Instance-Dependent Fixed-Budget Pure Exploration in Reinforcement Learning

Yeongjong Kim
Department of Mathematics
POSTECH
kim.yj@postech.ac.kr

Yeoneung Kim
Department of Applied Artificial Intelligence
Seoultech
kimyeoneung@gmail.com

Kwang-Sung Jun
Department of Computer Science
University of Arizona
kjun@cs.arizona.edu

Abstract

We study the problem of fixed-budget pure exploration in reinforcement learning. The goal is to identify a near-optimal policy, given a fixed budget on the number of interactions with the environment. Unlike the standard PAC setting, we do not require the target error level ε and failure rate δ as input. We propose novel algorithms and provide, to the best of our knowledge, the first instance-dependent theoretical guarantee for this setting. Our analysis yields an ε-correctness guarantee with instance-dependent probability, characterizing the budget requirements in terms of the problem-specific hardness of exploration. As a core component of our analysis, we derive an ε-good guarantee for the multiple bandit problem—solving multiple multi-armed bandit instances simultaneously—which may be of independent interest. To enable our analysis, we also develop tools for reward-free exploration under the fixed-budget setting, which we believe will be useful for future work in this area.

1 Introduction

Reinforcement Learning (RL) theory [1] has been studied under two main objectives: regret minimization and policy identification, also known as pure exploration. While the former focuses on maximizing cumulative reward during learning, the latter aims to identify a near-optimal policy without concern for rewards gained during learning. A substantial body of work on policy identification has focused on the fixed-confidence setting [16]. This line of research, often referred to as Probably Approximately Correct (PAC) RL, requires the algorithm to spend as many samples as possible until it can find an ε-optimal policy with probability at least $1 - \delta$. Specifically, the algorithm is required to *verify* itself that the returned arm is indeed ε-optimal policy – otherwise, it is not a fixed confidence algorithm. Due to the verification requirement, both ε and δ are input to the algorithm. Thus, the analysis must be done for the correctness of the *verification* (i.e., proving that the returned arm is indeed an ε-optimal policy) as well as the sample complexity (i.e., proving how many samples are taken before stopping).

However, the fixed-confidence setting is not the only way to perform policy identification. The fixed-budget setting has been popular in multi-armed bandits [9, 4]. In this setting, the learner is given a fixed number of interactions with the environment as a budget and is required to output a good policy after exhausting the budget. This setting has numerous merits. First, this setting is arguably more practical because the user of the algorithm can control the budget explicitly. In contrast, the fixed confidence setting assumes that the algorithm can use as many samples as possible (though less is preferred). When stopped forcefully to satisfy practical constraints, it is hard to guarantee the

First Exploration in AI Today Workshop at ICML (EXAIT at ICML 2025).

quality of the returned policy. Second, the fixed budget setting has potential to guarantee a better sample complexity because there is no verification requirement (i.e., the algorithm itself certifies that the returned policy is ε-optimal). This was true for multi-armed bandits where instant-dependent accelerated rates can be obtained as a function of how many good arms there are, and also a data-poor regime guarantee can be obtained, meaning that where a nontrivial performance guarantee is obtained even if the sampling budget is smaller than the number of arms, depending on the problem instance [26]. These bounds are not likely to be obtained in the fixed confidence setting due to the verification requirement unless extra knowledge about the best arm is known such as Chaudhuri and Kalyanakrishnan [6]. While the ε-correctness verification from the fixed-confidence setting can be necessary in mission-critical applications, there are many applications that do not require such a guarantee, in which case the parameters ε and δ becomes a cumbersome hyperparameter.

Despite the desirable properties of the fixed-budget setting in bandit problems, its counterpart in MDPs remains largely unexplored to our knowledge. In this paper, we take the first step at studying fixed-budget policy identification in MDPs, providing new theoretical insights and algorithms that bridge this gap. Specifically, a fixed budget algorithm is required to take in a episode budget B and return a policy $\hat{\pi}$ at the end of B-th episode. Our central interest is to upper bound the probability that the algorithm returns an ε-optimal policy as an exponentially decaying function of the budget B and instance-dependent quantities, *simultaneously for all* $\varepsilon \geq \varepsilon'$ for some budget dependent threshold ε'. In other words, the degree of suboptimality of the learned policy $\hat{\pi}$ is a random variable, and we are characterizing its distribution, in particular its tail behavior.

Contributions. Our main contributions are as follows:

- We propose a novel algorithm, **BREA** (Backward Reachability Estimation and Action elimination), which is, to the best of our knowledge, the first fixed-budget pure exploration algorithm for episodic MDPs with instance-dependent guarantees. The algorithm does not require the accuracy level ε as an input, and does not assume the uniqueness of the optimal action.
- For the first time, we establish an ε-correctness guarantee for the SAR algorithm in the multiple bandit setting, extending its original best arm identification result [5]. This extension may be of independent interest.
- We develop algorithmic and analytical tools that are broadly useful for reward-free exploration in the fixed-budget setting.

2 Preliminaries

Finite-horizon MDP. We consider a finite-horizon non-stationary Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=0}^{H-1}, \{R_h\}_{h=1}^{H})$, where $\mathcal{S}$ is a finite set of states of size S, $\mathcal{A}$ is a finite set of actions of size A, $H \in \mathbb{N}$ is the horizon, $P_0 \in \Delta(\mathcal{S})$ is the initial distribution, $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, and $R_h : \mathcal{S} \times \mathcal{A} \to \Delta([0,1])$ is the random rewards with $\mathbb{E}[R_h(s,a)] = r_h(s,a)$. $\{P_h\}_{h=0}^{H-1}$ and $\{R_h\}_{h=1}^{H}$ are unknown to the learner.

The initial state s_1 is drawn from the initial distribution P_0. At each step h, taking action a_h in state s_h results in a next state s_{h+1} sampled from the transition kernel $P_h(\cdot \mid s_h, a_h)$. A trajectory $\{(s_h, a_h, R_h(s_h, a_h))\}_{h=1}^{H}$ is called an *episode*, and when the learner reaches the end of the episode, a new episode begins.

A *policy* $\pi = (\pi_1, \ldots, \pi_H)$ is a sequence of decision rules $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ for each step $h \in [H]$. The Q-value function of a policy π at step $h \in [H]$ is defined as

$$Q_h^\pi(s,a) := \mathbb{E}_\pi[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a]$$

and it represents the expected reward obtained by choosing action a in state s at step h and choosing the subsequent actions according to the policy π. The value function of π at step h is defined as

$$V_h^\pi(s) = \mathbb{E}_\pi[Q_h^\pi(s, \pi_h(s))]$$

and it represents the expected reward obtained by choosing actions according to the policy π starting in state s at step h. We also define $V_0^\pi := \mathbb{E}_{s\sim P_0}[V_1^\pi(s)]$. The optimal Q-value function, optimal value function are defined as

$$Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a), \quad V_h^*(s) = \sup_\pi V_h^\pi(s), \quad V_0^* = \sup_\pi V_0^\pi.$$

We denote the optimal policy by π^*, which satisfies $V_0^{\pi^*} = V_0^*$.

Pure exploration under the fixed budget setting. In pure exploration under the fixed budget setting, the goal is to identify an optimal policy π^* (or near-optimal) based on a limited interaction budget. Specifically, the learner is allowed to execute a total of B episodes and must return a single policy $\hat{\pi}$ at the end. The performance is measured by the simple regret, which is defined as

$$V_0^* - V_0^{\hat{\pi}}.$$

A policy $\hat{\pi}$ is called ε*-good* if $V_0^* - V_0^{\hat{\pi}} \leq \varepsilon$. In this paper, we propose an algorithm and prove their performance guarantee by showing some instance-dependent upper bounds of the failure probability

$$\mathbb{P}(V_0^* - V_0^{\hat{\pi}} > \varepsilon).$$

Instance-dependent quantities. To capture the instance-dependent complexity of the problem, we need the notion of *suboptimality gaps* defined as

$$\begin{aligned} \Delta_h(s,a) &:= V_h^*(s) - Q_h^*(s,a), \\ \Delta_h^\pi(s,a) &:= V_h^\pi(s) - Q_h^\pi(s,a). \end{aligned}$$

For our analysis, we also denote

$$\bar{\Delta}_h(s,a) := \begin{cases} \Delta_h(s,a), & \text{if } Q_h^*(s,a) < V_h^*(s) \\ \Delta_h(s,a'), & \text{if } Q_h^*(s,a) = V_h^* \text{ and } a' \text{ is the second best action,} \end{cases}$$

$$\bar{\Delta}_h^\pi(s,a) := \begin{cases} \Delta_h^\pi(s,a), & \text{if } Q_h^\pi(s,a) < V_h^\pi(s) \\ \Delta_h^\pi(s,a'), & \text{if } Q_h^\pi(s,a) = V_h^\pi \text{ and } a' \text{ is the second best action with respect to } \pi. \end{cases}$$

Thus, if the optimal action in $s \in \mathcal{S}$ at step h is unique, $\bar{\Delta}_h(s,a) > 0$ for all $a \in \mathcal{A}$. In contrast, if there are multiple optimal actions in $s \in \mathcal{S}$ at step h, $\bar{\Delta}_h(s,a) = 0$ for all optimal actions a. Similar results hold for $\bar{\Delta}_h(s,a)$ as well.

In MDP, the probability of reaching each state or action is important. Let π be a policy, $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], \mathcal{Z} \subset \mathcal{S} \times \mathcal{A}$, we use the following notations:

$$\begin{aligned} w_h^\pi(s) &= \mathbb{P}_\pi[s_h = s], & w_h^\pi(s,a) &= \mathbb{P}_\pi[s_h = s, a_h = a], & w_h^\pi(\mathcal{Z}) &= \mathbb{P}_\pi[(s_h, a_h) \in \mathcal{Z}], \\ W_h(s) &= \sup_\pi w_h^\pi(s), & W_h(s,a) &= \sup_\pi w_h^\pi(s,a), & W_h(\mathcal{Z}) &= \sup_\pi w_h^\pi(\mathcal{Z}). \end{aligned}$$

We refer to $w_h^\pi(\cdot)$ as the *occupancy measure* and $W_h(\cdot)$ as the *reachability*. Using these notions, we define the *controllability* of MDP at step h as

$$C_h := \sup_\pi \sum_{s, W_h(s) > 0} \frac{w_h^\pi(s)}{W_h(s)}.$$

Then, we have

$$1 = \sup_\pi \sum_{s, W_h(s) > 0} w_h^\pi(s) \leq C_h = \sup_\pi \sum_{s, W_h(s) > 0} \frac{w_h^\pi(s)}{W_h(s)} \leq \sum_{s, W_h(s) > 0} \sup_\pi \frac{w_h^\pi(s)}{W_h(s)} \leq S.$$

We can see that $C_h = 1$ if $W_h(s) = 0$ or 1 for any state s i.e. the learner can reach $s_h = s$ with probability 1 by some policy for any reachable state s. On the other hand, $C_h = S$ if $w_h^\pi(s) = W_h(s) > 0$ for any state $s \in \mathcal{S}$, any policy π i.e. the learner cannot control the occupancy measure by varying policy and all states are reachable. Therefore, intuitively, a larger C_h indicates that the MDP is more difficult to control at step h.

3 The BREA algorithm

There are inherent difficulties in the fixed budget setting. First, while it is relatively straightforward to analyze algorithms in the fixed confidence setting using concentration bounds such as Hoeffding bound or Bernstein bound with a prespecified confidence level δ, it is much more challenging in the fixed budget setting, where neither the confidence level δ nor the accuracy level ε is known in advance. Second, whereas the fixed-confidence setting typically allows for a potentially excessive number of samples before termination (depending on the confidence level), the fixed-budget setting strictly limits the algorithm to a finite number of samples. Third, in the fixed-budget setting, the algorithm does not know in advance how to allocate the budget across different states s and time steps h. As a result, budget is often distributed uniformly, and the analysis must rely on nontrivial probabilistic arguments. In this section, we present how we design and analyze our algorithm to overcome the aforementioned difficulties.

At step h, each state s can be treated as a bandit problem, where the expected reward of each action a is given by $Q_h^*(s,a)$. If we aim to learn the exact optimal policy maximizing $Q_h^*(s,a)$, we need to sample trajectories $s_{h+1}, a_{h+1}, \dots, s_H, a_H$ generated under the optimal policy $\{\pi_{h'}^*\}_{h'=h+1}^H$, which is unknown. Fortunately, since our goal is to learn an approximately optimal policy, the following proposition shows that it suffices to use a suitably accurate policy $\{\hat\pi_{h'}\}_{h'=h+1}^H$ for sampling in order to learn $\hat\pi_h$.

Proposition 1. *[23, Lemma B.1] Assume that some deterministic policy $\hat\pi$ satisfies $\Delta_h^{\hat\pi}(s, \hat\pi_h(s)) \le \varepsilon_h(s)$ for any $h' \le h \le H$ and any $s \in \mathcal{S}$. Then, for any policy π',*

$$\sum_s w_{h'}^{\pi'}(s)\left(V_{h'}^*(s) - V_{h'}^{\hat\pi}(s)\right) \le \sum_{h=h'}^{H} \sup_\pi \sum_s w_h^\pi(s)\varepsilon_h(s).$$

Note that $\Delta_h^{\hat\pi}(s,a)$ depends only on the future policies $\{\hat\pi_{h'}\}_{h'=h+1}^H$, implying that we must determine them before learning $\hat\pi_h(s)$. By this observation, our learning proceeds backward from H to 1.

If we assume that the hypothesis of the previous proposition holds with $h' = 1$ and $\varepsilon_h(s) := \frac{\varepsilon}{C_h H W_h(s)}$, then the proposition says

$$\begin{aligned} V_0^* - V_0^{\hat\pi} &\le \sum_{h=1}^{H} \sup_\pi \sum_s w_h^\pi(s)\varepsilon_h(s) \\ &= \sum_{h=1}^{H} \sup_\pi \sum_s w_h^\pi(s)\frac{\varepsilon}{C_h H W_h(s)} \\ &= \sum_{h=1}^{H} \frac{\varepsilon}{H} \qquad \text{(definition of } C_h\text{)} \\ &= \varepsilon. \end{aligned}$$

Therefore, we design our algorithm to identify a $\Theta(\frac{\varepsilon}{C_h H W_h(s)})$-good action for each relevant state s. The precise definition of "relevant state" will be given in the analysis. We again emphasize that ε is not an input to our algorithm and can be chosen arbitrarily for the purpose of analysis. Our algorithm, which will be detailed in the following subsections, consists of two key components: estimating the reachability $W_h(s)$ and eliminating actions. We introduce the following notation, which will be used in the statements of upcoming results.

$$\varepsilon_B := (1 + \frac{\log(2)B}{c(B)})^{-0.6321}$$

denotes an error threshold that depends on the budget B. The factor

$$C_{\mathrm{L2E}}(B) = \tilde{O}(\mathrm{poly}(S, A, H)),$$

which arises from the first part of our algorithm, is formally defined in the Appendix C.

3.1 Reachability estimation

In the first part of the algorithm, we estimate the reachability $W_h(s)$ of each state s at step h. To this end, we execute reward-free exploration. One notable benefit of reward-free exploration is that it only needs to be run once, after which the collected data can be applied to a variety of downstream reward functions. More specifically, we reset the reward as $R_{h'}(s', a') = \begin{cases} 1, & \text{if } (s', a', h') = (s, 1, h), \\ 0, & \text{otherwise.} \end{cases}$, where we arbitrarily fix an action and denote it by 1. With this reset reward, an optimal policy maximizes the visitation probability of $(s, 1)$ at step h. Therefore, $V_0^* = W_h(s, 1) = W_h(s)$. To approximate such an optimal policy, we employ a regret minimization algorithm, StrongEuler [19].

More generally, the reachability $W_h(\mathcal{X})$ of any subset $\mathcal{X} \subset \mathcal{S} \times \mathcal{A}$ can be estimated in the same manner. We formalize this in Algorithm 1, which we refer to as **FB-L2E**, short for *Fixed-Budget Learn2Explore*. It is a fixed-budget variant of the Learn2Explore algorithm introduced in Wagenmaker

Algorithm 1 **F**ixed **B**udget **L**earn to **E**xplore (FB-L2E)

```
function FB-L2E($\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, step $h$, budget $B$)
    if $|\mathcal{X}| = 0$ then
        return $\{(\emptyset, \emptyset, 0)\}$
    end if
    $J \leftarrow \lceil 0.6321 \log_2(1 + \frac{\log(2)B}{c(B)}) \rceil$ ($c(B)$ is defined in Appendix C)
    for $j = 1$ to $J$ do
        $L_j \leftarrow 2^{J-j}, \quad \delta_j \leftarrow (\frac{1}{8SAH})^{0.6321 L_j \log\log(8SAH)}$
        $K_j \leftarrow K_j(\delta_j, SAH\delta_j)$ ($K_j$ is defined in Appendix C)
        $N_j \leftarrow K_j / (4|\mathcal{X}| \cdot 2^j)$
        $(\mathcal{X}_j, \Pi_j) \leftarrow$ FINDEXPLORABLESETS($\mathcal{X}, h, \delta, K_j, N_j$)
        $\mathcal{X} \leftarrow \mathcal{X} \setminus \mathcal{X}_j$
    end for
    return $\{(\mathcal{X}_j, \Pi_j, N_j)\}_{j=1}^{J}$
end function

function FINDEXPLORABLESETS($\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, step $h$, confidence $\delta$, epochs $K$, samples $N$)
    $r_h^1(s,a) \leftarrow 1$ if $(s,a) \in \mathcal{X}$, else 0
    $N(s,a,h) \leftarrow 0, \mathcal{Y} \leftarrow \emptyset, \Pi \leftarrow \emptyset, j \leftarrow 1$
    for $k = 1$ to $K$ do
        // StrongEuler is as defined in Simchowitz and Jamieson [19]
        Run STRONGEULER($\delta$) on reward $r_h^j$ to get trajectory $\{(s_h^k, a_h^k, h)\}_{h=1}^{H}$ and policy $\pi_k$
        $N(s_h^k, a_h^k) \leftarrow N(s_h^k, a_h^k) + 1, \quad \Pi \leftarrow \Pi \cup \{\pi_k\}$
        if $N(s_h^k, a_h^k) \geq N$, $(s_h^k, a_h^k) \in \mathcal{X}$ and $(s_h^k, a_h^k) \notin \mathcal{Y}$ then
            $\mathcal{Y} \leftarrow \mathcal{Y} \cup (s_h^k, a_h^k)$
            $r_h^{j+1}(s,a) \leftarrow 1$ if $(s,a) \in \mathcal{X} \setminus \mathcal{Y}$, else 0
            $j \leftarrow j + 1$
            Restart STRONGEULER($\delta$)
        end if
    end for
    return $\mathcal{Y}, \Pi$
end function
```

et al. [23], which itself is inspired by Zhang et al. [25], Brafman and Tennenholtz [3]. Algorithm 1 satisfies the following guarantee, the proof of which is deferred to the AppendixC.

Theorem 3.1. *Consider running Algorithm 1 with sufficiently large budget $B \geq c(B)$. Then, the following statements hold.*

1. *The total budget used is at most B.*

2. *For any $\varepsilon \geq 2SH^2\varepsilon_B$, with probability at least $1 - \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(B)}\right)\right)$,*

   (1) The reachability of each set $\mathcal{X}_i$ satisfies

   $$\frac{|\mathcal{X}_i|}{|\mathcal{X}|} \cdot 2^{-i-3} \leq W_h(\mathcal{X}_i) \leq 2^{-i+1} \quad \textit{for all } i \leq i_\varepsilon := \left\lceil \log_2\left(\frac{2SH^2}{\varepsilon}\right)\right\rceil,$$

   (2) The remaining elements, $\bar{\mathcal{X}} := \mathcal{X} \setminus \cup_{i=1}^{i_\varepsilon} \mathcal{X}_i$ satisfy

   $$\sup_\pi \sum_{(s,a)\in\bar{\mathcal{X}}} w_h^\pi(s,a) \leq \frac{\varepsilon}{2SH^2}.$$

   (3) Moreover, for any $i \leq i_\varepsilon$, if each policy in Π_i is executed A times, then every state-action pair $(s,a) \in \mathcal{X}_i$ is visited at least $\frac{1}{8}AN_i$ times.

Remark 2. *Theorem 3.1 crucially relies on the fact that* STRONGEULER *achieves a regret bound with $\log\frac{1}{\delta}$ dependence. However, when the target set is a singleton, i.e., $\mathcal{X} = (s,a)$, similar results can be obtained by combining alternative regret minimization algorithms—such as* EULER *[24], which exhibits a $\log^3\frac{1}{\delta}$ dependence—with a boosting technique. See the Appendix C for the details.*

3.2 Action elimination

In the second part of our algorithm, we iteratively sample trajectories, compute empirical Q-function of state-action pairs, and eliminate suboptimal actions. For this purpose, we employ a multiple bandit algorithm, *Successive Accepts and Rejects* (**SAR**), proposed by Bubeck et al. [5], and, for the first time, provide an ε-correctness guarantee for this algorithm. By employing this algorithm to our main algorithm, we are able to reduce the dependency on S compared to applying its multi-armed bandit counterpart.

Multiple bandit problem. Consider M instances of multi-armed bandit problems, each with K arms. Each arm i in instance m yields stochastic rewards supported on $[0, \sigma]$, with mean $\mu_{m,i}$, ordered such that $\mu_{m,1} \geq \cdots \geq \mu_{m,K}$. We denote each bandit-arm pair by (m, i), where $m \in [M]$ and $i \in [K]$. The objective is to identify a good arm in each instance $m \in [M]$ under a total budget of B pulls.

We now define some notations. Let $\widehat{\mu}_{m,i}(n)$ denote the empirical mean reward of arm i in instance m after n pulls. Define the suboptimality gap as

$$\bar{\Delta}_{m,i} := \begin{cases} \mu_{m,1} - \mu_{m,2}, & \text{if } i = 1, \\ \mu_{m,1} - \mu_{m,i}, & \text{if } i \in \{2, \ldots, K\}. \end{cases}$$

We enumerate all gaps $\bar{\Delta}_{m,i}$ over all $(m, i) \in [M] \times [K]$ in increasing order as

$$\bar{\Delta}_{(1)} \leq \bar{\Delta}_{(2)} \leq \cdots \leq \bar{\Delta}_{(MK)}.$$

Let $g(\varepsilon) := \left|\left\{(m, i) \in [M] \times [K] : \mu_{m,1} - \mu_{m,i} \leq \varepsilon\right\}\right|$ for any $\varepsilon > 0$, and define the harmonic log term

$$\overline{\log}(MK) := \frac{1}{2} + \sum_{i=2}^{MK} \frac{1}{i}.$$

For each $k \in [MK - 1]$, define

$$n_k(B, M, K) := \left\lceil \frac{1}{\overline{\log}(MK)} \cdot \frac{B - MK}{MK + 1 - k} \right\rceil. \tag{1}$$

The SAR algorithm is summarized in Algorithm 2. By leveraging the ranking of empirical gaps, SAR adaptively distributes the budget across bandit instances, solving the multiple bandit problem efficiently. We present a theoretical guarantee for its ability to identify ε-good arms.

Theorem 3.2. *If we run Algorithm 2 with budget $B \geq MK$, then the total number of budget used is at most B and the following holds for any $\varepsilon \geq 0$:*

$$\mathbb{P}(\exists m \in [M] : \mu_{m,1} - \mu_{m,J(m)} > \varepsilon) \leq 2M^2K^2 \exp\left(-\frac{B - MK}{128\sigma^2\, \overline{\log}(MK) \cdot \sum_{i \in [MK]} (\bar{\Delta}_{(i)} \vee \varepsilon)^{-2}}\right).$$

When $\varepsilon = 0$, Theorem 3.2 recovers the best arm identification result [5]. The proof of Theorem 3.2 is deferred to Appendix D.

3.3 Overview of the BREA algorithm

We combine the two mechanisms described above to construct our main algorithm. The algorithm proceeds in a backward manner over steps $h = H, H - 1, \ldots, 1$. At each step h, the first half of the budget is devoted to estimating the reachability $W_h(s)$ for each state s, while the second half applies the SAR mechanism to eliminate suboptimal actions. The whole algorithm is described in Algorithm 4 in the appendix.

In general MDPs, the stochasticity of the transition kernel prevents us from freely collecting arbitrary state-action samples. However, Theorem 3.1 ensures that, with high probability, the policies stored during the reachability estimation phase yield sufficient samples for each relevant state-action pair. Under this event, the SAR mechanism is expected to perform reliably. We now present our main theorem; its proof is provided in the Appendix E.

Algorithm 2 **S**uccessive **A**ccept and **R**eject (SAR) for the multiple bandit

```
input: Budget $B$
$A_1 \leftarrow \{(1,1),\ldots,(M,K)\}$, $n_0 \leftarrow 0$
for $k = 1$ to $MK-1$ do
    $n_k \leftarrow n_k(B,M,K)$ (as defined in (1))
    $\forall (m,i) \in A_k$,   pull $(m,i)$ for $n_k - n_{k-1}$ times
    $\forall m$,   $\hat{1}_m \leftarrow \arg\max_{i:(m,i)\in A_k} \hat{\mu}_{m,i}(n_k)$ (Break ties arbitrarily)
    if $\exists m$ such that $\hat{1}_m$ is the last active arm in $m$ then
        $J_m = \hat{1}_m$ (Accept)
        $A_{k+1} \leftarrow A_k \setminus \{(m,\hat{1}_m)\}$ (Deactivate)
    else
        $(m_k,i_k) \leftarrow \arg\max_{(m,i)\in A_k} \left(\hat{\mu}_{m,\hat{1}_m}(n_k) - \hat{\mu}_{m,i}(n_k)\right)$ (Break ties arbitrarily)
        $A_{k+1} \leftarrow A_k \setminus \{(m_k,i_k)\}$ (Reject and deactivate)
    end if
end for
$J_m \leftarrow i$ for $A_{MK} = \{(m,i)\}$
return $\{(m,J_m)\}_{m=1}^{M}$
```

Theorem 3.3. *If we run Algorithm 4 with a sufficiently large budget B, then the total number of budgets used is at most B. Moreover, for any $\varepsilon \geq 2SH^2\varepsilon_{\frac{B}{2SH}}$,*

$$\mathbb{P}\left(V_0^* - V_0^{\hat{\pi}} > \varepsilon\right) \leq \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(\frac{B}{2SH})}\right)\right) + \exp\left(-\tilde{\Theta}\left(\frac{B}{H^5 \max_{h\in[H]} C_h^2 \sum_{s\in\mathcal{S}} W_h(s)^{-1} \sum_{a\in\mathcal{A}}(\Delta_h(s,a) \vee \frac{\varepsilon}{W_h(s)})^{-2}}\right)\right).$$

Remark 3. *From Theorem 3.3, we can derive the sample complexity required by BREA to identify an ε-correct policy with probability at least $1-\delta$, given by*

$$\tau_{\varepsilon,\delta} = \tilde{\Theta}\left(\frac{C_{\mathrm{L2E}}\left(\frac{B}{2SH}\right)}{\varepsilon} + H^5 \max_{h\in[H]} C_h^2 \sum_{s\in\mathcal{S}} \frac{1}{W_h(s)} \sum_{a\in\mathcal{A}} \frac{1}{\left(\Delta_h(s,a) \vee \frac{\varepsilon}{W_h(s)}\right)^2}\right) \log\frac{1}{\delta}.$$

The first term inside $\tilde{\Theta}$ is a lower-order term. The second term inside $\tilde{\Theta}$ becomes $\sum_{a\in\mathcal{A}} \frac{1}{(\Delta(a)\vee\varepsilon)^2}$ for multi-armed bandits ($S = H = 1$). This is consistent with known results in the bandit literature ([9, 2, 14]). It is also noteworthy that our sample complexity is deterministic while the sample complexity of PAC RL algorithm typically is guaranteed with probability at least $1-\delta$.

Our sample complexity involves a $H^5 \max_h$ term, in contrast to the $H^4 \sum_h$ dependence that appears in PAC RL literature ([23, 22, 21]). This difference stems from the inherent difficulty of the fixed budget setting, where the algorithm does not know in advance how to distribute the budget across different h. A similar issue regarding the dependency on S could be resolved by employing a multiple bandit algorithm instead of a multi-armed bandit algorithm.

4 Conclusion

In this paper, we have explored the fixed-budget setting of the pure exploration MDP, which is surprisingly underexplored in RL theory. While our results establish the first fully instance-dependent guarantee in the fixed budget setting, these are just the beginning. First, it would be great to see what kind of instance-dependent acceleration can be proven in MDP, which should be possible given that accelerated rates were possible in bandits as a function of the number of good arms [15, 26]. Second, similarly, it would be interesting to explore what kind of data-poor regime guarantees are attainable – again, such bounds are available in the bandit setting [15, 26]. Third, we believe the factor H^2 in the sample complexity may be improved by leveraging variance-dependent concentration bounds. Finally, it would be interesting to extend our setting to the function approximation setting.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00219980).

References

[1] Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

[2] Audibert, J.-Y., Bubeck, S., and Munos, R. Best Arm Identification in Multi-Armed Bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2010.

[3] Brafman, R. I. and Tennenholtz, M. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3(null):213–231, Mar. 2003. ISSN 1532-4435. doi: 10.1162/153244303765208377. URL https://doi.org/10.1162/153244303765208377.

[4] Bubeck, S., Munos, R., and Stoltz, G. Pure Exploration in Multi-armed Bandits Problems. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, pages 23–37, 2009.

[5] Bubeck, S., Wang, T., and Viswanathan, N. Multiple Identifications in Multi-Armed Bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 258–265, 2013.

[6] Chaudhuri, A. R. and Kalyanakrishnan, S. PAC identification of a bandit arm relative to a reward quantile. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1777–1783, 2017.

[7] Dann, C., Marinov, T. V., Mohri, M., and Zimmert, J. Beyond value-function gaps: improved instance-dependent regret bounds for episodic reinforcement learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[8] Even-dar, E., Mannor, S., and Mansour, Y. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 255–270, 2002.

[9] Even-Dar, E., Mannor, S., and Mansour, Y. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

[10] Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 998–1027, 2016.

[11] Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

[12] Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jin20d.html.

[13] Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. PAC Subset Selection in Stochastic Multi-armed Bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 655–662, 2012.

[14] Karnin, Z., Koren, T., and Somekh, O. Almost Optimal Exploration in Multi-Armed Bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1238–1246, 2013.

[15] Katz-Samuels, J. and Jamieson, K. The True Sample Complexity of Identifying Good Arms. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1781–1791, 2020.

[16] Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

[17] Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, Dec. 2004. ISSN 1532-4435.

[18] Orabona, F. and Pal, D. Coin Betting and Parameter-Free Online Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 577–585, 2016.

[19] Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/10a5ab2db37feedfdeaab192ead4ac0e-Paper.pdf`.

[20] Tirinzoni, A., Al-Marjani, A., and Kaufmann, E. Near instance-optimal pac reinforcement learning for deterministic mdps. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[21] Tirinzoni, A., Al-Marjani, A., and Kaufmann, E. Optimistic pac reinforcement learning: the instance-dependent view. In Agrawal, S. and Orabona, F., editors, *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 1460–1480. PMLR, 20 Feb–23 Feb 2023. URL `https://proceedings.mlr.press/v201/tirinzoni23a.html`.

[22] Wagenmaker, A. and Jamieson, K. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[23] Wagenmaker, A. J., Simchowitz, M., and Jamieson, K. Beyond no regret: Instance-dependent pac reinforcement learning. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 358–418. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/wagenmaker22a.html`.

[24] Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/zanette19a.html`.

[25] Zhang, Z., Du, S., and Ji, X. Near optimal reward-free reinforcement learning. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12402–12412. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/zhang21e.html`.

[26] Zhao, Y., Stephens, C., Szepesvári, C., and Jun, K.-S. Revisiting simple regret: Fast rates for returning a good arm. *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Notation. For a positive integer n, we write $[n] := \{1, 2, \ldots, n\}$. We use $f = \tilde{\Theta}(g)$ to denote that the ratio $\frac{f}{g}$ is bounded both above and below by polylogarithmic functions. We define $\min^{+}_{x \in X} f(x) := \min_{x \in X : f(x) > 0} f(x)$. We use poly$(\cdot)$ to denote a polynomial in the variables inside the parentheses. We write $\log$ for natural logarithm and $\log_2$ for binary logarithm.

A Related work

Given the breadth of the literature on each topic, we focus on introducing only the most recent and relevant works.

Instance-dependent regret minimization in episodic MDPs. Zanette and Brunskill [24] proposed the EULER algorithm and proved a regret bound of $\sqrt{SAK \min\{\mathbb{Q}_\star H, \mathcal{G}^2\}}$, where $\mathbb{Q}_\star, \mathcal{G}$ are instance dependent term. Soon after, Simchowitz and Jamieson [19] proposed STRONGEULER algorithm and proved a gap-dependent regret bound for episodic tabular MDPs, showing that optimistic algorithms can achieve $O\big(\sum_{s,a,h} \frac{\log T}{\Delta_h(s,a)}\big)$ regret. This result, obtained via a novel "clipped" regret decomposition, smoothly interpolates between instance-dependent $O(\log T)$ growth and the worst-case $O(\sqrt{T})$ rate, without requiring simplifying assumptions like a bounded mixing time. Dann et al. [7] further refined these bounds by defining value-function gaps that ignore states never visited by an optimal policy. Finally, we note that any low-regret algorithm can be converted into a high-probability guarantee on near-optimal performance via an online-to-batch conversion. For detailed explanations, see Jin et al. [11]. However, recent studies ([23, 21]) suggest that algorithms for minimizing regret cannot be instance-optimal for identifying good policies, motivating specialized algorithms that explore more strategically than standard optimism.

Instance-dependent episodic PAC RL. The history of instance-dependent episodic PAC RL is not very long. Wagenmaker et al. [23] proposed a planning-based algorithm, MOCA, and analyzed its instance-dependent sample complexity. Tirinzoni et al. [20] provided an instance-dependent lower bound for deterministic MDPs and proposed the EPRL algorithm, which has an upper bound of sample complexity matching the lower bound up to a H^2 factor and logarithmic terms. Wagenmaker and Jamieson [22] considered finite horizon linear MDPs, a superset of tabular MDPs. They proposed the PEDEL algorithm, which takes a policy set as an input, and analyzed its sample complexity. Tirinzoni et al. [21] proved, for the first time, an instance-dependent sample complexity of an optimistic algorithm, BPI-UCRL.

Instance-dependent pure exploration in multi-armed bandits. The problem of pure exploration in multi-armed bandits (a special case of RL with $S = H = 1$) has a rich history and is typically studied in two frameworks: the fixed-confidence ((ε, δ)-PAC) setting and the fixed-budget setting.

In the fixed-confidence setting, the goal is to identify an arm whose mean reward is within ε of the optimal arm's mean with probability at least $1 - \delta$, while minimizing the number of samples (pulls). Even-dar et al. [8] initiated this line of work by proposing the Successive Elimination algorithm, which guarantees an optimal arm with probability $1 - \delta$ using distribution-dependent samples ((ε, δ)-sample complexity). Mannor and Tsitsiklis [17] later provided a distribution-dependent lower bound on the (ε, δ)-sample complexity. Kalyanakrishnan et al. [13] proposed the LUCB algorithm and analyzed their sample complexity. Karnin et al. [14] introduced the Exponential-Gap Elimination algorithm, removed unnecessary log factors and attained near-optimal sample complexity in the fixed-confidence regime. Garivier and Kaufmann [10] gave a tighter lower bound and proposed an algorithm, Track and Stop, which exactly hits the lower bound asymptotically.

In the fixed-budget setting, the learner is given a total sampling budget T and aims to maximize the probability of identifying the best arm by time T. Here, the results are often characterized by the exponential rate at which the failure probability decays with T. Audibert et al. [2] studied this setting and proposed the Successive Rejects algorithm, proving that its error probability decays at an optimal rate, up to logarithmic factors in the number of arms. Karnin et al. [14] proposed the Sequential Halving algorithm, proving that its error probability has an improved rate, which is optimal up to doubly logarithmic factors in the number of arms. Zhao et al. [26] provided a tighter analysis of the Sequential Halving algorithm and obtained an accelerated decay rate of ε-error probability.

B Properties of MDP

Although the statements and proofs of the lemmas in this section are nearly identical to those in the appendix of Wagenmaker et al. [23], we include them here for completeness.

Lemma 4. *Assume that some deterministic policy* $\hat{\pi}$ *satisfies* $\Delta_h^{\hat{\pi}}(s, \hat{\pi}_h(s)) \le \varepsilon_h(s)$ *for any* $h' \le h \le H$ *and any* $s \in \mathcal{S}$. *Then, for any policy* π',

$$\sum_s w_{h'}^{\pi'}(s) \left(V_{h'}^*(s) - V_{h'}^{\hat{\pi}}(s) \right) \le \sum_{h=h'}^{H} \sup_{\pi} \sum_s w_h^{\pi}(s) \varepsilon_h(s).$$

Proof. The proof proceeds by backward induction on h'. When $h' = H$, the statement trivially holds. Assume that

$$\sum_s w_{h'}^{\pi'}(s) \left(V_{h'}^*(s) - V_{h'}^{\hat{\pi}}(s) \right) \le \sum_{h=h'}^{H} \sup_{\pi} \sum_s w_h^{\pi}(s) \varepsilon_h(s)$$

holds for step $h' > 1$ and any policy π. Assume further that

$$\Delta_{h'-1}^{\hat{\pi}}(s, \hat{\pi}(s)) \le \varepsilon_{h'-1}(s).$$

By definition,

$$\begin{aligned} V_{h'-1}^*(s) - V_{h'-1}^{\hat{\pi}}(s) &= Q_{h'-1}^*(s, \pi_{h'-1}^*(s)) - Q_{h'-1}^{\hat{\pi}}(s, \hat{\pi}_{h'-1}(s)) \\ &= \underbrace{Q_{h'-1}^*(s, \pi_{h'-1}^*(s)) - Q_{h'-1}^{\hat{\pi}}(s, \pi_{h'-1}^*(s))}_{(1)} + \underbrace{Q_{h'-1}^{\hat{\pi}}(s, \pi_{h'-1}^*(s)) - \max_a Q_{h'-1}^{\hat{\pi}}(s, a)}_{(2)} \\ &\quad + \underbrace{\max_a Q_{h'-1}^{\hat{\pi}}(s, a) - Q_{h'-1}^{\hat{\pi}}(s, \hat{\pi}_{h'-1}(s)).}_{(3)} \end{aligned}$$

It is obvious that (2) ≤ 0 and (3) $= \Delta_{h'-1}^{\hat{\pi}}(s, \hat{\pi}_{h'-1}(s)) \le \varepsilon_{h'-1}(s)$ by our assumption. Furthermore,

$$(1) = \sum_{s'} P_{h'-1}(s'|s, \pi_{h'-1}^*(s))(V_{h'}^*(s') - V_{h'}^{\hat{\pi}}(s')).$$

Then, for any policy π',

$$\begin{aligned} \sum_s w_{h'-1}^{\pi'}(s)(V_{h'-1}^*(s) - V_{h'-1}^{\hat{\pi}}(s)) &\le \sum_s \sum_{s'} w_{h'-1}^{\pi'}(s) P_{h'-1}(s'|s, \pi_{h'-1}^*(s))(V_{h'}^*(s') - V_{h'}^{\hat{\pi}}(s')) \\ &\quad + \sum_s w_{h'-1}^{\pi'}(s) \varepsilon_{h'-1}(s) \\ &= \sum_s w_{h'}^{\pi''}(s)(V_{h'}^*(s) - V_{h'}^{\hat{\pi}}(s)) + \sum_s w_{h'-1}^{\pi''}(s) \varepsilon_{h'-1}(s) \\ &\le \sum_{h=h'-1}^{H} \sup_{\pi} \sum_s w_h^{\pi}(s) \varepsilon_h(s), \end{aligned}$$

where π'' is a policy that is equal to π' in step $1, \dots, h'-2$ and equal to π^* in step $h'-1, \dots, H$, the last inequality follows by the induction hypothesis. □

Lemma 5. *Assume* $\sup_{\pi} \sum_s w_{h+1}^{\pi}(s) \left(V_{h+1}^*(s) - V_{h+1}^{\hat{\pi}}(s) \right) \le \varepsilon$. *Then*

$$|\Delta_h(s, a) - \Delta_h^{\hat{\pi}}(s, a)| \le \varepsilon / W_h(s).$$

Proof.

$$\begin{aligned} |\Delta_h(s, a) - \Delta_h^{\hat{\pi}}(s, a)| &= |V_h^*(s) - Q_h^*(s, a) - (\max_{a'} Q_h^{\hat{\pi}}(s, a') - Q_h^{\hat{\pi}}(s, a))| \\ &\le \max\{|V_h^*(s) - \max_{a'} Q_h^{\hat{\pi}}(s, a')|, |Q_h^{\hat{\pi}}(s, a) - Q_h^*(s, a)|\}, \end{aligned}$$

where the last inequality follows since

$$V_h^*(s) - Q_h^*(s, a) - (\max_{a'} Q_h^{\hat{\pi}}(s, a') - Q_h^{\hat{\pi}}(s, a)) \le V_h^*(s) - \max_{a'} Q_h^{\hat{\pi}}(s, a')$$

and

$$-(V_h^*(s) - Q_h^*(s,a) - (\max_{a'} Q_h^{\hat{\pi}}(s,a') - Q_h^{\hat{\pi}}(s,a))) \leq Q_h^*(s,a) - Q_h^{\hat{\pi}}(s,a).$$

We can write

$$Q_h^*(s,a) = r_h(s,a) + \sum_{s'} P_h(s'|s,a) V_{h+1}^*(s'),$$
$$Q_h^{\hat{\pi}}(s,a) = r_h(s,a) + \sum_{s'} P_h(s'|s,a) V_{h+1}^{\hat{\pi}}(s').$$

Then we have

$$\begin{aligned} Q_h^*(s,a) - Q_h^{\hat{\pi}}(s,a) &= \sum_{s'} P_h(s'|s,a)(V_{h+1}^*(s') - V_{h+1}^{\hat{\pi}}(s')) \\ &= \frac{1}{W_h(s)} \sum_{s'} W_h(s) P_h(s'|s,a)(V_{h+1}^*(s') - V_{h+1}^{\hat{\pi}}(s')) \\ &\leq \frac{1}{W_h(s)} \sup_{\pi} \sum_{s'} w_{h+1}^{\pi}(s')(V_{h+1}^*(s') - V_{h+1}^{\hat{\pi}}(s')) \leq \frac{\varepsilon}{W_h(s)}. \end{aligned} \quad (2)$$

Let $a_1 := \arg\max_a Q_h^*(s,a)$. Then

$$\begin{aligned} V_h^*(s) - \max_{a'} Q_h^{\hat{\pi}}(s,a') &= \max_{a'} Q_h^*(s,a') - \max_{a'} Q_h^{\hat{\pi}}(s,a') = Q_h^*(s,a_1) - \max_{a'} Q_h^{\hat{\pi}}(s,a') \\ &= Q_h^*(s,a_1) - Q_h^{\hat{\pi}}(s,a_1) + Q_h^{\hat{\pi}}(s,a_1) - \max_{a'} Q_h^{\hat{\pi}}(s,a') \leq \frac{\varepsilon}{W_h(s)}. \end{aligned} \quad (3)$$

By (2), (3), the lemma follows. □

C Analysis of FB-L2E

C.1 Analysis of FINDEXPLORABLESETS

The overall analysis is similar to that of Wagenmaker et al. [23]. However, the details should be changed as we use STRONGEULER instead of EULER. We begin with a regret bound of STRONGEULER. Throughout this section, let $M := (SAH^2)^2$.

Lemma 6. *If we run* STRONGEULER *with confidence parameter δ for K episodes, with probability at least $1-\delta$,*

$$\sum_{k=1}^{K} V_0^* - \sum_{k=1}^{K} V_0^{\pi_k} \leq c_{\mathrm{se}} \sqrt{SAH^2 V_0^* K \log(HK) \log(\frac{MHK}{\delta})} + c_{\mathrm{se}} S^2 A H^6 \log(HK) \log(\frac{MHK}{\delta}),$$

where $M = (SAH^2)^2$ and c_{se} is a universal constant.

Proof. In Simchowitz and Jamieson [19, Theorem 2.4], the regret bound up to a universal constant is presented as

$$\sqrt{SA\bar{H}_T T \log(\frac{mT}{\delta})} + SAH^4(S \vee H)\log(\frac{mT}{\delta}) \min\{\log(\frac{mT}{\delta}), \log(\frac{mH}{\Delta_{\min}})\},$$

where $\Delta_{\min} = \min^+_{s,a,h} \Delta_h(s,a)$, $T = HK$, $m = (SAH)^2$, and $\bar{H}_T \leq \frac{\mathcal{G}^2}{H}\log(T)$. Here, $\mathcal{G}$ is a constant such that the reward of one episode of our MDP is bounded by $\mathcal{G}$. We can reduce this $\frac{\mathcal{G}^2}{H}$ term to $\frac{V_0^*}{4H}$ by using the argument used in the proof of Jin et al. [12, Lemma 3.4] and Wagenmaker et al. [23, Lemma D.4]. Thus, the regret bound (up to a universal constant) of STRONGEULER is given as

$$\sqrt{SAV_0^* T \log(T) \log(\frac{mT}{\delta})} + SAH^4(S \vee H)\log(\frac{mT}{\delta}) \min\{\log(\frac{mT}{\delta}), \log(\frac{mH}{\Delta_{\min}})\},$$

The second term is derived from their Simchowitz and Jamieson [19, Claim C.3]. In the proof of Simchowitz and Jamieson [19, Claim C.3], we can just bound

$$\log(1 + \frac{N \wedge n_{\mathrm{end}}}{n_0}) \leq \log(1+T)$$

since $N \leq T, n_0 \geq 1$. By using this bound, we get a regret bound of

$$\sqrt{SAV_0^* T \log(T) \log(\frac{mT}{\delta})} + SAH^4(S \vee H) \log(\frac{mT}{\delta}) \log(T).$$

Although this bound only applies to stationary MDPs, stationary MDPs can represent non-stationary MDPs by augmenting states s to (s, h). In this case, the effective number of states is SH. Thus, by substituting SH in to S, HK into T, the lemma follows. □

We now define the important quantities

$$\begin{aligned}
C_K(\delta, \delta_{\mathrm{samp}}, i) := \max \Bigg\{ & 432c_{\mathrm{se}}^2 S^3 A^2 H^6 (i+6)^2 \log^2(2 \cdot 2 \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^7 M(i+6), \\
& 432c_{\mathrm{se}}^2 S^3 A^2 H^6 \log(\frac{1}{\delta})(i+3) \log(2 \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^7 \log(\frac{1}{\delta})(i+3)), \\
& 24 \log(\frac{4}{\delta}), \quad 2^{11} S^2 A^2 \log(\frac{4SAH}{\delta_{\mathrm{samp}}}) \Bigg\}, \\
K_i(\delta, \delta_{\mathrm{samp}}) := \lceil 2^i & C_K(\delta, \delta_{\mathrm{samp}}, i) \rceil.
\end{aligned} \tag{4}$$

and prove the following property.

Lemma 7. *Let* $C_{\mathcal{R}} := 2c_{\mathrm{se}} S^3 A^2 H^6 \log(HK_i) \log(\frac{2MHK_i}{\delta}) + 2 \log \frac{4}{\delta}$ *and* $K_i = K_i(\delta, \delta_{\mathrm{samp}})$. *Then,*

$$K_i \geq 2^i \max\{4C_{\mathcal{R}}, 144c_{\mathrm{se}}^2 S^2 A^2 H^2 \log(HK_i) \log(\frac{2MHK_i}{\delta})\}.$$

Proof. For any $i, j > 0$ and $C > 0$, if $x \geq C^i (i+3j)^j \log^j(C(i+3j))$, then $x \geq C^i \log^j x$ since

$$\begin{aligned}
C^i \log^j x = C^i \log^j [C^i (i+3j)^j \log^j (C(i+3j))] &\leq C^i \log^j [C^{i+j} (i+3j)^{2j}] \\
&\leq C^i (i+3j)^j \log^j [C(i+3j)] \\
&= x
\end{aligned}$$

Since

$$2MHK_i \geq 2^i \cdot 2 \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^7 M(i+6)^2 \log^2(2 \cdot 2 \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^7 M(i+6),$$

we have

$$K_i \geq 2^i \cdot 2 \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^6 \log^2(2MHK_i).$$

Since

$$HK_i \geq 2^i \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^7 \log(\frac{1}{\delta})(i+3) \log(2 \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^7 \log(\frac{1}{\delta})(i+3)),$$

we have

$$K_i \geq 2^i \cdot 432c_{\mathrm{se}}^2 S^3 A^2 H^6 \log(HK_i) \log(\frac{1}{\delta}).$$

We also have $K_i \geq 2^i \cdot 24 \log(\frac{4}{\delta})$. Combining these three, we have

$$K_i \geq 2^i \left(144c_{\mathrm{se}}^2 S^3 A^2 H^6 (\log^2(2MHK_i) + \log(HK_i) \log(\frac{1}{\delta}) + 8 \log(\frac{4}{\delta}) \right),$$

which easily implies

$$\begin{aligned}
K_i &\geq 2^i \cdot 144c_{\mathrm{se}}^2 S^2 A^2 H^2 \log(HK_i) \log(\frac{2MHK_i}{\delta}), \\
K_i &\geq 8^i \left(2c_{\mathrm{se}} S^3 A^2 H^6 \log(HK_i) \log(\frac{2MHK_i}{\delta}) + 8 \log(\frac{4}{\delta}) \right) = 4C_{\mathcal{R}}.
\end{aligned}$$

□

Throughout the rest of this subsection, we consider running

$$\textsc{FindExplorableSets}(\mathcal{X}, h, \delta, K_i := K_i(\delta, \delta_{\mathrm{samp}}), N_i := \frac{K_i}{4|\mathcal{X}|2^i})$$

(defined in Algorithm 1) with some $\mathcal{X} \subset \mathcal{S} \times \mathcal{A}$ satisfying

$$W_h(\mathcal{X}) \leq 2^{-i+1}.$$

Let $\mathcal{X}_i \subset \mathcal{X}$, Π_i be the output. We introduce the following notations. Let K_{ij} denote the total number of episodes taken for j, where the index j changes when the reward r_h^j is reset. Let m_i denote the number of j. Thus, we have

$$\sum_{j=1}^{m_i} K_{ij} = K_i.$$

Let $V_0^{*,ij}$ denote the optimal value function on the reward function r_h^j, $V_0^{k,ij}$ denote the value function for the policy π_k on the reward function r_h^j. Then,

$$V_0^{k,ij} \leq V_0^{*,ij} \leq \sup_\pi \mathbb{E}_\pi[\mathbb{I}\{(s_h, a_h) \in \mathcal{X}\}] = W_h(\mathcal{X}) \leq 2^{-(i-1)}.$$

Now we define some events.

$$\begin{aligned}
\mathcal{C}_{1,\delta} = \Big\{ & \sum_{j=1}^{m_i} \Big(\sum_{k=1}^{K_{ij}} V_0^{*,ij} - \sum_{k=1}^{K_{ij}} V_0^{k,ij} \Big) \leq 2c_{\text{se}} \sqrt{S^2 A^2 H^2 V_0^{*,i1} K_i \log(HK_i) \log(\frac{MHK_i}{\delta})} \\
& + 2c_{\text{se}} S^3 A^2 H^6 \log(HK_i) \log(\frac{MHK_i}{\delta}) \Big\}, \\
\mathcal{C}_{2,\delta} = \Big\{ & \left| \sum_{j=1}^{m_i} \sum_{k=1}^{K_{ij}} \sum_{h=1}^{H} R_h^j(s_h^{j,k}, a_h^{j,k}) - \sum_{j=1}^{m_i} \sum_{k=1}^{K_{ij}} V_0^{k,ij} \right| \leq \sqrt{4K_i 2^{-i} \log \frac{2}{\delta}} + 2 \log \frac{2}{\delta} \Big\}, \\
\mathcal{D}_{1,\delta} = \Big\{ & \forall (s,a) \in \mathcal{X}, \left| \sum_{k=1}^{K_i} w_h^{\pi_k}(s,a) - \sum_{k=1}^{K_i} \mathbb{I}_{\{(s_h^k, a_h^k) = (s,a)\}} \right| \leq \sqrt{2K_i W_h(s) \log \frac{2}{\delta}} + 2 \log \frac{2}{\delta} \Big\} \\
& \text{for the process during the algorithm,} \\
\mathcal{D}_{2,\delta} = \Big\{ & \forall (s,a) \in \mathcal{X}_i, \left| \sum_{k=1}^{K_i} w_h^{\pi_k}(s,a) - \sum_{k=1}^{K_i} \mathbb{I}_{\{(s_h^k, a_h^k) = (s,a)\}} \right| \leq \sqrt{2K_i W_h(s) \log \frac{2}{\delta}} + 2 \log \frac{2}{\delta} \Big\} \\
& \text{for the process during the replay.}
\end{aligned}$$

Freedman's inequality is stated below for use in subsequent analysis.

Lemma 8 (Freedman's inequality). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \mathcal{F}$ be a filtration of σ-algebra. Let $\{X_i\}_i$ be random variables such that X_i is $\mathcal{F}_i$-measurable,*

$$\begin{aligned}
|X_i| &\leq M, \\
\mathbb{E}[X_n | \mathcal{F}_{n-1}] &= 0, \\
\mathbb{E}[X_n^2 | \mathcal{F}_{n-1}] &\leq V_n
\end{aligned}$$

for constants V_n. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$|\sum_{i=1}^n X_i| < 2M \log \frac{2}{\delta} + \sqrt{2 \sum_{i=1}^n V_n \log \frac{2}{\delta}}.$$

We state properties of the events defined above.

Lemma 9. *If $\delta \in (0,1)$ is the third argument of FindExplorableSets,*

$$\mathbb{P}(\mathcal{C}_{1,\delta/2}) \geq 1 - \delta/2.$$

Proof. For any fixed K and j,

$$\begin{aligned}
\Big(\sum_{k=1}^{K} V_0^{*,ij} - \sum_{k=1}^{K} V_0^{k,ij} \Big) | \mathcal{F}_{j-1} \leq{} & c_{\text{se}} \sqrt{SAH^2 V_0^{*,i1} K \log(HK) \log(\frac{MHK}{\delta})} \\
& + c_{\text{se}} S^2 A H^6 \log(HK) \log(\frac{MHK}{\delta})
\end{aligned}$$

with probability at least $1-\delta$, where $\mathcal{F}_{j-1}$ is the filtration up to iteration j, and we used $V_0^{*,ij} \leq V_0^{*,i1}$ for all j since the reward function can only decrease as j increases. FindExplorableSets stops and restarts STRONGEULER if the relevant condition is met, but this is a random stopping condition. Thus, to guarantee that the regret bound holds for any possible value of this stopping time, we union bound over all possible values. Since `FindExplorableSets` runs for at most K_i episodes, we union bound over K_i stopping times. We then have

$$\Big(\sum_{k=1}^{K} V_0^{*,ij} - \sum_{k=1}^{K} V_0^{k,ij}\Big)|\mathcal{F}_{j-1} \le 2c_{\mathrm{se}}\sqrt{SAH^2V_0^{*,i1}K\log(HK_i)\log(\frac{2MHK_i}{\delta})}$$
$$+ 2c_{\mathrm{se}}S^2AH^6\log(HK_i)\log(\frac{2MHK_i}{\delta})$$

for all $K \in [K_i]$ with probability at least $1-\frac{\delta}{2SA}$. Since $m_i \le SA$, union bounding over all j we then have that, with probability at least $1-\delta/2$,

$$\begin{aligned}\sum_{j=1}^{m_i}\Big(\sum_{k=1}^{K_{ij}} V_0^{\star,ij} - \sum_{k=1}^{K_{ij}} V_0^{k,ij}\Big) &\le \sum_{j=1}^{m_i} 2c_{\mathrm{se}}\sqrt{SAH^2V_0^{*,i1}K_{ij}\log(HK_i)\log(\frac{2MHK_i}{\delta})} \\ &\quad + 2c_{\mathrm{se}}S^3A^2H^6\log(HK_i)\log(\frac{2MHK_i}{\delta}) \\ &\le 2c_{\mathrm{se}}\sqrt{S^2A^2H^2V_0^{*,i1}K_i\log(HK_i)\log(\frac{2MHK_i}{\delta})} \\ &\quad + 2c_{\mathrm{se}}S^3A^2H^6\log(HK_i)\log(\frac{2MHK_i}{\delta}),\end{aligned}$$

where the last inequality follows from Jensen's inequality. □

Lemma 10. *For any $\delta \in (0,1)$,*

$$\mathbb{P}(\mathcal{C}_{2,\delta}) \ge 1-\delta.$$

Proof. For each $k \in [K_i]$, we have that $X_k := \sum_{h=1}^{H} R_h(s_h^k, a_h^k) \sim \mathrm{Bernoulli}(V_0^{\pi_k})$. Then $|X_k - V_0^{\pi_k}| \le 1$, $\mathbb{E}[(X_k - V_0^{\pi_k})^2|\mathcal{F}_{k-1}] = V_0^{\pi_k}(1-V_0^{\pi_k}) \le V_0^{\pi_k} \le W_h(\mathcal{X}) \le 2^{-i+1}$. Thus, if we apply Lemma 8, we obtain the statement. □

Lemma 11. *For any $\delta \in (0,1)$,*

$$\mathbb{P}(\mathcal{D}_{1,\delta}) \ge 1 - |\mathcal{X}|\delta \ge 1 - SA\delta.$$

Proof. Since $X_k := \mathbb{I}_{\{(s_h^k,a_h^k)=(s,a)\}} \sim \mathrm{Bernoulli}(w_h^{\pi_k}(s,a))$,

$$\mathbb{E}[(X_k - w_h^{\pi_k}(s,a))^2|\mathcal{F}_{k-1}] = w_h^{\pi_k}(s,a)(1-w_h^{\pi_k}(s,a)) \le w_h^{\pi_k}(s,a) \le W_h(s).$$

By Lemma 8, we have that

$$\left|\sum_{k=1}^{K_i} w_h^{\pi_k}(s,a) - \sum_{k=1}^{K_i}\mathbb{I}_{\{(s_h^k,a_h^k)=(s,a)\}}\right| \le \sqrt{2K_iW_h(s)\log\frac{2}{\delta}} + 2\log\frac{2}{\delta}$$

with probability at least $1-\delta$. Union bounding over $\mathcal{X}$ leads to the statement. □

Lemma 12. *For any $\delta \in (0,1)$,*

$$\mathbb{P}(\mathcal{D}_{2,\delta}) \ge 1 - |\mathcal{X}_i|\delta \ge 1 - SA\delta.$$

Proof. Since $X_k := \mathbb{I}_{\{(s_h^k,a_h^k)=(s,a)\}} \sim \mathrm{Bernoulli}(w_h^{\pi_k}(s,a))$,

$$\mathbb{E}[(X_k - w_h^{\pi_k}(s,a))^2|\mathcal{F}_{k-1}] = w_h^{\pi_k}(s,a)(1-w_h^{\pi_k}(s,a)) \le w_h^{\pi_k}(s,a) \le W_h(s).$$

By Lemma 8 and union bound over $\mathcal{X}_i$, the statement follows. □

Lemma 13. *If $\delta \in (0,1)$ is the third argument of `FindExplorableSets`, the event $\mathcal{C}_{1,\delta/2} \cap \mathcal{C}_{2,\delta/2}$ implies*

$$W_h(\mathcal{X}\setminus\mathcal{X}_i) \le 2^{-i}.$$

Proof. Putting Lemma 9, 10 and union bounding over these events, we have that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}}\sum_{h=1}^{H} R_h^j(s_h^{j,k},a_h^{j,k}) &\geq \sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}} V_0^{\star,ij} - \sqrt{4K_i 2^{-i}\log\frac{4}{\delta}} \\ &\quad - 2c_{\mathrm{se}}\sqrt{S^2A^2H^2V_0^{*,i1}K_i\log(HK_i)\log(\frac{2MHK_i}{\delta})} - C_{\mathcal{R}}\end{aligned}$$

where we denote

$$C_{\mathcal{R}} := 2c_{\mathrm{se}}S^3A^2H^6\log(HK_i)\log(\frac{2MHK_i}{\delta}) + 2\log\frac{4}{\delta}.$$

Assume that $V_0^{*,im_i} > 2^{-i}$. Using that the reward decreases monotonically so $V_0^{*,im_i} \leq V_0^{*,ij}$ for any $j \leq m_i$, we can lower bound the above as

$$\begin{aligned}&\geq 2^{-i}K_i - \sqrt{4K_i2^{-i}\log\frac{4}{\delta}} - 2c_{\mathrm{se}}\sqrt{S^2A^2H^2V_0^{*,i1}K_i\log(HK_i)\log(\frac{2MHK_i}{\delta})} - C_{\mathcal{R}} \\ &\geq 2^{-i}K_i - 3c_{\mathrm{se}}\sqrt{S^2A^2H^22^{-i}K_i\log(HK_i)\log(\frac{2MHK_i}{\delta})} - C_{\mathcal{R}}\end{aligned}$$

where the second inequality follows since $V_0^{*,i1} \leq 2^{-i+1}$ and $\sqrt{4K_i2^{-i}\log\frac{4}{\delta}}$ will then be dominated by the regret term. Lemma 7 gives

$$K_i \geq 2^i \max\left\{4C_{\mathcal{R}}, 144c_{\mathrm{se}}^2S^2A^2H^2\log(HK_i)\log(\frac{2MHK_i}{\delta})\right\}$$

which implies

$$\frac{1}{4}2^{-i}K_i - C_{\mathcal{R}} \geq 0$$

and

$$\begin{aligned}&\frac{1}{4}2^{-i}K_i - 3c_{\mathrm{se}}\sqrt{S^2A^2H^22^{-i}K_i\log(HK_i)\log(\frac{2MHK_i}{\delta})} \\ &\quad\geq \frac{2^i\cdot 144c_{\mathrm{se}}^2S^2A^2H^2\log(HK_i)\log(\frac{2MHK_i}{\delta})}{4\cdot 2^i} \\ &\quad - 3c_{\mathrm{se}}\sqrt{S^2A^2H^22^{-i}\log(HK_i)\log(\frac{2MHK_i}{\delta})\cdot 2^i144c_{\mathrm{se}}^2S^2A^2H^2\log(HK_i)\log(\frac{2MHK_i}{\delta})} \\ &\quad = 0.\end{aligned}$$

Thus, we can lower bound the above as

$$2^{-i}K_i - 3c_{\mathrm{se}}\sqrt{S^2A^2H^22^{-i}K_i\log(HK_i)\log(\frac{2MHK_i}{\delta})} - C_{\mathcal{R}} \geq \frac{1}{2}2^{-i}K_i.$$

Note that we can collect a total reward of at most $|\mathcal{X}|N_i$. However, by our choice of

$$N_i = K_i/(4|\mathcal{X}|\cdot 2^i),$$

we have that

$$|\mathcal{X}|N_i = \frac{1}{4\cdot 2^i}K_i < \frac{1}{2\cdot 2^i}K_i.$$

This is a contradiction. Thus, we must have that $W_h(\mathcal{X}\setminus\mathcal{X}_i) \leq V_0^{*,im_i} \leq 2^{-i}$. □

Lemma 14. *The event $\mathcal{C}_{\delta/2}$ with $\delta \geq \frac{\delta_{\mathrm{samp}}}{SAH}$ implies*

$$W_h(\mathcal{X}) \geq \frac{|\mathcal{X}_i|}{2^{i+3}|\mathcal{X}|}.$$

Proof.

$$N_i|\mathcal{X}_i| \leq \sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}} R_h^j(s_h^{j,k},a_h^{j,k}) \leq \sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}} V_0^{k,ij} + \sqrt{4K_i2^{-i}\log\frac{4}{\delta}} + 2\log\frac{4}{\delta}$$

$$
\begin{aligned}
&\leq K_i W_h(\mathcal{X}) + \sqrt{4K_i 2^{-i} \log \frac{4}{\delta}} + 2\log\frac{4}{\delta} \\
&\leq K_i W_h(\mathcal{X}) + \frac{K_i}{2^{i+4}SA} + \frac{K_i}{2^{i+10}SA} \\
&\leq K_i W_h(\mathcal{X}) + \frac{K_i}{2^{i+3}SA},
\end{aligned}
$$

where the forth inequality follows from $K_i \geq 2^{i+11} S^2 A^2 \log \frac{4SAH}{\delta_{\mathrm{samp}}}$. Then,

$$
W_h(\mathcal{X}) \geq \frac{N_i|\mathcal{X}_i|}{K_i} - \frac{1}{2^{i+3}SA} = \frac{|\mathcal{X}_i|}{2^{i+2}|\mathcal{X}|} - \frac{1}{2^{i+3}SA} \geq \frac{|\mathcal{X}_i|}{2^{i+3}|\mathcal{X}|}.
$$

□

Lemma 15. *The event $\mathcal{D}_{1,\delta} \cap \mathcal{D}_{2,\delta}$ with $\delta \geq \frac{\delta_{\mathrm{samp}}}{2SAH}$ implies that after rerunning each policy in Π_i once, the number of samples collected for each $(s,a) \in \mathcal{X}_i$ is at least $\frac{1}{4}N_i$.*

Proof. Let $\mathbb{I}^1, \mathbb{I}^2$ denote the indicator of an event during `FindExplorableSets`, and an event during rerunning policies respectively. For a pair $(s,a) \in \mathcal{X}_i$, we have

$$
\sum_{k=1}^{K_i} \mathbb{I}^1_{\{(s_h^k, a_h^k)=(s,a)\}} - \sum_{k=1}^{K_i} w_h^{\pi_k}(s,a) \leq \sqrt{2K_i W_h(s) \log\frac{2}{\delta}} + 2\log\frac{2}{\delta}
$$

$$
\sum_{k=1}^{K_i} w_h^{\pi_k}(s,a) - \sum_{k=1}^{K_i} \mathbb{I}^2_{\{(s_h^k, a_h^k)=(s,a)\}} \leq \sqrt{2K_i W_h(s) \log\frac{2}{\delta}} + 2\log\frac{2}{\delta}
$$

Then the number of samples of (s,a) collected during the rerunning satisfies

$$
\begin{aligned}
\sum_{k=1}^{K_i} \mathbb{I}^2_{\{(s_h^k, a_h^k)=(s,a)\}} &\geq \sum_{k=1}^{K_i} \mathbb{I}^1_{\{(s_h^k, a_h^k)=(s,a)\}} - 2\sqrt{2K_i W_h(s) \log\frac{2}{\delta}} - 4\log\frac{2}{\delta} \\
&\geq N_i - 2\sqrt{2K_i W_h(s) \log\frac{2}{\delta}} - 4\log\frac{2}{\delta} \\
&\geq N_i - 2\sqrt{2^{-i+2} K_i \log\frac{2}{\delta}} - 4\log\frac{2}{\delta} \\
&\geq N_i - \frac{K_i}{2^{i+3.5}SA} - \frac{K_i}{2^{i+9}S^2A^2} \\
&\geq N_i - \frac{K_i}{2^{i+2.5}SA} \\
&\geq N_i - \frac{K_i}{2^{i+2.5}|\mathcal{X}|} = N_i(1 - \frac{1}{\sqrt{2}}) \geq \frac{1}{4}N_i,
\end{aligned}
$$

where the forth inequality follows from $\delta \geq \frac{\delta_{\mathrm{samp}}}{4SAH}$ and $\log\frac{2SAH}{\delta_{\mathrm{samp}}} \leq \frac{K_i}{2^{i+11}S^2A^2}$. □

Lemma 16. *The event $\mathcal{D}_{1,\delta}$ with $\delta \geq \frac{\delta_{\mathrm{samp}}}{2SAH}$ implies*

$$
W_h(s) > \frac{1}{2^{i+3}|\mathcal{X}|} \textit{ for each } (s,a) \in \mathcal{X}_i, \quad W_h(\mathcal{X}_i) > \frac{|\mathcal{X}_i|}{2^{i+3}|\mathcal{X}|}.
$$

Proof. In the proof of the previous lemma, we showed that

$$
\sqrt{2^{-i+2} K_i \log\frac{2}{\delta}} + 2\log\frac{2}{\delta} \leq \frac{N_i}{2\sqrt{2}} < \frac{N_i}{2}
$$

when $\delta \geq \frac{\delta_{\mathrm{samp}}}{2SAH}$. Using this, we have

$$
N_i \leq \sum_{k=1}^{K_i} \mathbb{I}^1_{\{(s_h^k, a_h^k)=(s,a)\}} \leq \sum_{k=1}^{K_i} w_h^{\pi_k}(s,a) + \sqrt{2^{-i+2} K_i \log\frac{2}{\delta}} + 2\log\frac{2}{\delta} < K_i W_h(s) + \frac{N_i}{2}
$$

for each $(s,a) \in \mathcal{X}_i$. Thus,

$$W_h(s) > \frac{N_i}{2K_i} = \frac{1}{2^{i+3}|\mathcal{X}|}.$$

On the other hand,

$$|\mathcal{X}_i|N_i \le \sum_{(s,a)\in\mathcal{X}_i} \sum_{k=1}^{K_i} \mathbb{I}^1_{\{(s_h^k,a_h^k)=(s,a)\}} \le \sum_{k=1}^{K_i} w_h^{\pi_k}(\mathcal{X}_i) + |\mathcal{X}_i| \left(\sqrt{2^{-i+2}K_i \log\frac{2}{\delta}} + 2\log\frac{2}{\delta}\right) < K_i W_h(\mathcal{X}_i) + \frac{|\mathcal{X}_i|N_i}{2}.$$

Thus,

$$W_h(\mathcal{X}_i) > \frac{|\mathcal{X}_i|N_i}{2K_i} = \frac{|\mathcal{X}_i|}{2^{i+3}|\mathcal{X}|}.$$

□

We finally give a guarantee of `FindExplorableSets`.

Theorem C.1. *If we run*

$$\textit{FindExplorableSets}(\mathcal{X}, h, \delta, K_i = K_i(\delta, \delta_{\text{samp}} = SAH\delta), N_i = \frac{K_i}{4|\mathcal{X}|2^i})$$

for a subset $\mathcal{X} \subset \mathcal{S} \times \mathcal{A}$ with $W_h(\mathcal{X}) \le 2^{-i+1}$ and returns subset $\mathcal{X}_i \subset \mathcal{X}$, policy set Π_i, then

1. *$W_h(\mathcal{X} \setminus \mathcal{X}_i) \le 2^{-i}$ with probability at least $1-\delta$.*
2. *With probability at least $1 - SA\delta$,*
   (1) *If we rerun each policy in Π_i once, the number of samples collected for each $(s,a) \in \mathcal{X}_i$ is at least $\frac{1}{4}N_i$.*
   (2) *$W_h(s) > \frac{1}{2^{i+3}|\mathcal{X}|}$ for each $(s,a) \in \mathcal{X}_i$ and $W_h(\mathcal{X}_i) > \frac{|\mathcal{X}_i|}{2^{i+3}|\mathcal{X}|}$.*

Proof. By Lemma 9, 10, 11, 12, 13, 15, and 16, the theorem follows. □

C.2 Proof of Theorem 3.1

Before proving Theorem 3.1, we introduce a useful lemma related to the Lambert W-function. The Lambert function $W(s) : [0,\infty) \to [0,\infty)$ is defined by

$$x = W(x)\exp(W(x)), \quad \text{for } x \ge 0.$$

Then the following holds.

Lemma 17. *[18, Lemma 17]*

$$0.6321 \log(1+x) \le W(x) \le \log(1+x) \textit{ for } x \ge 0.$$

We define

$$c(B) = 4JC_K(\frac{1}{8SAH}, \frac{1}{8}, J) = \text{poly}(S, A, H, \log(B)),$$
$$C_{\text{L2E}}(B) = SH^2 c(B). \tag{5}$$

Recall that C_K was defined in (4). We now give a proof of Theorem 3.1

Theorem C.2 (Theorem 3.1)**.** *Consider running Algorithm 1 with sufficiently large budget $B \ge c(B)$. Then, the following statements hold.*

1. *The total budget used is at most B.*
2. *For any $\varepsilon \ge 2SH^2\varepsilon_B$, with probability at least $1 - \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\text{L2E}}(B)}\right)\right)$,*
   (1) *The reachability of each set $\mathcal{X}_i$ satisfies*
   $$\frac{|\mathcal{X}_i|}{|\mathcal{X}|} \cdot 2^{-i-3} \le W_h(\mathcal{X}_i) \le 2^{-i+1} \quad \textit{for all } i \le i_\varepsilon := \left\lceil \log_2\left(\frac{2SH^2}{\varepsilon}\right)\right\rceil,$$
   (2) *The remaining elements, $\bar{\mathcal{X}} := \mathcal{X} \setminus \cup_{i=1}^{i_\varepsilon} \mathcal{X}_i$ satisfy*
   $$\sup_\pi \sum_{(s,a)\in\bar{\mathcal{X}}} w_h^\pi(s,a) \le \frac{\varepsilon}{2SH^2}.$$

(3) *Moreover, for any $i \leq i_\varepsilon$, if each policy in Π_i is executed A times, then every state-action pair $(s, a) \in \mathcal{X}_i$ is visited at least $\frac{1}{8}AN_i$ times.*

Proof. We first prove that the total budget used is at most B. Let $\delta = \frac{1}{8SAH}$. By the definition of δ_i,

$$\begin{aligned}\log\frac{1}{\delta_i} &= 0.6321 L_i \log\frac{1}{\delta}\cdot\log\log\frac{1}{\delta}\\ &\leq 1 + 0.6321 L_i \log\frac{1}{\delta}\cdot\log\log\frac{1}{\delta}\\ &\leq (1 + L_i\log\frac{1}{\delta}\cdot\log\log\frac{1}{\delta})^{0.6321}\\ &\leq \exp(W(L_i\log\frac{1}{\delta}\cdot\log\log\frac{1}{\delta})).\end{aligned}$$

Thus,

$$\log\frac{1}{\delta_i}\cdot\log\log\frac{1}{\delta_i} \leq W(L_i\log\frac{1}{\delta}\cdot\log\log\frac{1}{\delta})\exp(W(L_i\log\frac{1}{\delta}\cdot\log\log\frac{1}{\delta})) = L_i\log\frac{1}{\delta}\cdot\log\log\frac{1}{\delta}. \tag{6}$$

The total budget used is

$$\begin{aligned}\sum_{j=1}^{J} K_j(\delta_j, SAH\delta_j) &\leq \sum_{j=1}^{J} 2^{j+1} C_K(\delta_j, SAH\delta_j, J)\\ &\leq \sum_{j=1}^{J} 2^{J+1} C_K(\frac{1}{8SAH}, \frac{1}{8}, J)\\ &\leq 2J(1 + \frac{\log(2)B}{c(B)})^{0.6321} C_K(\frac{1}{8SAH}, \frac{1}{8}, J)\\ &\leq 2J(1 + \frac{B}{c(B)}) C_K(\frac{1}{8SAH}, \frac{1}{8}, J),\end{aligned}$$

where the second inequality follows from (6) and that C_K has $\log(\frac{1}{\delta})\log\log(\frac{1}{\delta})$ dependence. If $B \geq c(B)$, then the above is bounded by

$$\frac{4JB}{c(B)} C_K(\frac{1}{8SAH}, \frac{1}{8}, J) = B.$$

We now prove the second part. By union bounding Theorem C.1 over $i = 1, 2, \ldots i_\varepsilon$, (1) hold with probability at least

$$1 - \sum_{i=1}^{i_\varepsilon} \delta_i \geq 1 - i_\varepsilon \delta_{i_\varepsilon}.$$

Here, $\delta_{i_\varepsilon} = \exp(-\tilde{\Theta}(L_{i_\varepsilon}))$ by the definition and

$$L_{i_\varepsilon} = 2^{J-i_\varepsilon} \geq \frac{\varepsilon}{4SH^2}(1+\frac{\log(2)B}{c(B)})^{0.6321} \geq \frac{\varepsilon}{4SH^2}(1+0.6321\frac{\log(2)B}{c(B)}) \geq \frac{\varepsilon}{4SH^2}\cdot 0.6321\frac{\log(2)B}{c(B)}.$$

Thus, (1) holds with probability at least $1 - \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(B)}\right)\right)$. Similarly, (2) holds with probability at least

$$1 - SA\sum_{i=1}^{i_\varepsilon}\delta_i.$$

Since SA becomes $\log(SA)$ when moving into the exponential, (2) also holds with probability at least $1 - \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(B)}\right)\right)$. We next compute the probability that (3) holds. For simplicity, let's consider the level $i = i_\varepsilon$, in which the failure probability $SA\delta_{i_\varepsilon}$ is dominant. For the collection of samples via rerunning policies to be successful, we need both $\mathcal{D}_{1,\delta_{i_\varepsilon}}$ and $\mathcal{D}_{2,\delta_{i_\varepsilon}}$ to hold. $\mathcal{D}_{1,\delta_{i_\varepsilon}}$ holds with probability at least $1 - \frac{SA\delta_{i_\varepsilon}}{2}$. On the event $\mathcal{D}_{1,\delta_{i_\varepsilon}}$, consider rerunning each policy in Π_{i_ε} for A times. By Lemma 18, with probability $1 - \exp(-\frac{1}{2}A\log(\frac{1}{eSA\delta_{i_\varepsilon}}))$, at least for $\frac{A}{2}$ trials of repetition, we collect $\frac{N_{i_\varepsilon}}{4}$ samples of each $(s,a) \in \mathcal{X}_{i_\varepsilon}$, which means we collect at least $\frac{AN_{i_\varepsilon}}{8}$ samples of each $(s,a) \in \mathcal{X}_{i_\varepsilon}$. Thus, the probability that there exists some $(s,a) \in \mathcal{X}_{i_\varepsilon}$, the sample number of which is less than $\frac{AN_{i_\varepsilon}}{8}$ is

$$\exp\left(-\frac{1}{2}A\log(\frac{1}{eSA\delta_{i_\varepsilon}})\right) = \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon AB}{C_{\mathrm{L2E}}(B)}\right)\right).$$

However, the failure probability of $\mathcal{D}_{1,\delta_{i_\varepsilon}}$ is already $\exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(B)}\right)\right)$, which is more dominant. Thus, (3) also holds with probability $1-\exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(B)}\right)\right)$. The theorem is proven. □

C.3 Boosting technique

In this subsection, we develop an alternative algorithm of FB-L2E. The core mechanism of this alternative is the boosting technique, which repeatedly executes independent trials. The number of repetitions and the failure probability is in the exponential relationship as we can see in the following lemma.

Lemma 18. *Let $\mathcal{E}$ be an event from a random trial such that $\mathbb{P}(\mathcal{E}) \le \delta$ Let $\alpha \in (\delta, 1)$. Let N be the number of trials where $\mathcal{E}$ is true out of L trials. Assume $\alpha > \delta$. Then,*

$$\mathbb{P}(\frac{N}{L} \ge \alpha) \le \exp\left(-\alpha L \ln\left(\frac{\alpha}{e\delta}\right)\right)$$

Proof. Recall the KL divergence based concentration inequality where $\hat{\mu}_n$ is the sample mean of n Bernoulli i.i.d. random variables with head probability μ:

$$\mathbb{P}(\hat{\mu}_n - \mu \ge \varepsilon) \le \exp(-n\mathsf{kl}(\mu+\varepsilon,\mu))\ .$$

Note that N/L can be viewed as the sample mean of Bernoulli trials with $\mu := \mathbb{P}(\mathcal{E})$. Then,

$$\begin{aligned}\mathbb{P}(N \ge \alpha L) &= \mathbb{P}(\frac{N}{L} \ge \alpha)\\ &= \mathbb{P}(\frac{N}{L} - \mu \ge \alpha - \mu)\\ &\le \exp(-L\mathsf{kl}(\alpha,\mu))\\ &= \exp\left(-L\left(\alpha\ln(\alpha/\mu) + (1-\alpha)\ln\frac{1-\alpha}{1-\mu}\right)\right)\\ &\overset{(a)}{\le} \exp\Big(-L\left(\alpha\ln(\alpha/\mu) - \alpha\right)\Big)\\ &\le \exp\Big(-L\left(\alpha\ln(\alpha/\delta) - \alpha\right)\Big)\end{aligned}$$

where (a) is by the following derivation:

$$\begin{aligned}(1-\alpha)\ln\frac{1-\alpha}{1-\mu} &= -(1-\alpha)\ln\frac{1-\mu}{1-\alpha}\\ &= -(1-\alpha)\ln\left(1+\frac{\alpha-\mu}{1-\alpha}\right)\\ &\ge -(\alpha-\mu)\\ &\ge -\alpha\end{aligned}$$

□

The alternative algorithm, FB-L2E-BS is described in Algorithm 0. Although it only applies to singleton subsets (subset of size 1), one can flexibly change the regret minimization algorithm in FINDEXPLORABLESETS. It was crucial for our result that the regret bound of STRONGEULER has $\log(\frac{1}{\delta})$ dependence. However, for FB-L2E-BS, we can use algorithms such as EULER, which has $\log^3(\frac{1}{\delta})$ dependence in the lower order term.

Algorithm 3 **F**ixed **B**udget **L**earn to **E**xplore with **B**oosting for **S**ingleton (FB-L2E-BS)

```
function FB-L2E-BS($\mathcal{X} = \{(s,a)\} \subseteq \mathcal{S} \times \mathcal{A}$, step $h$, budget $B$)
    if $|\mathcal{X}| = 0$ then
        return $\{(\emptyset, \emptyset, 0, )\}$
    end if
    $J \leftarrow \lceil 0.6321 \log_2(1 + \frac{\log(2)B}{c(B)}) \rceil$
    for $j = 1, \ldots, J$ do
        $K_j \leftarrow K_j(\frac{1}{8SAH}, \frac{1}{8}), \quad N_j \leftarrow K_j/(4|\mathcal{X}| \cdot 2^j), \quad L_j \leftarrow 2^{J-j}$
        for $m = 1, \ldots, L_j$ do
            $\mathcal{Y}_{j,m}, \Pi_{j,m} = \texttt{FindExplorableSets}(\mathcal{X}, h, \frac{1}{8SAH}, K_j, N_j)$
        end for
        Calculate the votes: $\forall (s,a) \in \mathcal{X}, v_{s,a} \leftarrow \sum_{m=1}^{L_j} \mathbb{1}\{(s,a) \in \mathcal{Y}_{j,m}\}$.
        Filter out only if chosen at least half the time: $\mathcal{X}_j \leftarrow \{(s,a) \mid v_{s,a} \geq L_j/2\}$
        $\Pi_j = \cup_{m=1}^{L_j} \Pi_{j,m}$
        $\mathcal{X} \leftarrow \mathcal{X} \backslash \mathcal{X}_j$
    end for
    return $\{(\mathcal{X}_j, \Pi_j, N_j)\}_{j=1}^{J}$
end function
```

We briefly argue that the statements of Theorem 3.1 also hold for FB-L2E-BS used for singleton subset. The total budget used is

$$\begin{aligned}\sum_{j=1}^{J} 2^{J-j} K_j &= J 2^{J+1} C_K(\frac{1}{8SAH}, \frac{1}{8}, J) \\ &\leq 2J(1 + \frac{\log(2)B}{c(B)})^{0.6321} C_K(\frac{1}{8SAH}, \frac{1}{8}, J) \\ &\leq 2J(1 + \frac{\log(2)B}{c(B)}) C_K(\frac{1}{8SAH}, \frac{1}{8}, J).\end{aligned}$$

If $B \geq c(B)$, then the above is bounded by

$$\frac{4JB}{c(B)} C_K(\frac{1}{8SAH}, \frac{1}{8}, J) = B.$$

Let $\delta = \frac{1}{8SAH}, \delta_{\text{samp}} = \frac{1}{8}$. The crucial part for other statements in Theorem 3.1, was to make the failure probability of the j-th iteration in the form of

$$(c_1 \delta)^{c_2 L_j} \tag{7}$$

for some constant c_1, c_2, which was done by defining δ_i as this form in FB-L2E. Once we get (7), the dominant term becomes $(c_1\delta)^{c_2 L_{i_\varepsilon}} = \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\text{L2E}}(B)}\right)\right)$. We show that (7) can also be obtained for FB-L2E-BS.

Assume $W_h(s) \in (2^{-i}, 2^{-i+1}]$ for $i \leq J$. Let's call i as the *reachable index* of s at h. Let $\mathcal{N}_j$ be the event that (s,a) is not filtered in j-th boosted FES. By Lemma 13,

$$\mathbb{P}\left((s,a) \text{ is not filtered in } i\text{-th step by a single FES} | \cap_{j=1}^{i-1} \mathcal{N}_j\right) \leq \delta.$$

If we apply Lemma 18, we obtain the form of (7) as

$$\mathbb{P}\left(\cap_{j=1}^{i} \mathcal{N}_j\right) \leq \mathbb{P}\left(\mathcal{N}_i | \cap_{j=1}^{i-1} \mathcal{N}_j\right) \leq \exp\left(-\frac{1}{2} L_i \log \frac{1}{2e\delta}\right).$$

We say that (s,a) is *upper well-filtered* at h if (s,a) is filtered in the index j for some $j \leq i$.

Now we consider the j-th boosted FES for some $j \leq i - 4$. By Lemma 11, 16,

$$\mathbb{P}\left((s,a) \text{ is filtered in } j\text{-th step by a single FES} | \cap_{k=1}^{j-1} \mathcal{N}_k\right) \leq \frac{\delta_{\text{samp}}}{2SAH}.$$

Thus, by Lemma 18, we obtain the form of (7) as

$$\mathbb{P}\left(\cap_{k=1}^{j-1} \mathcal{N}_k, \quad \mathcal{N}_j^{\mathsf{c}}\right) \leq \mathbb{P}\left(\mathcal{N}_j^{\mathsf{c}} | \cap_{k=1}^{j-1} \mathcal{N}_k\right) \leq \exp\left(-\frac{1}{2} L_j \log \frac{SAH}{e\delta_{\text{samp}}}\right).$$

We say that (s, a) is *lower well-filtered* at h if (s, a) is not filtered in the indices j with $j \leq i - 4$. We also say that (s, a) is *well-filtered* at h if (s, a) is both upper and lower well-filtered at h. We have

$$\mathbb{P}\left((s,a) \text{ is not lower well-filtered at } h\right) \leq \sum_{j=1}^{i-4} \exp\left(-\frac{1}{2} L_j \log \frac{SAH}{e\delta_{\text{samp}}}\right) \leq i \exp\left(-\frac{1}{2} L_i \log \frac{SAH}{e\delta_{\text{samp}}}\right).$$

Thus, we have

$$\mathbb{P}\left((s,a) \text{ is well-filtered at } h\right) \geq 1 - \exp\left(-\frac{1}{2} L_i \log \frac{1}{2e\delta}\right) - i \exp\left(-\frac{1}{2} L_i \log \frac{SAH}{e\delta_{\text{samp}}}\right).$$

Recall that $\varepsilon \geq 2SH^2\varepsilon_B$ and $i_\varepsilon := \lceil \log_2(\frac{2SH^2}{\varepsilon}) \rceil$. We define the set

$$\mathcal{S}_\varepsilon = \{(s, h) : \text{the reachable index of } s \text{ at } h \leq i_\varepsilon\}$$

and the event

$$\mathcal{M}_\varepsilon = \{(s, a) \text{ is well-filtered at } h \text{ for all } (s, h) \in S_\varepsilon\}.$$

By using the monotonicity of L_i and union bound, we have the following.

Lemma 19.

$$\begin{aligned}\mathbb{P}(\mathcal{M}_\varepsilon) &\geq 1 - SH \exp\left(-\frac{1}{2} L_{i^*} \log \frac{1}{2e\delta}\right) - SHi_\varepsilon \exp\left(-\frac{1}{2} L_{i^*} \log \frac{SAH}{e\delta_{\text{samp}}}\right) \\ &= 1 - \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\text{L2E}}(B)}\right)\right).\end{aligned}$$

Let $W_h(S) \in (2^{-i}, 2^{-i+1}]$ and assume that $\mathcal{D}_{1,\delta}$ happened for at least $\frac{L_j}{2}$, where j is the index that (s, a) is filtered. We denote the number of (s, a) samples at horizon h when running each policy in a policy set Π A times as $N_\Pi^A(s, a, h)$. Let $I \subset [L_j]$ be the set of indices that $\mathcal{D}_{1,\delta}$ happened, which means $|I| \geq L_j/2$. Assume $m \in I$. If we rerun each policy in $\Pi_{j,m}$ once,

$$\mathbb{P}(\# \text{ of } (s,a) \text{ samples at horizon } h < \frac{1}{4}N_j) \leq \frac{\delta_{\text{samp}}}{H}$$

by Lemma 15. Now consider rerunning each policy in $\Pi_{j,m}$ A times. Since running policies are independent, we can think of the process as A repetition of running each policy in $\Pi_{j,m}$ once. Thus, we get

$$\mathbb{P}(N_{\Pi_{j,m}}^A(s,a,h) < \frac{1}{8}AN_j) \leq \mathbb{P}(\sum_{i=1}^{A} \mathbb{I}^i_{\{N^1_{\Pi_{j,m}}(s,a,h) < \frac{1}{4}N_j\}} \geq \frac{A}{2}) \leq \exp(-\frac{A}{2}\ln(H/2e\delta_{\text{samp}})),$$

where $\mathbb{I}^i$ is the indicator function for i-th repetition of running each policy in $\Pi_{j,m}$ and the second inequality follows from Lemma 18. If we rerun each policy in Π_j A times,

$$\begin{aligned}\mathbb{P}(N_{\Pi_j}^A < \frac{1}{32}AN_jL_j) \leq \mathbb{P}(\sum_{m \in I} \mathbb{I}_{\{N^A_{\Pi_{j,m}}(s,a,h) < \frac{1}{8}AN_j\}} \geq \frac{|I|}{2}) &\leq \exp(-\frac{Y}{2}\ln(1/2e\exp(-\frac{A}{2}\ln(H/2e\delta_{\text{samp}})))) \\ &\leq \exp(-\frac{L_j}{4}\ln(1/2e\exp(-\frac{A}{2}\ln(H/2e\delta_{\text{samp}})))) \\ &\leq \exp\left(-\tilde{\Theta}\left(AL_j\right)\right)\end{aligned}$$

by Lemma 18. If this happens, let's say that (s, a) is *well-collected* at horizon h for A repetition. However, the failure probability

$$\mathbb{P}(\mathcal{D}_{1,\delta} \text{ happened less than } \frac{L_j}{2}) \leq \exp\left(-\tilde{\Theta}\left(L_j\right)\right),$$

which is more dominant. Thus, the following holds.

Lemma 20. *Consider s whose reachable index at h is $i \leq i_\varepsilon$. If we replay policies saved for (s, a) A times, the number T_{hs} of (s, a) samples we get satisfies*

$$\mathbb{P}\left(T_{hs} < \frac{AN_iL_i}{16}\right) \leq \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\text{L2E}}(B)}\right)\right).$$

D Analysis of SAR

Fix $\varepsilon \geq 0$. We say that an arm i of a bandit m is *ε-good* if $\mu_{m,1} - \mu_{m,i} \leq \varepsilon$. An arm is *$\varepsilon$-bad* if it is not ε-good. Let $g_m(\varepsilon)$ denote the number of ε-good arms in bandit m. We write $k^* := \max\left\{k : \bar{\Delta}_{(KM+1-k)} > \varepsilon\right\}$ and define the following two key events:

$$\mathcal{E}_1 = \left\{\forall k \in [k^*], \quad \tfrac{\varepsilon}{2}\text{-good pairs are not rejected at the end of phase } k\right\}$$

$$\mathcal{E}_2 = \Big\{\forall k \in [(k^*+1), \ldots, K], \quad \text{for every active bandit } m \text{ containing an } \varepsilon\text{-bad arm}$$
$$\text{at the beginning of phase } k, \text{ an } \tfrac{\varepsilon}{2}\text{-good arm in bandit } m \text{ is not rejected}\Big\}$$

We first show that the intersection of these two events leads to a successful good arm identification for every bandit.

Lemma 21. *Suppose $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Then for every $m \in [M]$, the accepted arm is ε-good.*

Proof. Suppose the conclusion is not true; i.e., there exists a bandit m for which an ε-bad arm (m, b) has been accepted. Then, there exists a phase $k \in [KM-1]$ where the best arm $(m, 1)$ is rejected from bandit m. Due to $\mathcal{E}_1$ and the fact that arm $(m, 1)$ is an $\frac{\varepsilon}{2}$-good arm, we know $k \geq k^*+1$. Now, at the beginning of phase k, the bandit m must contain both (m, b) and $(m, 1)$. However, due to $\mathcal{E}_2$, the arm $(m, 1)$ cannot be rejected, which contradicts our supposition. □

Furthermore, consider the following event

$$\mathcal{E}_0 = \left\{\forall m \in [M], \forall i \in [K], \forall k \in [MK-1], \left|\hat{\mu}_{m,i}(n_k) - \mu_{m,i}\right| < \frac{1}{8}(\bar{\Delta}_{(MK+1-k)} \vee \bar{\Delta}_{(g(\varepsilon)+1)})\right\}$$

Lemma 22. $\mathcal{E}_0 \implies \mathcal{E}_1 \cap \mathcal{E}_2$

Proof. Assume $\mathcal{E}_0$. To show $\mathcal{E}_1$, it suffices to show that, for every $k \in [k^*]$, if no $\frac{\varepsilon}{2}$-good arm was rejected before phase k then no $\frac{\varepsilon}{2}$-good arm will be rejected in phase k (i.e., either accepts an arm or rejects a non-$\frac{\varepsilon}{2}$-good arm).

So, let $k \in [k^*]$, which implies that $\bar{\Delta}_{(MK+1-k)} > \varepsilon$ by definition, and assume that no $\frac{\varepsilon}{2}$-good arm was rejected before phase k. Furthermore, $\mathcal{E}_1$ is trivially true if the phase k accepts an arm. Thus, it suffices to assume that the phase k does not accept an arm.

We claim that, at the beginning of phase k, there exists an arm $(\bar{m}, \bar{i}) \in S$ such that

$$\mu_{\bar{m},1} - \mu_{\bar{m},\bar{i}} \geq \bar{\Delta}_{(MK+1-k)}\ .$$

Hereafter, we omit (n_k) from $\hat{\mu}_{\cdot,\cdot}(n_k)$. To prove this claim, first note that there exists $(m', i') \in S$ such that

$$\bar{\Delta}_{m',i'} \geq \bar{\Delta}_{(MK+1-k)}\ .$$

(To see this, first, confirm that this is true with equality if the arm $(MK+1-k)$ is rejected or accepted at each phase k; now, notice that if an arm other than $(MK+1-k)$ was rejected or accepted, then it only makes the equality into $\geq$.) Then, we have the following two cases:

- If $i' \neq 1$, then $\bar{\Delta}_{m',i'} = \mu_{m',1} - \mu_{m',i'}$ by definition, so we can take $\bar{m} = m'$ and $\bar{i} = i'$ to prove the claim.
- If $i' = 1$, then, since phase k does not accept an arm, there must exist another surviving arm $i'' \neq 1$ in bandit m'. Since $\bar{\Delta}_{m',i''} = \mu_{m',1} - \mu_{m',i''}$ and
$$\bar{\Delta}_{m',i''} \geq \bar{\Delta}_{m',2} = \bar{\Delta}_{m',1} = \bar{\Delta}_{m',i'} \geq \bar{\Delta}_{(MK+1-k)}\ ,$$
we can choose $\bar{m} = m'$ and $\bar{i} = i''$ to prove the claim.

Assume that $\mathcal{E}_1$ is false; i.e., an $\frac{\varepsilon}{2}$-good arm in bandit m is rejected. This implies that there exists an active bandit m such that

$$\exists g \in [g_m(\tfrac{\varepsilon}{2})] : \hat{\mu}_{m,\hat{1}_m} - \hat{\mu}_{m,g} \geq \hat{\mu}_{\bar{m},\hat{1}_{\bar{m}}} - \hat{\mu}_{\bar{m},\bar{i}}\ .$$

Note that, using $\mathcal{E}_0$ and $\mu_{m,\hat{1}_m} - \mu_{m,g} \leq \mu_{m,1} - \mu_{m,g} \leq \frac{\varepsilon}{2} < \frac{1}{2}\bar{\Delta}_{(MK+1-k)}$,

$$(\text{LHS}) = \hat{\mu}_{m,\hat{1}_m} - \mu_{m,\hat{1}_m} + \mu_{m,\hat{1}_m} - \mu_{m,g} + \mu_{m,g} - \hat{\mu}_{m,g}$$

$$
\begin{aligned}
&< \frac{\bar{\Delta}_{(MK+1-k)}}{8} + \frac{\bar{\Delta}_{(MK+1-k)}}{2} + \frac{\bar{\Delta}_{(MK+1-k)}}{8} \\
&= \frac{3}{4}\bar{\Delta}_{(MK+1-k)} \,.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
(\text{RHS}) &\geq \hat{\mu}_{\bar{m},1} - \hat{\mu}_{\bar{m},\bar{i}} && ((m,1) \in S \text{ since no } \tfrac{\varepsilon}{2}\text{-good arm rejected before phase } k) \\
&= \hat{\mu}_{\bar{m},1} - \mu_{\bar{m},1} + \mu_{\bar{m},1} - \mu_{\bar{m},\bar{i}} + \mu_{\bar{m},\bar{i}} - \hat{\mu}_{\bar{m},\bar{i}} \\
&> -\frac{1}{8}\bar{\Delta}_{(MK+1-k)} + \bar{\Delta}_{(MK+1-k)} - \frac{1}{8}\bar{\Delta}_{(MK+1-k)} \\
&\geq \frac{3}{4}\bar{\Delta}_{(MK+1-k)} \,.
\end{aligned}
$$

This is a contradiction.

We now prove $\mathcal{E}_2$. Suppose not; there exists a phase $k \geq k^* + 1$ and a bandit m active at the beginning of phase k where an $\frac{\varepsilon}{2}$-good arm (g,m) is rejected even if there was a surviving bad arm (b,m). This means that

$$
\hat{\mu}_{m,g} \leq \hat{\mu}_{m,b}
$$

On the other hand, note that $k \geq k^* + 1$ implies $\bar{\Delta}_{(MK+1-k)} \leq \bar{\Delta}_{(g(\varepsilon)+1)}$, so $\bar{\Delta}_{(MK+1-k)} \vee \bar{\Delta}_{(g(\varepsilon)+1)} = \bar{\Delta}_{(g(\varepsilon)+1)}$. Thus,

$$
\begin{aligned}
\hat{\mu}_{m,g} - \hat{\mu}_{m,b} &= \hat{\mu}_{m,g} - \mu_{m,g} + \mu_{m,g} - \mu_{m,b} + \mu_{m,b} - \hat{\mu}_{m,b} \\
&> -\frac{1}{8}\bar{\Delta}_{(g(\varepsilon)+1)} + \mu_{m,g} - \mu_{m,b} - \frac{1}{8}\bar{\Delta}_{(g(\varepsilon)+1)} && (\mathcal{E}_0) \\
&\geq -\frac{1}{8}\bar{\Delta}_{(g(\varepsilon)+1)} + \frac{1}{2}\bar{\Delta}_{(g(\varepsilon)+1)} - \frac{1}{8}\bar{\Delta}_{(g(\varepsilon)+1)} && (\text{definition of } g \text{ and } b) \\
&> 0
\end{aligned}
$$

This is a contradiction. □

Let

$$
H_1(\varepsilon) := \sum_{i=1}^{MK} \frac{1}{(\bar{\Delta}_{(i)} \vee \varepsilon)^2}, \quad H_2(\varepsilon) := \max_{i \geq g(\varepsilon)+1} \frac{i}{\bar{\Delta}_{(i)}^2}.
$$

We present a relation between these two gap-dependent quantities.

Lemma 23. *$H_2(\varepsilon) \leq H_1(\varepsilon) \leq \frac{g(\varepsilon)}{\varepsilon^2} + \log(\frac{MK}{g(\varepsilon)})H_2(\varepsilon)$.*

Proof. Let $i^* = \arg\max_{i \geq g(\varepsilon)+1} i\bar{\Delta}_i^{-2}$. Note that

$$
\begin{aligned}
H_1(\varepsilon) = \sum_{i \geq 1} (\bar{\Delta}_i \vee \varepsilon)^{-2} &\geq \sum_{i=1}^{g(\varepsilon)} \bar{\Delta}_{g(\varepsilon)+1}^{-2} + \sum_{i \geq g(\varepsilon)+1} \Delta_i^{-2} \\
&\geq \sum_{i=1}^{g(\varepsilon)} \bar{\Delta}_{g(\varepsilon)+1}^{-2} + \sum_{i=g(\varepsilon)+1}^{i^*} \bar{\Delta}_{i^*}^{-2} \\
&= \sum_{i=1}^{g(\varepsilon)} \bar{\Delta}_{g(\varepsilon)+1}^{-2} + (i^* - g(\varepsilon))\bar{\Delta}_{i^*}^{-2} \\
&= \sum_{i=1}^{g(\varepsilon)} \bar{\Delta}_{g(\varepsilon)+1}^{-2} + H_2(\varepsilon) - g(\varepsilon)\bar{\Delta}_{i^*}^{-2} \\
&\geq \sum_{i=1}^{g(\varepsilon)} \bar{\Delta}_{g(\varepsilon)+1}^{-2} + H_2(\varepsilon) - g(\varepsilon)\bar{\Delta}_{g(\varepsilon)+1}^{-2} \\
&\geq H_2(\varepsilon).
\end{aligned}
$$

For the right inequality,

$$\begin{aligned}
H_1(\varepsilon) = \sum_{i\geq 1} \frac{1}{i} i(\bar{\Delta}_i \vee \varepsilon)^{-2} = \\
&\leq \sum_{i=1}^{g(\varepsilon)} \frac{1}{i} i\varepsilon^{-2} + \sum_{i=g(\varepsilon)+1}^{MK} \frac{1}{i} H_2(\varepsilon) \\
&\leq \frac{g(\varepsilon)}{\varepsilon^2} + \log(\frac{MK}{g(\varepsilon)}) H_2(\varepsilon).
\end{aligned}$$

□

We are now ready to prove Theorem 3.2.

Theorem D.1 (Theorem 3.2)**.** If we run Algorithm 2 with $B \geq MK$, then the total number of budget used is at most B and

$$\begin{aligned}
\mathbb{P}(\exists m \in [M] : \mu_{m,1} - \mu_{m,J_B(m)} > \varepsilon) &\leq 2M^2K^2 \exp\left(-\frac{B - MK}{128\sigma^2\,\overline{\log}(MK) \cdot \max_{i\geq g(\varepsilon)+1} i\bar{\Delta}_{(i)}^{-2}}\right) \\
&\leq 2M^2K^2 \exp\left(-\frac{B - MK}{128\sigma^2\,\overline{\log}(MK) \cdot \sum_{i\in[MK]} (\bar{\Delta}_{(i)} \vee \varepsilon)^{-2}}\right).
\end{aligned}$$

Proof. For the first part, the total budget used is bounded as

$$\sum_{k=1}^{MK-1} n_k(B,M,K) + n_{MK-1}(B,M,K) \leq MK + \frac{B - MK}{\overline{\log}(MK)}\left(\frac{1}{2} + \sum_{k=1}^{MK-1} \frac{1}{MK+1-k}\right) = B,$$

where we used $\lceil x \rceil \leq 1 + x$ For the second part, it suffices to bound $\mathbb{P}(\overline{\mathcal{E}}_0)$ by Lemma 21 and Lemma 22. Fix a bandit m and an arm i. Then,

$$\begin{aligned}
&\mathbb{P}\left(\exists k \in [KM-1] : \left|\hat{\mu}_{m,i}(n_k) - \mu_{m,i}\right| \geq \frac{1}{8}(\bar{\Delta}_{(MK+1-k)} \vee \bar{\Delta}_{(g(\varepsilon)+1)})\right) \\
&\leq \sum_{k=1}^{KM-1} 2\exp\left(-\frac{n_k}{2\sigma^2} \cdot \frac{(\bar{\Delta}_{(MK+1-k)} \vee \bar{\Delta}_{(g(\varepsilon)+1)})^2}{64}\right) \\
&\leq \sum_{k=1}^{KM-1} 2\exp\left(-\frac{B - MK}{\overline{\log}(MK) \cdot (MK+1-k)} \frac{(\bar{\Delta}_{(MK+1-k)} \vee \bar{\Delta}_{(g(\varepsilon)+1)})^2}{128\sigma^2}\right) \\
&\leq 2MK \exp\left(-\frac{B - MK}{128\sigma^2\,\overline{\log}(MK) \cdot \max_{i\in[2..MK]} i(\bar{\Delta}_{(i)} \vee \bar{\Delta}_{(g(\varepsilon)+1)})^{-2}}\right) \\
&\leq 2MK \exp\left(-\frac{B - MK}{128\sigma^2\,\overline{\log}(MK) \cdot \max_{i\geq g(\varepsilon)+1} i\bar{\Delta}_{(i)}^{-2}}\right).
\end{aligned}$$

Taking a union bound over $m \in [M]$ and $i \in [K]$ and Lemma 23 completes the proof. □

Algorithm 4 **B**ackward **R**eachability **E**stimation and **A**ction elimination (BREA)

```
input: Budget $B$
$B' \leftarrow \lfloor \frac{B}{2SH} \rfloor, \quad J \leftarrow \lceil 0.6321 \log_2(1 + \frac{\log(2)B'}{c(B')}) \rceil$
$B'' \leftarrow \frac{B}{2HJ}$
for $h = H, H-1, \ldots, 1$ do
    $\mathcal{Z}_h \leftarrow \emptyset$
    for $s \in \mathcal{S}$ do $\{(\mathcal{X}_j^{sh}, \Pi_j^{sh}, N_j^{sh})\}_{j=1}^J \leftarrow$ FB-L2E($\{(s,1)\}, h, B'$) (1 is an arbitrary action)
        if $\mathcal{X}_h^{sh} = \{(s,1)\}$ for some $j \in [J]$ then
            $\widehat{W}_h(s) \leftarrow 2^{-j+1}, \quad \mathcal{Z}_h \leftarrow \mathcal{Z}_h \cup \{s\}$
        end if
    end for
    for $i = 1$ to $J$ do
        $\mathcal{Z}_{hi} \leftarrow \{s \in \mathcal{Z}_h : \widehat{W}_h(s) = 2^{-i+1}\}, \quad A_1 \leftarrow \mathcal{Z}_{hi} \times \mathcal{A},$
        $\forall (s,a) \in A_1, \quad N(s,a) \leftarrow 0, \quad T(s,a) \leftarrow 0, \quad T_0(s,a) \leftarrow 0, \quad Q(s,a) \leftarrow 0$
        for $k = 1$ to $|\mathcal{Z}_{hi}|A - 1$ do
            $n_k \leftarrow n_k(\lfloor B'' 2^{-i-2} \rfloor, |\mathcal{Z}_{hi}|, A)$ (as defined in (1))
            for $(s,a) \in A_k$ do
                $T_k(s,a) \leftarrow \lfloor \frac{n_k}{N_i^s h} \rfloor$
                Rerun each policy in $\Pi_i^{sh}$ for $T_k - T_{k-1}$ times
                for each time $t = T(s,a) + 1$ to $T_k(s,a)$ do
                    if $(s,a)$ is visited at step $h$ then
                        Take action $a$ and extend a trajectory using $\{\hat{\pi}_{h'}\}_{h'=h+1}^H$
                        $N(s,a) \leftarrow N(s,a) + 1$
                        $Q(s,a) \leftarrow Q(s,a) + \sum_{h'=h}^H R_{h'}^t(s_{h'}^t, a_{h'}^t)$
                    end if
                end for
                $\hat{Q}_h^{\hat{\pi}}(s,a) \leftarrow Q(s,a)/N(s,a)$ if $N(s,a) > 0$ else $0$
                $T(s,a) \leftarrow T_k(s,a)$
            end for
            if $\exists$ state $s$ with unique surviving pair $(s,a)$ in $A_k$ then
                $\hat{\pi}_h(s) \leftarrow a, \quad A_{k+1} \leftarrow A_k \setminus \{(s,a)\}$
            else
                $\forall (s,a) \in A_k, \quad \widehat{\Delta}_h^{\hat{\pi}}(s,a) \leftarrow \max_{a:(s,a) \in A_k} \hat{Q}_h^{\hat{\pi}}(s,a) - \hat{Q}_h^{\hat{\pi}}(s,a)$
                $(s', a') \leftarrow \arg\max_{(s,a) \in A_k} \widehat{\Delta}_h^{\hat{\pi}}(s,a)$ (Break ties arbitrarily)
                $A_{k+1} \leftarrow A_k \setminus \{(s', a')\}$
            end if
        end for
        $\hat{\pi}(s) \leftarrow a$ for $A_{|\mathcal{Z}_{hi}|A} = \{(s,a)\}$
    end for
    For each $s \in \mathcal{S} \setminus \mathcal{Z}_h$, set $\hat{\pi}_h(s)$ as any action
end for
return $\hat{\pi}$
```

E Proof of Theorem 3.3

In this section, we provide an analysis of BREA. For convenience, we write the full algorithm again. Recall that $\varepsilon \geq 2SH^2\varepsilon_B$ and $i_\varepsilon = \lceil \log_2(\frac{2SH^2}{\varepsilon})\rceil$ We define the events

$$\begin{aligned}
\mathcal{M}_{h,\varepsilon} =& \Big\{ \text{For any } s \in \mathcal{S}, \\
&\quad \text{FB-L2E}(\{(s,1)\}, h, B') \text{ outputs } \mathcal{X}_i = \{(s,1)\} \text{ for some } i \leq i_\varepsilon \implies 2^{-i-3} \leq W_h(s) \leq 2^{-i+1}, \\
&\quad \text{FB-L2E}(\{(s,1)\}, h, B') \text{ outputs } \mathcal{X}_i = \emptyset \text{ for all } i \in [i_\varepsilon] \implies W_h(s) \leq \frac{\varepsilon}{2SH^2} \Big\}, \\
\mathcal{M}_\varepsilon =& \cup_{h=1}^H \mathcal{M}_{h,\varepsilon}, \\
\mathcal{L}_{h,\varepsilon} =& \Big\{ \text{For any } i \leq i_\varepsilon \text{ and any phase } k \in [|\mathcal{Z}_{hi}|A - 1], \\
&\quad \text{each } (s,a) \in A_k \text{ is collected at least } \lfloor \frac{n_k}{N_i^{sh}} \rfloor \frac{N_i^{sh}}{8} \text{ times} \Big\}, \\
\mathcal{L}_\varepsilon =& \cup_{h=1}^H \mathcal{L}_{h,\varepsilon} \\
\mathcal{E}_h =& \Big\{ \Delta_h^{\hat{\pi}}(s, \hat{\pi}_h(s)) \leq \frac{\varepsilon}{2C_h H W_h(s)} \text{ for all } s \in \cup_{i=1}^{i_\varepsilon} \mathcal{Z}_{hi} \Big\}.
\end{aligned} \tag{8}$$

Before proving Theorem 3.3, we provide lemmas that will give us a relation between the suboptimality gap and its empirical estimate.

Lemma 24. *Let $0 < a \leq b$ and assume $f_1, f_2 \geq 0$ satisfy $|f_1 - f_2| \leq b$. Then*

$$(f_1 \vee a)^{-2} \leq (\frac{a}{2b} f_2 \vee a)^{-2}.$$

Proof. If $f_1 \leq a$, then $(f_1 \vee a)^{-2} = a^{-2}$. On the other hand, $f_2 \leq f_1 + b \leq a + b \leq 2b$. Thus,

$$(f_1 \vee a)^{-2} = a^{-2} = (\frac{a}{2b} f_2 \vee a)^{-2}.$$

If $f_1 > a$, then $(f_1 \vee a)^{-2} = f_1^{-2} < a^{-2}$. Also, $f_2 \leq f_1 + b < f_1 + \frac{f_1}{a} b = f_1(1 + \frac{b}{a}) \leq \frac{2b}{a} f_1$. Thus,

$$(f_1 \vee a)^{-2} = f_1^{-2} < (\frac{a}{2b} f_2 \vee a)^{-2}.$$

□

Lemma 25. *On $\cap_{h'=H}^{h+1} \mathcal{E}_{h'} \cap \mathcal{M}_\varepsilon \cap \mathcal{L}_\varepsilon$, we have*

$$(\Delta_h^{\hat{\pi}}(s,a) \vee \frac{\varepsilon}{2C_h H W_h(s)})^{-2} \leq 16 C_h^2 H^2 (\Delta_h(s,a) \vee \frac{2\varepsilon}{W_h(s)})^{-2}.$$

Proof. By Lemma 4, for any policy π'

$$\begin{aligned}
\sum_s w_{h+1}^{\pi'}(s)(V_{h+1}^*(s) - V_{h+1}^{\hat{\pi}}(s)) &\leq \sum_{h'=h+1}^H \sup_\pi \sum_s w_{h'}^\pi(s) \varepsilon_h(s) \\
&\leq \sum_{h'=h+1}^H \sup_\pi \sum_{i \leq i_\varepsilon} \sum_{s \in \mathcal{Z}_{hi}} w_{h'}^\pi(s) \frac{\varepsilon}{2C_h H W_h(s)} + H \sum_{h'=h+1}^H \sum_{s \notin \cup_{i \leq i_\varepsilon} \mathcal{Z}_{hi}} \sup_\pi w_{h'}^\pi(s) \\
&\leq \sum_{h'=h+1}^H \frac{\varepsilon}{2H} + \sum_{h'=h+1}^H SH \frac{\varepsilon}{2SH^2} \\
&\leq \varepsilon.
\end{aligned}$$

By Lemma 5,

$$|\Delta_h(s,a) - \Delta_h^{\hat{\pi}}(s,a)| \leq \frac{\varepsilon}{W_h(s)}.$$

By applying Lemma 24 with $f_1 = \Delta_h^{\hat{\pi}}(s,a), f_2 = \Delta_h(s,a), a = \frac{\varepsilon}{2C_h H W_h(s)}, b = \frac{\varepsilon}{W_h(s)}$, the proof is done. □

Theorem E.1 (Theorem 3.3)**.** *If we run Algorithm 4 with sufficiently large budget B, then the total number of budget used is at most B. Moreover, for any $\varepsilon \geq 2SH^2\varepsilon_{\frac{B}{2SH}}$,*

$$\mathbb{P}\left(V_0^* - V_0^{\hat{\pi}} > \varepsilon\right) \leq \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(\frac{B}{2SH})}\right)\right) + \exp\left(-\tilde{\Theta}\left(\frac{B}{H^5 \max_{h\in[H]} C_h^2 \sum_{s\in\mathcal{S}} W_h(s)^{-1} \sum_{a\in\mathcal{A}}(\bar{\Delta}_h(s,a) \vee \frac{\varepsilon}{W_h(s)})^{-2}}\right)\right).$$

Proof. The budget used from the first part is

$$SH\lfloor\frac{B}{2SH}\rfloor \leq \frac{B}{2}$$

by Theorem 3.1. For the second part, we use

$$\begin{aligned}&\sum_{i=1}^{|\mathcal{Z}_{hi}|A-1} T_i(s,a) + T_{|\mathcal{Z}_{hi}|A-1}(s,a)\\ \leq&\frac{1}{N_i}\left(\sum_{i=1}^{|\mathcal{Z}_{hi}|A-1} n_i + n_{|\mathcal{Z}_{hi}|A-1}\right)\\ \leq&\frac{\lfloor B''2^{-i-2}\rfloor}{2^{i+2}} \leq B'' = \frac{B}{2HJ} && \text{(Theorem 3.2)}\end{aligned}$$

for each multiple bandit $\mathcal{Z}_{hi}$. Thus, the budget used in the second part is at most $\frac{B}{2}$, the total budget used is at most B.

We now prove the probability bound. By Theorem 3.1, we have

$$\begin{aligned}\mathbb{P}(\mathcal{M}_\varepsilon^{\mathsf{c}}) &\leq SH\exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(\frac{B}{2SH})}\right)\right) = \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(\frac{B}{2SH})}\right)\right),\\ \mathbb{P}(\mathcal{L}_\varepsilon^{\mathsf{c}}) &\leq S^2A^2H\exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(\frac{B}{2SH})}\right)\right) = \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(\frac{B}{2SH})}\right)\right).\end{aligned} \tag{9}$$

We can decompose the probability as

$$\begin{aligned}\mathbb{P}(V_0^* - V_0^{\hat{\pi}} > \varepsilon) &\leq \mathbb{P}(V_0^* - V_0^{\hat{\pi}} > \varepsilon, \mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon) + \mathbb{P}(\mathcal{M}_\varepsilon^{\mathsf{c}}) + \mathbb{P}(\mathcal{L}_\varepsilon^{\mathsf{c}})\\ &\leq \mathbb{P}(V_0^* - V_0^{\hat{\pi}} > \varepsilon, \mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon) + \exp\left(-\tilde{\Theta}\left(\frac{\varepsilon B}{C_{\mathrm{L2E}}(\frac{B}{2SH})}\right)\right).\end{aligned} \tag{10}$$

Assume that $\mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon, \{\mathcal{E}_h\}_{h=1}^H$ holds. Then, by Lemma 4,

$$\begin{aligned}V_0^* - V_0^{\hat{\pi}} &\leq \sum_{h=1}^H \sup_\pi \sum_s w_h^\pi(s)\varepsilon_h(s)\\ &\leq \sum_{h=1}^H \sup_\pi \sum_{i\leq i_\varepsilon}\sum_{s\in\mathcal{Z}_{hi}} w_h^\pi(s)\frac{\varepsilon}{2C_hHW_h(s)} + H\sum_{h=1}^H \sum_{s\notin\cup_{i\leq i_\varepsilon}\mathcal{Z}_{hi}} \sup_\pi w_h^\pi(s)\\ &\leq \sum_{h=1}^H \frac{\varepsilon}{2H} + SH^2\frac{\varepsilon}{2SH^2}\\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,\end{aligned}$$

where the second inequality follows from the definition of C_h. Thus, we have

$$\mathbb{P}(V_0^* - V_0^{\hat{\pi}} > \varepsilon, \mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon) \leq \sum_{h=1}^H \mathbb{P}(\mathcal{E}_h^{\mathsf{c}}, \mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon, \cap_{h'=H}^{h+1}\mathcal{E}_{h'}). \tag{11}$$

We try to bound $\mathbb{P}(\mathcal{E}_h^{\mathsf{c}}, \mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon, \cap_{h'=H}^{h+1}\mathcal{E}_{h'})$.

On the event $\mathcal{L}_\varepsilon$, every multiple bandit instance $\mathcal{Z}_{hi}$ effectively collects samples so that SAR with budget $\Theta(\frac{B}{2HJ}2^{-i-2})$ is run. On the event $\mathcal{M}_\varepsilon$, this is $\Theta(\frac{BW_h(s)}{HJ}) = \Theta(\frac{BW_h(s)}{H})$. By Theorem 3.2, we have

$$
\begin{aligned}
&\mathbb{P}\left(\Delta_h^{\hat{\pi}}(s, \hat{\pi}_h(s)) > \frac{\varepsilon}{2C_h H W_h(s)} \text{ for some } s \in \mathcal{Z}_{hi}, \mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon, \cap_{h'=H}^{h+1} \mathcal{E}_{h'} | \mathcal{F}_{h+1}\right) \\
&\leq \exp\left(-\tilde{\Theta}\left(\frac{B}{H^3 \sum_{(s,a)\in \mathcal{Z}_{hi}\times\mathcal{A}} W_h(s)^{-1}(\Delta_h^{\hat{\pi}}(s,a) \vee \frac{\varepsilon}{2C_h H W_h(s)})^{-2}}\right)\right) \\
&\leq \exp\left(-\tilde{\Theta}\left(\frac{B}{C_h^2 H^5 \sum_{(s,a)\in \mathcal{Z}_{hi}\times\mathcal{A}} W_h(s)^{-1}(\Delta_h(s,a) \vee \frac{\varepsilon}{W_h(s)})^{-2}}\right)\right),
\end{aligned}
$$

where the second inequality follows from Lemma 25, $\mathcal{F}_{h+1}$ is a filtration up to learning in step $h+1$. Thus, we have

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_h^{\mathsf{c}}, \mathcal{M}_\varepsilon, \mathcal{L}_\varepsilon, \cap_{h'=H}^{h+1}\mathcal{E}_{h'}) &\leq i_\varepsilon \exp\left(-\tilde{\Theta}\left(\frac{B}{H^5 \max_h C_h^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} W_h(s)^{-1}(\Delta_h(s,a) \vee \frac{\varepsilon}{W_h(s)})^{-2}}\right)\right) \\
&= \exp\left(-\tilde{\Theta}\left(\frac{B}{H^5 \max_h C_h^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} W_h(s)^{-1}(\Delta_h(s,a) \vee \frac{\varepsilon}{W_h(s)})^{-2}}\right)\right).
\end{aligned}
$$

If we plug this into (11) and (10), we get the probability bound of the theorem. □