PURE EXPLORATION FOR A GOOD POLICY IN REINFORCEMENT LEARNING WITH BANDIT FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Pure exploration in episodic Reinforcement Learning has primarily focused on Best Policy Identification (BPI), which seeks to identify a (near)-optimal policy with high confidence. Motivated by practical settings where a "good enough" policy suffices, we study an alternate objective of Good Policy Identification (GPI). For a given reward threshold μ_0, GPI only requires identifying a policy with expected reward in an episode at least μ_0 if such a policy exists (positive instance), or declaring None if no such policy exists (negative instance). We formalize GPI under the fixed-confidence setting. We require the output to be correct with probability $\geq 1-\delta$, and seek to minimize the expected sample complexity, which is the expected number of episodes explored for the output. We propose a novel algorithm BEE-GPI, and derive theoretically-grounded upper bounds on its sample complexity for positive and negative instances. Notably, for positive instances, the coefficient of $\log 1/\delta$ in our upper bound is $O(H^2/(V^* - \mu_0)^2)$, where H is the episode length and V^* is the optimal expected reward in an episode. The coefficient does not depend on the action and state space sizes otherwise, in sharp contrast to the sample complexity in BPI. We further establish lower bound results to show the near-optimality of BEE-GPI and the necessity of the $1/(V^*-\mu)^2$ term. Numerical experiments further validate the efficiency of our approach.

1 INTRODUCTION

Pure exploration is a fundamental objective in Reinforcement Learning (RL), where a learning agent seeks to identify a policy that satisfies specific performance criteria. Unlike the regret minimization framework, pure exploration prioritizes the probability of correct identification and sample efficiency over total episodes. To date, most research in this field has concentrated on Best Policy Identification (BPI) or ϵ-Best Policy Identification (ϵ-PI), leaving the problem of Good Policy Identification (GPI) largely unexplored. In an episodic Markov Decision Process (MDP) with unknown transition dynamics, GPI requires the agent to identify a policy whose value function exceeds a pre-specified threshold μ_0, if such a policy exists, or to return None otherwise.

Many real-world applications necessitate finding a "good enough" policy rather than a nearly optimal one, particularly when precise pairwise comparisons are computationally or practically difficult. For instance, in medical informatics, a new treatment protocol is typically evaluated against a known *Standard of Care* (with known utility μ_0). The objective is to identify a treatment sequence that ensures a patient recovery rate or "Years of Life Gained" whose utility is $\geq \mu_0$. In such cases, identifying the absolute best treatment with a high confidence $1-\delta$ is often unnecessary, especially when the performance differences among top-tier policies are marginal. Similarly, in cloud computing, service providers use RL for traffic routing and resource allocation. Rather than seeking an absolutely optimal routing rule, providers often aim to ensure reliability and efficiency levels that exceed a specified Service Level Agreement. Furthermore, institutional investors frequently evaluate whether a new trading strategy meets a minimum Sharpe ratio or Return-on-Investment (ROI) benchmark before deployment. This benchmark μ_0 is often derived from established indices with extensive historical data. In these scenarios, the primary concern is determining whether a new alternative is competitive relative to the benchmark, rather than finding the best strategy among all possibile ones.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Main Contributions. We provide three primary contributions. First, we formalize the Good Policy Identification (GPI) problem in the fixed-confidence setting. Given a failure probability tolerance parameter $\delta \in (0,1)$, the learning agent must identify a policy with expected reward at least threshold μ_0 (or correctly identify its non-existence) with probability $\geq 1-\delta$. Second, we develop BEE-GPI, a novel δ-PAC algorithm that utilizes a early-stopping BPI sub-routine as an oracle to achieve efficient sample complexity. We provide a rigorous upper bound on its expected sample complexity, demonstrating that for positive instances, the coefficient of $\log(1/\delta)$ shrinks with the gap $(V^* - \mu_0)$, but is independent of the state and action space sizes. Third, we establish lower bounds for the GPI problem. The resulting gap between our upper and lower bounds is limited to a polynomial factor of S, A, and H, indicating that BEE-GPI is nearly optimal.

Notation. Denote $\mathbb{N}, \mathbb{R}$ as the sets of positive integers and real numbers respectively. For $n \in \mathbb{N}$, denote $[n] = \{1, \ldots, n\}$. We abbreviate a random variable R following the probability distribution p as $R \sim p$.

Problem Formulation. An instance of the Good Policy Identification (GPI) problem is specified by the tuple $\nu = (\mathcal{S}, \mathcal{A}, \mu_0, \delta, H, R, P, \mathsf{p})$. The sets $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, respectively. We denote S, A as the respective cardinalities of $\mathcal{S}, \mathcal{A}$. The scalar $\mu_0 \in \mathbb{R}$ is the reward threshold, which is the expected total reward in an episode that the learning agent aims to accrue. The error parameter $\delta \in (0,1)$ is the tolerance level on the probability of an incorrect output. The integer H represents the episode horizon. The transition kernels and reward function are respectively denoted by $P = \{p_h(\cdot|s,a)\}_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}$ and $R = \{r_h(s,a)\}_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}$, respectively. At round $h \in [H]$ when the agent is at state $s \in \mathcal{S}$ and takes action $a \in \mathcal{A}$, the agent earns a deterministic reward $r_h(s,a) \in [0,1]$, and transits to states s' with probability $p_h(s'|s,a)$. The initial state S_1 is distributed according to the initial distribution p. We assume $\mathcal{S}, \mathcal{A}, \mu_0, \delta, H, R$ are known to the agent, while P, p are not known. To facilitate our discussion on P, p, we prepend each episode with a fictitious time round 0. Round 0 is associated with one fictitious s_0 and the single fictitious action a_0, which carry zero reward and yield state transition $p_0(\cdot|s_0,a_0) = \mathsf{p}(\cdot)$.

A policy π is expressed as $(\pi_h)_{h=1}^H \in \Pi^H$, where $\Pi = \{f \mid f : \mathcal{S} \to \mathcal{A}\}$ consists of mapping from the current state to an action. Thus, Π^H is the space of all deterministic policies, and π_h determines the action in round $h \in [H]$. The value function of π from round h to the end of the episode is

$$V_h^\pi(s_h) := \mathbb{E}^\pi \left[\sum_{\ell=h}^{H} r_\ell(s_\ell, \pi_\ell(s_\ell)) \mid s_h \right],$$

where $s_{\ell+1} \sim p_\ell(\cdot|s_\ell, \pi_\ell(s_\ell))$. Let $\pi^* \in \Pi^H$ denote an optimal policy. By Bellman optimality, we have $V_h^{\pi^*}(s) \geq V_h^\pi(s)$ for any policy π and $s \in \mathcal{S}, h \in [H]$. We abbreviate $V_0^*(s_0) = V_0^{\pi^*}(s_0)$ as the optimal total expected reward in an episode, and occasionally append the instance ν to denote $V_0^*(s_0 \mid \nu)$ to emphasize the dependence on ν.

Dynamics. The agent's algorithm is characterized by a sampling rule $\{\pi^t\}_{t=1}^\infty$, a stopping time τ, and a recommendation rule $\hat{\pi}$. When episode $t \in \mathbb{N}$ begins, the algorithm selects a policy $\pi^t = (\pi_h^t)_{h=1}^H \in \Pi^H$ based on the collected data $\mathcal{D}_{t-1}$ in episodes $1, \ldots, t-1$. Executing π^t on the instance generates the trajectory $z_t = (s_0, a_0, S_{t,1}, A_{t,1}, r_1(S_{t,1}, A_{t,1}), \ldots, S_{t,H}, A_{t,H}, r_H(S_{t,H}, A_{t,H}))$, where $S_{t,h} \sim p_{h-1}(\cdot|S_{t,h-1}, \pi_{h-1}^t(S_{t,h-1}))$. This trajectory is added to the history, i.e. $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{z_t\}$, where we initialize $\mathcal{D}_0 = \emptyset$. At the end of an episode, the agent may choose to terminate the algorithm. We define the stopping time τ as the index of the terminating episode. Upon termination, the agent outputs an answer $\hat{\pi} \in \Pi^H \cup \{\mathsf{None}\}$ based on $\mathcal{D}_\tau$.

The agent aims to output a policy $\hat{\pi} \in \Pi^H$ satisfying $V_0^{\hat{\pi}}(s_0) \geq \mu_0$ if μ_0 can be achieved, or to output None if the agent concludes that $V_0^*(s_0) < \mu_0$. We formalize the aim in the fixed confidence setting with some definitions that aid our discussions:

Definition 1 (Positive and Negative Instances)**.** *A GPI instance ν is a positive instance if $V_0^*(s_0 \mid \nu) > \mu_0$, and is a negative instance if $V_0^*(s_0 \mid \nu) < \mu_0$.*

Definition 2 (Qualified Policy)**.** *For a positive instance ν, we say that a policy π is qualified for ν, if $V_0^\pi(s_0 \mid \nu) \geq \mu_0$.*

The correctness of the output is qualified by being δ-Probably Approximately Correct (δ-PAC):

Definition 3 (δ-PAC)**.** *An algorithm is called δ-PAC, if it can guarantee*

$$\begin{cases} \Pr(V_0^{\hat{\pi}_\tau}(s_0 \mid \nu) \geq \mu_0, \tau < +\infty) \geq 1-\delta & \text{if } V_0^*(s_0 \mid \nu) > \mu_0 \\ \Pr(\hat{\pi}_\tau = \textsf{None}, \tau < +\infty) \geq 1-\delta & \text{if } V_0^*(s_0 \mid \nu) < \mu_0 \end{cases}$$

holds for any positive or negative Good Policy Identification instance ν.

Degenne & Koolen (2019) have proved that, on an instance ν with $V_0^*(s_0|\nu) = \mu_0$, any δ-PAC algorithm suffers from $\mathbb{E}_\nu[\tau] = \infty$, even when we restrict ν to the case of $|\mathcal{S}| = H = 1$. Thus, we only consider positive or negative instances in this paper.

Objective. The goal of the agent is to design a δ-PAC algorithm $(\{\pi^t\}_{t=1}^\infty, \tau, \hat{\pi})$ that minimizes the expected sampling complexity $\mathbb{E}_\nu[\tau]$.

2 Literature Review

A popular objective of pure exploration in Reinforcement Learning (RL) is to identify an optimal or ϵ-optimal policy with high confidence. This line of research, often termed Best Policy Identification (BPI), was pioneered by Fiechter (1994). Much of the subsequent work assumes the availability of a generative model, which allows the agent to interact with the environment by querying arbitrary state-action pairs Kearns & Singh (1998); Gheshlaghi Azar et al. (2013); Sidford et al. (2018; 2023); Agarwal et al. (2020); Al Marjani et al. (2021); Li et al. (2024). While the generative setting is theoretically insightful, a more practical and challenging branch of BPI focuses on the online (non-generative) setting, where the agent must navigate the MDP through sequential episodes to collect data. In this context, Dann & Brunskill (2015) proposed the UCFH algorithm based on the principle of optimism under uncertainty. This was followed by BPI-UCBVI Ménard et al. (2021), which improved the dependence on the state space S by adapting minimax-optimal regret algorithms. Al Marjani & Proutiere (2021) study BPI for RL in non-episodic MDPs. More recently, research has pivoted toward instance-dependent sample complexity. While algorithms like those in Wagenmaker et al. (2022) and Narang et al. (2024) achieve state-of-the-art non-asymptotic bounds, they are often computationally intensive. To bridge this gap, computationally efficient but theoretically suboptimal optimistic sampling rules have been analyzed by Kaufmann et al. (2021); Ménard et al. (2021); Tirinzoni et al. (2023).

Despite the extensive literature on BPI and ϵ-PI, these methods cannot be directly applied to the Good Policy Identification (GPI) problem. Recall that for ϵ-PI, the goal is to identify a policy whose expected reward is at least $V_0^*(s_0) - \epsilon$. The difficulty lies in the fact that $V_0^*(s_0)$ is unknown; consequently, an agent cannot pre-determine an appropriate ϵ for an ϵ-PI algorithm to ensure that the identified policy achieves the threshold μ_0. Furthermore, the lower bounds established for ϵ-PI Domingues et al. (2021) do not apply under the threshold-based requirements of GPI.

In the simplified case where $S = 1$ and $H = 1$, GPI reduces to the multi-armed bandit problem, specifically the "1-identification" or "any-low(high)" problem studied in Degenne & Koolen (2019); Katz-Samuels & Jamieson (2020); Li & Cheung (2025; 2026), also covered but not mainly focused in Kano et al. (2017); Jourdan et al. (2026). While these works provide different strategies for the bandit setting, the transition to the full episodic MDP setting remains unexplored.

3 Algorithms and Sample Complexity Upper Bounds

We propose the algorithm dubbed Balanced Exploration-Exploitation for Good Policy Identification (BEE-GPI), displayed in Algorithm 1. We elaborate on BEE-GPI in Section 3.1, and discuss its theoretical performance guarantees in Section 3.2.

3.1 Algorithm BEE-GPI

BEE-GPI, displayed in Algorithm 1, proceeds in phases. Each phase $k = 1, 2, \ldots$ consists of an adaptive exploration stage (Line 3) that estimates P, p, and potentially outputs a policy $\hat{\pi}_k$. The exploration stage is followed by a decision branching (Line 4 to 10), which leads to three possibilities: (1) the termination of BEE-GPI, or (2) an exploitation stage for policy evaluation (Lines 10-15), or

(3) proceeding to the next phase $k+1$. We elaborate on them in the following three main steps, preceded by a preliminary set-up step 0.

Algorithm 1 Balanced Exploration-Exploitation for Good Policy Identification (BEE-GPI)

1: **Initialize:** For $k \in \mathbb{N}$, $\epsilon_k = 1/2^k$, $\delta_k = 1/3^k$ and $\alpha_k = 5^k$. Initialize $\mathcal{H}^{\text{ee}} = \mathcal{H}^{\text{et}} = \emptyset, C = 1.01$.
2: **for** Phase $k = 1, 2, \ldots$ **do**
3: **Explore:** Call Algorithm 2 with $k \leftarrow k$, $\delta \leftarrow \delta_k$, $C \leftarrow C$, $\mathcal{H}^{\text{ee}} \leftarrow$ current exploration history $\mathcal{H}^{\text{ee}}$, exploration budget $T_k^{\text{ee}} \leftarrow (5)$,
Return Updated exploration history $\mathcal{H}^{\text{ee}}$, output $\hat{\pi}_k \in \Pi^H \cup \{\textsf{None}, \textsf{Not Completed}\}$
4: **if** $\hat{\pi}_k = \textsf{Not Completed}$ **then**
5: Go to the next phase $k+1$
6: **else if** $\hat{\pi}_k = \textsf{None}$ and $\delta_k \geq \delta/10$ **then**
7: Go to the next phase $k+1$
8: **else if** $\hat{\pi}_k = \textsf{None}$ and $\delta_k < \delta/10$ **then**
9: **Return** None, and declare the instance as negative.
10: **else if** $\hat{\pi}_k \in \Pi^H$ **then**
11: **Exploit:** Denote $\{X_i^{k,\text{et}}\}_{i=1}^{+\infty}$ as i.i.d. random variables, where the $X_i^{k,\text{et}}$ is the total reward collected in executing policy $\hat{\pi}_k$ for an episode in the i-th trial. $\{X_i^{k,\text{et}}\}_{i=1}^{+\infty}$ are independent of $\mathcal{H}^{\text{ee}}, \mathcal{H}^{\text{et}}$, and denote $\hat{V}_{\text{et}}^{\hat{\pi}_k,N} = \frac{\sum_{i=1}^N X_i^{k,\text{et}}}{N}$. Initialize $N = 0$.
12: **while** $N \leq 100 \cdot \frac{\log \frac{\alpha_k}{\delta} + \log\log \frac{24H^2}{\epsilon_k}}{\epsilon_k} - 1$ **do**
13: Execute $\hat{\pi}_k$ for one episode, and observe the total reward $X_{N+1}^{k,\text{et}}$, $N \leftarrow N+1$.
14: **if** $\hat{V}_{\text{et}}^{\hat{\pi}_k,N} - \sqrt{\frac{H^2 \log[2\alpha_k(\log_2 2N)^2/\delta]}{N}} \geq \mu_0$ **then**
15: Output $\hat{\pi} = \hat{\pi}_k$, and declare the instance as positive.
16: **end if**
17: **end while**
18: $\mathcal{H}^{\text{et}} \leftarrow \mathcal{H}^{\text{et}} \cup \{X_i^{k,\text{et}}\}_{i=1}^N$
19: **end if**
20: **end for**

Step 0: Initialization. BEE-GPI starts by initializing the hyper-parameters in Line 1. For the exploration stage in phase k, the parameter ϵ_k upper-bounds the optimality gap of the potentially output policy $\hat{\pi}_k$, and δ_k is the tolerance level on the probability of an incorrect output. The parameter α_k is for tuning the length of the potential exploration stage in phase k. The set $\mathcal{H}^{\text{ee}}$ is for accumulating the observed sample trajectories during the exploration stages. $\mathcal{H}^{\text{et}}$ is for collecting the total realized rewards by executing a candidate policy $\hat{\pi}_k$, during the exploitation stages. The absolute constant C is for tuning the stopping condition of the exploration stage.

Step 1: Exploration Stage. The exploration stage involves invoking Algorithm 2, which follows the optimistic sampling strategy in the BPI-UCRL framework Kaufmann et al. (2021). Different from Kaufmann et al. (2021), Algorithm 2 involves a novel early-stopping mechanism, which allows saving on the sample complexity for positive instances, and hence dubbed Early-Stopping BPI-UCRL (ES-BPI-UCRL).

We first describe the optimistic strategy, which is warm-started with the accumulated exploration trajectory set $\mathcal{H}^{\text{ee}}$. To ease the notation, we denote $\mathcal{H}^{\text{ee}} = \{z_q\}_{q=1}^t$, where t is the total number of episodes in the previous exploration stages. Following Kaufmann et al. (2021), we optimistically estimate the V function by the following sequence of construction. For each $s \in \mathcal{S}, a \in \mathcal{A}$, we set $n_h^t(s,a,) = \sum_{q=1}^t \mathbf{1}(S_{q,h} = s, A_{q,h} = a)$. For each $h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$, set the empirical transitional kernel after t episode as $\hat{p}_h^t(s'|s,a) = \frac{1}{n_h^t(s,a)} \sum_{q=1}^t \mathbf{1}(S_{q,h} = s, A_{q,h}, S_{q,h+1} = s')$ if $n_h^t(s,a) > 0$, and set $\hat{p}_h^t(s'|s,a) = 1/|\mathcal{S}|$ otherwise. We define the exploration bonus $\beta_p(t,\delta) = \log \frac{2SAH}{\delta} + (S-1)\log\big(e(1+\frac{t}{S-1})\big)$, following Kaufmann et al. (2021). We further define the confidence sets for the transition kernels as: $\mathcal{C}_h^t(s,a;\delta) := \Big\{ q(\cdot|s,a) \in \Delta^{\mathcal{S}} : \text{KL}\big(\widehat{p}_h^t(\cdot|s,a), q(\cdot|s,a)\big) \leq \frac{\beta_p(n_h^t(s,a),\delta)}{n_h^t(s,a)} \Big\}$. For each pair $(s,h) \in \mathcal{S} \times [H]$, $\delta \in (0,1)$, we define the

optimistic value function $\overline{V}_h^t$ and the corresponding optimistic policy $\overline{\pi}^t(\cdot;\delta) = (\overline{\pi}_h^t(\cdot;\delta))_{h=1}^H$ as:

$$\overline{Q}_h^t(s,a;\delta) := r_h(s,a) + \max_{\bar{p}_h(\cdot|s,a)\in\mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \bar{p}_h(s'|s,a)\overline{V}_{h+1}^t(s';\delta) \tag{1}$$

$$\overline{V}_h^t(s;\delta) := \max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a;\delta), \quad \overline{\pi}_h^t(s;\delta) \in \operatorname*{argmax}_{a\in\mathcal{A}} \left\{\overline{Q}_h^t(s,a;\delta)\right\} \tag{2}$$

where $\overline{Q}_{H+1}^t(s,a;\delta) := 0, \forall s \in \mathcal{S}, a \in \mathcal{A}, \delta \in (0,1)$. The optimistic $\overline{Q}_h^t$ guides our adaptive exploration in each episode. Upon the conclusion of an episode, the exploration history $\mathcal{H}^{\text{ee}}$ and the episode index t are updated. Then, the oracle checks whether any of the three stopping criteria are met (Line 3).

We next discuss the novel early-stopping mechanism of ES-BPI-UCRL, Recall $\overline{\pi}^t(\cdot;\delta) = \{\overline{\pi}_h^t(\cdot;\delta)\}_{h=1}^H$ defined in (2). We set a lower confidence bound on the value function:

$$\underline{V}_h^{t,\overline{\pi}^t(\cdot;\delta)}(s;\delta) := r_h(s,\overline{\pi}_h^t(s;\delta)) + \min_{\underline{p}_h\in\mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \underline{p}_h(s'|s,\overline{\pi}_h^t(s;\delta))\underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta)}(s';\delta),$$

where we initialize $\underline{V}_{H+1}^{t,\overline{\pi}^t(\cdot;\delta)}(s;\delta) = 0$ for all s. These definitions allow us to specify the following two stopping criteria for the oracle routine (Algorithm 2) with tolerance level δ_k:

$$\underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) - (C-1)(\overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k)) > \mu_0, \tag{3}$$

$$\overline{V}_0^t(s_0;\delta_k) < \mu_0. \tag{4}$$

The left hand side of (3) is a carefully crafted lower confidence bound of the value function of policy $\overline{\pi}^t(\cdot;\delta)$, while the left hand side of (4) represents an upper confidence bound of the value function of policy $\overline{\pi}^t(\cdot;\delta)$, which is also the upper confidence bound of $V_0^*(s_0)$. The stopping criteria also involve a budget constraint T_k^{ee} on the cumulative number of exploration episodes conducted so far:

$$T_k^{\text{ee}} = \frac{(H+1)^2SA\log(2SAH/\delta_k)}{\epsilon_k} + \frac{(H+1)^2S^2A}{\epsilon_k}\log\left(\frac{(H+1)^2S^2A\log(2SAH/\delta_k)}{\epsilon_k}\right). \tag{5}$$

The oracle checks whether With probability $1-\delta_k$: (i) condition (3) implies $\overline{\pi}^{t_k}(\cdot;\delta_k)$ is a qualified policy ($\hat{\pi}_k = \overline{\pi}^{t_k}(\cdot;\delta_k)$); (ii) condition (4) implies no qualified policy exists ($\hat{\pi}_k = \textsf{None}$); and (iii) exceeding the budget ($t > T_k^{\text{ee}} - 1$) indicates that the current exploration stage was insufficient.

Algorithm 2 Exploration Oracle: Early-Stopping BPI-UCRL (ES-BPI-UCRL)

1: **Input:** Phase index k, Error tolerance level δ_k, tuning parameter $C > 1$, accumulated exploration stage trajectories $\mathcal{H}^{\text{ee}}$, total budget T_k^{ee}.
2: **Initialize:** round index $t = |\mathcal{H}^{\text{ee}}|$, $n_h^t(s,a) = \sum_{q=1}^t \mathbf{1}(S_{q,h} = s, A_{q,h} = a)$, $\{S_{q,h}, A_{q,h}\}_{q\in[t],h\in[H]}$ are the trajectories stored in $\mathcal{H}^{\text{ee}}$.
3: **while** $\neg$(3) and $\neg$(4) and $t \le T_k^{\text{ee}} - 1$ **do**
4: **for** round $h = 0, 1, 2\cdots, H$ **do**
5: Execute $\overline{\pi}_h^t(S_{t+1,h};\delta_k)$ by taking action $A_{t+1,h} = \max_a \overline{Q}_h^t(S_{t+1,h},a;\delta_k)$ ▷ Note $S_{t+1,0} = s_0, A_{t+1,0} = a_0$
6: Observe reward $r_h(S_{t+1,h}, A_{t+1,h})$ and next state $S_{t+1,h+1}$
7: **end for**
8: Update $\mathcal{H}^{\text{ee}} \leftarrow \mathcal{H}^{\text{ee}} \cup \{(s_0, a_0, \{S_{t+1,h}, A_{t+1,h}, r_h(S_{t+1,h}, A_{t+1,h})\}_{h=1}^H)\}$
9: Update $t \leftarrow t+1$.
10: **end while**
11: Denote $t_k \leftarrow t$ for analysis. ▷ To clarify the definition of t_k, $t_k = |\mathcal{H}^{\text{ee}}|$ also holds
12: **if** (3) holds **then**
13: Output a policy $\hat{\pi}_k = (\overline{\pi}_h^t(\cdot|\delta_k))_{h=1}^H$
14: **else if** (4) holds **then**
15: Output $\hat{\pi}_k = \textsf{None}$
16: **else if** $t > T_k^{\text{ee}} - 1$ **then**
17: Output $\hat{\pi}_k = \textsf{Not Completed}$
18: **end if**

Step 2: Decision Branching. Based on the output $\hat{\pi}_k$ and the current δ_k, Algorithm 1 executes one of four logic branches. (i) If $\hat{\pi}_k = \textsf{Not Completed}$ (Line 4), the phase k exploration stage is inconclusive, and BEE-GPI proceeds phase $k+1$ with an increased budget T^{ee}_{k+1} and decreased tolerance level δ_{k+1} to continue exploration. (ii) If $\hat{\pi}_k = \textsf{None}$ and $\delta_k \geq \delta/10$, BEE-GPI still advances to phase $k+1$ but not terminate with a result. This ensures the δ-PAC requirement is met, as BEE-GPI only recommends None when the confidence is sufficiently high ($\delta_k < \delta/10$, the third branch at Line 8). (iv) If $\hat{\pi}_k$ is a policy (Line 10), BEE-GPI proceeds to the exploitation phase.

Step 3: Exploitation and Verification. The exploitation phase validates whether the candidate policy $\hat{\pi}_k$ is indeed qualified. Treating the total reward per episode as a random variable, the algorithm constructs a confidence interval with a coverage probability of $1 - \frac{\pi^2}{6}\frac{\delta}{\alpha_k}$. By collecting N independent samples, the algorithm evaluates the lower confidence bound: $\hat{V}^{\hat{\pi}_k,N}_{\text{et}} - \sqrt{H^2 \log(2\alpha_k(\log_2 2N)^2/\delta)/N}$. To ensure efficient sample complexity, we require that $N \leq T^{\text{et}}_k - 1$, where

$$T^{\text{et}}_k = 100 \cdot \frac{\log(\alpha_k/\delta) + \log\log(24H^2/\epsilon_k)}{\epsilon_k}. \tag{6}$$

If the lower confidence bound exceeds μ_0 within the budget $N \leq T^{\text{et}}_k - 1$, BEE-GPI recommends $\hat{\pi}_k$ as a qualified policy. Otherwise, it returns to the exploration phase in phase $k+1$.

3.2 Performance Guarantee of BEE-GPI, Algorithm 1

We now establish the theoretical performance of Algorithm 1. We first prove that our algorithm satisfies the standard PAC requirement.

Theorem 1. *Algorithm 1 is δ-PAC.*

The following theorems provide upper bounds on the expected sample complexity $\mathbb{E}[\tau]$ for both positive and negative instances.

Theorem 2. *On a positive instance ν where $V^*_0(s_0) > \mu_0$, BEE-GPI (Algorithm 1) satisfies*

$$\mathbb{E}[\tau] \leq O\left(\frac{H^2 \log\frac{1}{\delta}}{(V^*_0(s_0) - \mu_0)^2}\right) + O\left(\frac{(H+1)^4 SA \log\frac{SAH}{(V^*_0(s_0)-\mu_0)^2} + (H+1)^4 S^2 A \log(HSA)}{(V^*_0(s_0) - \mu_0)^2}\right).$$

Theorem 3. *On a negative instance ν where $V^*_0(s_0) < \mu_0$, BEE-GPI (Algorithm 1) satisfies*

$$\mathbb{E}[\tau] \leq O\left(\frac{(H+1)^4[SA \log\frac{1}{\delta} + S^2 A \log\log\frac{1}{\delta}]}{(V^*_0(s_0) - \mu_0)^2}\right)$$
$$+ O\left(\frac{(H+1)^4 SA \log\frac{H^2}{(V^*_0(s_0)-\mu_0)^2}}{(V^*_0(s_0) - \mu_0)^2} + \frac{(H+1)^4 S^2 A \log\log\left(\frac{H^3 SA}{(V^*_0(s_0)-\mu_0)^2}\right)}{(V^*_0(s_0) - \mu_0)^2}\right).$$

Theorems 1, 2, 3 are proved in Appendix B.1, and we provide their sketch proof after their discussions. The complexity results reveal an inherent asymmetry between positive and negative GPI instances, stemming from the mechanics of our exploration-exploitation stages. For positive instances, the sample complexity is decomposed into a δ-dependent term $O\left(\frac{H^2\log(1/\delta)}{(V^*_0(s_0)-\mu_0)^2}\right)$ and a δ-independent term $O\left(\frac{\text{poly}(S,A,H)}{(V^*_0(s_0)-\mu_0)^2}\right)$. Intuitively, the exploration stage first identifies a candidate policy $\hat{\pi}_k$ that is "sufficiently good" (i.e., $V^{\hat{\pi}_k}_0(s_0) - \mu_0 \geq \frac{C-1}{C}(V^*_0(s_0) - \mu_0)$) at a low confidence level, guaranteed by the stopping rule (3). The effect of adapting the stopping rules is in Lemma 13. The exploitation stage then uses $O\left(\frac{H^2\log(1/\delta)}{(V^*_0(s_0)-\mu_0)^2}\right)$ samples to verify this candidate with high confidence. Notably, the coefficient of $\log(1/\delta)$ is independent of S and A, which is a significant improvement over standard ϵ-optimal BPI reductions by taking $\epsilon = V^*_0(s_0) - \mu_0$, where the coefficient of $\log(1/\delta)$ is typically polynomial in H, S, A and $1/\epsilon^2 = 1/(V^*_0(s_0) - \mu_0)^2$.

For negative instances, the algorithm must certify that *every* policy is unqualified. This requires the oracle to globally explore the MDP to ensure the upper confidence bound of the optimal value

function is below μ_0. Because this verification requires checking all state-action pairs (h, s, a), the $\log(1/\delta)$ term remains coupled with a polynomial of S and A.

Sketch Proof of Theorems 1, 2, 3. We define $\mathcal{E}^{\text{cnt}}(\delta_k)$ and $\mathcal{E}(\delta_k)$ as the concentration events respectively on n_h^t, $\hat{p}_h^t$ for the k-th exploration stage, and $\mathcal{E}_k^{\text{et}}$ as the concentration event for the exploitation stage at phase k on evaluating the objective value under $\hat{\pi}^k$. For a formal definition of the concentration event, we refer the reader to Appendix B.1. A key feature of our algorithm is that sampling trajectories are shared across all exploration periods; consequently, these events are nested, i.e., $\mathcal{E}^{\text{cnt}}(\delta_k) \subset \mathcal{E}^{\text{cnt}}(\delta_{k+1})$ and $\mathcal{E}(\delta_k) \subset \mathcal{E}(\delta_{k+1})$.

To analyze the termination round of the algorithm, we introduce the random variable κ, representing the first phase in which the exploration concentration events hold:

$$\kappa = \min\{k : \mathcal{E}(\delta_k) \cap \mathcal{E}^{\text{cnt}}(\delta_k) \text{ hold}\}.$$

Furthermore, we define $\kappa_{\text{end}}^{\text{pos}}$ and $\kappa_{\text{end}}^{\text{neg}}$ as the indices of the phases by which the algorithm is guaranteed to terminate for positive and negative instances, respectively:

$$\kappa_{\text{end}}^{\text{pos}} = \min\left\{k : k \geq \max\left\{\kappa, O(1)\log_2 \frac{H^2}{(V_0^*(s_0) - \mu_0)^2}\right\}, \mathcal{E}_k^{\text{et}} \text{ holds}\right\},$$

$$\kappa_{\text{end}}^{\text{neg}} = \max\left\{\kappa, O(1)\log_2 \frac{H^2}{(V_0^*(s_0) - \mu_0)^2}, \log_3 \frac{O(1)}{\delta}\right\}.$$

Lemma 7 (Appendix B.2) establishes that Algorithm 1 terminates no later than the conclusion of phase $\kappa_{\text{end}}^{\text{pos}}$ (for positive ν) or $\kappa_{\text{end}}^{\text{neg}}$ (for negative ν). This lemma provides the foundation for our three main results.

Correctness (δ-PAC Guarantee). First, we establish that the algorithm terminates almost surely. Specifically, the probability $\Pr_\nu(\tau = +\infty)$ is bounded by the probability that the stopping phases κ_{end} are infinite. Since concentration theorems ensure that $\Pr_\nu(\kappa \geq k) \leq \delta_{k-1}$ and the conditional probability $\Pr_\nu(\kappa_{\text{end}}^{\text{pos}} - \kappa \geq k \mid \kappa < +\infty) \leq \frac{\pi^2}{6\alpha_{k-1}}$, we conclude that $\Pr_\nu(\tau < +\infty) = 1$. Given $\tau < +\infty$ almost surely, Theorem 1 follows by showing that the conditions for recommending a policy or outputting None only lead to wrong output when a concentration event is violated. By applying a union bound over all phases, we ensure the total error probability is at most δ.

Sample Complexity. We bound the expected sample complexity. Now, Lemma 18 provides implicit upper bounds for $t_{\kappa_{\text{end}}^{\text{pos}}} - t_{\kappa-1}$ and $t_{\kappa_{\text{end}}^{\text{neg}}} - t_{\kappa-1}$. Combining this with the inequality $t_{\kappa-1} \leq T_{\kappa-1}^{\text{ee}}$, we derive an upper bound for conditional expectation $\mathbb{E}[\tau|\kappa]$. This conditional bound is formulated as $\text{poly}(S, A, H, \log(1/\delta), (V_0^*(s_0) - \mu_0)^{-2}) \cdot \frac{\log \frac{\alpha_\kappa \epsilon_\kappa}{\delta_\kappa}}{\epsilon_\kappa}$, where $\text{poly}(S, A, H, \log(1/\delta), (V_0^*(s_0) - \mu_0)^{-2})$ is the right-hand side of inequalities in Theorem 2, 3. Finally, we compute the total expectation using the law of total expectation: $\mathbb{E}[\tau] = \mathbb{E}[\mathbb{E}[\tau|\kappa]]$. $\Pr_\nu(\kappa \geq k) \leq \delta_{k-1}$ implies

$$\mathbb{E}\left[\frac{\log \frac{\alpha_\kappa \epsilon_\kappa}{\delta_\kappa}}{\epsilon_\kappa}\right] \leq \sum_{k=1}^{+\infty} \delta_{k-1} \cdot \frac{\log \frac{\alpha_k \epsilon_k}{\delta_k}}{\epsilon_k} = O(1),$$

which concludes Theorems 2 and 3.

4 Sample Complexity Lower Bounds for δ-PAC Algorithms

We establish lower bounds on the sample complexity of δ-PAC algorithms on positive and negative instances. Due to space constraints, we present the core definitions and the sketch of the main ideas here, while deferred formal proofs and complete notations are provided in Appendix C.1.

4.1 Positive Instances

We construct a class of difficult MDP instances termed "Uniform Instances."

Definition 4 (Sketch Definition of a Uniform Instance, specified by (S, A, H, r, ϵ))**.** *Given even number H, integer $S \geq 3$, integer $A \geq 2$, positive real number r, ϵ such that $\frac{1}{4} < r < r + \epsilon <$*

$\frac{3}{4}$. *We define a Uniform Instance with* $\mathcal{S} = [S-2] \cup \{s_{good}, s_{bad}\}$, $\mathcal{A} = [A]$, *reward function* $r_h(s,a) = 1$ *iff* $s = s_{good}$ *else* 0. *For all* $h, a \in [H] \times \mathcal{A}$, $p_h(\cdot|s_{good}, a)$ $p_h(\cdot|s_{bad}, a)$ *are both uniform distribution on* $[S-2]$. *For all* $h, s \in [H] \times [S-2]$, $p_h(\cdot|s,a)$ *is a two-point distribution on* $\{s_{bad}, s_{good}\}$ *with* $p_h(s_{good}|s, a=1) = r + \epsilon$, $p_h(s_{good}|s, a \geq 2) = r$. *The initial distribution* $p_0(\cdot)$ *is uniform on* $[S-2]$.

The construction ensures that at any odd round h, the state $S_{t,h}$ follows a uniform distribution over the $[S-2]$, regardless of the policy. At even round, the agent is forced into either s_{good} or s_{bad}. The definition of a Uniform Instance assumes an agent can only collect positive reward at the state s_{good}. To maximize reward, the agent must identify the "optimal" action (action 1) at each state $s \in [S-2]$. To characterize the difficulty of identifying this optimal action, we consider the set of all possible permutations of the action space.

Definition 5 (Permutation of an Instance ν)**.** *Given an 1-policy-identification instance* ν *with reward function* $\{r_h^\nu(s,a)\}_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}$ *and transition probability kernel* $\{p_h^\nu(s'|s,a)\}_{(h,s's,a)\in[H]\times\mathcal{S}\times\mathcal{S}\times\mathcal{A}}$, *and a set of permutations* $\sigma = \{\sigma_{s,h}\}_{(s,h)\in\mathcal{S}\times[H]}$, *each* $\sigma_{s,h}$ *is a bijection* $\mathcal{A} \to \mathcal{A}$, *we define a permutation of* ν *as instance* ν_σ, *with reward function* $r_h^{\nu_\sigma}\big(s, \sigma_{s,h}(a)\big) = r_h^\nu(s,a)$, $p_h^{\nu_\sigma}(s'|s,\sigma(a)) = p_h^\nu(s'|s,a)$.

For a Uniform Instance ν, the optimal policy for its permutation ν_σ is $\pi_h(s) = \sigma_{s,h}(1)$, with optimal value $\frac{H(r+\epsilon)}{2}$. In addition, the value function of a policy π on instance ν_σ is

$$V_0^\pi(s_0 \mid \nu_\sigma) = \frac{\sum_{h:\text{odd}} \sum_{s=1}^{S-2} \mathbb{1}(\pi_h(s) = \sigma_{s,h}(1))(r+\epsilon) + \mathbb{1}(\pi_h(s) \neq \sigma_{s,h}(1))r}{S-2}$$

It is evident to see $\frac{Hr}{2} \leq V_0^\pi(s_0; \nu_\sigma) \leq \frac{H(r+\epsilon)}{2}$. To explore for a qualified policy, the agent has to guarantee sufficient exploration for the optimal action $\sigma_{s,h}(1)$ on a fraction of pair $\{(h,s) : h\text{:odd}, s \in [S-2]\}$, leading to the following Theorem.

Theorem 4. *Consider any* δ*-PAC alg and any Uniform Instance* ν *specified by* (S, A, H, r, ϵ), *with* μ_0 *satisfying* $\frac{3H(r+\epsilon)}{8} + \frac{Hr}{8} > \mu_0 > \frac{Hr}{2}$, $S \geq 100, H \geq 10$, $A \geq 100$, $(S-2)H$ *is divisible by* 16, $0 < \epsilon < \frac{1}{200\sqrt{5}}$. *Denote* $\Gamma := \{\sigma = \{\sigma_{s,h}\}_{(s,h)\in\mathcal{S}\times H} : \forall s, h, \sigma_{s,h}$ *is a bijection* $\mathcal{A} \to \mathcal{A}\}$, *the set of all the possible permutations on* ν, *it holds that*

$$\max_{\sigma\in\Gamma} \mathbb{E}_{\nu_\sigma, alg}[\tau] \geq \Omega\left(\frac{H \log\frac{1}{\delta}}{(V_0^*(s_0 \mid \nu_\sigma) - \mu_0)^2} + \frac{H^2SA}{(V_0^*(s_0 \mid \nu_\sigma) - \mu_0)^2}\right).$$

Theorem 4 is proved in Appendix C.2. The lower bound suggests that the upper bound in Theorem 2 is near-optimal from the minimax perspective. Specifically, the gap in the δ-dependent term is at most a factor of H, while the δ-independent term remains within poly$(H, S, \log(A))$.

4.2 Negative Instances

We first introduce the sketch definition of Tree Instance, with full definition in the Appendix C.1.

Definition 6 (Sketch Definition of a Tree Instance, specified by (S, A, H, r))**.** *Consider integer* S, A, H *such that* $S = 2^N + 1$ *for some integer* $N \geq 1$, *and* $A \geq 3$, $H - 1 \geq 6N = 6\log_2(S-1)$, *and a real number* $r \in (\frac{3}{8}, \frac{5}{8})$. *Take* $\mathcal{S} = [2^N - 1] \cup \{s_{good}, s_{bad}\}$, $\mathcal{A} = [A]$, *and consider a binary tree with nodes set* $[2^N - 1]$. s_{good}, s_{bad} *are absorbed states, while for state* $s \in [2^{N-1} - 1]$, *action 1 will go to state* $2s$, *action 2 will go to state* $2s+1$, *for action* $a \geq 3$, $p_h(s_{good}|s, a \geq 3) = r, p_h(s_{bad}|s, a \geq 3) = 1 - r$. *For state* $s \in [2^{N-1}, 2^N - 1]$, $p_h(s_{good}|s, a \geq 1) = r$, $p_h(s_{bad}|s, a \geq 1) = 1 - r$. *The reward function is* $r_h(s,a) = 1$ *iff* $s = s_{good}$ *else* 0. *The initial state is* 1, *with* $p_0(1) = 1, p_0(s) = 0, \forall s \neq 1$.

In a Tree Instance, an agent only collects rewards upon reaching s_{good}. To maximize the total reward in an episode, the agent must determine the round h and state s at which to transition to the absorbing states. For a threshold $\mu_0 > (H-1)r$, the instance is negative. To correctly identify it as a negative instance, the agent is forced to explore all state-action pairs (h, s, a) for $a \geq 3$ to ensure that no hidden "high-reward" transitions exist. This leads to the following lower bound:

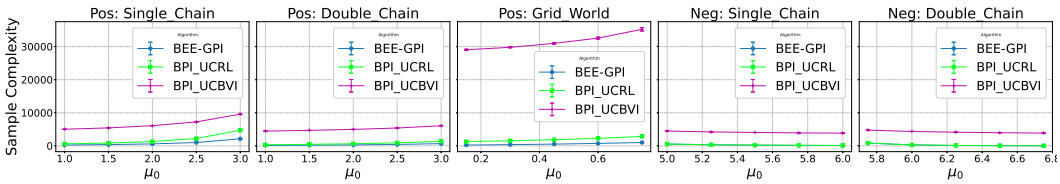


Figure 1: Numerical comparison between BEE-GPI and BPI-UCRL (the lower the better).

Theorem 5. *Apply any δ-PAC algorithm to a Tree instance ν specified by (S, A, H, r) with $S, A, H \geq 4$ and $\frac{3}{4}(H - \log_2(S-3)) > \mu_0 > (H-1)r$, we have*

$$\mathbb{E}_\nu \tau \geq \Omega\left(\frac{HSA\log\frac{1}{\delta}}{(\mu_0 - V_0^*(s_0 \mid \nu))^2}\right).$$

Full proof is in Appendix C.3. Comparing Theorem 5 with our upper bound in Theorem 3, we observe that the gap is restricted to a factor of $H^3 \log \frac{SAH}{(\mu_0 - V_0^*)^2} + S\log\log(1/\delta)$. Since $\log\log(1/\delta) < \log(1/\delta)$ for all $\delta \in (0,1)$, the gap between our upper and lower bounds is bounded by a polynomial in S, A, and H. Thus, our proposed algorithm BEE-GPI achieves a nearly optimal sample complexity.

5 Numerical Experiments

We provide numerical evidence to evaluate the empirical efficiency of our proposed BEE-GPI. We consider 3 environment types, including Single Chain, Double Chain, Grid World, which are from Kaufmann et al. (2021). By tuning μ_0 while fixing δ, H, $\mathcal{S}$, $\mathcal{A}$ and $p_h(\cdot|\cdot)$ for each environment types, we set up 5 experiment groups: 3 positive cases and 2 negative cases. Details of the environment setup is in the appendix E.

We compare our approach against two ϵ-optimal BPI algorithm: BPI_UCRL in Kaufmann et al. (2021) and BPI_UCBVI in Ménard et al. (2021). To adapt these two algorithms for the GPI task, we provide it with the oracle parameter $\epsilon = V_0^* - \mu_0$, a significant prior-knowledge advantage. For the negative case, we compare the upper confidence bound of V^*, maintained by BPI-UCRL and BPI-UCBVI, with μ_0 as the stopping rule. We run each experiment with 20 independent trials, and deploy them on an Intel i5-14500 CPU over a total computation time of approximately 72 hours.

Figure 1 presents the empirical stopping times averaged for each setting. The error bars represent three standard deviations (3σ), though they are notably small due to the low variance of termination round. As illustrated in Figure 1, BEE-GPI consistently outperforms the benchmarks in the positive instance, by requiring significantly fewer episodes to identify a qualified policy across all tested thresholds. While in the negative instances, the curves of BEE-GPI and BPI_UCRL nearly overlaps, as the BEE-GPI calls BPI_UCRL as an Oracle for exploration. Algorithm BPI_UCBVI spends more episodes to terminate compared to these two algorithm. These observations validate our theoretical contribution in Theorem 2: by decoupling the $\log(1/\delta)$ coefficient from the state and action space sizes (S, A), BEE-GPI achieves a tighter sample complexity than a standard reductions to the state-of-the-art ϵ-optimal BPI algorithms. What's more, BEE-GPI remains competitive in the negative instances.

6 Conclusion

We formalized the Good Policy Identification (GPI) problem and proposed the BEE-GPI algorithm. We establish upper and lower bounds on sample complexity, demonstrating that BEE-GPI is minimax optimal up to a polynomial factor of S, A, and H. Critically, we showed that for positive instances, our algorithm decouples the $\log 1/\delta$ term from the state-action space size, an efficiency gain validated by our numerical experiments. Future research includes deriving tighter instance-dependent bounds and extending this threshold-based framework to large-scale MDPs using function approximation.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing and editing the code for numeric experiments. We have not used generative AI tools for the generation of proof. Additionally, we used generative AI tools for suggesting experimental parameters. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided proof details in the Appendix, and source-code for reproduction of numeric results.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 67–83. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/agarwal20b.html.

Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes. In *International Conference on Machine Learning*, pp. 7459–7468. PMLR, 2021.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34:25852–25864, 2021.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 578–598. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the seventh annual conference on Computational learning theory*, pp. 88–97, 1994.

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL http://jmlr.org/papers/v11/jaksch10a.html.

Marc Jourdan, Andrée Delahaye-Duriez, and Clémence Réda. An anytime algorithm for good arm identification. *Journal of Machine Learning Research*, 27(19):1–90, 2026.

Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, (2), 2017.

Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1781–1791. PMLR, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021.

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

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024.

Zitian Li and Wang Chi Cheung. Near optimal non-asymptotic sample complexity of 1-identification. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34144–34184. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25f.html.

Zitian Li and Wang Chi Cheung. Closing the gap on the sample complexity of 1-identification. *arXiv preprint arXiv:2601.15620*, 2026.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Adhyyan Narang, Andrew Wagenmaker, Lillian Ratliff, and Kevin G Jamieson. Sample complexity reduction via policy difference estimation in tabular reinforcement learning. *Advances in Neural Information Processing Systems*, 37:22772–22826, 2024.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. *Naval Research Logistics (NRL)*, 70(5):423–442, 2023.

Andrea Tirinzoni, Aymen Al-Marjani, and Emilie Kaufmann. Optimistic pac reinforcement learning: the instance-dependent view. In *International Conference on Algorithmic Learning Theory*, pp. 1460–1480. PMLR, 2023.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pp. 358–418. PMLR, 2022.

A NOTATION

In this section, We list the notation used in Algorithms 1, 2, and we supplement the details on the notation in the analysis in Appendix B.

- $r_h(s,a)$: mean reward of taking action a in state s at round h.
- $p_h(\cdot|s,a)$: true transition kernel when we take action a in state s at round h.
- $p_h^\pi(s,a)$: Execute a given policy π on the MDP. Denote S_h as the random state the agent locate on round h, we define $p_h^\pi(s,a) = \Pr(S_h = s, \pi_h(s) = a)$.
- $S_{t,h}$: realized state round h at t th episode.
- $\beta^{\text{cnt}}(\delta) = \log \frac{2SAH}{\delta}$, $\beta_p(t,\delta) = \log \frac{2SAH}{\delta} + (S-1)\log\left(e(1+\frac{t}{S-1})\right)$

- $\hat{p}_h^t(\cdot|s,a)$ as empirical transition kernel after executing t episodes during the exploration stages. $\hat{p}_h^t(s'|s,a) = \frac{1}{n_h^t(s,a)} \sum_{q=1}^t \mathbf{1}(S_{q,h} = s, A_{q,h}, S_{q,h+1} = s')$ if $n_h^t(s,a) > 0$, and set $\hat{p}_h^t(s'|s,a) = 1/|\mathcal{S}|$ otherwise.
- $n_h^t(s,a) := \sum_{q=1}^t \mathbf{1}(S_{q,h} = s, A_{q,h} = a)$, visiting times of pair (s,a,h) after executing t episode during the exploration stages. $n_h^t(s',s,a) := \sum_{q=1}^t \mathbf{1}(S_{q,h+1} = s, S_{q,h} = s, A_{q,h} = a)$, visiting times of pair (s,a,h) and the next state is s' after executing t episodes. Here we denote $\mathcal{H}^{\text{ee}} = \{S_{q,h}, A_{q,h}\}_{q\in[t],h\in[H]}$.
- $\bar{n}_h^t(s,a)$: Denote k_q as the stage index at q-th sampled episode, and $\bar{\pi}^q := \bar{\pi}^q(\cdot;\delta_{k_q})$, defined in (2). We define $\bar{n}_h^t(s,a) := \sum_{q=0}^{t-1} p_h^{\bar{\pi}^q}(s,a)$. We define $\bar{\pi}^0$ by (2) with $\mathcal{C}_h^0(s,a;\delta_1) := \Delta^{\mathcal{S}}$.
- $V_0^*(s_0|\nu)$ optimal value function of instance ν. If we only mention one instance ν, we may use $V_0^*(s_0)$ for simplicity.

What's more, we introduce the confidence interval for each pair (s,a,h)

$$\mathcal{C}_h^t(s,a;\delta) := \left\{ q(\cdot|s,a) \in \Delta^{\mathcal{S}} : \mathrm{KL}\big(\widehat{p}_h^t(\cdot|s,a), q(\cdot|s,a)\big) \le \frac{\beta_p(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}.$$

Following Al Marjani & Proutiere (2021), we define the same optimistic and pessimistic approximation value function, for each (h,s,a).

$$\begin{aligned}
\overline{Q}_h^t(s,a;\delta) :=& r_h(s,a) + \max_{\bar{p}_h(\cdot|s,a)\in\mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \bar{p}_h(s'|s,a)\overline{V}_{h+1}^t(s';\delta) \\
\overline{V}_h^t(s;\delta) :=& \max_a \overline{Q}_h^t\big(s,a;\delta\big) \\
\overline{V}_{H+1}^t(s;\delta) :=& 0 \\
\bar{p}_h^t(s,a;\delta) \in& \arg\max_{\bar{p}_h\in\mathcal{C}_h^t(s,a,\delta)} \sum_{s'} \bar{p}_h(s'|s,a)\overline{V}_{h+1}^t(s';\delta) \\
\bar{\pi}_h^t(s;\delta) \in& \arg\max_a \overline{Q}_h^t\big(s,a;\delta\big) \\
\overline{V}_0^t(s_0;\delta) =& \max_{\bar{p}_0\in\mathcal{C}_0^t(s_0;\delta)} \sum_{s'} \bar{p}_0(s'|s_0,a_0)\overline{V}_1^t(s';\delta)
\end{aligned}$$

$$\begin{aligned}
\underline{Q}_h^t(s,a;\delta) :=& r_h(s,a) + \min_{\underline{p}_h\in\mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \underline{p}_h(s'|s,a)\underline{V}_{h+1}^t(s';\delta) \\
\underline{V}_h^t(s;\delta) :=& \max_a \underline{Q}_h^t(s,a;\delta) \\
\underline{V}_{H+1}^t(s;\delta) :=& 0 \\
\underline{p}_h^t(s,a;\delta) \in& \arg\min_{\underline{p}\in\mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \underline{p}_h(s'|s,a)\underline{V}_{h+1}^t(s';\delta) \\
\underline{\pi}_h^t(s;\delta) \in& \arg\max_a \underline{Q}_h^t\big(s,a;\delta\big) \\
\underline{V}_0^t(s_0;\delta) =& \min_{\underline{p}_0\in\mathcal{C}_0^t(s_0;\delta)} \sum_{s'} \bar{p}_0(s'|s_0,a_0)\underline{V}_1^t(s';\delta)
\end{aligned}$$

$$\begin{aligned}
\overline{Q}_h^{t,\pi}(s,a;\delta) :=& r_h(s,a) + \max_{\bar{p}_h\in\mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \bar{p}_h(s'|s,a)\overline{V}_{h+1}^{t,\pi}(s';\delta) \\
\overline{V}_h^{t,\pi}(s;\delta) :=& \overline{Q}_h^{t,\pi}\big(s,\pi(s);\delta\big) \\
\overline{V}_{H+1}^{t,\pi}(s;\delta) :=& 0 \\
\bar{p}_h^{t,\pi}(s,a;\delta) \in& \arg\max_{\bar{p}_h\in\mathcal{C}_h^t(s,a,\delta)} \sum_{s'} \bar{p}_h(s'|s,a)\overline{V}_{h+1}^{t,\pi}(s';\delta)
\end{aligned}$$

$$\begin{aligned}
\underline{Q}_h^{t,\pi}(s,a;\delta) :=& r_h(s,a) + \min_{\underline{p}_h \in \mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \underline{p}_h(s'|s,a)\underline{V}_{h+1}^{t,\pi}(s';\delta) \\
\underline{V}_h^{t,\pi}(s;\delta) :=& \underline{Q}_h^{t,\pi}(s,\pi(s);\delta) \\
\underline{V}_{H+1}^{t,\pi}(s;\delta) :=& 0 \\
\underline{p}_h^{t,\pi}(s,a;\delta) \in& \arg\min_{\underline{p} \in \mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \underline{p}_h(s'|s,a)\underline{V}_{h+1}^{t,\pi}(s';\delta).
\end{aligned}$$

Finally, we introduce the commonly used notation Q and V, noted as the "True" value function.

$$\begin{aligned}
Q_h^{\pi}(s,a) :=& r_h(s,a) + \sum_{s'} p_h(s'|s,a)V_{h+1}(s') \\
V_h^{\pi}(s) :=& Q_h^{\pi}(s,\pi(s)) \\
V_{H+1}^{\pi}(s) :=& 0 \\
V_0^{\pi}(s_0) =& \sum_{s'} p_0(s'|s_0,a_0)V_1^{\pi}(s').
\end{aligned}$$

B Performance Guarantee of Algorithm 1

B.1 Proof of Main Theorems

Recall we take default value $\delta_k = \frac{1}{3^k}$ in Algorithm 1. We define

$$\mathcal{E}^{\text{cnt}}(\delta) = \left\{ \forall t \in \mathbb{N}^*, \forall (s,a,h) : n_h^t(s,a) \geq \frac{1}{2}\bar{n}_h^t(s,a) - \beta^{\text{cnt}}(\delta) \right\} \tag{7}$$

$$\mathcal{E}(\delta) = \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a,h), \text{KL}(\hat{p}_h^t(\cdot|s,a), p_h(\cdot|s,a)) \leq \frac{\beta_p(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\} \tag{8}$$

$$\kappa = \min\{k : \mathcal{E}(\delta_k) \cap \mathcal{E}^{\text{cnt}}(\delta_k) \text{ hold}\} \tag{9}$$

$$\mathcal{E}_k^{\text{et}} = \left\{ \forall N > 0, \left| \hat{V}_{\text{et}}^{\hat{\pi}_k,N} - V_0^{\hat{\pi}_k} \right| \leq \sqrt{\frac{4H^2 \log \frac{2\alpha_k(\log_2 2N)^2}{\delta}}{N}} \right\}, \tag{10}$$

where $\mathcal{E}^{\text{cnt}}(\delta), \mathcal{E}(\delta)$ are only defined for the samples collected in the exploration period. Recall A key feature of Algorithm 1 is that sampling trajectories are shared across all exploration periods; consequently, $\{\mathcal{E}^{\text{cnt}}(\delta_k), \mathcal{E}(\delta_k)\}_{k=1}^{+\infty}$ are nested, i.e., $\mathcal{E}^{\text{cnt}}(\delta_k) \subset \mathcal{E}^{\text{cnt}}(\delta_{k+1})$ and $\mathcal{E}(\delta_k) \subset \mathcal{E}(\delta_{k+1})$.

We first prove the output correctness of Algorithm 1.

Theorem (Restatement of Theorem 1)**.** *Algorithm 1 is δ-PAC.*

Proof of Theorem 1. By the Lemma 8, we know $\Pr_\nu(\tau < +\infty) = 1$ holds no matter ν is positive or negative. The remaining work is to prove $\Pr_\nu(\hat{\pi} \neq \textsf{None}, V_0^{\hat{\pi}}(s_0) > \mu_0) \geq 1 - \delta$ for any positive instance ν and $\Pr_\nu(\hat{\pi} = \textsf{None}) \geq 1 - \delta$ for any negative instance ν. Or equivalently, $\Pr_\nu(\hat{\pi} = \textsf{None} \text{ or } V_0^{\hat{\pi}}(s_0) < \mu_0) < \delta$ holds for any positive instance ν, and $\Pr_\nu(\hat{\pi} \neq \textsf{None}) < \delta$

If ν is positive, we have

$$\Pr_\nu(\hat{\pi} = \textsf{None} \text{ or } V_0^{\hat{\pi}}(s_0) < \mu_0) \leq \Pr_\nu(\hat{\pi} = \textsf{None}) + \Pr_\nu(\hat{\pi} \neq \textsf{None}, V_0^{\hat{\pi}}(s_0) < \mu_0). \tag{11}$$

We are going to prove $\Pr_\nu(\hat{\pi} = \textsf{None}) \leq (1 + \frac{\pi^2}{3})\frac{3\delta}{10}$ and $\Pr_\nu(\exists k \in \mathbb{N}, \hat{\pi} \neq \textsf{None}, V_0^{\hat{\pi}}(s_0) \leq \frac{\pi^2\delta}{24}$.

By the Line (8) in Algorithm 1, we have

$$\begin{aligned}
&\Pr_{\nu}(\hat{\pi} = \textsf{None}) \\
\leq &\Pr_{\nu}\left(\exists \delta_k \leq \frac{\delta}{10}, \exists t \in \mathbb{N}, \overline{V}_0^{\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) < \mu_0\right) \\
\leq &\Pr_{\nu}\left(\exists \delta_k \leq \frac{\delta}{10}, \exists t \in \mathbb{N}, \overline{V}_0^{\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) < V_0^*(s_0)\right) \\
\leq &\Pr_{\nu}\left(\exists k \geq \lceil \log_3 \frac{10}{\delta} \rceil, \kappa \geq k\right) \qquad (12) \\
\leq &\frac{1}{3^{\lceil \log_3 \frac{10}{\delta} \rceil - 1}}. \qquad (13) \\
\leq &\frac{3\delta}{10}. \qquad (14)
\end{aligned}$$

Step (12) is by the Proposition 12 and Step (13) by the Lemma 10.

By the Line (14) in Algorithm 1, we have

$$\begin{aligned}
&\Pr_{\nu}(\exists k \in \mathbb{N}, \hat{\pi}_k \text{ is a policy}, V_0^{\hat{\pi}_k}(s_0) < \mu_0, \hat{\pi} = \hat{\pi}_k) \\
\leq &\Pr_{\nu}\left(\exists k \in \mathbb{N}, \exists N \in \mathbb{N}, \hat{V}_{\text{et}}^{\hat{\pi}_k, N} - \sqrt{\frac{4H^2 \log \frac{2\alpha_k (\log_2 2N)^2}{\delta}}{N}} \geq \mu_0\right),
\end{aligned}$$

where $\hat{V}_{\text{et}}^{\hat{\pi}_k, N} = \frac{\sum_{i=1}^N X_i^{(k)}}{N}$, $\{X_i^{(k)}\}_{i=1}^{+\infty}$ are i.i.d random variables, denoting the rewards collected by executing policy $\hat{\pi}_k$ during a single episode. The mean value for each $X_i^{(k)}$ is $V_0^{\hat{\pi}_k}(s_0) < \mu_0$. Since $X_i^{(k)}$ is bounded by the interval $[0, H]$, we can assert $X_i^{(k)} - V_0^{\hat{\pi}_k}(s_0)$ is $\frac{H^2}{4}$-subgaussin. By the union bound, we have

$$\begin{aligned}
&\Pr_{\nu}(\exists k \in \mathbb{N}, \hat{\pi} \neq \textsf{None}, V_0^{\hat{\pi}}(s_0) < \mu_0) \\
\leq &\sum_{k=1}^{+\infty} \Pr_{\nu}\left(\exists N \in \mathbb{N}, \frac{\sum_{i=1}^N X_i^{(k)}}{N} - \sqrt{\frac{H^2 \log \frac{2\alpha_k (\log_2 2N)^2}{\delta}}{N}} \geq \mu_0\right) \\
\leq &\sum_{k=1}^{+\infty} \Pr_{\nu}\left(\exists N \in \mathbb{N}, \frac{\sum_{i=1}^N X_i^{(k)}}{N} - \sqrt{\frac{H^2 \log \frac{2\alpha_k (\log_2 2N)^2}{\delta}}{N}} \geq V_0^{\hat{\pi}_k}(s_0)\right) \\
\leq &\sum_{k=1}^{+\infty} \frac{\pi^2}{6} \frac{\delta}{\alpha_k} \\
\leq &\frac{\pi^2 \delta}{6} \frac{1}{4}. \qquad (15)
\end{aligned}$$

The second last step is by the Lemma 23 and the last step is by the default setting $\alpha_k = \frac{1}{5^k}$. Substitute (15) and (14) into (11), we have

$$\Pr_{\nu}(\hat{\pi} = \textsf{None} \text{ or } V_0^{\hat{\pi}}(s_0) < \mu_0) \leq \frac{\pi^2 \delta}{6} \frac{1}{4} + \frac{3\delta}{10} < \delta.$$

If ν is negative, we can follow the calculation of (15), and conclude

$$\Pr_{\nu}(\hat{\pi} \neq \textsf{None}) \leq \frac{\pi^2 \delta}{6} \frac{1}{4} < \delta.$$

□

Then, we turn to prove the upper bound of $\mathbb{E}\tau$, by applying Algorithm 1.

Theorem (Restatement of Theorem 2)**.** *Apply Algorithm 1 to a positive instance ν with $V_0^*(s_0) > \mu_0$, we have*

$$\mathbb{E}\tau \le O\left(\frac{H^2\log\frac{1}{\delta} + (H+1)^4SA\log\frac{SAH}{(V_0^*(s_0)-\mu_0)^2} + (H+1)^4S^2A\log(HSA)}{(V_0^*(s_0)-\mu_0)^2}\right)$$

Proof of Theorem 2. Denote $L^{\text{pos}} = \max\{\kappa, \lceil\log_2\frac{9600(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}\rceil\}$, and define

$$\kappa_{\text{end}} = \min\left\{k : k \ge L^{\text{pos}}, \mathcal{E}_k^{\text{et}} \text{ holds}\right\}.$$

By the Lemma 7, We know Algorithm 1 will terminate no later than the end of phase κ_{end}.

Given $\kappa < +\infty$, and index $k \ge 1$, we know

$$\begin{aligned}
&\Pr(\kappa_{\text{end}} \ge L^{\text{pos}} + k|\kappa)\\
\le&\Pr\left((\neg\mathcal{E}^{\text{et}}_{L^{\text{pos}}}) \cap \cdots \cap (\neg\mathcal{E}^{\text{et}}_{L^{\text{pos}}+k-1})|\kappa\right)\\
\le&\Pr(\neg\mathcal{E}^{\text{et}}_{L^{\text{pos}}+k-1}|\kappa)\\
\le&\frac{\pi^2\delta}{6\alpha_{L^{\text{pos}}+k-1}}
\end{aligned} \tag{16}$$

The last step is by the Lemma 23.

Denote τ^{ee}, τ^{et} as the total collected episodes by the exploration and exploitation oracle. Theorem 20 suggests that

$$\begin{aligned}
\tau^{\text{ee}} \le&\frac{101(H+1)^4SA\log\frac{2SAH}{\delta_{\kappa_{\text{end}}}}}{\epsilon^{\text{pos}2}} + T^{\text{ee}}_{\kappa-1} +\\
&\frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(\frac{402(H+1)^4S^2A\log\frac{2SAH}{\delta_{\kappa_{\text{end}}}}}{\epsilon^{\text{pos}2}}\right) + \frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right)\\
\tau^{\text{et}} \le&\sum_{k=1}^{\kappa_{\text{end}}} 100\cdot\frac{\log\frac{\alpha_k}{\delta} + \log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}
\end{aligned}$$

where $\epsilon^{\text{pos}} = \frac{V_0^*(s_0)-\mu_0}{C}$. For simplicity, we denote $V_0^* := V_0^*(s_0)$ in the following proof. We first derive an upper bound for $\mathbb{E}\tau^{\text{ee}}$. By (16), we know $\Pr(\kappa_{\text{end}} - L^{\text{pos}} + k|\kappa) \le \frac{\pi^2\delta}{6\alpha_{L^{\text{pos}}+k-1}}, k \ge 1$, thus, we can conclude

$$\begin{aligned}
&\mathbb{E}[\tau^{\text{ee}}|\kappa]\\
\le&\frac{101(H+1)^4SA\log\frac{2SAH}{\delta_{L^{\text{pos}}}}}{\epsilon^{\text{pos}2}} + \sum_{k=1}^{+\infty}\frac{\pi^2}{6}\frac{\delta}{5^{k-1}}\frac{101(H+1)^4SA\log\frac{2SAH}{\delta_{L^{\text{pos}}+k}}}{\epsilon^{\text{pos}2}}\\
&T^{\text{ee}}_{\kappa-1} + \frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(\frac{402(H+1)^4S^2A\log\frac{2SAH}{\delta_{L^{\text{pos}}}}}{\epsilon^{\text{pos}2}}\right) +\\
&\sum_{k=1}^{+\infty}\frac{\pi^2}{6}\frac{\delta}{5^{k-1}}\frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(\frac{402(H+1)^4S^2A\log\frac{2SAH}{\delta_{L^{\text{pos}}+k}}}{\epsilon^{\text{pos}2}}\right) +\\
&\frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right)\\
\le&O(1)\underbrace{\frac{(H+1)^4SA\log\frac{2SAH}{\delta_{L^{\text{pos}}}}}{\epsilon^{\text{pos}2}}}_{\dagger 1} + O(1)\underbrace{\frac{(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(\frac{(H+1)^4S^2A\log\frac{2SAH}{\delta_{L^{\text{pos}}}}}{\epsilon^{\text{pos}2}}\right)}_{\dagger 2} +\\
&\underbrace{\frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right) + T^{\text{ee}}_{\kappa-1}}_{\dagger 3}
\end{aligned}$$

To move forward, we take expectation on †1, †2 and †3. By the Lemma 9, we have

$$
\begin{aligned}
&\mathbb{E}\frac{(H+1)^4SA\log\frac{2SAH}{\delta_{L^{\text{pos}}}}}{\epsilon^{\text{pos}2}}\\
\leq&\mathbb{E}\left(\frac{(H+1)^4SA\log\frac{2SAH}{\delta_\kappa}}{\epsilon^{\text{pos}2}}+\frac{(H+1)^4SA\log\frac{2SAH}{\delta_{\lceil\log_2\frac{9600H^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}\rceil}}}{\epsilon^{\text{pos}2}}\right)\\
\leq&\sum_{k=1}^{+\infty}\frac{1}{3^{k-1}}\cdot\frac{(H+1)^4SA\log(2SAH\cdot 3^k)}{\epsilon^{\text{pos}2}}+O\left(\frac{(H+1)^4SA\log\frac{SAHC^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}}{(V_0^*-\mu_0)^2}\right)\\
\leq&O\left(\frac{(H+1)^4SA\log\frac{SAHC^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}}{(V_0^*(s_0)-\mu_0)^2}\right).
\end{aligned}
\tag{17}
$$

$$
\begin{aligned}
&\mathbb{E}\frac{(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(\frac{(H+1)^4S^2A\log\frac{2SAH}{\delta_{L^{\text{pos}}}}}{\epsilon^{\text{pos}2}}\right)\\
\leq&\frac{(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(\frac{(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\right)+\mathbb{E}\frac{(H+1)^4S^2A\log\log\frac{2SAH}{\delta_{L^{\text{pos}}}}}{\epsilon^{\text{pos}2}}\\
\leq&\frac{(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(\frac{(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\right)+\sum_{k=1}^{+\infty}\frac{1}{3^{k-1}}\frac{(H+1)^4S^2A\log\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}}+\\
&\frac{(H+1)^4S^2A\log\log\frac{2SAH}{\delta_{\lceil\log_2\frac{9600(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}\rceil}}}{\epsilon^{\text{pos}2}}\\
\leq&O\left(\frac{C^2(H+1)^4S^2A}{(V_0^*(s_0)-\mu_0)^2}\log\left(\frac{C^2(H+1)^4S^2A\log\frac{C^2(H+1)^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}}{(V_0^*(s_0)-\mu_0)^2}\right)\right)
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
&\mathbb{E}\frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\log\left(2T_{\kappa-1}^{\text{ee}}\right)+T_{\kappa-1}^{\text{ee}}\\
\leq&\sum_{k=1}^{+\infty}\frac{1}{3^{k-1}}\left(\frac{(H+1)^2SA\log\frac{2SAH}{\delta_k}}{\epsilon_k}+\frac{(H+1)^2S^2A}{\epsilon_k}\log\left(\frac{(H+1)^2S^2A\log\frac{2SAH}{\delta_k}}{\epsilon_k}\right)\right)+\\
&\frac{400(H+1)^4S^2A\log 2}{\epsilon^{\text{pos}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{pos}2}}\cdot\\
&\sum_{k=1}^{+\infty}\frac{1}{3^{k-1}}\log\left(\frac{(H+1)^2SA\log\frac{2SAH}{\delta_k}}{\epsilon_k}+\frac{(H+1)^2S^2A}{\epsilon_k}\log\left(\frac{(H+1)^2S^2A\log\frac{2SAH}{\delta_k}}{\epsilon_k}\right)\right)\\
\leq&O\Big(\frac{C^2(H+1)^4S^2A\log(SAH)}{(V_0^*(s_0)-\mu_0)^2}+(H+1)^2S^2A\log(SAH)\Big)\\
\leq&O\Big(\frac{C^2(H+1)^4S^2A\log(SAH)}{(V_0^*(s_0)-\mu_0)^2}\Big).
\end{aligned}
\tag{19}
$$

The last step is by the fact that $(V_0^*(s_0)-\mu_0)^2\leq H^2$. Sum up (17), (18) (19) and plug in $C=1.01$, we can conclude

$$
\begin{aligned}
&\mathbb{E}\tau^{\text{ee}}\\
=&\mathbb{E}\mathbb{E}[\tau^{\text{ee}}|\kappa]\\
\leq&O\left(\frac{(H+1)^4SA\log\frac{SAH}{(V_0^*(s_0)-\mu_0)^2}}{(V_0^*(s_0)-\mu_0)^2}+\frac{(H+1)^4S^2A\log(HSA)}{(V_0^*(s_0)-\mu_0)^2}\right)
\end{aligned}
\tag{20}
$$

The remaining work is to derive an upper bound for $\mathbb{E}\tau^{\text{et}}$, we also first derive an upper bound for $\mathbb{E}[\tau^{\text{et}}|\kappa]$, which is

$$
\begin{aligned}
\mathbb{E}[\tau^{\text{et}}|\kappa] \leq & \mathbb{E}\left[\sum_{k=1}^{\kappa_{\text{end}}} 100 \cdot \frac{\log\frac{\alpha_k}{\delta} + \log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}|\kappa\right] \\
\leq & \mathbb{E}\left[200 \cdot \frac{\log\frac{\alpha_{\kappa_{\text{end}}}}{\delta} + \log\log\frac{24H^2}{\epsilon_{\kappa_{\text{end}}}}}{\epsilon_{\kappa_{\text{end}}}}|\kappa\right] \\
= & \mathbb{E}\left[200 \cdot \frac{\log\frac{\alpha_{\kappa_{\text{end}}-L^{\text{pos}}}\cdot\alpha_{L^{\text{pos}}}}{\delta} + \log\log\frac{24H^2}{\epsilon_{\kappa_{\text{end}}-L^{\text{pos}}}\cdot\epsilon_{L^{\text{pos}}}}}{\epsilon_{\kappa_{\text{end}}-L^{\text{pos}}}\cdot\epsilon_{L^{\text{pos}}}}|\kappa\right] \\
\leq & \frac{200\log\frac{\alpha_{L^{\text{pos}}}}{\delta} + 200\log\log\frac{24H^2}{\epsilon_{L^{\text{pos}}}}}{\epsilon_{L^{\text{pos}}}} + \sum_{k=1}^{+\infty}\frac{200\pi^2\delta}{6\alpha_{k-1}}\cdot\frac{\log\frac{\alpha_k\cdot\alpha_{L^{\text{pos}}}}{\delta} + \log\log\frac{24H^2}{\epsilon_k\cdot\epsilon_{L^{\text{pos}}}}}{\epsilon_k\cdot\epsilon_{L^{\text{pos}}}} \qquad (21) \\
= & \frac{200\log\frac{\alpha_{L^{\text{pos}}}}{\delta} + 200\log\log\frac{24H^2}{\epsilon_{L^{\text{pos}}}}}{\epsilon_{L^{\text{pos}}}} + \sum_{k=1}^{+\infty}\frac{200\pi^2\delta\cdot 2^k}{6\cdot 5^{k-1}}\cdot\frac{k\log 5 + \log\frac{\alpha_{L^{\text{pos}}}}{\delta} + \log\log\frac{24H^2\cdot 2^k}{\epsilon_{L^{\text{pos}}}}}{\epsilon_{L^{\text{pos}}}} \\
\leq & O\left(\frac{\log\frac{\alpha_{L^{\text{pos}}}}{\delta} + \log\log\frac{24H^2}{\epsilon_{L^{\text{pos}}}}}{\epsilon_{L^{\text{pos}}}}\right).
\end{aligned}
$$

Step (21) is from the (16). To derive an upper bound for $\mathbb{E}\tau^{\text{et}}$, we can conduct the following calculation

$$
\begin{aligned}
& \mathbb{E}\tau^{\text{et}} \\
\leq & \mathbb{E}\mathbb{E}[\tau^{\text{et}}|\kappa] \\
\leq & O\left(\mathbb{E}\frac{\log\frac{\alpha_{L^{\text{pos}}}}{\delta} + \log\log\frac{H^2}{\epsilon_{L^{\text{pos}}}}}{\epsilon_{L^{\text{pos}}}}\right) \\
\leq & O\left(\mathbb{E}\frac{\log\frac{\alpha_\kappa}{\delta} + \log\log\frac{24H^2}{\epsilon_\kappa}}{\epsilon_\kappa}\right) + O\left(\frac{\log\frac{\alpha_{\lceil\log_2\frac{9600(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}\rceil}}{\delta} + \log\log\frac{24H^2}{\epsilon_{\lceil\log_2\frac{9600(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}\rceil}}}{\epsilon_{\lceil\log_2\frac{9600(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}\rceil}}\right) \\
\leq & O\left(\log\frac{\log H}{\delta} + \sum_{k=1}^{+\infty}\delta_{k-1}\frac{\log\frac{\alpha_k}{\delta} + \log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}\right) + \\
& O\left(\frac{(H+1)^2C^2}{(C-1)^2}\cdot\frac{\log\frac{1}{\delta} + \log\frac{(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}}{(V_0^*(s_0)-\mu_0)^2}\right) \\
\leq & O\left(\frac{(H+1)^2C^2}{(C-1)^2}\cdot\frac{\log\frac{1}{\delta} + \log\frac{(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2}}{(V_0^*-\mu_0)^2} + \log\frac{\log H}{\delta}\right).
\end{aligned}
$$

The second last step is by the Lemma 9. Plug in $C = 1.01$, we can derive

$$
\mathbb{E}\tau^{\text{et}} \leq O\left(\frac{H^2\log\frac{1}{\delta} + H^2\log\frac{H^2}{(V_0^*(s_0)-\mu_0)^2}}{(V_0^*(s_0)-\mu_0)^2} + \log\frac{\log H}{\delta}\right) \qquad (22)
$$

Summing up (20) and (22), by the fact that $(V_0^*(s_0)-\mu_0)^2 \leq H^2$, we have

$$
\mathbb{E}\tau \leq O\left(\frac{H^2\log\frac{1}{\delta} + (H+1)^4SA\log\frac{SAH}{(V_0^*(s_0)-\mu_0)^2} + (H+1)^4S^2A\log(HSA)}{(V_0^*(s_0)-\mu_0)^2}\right).
$$

□

Theorem (Restatement of Theorem 3)**.** *Apply algorithm 1 to a negative instance ν with $V_0^*(s_0) < \mu_0$, we have*

$$\begin{aligned}&\mathbb{E}\tau\\ \leq& O\left(\frac{(H+1)^4SA(\log\frac{1}{\delta}+\log\frac{H^2}{(V_0^*(s_0)-\mu_0)^2})}{(V_0^*(s_0)-\mu_0)^2}\right)+\\ &O\left(\frac{(H+1)^4S^2A\log\left(\log(SAH)+\log\frac{1}{\delta}+\log\frac{H^2}{(V_0^*(s_0)-\mu_0)^2}\right)}{(V_0^*(s_0)-\mu_0)^2}\right).\end{aligned}$$

Proof of Theorem 3. Denote τ^{ee}, τ^{et} as the total collected episodes by the exploration and exploitation oracle. By the Lemma 7, we know Algorithm 1 will terminate no later than the end of phase $\max\{\lceil\log_2\frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2}\rceil,\lceil\log_3\frac{10}{\delta}\rceil,\kappa\}=:L^{\text{neg}}$. The remaining work is to prove upper bounds for τ^{ee} and τ^{et}, at the end of phase L^{neg}.

For τ^{et}, by the Theorem 20, we can conclude at the end of phase index $k\geq\kappa$, the Exploration Oracle(Line 3 in Algorithm 1) will always output None or Not Completed. This observation implies Algorithm 1 will not enter the exploitation period at Line 11 for phase index $k\geq\kappa$, suggesting that

$$\tau^{\text{et}}\leq\sum_{k=1}^{\kappa-1}100\cdot\frac{\log\frac{\alpha_k}{\delta}+\log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}. \tag{23}$$

For τ^{ee}, by the Theorem 20, we have

$$\begin{aligned}\tau^{\text{ee}}\leq&\frac{101(H+1)^4SA\log\frac{2SAH}{\delta_{L^{\text{neg}}}}}{\epsilon^{\text{neg}2}}+T^{\text{ee}}_{\kappa-1}+\\ &\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\frac{402(H+1)^4S^2A\log\frac{2SAH}{\delta_{L^{\text{neg}}}}}{\epsilon^{\text{neg}2}}\right)+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right).\end{aligned} \tag{24}$$

The remaining is to take expectation on both sides of (23), (24) and derive upper bounds. For (23), by the Lemma 6, we have

$$\begin{aligned}\mathbb{E}\tau^{\text{et}}\leq&\mathbb{E}\sum_{k=1}^{\kappa-1}100\cdot\frac{\log\frac{\alpha_k}{\delta}+\log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}\\ \leq&\mathbb{E}200\cdot\frac{\log\frac{\alpha_{\kappa-1}}{\delta}+\log\log\frac{24H^2}{\epsilon_{\kappa-1}}}{\epsilon_{\kappa-1}}\\ \leq&\sum_{k=1}^{+\infty}\frac{200}{3^{k-1}}\frac{\log\frac{\alpha_{k-1}}{\delta}+\log\log\frac{24H^2}{\epsilon_{k-1}}}{\epsilon_{k-1}}\\ \leq&O\left(\log\frac{\log H}{\delta}\right).\end{aligned} \tag{25}$$

The second step is by taking $\epsilon_k=\frac{1}{2^k}$ and $\left(2\cdot\frac{\log\frac{\alpha_{k+1}}{\delta}+\log\log\frac{24H^2}{\epsilon_{k+1}}}{\epsilon_{k+1}}\right)/\left(2\cdot\frac{\log\frac{\alpha_k}{\delta}+\log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}\right)\geq 2$. The second last step is by the Lemma 10.

For (24), we have

$$\begin{aligned}
&\mathbb{E}\tau^{\text{ee}} \\
\leq &\underbrace{\mathbb{E}\left(\frac{101(H+1)^4SA\log(\frac{1}{\delta_{L^{\text{neg}}}})}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\log\frac{2SAH}{\delta_{L^{\text{neg}}}}\right)\right)}_{\dagger 4}+ \\
&\underbrace{\mathbb{E}\left(T^{\text{ee}}_{\kappa-1}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right)\right)}_{\dagger 5}+ \\
&\frac{101(H+1)^4SA\log(2SAH)}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\frac{402(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\right) \qquad (26)
\end{aligned}$$

For †4,

$$\begin{aligned}
&\mathbb{E}\frac{101(H+1)^4SA\log(\frac{1}{\delta_{L^{\text{neg}}}})}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\log\frac{2SAH}{\delta_{L^{\text{neg}}}}\right) \\
\leq &\frac{101(H+1)^4SA\log(\frac{1}{\delta_{\lceil\log_2\frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2}\rceil}})}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\log\frac{2SAH}{\delta_{\lceil\log_2\frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2}\rceil}}\right)+ \\
&\frac{101(H+1)^4SA\log(\frac{1}{\delta_{\lceil\log_3\frac{10}{\delta}\rceil}})}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\log\frac{2SAH}{\delta_{\lceil\log_3\frac{10}{\delta}\rceil}}\right)+ \\
&\mathbb{E}\frac{101(H+1)^4SA\log(\frac{1}{\delta_\kappa})}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\log\frac{2SAH}{\delta_\kappa}\right). \qquad (27)
\end{aligned}$$

By Lemma 10, we can derive

$$\begin{aligned}
&\mathbb{E}\frac{101(H+1)^4SA\log(\frac{1}{\delta_\kappa})}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\log\frac{2SAH}{\delta_\kappa}\right) \\
\leq &\sum_{k=1}^{+\infty}\delta_{k-1}\left(\frac{101(H+1)^4SA\log(\frac{1}{\delta_k})}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(\log\frac{2SAH}{\delta_k}\right)\right) \\
\leq &O\left(\frac{(H+1)^4S^2A\log\left(\log(SAH)\right)}{\epsilon^{\text{neg}2}}\right).
\end{aligned}$$

Following (27), we can derive

$$\begin{aligned}
&\dagger 4 \\
\leq &O\left(\frac{(H+1)^4SA\log\frac{1}{\delta}}{(V_0^*(s_0)-\mu_0)^2}+\frac{(H+1)^4S^2A\log\left(\log(SAH)+\log\frac{1}{\delta}+\lceil\log\frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2}\rceil\right)}{(V_0^*(s_0)-\mu_0)^2}\right)+ \\
&O\left(\frac{(H+1)^4SA\lceil\log\frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2}\rceil}{(V_0^*(s_0)-\mu_0)^2}\right) \qquad (28)
\end{aligned}$$

For †5, by Lemma 10

$$
\begin{aligned}
&\mathbb{E}\left(T^{\text{ee}}_{\kappa-1}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right)\right)\\
\leq&\sum_{k=1}^{+\infty}\delta_{k-1}\left(\frac{(H+1)^2SA\log\frac{2SAH}{\delta_k}}{\epsilon_k}+\frac{(H+1)^2S^2A}{\epsilon_k}\log\left(\frac{(H+1)^2S^2A\log\frac{2SAH}{\delta_k}}{\epsilon_k}\right)\right)+\\
&\frac{400(H+1)^4S^2A\log 2}{\epsilon^{\text{neg}2}}+\frac{400(H+1)^4S^2A}{\epsilon^{\text{neg}2}}\cdot\\
&\sum_{k=1}^{+\infty}\delta_{k-1}\log\left(\frac{(H+1)^2SA\log\frac{2SAH}{\delta_k}}{\epsilon_k}+\frac{(H+1)^2S^2A}{\epsilon_k}\log\left(\frac{(H+1)^2S^2A\log\frac{2SAH}{\delta_k}}{\epsilon_k}\right)\right)\\
\leq&O\left(\frac{(H+1)^4S^2A\log(SAH)}{(V_0^*(s_0)-\mu_0)^2}+(H+1)^2S^2A\log(SAH)\right)\\
\leq&O\left(\frac{(H+1)^4S^2A\log(SAH)}{(V_0^*(s_0)-\mu_0)^2}\right).
\end{aligned}
\tag{29}
$$

The last step is by the fact that $(V_0^*(s_0)-\mu_0)^2\leq H^2$.

Substitute †4, †5 in (26) with (28) and (29), we prove

$$
\begin{aligned}
\mathbb{E}\tau^{\text{ee}}\leq&O\left(\frac{(H+1)^4SA(\log\frac{1}{\delta}+\log\frac{H^2}{(V_0^*(s_0)-\mu_0)^2})}{(V_0^*(s_0)-\mu_0)^2}\right)+\\
&O\left(\frac{(H+1)^4S^2A\log\left(\log(SAH)+\log\frac{1}{\delta}+\log\frac{H^2}{(V_0^*(s_0)-\mu_0)^2}\right)}{(V_0^*(s_0)-\mu_0)^2}\right).
\end{aligned}
\tag{30}
$$

Combine (25) and (30), we have

$$
\begin{aligned}
\mathbb{E}\tau\leq&O\left(\frac{(H+1)^4SA(\log\frac{1}{\delta}+\log\frac{H^2}{(V_0^*(s_0)-\mu_0)^2})}{(V_0^*(s_0)-\mu_0)^2}\right)+\\
&O\left(\frac{(H+1)^4S^2A\log\left(\log(SAH)+\log\frac{1}{\delta}+\log\frac{H^2}{(V_0^*(s_0)-\mu_0)^2}\right)}{(V_0^*(s_0)-\mu_0)^2}\right).
\end{aligned}
$$

□

B.2 Auxiliary Lemmas

In this section, we provide the proof of the required Lemmas in section B.1. The first one is about the maximum episodes collected in exploration and exploitation periods, up to the end of phase k.

Lemma 6. *Each time Algorithm 1 enters Line 20, i.e. the end of phase k, we have $|\mathcal{H}^{ee}|\leq T^{ee}_k, |\mathcal{H}^{et}|\leq\sum_{\ell=1}^{k}100\cdot\frac{\log\frac{\alpha_\ell}{\delta}+\log\log\frac{24H^2}{\epsilon_\ell}}{\epsilon_\ell}$ hold with certainty.*

Proof of Lemma 6. $|\mathcal{H}^{\text{ee}}|\leq T^{\text{ee}}_k$ is a direct conclusion from the condition $t\leq T^{\text{ee}}_k-1$ at the Line 3 of Algorithm 2.

Regarding $|\mathcal{H}^{\text{et}}|$, its upper bound is from the condition $N\leq 100\cdot\frac{\log\frac{\alpha_k}{\delta}+\log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}$ at the Line 12 of Algorithm 1. Since exploitation period k allows at most collecting $100\cdot\frac{\log\frac{\alpha_k}{\delta}+\log\log\frac{24H^2}{\epsilon_k}}{\epsilon_k}$ episodes, $|\mathcal{H}^{\text{et}}|\leq\sum_{\ell=1}^{k}100\cdot\frac{\log\frac{\alpha_\ell}{\delta}+\log\log\frac{24H^2}{\epsilon_\ell}}{\epsilon_\ell}$ must hold. □

The following is a basis Lemma, asserting the Algorithm must terminate no later than the end of specific phases.

Lemma 7. *Recall the definition* $\mathcal{E}_k^{et} = \left\{ \forall N > 0, \left| \hat{V}_{et}^{\hat{\pi}_k, N} - V_0^{\hat{\pi}_k} \right| \leq \sqrt{\frac{4H^2 \log \frac{2\alpha_k (\log_2 2N)^2}{\delta}}{N}} \right\}$.

Apply Algorithm 1 to an instance ν

- *If the instance* ν *is positive, denote* $L^{pos} = \max\{\kappa, \lceil \log_2 \frac{9600(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2} \rceil\}$*, and define*

$$\kappa_{end} = \min\left\{k : k \geq L^{pos}, \mathcal{E}_k^{et} \text{ holds}\right\}.$$

  We assert the Algorithm 1 must terminate no later than the end of phase κ_{end}*.*

- *If the instance* ν *is negative, denote* $L^{neg} := \max\{\lceil \log_2 \frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2} \rceil, \lceil \log_3 \frac{10}{\delta} \rceil, \kappa\}$*.We assert the Algorithm 1 must terminate no later than the end of phase* L^{neg}*.*

Proof of Lemma 7. We split the proof into two parts by discussing ν is is either positive or negative.

If ν is positive, by the Theorem 20, we can conclude at the end of phase index $k \geq \max\{\lceil \log_2 \frac{9600(H+1)^2C^2}{(V_0^*(s_0)-\mu_0)^2} \rceil, \kappa\}$, the Exploration Oracle(Line 3 in Algorithm 1) will always output $\hat{\pi}_k$ such that $V_0^{\hat{\pi}_k}(s_0) - \mu_0 \geq \frac{C-1}{C}\Big(V_0^*(s_0) - \mu_0\Big)$.

By the Lemma 22 and 24, for any phase index $k \geq \lceil \log_2 \frac{H^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2} \rceil$, if the Exploration Oracle returns a $\hat{\pi}_k$ satisfying that $V_0^{\hat{\pi}_k}(s_0) - \mu_0 \geq \frac{C-1}{C}(V_0^* - \mu_0)$ and the concentration event $\mathcal{E}_k^{\text{et}}$ holds, Algorithm 1 will enter Line 15 and take $\hat{\pi}_k$ as the final output. Recall the definition $\kappa_{\text{end}} = \min\{k : k \geq L^{\text{pos}}, \mathcal{E}_k^{\text{et}} \text{ holds}\}$, we know Algorithm 1 will terminate at the end of phase κ_{end}.

If ν is negative, by the Theorem 20, we can conclude at the end of phase index $k \geq \max\{\lceil \log_2 \frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2} \rceil, \kappa\}$, the Exploration Oracle(Line 3 in Algorithm 1) will always output None. Because of the condition of Line 8 in Algorithm 1, we can assert the algorithm will terminate no later than the end of phase $\max\{\lceil \log_2 \frac{9600(H+1)^2}{(V_0^*(s_0)-\mu_0)^2} \rceil, \lceil \log_3 \frac{10}{\delta} \rceil, \kappa\} =: L^{\text{neg}}$. □

The following Lemma is a direct conclusion from Lemma 7, which guarantee Algorithm 1 must terminate almost surely.

Lemma 8. *Apply Algorithm 1 to an instance* ν*. If* $V_0^*(s_0) \neq \mu_0$*, which means* ν *is either positive or negative, we have* $\Pr(\tau < +\infty) = 1$.

Proof of Lemma 8. Similar to the Lemma 7, we split the proof into two parts by discussing ν is either positive or negative.

If ν is positive, by the Lemma 7, we have

$$\begin{aligned}\Pr(\tau = +\infty) &\leq \Pr(\kappa_{\text{end}} = +\infty) \\ &\leq \Pr(\kappa = +\infty) + \Pr(\kappa_{\text{end}} - L^{\text{pos}} = +\infty | \kappa < +\infty),\end{aligned}$$

where $L^{\text{pos}} = \max\{\kappa, \lceil \log_2 \frac{9600(H+1)^2C^2}{(C-1)^2(V_0^*(s_0)-\mu_0)^2} \rceil\}$. By the Lemma 10, we have $\Pr(\kappa = +\infty) = 0$. In addition, Lemma 23 suggests that

$$\begin{aligned}&\Pr(\kappa_{\text{end}} \geq L^{\text{pos}} + k | \kappa) \\ \leq &\Pr\left((\neg \mathcal{E}_{L^{\text{pos}}}^{\text{et}}) \cap \cdots \cap (\neg \mathcal{E}_{L^{\text{pos}}+k-1}^{\text{et}}) | \kappa\right) \\ \leq &\Pr(\neg \mathcal{E}_{L^{\text{pos}}+k-1}^{\text{et}} | \kappa) \\ \leq &\frac{\pi^2 \delta}{6\alpha_{L^{\text{pos}}+k-1}} \qquad (31) \\ \leq &\frac{\pi^2 \delta}{6\alpha_{k-1}} \qquad (32)\end{aligned}$$

By the (32), we have $\Pr(\kappa_{\text{end}} - L^{\text{pos}} = +\infty | \kappa < +\infty) = 0$. Thus, we have $\Pr(\tau = +\infty) = 0$, which means $\Pr(\tau < +\infty) = 1$.

If ν is negative, by the Lemma 7 and 9, we have

$$\Pr(\tau = +\infty) \leq \Pr(\kappa = +\infty) = 0,$$

which suggests $\Pr(\tau < +\infty) = 1$. □

B.3 Performance Guarantee of Algorithm 2

This section mainly follows the existing analysis of Al Marjani & Proutiere (2021), which also provides required Lemmas for Appendix B.1. For self-completeness, we still provide detailed proof. In the following proof, we follow the convention that takeing $\sum_{n=n_1}^{n_2} a_n = 0$ for non-negative $\{a_n\}_{n=1}^{+\infty}$ and $n_1 > n_2 \geq 0; n_1, n_2 \in \mathbb{Z}$.

By the same argument as the Lemma 10 in Kaufmann et al. (2021), we can conclude

Lemma 9. $\Pr\left(\mathcal{E}(\delta) \cap \mathcal{E}^{cnt}(\delta)\right) \geq 1 - \delta$ *holds for all* $\delta \in (0,1)$*, where* $\mathcal{E}^{cnt}(\delta)$*,* $\mathcal{E}(\delta)$ *are defined in (7), (8).*

Proof of Lemma 9. Lemma 10 of Kaufmann et al. (2021) asserts

$$\begin{aligned}\Pr\left(\mathcal{E}(\delta)\right) \geq& 1 - \frac{1}{2}\delta \\ \Pr\left(\mathcal{E}^{\text{cnt}}(\delta)\right) \geq& 1 - \frac{1}{2}\delta.\end{aligned}$$

□

Given Lemma 9, the following Lemma is obvious.

Lemma 10. $\Pr(\kappa \geq k) \leq \delta_{k-1}$.

Proof of Lemma 10. Recall the definition of $\kappa := \min\{k : \mathcal{E}(\delta_k) \cap \mathcal{E}^{\text{cnt}}(\delta_k) \text{ hold}\}$. Evident to see $\Pr(\kappa \geq k) \leq \Pr(\mathcal{E}(\delta_{k-1}) \cap \mathcal{E}^{\text{cnt}}(\delta_{k-1})$ don't hold). Then Lemma 9 implies Lemma 10. □

The following lemmas shows the optimistic and pessimistic estimation defined in Appendix A indeed contains the value function or Q function.

Lemma 11. *For any round* t *during phase* $k \geq \kappa$*, and any policy* π*, we have*

$$\begin{aligned}\underline{Q}_h^{t,\pi}(s,a;\delta_k) \leq Q_h^{\pi}(s,a) \leq \overline{Q}_h^{t,\pi}(s,a;\delta_k) \\ \underline{Q}_h^{t}(s,a;\delta_k) \leq Q_h^{*}(s,a) \leq \overline{Q}_h^{t}(s,a;\delta_k).\end{aligned}$$

Proof of Lemma 11. We prove Lemma 11 by Induction. Evident to see

$$\begin{aligned}\underline{Q}_{H+1}^{t,\pi}(s,a;\delta) = Q_{H+1}^{\pi}(s,a) = \overline{Q}_{H+1}^{t,\pi}(s,a;\delta) = 0 \\ \underline{Q}_{H+1}^{t}(s,a;\delta) = Q_{H+1}^{*}(s,a) = \overline{Q}_{H+1}^{t}(s,a;\delta) = 0.\end{aligned}$$

Now, assume the conclusion holds for $h+1, h+2, \cdots, H+1$, we turn to work on the case h. To prove $\underline{Q}_h^{t,\pi}(s,a;\delta) \leq Q_h^{\pi}(s,a)$, we can derive

$$\begin{aligned}\underline{Q}_h^{t,\pi}(s,a;\delta) =& r_h(s,a) + \min_{\underline{p}_h \in \mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \underline{p}_h(s'|s,a)\underline{V}_{h+1}^{t,\pi}(s';\delta) \\ \leq& r_h(s,a) + \sum_{s'} p_h(s'|s,a)\underline{V}_{h+1}^{t,\pi}(s';\delta) \\ =& r_h(s,a) + \sum_{s'} p_h(s'|s,a)\underline{Q}_{h+1}^{t,\pi}(s',\pi(s');\delta) \\ \overset{\text{Induction}}{\leq}& r_h(s,a) + \sum_{s'} p_h(s'|s,a)Q_{h+1}^{\pi}(s',\pi(s')) \\ =& Q_h^{\pi}(s,a).\end{aligned}$$

To prove $\overline{Q}_h^{t,\pi}(s,a;\delta) \geq Q_h^{\pi}(s,a)$, we can derive

$$\begin{aligned}
\overline{Q}_h^{t,\pi}(s,a;\delta) =& r_h(s,a) + \max_{\bar{p}_h \in \mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \bar{p}_h(s'|s,a)\overline{V}_{h+1}^{t,\pi}(s';\delta) \\
\geq& r_h(s,a) + \sum_{s'} p_h(s'|s,a)\overline{V}_{h+1}^{t,\pi}(s';\delta) \\
=& r_h(s,a) + \sum_{s'} p_h(s'|s,a)\overline{Q}_{h+1}^{t,\pi}(s',\pi(s');\delta) \\
\overset{\text{Induction}}{\geq}& r_h(s,a) + \sum_{s'} p_h(s'|s,a)Q_{h+1}^{\pi}(s',\pi(s')) \\
=& Q_h^{\pi}(s,a).
\end{aligned}$$

To prove $\underline{Q}_h^t(s,a;\delta) \leq Q_h^*(s,a)$, we can derive

$$\begin{aligned}
\underline{Q}_h^t(s,a;\delta) =& r_h(s,a) + \min_{\underline{p}_h \in \mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \underline{p}_h(s'|s,a)\underline{V}_{h+1}^t(s';\delta) \\
\leq& r_h(s,a) + \sum_{s'} p_h(s'|s,a)\underline{V}_{h+1}^t(s';\delta) \\
=& r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_a \underline{Q}_{h+1}^t(s,a;\delta) \\
\overset{\text{Induction}}{\leq}& r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_a Q_{h+1}^*(s,a) \\
=& r_h(s,a) + \sum_{s'} p_h(s'|s,a)V_{h+1}^*(s) \\
=& Q_h^*(s,a).
\end{aligned}$$

To prove $\overline{Q}_h^t(s,a;\delta) \geq Q_h^*(s,a)$, we can derive

$$\begin{aligned}
\overline{Q}_h^t(s,a;\delta) =& r_h(s,a) + \max_{\bar{p}_h \in \mathcal{C}_h^t(s,a;\delta)} \sum_{s'} \bar{p}_h(s'|s,a)\overline{V}_{h+1}^t(s';\delta) \\
\geq& r_h(s,a) + \sum_{s'} p_h(s'|s,a)\overline{V}_{h+1}^t(s';\delta) \\
=& r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_a \overline{Q}_{h+1}^t(s',a;\delta) \\
\overset{\text{Induction}}{\geq}& r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_a Q_{h+1}(s',a) \\
=& r_h(s,a) + \sum_{s'} p_h(s'|s,a)V_{h+1}^*(s) \\
=& Q_h^*(s,a).
\end{aligned}$$

Then, we complete the induction. □

Given Lemma 11, the following proposition is evident.

Proposition 12. *For any round t during phase $k \geq \kappa$, we have*

$$\max_a \underline{Q}_h^t(s,a;\delta_k) \leq \max_a Q_h^*(s,a) \leq \max_a \overline{Q}_h^t(s,a;\delta_k), \forall s,h$$

which is equivalent to $\underline{V}_h^t(s;\delta) \leq V_h^*(s) \leq \overline{V}_h^t(s;\delta)$, $\forall s,h$. Further, for any policy π,

$$\underline{Q}_h^{t,\pi}(s,\pi_h(s);\delta_k) \leq Q_h^{\pi}(s,\pi_h(s)) \leq \overline{Q}_h^{t,\pi}(s,\pi_h(s);\delta_k), \forall s,h$$

which is equivalent to $\underline{V}_h^{t,\pi}(s;\delta) \leq V_h^{\pi}(s) \leq \overline{V}_h^{t,\pi}(s;\delta)$, $\forall s,h$.

Following the Lemma 3 in Kaufmann et al. (2021), we have

Lemma 13. *For any round t during phase $k \geq \kappa$ in Algorithm 1, meaning that $\mathcal{E}(\delta_k) \cap \mathcal{E}^{cnt}(\delta_k)$ holds, we have*

- *if the instance is positive, the output $\hat{a}$ from Algorithm 2 is either policy $\overline{\pi}^t(\cdot; \delta_k)$, satisfying that*

$$V_0^{\overline{\pi}^t}(s_0) - \mu_0 \geq \frac{C-1}{C}\Big(V_0^*(s_0) - \mu_0\Big)$$

  or "Not Complete"

- *If the instance is negative, the output $\hat{a}$ from Algorithm 2 is either "Negative" or Not Completed.*

Proof of Lemma 13. For simpliciy, we denote $\overline{V}^t := \overline{V}_0^t(s_0; \delta_k)$, $\underline{V}^{t,\overline{\pi}^t} := \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0; \delta_k)$, $V^* := V_0^*(s_0)$, $V^{\overline{\pi}^t} := V_0^{\overline{\pi}^t(\cdot;\delta_k)}(s_0)$ in the following proof.

If the instance is positive, Proposition 12 suggests that $\overline{V}^t \geq V^* > \mu_0$, suggesting that Algorithm 2 can only leave the loop at the Line 3 when (3) holds or $t \geq T_k^{\text{ee}}$. If $t \geq T_k^{\text{ee}}$ holds, $\hat{a} =$ Not Completed. If (3) holds, $\hat{a}$ must be $\overline{\pi}^t$, notice that

$$\begin{aligned}
&\mu_0 < \underline{V}^{t,\overline{\pi}^t} - (C-1)(\overline{V}^t - \underline{V}^{t,\overline{\pi}^t}) \\
\Leftrightarrow &\mu_0 < \underline{V}^{t,\overline{\pi}^t} + (\overline{V}^t - \underline{V}^{t,\overline{\pi}^t}) - C(\overline{V}^t - \underline{V}^{t,\overline{\pi}^t}) \\
\Leftrightarrow &C(\overline{V}^t - \underline{V}^{t,\overline{\pi}^t}) < \overline{V}^t - \mu_0. \qquad (33)
\end{aligned}$$

Meanwhile, Lemma 11 and Proposition 12 imply $V^{\overline{\pi}^t}, V^* \in [\underline{V}^{t,\overline{\pi}^t}, \overline{V}^t]$, suggesting

$$\begin{aligned}
&(\overline{V}^t - \mu_0) \leq (\overline{V}^t - \underline{V}^{t,\overline{\pi}^t}) + (V^{\overline{\pi}^t} - \mu_0) \qquad (34) \\
\Rightarrow &(\overline{V}^t - \mu_0) - (V^{\overline{\pi}^t} - \mu_0) \leq (\overline{V}^t - \underline{V}^{t,\overline{\pi}^t}). \qquad (35)
\end{aligned}$$

Replace $\overline{V}^t - \underline{V}^t$ in (33) with (35), we can conclude

$$\begin{aligned}
&C(\overline{V}^t - \underline{V}^t) < \overline{V}^t - \mu_0 \\
\Rightarrow &C\Big((\overline{V}^t - \mu_0) - (V^{\overline{\pi}^t} - \mu_0)\Big) < \overline{V}^t - \mu_0 \\
\Leftrightarrow &(C-1)(\overline{V}^t - \mu_0) < C(V^{\overline{\pi}^t} - \mu_0) \\
\Rightarrow &(C-1)(V^* - \mu_0) < C(V^{\overline{\pi}^t} - \mu_0) \\
\Leftrightarrow &\frac{C-1}{C}(V^* - \mu_0) < V^{\overline{\pi}^t} - \mu_0.
\end{aligned}$$

The last line must complete the proof for the case that the instance is positive.

If the instance is negative, Proposition 12 suggests that $\underline{V}^t \leq V^* < \mu_0$, suggesting that Algorithm 2 can only leave the loop at the Line 3 when (4) holds or $t \geq T_k^{\text{ee}}$. If $t \geq T_k^{\text{ee}}$, $\hat{a} =$ Not Completed. If (4) holds, $\hat{a} =$ None, which completes the proof for negative case. □

The following Lemma builds the suggests a necessary condition for Algorithm 2 to continue sampling.

Lemma 14. *For any round t during phase $k \geq \kappa$, meaning that $\mathcal{E}(\delta_k) \cap \mathcal{E}^{cnt}(\delta_k)$ holds, if Algorithm 2 doesn't terminate after sampling t-th episode,*

- *if the instance is positive, we have*

$$\frac{V_0^*(s_0) - \mu_0}{C} \leq \overline{V}_0^t(s_0; \delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0; \delta_k)$$

- *if the instance is negative, we have*

$$\mu_0 - V_0^*(s_0) \le \overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k)$$

Proof of Lemma 14. If the instance is positive, Algorithm 2 doesn't terminate after sampling t-th episode means

$$\begin{aligned}
&\underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) - (C-1)(\overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k)) \le \mu_0 \\
\Leftrightarrow & C(\overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k)) \ge \overline{V}_0^t(s_0;\delta_k) - \mu_0 \\
\Rightarrow & C(\overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k)) \ge V_0^*(s_0) - \mu_0 \\
\Leftrightarrow & \overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) \ge \frac{V_0^*(s_0)-\mu_0}{C}.
\end{aligned}$$

The second last line is by $\overline{V}_0^t(s_0;\delta_k) \ge V_0^*(s_0)$, guarantee by the $\mathcal{E}(\delta_k) \cap \mathcal{E}^{\text{cnt}}(\delta_k)$, Lemma 11 and Proposition 12.

If the instance is negative, Algorithm 2 doesn't terminate after sampling t-th episode means $\overline{V}_0^t(s_0;\delta_k) \ge \mu_0$, which is equivalent to

$$\begin{aligned}
&\overline{V}_0^t(s_0;\delta_k) \ge \mu_0 \\
\Leftrightarrow & \overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) \ge \mu_0 - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) \\
\Rightarrow & \overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) \ge \mu_0 - V_0^*(s_0).
\end{aligned}$$

The last line is by $V_0^*(s_0) \ge V_0^{\overline{\pi}^t(\cdot;\delta_k)}(s_0) \ge \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k)$, guarantee by the $\mathcal{E}(\delta_k) \cap \mathcal{E}^{\text{cnt}}(\delta_k)$, Lemma 11 and Proposition 12. □

To utilize Lemma 14 to prove an implicit upper bound for the length of a phse $k \ge \kappa$, we introduce the following Lemma 15. Lemma 15 can be considered an intermediate step for the proof of implicit upper bound in Lemma 18.

Lemma 15. *For any round t during phase $k \ge \kappa$, $\forall s, a, h$, we have*

$$\begin{aligned}
&\overline{Q}_h^t(s,a;\delta_k) - \underline{Q}_h^{t,\overline{\pi}^t(\cdot;\delta_k)}(s,a;\delta_k) \\
\le & 4(H-h)\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}}, 1\right\} + \sum_{s'} p_h(s'|s,a)\big(\overline{V}_{h+1}^t(s';\delta_k) - \underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s';\delta_k)\big).
\end{aligned}$$

Proof of Lema 15. By the definition of $\overline{Q}_h^t(s,a)$, $\underline{Q}_h^{t,\overline{\pi}^t(\cdot;\delta_k)}(s,a)$, we have

$$\begin{aligned}
&\overline{Q}_h^t(s,a) - \underline{Q}_h^{t,\overline{\pi}^t(\cdot;\delta_k)}(s,a) \\
=& \sum_{s'} \overline{p}_h^t(s'|s,a)\overline{V}_{h+1}^t(s';\delta_k) - \sum_{s'} \underline{p}_h^{t,\overline{\pi}^t}(s'|s,a)\underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s';\delta_k) \\
=& \sum_{s'} \overline{p}_h^t(s'|s,a)\overline{V}_{h+1}^t(s';\delta_k) - \sum_{s'} p_h(s'|s,a)\overline{V}_{h+1}^t(s';\delta_k) + \\
& \sum_{s'} p_h(s'|s,a)\overline{V}_{h+1}^t(s';\delta_k) - \sum_{s'} p_h(s'|s,a)\underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s';\delta_k) + \\
& \sum_{s'} p_h(s'|s,a)\underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s';\delta_k) - \sum_{s'} \underline{p}_h^{t,\overline{\pi}^t(s';\delta_k)}(s'|s,a)\underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s';\delta_k) \\
\le& \|\overline{p}_h^t(\cdot|s,a) - p_h(\cdot|s,a)\|_1(H-h) + \sum_{s'} p_h(s'|s,a)\big(\overline{V}_{h+1}^t(s';\delta_k) - \underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s';\delta_k)\big) + \\
& \|p_h(\cdot|s,a) - \underline{p}_h^{t,\overline{\pi}^t(s';\delta_k)}(\cdot|s,a)\|_1(H-h) \qquad (36) \\
\le& 4(H-h)\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}}, 1\right\} + \sum_{s'} p_h(s'|s,a)\big(\overline{V}_{h+1}^t(s';\delta_k) - \underline{V}_{h+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s';\delta_k)\big) \qquad (37)
\end{aligned}$$

Step (36) is by the fact that $\overline{V}^t_{h+1}(s';\delta_k), \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_{h+1}(s';\delta_k) \le H$. Step (37) is by the fact that $\|p(\cdot)-q(\cdot)\|_1 \le 2$, and the Pinsker's Inequality $\|p(\cdot)-q(\cdot)\|_1 \le \sqrt{\frac{1}{2}\mathrm{KL}(p,q)}$, as $k \ge \kappa$ suggests that $\overline{p}^t_h(\cdot|s,a), \underline{p}^{t,\overline{\pi}^t}_h(\cdot|s,a) \in \mathcal{C}^t_h(s,a;\delta_k)$. □

Given Lemma 15, we can immediately derive the following proposition by taking $a = \overline{\pi}^t_h(s;\delta_k)$.

Proposition 16. *For any round t during phase $k \ge \kappa$, $\forall s, a, h$, we have*

$$\begin{aligned}
&\overline{V}^t_h(s;\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_h(s;\delta_k) \\
\le& 4(H-h)\min\left\{\sqrt{\frac{\beta_p(n^t_h(s,\overline{\pi}^t_h(s)),\delta_k)}{n^t_h(s,\overline{\pi}^t_h(s))}},1\right\} + \\
&\sum_{s'} p_h(s'|s,\overline{\pi}^t_h(s;\delta_k))\big(\overline{V}^t_{h+1}(s';\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_{h+1}(s';\delta_k)\big).
\end{aligned}$$

By repeatedly apply Proposition 16, we can prove an upper bound for the $\overline{V}^t_0(s_0;\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_0(s_0;\delta_k)$, which is the following Lemma.

Lemma 17. *For any round t during phase $k \ge \kappa$, meaning that $\mathcal{E}(\delta) \cap \mathcal{E}^{cnt}(\delta)$ holds, we have*

$$\begin{aligned}
&\overline{V}^t_0(s_0;\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_0(s_0;\delta_k) \\
\le& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + 4\sum_{h=1}^{H}(H-h)\sum_{s',a'} p^{\overline{\pi}^t}_h(s',a')\min\left\{\sqrt{\frac{\beta_p(n^t_h(s',a'),\delta_k)}{n^t_h(s',a')}},1\right\}
\end{aligned}$$

Proof of Lemma 17. Given $\overline{\pi}^t(\cdot;\delta_k)$ and the definition of $p^{\overline{\pi}^t(\cdot;\delta_k)}_h(s',a') := \Pr(S_{t,h} = s, \overline{\pi}^t_h(s;\delta_k) = a)$.

We use induction to prove

$$\begin{aligned}
&\overline{V}^t_0(s_0;\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_0(s_0;\delta_k) \\
\le& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + 4\sum_{h=1}^{h'-1}(H-h)\sum_{s',a'} p^{\overline{\pi}^t}_h(s',a')\min\left\{\sqrt{\frac{\beta_p(n^t_h(s',a'),\delta_k)}{n^t_h(s',a')}},1\right\} + \\
&\sum_{s',a'} p^{\overline{\pi}^t}_{h'}(s',a')\big(\overline{V}^t_{h'}(s';\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_{h'}(s';\delta_k)\big) \qquad (38)
\end{aligned}$$

holds for all $h' = 1, \cdots, H+1$.

For $h' = 1$, by the Proposition 16, we know

$$\begin{aligned}
&\overline{V}^t_0(s_0;\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_0(s_0;\delta_k) \\
\le& 4H\min\left\{\sqrt{\frac{\beta_p(n^t_h(s_0,a_0),\delta_k)}{n^t_h(s_0,a_0)}},1\right\} + \sum_{s'} p_0(s')\big(\overline{V}^t_1(s';\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_1(s';\delta_k)\big) \\
=& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + \sum_{s'} p_0(s')\big(\overline{V}^t_1(s';\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_1(s';\delta_k)\big) \\
=& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + \sum_{s',a'} p_0(s')\mathbb{1}(a' = \overline{\pi}^t_1(s'))\big(\overline{V}^t_1(s';\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_1(s';\delta_k)\big) \\
=& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + \sum_{s',a'} p^{\overline{\pi}^t}_1(s',a')\big(\overline{V}^t_1(s';\delta_k) - \underline{V}^{t,\overline{\pi}^t(\cdot;\delta_k)}_1(s';\delta_k)\big).
\end{aligned}$$

The third line is by $n^t_h(s_0,a_0) = t$ holds for all t, the last line is by the definition $p^{\overline{\pi}^t}_1(s,a) = \Pr(S_1 = s, \overline{\pi}^t_h(s) = a)$.

Given (38) holds for h', apply the Proposition 16 again, we can derive

$$
\begin{aligned}
&\overline{V}_0^t(s_0;\delta_k) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_k) \\
\leq& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + 4\sum_{h=1}^{h'-1}(H-h)\sum_{s',a'} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s',a')\min\left\{\sqrt{\frac{\beta_p(n_h^t(s',a'),\delta_k)}{n_h^t(s',a')}},1\right\} + \\
&\sum_{s',a'} p_{h'}^{\overline{\pi}^t(\cdot;\delta_k)}(s',a')\left(4(H-h')\min\left\{\sqrt{\frac{\beta_p(n_{h'}^t(s',\overline{\pi}_{h'}^t(s';\delta_k)),\delta_k)}{n_{h'}^t(s',\overline{\pi}_{h'}^t(s';\delta_k))}},1\right\} + \right. \\
&\left.\sum_{s''} p_{h'}(s''|s',\overline{\pi}_{h'}^t(s';\delta_k))\big(\overline{V}_{h'+1}^t(s'';\delta_k) - \underline{V}_{h'+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s'';\delta_k)\big)\right) \\
=& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + 4\sum_{h=1}^{h'}(H-h)\sum_{s',a'} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s',a')\min\left\{\sqrt{\frac{\beta_p(n_h^t(s',a'),\delta_k)}{n_h^t(s',a')}},1\right\} + \\
&\sum_{s',a'}\sum_{s''} p_{h'}^{\overline{\pi}^t(\cdot;\delta_k)}(s',a')p_{h'}(s''|s',\overline{\pi}_{h'}^t(s';\delta_k))\big(\overline{V}_{h'+1}^t(s'';\delta_k) - \underline{V}_{h'+1}^{t,\overline{\pi}^t}(s'';\delta_k)\big) \\
=& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + 4\sum_{h=1}^{h'}(H-h)\sum_{s',a'} p_h^{\overline{\pi}^t}(s',a')\min\left\{\sqrt{\frac{\beta_p(n_h^t(s',a'),\delta_k)}{n_h^t(s',a')}},1\right\} + \\
&\sum_{s'}\sum_{s''} p_{h'}^{\overline{\pi}^t(\cdot;\delta_k)}(s',\overline{\pi}_{h'}^t(s';\delta_k))p_{h'}(s''|s',\overline{\pi}_{h'}^t(s';\delta_k))\big(\overline{V}_{h'+1}^t(s'';\delta_k) - \underline{V}_{h'+1}^{t,\overline{\pi}^t}(s'';\delta_k)\big) \\
=& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + 4\sum_{h=1}^{h'}(H-h)\sum_{s',a'} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s',a')\min\left\{\sqrt{\frac{\beta_p(n_h^t(s',a'),\delta_k)}{n_h^t(s',a')}},1\right\} + \\
&\sum_{s''}\Pr_{\overline{\pi}^t(\cdot;\delta_k)}(S_{h'+1}=s'')\big(\overline{V}_{h'+1}^t(s'';\delta_k) - \underline{V}_{h'+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s'';\delta_k)\big) \\
=& 4H\min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}},1\right\} + 4\sum_{h=1}^{h'}(H-h)\sum_{s',a'} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s',a')\min\left\{\sqrt{\frac{\beta_p(n_h^t(s',a'),\delta_k)}{n_h^t(s',a')}},1\right\} + \\
&\sum_{s',a'} p_{h'+1}^{\overline{\pi}^t(\cdot;\delta_k)}(s',a')\big(\overline{V}_{h'+1}^t(s'';\delta_k) - \underline{V}_{h'+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s'';\delta_k)\big),
\end{aligned}
$$

which completed the induction and (38) holds for all $h' = 1, 2, \cdots, H+1$.

By taking $h' = H$ and utilizing the fact that $\overline{V}_{H+1}^t(s;\delta_k) = 0, \underline{V}_{H+1}^{t,\overline{\pi}^t(\cdot;\delta_k)}(s;\delta_k) = 0, \forall s$, we complete the proof of Lemma 17. □

Definition 7. *Denote t_{k-1} as the starting t index of phase k for phase index $k \geq 1$. For $k \geq 2$, t_{k-1} is a random variable. For consistency, we set $t_0 = 0$.*

What's more, we know ¬(3) and ¬(4) must hold for episode index $t = t_{k-1}, t_{k-1}+1, \cdots, t_k - 1$. It is indeed possible that $t_{k-1} = t_k$ for some sampling path $\{\{S_{t,h}, A_{t,h}\}_{h=1}^H\}_{t=1}^{+\infty}$ and index $k \in \mathbb{N}$. In this case, the history $\mathcal{H}^{\text{ee}}$ with length $|\mathcal{H}^{\text{ee}}| = t_{k-1}$ guarantees that $\underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_{k+1}) - (C-1)(\overline{V}_0^t(s_0;\delta_{k+1}) - \underline{V}_0^{t,\overline{\pi}^t(\cdot;\delta_k)}(s_0;\delta_{k+1})) > \mu_0$ or $\overline{V}_0^t(s_0;\delta_{k+1}) < \mu_0$. Following this idea, we can combine Lemma 14 and 17 to derive upper bounds for the length of each phase.

Lemma 18. *Consider any phase index $k \geq \kappa$. If the instance is positive, we have*

$$
\begin{aligned}
&(t_k - t_{k-1})\frac{V_0^*(s_0) - \mu}{C} \\
\leq& \sum_{t=t_{k-1}}^{t_k-1}\sum_{h=0}^{H}\sum_{(s,a)} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s,a)4(H-h)\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}},1\right\}.
\end{aligned}
$$

If the instance is negative, we have

$$(t_k - t_{k-1})(\mu - V_0^*(s_0)) \leq \sum_{t=t_{k-1}}^{t_k-1} \sum_{h=0}^{H} \sum_{(s,a)} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s,a) 4(H-h) \min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}}, 1\right\}.$$

Proof of Lemma 18. If the instance is positive, by Lemma 14 and 17, we can conclude

$$\frac{V_0^*(s_0) - \mu_0}{C} \leq 4H \min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}}, 1\right\} + 4\sum_{h=1}^{H}(H-h) \sum_{s',a'} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s',a') \min\left\{\sqrt{\frac{\beta_p(n_h^t(s',a'),\delta_k)}{n_h^t(s',a')}}, 1\right\}$$

holds for all $t = t_{k-1}, t_{k-1}+1, \cdots .t_k - 1$. Summing up the above inequalities for $t = t_{k-1}, t_{k-1}+1, \cdots t_k - 1$ leads to the inequality for the positive case.

If the instance is negative, by Lemma 14 and 17, we can conclude

$$\mu_0 - V_0^*(s_0) \leq 4H \min\left\{\sqrt{\frac{\beta_p(t,\delta_k)}{t}}, 1\right\} + 4\sum_{h=1}^{H}(H-h) \sum_{s',a'} p_h^{\overline{\pi}^t(\cdot;\delta_k)}(s',a') \min\left\{\sqrt{\frac{\beta_p(n_h^t(s',a'),\delta_k)}{n_h^t(s',a')}}, 1\right\}$$

holds for all $t = t_{k-1}, t_{k-1}+1, \cdots .t_k - 1$. Summing up the above inequalities for $t = t_{k-1}, t_{k-1}+1, \cdots t_k - 1$ leads to the inequality for the negative case. □

To derive an upper bound for t_k, we need are going to prove an upper bound for the right-hand side of inequalities in Lemma 18. The following Lemma will be useful.

Lemma 19 (Lemma 7 in Kaufmann et al. (2021))**.** *For any round t during phase $k \geq \kappa$, any pair (s,a,h)*

$$\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}}, 1\right\} \leq 2\sqrt{\frac{\beta_p(\bar{n}_h^t(s,a),\delta_k)}{\max\{\bar{n}_h^t(s,a),1\}}}$$

holds for all $t < \tau$.

Proof of Lemma 19. Recall the definition $\beta^{\text{cnt}}(\delta) = \log\frac{2SAH}{\delta}$, $\beta_p(t,\delta) = \log\frac{2SAH}{\delta} + (S-1)\log\left(e(1+\frac{t}{S-1})\right)$. $k \geq \kappa$ suggests that $n_h^t(s,a) \geq \frac{1}{2}\bar{n}_h^t(s,a) - \beta^{\text{cnt}}(\delta_k)$ holds for all s,a,h.

We prove Lemma 19 in two cases, $\beta^{cnt}(\delta_k) \leq \frac{1}{4}\bar{n}_h^t(s,a)$ or $\beta^{cnt}(\delta_k) > \frac{1}{4}\bar{n}_h^t(s,a)$.

First, if $\beta^{cnt}(\delta_k) \leq \frac{1}{4}\bar{n}_h^t(s,a)$, we have $n_h^t(s,a) \geq \frac{1}{4}\bar{n}_h^t(s,a)$. Notice that $\beta_p(t,\delta) = \log\frac{2SAH}{\delta} + (S-1)\log\left(S-1+t\right) + (S-1)\log\frac{e}{S-1}$, we can derive its partial derivative regarding t,

$$\frac{\partial\frac{\beta_p(t,\delta)}{t}}{\partial t} = -\frac{\log\frac{2SAH}{\delta}}{t^2} + \frac{(S-1)\left(\frac{1}{S-1+t} - \log(S-1+t)\right)}{t^2}$$

For $t \geq 1$, $\frac{1}{S-1+t} - \log(S-1+t) \leq \frac{1}{S} - \log S < 0$ holds for $S \geq 2$. We can conclude $\frac{\partial\frac{\beta_p(t,\delta)}{t}}{\partial t} < 0$ for $t \geq 1$. Since $\beta^{cnt}(\delta_k) > 1$ holds for all k, we can conclude $\frac{1}{4}\bar{n}_h^t(s,a) > 1$. Together with $n_h^t(s,a) \geq \frac{1}{4}\bar{n}_h^t(s,a)$, we have

$$\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}}, 1\right\} \leq \sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}} \leq \sqrt{\frac{\beta_p(\frac{1}{4}\bar{n}_h^t(s,a),\delta_k)}{\frac{1}{4}\bar{n}_h^t(s,a)}} \leq 2\sqrt{\frac{\beta_p(\bar{n}_h^t(s,a),\delta_k)}{\bar{n}_h^t(s,a)}}.$$

Again, by the fact that $\frac{4\beta_p(\bar{n}_h^t(s,a),\delta_k)}{\bar{n}_h^t(s,a)} = \frac{4\beta_p(\bar{n}_h^t(s,a),\delta_k)}{\max\{\bar{n}_h^t(s,a),1\}}$, we can conclude

$$\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}},1\right\} \le 2\sqrt{\frac{\beta_p(\bar{n}_h^t(s,a),\delta_k)}{\max\{\bar{n}_h^t(s,a),1\}}},$$

which suggests the Lemma holds.

Second, if $\beta^{cnt}(\delta_k) > \frac{1}{4}\bar{n}_h^t(s,a)$, we have $\frac{4\beta^{cnt}(\delta_k)}{\bar{n}_h^t(s,a)} \ge 1$. Meanwhile, $\beta^{cnt}(\delta_k) \ge 1$ by the definition, suggesting that

$$\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}},1\right\} \le 1 \le \sqrt{\frac{4\beta^{cnt}(\delta_k)}{\max\{\bar{n}_h^t(s,a),1\}}} = 2\sqrt{\frac{\beta^{cnt}(\delta_k)}{\max\{\bar{n}_h^t(s,a),1\}}} \tag{39}$$

By the definition of β_p and β^{cnt}, we know $\beta_p(t,\delta_k) > \beta^{\text{cnt}}(\delta_k)$ holds for all $t > 0$. Together with (39), we conclude

$$\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}},1\right\} \le 2\sqrt{\frac{\beta^{cnt}(\delta_k)}{\max\{\bar{n}_h^t(s,a),1\}}} < 2\sqrt{\frac{\beta(\bar{n}_h^t(s,a),\delta_k)}{\max\{\bar{n}_h^t(s,a),1\}}},$$

which means the Lemma still holds. □

Given all the above Lemmas, we are ready to prove a main theorem, showing that Algorithm 2 is able to output the correct prediction for large enough phase index k, with an upper bound of t_k.

Theorem 20. *Apply Algorithm 2 on an instance ν.*

If ν is positive, for any phase index $k \ge \max\{\kappa, \lceil\log_2 \frac{9600(H+1)^2}{\epsilon^{pos2}}\rceil\}$, Algorithm 2 will output a policy $\hat{\pi}_k$ satisfying

$$V_0^{\hat{\pi}_k}(s_0) - \mu_0 \ge \frac{C-1}{C}\Big(V_0^*(s_0) - \mu_0\Big)$$

at the end of phase k, with

$$\begin{aligned} t_k \le & \frac{101(H+1)^4SA\log\frac{2SAH}{\delta_k}}{\epsilon^{pos2}} + T_{\kappa-1}^{ee} + \\ & \frac{400(H+1)^4S^2A}{\epsilon^{pos2}}\log\left(\frac{402(H+1)^4S^2A\log\frac{2SAH}{\delta_k}}{\epsilon^{pos2}}\right) + \frac{400(H+1)^4S^2A}{\epsilon^{pos2}}\log\left(2T_{\kappa-1}^{ee}\right). \end{aligned} \tag{40}$$

If ν is negative, for any phase index $k \ge \max\{\kappa, \lceil\log_2 \frac{9600(H+1)^2}{\epsilon^{neg2}}\rceil\}$, Algorithm 2 will output None *at the end of phase k, with*

$$\begin{aligned} t_k \le & \frac{101(H+1)^4SA\log\frac{2SAH}{\delta_k}}{\epsilon^{neg2}} + T_{\kappa-1}^{ee} + \\ & \frac{400(H+1)^4S^2A}{\epsilon^{neg2}}\log\left(\frac{402(H+1)^4S^2A\log\frac{2SAH}{\delta_k}}{\epsilon^{neg2}}\right) + \frac{400(H+1)^4S^2A}{\epsilon^{neg2}}\log\left(2T_{\kappa-1}^{ee}\right) \end{aligned} \tag{41}$$

Here $\epsilon^{pos} = \frac{V_0^*(s_0)-\mu_0}{C}$, and $\epsilon^{neg} = \mu_0 - V_0^*(s_0)$.

Proof of Theorem 20. We temporarily focus on the positive case. The analysis for the negative case is similar.

From Lemma 18, we know

$$(t_k - t_{k-1})\epsilon^{\text{pos}} \le \sum_{t=t_{k-1}}^{t_k-1}\sum_{h=0}^{H}\sum_{(s,a)} p_h^{\overline{\pi}^t}(s,a)4(H-h)\min\left\{\sqrt{\frac{\beta_p(n_h^t(s,a),\delta_k)}{n_h^t(s,a)}},1\right\} \tag{42}$$

holds for all $k \geq \kappa$. Recall we denote t_{k-1} as the starting t index of stage k, we first show (40) holds for $k \geq \kappa$. In the following, we will consider consider multiple phase indexes $\ell = \kappa, \kappa+1, \cdots, k$. For simplicity, we will use $\overline{\pi}^t$ as the shorthand of $\overline{\pi}^t(\cdot; \delta_\ell)$ in the following calculation.

By summing up (42) for stage index $\ell = \kappa, \kappa+1, \cdots, k$, We have

$$
\begin{aligned}
&(t_k - t_{\kappa-1})\epsilon^{\text{pos}} \\
\leq &\sum_{\ell=\kappa}^{k} \sum_{t=t_{\ell-1}}^{t_\ell - 1} \sum_{h=0}^{H} \sum_{(s,a)} p_h^{\overline{\pi}^t(\cdot;\delta_\ell)}(s,a) 4(H-h) \min\left\{ \sqrt{\frac{\beta_p(n_h^t(s,a),\delta_\ell)}{n_h^t(s,a)}}, 1 \right\} \\
\overset{\text{Lemma 19}}{\leq} &\sum_{\ell=\kappa}^{k} \sum_{t=t_{\ell-1}}^{t_\ell - 1} \sum_{h=0}^{H} \sum_{(s,a)} p_h^{\overline{\pi}^t}(s,a) 8(H-h) \sqrt{\frac{\beta_p(\bar{n}_h^t(s,a),\delta_\ell)}{\max\{\bar{n}_h^t(s,a),1\}}} \\
\leq &8\sqrt{\beta_p(t_k,\delta_k)} \sum_{h=0}^{H} (H-h) \sum_{(s,a)} \sum_{\ell=\kappa}^{k} \sum_{t=t_{\ell-1}}^{t_\ell-1} p_h^{\overline{\pi}^t}(s,a) \sqrt{\frac{1}{\max\{\bar{n}_h^t(s,a),1\}}} \\
= &8\sqrt{\beta_p(t_k,\delta_k)} \sum_{h=0}^{H} (H-h) \sum_{(s,a)} \sum_{\ell=\kappa}^{k} \sum_{t=t_{\ell-1}}^{t_\ell-1} \\
&p_h^{\overline{\pi}^t}(s,a) \sqrt{\frac{1}{\max\{\sum_{v=0}^{t_{\kappa-1}-1} p_h^{\overline{\pi}^v}(s,a) + \sum_{v=t_{\kappa-1}}^{t-1} p_h^{\overline{\pi}^v}(s,a), 1\}}} \\
= &8\sqrt{\beta_p(t_k,\delta_k)} \sum_{h=0}^{H} (H-h) \sum_{(s,a)} \sum_{t=t_{\kappa-1}}^{t_k-1} \\
&p_h^{\overline{\pi}^t}(s,a) \sqrt{\frac{1}{\max\{\sum_{v=0}^{t_{\kappa-1}-1} p_h^{\overline{\pi}^v}(s,a) + \sum_{v=t_{\kappa-1}}^{t-1} p_h^{\overline{\pi}^v}(s,a), 1\}}} \\
\overset{\text{Lemma 41}}{\leq} &8(1+\sqrt{2})\sqrt{\beta_p(t_k,\delta_k)} \sum_{h=0}^{H} (H-h) \sum_{(s,a)} \sqrt{\sum_{v=t_{\kappa-1}}^{t_k-1} p_h^{\overline{\pi}^v}(s,a)}.
\end{aligned} \tag{43}
$$

To get the last step, we apply Lemma 41 by taking $M = \sum_{v=0}^{t_{\kappa-1}-1} p_h^v(s,a)$. By the Cauchy-Schwarz Inequality,

$$
\sum_{(s,a)} \sqrt{\sum_{v=t_{\kappa-1}}^{t_k-1} p_h^v(s,a)} \leq \sqrt{\left(\sum_{s,a} \sum_{v=t_{\kappa-1}}^{t_k-1} p_h^v(s,a) \right) \left(\sum_{(s,a)} 1 \right)} = \sqrt{SA(t_k - t_{\kappa-1})}.
$$

Following (43), we conclude

$$
\begin{aligned}
&(t_k - t_{\kappa-1})\epsilon^{\text{pos}} \\
\leq &8(1+\sqrt{2})\sqrt{SA\beta_p(t_k,\delta_k)} \sum_{h=0}^{H} (H-h)\sqrt{t_k - t_{\kappa-1}} \\
= &4(1+\sqrt{2})\sqrt{SA\beta_p(t_k,\delta_k)} H(H+1)\sqrt{t_k - t_{\kappa-1}}.
\end{aligned} \tag{44}
$$

By the definition $\beta_p(t,\delta) = \log\frac{2SAH}{\delta} + (S-1)\log\left(e(1+\frac{t}{S-1})\right)$, (44) suggests that

$$
\begin{aligned}
&(t_k - t_{\kappa-1})\epsilon^{\text{pos}} \le 4(1+\sqrt{2})H(H+1)\sqrt{\log\frac{2SAH}{\delta_k} + S\log(et_k)}\sqrt{SA}\sqrt{t_k - t_{\kappa-1}} \\
\Leftrightarrow &(t_k - t_{\kappa-1})\epsilon^{\text{pos}2} \le 100(H+1)^4 SA\log\frac{2SAH}{\delta_k} + 100(H+1)^4 S^2 A\log(et_k) \\
\Leftrightarrow &t_k \le \frac{100(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + t_{\kappa-1} + \frac{100(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}} + \frac{100(H+1)^4 S^2 A\log t_k}{\epsilon^{\text{pos}2}} \\
\overset{\text{Lemma 38}}{\Rightarrow} &t_k \le \frac{100(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + t_{\kappa-1} + \frac{100(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}} + \\
&\frac{300(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(\frac{100(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + t_{\kappa-1} + \frac{100(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\right) \\
\Rightarrow &t_k \le \frac{100(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + t_{\kappa-1} + \\
&\frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(\frac{200(H+1)^4 S^2 A\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + t_{\kappa-1}\right) \\
\overset{\text{Lemma 6}}{\Rightarrow} &t_k \le \frac{100(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + T^{\text{ee}}_{\kappa-1} + 1 \\
&\frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(\frac{200(H+1)^4 S^2 A\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + T^{\text{ee}}_{\kappa-1} + 1\right) \\
\overset{\text{Lemma 39}}{\Rightarrow} &t_k \le \frac{101(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + T^{\text{ee}}_{\kappa-1} + \\
&\frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(\frac{402(H+1)^4 S^2 A\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}}\right) + \frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right).
\end{aligned}
$$

For phase index $k \ge \max\{\kappa, \lceil\log_2\frac{9600(H+1)^2}{\epsilon^{\text{pos}2}}\rceil\}$, by the Lemma 21, we can conclude

$$
\begin{aligned}
&T^{\text{ee}}_k \\
\ge &\frac{101(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + T^{\text{ee}}_{k-1} + \\
&\frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(\frac{402(H+1)^4 S^2 A\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}}\right) + \frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(2T^{\text{ee}}_{k-1}\right) \\
\ge &\frac{101(H+1)^4 SA\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}} + T^{\text{ee}}_{\kappa-1} + \\
&\frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(\frac{402(H+1)^4 S^2 A\log\frac{2SAH}{\delta_k}}{\epsilon^{\text{pos}2}}\right) + \frac{400(H+1)^4 S^2 A}{\epsilon^{\text{pos}2}}\log\left(2T^{\text{ee}}_{\kappa-1}\right).
\end{aligned}
$$

That means Algorithm 2 will fulfill (3) and quit the loop at Line 3. By the Lemma 13, we can conclude Algorithm 2 will output a policy $\hat{\pi}_k$ satisfying $V_0^{\hat{\pi}_k}(s_0) - \mu_0 \ge \frac{C-1}{C}\left(V_0^*(s_0) - \mu_0\right)$ at the end of phase k.

For the case of negative instance, the analysis is nearly the same. We can replace ϵ^{pos} with $\epsilon^{\text{neg}} = \mu_0 - V_0^*(s_0)$. By the Lemma 13, we complete the proof. □

The upcoming Lemma 21 shows the minimum phase index for sufficiently large T^{ee}_k..

Lemma 21. *For any $\epsilon > 0$, $k \geq \lceil \log_2 \frac{9600(H+1)^2}{\epsilon^2} \rceil$ implies*

$$\begin{aligned}
&T_k^{ee} \\
\geq &\frac{101(H+1)^4 SA \log \frac{2SAH}{\delta_k}}{\epsilon^2} + T_{k-1}^{ee} + \\
&\frac{400(H+1)^4 S^2 A}{\epsilon^2} \log\left(\frac{402(H+1)^4 S^2 A \log \frac{2SAH}{\delta_k}}{\epsilon^2}\right) + \frac{400(H+1)^4 S^2 A}{\epsilon^2} \log\left(2T_{k-1}^{ee}\right).
\end{aligned}$$

Proof. We can validate the lemma by conducting direct calculation. Recall the definition of T_k^{ee},

$$T_k^{\text{ee}} = \frac{(H+1)^2 SA \log \frac{2SAH}{\delta_k}}{\epsilon_k} + \frac{(H+1)^2 S^2 A}{\epsilon_k} \log\left(\frac{(H+1)^2 S^2 A \log \frac{2SAH}{\delta_k}}{\epsilon_k}\right),$$

and we take $\epsilon_k = \frac{1}{2^k}$. We can conclude $k \geq \lceil \log_2 \frac{9600(H+1)^2}{\epsilon^2} \rceil$ implies

$$T_k^{\text{ee}} \geq \frac{404(H+1)^4 SA \log \frac{2SAH}{\delta_k}}{\epsilon^2} + \frac{3200(H+1)^4 S^2 A}{\epsilon^2} \log\left(\frac{9600(H+1)^4 S^2 A \log \frac{2SAH}{\delta_k}}{\epsilon^2}\right).$$

We can further conclude

$$\frac{1}{4} T_k^{\text{ee}} \geq \frac{101(H+1)^4 SA \log \frac{2SAH}{\delta_k}}{\epsilon^2} + \frac{400(H+1)^4 S^2 A}{\epsilon^2} \log\left(\frac{402(H+1)^4 S^2 A \log \frac{2SAH}{\delta_k}}{\epsilon^2}\right) \tag{45}$$

$$2T_k^{\text{ee}} \geq \frac{2 \cdot 3200(H+1)^4 S^2 A}{\epsilon^2} \log\left(\frac{3 \cdot 3200(H+1)^4 S^2 A}{\epsilon^2}\right) \tag{46}$$

By the Lemma 40, (46) implies

$$\begin{aligned}
&2T_k^{\text{ee}} \geq \frac{2 \cdot 3200(H+1)^4 S^2 A}{\epsilon^2} \log\left(\frac{3 \cdot 3200(H+1)^4 S^2 A}{\epsilon^2}\right) \\
\Rightarrow &2T_k^{\text{ee}} \geq \frac{3200(H+1)^4 S^2 A}{\epsilon^2} \log\left(2T_k^{\text{ee}}\right) \\
\Rightarrow &\frac{1}{4} T_k^{\text{ee}} \geq \frac{400(H+1)^4 S^2 A}{\epsilon^2} \log\left(2T_{k-1}^{\text{ee}}\right).
\end{aligned} \tag{47}$$

Meanwhile, $\epsilon_k = \frac{1}{2^k}$ implies

$$\frac{1}{2} T_k^{\text{ee}} \geq T_{k-1}^{\text{ee}}. \tag{48}$$

Sum up (45),(47) and (48), we have

$$\begin{aligned}
&T_k^{\text{ee}} \\
\geq &\frac{101(H+1)^4 SA \log \frac{2SAH}{\delta_k}}{\epsilon^2} + T_{k-1}^{\text{ee}} + \\
&\frac{400(H+1)^4 S^2 A}{\epsilon^2} \log\left(\frac{402(H+1)^4 S^2 A \log \frac{2SAH}{\delta_k}}{\epsilon^2}\right) + \frac{400(H+1)^4 S^2 A}{\epsilon^2} \log\left(2T_{k-1}^{\text{ee}}\right).
\end{aligned}$$

□

B.4 Performance Guarantee of Exploitation Stage

In this subsection, we present the performance Guarantee of the exploitation stage, i.e. Line 11 to 17 in Algorithm 1.

The following Lemma shows the minimum phase index for sufficiently large T_k^{et}.

Lemma 22.

$$k \geq \lceil \log_2 \frac{H^2}{\omega^2 (V_0^*(s_0) - \mu_0)^2} \rceil$$
$$\Rightarrow 100 \cdot \frac{\log \frac{\alpha_k}{\delta} + \log\log \frac{1}{\epsilon_k}}{\epsilon_k} \geq \frac{28H^2 \log \frac{\alpha_k}{\delta} + 16H^2 \log\left(\log\left(\frac{24H^2}{\omega^2 (V_0^*(s_0) - \mu_0)^2}\right)\right)}{\omega^2 (V_0^*(s_0) - \mu_0)^2},$$

where ω is an arbitrary constant in $(0, 1)$.

Proof. We can validate Lemma 22 by straight forward calculation.

$$k \geq \lceil \log_2 \frac{H^2}{\omega^2 (V_0^*(s_0) - \mu_0)^2} \rceil$$
$$\Rightarrow \frac{1}{\epsilon_k} \geq \frac{H^2}{\omega^2 (V_0^*(s_0) - \mu_0)^2}$$
$$\Rightarrow 100 \cdot \frac{\log \frac{\alpha_k}{\delta} + \log\log \frac{1}{\epsilon_k}}{\epsilon_k} > \frac{28H^2 \log \frac{\alpha_k}{\delta} + 16H^2 \log\left(\log\left(\frac{24H^2}{\omega^2 (V_0^*(s_0) - \mu_0)^2}\right)\right)}{\omega^2 (V_0^*(s_0) - \mu_0)^2}$$

□

The upcoming Lemma 23, 24 are to prove sufficient conditions that Algorithm 1 will accept a candidate policy $\hat{\pi}_k$.

Lemma 23. *Given a policy $\hat{\pi}$, denote $\hat{V}_0^{\hat{\pi},N}$ as the empirical estimation of $V_0^{\hat{\pi}}(s_0)$ by executing $\hat{\pi}$ N times, we have*

$$\Pr\left(\forall N \in \mathbb{N}, |\hat{V}_0^{\hat{\pi},N} - V_0^{\hat{\pi}}(s_0)| < \sqrt{\frac{H^2 \log \frac{2(\log_2 2N)^2}{\delta}}{N}}\right) \geq 1 - \frac{\pi^2}{6}\delta$$

holds for all $\delta > 0$.

Proof. Denote X_i as the total random reward collected by executing the policy $\hat{\pi}$ in i th time. We have $\hat{V}_0^{\hat{\pi},N} = \frac{\sum_{i=1}^N X_i}{N}$. From the assumption that the reward is always bounded in $[0, 1]$, we know X_i is $\frac{H^2}{4}$-subgaussian. By the Lemma 43, we complete the proof. □

Lemma 24. *Consider a positive instance with $V_0^*(s_0) > \mu_0$. Assume policy $\hat{\pi}_k$ satisfies $V_0^{\hat{\pi}_k}(s_0) - \mu_0 \geq \omega\Big(V_0^*(s_0) - \mu_0\Big)$ for some $\omega > 0$ and $k \geq \lceil \log_2 \frac{H^2}{\omega^2 (V_0^*(s_0) - \mu_0)^2} \rceil$. If the concentration event*

$$\mathcal{E}_k^{et} = \left\{\forall N \in \mathbb{N}, |\hat{V}_0^{\hat{\pi}_k,N} - V_0^{\hat{\pi}_k}(s_0)| < \sqrt{\frac{H^2 \log \frac{2\alpha_k(\log_2 2N)^2}{\delta}}{N}}\right\}$$

holds, where $\hat{V}_0^{\hat{\pi}_k,N}$ is the empirical mean value of $V_0^{\hat{\pi}_k}(s_0)$ by executing $\hat{\pi}_k$ N times, Line 11 to 17 in Algorithm 1 will enter Line 15 and output $\hat{\pi}_k$.

Proof. Given the event $\mathcal{E}_k^{\text{et}}$, we have

$$\hat{V}_0^{\hat{\pi}_k,N} - \sqrt{\frac{H^2 \log \frac{2\alpha_k(\log_2 2N)^2}{\delta}}{N}} < V_0^{\hat{\pi}_k}(s_0) < \hat{V}_0^{\hat{\pi}_k,N} + \sqrt{\frac{H^2 \log \frac{2\alpha_k(\log_2 2N)^2}{\delta}}{N}}$$

holds for all $N \in \mathbb{N}$. Thus, we can conclude

$$\begin{aligned}
\Rightarrow N \geq & \frac{28H^2 \log \frac{\alpha_k}{\delta} + 16H^2 \log\left(\log\left(\frac{24H^2}{\omega^2(V_0^*(s_0)-\mu_0)^2}\right)\right)}{\omega^2(V_0^*(s_0)-\mu_0)^2} \\
\Rightarrow N \geq & \frac{28H^2 \log \frac{\alpha_k}{\delta} + 16H^2 \log\left(\log\left(\frac{24H^2}{(V_0^{\hat{\pi}_k}(s_0)-\mu_0)^2}\right)\right)}{(V_0^{\hat{\pi}_k}(s_0)-\mu_0)^2} \\
\Rightarrow & V_0^{\hat{\pi}_k}(s_0) - 2\sqrt{\frac{H^2 \log \frac{2\alpha_k(\log_2 2N)^2}{\delta}}{N}} > \mu_0 \\
\Rightarrow & \hat{V}_0^{\hat{\pi}_k, N} - \sqrt{\frac{H^2 \log \frac{2\alpha_k(\log_2 2N)^2}{\delta}}{N}} > \mu_0
\end{aligned}$$

The second last step is by the Lemma 42 and the last step is by the event $\mathcal{E}_k^{\text{et}}$. Meanwhile, by the Lemma 22, we know $100H^2 \cdot \frac{\log \frac{\alpha_k}{\delta} + \log\log \frac{1}{\epsilon_k}}{\epsilon_k} > \frac{28H^2 \log \frac{\alpha_k}{\delta} + 16H^2 \log\left(\log\left(\frac{24H^2}{(V_0^{\hat{\pi}_k}(s_0)-\mu_0)^2}\right)\right)}{(V_0^{\hat{\pi}_k}(s_0)-\mu_0)^2}$, that means the Algorithm 1 will enter Line 15 and output $\hat{\pi}_k$.. □

C Proof of Lower Bound

C.1 Full Definition of Instances

In this subsection, we present the full definition of Uniform Instance and Tree Instance.

Definition 8 (Uniform Instance specified by (S, A, H, r, ϵ))**.** *Given integers $S, H \geq 10$, $A > 100$, H is even, $(S-2)H$ is divisible by 16, $0 < \epsilon < \frac{1}{200\sqrt{5}}$, $\frac{1}{4} < r < r + \epsilon < \frac{3}{4}$, we define $\mathcal{S} = [S-2] \cup \{s_{good}, s_{bad}\}, \mathcal{A} = [A], \mathcal{H} = [H]$. Define a Meaningless instance ν with reward function $r_h(s,a) = 0, \forall (h,s,a) \in [H] \times [S-2] \times \mathcal{A}$, $r_h(s_{good}, a) = 1, \forall (h,a) \in [H] \times \mathcal{A}$, $r_h(s_{bad}, a) = 0, \forall (h,a) \in [H] \times \mathcal{A}$; initial distribution $p_0(s) = \frac{1}{S-2}, \forall s \in [S-2]$, $p_0(s) = 0, \forall s \in \{s_{good}, s_{bad}\}$; transition probability $p_h(s'|s,1) = \begin{cases} r+\epsilon & s' = s_{good} \\ 1-r-\epsilon & s' = s_{bad} \\ 0 & else \end{cases}, \forall (h,s) \in [H] \times [S-2]$; $p_h(s'|s,a) = \begin{cases} r & s' = s_{good} \\ 1-r & s' = s_{bad} \\ 0 & else \end{cases}, \forall (h,s) \in [H] \times [S-2], a \in \{2,3,\cdots,A\}$; $p_h(s'|s,a) = \begin{cases} \frac{1}{S-2} & s' \in [S-2] \\ 0 & else \end{cases}, \forall h \in [H], s \in \{s_{good}, s_{bad}\}, a \in \{1,2,3,\cdots,A\}$*

Definition 9 (Tree Instance specified by (S, A, H, r))**.** *Consider integer S, A, H such that $S - 1 = 2^N$ for some integer $N \geq 1$, and $A \geq 3$, $H - 1 \geq 6N = 6\log_2(S-1)$, and a real value $r \in (\frac{3}{8}, \frac{5}{8})$. We consider a binary tree, with state space $\mathcal{S} = [2^N - 1] \cup \{s_{good}, s_{bad}\}$. The transition probability is*

$$\begin{aligned}
p_h(2s|s, a=1) =& 1, \forall s \leq 2^{N-1} - 1, \forall h \in [H] \\
p_h(2s+1|s, a=2) =& 1, \forall s \leq 2^{N-1} - 1, \forall h \in [H] \\
p_h(s_{good}|s, a \geq 3) =& r, \forall s \leq 2^{N-1} - 1, \forall h \in [H] \\
p_h(s_{bad}|s, a \geq 3) =& 1 - r, \forall s \leq 2^{N-1} - 1, \forall h \in [H] \\
p_h(s_{good}|s, a) =& r, \forall 2^N - 1 \geq s \geq 2^{N-1}, \forall a \in [A], \forall h \in [H] \\
p_h(s_{bad}|s, a) =& 1 - r, \forall 2^N - 1 \geq s \geq 2^{N-1}, \forall a \in [A], \forall h \in [H] \\
p_h(s_{good}|s_{good}, a) =& 1, \forall h \in [H], \forall a \in [A] \\
p_h(s_{bad}|s_{bad}, a) =& 1, \forall h \in [H], \forall a \in [A]
\end{aligned}$$

Assume we always start at the state 1, i.e. $\Pr_\nu(s_1^\pi = 1) = 1$ *holds for any policy* π*. The reward function is* $r_h(s,a) = 0, \forall s \in [2^N - 1] \cup \{s_{bad}\}$*;* $r_h(s_{good}, a) = 1, \forall h \in [H], a \in [A]$.

C.2 Proof of Lower Bound, for positive instances

Before illustrating the proof of Theorem 4, we first introduce the definition of symmetric algorithm.

Definition 10 (Symmetric Algorithm)**.** *Given a 1-policy-identification algorithm alg, an instance* ν*, two different permutations* $\sigma' = \{\sigma'_{s,h}\}_{(s,h)\in\mathcal{S}\times\mathcal{H}}$*,* $\sigma'' = \{\sigma''_{s,h}\}_{(s,h)\in\mathcal{S}\times\mathcal{H}}$*. We call the algorithm is symmetric, if for any* $T \in \mathbb{N}$*, any realization path* $\{(s_{t,h}, a_{t,h})_{h=1}^H\}_{t=1}^T$ *up to* T *episodes, any output policy* $\{a(s,h)\}_{(s,h)\in\mathcal{S}\times\mathcal{H}}$ *algorithm can fulfill*

$$\Pr_{alg,\nu_{\sigma'}}\left(\forall t\in[T], \forall h\in[H], S_{t,h} = s_{t,h}, A_{t,h} = \sigma'_{s_{t,h},h}(a_{t,h}); \forall s,h,\tau = T, \hat{\pi}_h(s) = \sigma'_{s,h}(a(s,h))\right)$$
$$= \Pr_{alg,\nu_{\sigma''}}\left(\forall t\in[T], \forall h\in[H], S_{t,h} = s_{t,h}, A_{t,h} = \sigma''_{s_{t,h},h}(a_{t,h}); \forall s,h,\tau = T, \hat{\pi}_h(s) = \sigma''_{s,h}(a(s,h))\right)$$

Here $S_{t,h}, A_{t,h}$ *denote the state and action we take at round* h*, episode* t.

The following Theorem suggests that we suffice to consider symmetric algorithm for lower bounds

Theorem 25. *Denote* Γ *as the set of all the possible permutations on* ν*. For any instance* ν *and any* δ*-PAC alg, we can find a* δ*-PAC and symmetric* $\tilde{alg}$*, such that* $\max_{\sigma\in\Gamma}\mathbb{E}_{alg,\nu_\sigma}\tau \geq \mathbb{E}_{\tilde{alg},\nu}\tau$.

Given Theorem 25, we suffice to prove the following theorem to show Theorem 4 holds.

Theorem 26. *For any* δ*-PAC symmetric algorithm and any Uniform Instance specified by* (S, A, H, r, ϵ)*, with* $\frac{3H(r+\epsilon)}{8} + \frac{Hr}{8} > \mu_0 > \frac{Hr}{2}$*,* $\frac{3H(r+\epsilon)}{8} + \frac{Hr}{8} > \mu_0 > \frac{Hr}{2}$*,* $S, H \geq 10$*,* $A \geq 100$*,* $(S-2)H$ *is divisible by* 16*,* $0 < \epsilon < \frac{1}{200\sqrt{5}}$*, it holds that*

$$\mathbb{E}\tau \geq \Omega\left(\frac{H\log\frac{1}{\delta}}{(V_0^*(s_0|\nu) - \mu_0)^2} + \frac{SAH^2}{(V_0^*(s_0|\nu) - \mu_0)^2}\right)$$

We first present the proof of Theorem 25. After that, we split the proof of Theorem 26 into multiple steps.

Proof of Theorem 25. We define the symmetric $\tilde{\text{alg}}$ as follows. We choose a permutation $\sigma = \{\sigma_{s,h}\}_{(s,h)\in\mathcal{S}\times\mathcal{H}}$ from Γ uniformly and randomly, before the sampling process starts. Then apply the alg to the instance ν_σ to conduct the sampling process.

We first validate $\tilde{\text{alg}}$ is $\delta - PAC$. If the instance ν is positive, we have

$$\begin{aligned}
&\Pr_{\tilde{\text{alg}},\nu}(V_0^{\hat{\pi}_\tau}(s_0) \geq \mu_0, \tau < +\infty)\\
=&\sum_{\sigma\in\Gamma}\frac{\Pr_{\text{alg},\nu_\sigma}(V_0^{\hat{\pi}_\tau}(s_0) \geq \mu_0, \tau < +\infty)}{(A!)^{SH}}\\
\geq&\sum_{\sigma\in\Gamma}\frac{1-\delta}{(A!)^{SH}}\\
=&1-\delta.
\end{aligned}$$

The idea applies to the case that ν is negative.

Then we turn to show $\tilde{\text{alg}}$ is symmetric. Given permutation σ', σ'', since we choose $\sigma \in \Gamma$ in uniformly random manner, we know instances $(\nu_{\sigma'})_\sigma$ and $(\nu_{\sigma''})_\sigma$ share the same distribution as ν_σ. Thus, we can conclude

$$\begin{aligned}
&\Pr_{\text{alg},\nu_{\sigma'}}\left(\forall t\in[T], \forall h\in[H], S_{t,h} = s_{t,h}, A_{t,h} = \sigma'_{s_{t,h},h}(a_{t,h}); \forall s,h,\tau = T, \hat{\pi}_h(s) = \sigma'_{s,h}(a(s,h))\right)\\
=&\frac{\sum_{\sigma\in\Gamma}\Pr_{\text{alg},\nu_\sigma}\left(\forall t\in[T], \forall h\in[H], S_{t,h} = s_{t,h}, A_{t,h} = \sigma_{s_{t,h},h}(a_{t,h}); \forall s,h,\tau = T, \hat{\pi}_h(s) = \sigma_{s,h}(a(s,h))\right)}{(A!)^{SH}}\\
=&\Pr_{\text{alg},\nu_{\sigma''}}\left(\forall t\in[T], \forall h\in[H], S_{t,h} = s_{t,h}, A_{t,h} = \sigma''_{s_{t,h},h}(a_{t,h}); \forall s,h,\tau = T, \hat{\pi}_h(s) = \sigma''_{s,h}(a(s,h))\right),
\end{aligned}$$

suggesting that $\tilde{\text{alg}}$ is symmetric.

The last step is to prove $\max_{\sigma\in\Gamma}\mathbb{E}_{\text{alg},\nu_\sigma}\tau \geq \mathbb{E}_{\tilde{\text{alg}},\nu}\tau$. From the definition of algorithm $\tilde{\text{alg}}$, we know

$$\mathbb{E}_{\tilde{\text{alg}},\nu}\tau = \frac{\sum_{\sigma\in\Gamma}\mathbb{E}_{\text{alg},\nu_\sigma}\tau}{(A!)^{SH}} \leq \max_{\sigma\in\Gamma}\mathbb{E}_{\text{alg},\nu_\sigma}\tau.$$

And we have completed the proof. □

To prove Theorem 26, we present the following two theorems, which is required to prove Theorem 26.

Theorem 27. *Given a Uniform Instance ν with μ_0 satisfying that $\frac{H(r+\epsilon)}{2} \geq \mu_0 > \frac{Hr}{2}$, for any δ-PAC alg, we have*

$$\mathbb{E}_\nu\tau \geq \Omega\left(\frac{H\log\frac{1}{2.4\delta}}{(V_0^*(s_0|\nu)-\mu_0)^2}\right).$$

Theorem 28. *Given a Uniform Instance ν with (S,A,H,r,ϵ), μ_0 satisfying that $\frac{H(r+\epsilon)}{2} > \mu_0 \geq \frac{H(r+\epsilon)}{4} + \frac{Hr}{4}$, $0<\delta<\frac{1}{2}$, for any δ-PAC and symmetric algorithm, we have*

$$\mathbb{E}\tau \geq \frac{\Omega(SA)}{\epsilon^2}.$$

We first present the proof of Theorem 27 and 28, which require Lemmas in section C.4. At the end of this subsection, we present the proof of Theorem 26.

Proof of Theorem 27. Define $r'(\eta) = \frac{\mu_0}{H} - \eta$ for arbitrary $\frac{\mu_0}{H} - r > \eta > 0$. We consider an alternative instance $\nu'(\eta)$ which is Uniform Instance with parameters $(S,A,H,r,\frac{2\mu_0}{H}-r-\eta)$. It is evident to see $V_0^*(s_0|\nu'(\eta)) = \mu_0 - \frac{H}{2}\eta < \mu_0$, suggesting that $\Pr_{\text{alg},\nu'(\eta)}(\hat{\pi}=\textsf{None}) > 1-\delta$.

Notice that $\Pr_{\text{alg},\nu}(\hat{\pi}=\textsf{None}) < \delta$, by the Lemma 1 in Kaufmann et al. (2016), we can conclude

$$\sum_{s=1}^{S-2}\sum_{h:\text{odd}}\text{kl}(r+\epsilon,\frac{2\mu_0}{H}-\eta)\mathbb{E}_{\text{alg},\nu}n_h^\tau(s,1) \geq \text{kl}(1-\delta,\delta) \tag{49}$$

where $\text{kl}(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$. From the definition of Simple Requirement, we know $\frac{1}{4} < r < \frac{2\mu_0}{H} - \eta < \frac{3}{4}$. By the Lemma 36, we know $\text{kl}(r+\epsilon,\frac{2\mu_0}{H}-r-\eta) \leq 3(r+\epsilon-\frac{2\mu_0}{H}+\eta)^2$ and (49) implies

$$\sum_{s=1}^{S-2}\sum_{h:\text{odd}}\mathbb{E}_{\text{alg},\nu}n_h^\tau(s,1) \geq \frac{\text{kl}(1-\delta,\delta)}{3(r+\epsilon-\frac{2\mu_0}{H}+\eta)^2}$$

$$\Rightarrow \frac{1}{H}\sum_{s=1}^{S}\sum_{h=1}^{H}\sum_{a=1}^{A}\mathbb{E}_{\text{alg},\nu}n_h^\tau(s,a) \geq \frac{H\text{kl}(1-\delta,\delta)}{12(V_0^*(s_0|\nu)-\mu_0+\frac{H}{2}\eta)^2}$$

By the fact that $\mathbb{E}_\nu\tau = \frac{1}{H}\sum_{s=1}^{S}\sum_{h=1}^{H}\sum_{a=1}^{A}\mathbb{E}_{\text{alg},\nu}n_h^\tau(s,a)$, we can conclude

$$\mathbb{E}_\nu\tau \geq \frac{H\log\frac{1}{2.4\delta}}{12(V_0^*(s_0|\nu)-\mu_0+\frac{H}{2}\eta)^2}$$

holds for all $\frac{\mu_0}{H} - r > \eta > 0$. Let $\eta\to 0$, we have completed the proof. □

Proof of Theorem 28. By the Lemma 31, we can find $\{(s^{(i)},h^{(i)})\}_{i=1}^{\frac{(S-2)H}{16}}$, where $s^{(i)} \in [S-2]$, $h^{(i)}\in[H]$ is odd, such that

$$\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{(i)}}(s^{(i)}) = 1, \tau < +\infty\right) \geq \frac{1}{7}, \forall i \in \{1,2,\cdots,\frac{(S-2)H}{16}\}.$$

By the Lemma 34, we can conclude

$$\mathbb{E}_{\nu,\tilde{\text{alg}}} n^{\tau}_{h^{(i)}}(s^{(i)}, a) \geq \frac{1}{2 \cdot 10^5 \epsilon^2}, \forall a \in \{2, \cdots, A\}, \forall i \in \{1, 2, \cdots, \frac{(S-2)H}{16}\}. \tag{50}$$

Then, we can conclude

$$\begin{aligned}
\mathbb{E}_{\nu,\tilde{\text{alg}}} \tau =& \frac{1}{H} \sum_{a=1}^{A} \sum_{s=1}^{S} \sum_{h=1}^{H} \mathbb{E}_{\nu,\tilde{\text{alg}}} n_h(s, a) \\
\geq& \frac{1}{H} \sum_{a=2}^{A} \sum_{i=1}^{\frac{(S-2)H}{16}} \mathbb{E}_{\nu,\tilde{\text{alg}}} n^{\tau}_{h^{(i)}}(s^{(i)}, a) \\
\geq& \frac{(A-1)(S-2)}{32 \cdot 10^5 \epsilon^2}.
\end{aligned}$$

The last line is by the (50), which completes the proof. □

Given Theorem 27 and 28, we are ready to prove Theorem 26.

Proof of Theorem 26. By Theorem 27 and 28, we can conclude

$$\mathbb{E}_{\nu} \tau \geq \Omega \left(\frac{H \log \frac{1}{2.4\delta}}{(V_0^*(s_0|\nu) - \mu_0)^2} + \frac{SA}{\epsilon^2} \right)$$

By the construction of Uniform Instance ν, we know $V_0^*(s_0|\nu) = \frac{H(r+\epsilon)}{2}$. Since $\frac{3H(r+\epsilon)}{8} + \frac{Hr}{8} > \mu_0 > \frac{Hr}{2}$, we can conclude

$$(V_0^*(s_0|\nu) - \mu_0)^2 \geq (\frac{H(r+\epsilon)}{2} - \frac{3H(r+\epsilon)}{8} - \frac{Hr}{8})^2 = (\frac{H\epsilon}{8})^2 = \frac{H^2\epsilon^2}{64}.$$

The last ste implies the lower bound

$$\mathbb{E}_{\nu} \tau \geq \Omega \left(\frac{H \log \frac{1}{2.4\delta}}{(V_0^*(s_0|\nu) - \mu_0)^2} + \frac{SAH^2}{(V_0^*(s_0|\nu) - \mu_0)^2} \right).$$

□

C.3 Proof of Lower Bound, for negative instances

Before delivering the proof of Theorem 5, we first present the following lemma to calculate the exact value of $V_0^*(s_0|\nu)$ for a Tree instance ν.

Lemma 29. *For a Tree Instance ν with parameter (S, A, H, r), $V_0^*(s_0|\nu) = (H-1)r$.*

Proof. Given a policy π, notice that $\{(s, a, h) : h : \Pr_{\nu,\pi}(\text{hitting } s \text{ at round } h) > 0, \pi_h(s) = a \geq 3\}$ contains at most 1 element. Denote the unique element as (s^π, a^π, h^π). We have $V_0^\pi(1) = (H - h^\pi)r \leq (H-1)r$. And the equality holds only when $\pi_{h=1}(s=1) = a$ for $a \geq 3$. □

Then, we are ready to prove Theorem 5.

Theorem (Restatement of Theorem 5)**.** *Apply any δ-PAC algorithm to a Tree instance specified by (S, A, H, r) with $S, A, H \geq 4$ and $\frac{3}{4}(H - \log_2(S-3)) > \mu_0 > (H-1)r$, we have*

$$\mathbb{E}_{\nu} \tau \geq \Omega \left(\frac{HAS \log \frac{1}{\delta}}{(\mu_0 - V_0^*(s_0|\nu))^2} \right).$$

Proof of Theorem 5. For any pair (s_0, a_0), where $1 \leq s_0 \leq 2^N - 1$, $A \geq a_0 \geq 3$, we define $h(s) = \lceil \log_2(s_0 + 1) \rceil$ and an instance $\nu_{s_0,a_0}(\eta)$, $1 - \frac{\mu_0}{H - h(s)} > \eta > 0$, such that ν_{s_0,a_0} only differs in $p^{\nu_{s_0,a_0}}_{h(s_0)}(s_{\text{good}}|s_0, a_0) = \frac{\mu_0}{H-h(s)} + \eta$ compared to ν.

We first validate the instance $\nu_{s_0,a_0}(\eta)$ is positive for any $\eta > 0$, by figuring out a qualified policy. From Lemma 30, we know there exists a policy π such that $\Pr_\nu(s^\pi_{h(s_0)} = s_0) = 1$. Now, we consider a policy π' such that $\pi'_h(s) = \begin{cases} a_0 & s = s_0, h = h(s_0) \\ \pi_h(s) & \text{else} \end{cases}$. Evident to see $V_0^{\pi'}(s_0; \nu_{s_0,a_0}(\eta)) = (H - h(s_0))\left(\frac{\mu_0}{H-h(s)} + \eta\right) = \mu_0 + \eta(H - h(s_0)) > \mu_0$.

Given $\nu_{s_0,a_0}(\eta)$ is positive and ν is negative, we can conclude $\Pr_{\nu,\text{alg}}(\hat{\pi} = \textsf{None}) \geq 1 - \delta$, $\Pr_{\nu_{s_0,a_0}(\eta),\text{alg}}(\hat{\pi} = \textsf{None}) \leq \delta$. Apply the Lemma 1 in Kaufmann et al. (2016), we have

$$\mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s_0)}(s_0, a_0)\text{kl}(r, \frac{\mu_0}{H - h(s)} + \eta) \geq \log \frac{1}{2.4\delta}.$$

Since the above inequality holds for arbitrary η, letting $\eta \to 0$ implies

$$\begin{aligned}
&\mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s_0)}(s_0, a_0)\text{kl}(r, \frac{\mu_0}{H - h(s)}) \geq \log \frac{1}{2.4\delta} \\
\Rightarrow &\mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s_0)}(s_0, a_0) \cdot 3\big(\frac{\mu_0}{H - h(s)} - r\big)^2 \geq \log \frac{1}{2.4\delta} && (51) \\
\Rightarrow &\mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s_0)}(s_0, a_0) \cdot 3\big(\frac{\mu_0}{H - \lceil \log_2(S-1) \rceil} - r\big)^2 \geq \log \frac{1}{2.4\delta} && (52) \\
\Rightarrow &\mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s_0)}(s_0, a_0) \cdot 3\big(\frac{\mu_0}{\frac{2}{3}(H-1)} - r\big)^2 \geq \log \frac{1}{2.4\delta} && (53) \\
\Rightarrow &\mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s_0)}(s_0, a_0) \geq \frac{\frac{4}{9}(H-1)^2 \log \frac{1}{2.4\delta}}{3(\mu_0 - (H-1)r)^2} && (54)
\end{aligned}$$

Step (51) is by the assumption $\frac{3}{4}(H - \log_2(S-3)) > \mu_0 > (H-1)r$, which suggests that $r, \frac{\mu_0}{H-h(s)} \in (\frac{1}{4}, \frac{3}{4})$. Lemma 37 concludes $\text{kl}(r, \frac{\mu_0}{H-h(s)}) \leq 3(\frac{\mu_0}{H-h(s)} - r)^2$. Step (52) is by $s_0 \leq S-2$ and $\frac{3}{4}(H - \log_2(S-3)) > \mu_0 > (H-1)r$ implies $\frac{\mu_0}{H - \lceil \log_2(S-1) \rceil} \geq \frac{\mu_0}{H-h(s)} > r$. Step (53) is by the definition of Tree Instance, which requires $H \geq 6N, 2^N = S - 3, N \geq 1$, implying

$$\begin{aligned}
&\lceil \log_2(S-1) \rceil \\
=&\lceil \log_2(2^N) \rceil \\
\leq &N \\
\leq &\frac{H}{6}.
\end{aligned}$$

The last line means $H - \lceil \log_2(S-1) \rceil \geq \frac{2(H-1)}{3}$, which is step (53).

Summing up (54) for $s_0 \in [S-2]$ and $a_0 \in \{3, \cdots, A\}$, we can conclude

$$\begin{aligned}
&\sum_{s=1}^{S-2} \sum_{a=3}^{A} \mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s)}(s, a) \geq \frac{4(S-2)(A-1)(H-1)^2 \log \frac{1}{2.4\delta}}{27(\mu_0 - (H-1)r)^2} \\
\Rightarrow &\mathbb{E}_{\nu,\text{alg}} \tau = \frac{1}{H} \sum_{h=1}^{H} \sum_{s=1}^{S} \sum_{a=1}^{A} \mathbb{E}_{\nu,\text{alg}} n^\tau_{h(s)}(s, a) \geq \Omega\left(\frac{HAS \log \frac{1}{\delta}}{(\mu_0 - V_0^*(s_0|\nu))^2}\right).
\end{aligned}$$

We have completed the proof. □

The following is an auxiliary Lemma, showing that there exists a shorted path from state 1 to s for any $s \in [S-2]$ in the Tree Instance.

Lemma 30. *Consider a Tree Instance ν specified by parameter (S, A, H, r). For any state $s \in [S-2]$, there exists a policy π, such that* $\Pr_\nu(s^\pi_{\lceil \log_2(s+1) \rceil} = s) = 1$.

Proof of Lemma 30. We prove the Lemma by induction. For $s = 1$, the lemma is true, as we always start at the state 1. Now, assume for all $s \in [2^n, 2^{n+1} - 1] \cap \mathbb{Z}$, the conclusion is true, we consider

a state $s \in [2^{n+1}, 2^{n+2}-1] \cap \mathbb{Z}$. Easy to see $\lfloor \frac{s}{2} \rfloor \in [2^n, 2^{n+1}-1]$. Thus, there exists a policy π, such that $\Pr_\nu(s^\pi_{\lceil \log_2(\lfloor \frac{s}{2} \rfloor + 1) \rceil} = \lfloor \frac{s}{2} \rfloor) = 1$. Since there exists an action $a_s \in \{1, 2\}$ in state $\lfloor \frac{s}{2} \rfloor$ such that $p_{\lceil \log_2(\lfloor \frac{s}{2} \rfloor + 1) \rceil}(s | \lfloor \frac{s}{2} \rfloor, a_s) = 1$, we can conclude there exists a policy π' such that $\Pr_\nu(s^{\pi'}_{\lceil \log_2(\lfloor \frac{s}{2} \rfloor + 1) \rceil + 1} = s) = 1$.

Notice that $2^n + 1 \le \lfloor \frac{s}{2} \rfloor + 1 \le 2^{n+1}$, suggesting that

$$\begin{aligned} &\lceil \log_2(\lfloor \frac{s}{2} \rfloor + 1) \rceil + 1 \\ =& n + 1 + 1 \\ =& n + 2 \\ =& \lceil \log_2(s+1) \rceil. \end{aligned}$$

The last line completes the induction, also the proof of Lemma 30. □

C.4 Auxiliary Lemmas in the Proof of Lower Bound, for Uniform Instances

In this section, We present auxiliary lemmas to support Theorem 28. Without extra description, we assume $\frac{3H(r+\epsilon)}{8} + \frac{Hr}{8} > \mu_0 > \frac{Hr}{2}$, $S, H \ge 10$, $A \ge 100$, $(S-2)H$ is divisible by 16, $0 < \epsilon < \frac{1}{200\sqrt{5}}$ always hold in the remaining subsection.

The first lemma is to show a δ-PAC algorithm must output policy that identifies action 1 as the optimal action for "enough" pairs of (s, h).

Lemma 31. *Apply a δ-PAC algorithm alg to a Uniform Instance ν with (S, A, H, r, ϵ). Assume μ_0 satisfies $\frac{H(r+\epsilon)}{2} > \mu_0 \ge \frac{H(r+\epsilon)}{4} + \frac{Hr}{4}$, $0 < \delta < \frac{1}{2}$, we can find $\{(s^{(i)}, h^{(i)})\}_{i=1}^{\frac{(S-2)H}{16}}$, where $s^{(i)} \in [S-2]$, $h^{(i)} \in [H]$ is odd, such that*

$$\Pr_{alg,\nu}\left(\hat{\pi}_{h^{(i)}}(s^{(i)}) = 1, \tau < +\infty\right) \ge \frac{1}{7}, \forall i \in \{1, 2, \cdots, \frac{(S-2)H}{16}\}$$

Proof of Lemma 31. Given a policy π on the Uniform Instance, we know

$$V_0^\pi(s_0) = \frac{\sum_{s \in [S-2], \text{odd } h}(r+\epsilon)\mathbb{1}(\pi_h(s) = 1) + \sum_{s \in [S-2], \text{odd } h} r\mathbb{1}(\pi_h(s) \ne 1)}{S-2}.$$

By the assumption that $\mu_0 \ge \frac{H(r+\epsilon)}{4} + \frac{Hr}{4}$ and the δ-PAC requirement, we know

$$\Pr_{\text{alg},\nu}\left(\sum_{s \in [S-2], \text{odd } h} \mathbb{1}\left(\hat{\pi}_h(s) = 1\right) \ge \frac{H(S-2)}{4}, \tau < +\infty\right) \ge 1 - \delta.$$

Denote $\{s^{(i)}, h^{(i)}\}_{i=1}^{SH}$ as a permutation of $[S-2] \times \{h \in [H] : h \text{ is odd}\}$, such that

$$\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{(i)}}(s^{(i)}) = 1, \tau < +\infty\right) \ge \Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{(i+1)}}(s^{(i+1)}) = 1, \tau < +\infty\right)$$

Define $\Lambda = \{C : C \subset [S-2] \times \{h \in [H] : h \text{ is odd}\}, |C| \ge \frac{(S-2)H}{4}\}$, $\Lambda_{\text{all}} = \{C : C \subset [S-2] \times \{h \in [H] : h \text{ is odd}\}\}$. Notice that

$$\begin{aligned} &\left\{\sum_{(s,h)} \mathbb{1}\left(\hat{\pi}_h(s) = 1\right) \ge \frac{H(S-2)}{4}\right\} \\ =& \cup_{C \in \Lambda}\left\{\forall (s,h) \in C, \hat{\pi}_h(s) = 1; \forall (s,h) \notin C, \hat{\pi}_h(s) \ne 1\right\} \end{aligned}$$

$\Pr_{\text{alg},\nu}\left(\sum_{(s,h)} \mathbb{1}\left(\hat{\pi}_h(s) = 1\right) \ge \frac{H(S-2)}{4}, \tau < +\infty\right) \ge 1 - \delta$ implies

$$\sum_{C \in \Lambda} \Pr_{\text{alg},\nu}\left(\forall (s,h) \in C, \hat{\pi}_h(s) = 1; \forall (s,h) \notin C, \hat{\pi}_h(s) \ne 1\right) \ge 1 - \delta. \tag{55}$$

By straight forward calculation,

$$
\begin{aligned}
&\sum_{s\in[S-2],\text{odd } h} \Pr_{\text{alg},\nu}\left(\hat{\pi}_h(s)=1,\tau<+\infty\right)\\
=&\sum_{s\in[S-2],\text{odd } h}\ \sum_{C:C\in\Lambda_{\text{all}},(s,h)\in C} \Pr_{\text{alg},\nu}\left(\forall(s',h')\in C,\hat{\pi}_{h'}(s')=1;\forall(s',h')\notin C,\hat{\pi}_{h'}(s')\neq 1,\tau<+\infty\right)\\
\geq&\sum_{s\in[S-2],\text{odd } h}\ \sum_{C:C\in\Lambda,(s,h)\in C} \Pr_{\text{alg},\nu}\left(\forall(s',h')\in C,\hat{\pi}_{h'}(s')=1;\forall(s',h')\notin C,\hat{\pi}_{h'}(s')\neq 1,\tau<+\infty\right)\\
=&\sum_{s\in[S-2],\text{odd } h}\ \sum_{C:C\in\Lambda} \mathbb{1}\Big((s,h)\in C\Big)\Pr_{\text{alg},\nu}\left(\forall(s',h')\in C,\hat{\pi}_{h'}(s')=1;\forall(s',h')\notin C,\hat{\pi}_{h'}(s')\neq 1,\tau<+\infty\right)\\
=&\sum_{C:C\in\Lambda}\ \sum_{s\in[S-2],\text{odd } h} \mathbb{1}\Big((s,h)\in C\Big)\Pr_{\text{alg},\nu}\left(\forall(s',h')\in C,\hat{\pi}_{h'}(s')=1;\forall(s',h')\notin C,\hat{\pi}_{h'}(s')\neq 1,\tau<+\infty\right)\\
\geq&\sum_{C:C\in\Lambda}\frac{(S-2)H}{4}\Pr_{\text{alg},\nu}\left(\forall(s',h')\in C,\hat{\pi}_{h'}(s')=1;\forall(s',h')\notin C,\hat{\pi}_{h'}(s')\neq 1,\tau<+\infty\right)\\
\geq&\frac{(S-2)H}{4}(1-\delta)\\
\geq&\frac{(S-2)H}{8}.
\end{aligned}
$$

The second last step is by (55) and the last step is by $\delta<\frac{1}{2}$.

Recall $\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{(i)}}(s^{(i)})=1,\tau<+\infty\right)$ is non-decreasing regarding i, we can conclude

$$
\begin{aligned}
&\sum_{s,h}\Pr_{\text{alg},\nu}\left(\hat{\pi}_h(s)=1,\tau<+\infty\right)\\
=&\sum_{i=1}^{(S-2)H/2}\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{(i)}}(s^{(i)})=1,\tau<+\infty\right)\\
=&\sum_{i=1}^{(S-2)H/16}\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{(i)}}(s^{(i)},\tau<+\infty)=1\right)+\sum_{i=(S-2)H/16+1}^{(S-2)H/2}\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{(i)}}(s^{(i)})=1,\tau<+\infty\right)\\
\leq&(S-2)H/16+\frac{7(S-2)H}{16}\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{((S-2)H/16+1)}}(s^{((S-2)H/16+1)})=1,\tau<+\infty\right).
\end{aligned}
$$

The last line suggests that

$$
\frac{(S-2)H}{8}\leq(S-2)H/16+\frac{7(S-2)H}{16}\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{((S-2)H/16+1)}}(s^{((S-2)H/16+1)})=1,\tau<+\infty\right),
$$

which means

$$
\frac{1}{7}\leq\Pr_{\text{alg},\nu}\left(\hat{\pi}_{h^{((S-2)H/16+1)}}(s^{((S-2)H/16+1)})=1,\tau<+\infty\right).
$$

From the non-decreasing order, we can conclude the Lemma. □

Lemma 31 shows there exists at least $\Omega(SH)$ pairs of (s,h) such that a δ-PAC algorithm will identify its optimal action with probability $1/7$. The following lemmas shows a δ-PAC and symmetric algorithm must distinguish arm 1 and other arms by pulling them $\Omega(\frac{1}{\epsilon^2})$ times in these pairs (s,h).

Lemma 32. *Apply a δ-PAC symmetric algorithm $\tilde{alg}$ to a Uniform Instance ν with (S,A,H,r,ϵ). If $\exists s_0\in[S-2]$ and even h_0 such that $\Pr_{\nu,\tilde{alg}}(\hat{\pi}_{h_0}(s_0)=1,\tau<+\infty)\geq\frac{1}{7}$, we have*

$$
\Pr_{\nu,\tilde{alg}}\left(\hat{\pi}_{h_0}(s_0)=1,n^{\tau}_{h_0}(s_0,1)+n^{\tau}_{h_0}(s_0,a)\geq\frac{1}{M\epsilon^2},\tau<+\infty\right)\geq\frac{1}{14},\forall a=2,\cdots,A.
$$

where $M=2000$.

Proof of Lemma 32. We first show for $a', a'' \in \{2, \cdots, A\}, a' \neq a''$, we have

$$\Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = a', \tau < +\infty) = \Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = a'', \tau < +\infty) \tag{56}$$

Consider a permutation $\sigma = \{\sigma_{s,h}\}_{i=1}^{SH}$ such that $\sigma_{s,h}(a) = \begin{cases} a' & s = s_0, h = h_0, a = a'' \\ a'' & s = s_0, h = h_0, a = a' \\ a & \text{else} \end{cases}$.

From the definition of Uniform Instance, we know $p_{h_0}(s'|s_0, a') = p_{h_0}(s'|s_0, a'')$, $r_{h_0}(s_0, a') = r_{h_0}(s_0, a'')$, which means $\nu_\sigma = \nu$. Meanwhile, by the definition of symmetric algorithm, we know

$$\Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = a', \tau < +\infty) = \Pr_{\nu_\sigma,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = a'', \tau < +\infty).$$

Together with the fact that $\nu_\sigma = \nu$, we complete the proof.

In the following, we assume $\frac{1}{M\epsilon^2}$ is an integer for simplicity. To derive Lemma 32, we prove it by contradiction. Assume

$$\Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = 1, n^\tau_{h_0}(s_0, 1) + n^\tau_{h_0}(s_0, a) \geq \frac{1}{M\epsilon^2}, \tau < +\infty) < \frac{1}{14}$$

holds for some $a \in \{2, 3, \cdots, A\}$. From the assumption that $\Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = 1, \tau < +\infty) \geq \frac{1}{7}$, we know

$$\Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = 1, n^\tau_{h_0}(s_0, 1) + n^\tau_{h_0}(s_0, a) < \frac{1}{M\epsilon^2}, \tau < +\infty) \geq \frac{1}{14} \tag{57}$$

Denote event $\mathcal{E}^{(1)} = \{\hat{\pi}_{h_0}(s_0) = 1, \tau < +\infty\}$, $\mathcal{E}^{(a)} = \{\hat{\pi}_{h_0}(s_0) = a, \tau < +\infty\}$, $\mathcal{E}_{\text{pull}} = \{n^\tau_{h_0}(s_0, 1) + n^\tau_{h_0}(s_0, a) < \frac{1}{M\epsilon^2}\}$. We consider a permutation $\sigma^{(a)}$ with $\sigma^{(a)}_{s,h}(\tilde{a}) = \begin{cases} 1 & s = s_0, h = h_0, \tilde{a} = a \\ a & s = s_0, h = h_0, \tilde{a} = 1 \\ \tilde{a} & \text{else} \end{cases}$. By the symmetry of $\tilde{\text{alg}}$ and (57), we can conclude

$$\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}}) \geq \frac{1}{14}. \tag{58}$$

Apply the Transportation Equality(Lemma 18 in Kaufmann et al. (2016)), which is based on the idea of Change of Measure, we have

$$\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}}) = \mathbb{E}_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} \mathbb{1}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}}) \exp\left(- \sum_{t=1}^{n^\tau_{h_0}(s_0,1)} \hat{\text{KL}}_{1,t} - \sum_{t=1}^{n^\tau_{h_0}(s_0,a)} \hat{\text{KL}}_{a,t} \right) \tag{59}$$

where $\hat{\text{KL}}_{1,t} = X_{1,t} \log \frac{r+\epsilon}{r} + (1 - X_{1,t}) \log \frac{1-r-\epsilon}{1-r}$, $\hat{\text{KL}}_{a,t} = X_{a,t} \log \frac{r}{r+\epsilon} + (1 - X_{1,t}) \log \frac{1-r}{1-r-\epsilon}$, $X_{1,t}|\{X_{1,s}\}_{s=1}^{t-1} \sim \text{Bern}(r + \epsilon)$, $X_{a,t}|\{X_{a,s}\}_{s=1}^{t-1} \sim \text{Bern}(r)$. $\{X_{1,s}\}_{s=1}^{+\infty}$ are independent with $\{X_{a,s}\}_{s=1}^{+\infty}$.

Denote

$$\begin{aligned} \text{kl}(x, y) =& x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y} \\ \mathcal{E}^{(1)}_{\text{kl-cnt}} =& \left\{ \max_{1 \leq n \leq \frac{1}{M\epsilon^2}} \sum_{t=1}^{n} (\hat{\text{KL}}_{1,t} - \text{kl}(r + \epsilon, r)) \leq \frac{1}{4} \right\} \\ \mathcal{E}^{(a)}_{\text{kl-cnt}} =& \left\{ \max_{1 \leq n \leq \frac{1}{M\epsilon^2}} \sum_{t=1}^{n} (\hat{\text{KL}}_{a,t} - \text{kl}(r, r + \epsilon)) \leq \frac{1}{4} \right\} \end{aligned}$$

By the Lemma 36, we know $\hat{\text{KL}}_{1,t} - \text{kl}(r + \epsilon, r)$ and $\hat{\text{KL}}_{a,t} - \text{kl}(r, r + \epsilon)$ are both $8\epsilon^2$-subgaussian. By the Lemma 35, since $2 \cdot \lfloor \frac{1}{M\epsilon^2} \rfloor \cdot 8\epsilon^2 \leq \frac{16}{M}$, we have $\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{(1)}_{\text{kl-cnt}}) \geq 1 - \exp(-\frac{M}{16}\frac{1}{16})$, $\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{(a)}_{\text{kl-cnt}}) \geq 1 - \exp(-\frac{M}{16}\frac{1}{16})$.

Following (59), we have

$$
\begin{aligned}
&\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}}) \\
\geq &\mathbb{E}_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} \mathbb{1}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}} \cap \mathcal{E}^{(1)}_{\text{kl-cnt}} \cap \mathcal{E}^{(a)}_{\text{kl-cnt}}) \exp\left(- \sum_{t=1}^{n^{\tau}_{h_0}(s_0,1)} \hat{\text{KL}}_{1,t} - \sum_{t=1}^{n^{\tau}_{h_0}(s_0,a)} \hat{\text{KL}}_{a,t} \right) \\
= &\mathbb{E}_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} \mathbb{1}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}} \cap \mathcal{E}^{(1)}_{\text{kl-cnt}} \cap \mathcal{E}^{(a)}_{\text{kl-cnt}}) \exp(-n^{\tau}_{h_0}(s_0,1)\text{kl}(r+\epsilon, r) - n^{\tau}_{h_0}(s_0,a)\text{kl}(r, r+\epsilon)) \\
&\exp\left(- \sum_{t=1}^{n^{\tau}_{h_0}(s_0,1)} \left(\hat{\text{KL}}_{1,t} - \text{kl}(r+\epsilon, r)\right) - \sum_{t=1}^{n^{\tau}_{h_0}(s_0,a)} \left(\hat{\text{KL}}_{a,t} - \text{kl}(r, r+\epsilon)\right) \right) \\
\geq &\mathbb{E}_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} \mathbb{1}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}} \cap \mathcal{E}^{(1)}_{\text{kl-cnt}} \cap \mathcal{E}^{(a)}_{\text{kl-cnt}}) \exp(-\frac{3\epsilon^2}{M\epsilon^2} - \frac{3\epsilon^2}{M\epsilon^2}) \exp\left(-\frac{1}{2}\right) \qquad (60) \\
\geq &\exp\left(-\frac{6}{M} - \frac{1}{2}\right) \left(\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} \left(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}}\right) - \Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} (\neg \mathcal{E}^{(1)}_{\text{kl-cnt}}) - \Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} (\neg \mathcal{E}^{(a)}_{\text{kl-cnt}}) \right) \\
\geq &\exp\left(-\frac{6}{M} - \frac{1}{2}\right) \left(\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}} \left(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}}\right) - 2\exp(-\frac{M}{256}) \right). \qquad (61)
\end{aligned}
$$

Step (60) is by the fact that under event $\mathcal{E}_{\text{pull}}$, $n^{\tau}_{h_0}(s_0,1), n^{\tau}_{h_0}(s_0,a) \leq \frac{1}{M\epsilon^2}$. Given this, event $\mathcal{E}^{(1)}_{\text{kl-cnt}}, \mathcal{E}^{(a)}_{\text{kl-cnt}}$ guarantee that

$$
\begin{aligned}
\sum_{t=1}^{n^{\tau}_{h_0}(s_0,1)} \left(\hat{\text{KL}}_{1,t} - \text{kl}(r+\epsilon, r)\right) &\leq \max_{1 \leq n \leq \frac{1}{M\epsilon^2}} \sum_{t=1}^{n} \left(\hat{\text{KL}}_{1,t} - \text{kl}(r+\epsilon, r)\right) \leq \frac{1}{4} \\
\sum_{t=1}^{n^{\tau}_{h_0}(s_0,a)} \left(\hat{\text{KL}}_{a,t} - \text{kl}(r, r+\epsilon)\right) &\leq \max_{1 \leq n \leq \frac{1}{M\epsilon^2}} \sum_{t=1}^{n} \left(\hat{\text{KL}}_{a,t} - \text{kl}(r, r+\epsilon)\right) \leq \frac{1}{4}
\end{aligned}
$$

In addition, Lemma 37 suggests that $\text{kl}(r+\epsilon, r), \text{kl}(r, r+\epsilon) \leq 3\epsilon^2$, since we require $\frac{1}{4} < r < r+\epsilon < \frac{3}{4}$ in the definition of Uniform Instance. Step (61) is by the inequalities $\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{(1)}_{\text{kl-cnt}}) \geq 1 - \exp(-\frac{M}{256})$, $\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{(a)}_{\text{kl-cnt}}) \geq 1 - \exp(-\frac{M}{256})$.

Combining (58) and (61), we have

$$
\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}^{(a)} \cap \mathcal{E}_{\text{pull}}) \geq \exp\left(-\frac{6}{M} - \frac{1}{2}\right) \left(\frac{1}{14} - 2\exp(-\frac{M}{256})\right) \geq \frac{1}{100},
$$

which implies $\Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = a, \tau < +\infty) \geq \frac{1}{100}$. (56) implies

$$
\Pr_{\nu,\tilde{\text{alg}}}\left(\hat{\pi}_{h_0}(s_0) \in \{2, \cdots, A\}, \tau < +\infty\right) \geq \frac{A}{100}.
$$

By the requirement of $A > 100$ in the definition of Uniform Instance, we have found a contradiction, suggesting we have proved Lemma 32. □

Based on Lemma 32, the following lemma claims that a symmetric algorithm is unable to distinguish action 1 and a when the total pulling times of action 1 and a is below $\frac{1}{2000\epsilon^2}$.

Lemma 33. *Apply a δ-PAC symmetric algorithm $\tilde{alg}$ to a Uniform Instance ν with (S, A, H, r, ϵ). If $\exists s_0 \in [S-2]$ and even h_0 such that $\Pr_{\nu,\tilde{alg}}(\hat{\pi}_{h_0}(s_0) = 1, \tau < +\infty) \geq \frac{1}{7}$, we have*

$$
\Pr_{\nu,\tilde{alg}}\left(n^{\tau}_{h_0}(s_0,a) \geq \frac{1}{2M\epsilon^2}, \tau < +\infty\right) \geq \frac{1}{50}, \forall a = 2, \cdots, A.
$$

where $M = 2000$.

Proof of Lemma 33. We first show for $a', a'' \in \{2, \cdots, A\}, a' \neq a''$, we have

$$\Pr_{\nu,\tilde{\text{alg}}}(n^{\tau}_{h_0}(s_0,a') \geq \frac{1}{2M\epsilon^2}, \tau < +\infty) = \Pr_{\nu,\tilde{\text{alg}}}(n^{\tau}_{h_0}(s_0,a'') \geq \frac{1}{2M\epsilon^2}, \tau < +\infty). \tag{62}$$

Similar to the statement at the proof of Lemma 32, consider a permutation $\sigma = \{\sigma_{s,h}\}_{i=1}^{SH}$ such that $\sigma_{s,h}(a) = \begin{cases} a' & s = s_0, h = h_0, a = a'' \\ a'' & s = s_0, h = h_0, a = a' \\ a & \text{else} \end{cases}$. From the definition of Uniform Instance, we know $p_{h_0}(s'|s_0,a') = p_{h_0}(s'|s_0,a'')$, $r_{h_0}(s_0,a') = r_{h_0}(s_0,a'')$, which means $\nu_\sigma = \nu$. Meanwhile, by the definition of symmetric algorithm, we know

$$\Pr_{\nu,\tilde{\text{alg}}}(n^{\tau}_{h_0}(s_0,a') \geq \frac{1}{2M\epsilon^2}, \tau < +\infty) = \Pr_{\nu_\sigma,\tilde{\text{alg}}}(n^{\tau}_{h_0}(s_0,a'') \geq \frac{1}{2M\epsilon^2}, \tau < +\infty).$$

Together with the fact that $\nu_\sigma = \nu$, we complete the proof.

(62) suggests that we suffice to prove $\Pr_{\nu,\tilde{\text{alg}}}(n^{\tau}_{h_0}(s_0,a) \geq \frac{1}{2M\epsilon^2}, \tau < +\infty) \geq \frac{1}{50}$ for $a = 2$. In the following proof, we fix $a = 2$ and we assume $\frac{1}{M\epsilon^2}$ is an integer for simplicity.

Define $\tilde{\tau}_{1,a} = \min\{t : n^t_{h_0}(s_0,1) + n^t_{h_0}(s_0,a) = \lfloor \frac{1}{M\epsilon^2} \} \rfloor$. From Lemma 32, we know

$$\begin{aligned}
&\Pr_{\nu,\tilde{\text{alg}}}(\tilde{\tau}_{1,a} \leq \tau < +\infty) \\
=&\Pr_{\nu,\tilde{\text{alg}}}(n^{\tau}_{h_0}(s_0,1) + n^{\tau}_{h_0}(s_0,a) \geq \frac{1}{M\epsilon^2}, \tau < +\infty) \\
\geq&\Pr_{\nu,\tilde{\text{alg}}}(\hat{\pi}_{h_0}(s_0) = 1, n^{\tau}_{h_0}(s_0,1) + n^{\tau}_{h_0}(s_0,a) \geq \frac{1}{M\epsilon^2}, \tau < +\infty) \\
\geq&\frac{1}{14}.
\end{aligned}$$

Denote $\mathcal{E}^{\text{pull}}_{1,a} = \{\tilde{\tau}_{1,a} \leq \tau < +\infty, n^{\tilde{\tau}_{1,a}}_{h_0}(s_0,1) \leq n^{\tilde{\tau}_{1,a}}_{h_0}(s_0,a)\}$ and $\mathcal{E}^{\text{pull}}_{a,1} = \{\tilde{\tau}_{1,a} \leq \tau < +\infty, n^{\tilde{\tau}_{1,a}}_{h_0}(s_0,a) \leq n^{\tilde{\tau}_{1,a}}_{h_0}(s_0,1)\}$. Evident to see

$$\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}^{\text{pull}}_{1,a}) + \Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}^{\text{pull}}_{a,1}) \geq \Pr_{\nu,\tilde{\text{alg}}}(\tilde{\tau}_{1,a} \leq \tau < +\infty) \geq \frac{1}{14}. \tag{63}$$

We consider a permutation $\sigma^{(a)}$ with $\sigma^{(a)}_{s,h}(\tilde{a}) = \begin{cases} 1 & s = s_0, h = h_0, \tilde{a} = a \\ a & s = s_0, h = h_0, \tilde{a} = 1 \\ \tilde{a} & \text{else} \end{cases}$. By the symmetry of $\tilde{\text{alg}}$, we know

$$\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}^{\text{pull}}_{1,a}) = \Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{\text{pull}}_{a,1}), \Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}^{\text{pull}}_{a,1}) = \Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{\text{pull}}_{1,a}). \tag{64}$$

Apply the Transportation Equality(Lemma 18 in Kaufmann et al. (2016)) on the event $\mathcal{E}^{\text{pull}}_{1,a}$, we can derive

$$\Pr_{\nu_{\sigma^{(a)}},\tilde{\text{alg}}}(\mathcal{E}^{\text{pull}}_{a,1}) = \mathbb{E}_{\nu,\tilde{\text{alg}}}\mathbb{1}(\mathcal{E}^{\text{pull}}_{a,1}) \exp\left(-\sum_{t=1}^{n^{\tilde{\tau}_{1,a}}_{h_0}(s_0,1)} \hat{\text{KL}}_{1,t} - \sum_{t=1}^{n^{\tilde{\tau}_{1,a}}_{h_0}(s_0,a)} \hat{\text{KL}}_{a,t}\right), \tag{65}$$

where $\hat{\text{KL}}_{1,t} = X_{1,t} \log \frac{r+\epsilon}{r} + (1 - X_{1,t}) \log \frac{1-r-\epsilon}{1-r}$, $\hat{\text{KL}}_{a,t} = X_{a,t} \log \frac{r}{r+\epsilon} + (1 - X_{1,t}) \log \frac{1-r}{1-r-\epsilon}$, $X_{1,t}|\{X_{1,s}\}_{s=1}^{t-1} \sim \text{Bern}(r+\epsilon)$, $X_{a,t}|\{X_{a,s}\}_{s=1}^{t-1} \sim \text{Bern}(r)$. $\{X_{1,s}\}_{s=1}^{+\infty}$ are independent with

$\{X_{a,s}\}_{s=1}^{+\infty}$. Similar to the proof of Lemma 32, we introduce

$$\begin{aligned}
\mathrm{kl}(x,y) =& x\log\frac{x}{y}+(1-x)\log\frac{1-x}{1-y}\\
\mathcal{E}_{\text{kl-cnt}}^{(1)} =& \left\{\max_{1\le n\le\frac{1}{M\epsilon^2}}\sum_{t=1}^{n}(\hat{\mathrm{KL}}_{1,t}-\mathrm{kl}(r+\epsilon,r))\le\frac{1}{4}\right\}\\
\mathcal{E}_{\text{kl-cnt}}^{(a)} =& \left\{\max_{1\le n\le\frac{1}{M\epsilon^2}}\sum_{t=1}^{n}(\hat{\mathrm{KL}}_{a,t}-\mathrm{kl}(r,r+\epsilon))\le\frac{1}{4}\right\}
\end{aligned}$$

Following (65) and (64), we have

$$\begin{aligned}
&\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}_{1,a}^{\text{pull}})\\
\ge&\mathbb{E}_{\nu,\tilde{\text{alg}}}\mathbb{1}(\mathcal{E}_{a,1}^{\text{pull}}\cap\mathcal{E}_{\text{kl-cnt}}^{(1)}\cap\mathcal{E}_{\text{kl-cnt}}^{(a)})\exp\left(-\sum_{t=1}^{n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1)}\hat{\mathrm{KL}}_{1,t}-\sum_{t=1}^{n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)}\hat{\mathrm{KL}}_{a,t}\right)\\
=&\mathbb{E}_{\nu,\tilde{\text{alg}}}\mathbb{1}(\mathcal{E}_{a,1}^{\text{pull}}\cap\mathcal{E}_{\text{kl-cnt}}^{(1)}\cap\mathcal{E}_{\text{kl-cnt}}^{(a)})\exp(-n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1)\mathrm{kl}(r+\epsilon,r)-n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)\mathrm{kl}(r,r+\epsilon))\\
&\exp\left(-\sum_{t=1}^{n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1)}\left(\hat{\mathrm{KL}}_{1,t}-\mathrm{kl}(r+\epsilon,r)\right)-\sum_{t=1}^{n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)}\left(\hat{\mathrm{KL}}_{a,t}-\mathrm{kl}(r,r+\epsilon)\right)\right)\\
\ge&\mathbb{E}_{\nu,\tilde{\text{alg}}}\mathbb{1}(\mathcal{E}_{a,1}^{\text{pull}}\cap\mathcal{E}_{\text{kl-cnt}}^{(1)}\cap\mathcal{E}_{\text{kl-cnt}}^{(a)})\exp(-\frac{3\epsilon^2}{M\epsilon^2}-\frac{3\epsilon^2}{M\epsilon^2})\exp(-\frac{1}{2}) \qquad (66)\\
\ge&\exp(-\frac{6}{M}-\frac{1}{2})\left(\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}_{a,1}^{\text{pull}})-\Pr_{\nu,\tilde{\text{alg}}}(\neg\mathcal{E}_{\text{kl-cnt}}^{(1)})-\Pr_{\nu,\tilde{\text{alg}}}(\neg\mathcal{E}_{\text{kl-cnt}}^{(a)})\right)\\
\ge&\exp(-\frac{6}{M}-\frac{1}{2})\left(\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}_{a,1}^{\text{pull}})-2\exp(-\frac{M}{256})\right) \qquad (67)
\end{aligned}$$

Step (66) is by the fact that $n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1), n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)\le n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1)+n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)\le\frac{1}{M\epsilon^2}$, which is guaranteed by the definition of stopping time $\tilde{\tau}_{1,a}$. And $\mathcal{E}_{\text{kl-cnt}}^{(1)},\mathcal{E}_{\text{kl-cnt}}^{(a)}$ guarantee

$$\exp\left(-\sum_{t=1}^{n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1)}\left(\hat{\mathrm{KL}}_{1,t}-\mathrm{kl}(r+\epsilon,r)\right)-\sum_{t=1}^{n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)}\left(\hat{\mathrm{KL}}_{a,t}-\mathrm{kl}(r,r+\epsilon)\right)\right)\ge\exp(\frac{1}{2}).$$

Step (67) is by the fact that $\Pr_{\nu,\tilde{\text{alg}}}(\neg\mathcal{E}_{\text{kl-cnt}}^{(1)}), \Pr_{\nu,\tilde{\text{alg}}}(\neg\mathcal{E}_{\text{kl-cnt}}^{(a)}), \le\exp(-\frac{M}{256})$, proved by Lemma 35 and 36.

Combining (63) and (67), we have

$$\begin{aligned}
&\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}_{1,a}^{\text{pull}})\ge\exp(-\frac{6}{M}-\frac{1}{2})\left(\frac{1}{14}-\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}_{1,a}^{\text{pull}})-2\exp(-\frac{M}{256})\right)\\
\Leftrightarrow&\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}_{1,a}^{\text{pull}})\ge\frac{\exp(-\frac{6}{M}-\frac{1}{2})\left(\frac{1}{14}-2\exp(-\frac{M}{256})\right)}{1+\exp(-\frac{6}{M}-\frac{1}{2})}\\
\Rightarrow&\Pr_{\nu,\tilde{\text{alg}}}(\mathcal{E}_{1,a}^{\text{pull}})\ge\frac{1}{50}\\
\Rightarrow&\Pr_{\nu,\tilde{\text{alg}}}\left(\tilde{\tau}_{1,a}\le\tau<+\infty, n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)+n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1)=\frac{1}{M\epsilon^2}, n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,a)\ge n_{h_0}^{\tilde{\tau}_{1,a}}(s_0,1)\right)\ge\frac{1}{50}\\
\Rightarrow&\Pr_{\nu,\tilde{\text{alg}}}\left(n_{h_0}^{\tau}(s_0,a)\ge\frac{1}{2M\epsilon^2},\tau<+\infty\right)\ge\frac{1}{50}
\end{aligned}$$

The last line suggests that we have proved the Lemma. □

Given Lemma 33, we are ready to prove the lower bounds for expected pulling times.

Lemma 34. *Apply a δ-PAC symmetric algorithm $a\tilde{l}g$ to a Uniform Instance ν with (S, A, H, r, ϵ). If $\exists s_0 \in [S-2]$ and even h_0 such that $\Pr_{\nu, a\tilde{l}g}(\hat{\pi}_{h_0}(s_0) = 1) \geq \frac{1}{7}$, we have*

$$\mathbb{E}_{\nu, a\tilde{l}g} n^{\tau}_{h_0}(s_0, a) \geq \frac{1}{100 M \epsilon^2}, \forall a \in \{2, \cdots, A\}$$

where $M = 2000$.

Proof of Lemma 34. By the Lemma 33, we know

$$\Pr_{\nu, \tilde{\text{alg}}} \left(n^{\tau}_{h_0}(s_0, a) \geq \frac{1}{2M\epsilon^2} \right) \geq \frac{1}{50}$$

holds for all $a \in \{2, \cdots, A\}$. We can prove Lemma 34 by applying the Markov Inequality. □

D TECHNICAL INEQUALITY

In this section, we present technical inequalities used by the above proof for Lemma and Theorems, including multiple inequalities and concentration theorems.

Lemma 35. *Consider Random variable $\{X_t\}_{t=1}^{+\infty}$ such that $X_t | \{X_s\}_{s=1}^{t-1}$ is σ^2-subgaussian with mean value 0. Then for an integer N, we have*

$$\Pr\left(\max_{1 \leq n \leq N} \sum_{t=1}^{n} X_t > x \right) \leq \exp(-\frac{x^2}{2N\sigma^2}).$$

Proof of Lemma 35. Take $\lambda = \frac{x}{N\sigma^2}$. We first validate $\{\exp(\lambda \sum_{t=1}^{n} X_t)\}_{n=1}^{+\infty}$ is a submartingale. By Jensen's Inequality,

$$\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n} X_t \right) | X_1, \cdots, X_{n-1} \right] \geq \exp\left(\lambda \sum_{t=1}^{n-1} X_t \right) \exp(\mathbb{E} \lambda X_n) = \exp\left(\lambda \sum_{t=1}^{n-1} X_t \right).$$

By direct calculation, we have

$$\begin{aligned}
&\Pr\left(\max_{1 \leq n \leq N} \sum_{t=1}^{n} X_t > x \right) \\
=& \Pr\left(\max_{1 \leq n \leq N} \exp\left(\lambda \sum_{t=1}^{n} X_t \right) > \exp(\lambda x) \right) \\
\leq& \frac{\mathbb{E} \exp\left(\lambda \sum_{t=1}^{N} X_t \right)}{\exp(\lambda x)} \qquad (68) \\
\leq& \frac{\exp(\frac{N\lambda^2\sigma^2}{2})}{\exp(\lambda x)} \\
=& \exp(-\frac{x^2}{2N\sigma^2}). \qquad (69)
\end{aligned}$$

Step (68) is by the maximal inequality for the submartingale. □

Lemma 36. *Assume $p, q \in (\frac{1}{4}, \frac{3}{4})$, and $X \sim$ Bern(p), let*

$$Z := X \log \frac{p}{q} + (1 - X) \log \frac{1-p}{1-q} - (p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}),$$

Z is $8(p-q)^2$-subgaussian, i.e. for any $\lambda \in \mathbb{R}$, we have $\mathbb{E}[\exp(\lambda Z) \leq \exp(4\lambda^2 (p-q)^2)$

Proof of Lemma 36. Not hard to see $\mathbb{E}Z = 0$ and

$$\begin{aligned} Z =& (X-p)\left(\log\frac{p}{q} - \log\frac{1-p}{1-q}\right) \\ =& (X-p)\log\frac{\frac{p}{1-p}}{\frac{q}{1-q}} \end{aligned}$$

Since $X - p$ is a $\frac{1}{4}$-subgaussian random variable, we suffice to show $\frac{1}{4}\left(\log\frac{\frac{p}{1-p}}{\frac{q}{1-q}}\right)^2 \leq 8(p-q)^2, \forall p, q \in (\frac{1}{4}, \frac{3}{4})$. Notice that

$$\begin{aligned} &\frac{1}{4}\left(\log\frac{\frac{p}{1-p}}{\frac{q}{1-q}}\right)^2 \leq 8(p-q)^2 \\ \Leftrightarrow &|\frac{\log\frac{p}{1-p} - \log\frac{q}{1-q}}{p-q}| \leq 2\sqrt{8} \\ \Leftarrow &\sup_{\frac{1}{4}\leq x\leq\frac{3}{4}} |\frac{d\log\frac{x}{1-x}}{dx}| \leq 2\sqrt{8} \\ \Leftrightarrow &\sup_{\frac{1}{4}\leq x\leq\frac{3}{4}} |\frac{1}{x} + \frac{1}{1-x}| \leq 2\sqrt{8} \end{aligned}$$

The second last step is by the median value theorem. Since $\sup_{\frac{1}{4}\leq x\leq\frac{3}{4}} |\frac{1}{x} + \frac{1}{1-x}| = |\frac{1}{x} + \frac{1}{1-x}|_{x=\frac{1}{4} \text{ or } \frac{3}{4}} = \frac{16}{3} < 2\sqrt{8}$, we have completed the proof. □

Lemma 37. *Define* $kl(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ *. For* x, y *such that* $\frac{1}{4} < x, y < \frac{3}{4}$*, we have*

$$kl(x,y) \leq 3(x-y)^2$$

Proof of Lemma 37. Given $x \in (\frac{1}{4}, \frac{4}{3})$, define $f(y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y} - 3(x-y)^2$. We have

$$\begin{aligned} f'(y) =& -\frac{x}{y} + \frac{1-x}{1-y} - 6(y-x) \\ =& \frac{-x+y}{y(1-y)} - 6(y-x) \\ =& (y-x)\left(\frac{1}{y(1-y)} - 6\right) \end{aligned}$$

Notice that $\frac{3}{16} \leq y(1-y) \leq \frac{1}{4}$ holds for $y \in (\frac{1}{4}, \frac{3}{4})$, which means $\frac{1}{y(1-y)} - 6 \leq 0$ holds for $y \in (\frac{1}{4}, \frac{3}{4})$. We can conclude $f'(y)$ is decreasing in the interval $(x, \frac{3}{4})$ and increasing in the interval $(\frac{1}{4}, x)$. That means $\max_{\frac{1}{4}<y<\frac{3}{4}} f(y) = f(x) = 0$, which proves the conclusion. □

Lemma 38. *Given any* $b \geq a \geq 2$*. If* $x > 0$ *satisfies* $x \leq b + a\log x$*, we can conclude* $x < b + 3a\log b$*.*

Proof. Prove by contradiction. Assume $x \geq b + 3a\log b$. Notice that $\frac{d(x - a\log x - b)}{dx} = 1 - \frac{a}{x}$, which implies $x - a\log x - b$ increases at the interval $(b + 3a\log b, +\infty)$. In addition, we have

$$\begin{aligned} & x - a\log x - b|_{x=b+3a\log b} \\ =& b + 3a\log b - a\log(b + 3a\log b) - b \\ =& 3a\log b - a\log(b + 3a\log b) \\ =& a\log\frac{b^3}{b + 3a\log b} \\ \geq& a\log\frac{b^3}{b + 3b\log b} \\ =& a\log\frac{b^2}{1 + 3\log b}. \end{aligned}$$

As $b^2 \geq 1 + 3\log b$ for any $b \geq 2$, ($2^2 \geq 1 + 3\log 2$), we can conclude $x - a\log x - b \geq 0$ holds for all $x \geq b + 3a\log b$, which implies a contradiction. □

Lemma 39. $\log(a+b) \leq \log(4ab) = \log(2a) + \log(2b)$ *for* $a, b \geq 1$

Proof. Notice that for $a, b \geq 1$, we have $4ab - a - b = a(2b-1) + b(2a-1) > 0$, we can conclude $\log(a+b) \leq \log(4ab)$. □

Lemma 40. *Given* $a \geq 1$*, we have* $x \geq 2a\log(3a) \Rightarrow x \geq a\log x$

Proof. Define $f(x) = x - a\log x$. By $\frac{df}{dx} = 1 - \frac{a}{x}$, we know f is increasing in the interval $(a, +\infty)$. By the fact that $a \geq 1$, we have $\log(3a) \geq 1$, further $2a\log(3a) \geq a$. Meanwhile, not hard to see

$$\begin{aligned}
&a \geq 1 \\
\Rightarrow &\log(3a) \geq \log(2a), \log(3a) \geq \log(\log(3a)) \\
\Rightarrow &2\log(3a) \geq \log(2a) + \log(\log(3a)) \\
\Leftrightarrow &2\log(3a) \geq \log(2a\log(3a)) \\
\Rightarrow &2a\log(3a) - a\log(2a\log(3a)) \geq 0.
\end{aligned}$$

We can conclude $f(2a\log(3a)) \geq 0$, which means $x \geq 2a\log(3a) \Rightarrow x \geq a\log x$. □

Lemma 41 (Lemma 19 in Jaksch et al. (2010))**.** *For* $0 \leq z_k \leq 1$*,* $M \geq 0$*, we have* $\sum_{k=1}^{n} \frac{z_k}{\sqrt{\max\{M + \sum_{i=1}^{k-1} z_i, 1\}}} \leq (\sqrt{2}+1)\sqrt{\sum_{i=1}^{n} z_i}$

Proof. Denote $Z_{k-1} := \max\left\{1, \sum_{i=1}^{k-1} z_i\right\}$. Trivial to see $Z_{k-1} \geq z_k$. We use induction over n. First, we show the lemma holds for $n = 1$. The reason is

$$\sum_{k=1}^{1} \frac{z_k}{\sqrt{\max\{M + \sum_{i=1}^{n-1} z_i, 1\}}} = \frac{z_1}{\max\{M, 1\}} \leq z_1 \leq \sqrt{z_1} < (\sqrt{2}+1)\sqrt{z_1}.$$

Then, we turn to work on two cases for a general n, given $n-1$ is correct, we consider two cases $\sum_{k=1}^{n-1} z_k \leq 1$ or $\sum_{k=1}^{n-1} z_k > 1$. If $\sum_{k=1}^{n-1} z_k \leq 1$, we have $\sum_{k=1}^{n} z_k = z_n + \sum_{k=1}^{n-1} z_k \leq 2$, which implies $\sum_{k=1}^{n} z_k \leq \sqrt{2}\sqrt{\sum_{k=1}^{n} z_k}$. It follows that

$$\begin{aligned}
&\sum_{k=1}^{n} \frac{z_k}{\sqrt{\max\{M + \sum_{i=1}^{k-1} z_i, 1\}}} \\
\leq &\sum_{k=1}^{n-1} z_k + z_n \\
\leq &\sqrt{2}\sqrt{\sum_{k=1}^{n} z_k} \\
< &(\sqrt{2}+1)\sqrt{\sum_{k=1}^{n} z_k}.
\end{aligned}$$

Then, we show that the lemma also holds for n, when $\sum_{k=1}^{n-1} z_k > 1$. By the induction hypothesis, we have

$$\begin{aligned}
&\sum_{k=1}^{n} \frac{z_k}{\max\{M + \sum_{i=1}^{k-1} z_i, 1\}} \\
\leq&(\sqrt{2}+1)\sqrt{\sum_{k=1}^{n-1} z_k + \frac{z_n}{\sqrt{\max\{\sum_{i=1}^{n-1} z_i, 1\}}}} \\
=&(\sqrt{2}+1)\sqrt{\sum_{k=1}^{n-1} z_k + \frac{z_n}{\sqrt{\sum_{i=1}^{n-1} z_i}}} \\
=&\sqrt{\left((\sqrt{2}+1)\sqrt{\sum_{k=1}^{n-1} z_k + \frac{z_n}{\sqrt{\sum_{i=1}^{n-1} z_i}}}\right)^2} \\
=&\sqrt{(\sqrt{2}+1)^2(\sum_{k=1}^{n-1} z_k) + 2(\sqrt{2}+1)z_n + \frac{z_n^2}{\sum_{k=1}^{n-1} z_k}} \\
\overset{z_n \leq 1, \sum_{k=1}^{n-1} z_k \geq 1}{\leq}&\sqrt{(\sqrt{2}+1)^2(\sum_{k=1}^{n-1} z_k) + 2(\sqrt{2}+1)z_n + z_n} \\
=&\sqrt{(2\sqrt{2}+3)(\sum_{k=1}^{n-1} z_k) + (2\sqrt{2}+3)z_n} \\
=&(\sqrt{2}+1)\sqrt{\sum_{k=1}^{n} z_k},
\end{aligned}$$

which means the induction holds. □

The remaining two lemmas are from Li & Cheung (2025). We don't repeat the proof as they are exactly the same.

Lemma 42 (Lemma D.2 in Li & Cheung (2025)). *For any $\Delta \in (0, 1], K \geq 2, \delta \in (0, \frac{1}{2}], C \geq 1$, we can conclude*

$$\begin{aligned}
t >& \frac{28C^2 \log \frac{2K}{\delta}}{\Delta^2} + \frac{16C^2 \log\left(\log\left(\frac{24C^2}{\Delta^2}\right)\right)}{\Delta^2} \\
\Rightarrow& C\sqrt{\frac{4 \log \frac{2K(\log_2 2t)^2}{\delta}}{t}} < \Delta
\end{aligned}$$

Lemma 43 (Lemma D.3 in Li & Cheung (2025)). *Assume $\{X_i\}_{i=1}^{+\infty}$ are i.i.d random variables with mean value μ. Assume $\{X_i - \mu\}_{i=1}^{+\infty}$ are σ^2-subgaussian. We have*

$$\Pr\left(\forall N \in \mathbb{N}, \left|\frac{\sum_{i=1}^{N} X_i}{N} - \mu\right| < \sqrt{\frac{4\sigma^2 \log \frac{2(\log_2 2N)^2}{\delta}}{N}}\right) \geq 1 - \frac{\pi^2}{6}\delta$$

holds for all $\delta > 0$.

E Details of Numeric Experiments

The details of 5 experiment groups are as follows. The environmental type Single Chain only differs in the starting state with Double Chain.

1. In the instances "Pos: Single_Chain", we set $\delta = 0.001$, $H = 8$, $\mathcal{S} = \{0, 1, 2, 3\}$, and $\mathcal{A} = \{\text{left}, \text{right}\}$. The transition dynamics are defined such that the "intended" direction occurs with probability 0.9 and the "reverse" direction with probability 0.1. Specifically,
   - $p_h(\max\{s-1, 0\} \mid s, \text{left}) = 0.9$, $p_h(\min\{s+1, 3\} \mid s, \text{left}) = 0.1$,
   - $p_h(\min\{s+1, 3\} \mid s, \text{right}) = 0.9$, $p_h(\min\{s-1, 0\} \mid s, \text{right}) = 0.1$,

   The agent always starts at state 0, $r_h(s, a) = \begin{cases} 1 & s = 3 \\ 0 & \text{else} \end{cases}$. Easy to calculate $V^* \approx 4.06$.

   We take $\mu_0 \in \{1, 1.5, 2, 2.5, 3\}$.
2. In the instances "Pos: Double_Chain", we set $\delta = 0.001$, $H = 8$, $\mathcal{S} = \{0, 1, 2, 3\}$, and $\mathcal{A} = \{\text{left}, \text{right}\}$. The transition dynamics are defined such that the "intended" direction occurs with probability 0.9 and the "reverse" direction with probability 0.1. Specifically,
   - $p_h(\max\{s-1, 0\} \mid s, \text{left}) = 0.9$, $p_h(\min\{s+1, 3\} \mid s, \text{left}) = 0.1$,
   - $p_h(\min\{s+1, 3\} \mid s, \text{right}) = 0.9$, $p_h(\min\{s-1, 0\} \mid s, \text{right}) = 0.1$,

   The agent always starts at state 1 with $r_h(s, a) = \begin{cases} 1 & s = 3 \\ 0 & \text{else} \end{cases}$. Easy to calculate $V^* \approx 5.04$.

   We take $\mu_0 \in \{1, 1.5, 2, 2.5, 3\}$.
3. In the instances "Pos: Grid_World", we set $\delta = 0.001$, $H = 5$, $\mathcal{S} = \{(i, j) : i, j \in \{0, 1, 2, 3\}\}$ and $\mathcal{A} = \{\text{up}, \text{left}, \text{right}, \text{down}\}$. The transition dynamics are defined such that the "intended" direction occurs with probability 0.9 and the "reverse" direction with probability 0.1. Specifically, at a state $s = (i, j)$, we define $s_{\text{up}} = (i, \max\{j-1, 0\})$, $s_{\text{right}} = (\min\{i+1, 3\}, j)$, $s_{\text{down}} = (i, \min\{j+1, 3\})$, $s_{\text{left}} = (\max\{i-1, 0\}, j)$. And
   - $p_h(s_{\text{up}} \mid s, \text{up}) = 0.95$, $p_h(s_{\text{down}} \mid s, \text{up}) = 0.05$.
   - $p_h(s_{\text{right}} \mid s, \text{right}) = 0.95$, $p_h(s_{\text{left}} \mid s, \text{right}) = 0.05$.
   - $p_h(s_{\text{down}} \mid s, \text{down}) = 0.95$, $p_h(s_{\text{up}} \mid s, \text{down}) = 0.05$.
   - $p_h(s_{\text{left}} \mid s, \text{left}) = 0.95$, $p_h(s_{\text{right}} \mid s, \text{left}) = 0.05$.

   The agent always starts at the state $(1, 1)$, with $r_h(s, a) = \begin{cases} 1 & s = (2, 2) \\ 0 & \text{else} \end{cases}$. Easy to calculate $V^* \approx 1.85$.

   We take $\mu_0 \in \{0.15, 0.3, 0.45, 0.6, 0.75\}$.
4. In the instances "Neg: Single_Chain", we set $\delta = 0.001$, $H = 8$, $\mathcal{S} = \{0, 1, 2, 3\}$, and $\mathcal{A} = \{\text{left}, \text{right}\}$. The transition dynamics are defined such that the "intended" direction occurs with probability 0.9 and the "reverse" direction with probability 0.1. Specifically,
   - $p_h(\max\{s-1, 0\} \mid s, \text{left}) = 0.9$, $p_h(\min\{s+1, 3\} \mid s, \text{left}) = 0.1$,
   - $p_h(\min\{s+1, 3\} \mid s, \text{right}) = 0.9$, $p_h(\min\{s-1, 0\} \mid s, \text{right}) = 0.1$,

   The agent always starts at state 0, $r_h(s, a) = \begin{cases} 1 & s = 3 \\ 0 & \text{else} \end{cases}$. Given these setting, $V^* \approx 4.06$.

   We take $\mu_0 \in \{5, 5.25, 5.5, 5.75, 6\}$.
5. In the instances "Neg: Double_Chain", we set $\delta = 0.001$, $H = 8$, $\mathcal{S} = \{0, 1, 2, 3\}$, and $\mathcal{A} = \{\text{left}, \text{right}\}$. The transition dynamics are defined such that the "intended" direction occurs with probability 0.9 and the "reverse" direction with probability 0.1. Specifically,
   - $p_h(\max\{s-1, 0\} \mid s, \text{left}) = 0.9$, $p_h(\min\{s+1, 3\} \mid s, \text{left}) = 0.1$,
   - $p_h(\min\{s+1, 3\} \mid s, \text{right}) = 0.9$, $p_h(\min\{s-1, 0\} \mid s, \text{right}) = 0.1$,

   The agent always starts at state 1, $r_h(s, a) = \begin{cases} 1 & s = 3 \\ 0 & \text{else} \end{cases}$. Given these setting, $V^* \approx 5.04$.

   We take $\mu_0 \in \{5.75, 6, 6.25, 6.5, 6.75\}$.

BPI_UCRL and our proposed BEE-GPI were implemented using the Academic MOSEK solver, to solve the convex exponential cone optimization problem in (1).