_{s,a,z} = 1$, $(p_{s_h,a_h} - p'_{s_h,a_h})W_{d-h}^\pi(\mathbf{1}_{s,a}, p', \mathbf{1}_{s_h,a_h}) = 0$. The proof is completed. ■

Lemma 31 *Let the initial distribution μ_1 and $s \in \mathcal{S}$ be fixed. Let π be a policy (which is possibly non-stationary). Suppose $p_{z,a} = \mathbf{1}_z$ for any a. Then we have that*

$$X_d^\pi(\{s\}, p, \mu_1) \geq X_d^\pi(\{s\}, \mathrm{cut}(p), \mu_1) - W_d^\pi(\mathbf{1}_z, p, \mu_1).$$

Proof [Proof of Lemma 31] Let $\bar{p}$ be defined as $\bar{p}_{s',a'} = p_{s',a'}$ for $s' \neq s$ and any a, and $\bar{p}_{s,a} = \mathbf{1}_{z_1}$ for any a, where $p_{z_1,a} = \mathbf{1}_{z_2}$ and $p_{z_2,a} = \mathbf{1}_{z_2}$ for any a. We also define $\bar{p}'$ by $\bar{p}' = p'_{s,a}$ for $s' \neq s$ and any a, and $\bar{p}'_{s,a} = \mathbf{1}_{z_1}$, $\bar{p}'_{z_1,a} = \mathbf{1}_{z_1}$ and $\bar{p}'_{z_2,a} = \mathbf{1}_{z_2}$ for any a. Then we have that

$$\begin{aligned}
X_d^\pi(\{s\}, p, \mu_1) &= W_d^\pi(\mathbf{1}_{z_1}, \bar{p}, \mu_1); \\
X_d^\pi(\{s\}, p', \mu_1) &= W_d^\pi(\mathbf{1}_{z_1}, \bar{p}', \mu_1). \qquad (106)
\end{aligned}$$

Using policy difference lemma (Lemma 32), and noting that $0 \leq W_{d'}^\pi(\mathbf{1}_{z_1}, p', \mu_1) \leq 1$ for any $0 \leq d' \leq d+1$, we obtain that

$$\begin{aligned}
W_d^\pi(\mathbf{1}_{z_1}, \bar{p}, \mu_1) &\geq W_d^\pi(\mathbf{1}_{z_1}, \bar{p}', \mu_1) - \mathbb{E}_{\bar{p},\pi}\left[\sum_{h=1}^{d-1} p_{s_h,a_h,z}\mathbb{I}[s_h \neq z]\right] \\
&\geq W_d^\pi(\mathbf{1}_{z_1}, \bar{p}', \mu_1) - W_d^\pi(\mathbf{1}_z, \bar{p}, \mu_1) \\
&\geq W_d^\pi(\mathbf{1}_{z_1}, \bar{p}', \mu_1) - W_d^\pi(\mathbf{1}_z, p, \mu_1), \qquad (107)
\end{aligned}$$

where the last line is by the fact that

$$W_d^\pi(\mathbf{1}_z, p, \mu_1) - W_d^\pi(\mathbf{1}_z, \bar{p}, \mu_1) = \mathbb{E}_{p,\pi}\left[\sum_{h=1}^{d-1}(p_{s_h,a_h} - \bar{p}_{s_h,a_h})W_d^\pi(\mathbf{1}_z, \bar{p}, \mathbf{1}_{s_h,a_h})\mathbb{I}[s_h \neq z]\right] \geq 0. \qquad (108)$$

■

Lemma 32 *Let p and p' be two different transition model. Let the reward r, policy π, horizon d and initial distribution μ_1 be fixed. It then holds that*

$$\begin{aligned}
&W_d^\pi(r,p,\mu_1) - W_d^\pi(r,p',\mu_1) \\
&= \sum_{h=1}^{d}\sum_{s,a}\mathbb{P}_{\pi,p}[(s_h,a_h)=(s,a)|s_1\sim\mu_1]\sum_{s'}(p_{s,a,s'}-p'_{s,a,s'})W_{d-h}^\pi(r,p',\boldsymbol{1}_{s'}).
\end{aligned} \tag{109}$$

Proof Let $\tilde{\mu}_i$ be the distribution of s_{i+1} under p for $i \geq 0$. Define $w_i = W_i^\pi(r,p,\mu_1) + W_{d-i}^\pi(r,p',\tilde{\mu}_i)$. Then we have that $w_d = W_d^\pi(r,p,\mu_1)$ and $w_0 = W_d^\pi(r,p',\mu_1)$. Then the proof is completed by noting that

$$\begin{aligned}
&w_{h+1} - w_h \\
&= W_{h+1}^\pi(r,p,\mu_1) + W_{d-h+1}^\pi(r,p',\tilde{\mu}_{h+1}) - W_h^\pi(r,p,\mu_1) + W_{d-h}^\pi(r,p',\tilde{\mu}_h) \\
&= \sum_{s,a}\tilde{\mu}_h(s)\pi_{h+1}(a|s)r_{h+1}(s,a) + W_{d-h-1}^\pi(r,p',\tilde{\mu}_{h+1}) - W_{d-h}^\pi(r,p',\tilde{\mu}_h) \\
&= \sum_{s,a}\tilde{\mu}_h(s)\pi_{h+1}(a|s)r_{h+1}(s,a) + W_{d-h-1}^\pi(r,p',\tilde{\mu}_{h+1}) \\
&\qquad - \sum_{s,a}\tilde{\mu}_h(s)\pi_{h+1}(a|s)(r_{h+1}(s,a) + \sum_{s'}p'_{s,a,s'}W_{d-h-1}^\pi(r,p',\mathbf{1}_s)) \\
&= \sum_{s,a}\tilde{\mu}_h(s)\pi_{h+1}(a|s)\sum_{s'}p_{s,a,s'}W_{d-h-1}^\pi(r,p',\mathbf{1}_s) - \sum_{s,a}\tilde{\mu}_h(s)\pi_{h+1}(a|s)\sum_{s'}p'_{s,a,s'}W_{d-h-1}^\pi(r,p',\mathbf{1}_s).
\end{aligned}$$

Here the second last inequality is by the fact that $\tilde{\mu}_{h+1}(s') = \sum_{s,a}\tilde{\mu}_h(s)\pi_{h+1}(a|s)p_{s,a,s'}$ ■