Gap-Dependent Unsupervised Exploration for Reinforcement Learning

Jingfeng Wu [1] **Vladimir Braverman** [1] **Lin F. Yang** [2]

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.

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 & 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., 2020; 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 fixed yet unknown during exploration, the problem is called *task-agnostic exploration* (TAE) (Zhang et al., 2020a), and if the reward function is allowed to be chosen arbitrary, 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). Existing algorithms for unsupervised RL exploration (Jin et al., 2020; Zhang et al., 2020a; Wu et al., 2020; 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 & Brunskill, 2015).

However, the above worst-case sample bounds can be *overly pessimistic* in practical scenarios, since the planning task is usually a benign instance. For example, the revealed reward in the planning phase could induce a constant *minimum nonzero sub-optimality gap* (or simply *gap*, which we denote as ρ, that measures 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 & Bartlett, 2007; Ortner & Auer, 2007; Ok et al., 2018). For the supervised RL setting, where a reward function is given in prior, 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 & Jamieson, 2019; Yang

[1]Department of Computer Science, Johns Hopkins University, Baltimore, USA [2]Department of Electrical and Computer Engineering, University of California, Los Angeles, USA. Correspondence to: Jingfeng Wu <uuujf@jhu.edu>, Lin F. Yang <linyang@ee.ucla.edu>.

ICML 2021 Workshop on Reinforcement Learning Theory.

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

et al., 2020; He et al., 2020; Xu et al., 2021). However, to the best of our knowledge, no unsupervised RL algorithm is known to actively take advantage of the gap information in the rewards, and the following question remains open:

Can unsupervised RL problems be solved faster when only targeting on tasks with constant gap?

A Case Study on Multi-Armed Bandit. To gain more intuition on this problem, 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 & 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 & Szepesvári, 2020), or Appendix D). This example provides positive evidence that a constant gap could accelerate unsupervised exploration for RL, too.

Our Contributions. In this paper, we study the *gap-dependent task-agnostic exploration* (gap-TAE) problem[2] 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., 2020); 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 gap, 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., 2020) 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.

Interestingly, our results imply that RL is *truly harder* than MAB in terms 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 is nearly as easy as learning MAB (Jiang & Agarwal, 2018; Wang et al., 2020a; Zhang et al., 2020c).

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. For simplicity we only consider *deterministic and bounded reward function* , which is denoted by $r = \{r_1, \ldots, r_H\}$ 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 $\pi := \{\pi_1, \ldots, \pi_H\}$, where each $\pi_h : \mathcal{S} \to [0,1]^{\mathcal{A}}$ is a potentially random policy at the h-th step. Fixing a policy π, the *Q-value function* and the *value function* are defined as

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

where $x_j \sim \mathbb{P}(\cdot \mid x_{j-1}, a_{j-1})$ and $a_j \sim \pi_j(x_j)$ for $j > h$. For an optimal policy $\pi^* \in \arg\max_\pi V_1^\pi(x_1)$, the 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.

Task-Agnostic Exploration. *Task-agnostic exploration* (TAE) (Zhang et al., 2020a) problems consist of two phases. In the *exploration phase*, the agent interacts with the environment to draw K trajectories generated by the unknown transition probability using its exploration policy. Note that in this phase the agent is unsupervised and cannot observe reward feedback. When the exploration phase is ended, the agent switches to the *planning phase*, and a reward function (independent of the collected dataset, or equivalently, determined before exploration but revealed after exploration) is revealed to the agent[3]. Then the agent needs to compute a *probably-approximately-correct* (PAC) policy based on the collected samples. Formally, a policy π is (ϵ, δ)-PAC, if

$$\mathbb{P}\left\{V_1^*(x_1) - V_1^\pi(x_1) > \epsilon\right\} < \delta,$$

where the probablity is over all the randomness in the procedure of producing π. The *sample complexity* of a TAE

[2] Our algorithm as well as its theoretic results can be easily extended to other unsupervised exploration problems, e.g., reward-free exploration, by a covering and union bound argument.

[3] In the original setup of TAE (Zhang et al., 2020b), N reward functions are revealed in the planning phase, and only bandit feedback is 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. However the presented results and the techniques are ready to be extended to these settings as well in a standard manner.

algorithm is measured by the number of trajectories K that the agent needs to collect during the exploration phase to guarantee being (ϵ, δ)-PAC in planning phase.

Sub-Optimality Gap. The *stage-dependent state-action sub-optimality gap* (Jaksch et al., 2010; Tewari & Bartlett, 2007; Ok et al., 2018; Simchowitz & Jamieson, 2019) for an MDP is defined as

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

Clearly, $\texttt{gap}_h(x, a) = 0$ if and only if a is an optimal action at state x and at the h-th decision step. Intuitively, when $\texttt{gap}_h(x, a) > 0$, $\texttt{gap}_h(x, a)$ characterizes the difficult to distinguish the sub-optimal action a from the optimal actions at state x and at the h-th step; and the larger $\texttt{gap}_h(x, a)$ is, the easier should it be distinguishing a from the optimal actions. We now define the *minimum non-zero stage-dependent state-action sub-optimality gap* as:

$$\rho := \min_{h,x,a} \{\texttt{gap}_h(x, a) : \texttt{gap}_h(x, a) > 0\}.$$

For simplicity, we often refer ρ to the *minimum sub-optimality gap* of an MDP, or simply *gap*, across the paper. Intuitively, an MDP with a constant ρ is easy to learn (Simchowitz & Jamieson, 2019; Yang et al., 2020) since (intuitively) constant number of visitations to an state-action pair suffices to distinguish whether it is optimal.

We will use the following *clip operator* (Simchowitz & Jamieson, 2019)

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

to clip a quantity smaller than ρ to 0.

Gap-Dependent Task-Agnostic Exploration. We are now ready to formally state the *gap-dependent task-agnostic exploration* (gap-TAE) problem. In this problem, an agent is provided with a parameter ρ to start the exploration phase of gap-TAE, where the agent explores the environment without the guidance of any reward. After the exploration phase, the agent is revealed with a reward, with which the ground truth MDP is guaranteed to have a minimum sub-optimality gap at least ρ. The agent then computes a policy under this reward. In the following sections, we will propose a sample efficient algorithm for gap-TAE (Section 3), and provide its sample complexity analysis (Section 4).

3. A Sample-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` (Algorithm 1) maintains an estimated maximum cumulative *bonus* based on the current empirical transition kernel (Algorithm 1, line 5), then explores the environment through executing a greedy policy according to this maximum cumulative bonus. Roughly speaking, the exploration bonus at a state-action pair is inversely proportional to the number of visitations to the state-action pair (Algorithm 1, line 4), 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. In this manner, `UCB-Clip` collects a dataset that contains sufficient samples at every state-action pair for the later planning task. In the planning phase, `UCB-Clip` (Algorithm 2) receives the reward, computes a sequence of *optimistic* estimations to the value function based on the collected samples (Algorithm 2, line 5), and outputs, uniformly at random, a greedy policy that corresponds to one of the optimistic estimated value functions (Algorithm 2, line 6).

`UCB-Clip` is inspired by two existing algorithms: `UCBVI` (Azar et al., 2017), a model-based algorithm that achieves minimax-optimal sample complexity for supervised RL problems, and `PF-UCB` (Wu et al., 2020), a variant of `UCBVI` that efficiently solves the 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 planning (based on samples collected in the exploration phase). Similar to `PF-UCB` but different from `UCBVI`, `UCB-Clip` (Algorithm 1) explores the unknown environment through a greedy policy that minimizes the maximum cumulative exploration bonus. Different from either `PF-UCB` or `UCBVI`, in the exploration phase, `UCB-Clip` adopts a *clipped bonus* (Algorithm 1, line 4).

The clipped 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 zero, when a state-action pair has been visited for sufficiently many times, i.e., when $N(x, a) \approxeq H^4\iota/\rho^2$ (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. Note that the considered MDP has a minimum sub-optimality gap, ρ, thus $N(x, a) \approxeq H^4\iota/\rho^2$ 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 thus saves opportunities for visiting the state-actions that have not yet been visited sufficiently. In this way, `UCB-Clip` accelerates TAE utilizing the sub-optimality gap parameter. We formally justify the above intuitions in the next section.

4. Theoretic Results

4.1. Upper and Lower Bounds

We first present the following two theorems to justify a sample-complexity upper bound for `UCB-Clip` and a sample-complexity lower bound for gap-TAE, respectively. The proofs are deferred to Appendices B and C, respectively.

Theorem 1 (An upper bound for `UCB-Clip`)**.** *Suppose `UCB-Clip` (Algorithm 1) runs for K episodes and collects a dataset $\mathcal{H}^K$. Let policy π be outputted by `UCB-Clip` (Algorithm 2) for an arbitrary input task that is independent of $\mathcal{H}^K$ and has a minimum sub-optimality gap ρ. Then with probability at least $1-\delta$, the planning error is bounded by*

$$V_1^*(x_1) - V_1^\pi(x_1) \leq \widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\rho K} + \frac{H^4S^2A}{K}\Big).$$

Theorem 2 (A lower bound for gap-TAE)**.** *Fix $S \geq 5, A \geq 2, H \geq 2 + \log_A S$. 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 that runs for K episodes, there exists some gap-TAE instances with minimum sub-optimality gap ρ, such that*

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

where the expectation is taken with respect to the randomness of choosing the MDP instances and the randomness of the algorithm.

Remark 1. According to Theorem 1, `UCB-Clip` only requires $\propto \widetilde{\mathcal{O}}(1/\epsilon)$ number of episodes to solve TAE when there is a constant minimum sub-optimality gap, which improves the existing, pessimistic rates $\propto \widetilde{\mathcal{O}}(1/\epsilon^2)$ achieved by the TAE algorithms that focus on worst-case instances (Zhang et al., 2020a; Wu et al., 2020). 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 minimum sub-optimality gap.

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.

4.2. Comparison with Multi-Armed Bandit

Theorems 1 and 2 show that for unsupervised exploration problems on an MDP, even when the instance is benign and has a constant minimum sub-optimality gap, 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) (Lattimore & Szepesvári, 2020). In particular, for an MAB with A arms and a constant minimum sub-optimality gap ρ between the expected rewards associated with the arms, 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., Theorem 33.1 in Lattimore & Szepesvári (2020), or Appendix D). These observations from gap-dependent unsupervised exploration establish a clear 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. $\propto \widetilde{\mathcal{O}}(1)$, i.e., *RL is significantly more challenging than MAB*. This is against an emerging wisdom from the supervised reinforcement learning theory, that RL is nearly as easy as solving MAB (Jiang & Agarwal, 2018; Wang et al., 2020a; Zhang et al., 2020c).

Let us take a deeper look at where MDP is harder than MAB. The hard instance in Figure 2 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 2 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 must run $\propto \Omega(1/\epsilon)$ episodes.

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. Concluding Remarks

In this paper we study sample-efficient algorithms for gap-dependent unsupervised exploration problems in RL. When the targeted planning tasks have a constant minimum non-zero sub-optimality gap, the proposed algorithm achieves a gap-dependent sample complexity upper bound that significantly improves the existing minimax bounds. Moreover, an

information-theoretic lower bound is provided to justify the tightness of the obtained upper bound. These results establish an interesting separation between RL and MAB (or RL with a simulator) in terms of gap-dependent unsupervised exploration problems.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sidford, A., Wang, M., Wu, X., and Ye, Y. 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*, pp. 770–787. SIAM, 2018b.

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

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

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

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. Is long horizon reinforcement learning more difficult than

short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020a.

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

Wu, J., Braverman, V., and Yang, L. F. Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *arXiv preprint arXiv:2011.13034*, 2020.

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

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

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

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

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

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

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

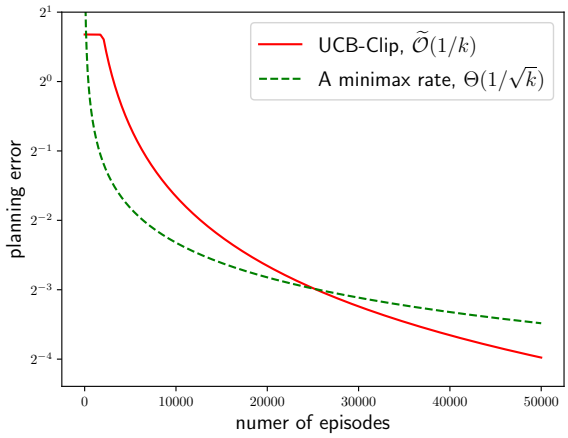

Figure 1. An illustration of the fast task-agnostic exploration achieved by `UCB-Clip`. In the experiment we simulate a random MDP with $H = 5, S = 10, A = 10$ and $\rho = 0.4$, and run `UCB-Clip` for $K = 50,000$ episodes. The red curve shows the planning error of `UCB-Clip`, and the green dotted curve 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 1 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.

B. Proof of the Upper Bound (Theorem 1)

Our proof is inspired by (Simchowitz & Jamieson, 2019) and (Wu et al., 2020).

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 $a \vee b := \max\{a, b\}$. If not noted otherwise, we use ϵ to denote the allowed error in the planning phase, and T to denote the number of samples collected by an algorithm during the exploration phase; in particular if the algorithm is run on some finite-horizon MDP with horizon length H, we use K to denote the number of trajectories collected, then $T = HK$. For simplicity, we use $\propto \widetilde{\mathcal{O}}(\cdot)$ and $\propto \Omega(\cdot)$ to emphasize the rates' dependence on ϵ and treat all other parameters as constants, e.g., $\widetilde{\mathcal{O}}(H^3SA/(\rho\epsilon)) \propto \widetilde{\mathcal{O}}(1/\epsilon)$, and $\propto \widetilde{\mathcal{O}}(1)$ means a rate that depends at most polylogarithmically on ϵ.

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

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

Algorithm 1 `UCB-Clip` (Exploration)

Require: gap parameter ρ, number of episodes K

initialize history $\mathcal{H}^0 = \emptyset$, $\iota = \log(2HS^2AK/\delta)$
for episode $k = 1, 2, \ldots, K$ **do**
  $N^k(x,a),\ \widehat{\mathbb{P}}^k(y \mid x,a) \leftarrow$ `Empi-Prob`$(\mathcal{H}^{k-1})$
  compute exploration bonus $c^k(x,a) := \texttt{clip}_{\frac{\rho}{2H}}\left[\sqrt{\frac{8H^2\iota}{N^k(x,a)}}\right] + \frac{120(S+H)H^3\iota}{N^k(x,a)} + \frac{240H^6S^2\iota^2}{(N^k(x,a))^2}$
  $\left\{\overline{Q}_h^k(x,a),\ \overline{V}_h^k(x)\right\}_{h=1}^H \leftarrow$ `UCB-Q-Value`$(\widehat{\mathbb{P}}^k, r=0, c^k)$
  receive initial state $x_1^k = x_1$
  for step $h = 1, 2, \ldots, H$ **do**
    take action $a_h^k = \arg\max_a \overline{Q}_h^k(x_h^k, a)$, and obtain a new state x_{h+1}^k
  end for
  update history $\mathcal{H}^k = \mathcal{H}^{k-1} \cup \{x_h^k, a_h^k\}_{h=1}^H$
end for
return History $\mathcal{H}^k$

Function `Empi-Prob`
Require: history $\mathcal{H}^{k-1}$
for $(x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ **do**
  $N^k(x,a,y) := \#\{(x,a,y) \in \mathcal{H}^{k-1}\}$, and $N^k(x,a) := \sum_y N^k(x,a,y)$
  if $N^k(x,a) > 0$ **then**
    $\widehat{\mathbb{P}}^k(y \mid x,a) = N^k(x,a,y)/N^k(x,a)$
  else
    $\widehat{\mathbb{P}}^k(y \mid x,a) = 1/S$
  end if
end for
return $N^k(x,a), \widehat{\mathbb{P}}^k(y \mid x,a)$

Function `UCB-Q-Value`
Require: empirical transition $\widehat{\mathbb{P}}^k$, reward function r, bonus function b^k
set $V_{H+1}^k(x) = 0$
for step $h = H, H-1, \ldots, 1$ **do**
  for $(x,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $Q_h^k(x,a) = \min\left\{H,\ r_h(x,a) + b^k(x,a) + \widehat{\mathbb{P}}_h^k V_{h+1}^k(x,a)\right\}$
    $V_h^k(x) = \max_{a \in \mathcal{A}} Q_h^k(x,a)$
  end for
end for
return $\left\{Q_h^k(x,a),\ V_h^k(x)\right\}_{h=1}^H$

Algorithm 2 `UCB-Clip` (Planning)

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

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

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

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

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

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

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

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

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

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

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 $Q_h^k(x, a) \geq Q_h^*(x, a)$ for every k, h, x, a.*

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

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

If $Q_h^k(x, a) = H$, then $Q_h^k(x, a) = H \geq Q_h^*(x, a)$; otherwise, we have that

$$\begin{aligned}
Q_h^k(x, a) - Q_h^*(x, a) &= r_h(x, a) + b^k(x, a) + \widehat{\mathbb{P}}^k V_{h+1}^k(x, a) - r_h(x, a) - \mathbb{P}V_{h+1}^*(x, a) \\
&= b^k(x, a) + \widehat{\mathbb{P}}^k V_{h+1}^k(x, a) - \mathbb{P}V_{h+1}^*(x, a) \\
&\geq b^k(x, a) + (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*(x, a) \qquad (\text{since } Q_{h+1}^k(x, a) \geq Q_{h+1}^*(x, a)) \\
&\geq 0 \qquad (\text{since } G_1 \text{ holds}).
\end{aligned}$$

These complete our induction. □

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

$$E_h^k(x, a) := Q_h^k(x, a) - \left(r_h(x, a) + \mathbb{P}V_{h+1}^k(x, a)\right). \tag{B.3}$$

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

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

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

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

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

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

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

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

inserting which to (B.4) we have that

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

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:*

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

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,

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

For the second inequality,

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

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,*

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

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

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

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

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

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

then we immediately see that

$$\begin{aligned} \ddot{V}_h^k(x) - V_h^{\pi^k}(x) &\geq \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t \geq h} \left(E_t^k(x_t, a_t) - \frac{\rho}{2H} \right) = V_h^k(x) - V_h^{\pi^k}(x) - \frac{H-h+1}{H} \cdot \frac{\rho}{2} \\ &\geq V_h^k(x) - V_h^{\pi^k}(x) - \frac{\rho}{2}. \end{aligned} \tag{B.8}$$

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

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

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

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

$$= \frac{1}{2}\left(V_1^*(x_1) - V_1^{\pi^k}(x_1)\right).$$

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

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

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

Proof. We proceed the proof as follows:

$$\begin{aligned}
V_1^*(x_1) - V_1^{\pi^k}(x_1) &\leq 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \texttt{clip}_{\frac{\rho}{2H}}[E_h^k(x_h,a_h)] \qquad \text{(use Lemma 5)} \\
&\leq 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \texttt{clip}_{\frac{\rho}{2H}} \left[\sqrt{\frac{2H^2\iota}{N^k(x,a)}} + \frac{HS\iota}{N^k(x,a)} \right. \\
&\qquad \left. + \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t\geq h+1} \left(\frac{4e^2H^3\iota}{N^k(x_t,a_t)} + \frac{8e^2H^5S^2\iota^2}{(N^k(x_t,a_t))^2} \right) \right] \qquad \text{(use Lemma 3)} \\
&\leq 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \left\{ \texttt{clip}_{\frac{\rho}{4H}} \left[\sqrt{\frac{2H^2\iota}{N^k(x_h,a_h)}} \right] + \frac{2HS\iota}{N^k(x_h,a_h)} \right. \\
&\qquad \left. + \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t\geq h+1} \left(\frac{8e^2H^3\iota}{N^k(x_t,a_t)} + \frac{16e^2H^5S^2\iota^2}{(N^k(x_t,a_t))^2} \right) \right\} \qquad \text{(use Lemma 11)} \\
&\leq 2 \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \left\{ \texttt{clip}_{\frac{\rho}{4H}} \left[\sqrt{\frac{2H^2\iota}{N^k(x_h,a_h)}} \right] + \frac{2HS\iota}{N^k(x_h,a_h)} \right\} \\
&\qquad + 2H \cdot \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{t=1}^{H} \left(\frac{8e^2H^3\iota}{N^k(x_t,a_t)} + \frac{16e^2H^5S^2\iota^2}{(N^k(x_t,a_t))^2} \right) \\
&= \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \left(\texttt{clip}_{\frac{\rho}{2H}} \left[\sqrt{\frac{8H^2\iota}{N^k(x_h,a_h)}} \right] + \frac{4HS\iota}{N^k(x_h,a_h)} + \frac{16e^2H^4\iota}{N^k(x_h,a_h)} + \frac{32e^2H^6S^2\iota^2}{(N^k(x_h,a_h))^2} \right) \\
&\leq \mathbb{E}_{\pi^k,\mathbb{P}} \sum_{h=1}^{H} \left(\texttt{clip}_{\frac{\rho}{2H}} \left[\sqrt{\frac{8H^2\iota}{N^k(x,a)}} \right] + \frac{120(H^3+S)H\iota}{N^k(x,a)} + \frac{240H^6S^2\iota^2}{(N^k(x,a))^2} \right).
\end{aligned}$$

□

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

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

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,*

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

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

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

Recursively applying the above relation, and using a fact that

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

we obtain that

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

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

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

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*

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

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:

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

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

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

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

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

□

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

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

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":

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

Then we have

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

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

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

The second term is bounded by

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

The first term is bounded by

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

$$\lesssim H\sqrt{\iota}\cdot SA\cdot\sqrt{\frac{H^4\iota}{\rho^2}+H\iota}\lesssim\frac{H^3SA\iota}{\rho}. \qquad \text{(integration trick)}$$

Summing up everything yields that

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

□

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

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

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:

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

□

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 $B\ge0$;*
- $\texttt{clip}_\rho[A_1+\cdots+A_m]\le2\left\{\texttt{clip}_{\frac{\rho}{2m}}[A_1]+\cdots+\texttt{clip}_{\frac{\rho}{2m}}[A_m]\right\}$.

Proof. The first three claims are easy to see by the definition of the clip operator. The last claim is from (Simchowitz & 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

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

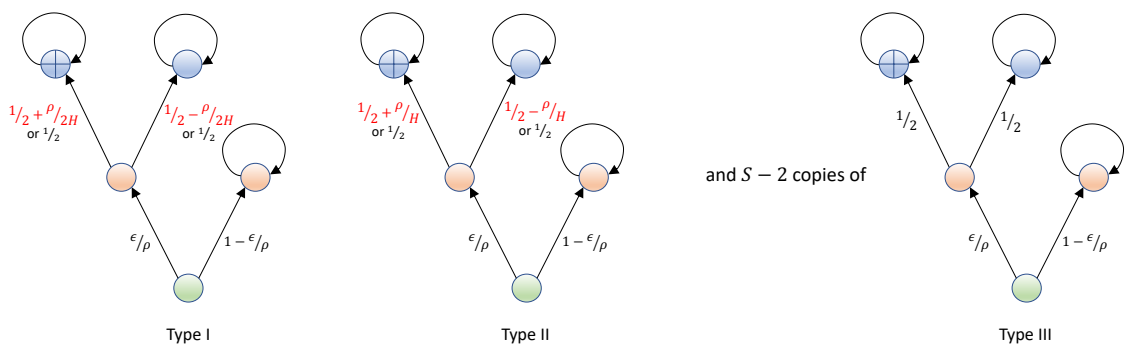

Figure 2. A hard-to-learn MDP example. States are denoted by circles. A state-determined reward function takes value 1 at the state denoted by a circle with a plus sign and takes value zero otherwise. The plot only shows the structure of the last three layers of the MDP, which consists of a Type I model, 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.

In sum, we have that

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

□

C. Proof of the Lower Bound (Theorem 2)

Lemma 12 ((Mannor & Tsitsiklis, 2004), Theorem 1)**.** *There exist positive constants c_1, c_2, ϵ_0, and δ_0, such that for every $n \geq 2$, $\epsilon \in (0, \epsilon_0)$, $\delta \in (0, \delta_0)$, 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*

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

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 $S \geq 5, A \geq 2, H \geq 2+\log_A S$. 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*

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

Proof. The hard example is constructed as in Figure 2. 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

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

which implies that

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

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 & 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 $T \approxeq \frac{A}{\rho^2}\log\frac{A}{\delta}$;*
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

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

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,

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

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

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

The above two inequalities imply that for every x,

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

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

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

which forces $a \in \pi^*_h(x)$ since otherwise $V^*_h(x) - Q^*_h(x,a) \geq \rho$. 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*

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

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

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

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

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