Gap-Dependent Unsupervised Exploration for Reinforcement Learning

Jingfeng Wu
Johns Hopkins Universit
uuujf@jhu.edu

Vladimir Braverman
Johns Hopkins University
vova@cs.jhu.edu

Lin F. Yang
University of California, Los Angeles
linyang@ee.ucla.edu

Abstract

For the problem of task-agnostic reinforcement learning (RL), an agent first collects samples from an unknown environment without the supervision of reward signals, then is revealed with a reward and is asked to compute a corresponding near-optimal policy. Existing approaches mainly concern the worst-case scenarios, in which no structural information of the reward/transition-dynamics is utilized. Therefore the best sample upper bound is $\propto \widetilde{\mathcal{O}}(1/\epsilon^2)$, where $\epsilon > 0$ is the target accuracy of the obtained policy, and can be overly pessimistic. To tackle this issue, we provide an efficient algorithm that utilizes a gap parameter, $\rho > 0$, to reduce the amount of exploration. In particular, for an unknown finite-horizon Markov decision process, the algorithm takes only $\widetilde{\mathcal{O}}(1/\epsilon \cdot (H^3SA/\rho + H^4S^2A))$ episodes of exploration, and is able to obtain an ϵ-optimal policy for a post-revealed reward with sub-optimality gap at least ρ, where S is the number of states, A is the number of actions, and H is the length of the horizon, obtaining a nearly *quadratic saving* in terms of ϵ. We show that, information-theoretically, this bound is nearly tight for $\rho < \Theta(1/(HS))$ and $H > 1$. We further show that $\propto \widetilde{\mathcal{O}}(1)$ sample bound is possible for $H = 1$ (i.e., multi-armed bandit) or with a sampling simulator, establishing a stark separation between those settings and the RL setting.

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

Unsupervised exploration is an emergent and challenging topic for reinforcement learning (RL) that inspires research interests in both application (Riedmiller et al., 2018; Finn and Levine, 2017; Xie et al., 2018, 2019; Schaul et al., 2015; Riedmiller et al., 2018) and theory (Hazan et al., 2018; Jin et al., 2020; Zhang et al., 2020a; Kaufmann et al., 2020; Ménard et al., 2020; Zhang et al., 2020b; Wu et al., 2021; Wang et al., 2020b). The formal formulation of an unsupervised RL problem consists of an *exploration phase* and a *planning phase* (Jin et al., 2020): in the exploration phase, an agent interacts with the unknown environment without the supervision of reward signals; then in the planning phase, the agent is prohibited to interact with the environment, and is required to compute a nearly optimal policy for some revealed reward function based on its exploration experiences. In particular, if the reward function is *independent* of the agent's exploration policy, the problem is called *task-agnostic exploration* (TAE) (Zhang et al., 2020a); and if the reward function is chosen *adversarially* according to the agent's exploration policy, the problem is called *reward-free exploration* (RFE) (Jin et al., 2020). The performance of an unsupervised exploration algorithm is measured by the *sample complexity*, i.e., the number of samples the algorithm needs to collect during the exploration phase in order to complete the planning task near-optimally up a small error (with high probability, see Section 2 for a formal definition). Existing algorithms for unsupervised RL exploration (Jin et al., 2020; Zhang et al., 2020a; Wu et al., 2021; Zhang et al., 2020b; Wang et al., 2020b) suffer a sample complexity (upper bounded by) $\propto \widetilde{\mathcal{O}}(1/\epsilon^2)$[1] for a target planning error tolerance ϵ. In a worst-case consideration, this rate, in terms of dependence on ϵ, is known to be unimprovable except for logarithmic factors (Jin et al., 2020; Dann and Brunskill, 2015).

[1]Here we use $\propto \widetilde{\mathcal{O}}(\cdot)$ to emphasize the rates' dependence on ϵ, where the other parameters are treated as constants. Similarly hereafter.

However, the above worst-case sample bounds can be *overly pessimistic* in practical scenarios, since the planning task is often known to be a benign instance, even though the exploration phase is conducted under no supervision from rewards. In particular, the reward function revealed in the planning phase could induce a constant *minimum nonzero sub-optimality gap* (or simply *gap*, that is the minimum gap between the best action and the second best action in the optimal Q-value function, and is defined formally in Section 2) (Tewari and Bartlett, 2007; Ortner and Auer, 2007; Ok et al., 2018) such that the planning task is essentially an "easy" one (Simchowitz and Jamieson, 2019). More importantly, a *reward-agnostic* gap parameter, e.g., an uniform lower bound on the gaps of possibly revealed reward functions, could be available to facilitate the exploration process. See following for an example.

An Example. Let us consider training an agent for Go-Game (vs. an unknown player), where the winning rule could be either the Chinese rule, the Japanese rule or the Korean rule. The agent will be rewarded 1 for winning and 0 for losing, and its goal is to explore without knowing the winning condition and to provide a solution to one/any of these rules specified afterward. Note that all these winning conditions have an uniform, constant gap lower bound (≈ 1, assuming that the opponent plays nearly deterministically). Note that this RL problem is still unsupervised as the winning condition is unknown during exploration; but the uniform gap lower bound could potentially be exploited to accelerate exploration.

Open Problem. In the supervised RL setting where the reward signals are available, a constant gap significantly improves the sample complexity bounds, e.g., from $\propto \widetilde{\mathcal{O}}(1/\epsilon^2)$ to $\propto \widetilde{\mathcal{O}}(1)$ (Jaksch et al., 2010; Simchowitz and Jamieson, 2019; Yang et al., 2020; He et al., 2020; Xu et al., 2021). However, the following question remains open for unsupervised RL:

Can unsupervised RL problems be solved faster when provided with a reward-agnostic gap parameter?

A Case Study on Multi-Armed Bandit. To gain more intuition, let us take a quick look at the (gap-dependent) unsupervised exploration problem for multi-armed bandit (MAB). In the worst-case setup, there is a minimax lower bound $\propto \Omega(1/\epsilon^2)$ for unsupervised exploration on an MAB instance (Mannor and Tsitsiklis, 2004), where ϵ is a small tolerance for the planning error. On the other hand, if the MAB instance has a constant gap, a rather simple *uniform exploration* strategy achieves $\propto \widetilde{\mathcal{O}}(1)$ sample complexity upper bound (see, e.g., Theorem 33.1 in (Lattimore and Szepesvári, 2020), or Appendix D). This example provides positive evidence that a constant gap parameter could accelerate unsupervised RL, too.

Our Contributions. In this paper, we study the *gap-dependent task-agnostic exploration* (gap-TAE) problem on a finite-horizon Markov decision process (MDP) with S states, A actions and $H \geq 2$ decision steps per episode. We consider a variant of upper-confidence-bound (UCB) algorithm that explores the unknown environment through a greedy policy that minimizes the cumulative exploration bonus (Zhang et al., 2020a; Wang et al., 2020b; Wu et al., 2021); our exploration bonus is of UCB-type, but is *clipped* according to the gap parameter. Theoretically, we show that $\widetilde{\mathcal{O}}(H^3SA/(\rho\epsilon) + H^4S^2A/\epsilon)$ number of trajectories is sufficient for the proposed algorithm to plan ϵ-optimally for a task with a gap parameter ρ, where $\epsilon > 0$ is the planning error parameter. This fast rate $\propto \widetilde{\mathcal{O}}(1/\epsilon)$ improves the existing, pessimistic rates $\propto \widetilde{\mathcal{O}}(1/\epsilon^2)$ (Zhang et al., 2020a; Wang et al., 2020b; Wu et al., 2021) significantly when $\epsilon \ll \rho$. Furthermore, we provide an information-theoretic lower bound, $\Omega(H^2SA/(\rho\epsilon))$, on the number of trajectories required to solve the problem of gap-TAE on MDPs with $H \geq 2$. This indicates that, for gap-TAE on MDP with $H \geq 2$, the $\propto \widetilde{\mathcal{O}}(1/\epsilon)$ rate achieved by our algorithm is nearly the best possible. These results naturally extend to other unsupervised RL settings.

Interestingly, our results imply that RL is *statistically harder* than MAB in the setting of gap-dependent unsupervised exploration. In particular, a finite-horizon MDP with $H = 1$ reduces to an MAB problem, where it is known that $\propto \widetilde{\mathcal{O}}(1)$ samples are sufficient for solving gap-TAE; however when $H \geq 2$ which corresponds to the general RL setting, our results show that at least $\propto \Omega(1/\epsilon)$ amount of samples are required for solving gap-TAE. This is against an emerging wisdom from the supervised RL theory, that RL ($H \geq 2$) is statistically as easy as learning MAB ($H = 1$) when the H factor is normalized, ignoring logarithmic factors (Jiang and Agarwal, 2018; Wang et al., 2020a; Zhang et al., 2020c).

Notations. For two functions $f(x) \geq 0$ and $g(x) \geq 0$ defined for $x \in [0, \infty)$, we write $f(x) \lesssim g(x)$ if $f(x) \leq c \cdot g(x)$ for some absolute constant $c > 0$; we write $f(x) \gtrsim g(x)$ if $g(x) \lesssim f(x)$; and we write $f(x) \eqsim g(x)$ if $f(x) \lesssim g(x) \lesssim f(x)$. Moreover, we write $f(x) = \mathcal{O}(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) < c$ for some absolute constant $c > 0$; we write $f(x) = \Omega(g(x))$ if $g(x) = \mathcal{O}(f(x))$; and we write $f(x) = \Theta(g(x))$ if $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$. To hide the logarithmic factors, we write $f(x) = \widetilde{\mathcal{O}}(g(x))$ if $f(x) = \mathcal{O}(g(x) \log^d x)$ for some absolute constant

$d > 0$. For $a, b \in \mathbb{R}$, we write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For a positive integer H, we define $[H] := \{1, 2, \dots, H\}$.

2 PROBLEM SETUP

Finite-Horizon MDP. We focus on *finite-horizon Markov decision process* (MDP), which is specified by a tuple, $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, x_1, r)$. $\mathcal{S}$ is a finite *state* set where $|\mathcal{S}| = S$. $\mathcal{A}$ is a finite *action* set where $|\mathcal{A}| = A$. H is the *length of the horizon.* $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to [0, 1]^{\mathcal{S}}$ is an *unknown, stationary transition probability.* Without lose of generality, we assume the MDP has fixed initial state x_1[2]. For simplicity we only consider *deterministic and bounded reward function*[3], which is denoted by $r = \{r_1, \dots, r_H\}$ where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function at the h-th step. A *policy* is represented by $\pi := \{\pi_1, \dots, \pi_H\}$, where each $\pi_h : \mathcal{S} \to [0, 1]^{\mathcal{A}}$ is a potentially random policy at the h-th step. For a policy π, the *Q-value function* and the *value function* are defined as

$$Q_h^\pi(x, a) := \mathbb{E}_{\pi,\mathbb{P}}\big[\sum_{t \ge h} r_t(x_t, a_t)\big| x_h = x, a_h = a\big],$$
$$V_h^\pi(x) := Q_h^\pi(x, \pi_h(x)),$$

where the trajectory is given by $x_t \sim \mathbb{P}(\cdot \mid x_{t-1}, a_{t-1})$ and $a_t \sim \pi_t(x_t)$ for $t > h$. The following well-known Bellman equation is worth mentioning:

$$Q_h^\pi(x, a) = r_h(x, a) + \mathbb{E}_{y \sim \mathbb{P}(\cdot|x,a)} V_{h+1}^\pi(y).$$

$\pi^* \in \arg\max_\pi V_1^\pi(x_1)$ is an optimal policy, and its induced optimal Q-value function and the optimal value function are denoted by $Q_h^*(x, a) := Q_h^{\pi^*}(x, a)$ and $V_h^*(x) := V_h^{\pi^*}(x)$, respectively.

Sub-Optimality Gap. Given an MDP, the *stage-dependent state-action sub-optimality gap* (see, e.g., Simchowitz and Jamieson (2019)) is defined as

$$\mathtt{gap}_h(x, a) := V_h^*(x) - Q_h^*(x, a) \ge 0.$$

Clearly, $\mathtt{gap}_h(x, a) = 0$ if and only if a is an optimal action at state x and at the h-th decision step. Intuitively, when $\mathtt{gap}_h(x, a) > 0$, $\mathtt{gap}_h(x, a)$ characterizes the difficulty to distinguish the sub-optimal action a from the optimal actions at state x and at the h-th step; and the larger $\mathtt{gap}_h(x, a)$ is, the easier should it be distinguishing a from the optimal actions. The *minimum sub-optimality gap* is then defined as

$$\mathtt{gap}_{\min} := \min_{h,x,a}\{\mathtt{gap}_h(x, a) : \ \mathtt{gap}_h(x, a) > 0\}. \quad (2.1)$$

Intuitively, an MDP with a constant $\mathtt{gap}_{\min}$ is easy to learn since constant number of visitations to an state-action pair suffices to distinguish whether it is optimal (Simchowitz and Jamieson, 2019).

Task-Agnostic Exploration. The problem of *task-agnostic exploration* (TAE) (Zhang et al., 2020a) involves an MDP environment $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, x_1)$ and a set of reward functions

$$\mathcal{R} \subset \{r : [H] \times \mathcal{S} \times \mathcal{A} \to [0, 1]\}$$

that could possibly contain infinitely many reward functions. An agent first decides exploration policies to collect K trajectories from the environment, in which process reward feedback is not observable. Then an *oblivious* chooses a reward function r from the candidate set $\mathcal{R}$ and reveals it to the agent [4], i.e., $r \in \mathcal{R}$ is selected *independently* from the agent's exploration policy. Then the agent needs to compute an (ϵ, δ)-*probably-approximately-correct* (PAC) policy π under the reward function r, which means:

$$\text{for an oblivious}^5\ r,\ \mathbb{P}\{V_1^*(x_1) - V_1^\pi(x_1) > \epsilon\} < \delta, \quad (2.2)$$

where the probability is over the randomness of trajectory collecting in the exploration phase. The *sample complexity* is measured by the number of trajectories K that the agent needs to collect in the exploration phase to guarantee being (ϵ, δ)-PAC in planning phase.

Here are a few remarks on the TAE setting. Firstly, note that the reward function r in TAE is oblivious and is not adversarial to the agent's exploration policy. Due to this non-adversarial nature, TAE can be achieved with a sample complexity $\propto \widetilde{\mathcal{O}}(S)$ (Zhang et al., 2020a; Wu et al., 2021), which is much cheaper than that required to estimate the transition kernel accurately ($\propto \mathcal{O}(S^2)$, the error is measured by total variation distance). On the other hand, if the reward

[2]We may as well consider MDPs with an external initial state x_0 with zero reward for all actions, and a transition $\mathbb{P}_0(\cdot \mid x_0, a) = \mathbb{P}_0(\cdot)$ for all action a, which is equivalent to our setting by letting the horizon length H be $H + 1$.

[3]For the sake of presentation, we focus on bounded deterministic reward functions in this work. The techniques can be readily extended to stochastic reward settings and the obtained bounds will match our presented ones.

[4]In Zhang et al. (2020b), N reward functions are selected during planning and only bandit signals are available. For the sake of presentation, we assume there is only 1 reward function and the agent is provided with the full-information of the reward function in the planning phase. Our results and techniques are ready to be extended to setting of Zhang et al. (2020b) in a standard manner.

[5]An oblivious reward means the reward is chosen independent of the agent's exploration policy, or equivalently, the reward is chosen prior to the beginning of the exploration phase. For example, a reward chosen uniformly at random from $\mathcal{R}$ is oblivious.

function is allowed to be chosen *adversarially* against the agent's exploration policy, the problem is known as *reward-free exploration* (RFE) (Jin et al., 2020); the counter part of condition (2.2) in RFE reads:

$$\text{for any } r \in \mathcal{R},\ \mathbb{P}\left\{V_1^*(x_1) - V_1^\pi(x_1) > \epsilon\right\} < \delta,$$

where the probability is over the randomness of trajectory collecting in the exploration phase. Due to the adversarial nature, a RFE algorithm must estimate the environment with high precision, and a $\propto \mathcal{O}(S^2)$ sample complexity is unavoidable (Jin et al., 2020). In the following paper we will focus on the TAE setting to show a more sample-efficient algorithm for benign TAE problems. Nonetheless, our algorithm naturally extends to other unsupervised RL settings (Jin et al., 2020; Wu et al., 2021) and the improved sample-efficiency for benign instances also holds (by a standard covering argument on value functions or reward functions). We refer the reader to Remark 3 in Section 5 for an example of applying our results in reward-free exploration.

Gap-Dependent Unsupervised RL. We now formally state the problem of *gap-dependent task-agnostic exploration* (gap-TAE). Gap-TAE is a benign TAE instance, in which we assume there is a constant, reward-agnostic gap parameter ρ such that

$$0 < \rho \le \mathtt{gap}_{\min}(r) \text{ for every } r \in \mathcal{R}, \tag{2.3}$$

where, with a slightly abuse of notation, $\mathtt{gap}_{\min}(r)$ refers to the minimum sub-optimality gap (2.1) induced by reward function r; moreover, the gap parameter ρ is known to the agent. Our focus is to study whether or not gap-TAE problems can be solved in a faster rate compared to the worst-case TAE problems.

We end this section with the definition of a *clip operator* (Simchowitz and Jamieson, 2019):

$$\mathtt{clip}_\rho[z] := z \cdot \mathbb{1}\left[z \ge \rho\right],\ \text{for } z \in \mathbb{R} \text{ and } \rho > 0,$$

which cuts a quantity smaller than ρ to 0.

3 AN EFFICIENT ALGORITHM

In this section, we introduce UCB-Clip for solving gap-TAE in a sample-efficient manner. The algorithm is formally presented as Algorithms 1 and 2, for exploration and planning, respectively.

In the exploration phase, UCB-Clip maintains an estimated maximum cumulative *bonus* based on the current empirical transition kernel (Algorithm 1, line 6), and explores the environment through executing a greedy policy that maximizes the cumulative bonus. Note that the exploration bonus at a state-action pair is inversely proportional to the number of visitations to the state-action pair (Algorithm 1, line 5), thus the exploration policy is encouraged to pay more visits to the state-actions that have not yet been visited for sufficient times, where its induced bonus is larger. This design permits UCB-Clip to dynamically explore the environment. In the planning phase, UCB-Clip receives an obliviously/independently revealed reward function, and computes an (averaging of a sequence

Algorithm 1 UCB-Clip (Exploration)

Require: gap parameter ρ, number of episodes K
1: initialize history $\mathcal{H}^0 = \emptyset$
2: set up a constant $\iota := \log(2HS^2AK/\delta)$
3: **for** episode $k = 1, 2, \ldots, K$ **do**
4: $N^k(x,a),\ \widehat{\mathbb{P}}^k(y \mid x,a) \leftarrow \mathtt{Empi\text{-}Prob}(\mathcal{H}^{k-1})$
5: compute exploration bonus $c^k(x,a) := \mathtt{clip}_{\frac{\rho}{2H}}\left[\sqrt{\frac{8H^2\iota}{N^k(x,a)}}\right] + \frac{120(S+H)H^3\iota}{N^k(x,a)} + \frac{240H^6S^2\iota^2}{(N^k(x,a))^2}$
6: $\{\overline{Q}_h^k(x,a), \overline{V}_h^k(x)\}_{h=1}^H \leftarrow \mathtt{UCB\text{-}QValue}(\widehat{\mathbb{P}}^k, 0, c^k)$
7: receive initial state $x_1^k = x_1$
8: **for** step $h = 1, 2, \ldots, H$ **do**
9: take action $a_h^k = \arg\max_a \overline{Q}_h^k(x_h^k, a)$
10: obtain a new state x_{h+1}^k
11: **end for**
12: update history $\mathcal{H}^k = \mathcal{H}^{k-1} \cup \{x_h^k, a_h^k\}_{h=1}^H$
13: **end for**
14: **return** History $\mathcal{H}^k$

15: **Function** Empi-Prob
16: **Require:** history $\mathcal{H}^{k-1}$
17: **for** $(x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ **do**
18: $N^k(x,a,y) := \#\{(x,a,y) \in \mathcal{H}^{k-1}\}$
19: $N^k(x,a) := \sum_y N^k(x,a,y)$
20: **if** $N^k(x,a) > 0$ **then**
21: $\widehat{\mathbb{P}}^k(y \mid x,a) = N^k(x,a,y)/N^k(x,a)$
22: **else**
23: $\widehat{\mathbb{P}}^k(y \mid x,a) = 1/S$
24: **end if**
25: **end for**
26: **return** $N^k(x,a), \widehat{\mathbb{P}}^k(y \mid x,a)$

27: **Function** UCB-QValue
28: **Require:** empirical transition $\widehat{\mathbb{P}}^k$, reward function r, bonus function b^k
29: set $V_{H+1}^k(x) = 0$
30: **for** step $h = H, H-1, \ldots, 1$ **do**
31: **for** $(x,a) \in \mathcal{S} \times \mathcal{A}$ **do**
32: $Q_h^k(x,a) = r_h(x,a) + b^k(x,a) + \widehat{\mathbb{P}}_h^k V_{h+1}^k(x,a)$
33: $Q_h^k(x,a) \leftarrow \min\{H,\ Q_h^k(x,a)\}$
34: $V_h^k(x) = \max_{a \in \mathcal{A}} Q_h^k(x,a)$
35: **end for**
36: **end for**
37: **return** $\left\{Q_h^k(x,a),\ V_h^k(x)\right\}_{h=1}^H$

of) *optimistic* estimation to the optimal value function based on the collected samples (Algorithm 2, line 5). The outputted policy equivalently corresponds to the (averaged) optimistic estimated value function (Algorithm 2, line 6).

Algorithm 2 `UCB-Clip` (Planning)

Require: history $\mathcal{H}^K$, reward function r

1: set up a constant $\iota := \log(2HS^2AK/\delta)$
2: **for** $k = 1, 2, \ldots, K$ **do**
3: $\quad N^k(x, a),\ \widehat{\mathbb{P}}^k(y \mid x, a) \leftarrow$ `Empi-Prob`$(\mathcal{H}^{k-1})$
4: $\quad$ compute planning bonus $b^k(x, a) := \sqrt{\frac{H^2\iota}{2N^k(x,a)}}$
5: $\quad \{Q_h^k(x, a),\ V_h^k(x)\}_{h=1}^H \leftarrow$ `UCB-QValue`$(\widehat{\mathbb{P}}^k, r, b^k)$
6: $\quad$ infer greedy policy $\pi_h^k(x) = \arg\max_a Q_h^k(x, a)$
7: **end for**
8: **return** π drawn uniformly from $\{\pi^1, \ldots, \pi^K\}$

`UCB-Clip` is inspired by two existing model-based algorithms: `UCBVI` (Azar et al., 2017) that achieves minimax optimal sample complexity for supervised RL problems, and `PF-UCB` (Wu et al., 2021) that efficiently solves preference-free exploration problems in the context of unsupervised multi-objective RL. Similar to both `UCBVI` and `PF-UCB`, in the planning phase `UCB-Clip` (Algorithm 2) adopts an *upper-confidence-bound* (UCB) type bonus to perform optimistic and model-based planning. Similar to `PF-UCB` but different from `UCBVI`, `UCB-Clip` (Algorithm 1) explores the unknown environment through a greedy policy that maximizes the cumulative exploration bonus. Different from either `PF-UCB` or `UCBVI`, `UCB-Clip` exploits the gap parameter by *clipping* the exploration bonus (Algorithm 1, line 5).

The clipped exploration bonus turns out to be a key ingredient for `UCB-Clip` to save samples for gap-TAE problems. Specifically, the leading-order term in the UCB-type bonus will be brute-force clipped to near zero when a state-action pair has been visited for sufficiently many times, i.e., $N(x, a) \approx H^4\iota/\rho^2$ (Algorithm 1, line 5). This will cause a sudden decrease of the bonus, and discourages the agent to continue to visit this state-action pair. Recall that the considered MDP has a sub-optimality gap at least ρ, thus $N(x, a) \approx H^4\iota/\rho^2$ amount of samples is already sufficient to distinguish sub-optimal actions from the optimal ones at the state-action pair. Then by clipping the bonus at such state-action pairs, `UCB-Clip` spends less unnecessary visitations to these pairs, and saves opportunities for visiting the state-actions that have not yet been visited sufficiently. In consequence, `UCB-Clip` accelerates benign TAE instances by exploiting the gap parameter. The above discussions are formally justified in the next section.

4 THEORETIC RESULTS

We turn to present our theoretical results. We will first show Theorem 1 that provides a sample-complexity upper bound for the proposed `UCB-Clip` algorithm, and Theorem 2 that provides a sample-complexity lower bound for the gap-TAE problem. Then we will discuss a novel statistical separation between RL vs. MAB based on these results.

4.1 Upper and Lower Bounds

Theorem 1 (An upper bound for `UCB-Clip`)**.** *Suppose that Algorithm 1 accepts a gap parameter ρ that satisfies (2.3), and runs for K episodes to collect a dataset $\mathcal{H}^K$. Let policy π be the output of Algorithm 2 for an oblivious input reward function $r \in \mathcal{R}$ that is independent of $\mathcal{H}^K$. Then with probability at least $1 - \delta$, the planning error is bounded by*

$$V_1^*(x_1) - V_1^\pi(x_1) \lesssim \frac{H^3SA}{\rho K} \cdot \log \frac{HSAK}{\delta} + \frac{H^4S^2A}{K} \cdot \log(HK) \cdot \log \frac{HSAK}{\delta}.$$

Theorem 2 (A lower bound for gap-TAE)**.** *Fix $S \ge 5, A \ge 2, H \ge 2 + \log_A S$. There exist positive constants $c_1, c_2, \rho_0, \delta_0$, such that for every $\rho \in (0, \rho_0)$, $\epsilon \in (0, \rho)$, $\delta \in (0, \delta_0)$, and for every (ϵ, δ)-PAC algorithm (see condition (2.2)) that runs for K episodes, there exists some gap-TAE instances with a gap parameter ρ that satisfies (2.3), such that*

$$\mathbb{E}[K] \ge c_1 \cdot \frac{H^2SA}{\rho\epsilon} \cdot \log \frac{c_2}{\delta},$$

where the expectation is taken with respect to the randomness of choosing the gap-TAE instance.

Remark 1. According to Theorem 1, `UCB-Clip` only requires $\propto \widetilde{\mathcal{O}}(1/\epsilon)$ number of episodes to solve TAE provided with a constant gap parameter, which improves the existing, pessimistic rates $\propto \widetilde{\mathcal{O}}(1/\epsilon^2)$ achieved by algorithms that focus on worst-case TAE instances (Zhang et al., 2020a; Wang et al., 2020b; Wu et al., 2021). Moreover, according to Theorem 2, this $\propto \widetilde{\mathcal{O}}(1/\epsilon)$ rate is nearly optimal upto logarithmic factors, which exhibits some fundamental limitations of the acceleration afforded by a constant gap parameter. A numerical simulation for the acceleration phenomenon is provided in Appendix A.

Remark 2. If $\rho \lesssim 1/(HS)$, the error upper bound in Theorem 1 is simplified to $\widetilde{\mathcal{O}}(H^3SA/(\rho K))$, i.e., `UCB-Clip` needs at most $K = \widetilde{\mathcal{O}}(H^3SA/\rho\epsilon)$ episodes to be (ϵ, δ)-PAC. In this regime, Theorem 2 suggests that `UCB-Clip` achieves a nearly optimal rate in terms of S, A, ρ and ϵ (or K) ignoring logarithmic factors.

Still, the dependence of H is improvable, which we leave as a future work.

Remark 3. As explained before, our algorithm can be applied to other unsupervised RL settings as well, e.g., reward-free exploration (Jin et al., 2020) with a gap parameter ρ that satisfies (2.3). In this case, the upper bound in Theorem 1 needs to be revised to

$$\widetilde{\mathcal{O}}\left(\frac{H^3S^2A}{\rho K}+\frac{H^4S^3A}{K}\right).$$

This is obtained by a standard converging and union bound argument on the set of all possible value functions, and from where the extra S factor stem[6][7]. Comparing to the worst-case optimal rate $\propto \mathcal{O}(1/\sqrt{K})$ for RFE (Jin et al., 2018; Zhang et al., 2020b), we again achieve a quadratic saving in terms of K for benign RFE instances with a constant ρ.

Proof Sketch of Theorem 1. We first look at the planning phase (Algorithm 2). Since the reward induces a sub-optimality gap at least ρ, with some computations we can obtain the following error estimation of the planning error per episode: for every k,

$$V_1^*(x_1)-V_1^{\pi^k}(x_1)\lesssim \mathbb{E}_{\pi^k,\mathbb{P}}\sum_{h=1}^{H}c^k(x_h,a_h), \qquad (4.1)$$

where π^k is the planning policy at the k-th episode and $c^k(x,a)$ is the *clipped* exploration bonus at the k-th episode. The right hand side of (4.1) can be further improved to have a uniform upper bound H per decision step. Note that the right hand side of (4.1) is the expected cumulative bonus over the trajectory induced by policy π^k and the *true* transition $\mathbb{P}$, and that the bonus contains lower order terms that control the error of an inaccurately estimated probability transition (Algorithm 1, line 5). Therefore, upto some constant factors, it can be bounded by the expected cumulative bonus over the trajectory induced by policy π^k and the *empirical* transition $\widehat{\mathbb{P}}^k$. The above analysis reflects (4.2). Moreover, (4.3) holds naturally according to Algorithm 1, since $\overline{V}_1^k(x_1)$ maximizes the cumulative bonus over the empirical transition by dynamic programming.

$$V_1^*(x_1)-V_1^{\pi^k}(x_1)\lesssim \mathbb{E}_{\pi^k,\widehat{\mathbb{P}}^k}\sum_{h=1}^{H}H\wedge c^k(x_h,a_h) \qquad (4.2)$$

$$\le \overline{V}_1^k(x_1). \qquad (4.3)$$

Finally, a standard regret analysis for the exploration phase shows that the total exploration values in the exploration phase is logarithmic, thanks to the clipped exploration bonus. Thus the total planning error is also logarithmic. The presented error upper bound is established by averaging over the K episodes. Some of proving techniques are motivated by (Simchowitz and Jamieson, 2019; Wu et al., 2021). Our key novelty is to carefully incorporate the clip operator in the bonus function and use that to build a connection between the planning and exploration phases. See Appendix B for more details.

Proof Sketch of Theorem 2. A hard instance that witnesses the lower bound is shown in Figure 1. The hard instance is motivated by (Mannor and Tsitsiklis, 2004; Dann and Brunskill, 2015). One can verify that this instance has a minimum sub-optimality gap $\rho/2$. Indeed, the only states that have sub-optimal actions are the *left orange states* in Type I model or Type II model. For the left orange state in Type I model, the optimal action has value $H/2+\rho/2$, but the sub-optimal ones have value $H/2$, where the gap is $\rho/2$. Similarly we can verify the gap in Type II model is ρ. Moreover, in order to be $\epsilon/3$-correct, the agent needs to identify the optimal action at the left orange states, otherwise it takes a sub-optimal action and incurs a value error at least $\epsilon/\rho\cdot\rho/2=\epsilon/2$. Now the problem is reduced to identify the best action at the left orange states. Let us ignore H, S, A factors and focus on ρ and ϵ. Then at the left orange states, the probability gap between the best action and the second best action is $\propto\Theta(\rho)$, thus $\propto\Omega(1/\rho^2)$ samples are needed to identify the optimal action. On the other hand, in each episode there is only ϵ/ρ chance to visit the left orange states, thus $\propto \Omega(1/(\rho\epsilon))$ episodes are needed to provide $\propto \Omega(1/\rho^2)$ samples at the left orange sates. This justifies the $\propto\Omega(1/(\rho\epsilon))$ rate in the lower bound. A complete proof is deferred to Appendix C.

4.2 Comparison with Multi-Armed Bandit

Theorems 1 and 2 show that for unsupervised RL problems, even when the instance is benign and has a constant gap parameter, a $\propto\Omega(1/\epsilon)$ sample complexity must be paid. This establishes a stark contrast to the gap-dependent unsupervised exploration problems on a *multi-armed bandit* (MAB). In particular, for an MAB with A arms and a minimum sub-optimality gap ρ (that is the gap between the expected rewards of the best action and the second-best action), a rather simple *uniform exploration* strategy, with $T=\mathcal{O}\big(\frac{A}{\rho^2}\log\frac{A}{\delta}\big)\propto\widetilde{\mathcal{O}}(1)$ pulls of the arms, is (ϵ,δ)-correct for identifying the best arm (see, e.g.,

[6]For more details, we refer the reader to event G_1 defiend in Appendix B. To apply our result in reward-free exploration, we need G_1 holds for every optimal value function V_h^*, which can be guaranteed by covering and union bound argument over all possible value functions, $V_h\in[0,H]^S$.

[7]Due to a similar reasoning, when applied to the task-agnostic setting with N selected planning tasks (Zhang et al., 2020a), the $\log(HSAK/\delta)$ factor in our bound needs to be modified to $\log(NHSAK/\delta)$.

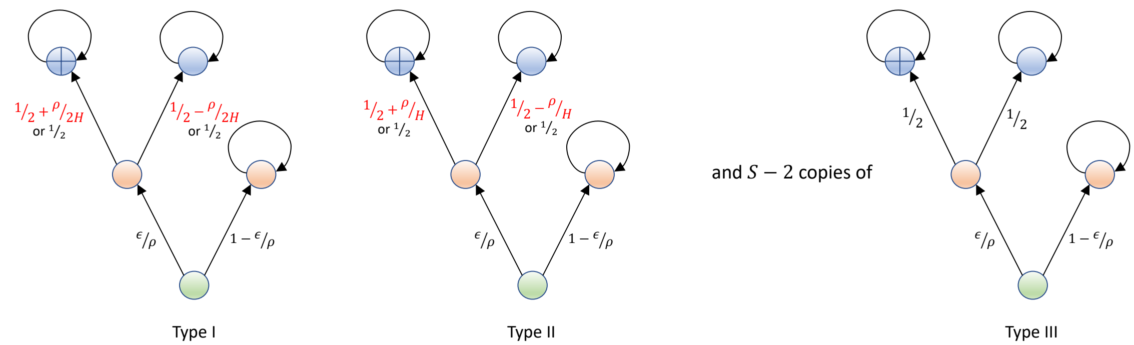

Figure 1: A hard-to-learn MDP example. States are denoted by circles. A state-determined reward function takes value 1 at the state denoted by a circle with a plus sign and takes value zero otherwise. The plot only shows the structure of the last three layers of the MDP, which consists of a Type I model, a Type II model, and $S-2$ Type III model. These models are connected by a $(\log_A S)$-layer tree with A-branches in each layer, and with deterministic and known transition. In the Type III model, from the green state and for all actions, it transits to the left orange state with probability ϵ/ρ, or to the self-absorbing right orange state with probability $1-\epsilon/\rho$. Then from the left orange state and for all actions, it transits to the two blue states evenly. The Type I (II) model is only different from the Type III model at the left orange state: in Type I (II) model, there exists and only one action such that it transits to the left blue state with probability $1/2+\rho/(2H)$ (with probability $1/2+\rho/H$), and to the right blue state otherwise. In other words, in the left orange states, an optimal action exists in the Type II model, a second-to-the-best action exists in the Type I model, and all other actions are equivalent and are sub-optimal.

Theorem 33.1 in Lattimore and Szepesvári (2020), or Appendix D). These observations from gap-dependent unsupervised exploration establish an interesting statistical separation between general RL (corresponds to MDP with $H \geq 2$) and MAB (corresponds to MDP with $H = 1$), with a sample complexity comparison $\propto \widetilde{\mathcal{O}}(1/\epsilon)$ vs. $\propto \widetilde{\mathcal{O}}(1)$. This separation suggests that *unsupervised RL is significantly more challenging than unsupervised MAB* even after normalizing the H factor. The conclusion goes against an emerging wisdom from the supervised RL theory, that RL is nearly as easy as MAB in the supervised setting, given that H factor is normalized (Jiang and Agarwal, 2018; Wang et al., 2020a; Zhang et al., 2020c).

Let us take a deeper look at where RL is harder than MAB. The hard instance in Figure 1 clearly illustrates the issue: *there could exist some important states in MDP that cannot be ignored, but are hard to reach in the same time*, e.g., ignoring the left orange states in Figure 1 would result in a $\Theta(\epsilon)$ error, but there is only $\Theta(\epsilon)$ chance to reach these states per episode, thus in order to reach the left orange states for at least constant times, an algorithm needs at least $\propto \Omega(1/\epsilon)$ number of trajectories.

To further verify our understanding, let us consider an MDP with a sampling simulator (Sidford et al., 2018a,b; Wang, 2017; Azar et al., 2013). The sampling simulator allows us to draw samples at any state-action pair, thus exempts the "hard-to-reach" states. The following theorem shows that for MDP with a sampling simulator, the gap-TAE problem can also be solved with $\propto \widetilde{\mathcal{O}}(1)$ samples, as in the case of MAB. A proof is included in Appendix D.

Theorem 3 (MDP with a sampling simulator)**.** *Suppose there is a sampling simulator for the MDP considered in the gap-TAE problem. Consider exploration with the uniformly sampling strategy, and planning with the dynamic programming method with the obtained empirical probability. If T samples are drawn, where*

$$T \geq \frac{2H^4SA}{\rho^2} \cdot \log \frac{2HSA}{\delta},$$

then with probability at least $1-\delta$, the obtained policy is optimal ($\epsilon = 0$).

5 ADDITIONAL DISCUSSIONS

The Gap Parameter. The gap parameter ρ is a hyperparameter for Algorithm 1. So long as the inputted gap parameter ρ is a reasonably large constant and satisfies condition (2.3), our algorithm achieves a fast, $\propto \widetilde{\mathcal{O}}(1/\epsilon)$ rate for TAE according to Theorem 1. If the inputted gap parameter ρ violates condition (2.3), Theorem 1 no longer directly holds; but one can easily show that

$$V_1^*(x_1) - V_1^\pi(x_1) = \mathbb{E}_{\pi,\mathbb{P}} \sum_{h=1}^{H} \big(V_h^*(x_h) - Q_h^*(x_h,a_h)\big)$$

$$\leq \mathbb{E}_{\pi,\mathbb{P}} \sum_{h=1}^{H} \texttt{clip}_\rho \left[V_h^*(x_h) - Q_h^*(x_h,a_h)\right] + H\rho \quad (5.1)$$

$$\leq \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\rho K} + \frac{H^4S^2A}{K} + H\rho\Big), \quad (5.2)$$

where (5.1) is by the definition of clip operator, and (5.2) is by applying Theorem 1 to the first term in (5.1) (that equals to the error on an MDP with a sub-optimality gap ρ). In this way, if the TAE instance is in fact "hard" in that condition (2.3) cannot be met for any reasonably large ρ, one can still choose a small $\rho \approx \epsilon/H$ and run UCB-Clip for $K = \widetilde{\mathcal{O}}(H^2SA/\rho^2 + H^3S^2A/\rho)$ episodes to obtain a policy that is $(H\rho, \delta)$-PAC. Note that this $\propto \widetilde{\mathcal{O}}(1/\epsilon^2)$ rate is minimax optimal ignoring logarithmic factors (and the H^2 factor) (Jin et al., 2020; Dann and Brunskill, 2015). An open question is: is there a sample-efficient algorithm for gap-TAE that does not need an inputted gap parameter?

The H Dependence. In the regime that $\rho \lesssim 1/(HS)$, our upper bound can potentially be improved for an H factor, comparing with the provided lower bound. Technically, this is because Algorithm 1 utilizes a Hoeffding-type bonus, which is known to be less tight compared with a Bernstein-type bonus (Azar et al., 2017). Hoeffding-type bonus has the benefits of being *reward-independent*, which allows us to construct a reward-independent, hence unsupervised, exploration policy that minimizes an upper bound of the per-episode, reward-dependent planning error. In contrast, Bernstein-type bonus is reward-dependent as it requires an (good) estimation to the value function which relies on reward signals. Since the reward signal is not available in the exploration phase, we do not see a clear way to adopt Bernstein-type bonus in our problem. We leave the issue of further tightening the H factor as a future work.

Removing the Gap-Independent Term. The bound presented in Theorem 1 has a gap-independent term, $\widetilde{\mathcal{O}}(H^4S^2A/K)$, which is still linear in $\widetilde{\mathcal{O}}(1/K)$ but is quadratic in S. This S^2-dependence appears in the sample-complexity lower-order term of all the model-based algorithms that we are aware of, e.g., (Azar et al., 2017; Zanette and Brunskill, 2019; Simchowitz and Jamieson, 2019), and could potentially be mitigated by model-free algorithms (Jin et al., 2018; Yang et al., 2020).

Visiting Ratio Based Approaches. In the context of reward-free exploration, Zhang et al. (2020b) extend a visiting ratio based approach that is initially proposed by Jin et al. (2020), and achieves a nearly minimax optimal sample complexity for reward-free exploration. The visiting ratio based method has the advantage of supporting plug-in planners (i.e., an approximate MDP solver given trasition matrix and reward) (Jin et al., 2020). Such approach can be adapted to obtain a $\widetilde{\mathcal{O}}(1/\epsilon)$ fast rate for gap-TAE as well; however, it involves a S^2-dependence on the obtained bound, which is sub-optimal in the context of task-agnostic exploration (see Zhang et al. (2020a); Wu et al. (2021) and Theorem 2).

A Refined Gap Dependence? For supervised RL, the finest gap-dependent sample complexity bound can accurately characterize the role of each state-action gap (Simchowitz and Jamieson, 2019; Xu et al., 2021). In contrast, for unsupervised RL, our derived gap-dependent bound is stated as a function of a gap parameter that reflects only the *minimum* nonzero state-action gap. We argue that obtaining a refined gap dependence is generally not possible in the unsupervised setting. In particular, let us consider a gap-TAE problem with two candidate rewards, where one of them is specified as in our lower bound construction (see Theorem 2 and Figure 1) and the other induces better state-action gaps; in the planning phase, one of the two rewards is selected uniformly at random. Then any algorithm for this problem instance will need to solve the hard instance in Theorem 2 with probability 1/2, therefore it must suffers from the "worst" gap parameter instead of the better state-action gaps.

6 RELATED WORKS

Supervised RL with Gap Dependence. In the literature of supervised RL, the state-action sub-optimality gap has been long embraced to characterize instance-dependent theoretic guarantees. Ortner and Auer (2007); Tewari and Bartlett (2007); Ok et al. (2018) study gap-dependent bounds in the asymptotic sense, and Jaksch et al. (2010) establish a finite time, gap-dependent bound. More recently, in the setting of finite-horizon MDP and for a broad class of UCB-type, model based algorithms, Simchowitz and Jamieson (2019) provide a finite, gap-dependent regret bound that comprehensively interpolates the minimax, gap-independent regret bound and a logarithmic, gap-dependent regret bound. Similar results are further built for model-free algorithms (Yang et al., 2020), linear MDP (He et al., 2020), and MDP with corruptions (Lykouris et al., 2019). More recently, Dann et al. (2021) improves the gap-dependent bounds in Simchowitz and Jamieson (2019) by considering the reachability of state-action pairs. In addition to this line, Jonsson et al. (2020) study the gap-dependent bound for a Monte-Carlo tree search algorithm, Xu et al. (2021) establish a fine-grained gap-dependent bound through a non-UCB type algorithm, and Al Marjani et al. (2021); Wagenmaker et al. (2021) investigate gap-dependent bound for the best-policy-identification problem[8]. However, all these results explores the environment under the guidance of an observable reward signal, thus are not applicable to the unsupervised exploration problems studied in this paper.

Unsupervised RL in the Worst Case. The arguably most typical unsupervised exploration problem is the reward-free exploration problem formalized by Jin et al. (2020), where the agent collects samples in the unsupervised fashion in order to be able to plan nearly optimally for arbitrary rewards. Prior to Jin et al. (2020), Hazan et al. (2018); Brafman and Tennenholtz (2002); Du et al. (2019) also study exploratory policies with certain covering properties; and following Jin et al. (2020), the sample complexity of reward-free exploration is further improved to nearly minimax optimal by Kaufmann et al. (2020); Ménard et al. (2020); Zhang et al. (2020b). Besides reward-free exploration problems, Zhang et al. (2020a); Wang et al. (2020b) introduce and study task-agnostic exploration problems where the planning reward is fixed but unknown during exploration, and Wu et al. (2021) study preference-free exploration problems in the context of multi-objective RL. Nonetheless, the above considerations of unsupervised exploration problems and their algorithms take no advantage of a gap parameter, and the obtained theoretic results are pessimistic and restricted by the worst cases.

Lenient Regret. Our results can also be understood from the viewpoint of *lenient error* (Merlis and Mannor, 2020), which is first introduced in the context of MAB and aims to capture a regret that tolerances small errors per decision step. This notion naturally generalizes to RL as follows:

$$\texttt{LenientError}(\pi) := \mathbb{E}_{\pi,\mathbb{P}} \sum_{h=1}^{H} \texttt{clip}_{\rho}\left[V_h^*(x_h) - Q_h^*(x_h, a_h)\right]. \tag{6.1}$$

Clearly, solving a gap-TAE problem with a gap parameter ρ under the error measured by $V_1^*(x_1) - V_1^{\pi}(x_1)$ is equivalent to solving a TAE problem under a lenient error measured by (6.1). From this viewpoint, an interesting future direction to extend our results to general gap function (see Definition 1 in Merlis and Mannor (2020)) beyond the $\texttt{clip}_{\rho}[\cdot]$ studied in this work.

[8]Similarly to our lower bound construction, Wagenmaker et al. (2021) also exploit a transition probability to show their gap-dependent lower bound. We emphasize that our works are concurrent and the similar idea is used for solving different problems.

7 CONCLUSION

In this paper we study sample-efficient algorithms for gap-dependent unsupervised exploration problems in RL. When the targeted planning tasks have a constant gap parameter, the proposed algorithm achieves a gap-dependent sample complexity upper bound that significantly improves the existing pessimistic bounds. Moreover, an information-theoretic lower bound is provided to justify the tightness of the obtained upper bound. These results establish an interesting statistical separation between RL and MAB (or RL with a simulator) in terms of gap-dependent unsupervised exploration problems.

Acknowledgements

We would like to thank the anonymous reviewers and area chairs for their helpful comments. This project was supported in part by DARPA-PA-20-02-11-ShELL-FP-007

References

Al Marjani, A., Garivier, A., and Proutiere, A. (2021). Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34.

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

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

Brafman, R. I. and Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231.

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

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*.

Dann, C., Marinov, T. V., Mohri, M., and Zimmert, J. (2021). Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. (2019). Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR.

Finn, C. and Levine, S. (2017). Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE.

Hazan, E., Kakade, S. M., Singh, K., and Van Soest, A. (2018). Provably efficient maximum entropy exploration. *arXiv preprint arXiv:1812.02690*.

He, J., Zhou, D., and Gu, Q. (2020). Logarithmic regret for reinforcement learning with linear function approximation. *arXiv preprint arXiv:2011.11566*.

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

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

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

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020). Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*.

Jonsson, A., Kaufmann, E., Ménard, P., Domingues, O. D., Leurent, E., and Valko, M. (2020). Planning in markov decision processes with gap-dependent sample complexity. *arXiv preprint arXiv:2006.05879*.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. (2020). Adaptive reward-free exploration. *arXiv preprint arXiv:2006.06294*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. (2019). Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*.

Mannor, S. and Tsitsiklis, J. N. (2004). The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648.

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

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. (2020). Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442*.

Merlis, N. and Mannor, S. (2020). Lenient regret for multi-armed bandits. *arXiv preprint arXiv:2008.03959*.

Ok, J., Proutiere, A., and Tranos, D. (2018). Exploration in structured reinforcement learning. *arXiv preprint arXiv:1806.00775*.

Ortner, P. and Auer, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, 19:49.

Riedmiller, M., Hafner, R., Lampe, T., Neunert, M., Degrave, J., Wiele, T., Mnih, V., Heess, N., and Springenberg, J. T. (2018). Learning by playing solving sparse reward tasks from scratch. In *International Conference on Machine Learning*, pages 4344–4353. PMLR.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). Universal value function approximators. In *International conference on machine learning*, pages 1312–1320.

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

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

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

Tewari, A. and Bartlett, P. L. (2007). Optimistic linear programming gives logarithmic regret for irreducible mdps. In *NIPS*, pages 1505–1512. Citeseer.

Wagenmaker, A., Simchowitz, M., and Jamieson, K. (2021). Beyond no regret: Instance-dependent pac reinforcement learning. *arXiv preprint arXiv:2108.02717*.

Wang, M. (2017). Randomized linear programming solves the discounted markov decision problem in nearly-linear (sometimes sublinear) running time. *arXiv preprint arXiv:1704.01869*.

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

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. (2020b). On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*.

Wu, J., Braverman, V., and Yang, L. F. (2021). Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *The 35th Conference on Neural Information Processing Systems*.

Xie, A., Ebert, F., Levine, S., and Finn, C. (2019). Improvisation through physical understanding: Using novel objects as tools with visual foresight. *arXiv preprint arXiv:1904.05538*.

Xie, A., Singh, A., Levine, S., and Finn, C. (2018). Few-shot goal inference for visuomotor learning and planning. In *Conference on Robot Learning*, pages 40–52. PMLR.

Xu, H., Ma, T., and Du, S. S. (2021). Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*.

Yang, K., Yang, L. F., and Du, S. S. (2020). *q*-learning with logarithmic regret. *arXiv preprint arXiv:2006.09118*.

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

Zhang, X., Singla, A., et al. (2020a). Task-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2006.09497*.

Zhang, Z., Du, S. S., and Ji, X. (2020b). Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*.

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

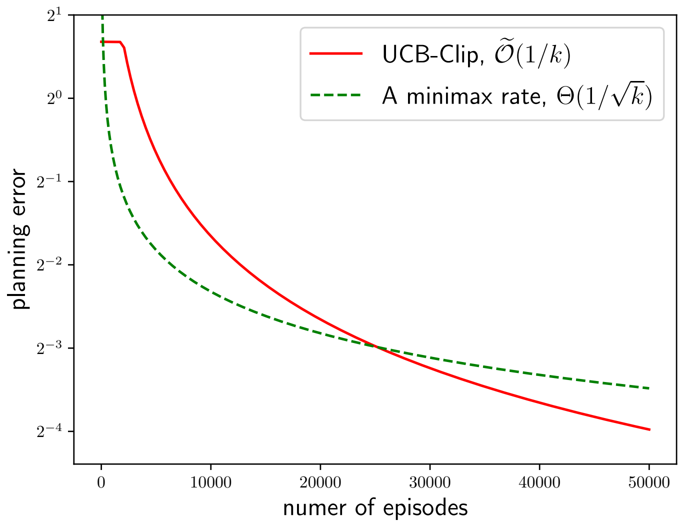

Figure 2: An illustration of the fast task-agnostic exploration achieved by UCB-Clip. In the experiment we simulate a random MDP with $H = 5, S = 10, A = 10$ and $\rho = 0.4$, and run UCB-Clip for $K = 50,000$ episodes. The red curve shows the planning error of UCB-Clip, and the green dotted curve shows a minimax error rate. The plot shows that UCB-Clip achieves an improved rate for gap-dependent task-agnostic exploration.

A NUMERICAL SIMULATIONS

Figure 2 illustrates the fast rate achieved by UCB-Clip for task-agnostic exploration on a benign MDP with constant minimum sub-optimality gap. The curve for UCB-Clip indicates the planning error of UCB-Clip when running on a random MDP with $H = 5, S = 10, A = 10, \rho = 0.4$ and $K = 50,000$. By comparing with the minimax rate, we observe that UCB-Clip solves task-agnostic exploration with a faster rate, when the task has a constant minimum sub-optimality gap.

In this experiment, the random MDP can be regarded as a Grid World. It is generated as follows. The reward is 1 at a state x^* and is 0 otherwise. The initial state is fixed $x_0 \neq x^*$ and cannot be revisited in the rest steps in one game. x^* can only be reached by taking an optimal action a^* at two states, x_0 and x^*, where $\mathbb{P}\{x^*|x_0, a^*\} = \rho$ and $\mathbb{P}\{x^*|x^*, a^*\} = 1$. Except for these constraints, the transition kernel is filled with random numbers uniformly distributed in $(0, 1)$ and is then properly normalized. This MDP has a gap parameter ρ. In experiment the exploration bonus in Algorithm 1 is simplified: only the first two terms are considered (which are leading terms) and absolute constants and logarithmic factors are set to be 1.

B PROOF OF THE UPPER BOUND (THEOREM 1)

Our proof is inspired by (Simchowitz and Jamieson, 2019) and (Wu et al., 2021).

Preliminaries. Let π^k be the planning policy at the k-th episode, i.e., a greedy policy that maximizes $Q_h^k(x, a)$. Let $\bar{\pi}^k$ be the exploration policy at the k-th episode, i.e., a greedy policy given that maximizes $\overline{Q}_h^k(x, a)$. Let $w_h^k(x, a) := \mathbb{P}\left\{(x_h, a_h) = (x, a) \mid \bar{\pi}^k, \mathbb{P}\right\}$ and $w^k(x, a) := \sum_h w_h^k(x, a)$. In the following, if not otherwise noted, we define $\iota := \log \frac{2HS^2AK}{\delta}$.

Consider the following good events

$$G_1 := \left\{\forall x, a, h, k, \ \left|(\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*(x, a)\right| \leq \sqrt{\frac{H^2}{2N^k(x, a)} \log \frac{2HSAK}{\delta}}\right\}, \tag{G_1}$$

$$G_2 := \left\{ \forall x, a, y, k,\ \left| \widehat{\mathbb{P}}^k(y \mid x, a) - \mathbb{P}(y \mid x, a) \right| \le \sqrt{\frac{2\mathbb{P}(y \mid x, a)}{N^k(x, a)} \log \frac{2S^2AK}{\delta}} + \frac{2}{3N^k(x, a)} \log \frac{2S^2AK}{\delta} \right\}, \tag{G_2}$$

$$G_3 := \left\{ \forall x, a, y, k,\ \left| \widehat{\mathbb{P}}^k(y \mid x, a) - \mathbb{P}(y \mid x, a) \right| \le \sqrt{\frac{2\widehat{\mathbb{P}}^k(y \mid x, a)}{N^k(x, a)} \log \frac{2S^2AK}{\delta}} + \frac{7}{3N^k(x, a)} \log \frac{2S^2AK}{\delta} \right\}, \tag{G_3}$$

$$G_4 := \left\{ \forall x, a, k,\ N^k(x, a) \ge \frac{1}{2} \sum_{j<k} w^j(x, a) - H \log \frac{HSA}{\delta} \right\}. \tag{G_4}$$

Lemma 1 (The probability of good events)**.** $\mathbb{P}\{G_1 \cap G_2 \cap G_3 \cap G_4\} \ge 1 - 4\delta$.

Proof. By Hoeffding's inequality and a union bound, we have that $\mathbb{P}\{G_1\} \ge 1 - \delta$.

By Bernstein's inequality, a union and that $1 - \mathbb{P}(y \mid x, a) \le 1$, we have that $\mathbb{P}\{G_2\} \ge 1 - \delta$.

By empirical Bernstein's inequality (Maurer and Pontil, 2009), a union bound and that $1 - \widehat{\mathbb{P}}^k(y \mid x, a) \le 1$, we have that $\mathbb{P}\{G_3\} \ge 1 - \delta$.

According to Lemma F.4 by (Dann et al., 2017) and a union bound, we have that $\mathbb{P}\{G_4\} \ge 1 - \delta$.

Finally, a union bound over the four events proves the claim. □

Planning Phase. Recall the planning bonus is set to be

$$b^k(x, a) := \sqrt{\frac{H^2\iota}{2N^k(x, a)}}. \tag{B.1}$$

Lemma 2 (Optimistic planning)**.** *If G_1 holds, then $Q^k_h(x, a) \ge Q^*_h(x, a)$ for every k, h, x, a.*

Proof. We prove it by induction. Clearly the hypothesis holds for $H + 1$; now suppose that $Q^k_{h+1}(x, a) \ge Q^*_{h+1}(x, a)$, and consider h. From Algorithm 2, we see that

$$Q^k_h(x, a) := H \wedge \left(r_h(x, a) + b^k(x, a) + \widehat{\mathbb{P}}^k V^k_{h+1}(x, a) \right). \tag{B.2}$$

If $Q^k_h(x, a) = H$, then $Q^k_h(x, a) = H \ge Q^*_h(x, a)$; otherwise, we have that

$$\begin{aligned} Q^k_h(x, a) - Q^*_h(x, a) &= r_h(x, a) + b^k(x, a) + \widehat{\mathbb{P}}^k V^k_{h+1}(x, a) - r_h(x, a) - \mathbb{P}V^*_{h+1}(x, a) \\ &= b^k(x, a) + \widehat{\mathbb{P}}^k V^k_{h+1}(x, a) - \mathbb{P}V^*_{h+1}(x, a) \\ &\ge b^k(x, a) + (\widehat{\mathbb{P}}^k - \mathbb{P}) V^*_{h+1}(x, a) \qquad (\text{since } Q^k_{h+1}(x, a) \ge Q^*_{h+1}(x, a)) \\ &\ge 0 \qquad (\text{since } G_1 \text{ holds}). \end{aligned}$$

These complete our induction. □

Let us denote the *optimistic surplus* (Simchowitz and Jamieson, 2019) as

$$E^k_h(x, a) := Q^k_h(x, a) - \left(r_h(x, a) + \mathbb{P}V^k_{h+1}(x, a) \right). \tag{B.3}$$

Lemma 3 (Optimistic surplus bound)**.** *If G_1, G_2 and G_3 hold, then for every k, h, x, a,*

$$E^k_h(x, a) \le H \wedge \left(\sqrt{\frac{2H^2\iota}{N^k(x, a)}} + \frac{HS\iota}{N^k(x, a)} + \mathbb{E}_{\pi^k, \mathbb{P}} \sum_{t \ge h+1} \left(\frac{4e^2H^3\iota}{N^k(x_t, a_t)} + \frac{8e^2H^5S^2\iota^2}{\left(N^k(x_t, a_t)\right)^2} \right) \right).$$

Proof. By (B.3) and (B.2) we have $E_h^k(x,a) \le Q_h^k(x,a) \le H$. As for the second bound, note that

$$\begin{aligned}
E_h^k(x,a) &= Q_h^k(x,a) - \big(r_h(x,a) + \mathbb{P}V_{h+1}^k(x,a)\big) \qquad \text{(use (B.3))}\\
&\le r_h(x,a) + b^k(x,a) + \widehat{\mathbb{P}}^k V_{h+1}^k(x,a) - r_h(x,a) - \mathbb{P}V_{h+1}^k(x,a) \qquad \text{(use (B.2))}\\
&= b^k(x,a) + (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*(x,a) + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_{h+1}^k - V_{h+1}^*)(x,a)\\
&\le 2b^k(x,a) + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_{h+1}^k - V_{h+1}^*)(x,a) \qquad \text{(use (B.1) and that } G_1 \text{ holds)}\\
&\le \sqrt{\frac{2H^2\iota}{N^k(x,a)}} + \frac{HS\iota}{N^k(x,a)} + \mathbb{P}\left(V_{h+1}^k - V_{h+1}^*\right)^2(x,a). \quad \text{(use Lemma 4)}
\end{aligned} \tag{B.4}$$

We next bound $V_h^k(x) - V_h^*(x)$ by

$$\begin{aligned}
&V_h^k(x) - V_h^*(x) \le Q_h^k(x,a) - Q_h^*(x,a) \qquad (\text{set } a = \pi^k(x))\\
&\le b^k(x,a) + \widehat{\mathbb{P}}^k V_{h+1}^k(x,a) - \mathbb{P}V_{h+1}^*(x,a) \qquad \text{(use (B.2))}\\
&= \Big(b^k + \mathbb{P}\left(V_{h+1}^k - V_{h+1}^*\right) + (\widehat{\mathbb{P}}^k - \mathbb{P})\left(V_{h+1}^k - V_{h+1}^*\right) + (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*\Big)(x,a)\\
&\le \Big(2b^k + \mathbb{P}\left(V_{h+1}^k - V_{h+1}^*\right) + (\widehat{\mathbb{P}}^k - \mathbb{P})\left(V_{h+1}^k - V_{h+1}^*\right)\Big)(x,a) \quad \text{(use (B.1) and } G_1)\\
&\le \sqrt{\frac{2H^2\iota}{N^k(x,a)}} + \left(1 + \frac{1}{H}\right)\mathbb{P}\left(V_{h+1}^k - V_{h+1}^*\right)(x,a) + \frac{2H^2S\iota}{N^k(x,a)}. \qquad \text{(use Lemma 4)}
\end{aligned}$$

Solving the recursion we obtain

$$V_h^k(x) - V_h^*(x) \le e \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t\ge h}\left(\sqrt{\frac{2H^2\iota}{N^k(x_t,a_t)}} + \frac{2H^2S\iota}{N^k(x_t,a_t)}\right),$$

where $x_h = x$. This implies that

$$\begin{aligned}
\left(V_h^k(x) - V_h^*(x)\right)^2 &\le \left(\mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t\ge h}\left(\sqrt{\frac{2e^2H^2\iota}{N^k(x_t,a_t)}} + \frac{2eH^2S\iota}{N^k(x_t,a_t)}\right)\right)^2\\
&\le \mathbb{E}_{\pi^k,\mathbb{P}}\left(\sum_{t\ge h}\sqrt{\frac{2e^2H^2\iota}{N^k(x_t,a_t)}} + \frac{2eH^2S\iota}{N^k(x_t,a_t)}\right)^2 \qquad ((\cdot)^2 \text{ is convex})\\
&\le H\cdot\mathbb{E}_{\pi^k,\mathbb{P}}\sum_{t\ge h}\left(\sqrt{\frac{2e^2H^2\iota}{N^k(x_t,a_t)}} + \frac{2eH^2S\iota}{N^k(x_t,a_t)}\right)^2 \quad \text{(Cauchy–Schwarz)}\\
&\le \mathbb{E}_{\pi^k,\mathbb{P}}\sum_{t\ge h}\left(\frac{4e^2H^3\iota}{N^k(x_t,a_t)} + \frac{8e^2H^5S^2\iota^2}{\left(N^k(x_t,a_t)\right)^2}\right),
\end{aligned}$$

inserting which to (B.4) we have that

$$E_h^k(x,a) \le \sqrt{\frac{2H^2\iota}{N^k(x,a)}} + \frac{HS\iota}{N^k(x,a)} + \mathbb{E}_{\pi^k,\mathbb{P}}\sum_{t\ge h+1}\left(\frac{4e^2H^3\iota}{N^k(x_t,a_t)} + \frac{8e^2H^5S^2\iota^2}{\left(N^k(x_t,a_t)\right)^2}\right),$$

where $(x_h, a_h) = (x,a)$. The two upper bounds on $E_h^k(x,a)$ together complete the proof. ◻

Lemma 4 (Bounds for the lower order term)**.** *If G_2 and G_3 hold, we have that for every V_1, V_2 such that $0 \le V_1(x) \le V_2(x) \le M$ and for every k, x, a, the following inequalities hold:*

$$\begin{aligned}
\left|(\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)(x,a)\right| &\le \mathbb{P}\left(V_2 - V_1\right)^2(x,a) + \frac{MS\iota}{N^k(x,a)};\\
\left|(\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)(x,a)\right| &\le \frac{1}{H}\mathbb{P}\left(V_2 - V_1\right)(x,a) + \frac{2MHS\iota}{N^k(x,a)};\\
\left|(\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)(x,a)\right| &\le \frac{1}{H}\widehat{\mathbb{P}}^k\left(V_2 - V_1\right)(x,a) + \frac{3MHS\iota}{N^k(x,a)}.
\end{aligned}$$

Proof. For simplicity let us denote $p(y) := \mathbb{P}(y \mid x,a)$ and $\hat{p}(y) := \widehat{\mathbb{P}}^k(y \mid x,a)$. For the first inequality,

$$
\begin{aligned}
&\left|(\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)(x,a)\right| \le \sum_{y\in\mathcal{S}} \left|\hat{p}^k(y) - p(y)\right| (V_2(y) - V_1(y)) \\
&\le \sum_{y\in\mathcal{S}} \left(\sqrt{\frac{2p(y)\iota}{N^k(x,a)}} + \frac{2\iota}{3N^k(x,a)}\right)(V_2(y) - V_1(y)) \qquad \text{(since } G_2 \text{ holds)} \\
&\le \sum_{y\in\mathcal{S}} \sqrt{\frac{2\iota}{N^k(x,a)}} \cdot \sqrt{p(y)\left(V_2(y) - V_1(y)\right)^2} + \frac{2MS\iota}{3N^k(x,a)} \\
&\le \sum_{y\in\mathcal{S}} \left(\frac{\iota}{N^k(x,a)} + p(y)\left(V_2(y) - V_1(y)\right)^2\right) + \frac{2MS\iota}{3N^k(x,a)} \qquad \text{(use } \sqrt{ab} \le a+b) \\
&\le \mathbb{P}\left(V_2 - V_1\right)^2(x,a) + \frac{MS\iota}{N^k(x,a)}.
\end{aligned}
$$

For the second inequality,

$$
\begin{aligned}
&\left|(\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)(x,a)\right| \le \sum_{y\in\mathcal{S}} \left|\hat{p}^k(y) - p(y)\right| (V_2(y) - V_1(y)) \\
&\le \sum_{y\in\mathcal{S}} \left(\sqrt{\frac{2p(y)\iota}{N^k(x,a)}} + \frac{2\iota}{3N^k(x,a)}\right)(V_2(y) - V_1(y)) \qquad \text{(since } G_2 \text{ holds)} \\
&\le \sum_{y\in\mathcal{S}} \left(\frac{p(y)}{H} + \frac{H\iota}{2N^k(x,a)} + \frac{2\iota}{3N^k(x,a)}\right)(V_2(y) - V_1(y)) \qquad \text{(use } \sqrt{ab} \le \tfrac{1}{2}(a+b)) \\
&\le \sum_{y\in\mathcal{S}} \frac{p(y)}{H}\left(V_2(y) - V_1(y)\right) + \frac{2MHS\iota}{N^k(x,a)} \\
&= \frac{1}{H}\mathbb{P}\left(V_2 - V_1\right)(x,a) + \frac{2MHS\iota}{N^k(x,a)}.
\end{aligned}
$$

The third inequality is proved in a same way as the second inequality, except that in the second step we use event G_3 rather than G_2. □

Lemma 5 (Half-clip trick)**.** *If G_1 holds, then for every k,*

$$
V_1^*(x_1) - V_1^{\pi^k}(x_1) \le 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \mathtt{clip}_{\frac{\rho}{2H}}[E_h^k(x_h,a_h)].
$$

Proof. Following (Simchowitz and Jamieson, 2019), let us consider

$$
\ddot{E}_h^k(x,a) := \mathtt{clip}_{\frac{\rho}{2H}}[E_h^k(x,a)] = E_h^k(x,a) \cdot \mathbb{1}\left[E_h^k(x,a) \ge \frac{\rho}{2H}\right] \ge 0 \tag{B.5}
$$

and

$$
\begin{cases}
\ddot{V}_h^k(x) := \ddot{Q}_h^k(x, \pi^k(x)), \\
\ddot{Q}_h^k(x,a) := r_h(x,a) + \ddot{E}_h^k(x,a) + \mathbb{P}\ddot{V}_{h+1}^k(x,a).
\end{cases}
$$

Notice that

$$
V_h^k(x) - V_h^{\pi^k}(x) = \mathbb{E}_{\pi^k.\mathbb{P}} \sum_{t\ge h} E_t^k(x_t,a_t), \tag{B.6}
$$

$$
\ddot{V}_h^k(x) - V_h^{\pi^k}(x) = \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t\ge h} \ddot{E}_t^k(x_t,a_t) \ge 0, \tag{B.7}
$$

then we immediately see that

$$
\ddot{V}_h^k(x) - V_h^{\pi^k}(x) \ge \mathbb{E}_{\pi^k.\mathbb{P}} \sum_{t\ge h} \left(E_t^k(x_t,a_t) - \frac{\rho}{2H}\right) = V_h^k(x) - V_h^{\pi^k}(x) - \frac{H-h+1}{H} \cdot \frac{\rho}{2}
$$

$$\geq V_h^k(x) - V_h^{\pi^k}(x) - \frac{\rho}{2}. \tag{B.8}$$

Given a sequence of a random trajectory $\{x_h, a_h\}_{h=1}^H$ induced by policy π^k and $\mathbb{P}$, let F_h be the event such that

$$F_h := \{a_h \notin \pi_h^*(x_h),\ \forall t < h,\ a_t \in \pi_t^*(x_t)\}.$$

Clearly $\{F_h\}_{h=1}^{H+1}$ are disjoint and form a partition for the sample space of the random trajectory induced by policy π^k. Then we have that

$$\begin{aligned}
V_1^*(x_1) - V_1^{\pi^k}(x_1) &= \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(V_1^*(x_1) - V_1^{\pi^k}(x_1) \right) + 0 \\
&= \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(V_h^*(x_h) - V_h^{\pi^k}(x_h) \right) \\
&= \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(\texttt{gap}_h(x_h, a_h) + Q_h^*(x_h, a_h) - V_h^{\pi^k}(x_h) \right),
\end{aligned}$$

where $\texttt{gap}_h(x_h, a_h) \geq \rho > 0$ under F_h (since $a_h \notin \pi_h^*(x_h)$). Similarly we have that

$$\begin{aligned}
\ddot{V}_1^k(x_1) - V_1^{\pi^k}(x_1) &= \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(\ddot{V}_1^k(x_1) - V_1^{\pi^k}(x_1) \right) + \mathbb{1}[F_{H+1}] \left(\ddot{V}_1^k(x_1) - V_1^{\pi^k}(x_1) \right) \\
&\geq \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(\ddot{V}_h^k(x_h) - V_h^{\pi^k}(x_h) \right) \qquad \text{(use (B.7) and (B.5))} \\
&\geq \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(V_h^k(x_h) - V_h^{\pi^k}(x_h) - \frac{\rho}{2} \right) \qquad \text{(use (B.8))} \\
&\geq \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(V_h^*(x_h) - V_h^{\pi^k}(x_h) - \frac{1}{2}\texttt{gap}_h(x_h, a_h) \right) \\
&\qquad \text{(use Lemma 2, and that } \texttt{gap}_h(x_h, a_h) \geq \rho \text{ under } F_h) \\
&= \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(\frac{1}{2}\texttt{gap}_h(x_h, a_h) + Q_h^*(x_h, a_h) - V_h^{\pi^k}(x_h) \right) \\
&\geq \frac{1}{2}\mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \mathbb{1}[F_h] \left(\texttt{gap}_h(x_h, a_h) + Q_h^*(x_h, a_h) - V_h^{\pi^k}(x_h) \right) \\
&\qquad \text{(note that } Q_h^*(x_h, a_h) - V_h^{\pi^k}(x_h) \geq 0) \\
&= \frac{1}{2} \left(V_1^*(x_1) - V_1^{\pi^k}(x_1) \right).
\end{aligned}$$

The above inequality plus (B.7) (B.5) completes the proof. ∎

Lemma 6. *If G_1, G_2 and G_3 hold, then*

$$\begin{aligned}
&V_1^*(x_1) - V_1^{\pi^k}(x_1) \leq \\
&\qquad \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \left(\texttt{clip}_{\frac{\rho}{2H}} \left[\sqrt{\frac{8H^2\iota}{N^k(x_h, a_h)}} \right] + \frac{120(H^3+S)H\iota}{N^k(x_h, a_h)} + \frac{240H^6S^2\iota^2}{(N^k(x_h, a_h))^2} \right).
\end{aligned}$$

Proof. We proceed the proof as follows:

$$V_1^*(x_1) - V_1^{\pi^k}(x_1) \leq 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^H \texttt{clip}_{\frac{\rho}{2H}}[E_h^k(x_h, a_h)] \qquad \text{(use Lemma 5)}$$

$$
\begin{aligned}
&\le 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \mathtt{clip}_{\frac{\rho}{2H}} \Bigg[\sqrt{\frac{2H^2\iota}{N^k(x,a)}} + \frac{HS\iota}{N^k(x,a)} \\
&\qquad + \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t\ge h+1} \left(\frac{4e^2H^3\iota}{N^k(x_t,a_t)} + \frac{8e^2H^5S^2\iota^2}{\left(N^k(x_t,a_t)\right)^2} \right) \Bigg] \qquad \text{(use Lemma 3)} \\
&\le 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \Bigg\{ \mathtt{clip}_{\frac{\rho}{4H}} \left[\sqrt{\frac{2H^2\iota}{N^k(x_h,a_h)}} \right] + \frac{2HS\iota}{N^k(x_h,a_h)} \\
&\qquad + \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t\ge h+1} \left(\frac{8e^2H^3\iota}{N^k(x_t,a_t)} + \frac{16e^2H^5S^2\iota^2}{\left(N^k(x_t,a_t)\right)^2} \right) \Bigg\} \qquad \text{(use Lemma 11)} \\
&\le 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \left\{ \mathtt{clip}_{\frac{\rho}{4H}} \left[\sqrt{\frac{2H^2\iota}{N^k(x_h,a_h)}} \right] + \frac{2HS\iota}{N^k(x_h,a_h)} \right\} \\
&\qquad + 2H \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t=1}^{H} \left(\frac{8e^2H^3\iota}{N^k(x_t,a_t)} + \frac{16e^2H^5S^2\iota^2}{\left(N^k(x_t,a_t)\right)^2} \right) \\
&= \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \left(\mathtt{clip}_{\frac{\rho}{2H}} \left[\sqrt{\frac{8H^2\iota}{N^k(x_h,a_h)}} \right] + \frac{4HS\iota}{N^k(x_h,a_h)} + \frac{16e^2H^4\iota}{N^k(x_h,a_h)} + \frac{32e^2H^6S^2\iota^2}{\left(N^k(x_h,a_h)\right)^2} \right) \\
&\le \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \left(\mathtt{clip}_{\frac{\rho}{2H}} \left[\sqrt{\frac{8H^2\iota}{N^k(x,a)}} \right] + \frac{120(H^3+S)H\iota}{N^k(x,a)} + \frac{240H^6S^2\iota^2}{\left(N^k(x,a)\right)^2} \right).
\end{aligned}
$$

□

Exploration Phase. Recall the exploration bonus in Algorithm 1 is defined as

$$
c^k(x,a) := \mathtt{clip}_{\frac{\rho}{2H}} \left[\sqrt{\frac{8H^2\iota}{N^k(x,a)}} \right] + \frac{120(H+S)H^3\iota}{N^k(x,a)} + \frac{240H^6S^2\iota^2}{\left(N^k(x,a)\right)^2}, \tag{B.9}
$$

and the exploration value function in Algorithm 1 is given by

$$
\begin{cases}
\overline{V}_h^k(x) = \max_a \overline{Q}_h^k(x,a), \\
\overline{Q}_h^k(x,a) = H \wedge \left(c^k(x,a) + \widehat{\mathbb{P}}^k \overline{V}_{h+1}^k(x,a) \right).
\end{cases} \tag{B.10}
$$

Let us also define the following population and empirical *bonus value functions* (for some policy π):

$$
\begin{cases}
\widetilde{V}_h^{k,\pi}(x) = \widetilde{Q}_h^{k,\pi}(x,\pi(x)), \\
\widetilde{Q}_h^{k,\pi}(x,a) = H \wedge \left(c^k(x,a) + \mathbb{P}\widetilde{V}_{h+1}^{k,\pi}(x,a) \right);
\end{cases} \tag{B.11}
$$

$$
\begin{cases}
\overline{V}_h^{k,\pi}(x) = \overline{Q}_h^{k,\pi}(x,\pi(x)), \\
\overline{Q}_h^{k,\pi}(x,a) = H \wedge \left(c^k(x,a) + \widehat{\mathbb{P}}^k \overline{V}_{h+1}^{k,\pi}(x,a) \right).
\end{cases} \tag{B.12}
$$

Lemma 7 (Exploration value function maximizes empirical bonus value functions)**.** *For every π and every k,h,x,a,*

$$
\overline{Q}_h^{k,\pi}(x,a) \le \overline{Q}_h^k(x,a) \quad \text{and} \quad \overline{V}_h^{k,\pi}(x) \le \overline{V}_h^k(x).
$$

Proof. Use induction and (B.10) (B.12). □

Lemma 8 (Planning error is upper bounded by population bonus value function)**.** *For every π^k*

$$
V_1^*(x_1) - V_1^{\pi^k}(x_1) \le \widetilde{V}_1^{k,\pi^k}(x_1). \tag{B.13}
$$

Proof. Let $\mathcal{A}_h$ be the σ-field generated by $\{x_1, a_1, \dots, x_h, a_h\}$ (induced by π^k and $\mathbb{P}$). For simplicity, denote $\mathbb{E}_{\geq h}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{A}_{h-1}]$, i.e., taking conditional expectation given a trajectory $\{x_1, a_1, \dots, x_{h-1}, a_{h-1}\}$. Then $\mathbb{E}_{\geq 1}[\cdot]$ is taking the full expectation. From (B.11) we obtain

$$\begin{aligned}
&\mathbb{E}_{\geq h}\left[\widetilde{Q}_h^{k,\pi^k}(x_h, a_h)\right] = \mathbb{E}_{\geq h}\left[H \wedge \left(c(x_h,a_h) + \mathbb{E}_{\geq h+1}\left[\widetilde{V}_{h+1}^{k,\pi^k}(x_{h+1})\right]\right)\right] \\
&= \mathbb{E}_{\geq h}\left[H \wedge \left(c(x_h,a_h) + \mathbb{E}_{\geq h+1}\left[\widetilde{Q}_{h+1}^{k,\pi^k}(x_{h+1}, a_{h+1})\right]\right)\right] \\
&= \mathbb{E}_{\geq h}\left[H \wedge \mathbb{E}_{\geq h+1}\left[c(x_h,a_h) + \widetilde{Q}_{h+1}^{k,\pi^k}(x_{h+1}, a_{h+1})\right]\right] \\
&\geq \mathbb{E}_{\geq h}\mathbb{E}_{t\geq h+1}\left[H \wedge \left(c(x_h,a_h) + \widetilde{Q}_{h+1}^{k,\pi^k}(x_{h+1}, a_{h+1})\right)\right] \qquad (H \wedge \cdot \text{ is concave}) \\
&= \mathbb{E}_{\geq h}\left[H \wedge \left(c(x_h,a_h) + \widetilde{Q}_{h+1}^{k,\pi^k}(x_{h+1}, a_{h+1})\right)\right].
\end{aligned}$$

Recursively applying the above relation, and using a fact that

$$H \wedge (a + H \wedge b) = H \wedge (a+b) \ \text{ for } a, b \geq 0,$$

we obtain that

$$\widetilde{V}_1^{k,\pi^k}(x_1) = \mathbb{E}_{\geq 1}\left[\widetilde{Q}_1^{k,\pi^k}(x_1, a_1)\right] \geq \mathbb{E}_{\geq 1}\left[H \wedge \sum_{h=1}^{H} c(x_h, a_h)\right] = \mathbb{E}_{\pi^k,\mathbb{P}}\left[H \wedge \sum_{h=1}^{H} c(x_h, a_h)\right].$$

Finally, by Lemma 6 and that $V_1^*(x_1) - V_1^{\pi^k}(x_1) \leq H = \mathbb{E}_{\pi^k,\mathbb{P}}[H]$, we have that

$$V_1^*(x_1) - V_1^{\pi^k}(x_1) \leq \mathbb{E}_{\pi^k,\mathbb{P}}\left[H \wedge \sum_{h=1}^{H} c(x_h, a_h)\right] \leq \widetilde{V}_1^{k,\pi^k}(x_1),$$

which completes our proof. □

Lemma 9 (Empirical vs. population bonus value functions)**.** *If G_2 and G_3 hold, then for every k and for every policy π, we have that*

$$\frac{1}{e} \cdot \overline{V}_1^{k,\pi}(x_1) \leq \widetilde{V}_1^{k,\pi}(x_1) \leq e \cdot \overline{V}_1^{k,\pi}(x_1).$$

Proof. For the second inequality, we only need to prove that for every k and π,

$$\widetilde{V}_h^{k,\pi}(x) \leq \left(1 + \frac{1}{H}\right)^{H-h+1} \cdot \overline{V}_h^{k,\pi}(x), \ \text{ for every } h, x.$$

We proceed by induction over h. For $H+1$ the hypothesis holds trivially as $\widetilde{V}_{H+1}^{k,\pi}(x) = 0 = \overline{V}_{H+1}^{k,\pi}(x)$. Now suppose the hypothesis holds for $h+1$, and let us consider h:

$$\begin{aligned}
\widetilde{V}_h^{k,\pi}(x) &= \widetilde{Q}_h^{k,\pi}(x, a) \qquad (\text{set } a = \pi(x)) \\
&\leq c^k(x,a) + \mathbb{P}\widetilde{V}_{h+1}^{\pi}(x,a) \qquad (\text{by (B.11)}) \\
&= c^k(x,a) + \widehat{\mathbb{P}}^k\widetilde{V}_{h+1}^{k,\pi}(x,a) + (\mathbb{P} - \widehat{\mathbb{P}}^k)\widetilde{V}_{h+1}^{k,\pi}(x,a) \\
&\leq c^k(x,a) + \left(1 + \frac{1}{H}\right)\widehat{\mathbb{P}}^k\widetilde{V}_{h+1}^{k,\pi}(x,a) + \frac{3H^2 S\iota}{N^k(x,a)} \qquad (\text{by Lemma 4 and } \widetilde{V}_{h+1}^{k,\pi} \leq H) \\
&\leq \left(1 + \frac{1}{H}\right)c^k(x,a) + \left(1 + \frac{1}{H}\right)\widehat{\mathbb{P}}^k\widetilde{V}_{h+1}^{k,\pi}(x,a). \qquad (\text{by (B.9)}) \\
&\leq \left(1 + \frac{1}{H}\right)c^k(x,a) + \left(1 + \frac{1}{H}\right)^{H-h+1}\widehat{\mathbb{P}}^k\overline{V}_{h+1}^{k,\pi}(x,a) \qquad (\text{by induction hypothesis}) \\
&\leq \left(1 + \frac{1}{H}\right)^{H-h+1} \cdot \left(c^k(x,a) + \widehat{\mathbb{P}}^k\overline{V}_{h+1}^{k,\pi}(x,a)\right),
\end{aligned}$$

moreover from (B.11) we have that $\widetilde{V}_h^{k,\pi}(x) \le H$. In sum, we have

$$\begin{aligned}\widetilde{V}_h^{k,\pi}(x) &\le \left(1+\frac{1}{H}\right)^{H-h+1} \cdot \left(H \wedge \left(c^k(x,a) + \widehat{\mathbb{P}}^k \overline{V}_{h+1}^{k,\pi}(x,a)\right)\right) \\ &= \left(1+\frac{1}{H}\right)^{H-h+1} \cdot \overline{Q}_{h+1}^{k,\pi}(x,a) = \left(1+\frac{1}{H}\right)^{H-h+1} \cdot \overline{V}_{h+1}^{k,\pi}(x).\end{aligned}$$

This completes our induction, and as a consequence proves the second inequality.

The first inequality is proved by repeating the above argument to show that for every k and π

$$\overline{V}_h^{k,\pi}(x) \le \left(1+\frac{1}{H}\right)^{H-h+1} \cdot \widetilde{V}_h^{k,\pi}(x), \text{ for every } h, x.$$

□

Lemma 10 (Exploration regret)**.** *If G_2, G_3 and G_4 hold, then*

$$\sum_{k=1}^{K} \overline{V}_1^k(x_1) \lesssim \frac{H^3 SA\iota}{\rho} + H^4 S^2 A\iota \cdot \log(HK).$$

Proof. Recall $\bar{\pi}^k$ is the exploration policy at the k-th episode, i.e., a greedy policy given by maximizing $\overline{Q}_h^k(x,a)$. Recall $w_h^k(x,a) := \mathbb{P}\left\{(x_h,a_h) = (x,a) \mid \bar{\pi}^k, \mathbb{P}\right\}$ and $w^k(x,a) = \sum_h w_h^k(x,a)$. Let us consider the following "good sets":

$$L^k := \Big\{(x,a) : \sum_{j<k} w^j(x,a) \ge H^3 S\iota\Big\}. \tag{B.14}$$

Then we have

$$\begin{aligned}&\sum_{k=1}^{K} \overline{V}_1^k(x_1) \le e \cdot \sum_{k=1}^{K} \widetilde{V}_1^{k,\bar{\pi}^k}(x_1) \qquad \text{(use Lemma 9)} \\ &\le e \cdot \sum_{k=1}^{K} \sum_{x,a\in L^k} \sum_{h=1}^{H} w_h^k(x,a) c^k(x,a) + e \cdot \sum_{k=1}^{K} \sum_{x,a\notin L^k} \sum_{h=1}^{H} w_h^k(x,a) H \qquad \text{(use (B.11))} \\ &= e \cdot \sum_{k=1}^{K} \sum_{x,a\in L^k} w^k(x,a) c^k(x,a) + e \cdot \sum_{k=1}^{K} \sum_{x,a\notin L^k} w^k(x,a) H \\ &\le e \cdot \sum_{k=1}^{K} \sum_{x,a\in L^k} w^k(x,a) \left(\mathtt{clip}_{\frac{\rho}{2H}}\left[\sqrt{\frac{8H^2\iota}{N^k(x,a)}}\right] + \frac{120(H+S)H^3\iota}{N^k(x,a)} + \frac{240H^6S^2\iota^2}{\left(N^k(x,a)\right)^2}\right) \\ &\quad + e \cdot \sum_{k=1}^{K} \sum_{x,a\notin L^k} w^k(x,a) H \qquad \text{(use (B.9))}.\end{aligned}$$

We then bound each terms using integration tricks and that G_4 holds. The fourth term is bounded by

$$e \sum_{k=1}^{K} \sum_{x,a\notin L^k} w^k(x,a) H \lesssim SA \cdot (H + H^3 S\iota) \cdot H \lesssim H^4 S^2 A\iota.$$

The third term is bounded by

$$\begin{aligned}&e \sum_{k=1}^{K} \sum_{x,a\in L^k} w^k(x,a) \frac{240H^6S^2\iota^2}{\left(N^k(x,a)\right)^2} \\ &\lesssim H^6 S^2 \iota^2 \sum_{k=1}^{K} \sum_{x,a} \frac{w^k(x,a)}{\left(\sum_{j<k} w^j(x,a) - 2H\iota\right)^2} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \ge H^3 S\iota\right] \quad \text{(use } G_4 \text{ and (B.14))}\end{aligned}$$

$$\lesssim H^6 S^2 \iota^2 \cdot SA \cdot \frac{1}{H^3 S\iota - 2H\iota} \lesssim H^3 S^2 A\iota. \qquad \text{(integration trick)}$$

The second term is bounded by

$$\begin{aligned}
& e\sum_{k=1}^{K} \sum_{x,a\in L^k} w^k(x,a)\frac{120(H+S)H^3\iota}{N^k(x,a)} \\
& \lesssim (H+S)H^3\iota \sum_{k=1}^{K}\sum_{x,a} \frac{w^k(x,a)}{\sum_{j<k} w^j(x,a) - 2H\iota} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \geq H^3 S\iota\right] \quad \text{(use } G_4 \text{ and (B.14))} \\
& \lesssim (H+S)H^3\iota \cdot SA \cdot \log(HK) \lesssim (H+S)H^3 SA\iota \cdot \log(HK). \qquad \text{(integration trick)}
\end{aligned}$$

The first term is bounded by

$$\begin{aligned}
& e\sum_{k=1}^{K} \sum_{x,a\in L^k} w^k(x,a)\, \texttt{clip}_{\frac{\rho}{2H}}\left[\sqrt{\frac{8H^2\iota}{N^k(x,a)}}\right] \\
&= e\sum_{k=1}^{K}\sum_{x,a} w^k(x,a)\cdot \sqrt{\frac{8H^2\iota}{N^k(x,a)}} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \geq H^3 S\iota\right] \cdot \mathbb{1}\left[N^k(x,a) \leq \frac{32H^4\iota}{\rho^2}\right] \\
&\lesssim H\sqrt{\iota}\cdot \sum_{k=1}^{K}\sum_{x,a} \frac{w^k(x,a)}{\sqrt{\sum_{j<k} w^j(x,a) - 2H\iota}} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \leq \frac{64H^4\iota}{\rho^2} + 2H\iota\right] \quad \text{(use } G_4\text{)} \\
&\lesssim H\sqrt{\iota}\cdot SA \cdot \sqrt{\frac{H^4\iota}{\rho^2} + H\iota} \lesssim \frac{H^3 SA\iota}{\rho}. \qquad \text{(integration trick)}
\end{aligned}$$

Summing up everything yields that

$$\sum_{k=1}^{K} \overline{V}_1^k(x_1) \lesssim \frac{H^3 SA\iota}{\rho} + (H+S)H^3 SA\iota \cdot \log(HK) + H^3 S^2 A\iota + H^4 S^2 A\iota \lesssim \frac{H^3 SA\iota}{\rho} + H^4 S^2 A\iota \cdot \log(HK).$$

□

Theorem 4 (Restatement of Theorem 1)**.** *With probability at least $1-\delta$, the planning error is bounded by*

$$V_1^*(x_1) - V_1^{\pi}(x_1) \lesssim \frac{H^3 SA}{\rho K} \cdot \log\frac{HSAK}{\delta} + \frac{H^4 S^2 A}{K} \cdot \log(HK) \cdot \log\frac{HSAK}{\delta}.$$

Proof. First by Lemma 1, we have that with probability at least $1-\delta$, G_1, G_2, G_3 and G_4 hold. Next we have the following:

$$\begin{aligned}
V_1^*(x_1) - V_1^{\pi}(x_1) &= \frac{1}{K}\sum_{k=1}^{K}\left(V_1^*(x_1) - V_1^{\pi^k}(x_1)\right) \qquad \text{(by Algorithm 2)} \\
&\leq \frac{1}{K}\sum_{k=1}^{K} \widetilde{V}_1^{k,\pi^k}(x_1) \qquad \text{(by Lemma 8)} \\
&\leq \frac{e}{K}\sum_{k=1}^{K} \overline{V}_1^{k,\pi^k}(x_1) \qquad \text{(by Lemma 9)} \\
&\leq \frac{e}{K}\sum_{k=1}^{K} \overline{V}_1^{k}(x_1) \qquad \text{(by Lemma 7)} \\
&\lesssim \frac{H^3 SA\iota}{\rho K} + \frac{H^4 S^2 A\iota\log(HK)}{K}. \qquad \text{(by Lemma 10)}
\end{aligned}$$

□

Lemma 11 (Properties of the clip operator)**.** *Let $\rho, \rho', a > 0$, then*

- $a \cdot \texttt{clip}_\rho[A] = \texttt{clip}_{a\rho}[a \cdot A]$;
- *Let $\rho \ge \rho'$ and $A \le A'$, then $A - \rho \le \texttt{clip}_\rho[A] \le \texttt{clip}_{\rho'}[A'] \le A'$;*
- $\texttt{clip}_\rho[A + B] \le \texttt{clip}_{\frac{\rho}{2}}[A] + 2B$ *for $B \ge 0$;*
- $\texttt{clip}_\rho[A_1 + \cdots + A_m] \le 2\left\{\texttt{clip}_{\frac{\rho}{2m}}[A_1] + \cdots + \texttt{clip}_{\frac{\rho}{2m}}[A_m]\right\}$.

Proof. The first three claims are easy to see by the definition of the clip operator. The last claim is from (Simchowitz and Jamieson, 2019), for which we provided a proof here for completeness. Without loss of generality, assume that $A_1 + \cdots + A_m \ge \rho$. Let us divide $\{A_i\}_{i=1}^m$ into two groups by examining whether or not $A_i \ge \frac{\rho}{2m}$. Without loss of generality, assume that

$$A_1, \ldots, A_k \ge \frac{\rho}{2m}, \qquad A_{k+1}, \ldots, A_m < \frac{\rho}{2m}.$$

The latter implies that $A_{k+1} + \cdots + A_m < \frac{\rho}{2m} \cdot (m - k) \le \frac{\rho}{2}$, then by $A_1 + \cdots + A_m \ge \rho$ we obtain that

$$A_1 + \cdots + A_k \ge \frac{\rho}{2} > A_{k+1} + \cdots + A_k,$$

In sum, we have that

$$\begin{aligned}
\text{RHS} &= 2\left\{\texttt{clip}_{\frac{\rho}{2m}}[A_1] + \cdots + \texttt{clip}_{\frac{\rho}{2m}}[A_m]\right\} \\
&= 2\{A_1 + \cdots + A_k\} \\
&\ge A_1 + \cdots + A_k + A_{k+1} + \cdots + A_m \\
&= \text{LHS}.
\end{aligned}$$

□

C PROOF OF THE LOWER BOUND (THEOREM 2)

Lemma 12 ((Mannor and Tsitsiklis, 2004), Theorem 1)**.** *There exist positive constants c_1, c_2, ϵ_0, and δ_0, such that for every $n \ge 2$, $\epsilon \in (0, \epsilon_0)$, $\delta \in (0, \delta_0)$, and for every (ϵ, δ)-correct policy, there exists some Bernoulli multi-armed bandit model with n arms, such that the policy needs at least T number of trials where*

$$\mathbb{E}[T] \ge c_1 \frac{n}{\epsilon^2} \log \frac{c_2}{\delta}.$$

In particular, the bandit model can be chosen such that one arm pays 1 w.p. $1/2 + \epsilon/2$, one arm pays 1 w.p. $1/2 + \epsilon$, and the rest arms pay 1 w.p. $1/2$.

Theorem 5 (Restatement of Theorem 2)**.** *Fix $S \ge 5, A \ge 2, H \ge 2 + \log_A S$. There exist positive constants $c_1, c_2, \rho_0, \delta_0$, such that for every $\rho \in (0, \rho_0)$, $\epsilon \in (0, \rho)$, $\delta \in (0, \delta_0)$, and for every (ϵ, δ)-correct policy, there exists some MDP instance with gap ρ, such that*

$$\mathbb{E}[K] \ge c_1 \frac{H^2 SA}{\epsilon\rho} \log \frac{c_2}{\delta}.$$

Proof. The hard example is constructed as in Figure 1. We prove such example witness our lower bound as follows.

Let us call all left orange states the bandit states. Let N_b be the number of visits to the bandit states. Then from the construction, we have that

$$\mathbb{E}[N_b] = \mathbb{E}[K] \cdot \frac{\epsilon}{\rho}.$$

We may without loss of generality image the bandit states as an entity, and at this entity, there are SA arms: one arm pays reward H w.p. $\frac{1}{2}+\frac{\rho}{2H}$, one arm pays reward H w.p. $\frac{1}{2}+\frac{\rho}{H}$, and the rest arms pay reward H w.p. $\frac{1}{2}$. Next, for any (ϵ,δ)-correct policy on the MDP, it induces a policy that is (ρ,δ)-correct policy on the above bandit model. By a linear scaling of the reward from H to 1, it is equivalent to a policy that is $(\frac{\rho}{H},\delta)$-correct in a stand hard-to-learn bandit model with SA-arms.

By Lemma 12, we must have that

$$\mathbb{E}[N_b] \geq c_1 \frac{H^2SA}{\epsilon^2} \log \frac{c_2}{\delta},$$

which implies that

$$\mathbb{E}[K] \geq c_1 \frac{H^2SA}{\epsilon\rho} \log \frac{c_2}{\delta}.$$

Clearly, the MDP discussed above has A actions, $2S$ states, $H+2+\log_A S$ length of the horizon. We next verify that the MDP has $\frac{\rho}{2}$ gap. Notice that except in the left orange states, all actions have the same consequence, therefore the gap is zero if the agent is not at a left orange state. When we are at the left orange state at the Type III model, there is no gap. When we are at the left orange state at the Type II model, the gap is $\left(\frac{1}{2}+\frac{\rho}{H}\right)\cdot H - \frac{1}{H}\cdot H = \rho$. When we are at the left orange state at the Type I model, the gap is $\left(\frac{1}{2}+\frac{\rho}{2H}\right)\cdot H - \frac{1}{H}\cdot H = \frac{\rho}{2}$. By a rescaling of the number of states, length of the horizon, MDP gap, and the absolute constants, the promised lower bound is established.

□

D GAP-DEPENDENT UNSUPERVISED EXPLORATION FOR MULTI-ARMED BANDIT AND MDP WITH A SAMPLING SIMULATOR

Multi-Armed Bandit. The following result is from Theorem 33.1 in (Lattimore and Szepesvári, 2020). For completeness, we restate the result and the proof here.

Lemma 13 (Uniform Exploration)**.** *Consider a Bernoulli bandit with A arms and a minimum non-zero expected reward gap $\rho > 0$. Consider the following policy: in the exploration phase an agent uniformly pulls each arm and collects rewards for $K = T/A$ rounds, and in the planning phase the agent chooses the arm with the highest empirical rewards. Then*

1. *the output is (ϵ,δ)-correct for $\epsilon < \rho$ if $T \eqsim \frac{A}{\rho^2}\log\frac{A}{\delta}$;*
2. *the expected error is at most $\mathbb{E}_\pi[V^* - V^\pi] \lesssim A\exp(-\rho^2 T/A) \propto \exp(-T)$.*

Proof. Let us denote the expected reward of an arm a as r_a, and denote the empirical reward of an arm a as $\widehat{R}_a = (R_a^1 + \cdots + R_a^K)/K$. Suppose a is the best arm, and a' is the arm with highest empirical reward, then

$$\begin{aligned}
\mathbb{P}\{a' \neq a\} &= \mathbb{P}\left\{\widehat{R}_{a'} > \widehat{R}_a\right\} \\
&= \mathbb{P}\left\{\left(\widehat{R}_{a'} - r_{a'}\right) - \left(\widehat{R}_a - r_a\right) > r_a - r_{a'}\right\} \\
&\leq \mathbb{P}\left\{\left(\widehat{R}_{a'} - r_{a'}\right) - \left(\widehat{R}_a - r_a\right) > \rho\right\} \\
&\lesssim A\exp(-\rho^2 K) \eqsim A\exp(-\rho^2 T/A).
\end{aligned}$$

The proof is completed by noting that $0 \leq r_a \leq 1$. □

MDP with a Sampling Simulator. Now we consider gap-TAE for MDP with a sampling simulator. The algorithm is simple: in the exploration phase we sample N data at each pair (x,a) and compute an empirical transition probability $\widehat{\mathbb{P}}$; in the planning phase we compute the optimal value function over $\widehat{\mathbb{P}}$, and output the induced greedy policy π. Mathematically speaking, the policy π is given by

$$\begin{cases}
\widehat{Q}_h^*(x,a) = r_h(x,a) + \widehat{\mathbb{P}}\widehat{V}_{h+1}^*(x,a), \\
\widehat{V}_h^*(x) = \max_a \widehat{Q}_h^*(x,a), \\
\pi_h(x) = \arg\max_a \widehat{Q}_h^*(x,a).
\end{cases}$$

Next we justify the sample complexity of this algorithm.

Lemma 14 (Good events)**.** *Consider the following two events*

$$G := \left\{\forall x, a, h,\ \left|(\widehat{\mathbb{P}} - \mathbb{P})V^*_{h+1}(x,a)\right| < \frac{\rho}{2H}\right\}, \qquad E := \left\{\forall x, h,\ V^*_h(x) - V^\pi_h(x) = 0\right\},$$

then G implies E.

Proof. Assume G holds, and define $E_h := \{\forall x,\ V^*_h(x) - V^\pi_h(x) = 0\}$. We prove E holds by induction over E_h for $h \in \{H+1, H, \dots, 1\}$. First E_{H+1} holds by definition. Next suppose that $E_{h+1}, \dots, E_{H+1}$ holds, and consider E_h. Since G holds, we have that for every x,

$$V^*_h(x) - \widehat{V}^{\pi^*}_h(x) = \mathbb{E}_{\pi,\widehat{\mathbb{P}}} \sum_{t\ge h} (\mathbb{P} - \widehat{\mathbb{P}})V^*_{t+1}(x_t, a_t) < (H+1-h)\cdot \frac{\rho}{2H} < \frac{\rho}{2}.$$

Since E_t holds for $t \ge h+1$, we have that for every x,

$$\begin{aligned}
\widehat{V}^\pi_h(x) - V^\pi_h(x) &= \mathbb{E}_{\pi,\widehat{\mathbb{P}}} \sum_{t\ge h} (\widehat{\mathbb{P}} - \mathbb{P})V^\pi_{t+1}(x_t, a_t) \\
&= \mathbb{E}_{\pi,\widehat{\mathbb{P}}} \sum_{t\ge h} (\widehat{\mathbb{P}} - \mathbb{P})V^*_{t+1}(x_t, a_t) \qquad (\text{since } E_t \text{ holds for } t \ge h+1) \\
&< (H+1-h)\cdot \frac{\rho}{2H} < \frac{\rho}{2}. \qquad (\text{since } G \text{ holds})
\end{aligned}$$

The above two inequalities imply that for every x,

$$\begin{aligned}
V^*_h(x) - V^\pi_h(x) &= V^*_h(x) - \widehat{V}^{\pi^*}_h(x) + \widehat{V}^{\pi^*}_h(x) - \widehat{V}^\pi_h(x) + \widehat{V}^\pi_h(x) - V^\pi_h(x) \\
&\le V^*_h(x) - \widehat{V}^{\pi^*}_h(x) + \widehat{V}^\pi_h(x) - V^\pi_h(x) < \rho,
\end{aligned}$$

which further yields that for every x and $a = \pi_h(x)$,

$$V^*_h(x) - Q^*_h(x,a) \le V^*_h(x) - V^\pi_h(x) < \rho,$$

which forces $a \in \pi^*_h(x)$ since otherwise $V^*_h(x) - Q^*_h(x,a) \ge \rho$. Therefore, we have that for every x,

$$V^*_h(x) - V^\pi_h(x) = Q^*_h(x,a) - Q^\pi_h(x,a) = \mathbb{P}\left(V^*_{h+1} - V^\pi_{h+1}\right)(x,a) = 0,$$

where the last equality holds since E_{h+1} holds, and by this we show that E_h holds, which completes our induction. □

Lemma 15 (Probability of the good event)**.** $\mathbb{P}\{G^c\} < 2HSA \cdot \exp\left(-\frac{\rho^2 N}{2H^4}\right)$.

Proof. This is by Hoeffding's inequality and a union bound over x, a, h. □

Theorem 6 (Restatement of Theorem 3)**.** *Suppose there is a sampling simulator for the MDP considered in the gap-TAE problem. Consider exploration with the uniformly sampling strategy, and planning with the dynamic programming method with the obtained empirical probability. If T samples are drawn, where*

$$T \ge \frac{2H^4SA}{\rho^2}\cdot \log\frac{2HSA}{\delta},$$

then with probability at least $1-\delta$, the obtained policy is optimal ($\epsilon = 0$).

Proof. Note that $T \ge \frac{2H^4SA}{\rho^2}\cdot\log\frac{2HSA}{\delta}$ implies that $N = T/(SA) \ge \frac{2H^4}{\rho^2}\cdot\log\frac{2HSA}{\delta}$, then by Lemma 15, we have that

$$\mathbb{P}\{G\} \ge 1-\delta,$$

then according to Lemma 14, the policy is optimal with probability at least $1-\delta$. □