Cooperative Online Learning in Stochastic and Adversarial MDPs

Tal Lancewicki [*1] **Aviv Rosenberg** [*1] **Yishay Mansour** [1 2]

Abstract

We study cooperative online learning in stochastic and adversarial Markov decision process (MDP). That is, in each episode, m agents interact with an MDP simultaneously and share information in order to minimize their individual regret. We consider environments with two types of randomness: *fresh* – where each agent's trajectory is sampled i.i.d, and *non-fresh* – where the realization is shared by all agents (but each agent's trajectory is also affected by its own actions). More precisely, with non-fresh randomness the realization of every cost and transition is fixed at the start of each episode, and agents that take the same action in the same state at the same time observe the same cost and next state. We thoroughly analyze all relevant settings, highlight the challenges and differences between the models, and prove nearly-matching regret lower and upper bounds. To our knowledge, we are the first to consider cooperative reinforcement learning (RL) with either non-fresh randomness or in adversarial MDPs.

1. Introduction

Cooperative multi-agent reinforcement learning (MARL; see Zhang et al. (2021a)) achieved impressive empirical success in many applications such as cyber-physical systems (Adler & Blue, 2002; Wang et al., 2016), finance (Lee et al., 2002; 2007) and sensor/communication networks (Cortes et al., 2004; Choi et al., 2009). Many of the theoretical work on MARL focus on Markov Games (MGs) (Jin et al., 2021), where agents have a shared state, the transition and cost is affected by all agents' actions, and usually the goal is to converge to an equilibrium. On the other hand, in cooperative learning (Lidard et al., 2021), agents do not affect each other and the notion of equilibrium becomes irrelevant. In this model, agents share information with each other and the goal here is to utilize the shared information in order to obtain a significant improvement upon single-agent performance. This model is a generalization of the extensively studied Multi-agent multi-armed bandit (MAB) model (Cesa-Bianchi et al., 2016) and reveals many new challenges that are unique to MDPs.

In this paper we initiate the study of two topics not addressed before in the Cooperative MARL literature. First, we differentiate between two types of randomness: *fresh* – where each agent's trajectory is sampled i.i.d, and *non-fresh* – where at any time the cost and transition kernel's randomness is shared by all agents. More precisely, if at the same time two different agents perform the same action in the same state, they observe the same cost and the same next state. Second, we consider cooperation in the challenging adversarial MDP setting that generalizes stochastic MDPs and allows to model temporal changes in the environment through costs that change arbitrarily and are chosen by an adversary.

While previous works focus mostly on fresh randomness, non-fresh randomness models are just as well-motivated since different agents might experience the same dynamics and rewards when visiting the same state simultaneously. Conceptually, non-fresh randomness models cases where the randomness is more a function of the time than the agent. For example, drones that fly together experience similar weather conditions and autonomous vehicles that drive on the same roads on the same time encounter the same traffic congestion. Moreover, the non-fresh randomness model is theoretically highly challenging, as we show in this paper. We indicate a gap in the lower bounds between fresh and non-fresh randomness and identify the weaknesses of current optimistic approaches in handling non-fresh randomness, thus requiring us to develop new algorithmic techniques.

Our main contributions can be summarized as follows. First, we derive multi-agent versions of known regret minimization algorithms in stochastic and adversarial MDPs, and thoroughly analyze their regret in the fresh randomness model. To complement our bounds, we formally prove matching lower bounds (the adversarial MDP with unknown transitions lower bound nearly-matches, as optimal regret has not been achieved even for a single agent). Second, we point to

[*]Equal contribution [1]Tel Aviv University [2]Google Research, Tel Aviv. Correspondence to: Tal Lancewicki <lancewicki@mail.tau.ac.il>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Table 1. Summary of our regret upper and lower bounds for m agents facing K episode interaction with an MDP that has S states, A actions and horizon H. The bounds ignore poly-logarithmic factors and lower order terms. (*) The algorithm requires $m = \sqrt{K}$ agents.

Algorithm	Regret	Lower Bound	Randomness	Cost	Transition
coop-ULCVI	$\sqrt{\frac{H^3SAK}{m}}$	$\sqrt{\frac{H^3SAK}{m}}$	fresh	stochastic	unknown
coop-O-REPS	$\sqrt{H^2K} + \sqrt{\frac{H^2SAK}{m}}$	$\sqrt{H^2K} + \sqrt{\frac{H^2SAK}{m}}$	fresh	adversarial	known
coop-UOB-REPS	$\sqrt{H^2K} + \sqrt{\frac{H^4S^2AK}{m}}$	$\sqrt{H^2K} + \sqrt{\frac{H^3SAK}{m}}$	fresh	adversarial	unknown
coop-ULCAE	$\sqrt{H^5SK} + \sqrt{\frac{H^7SAK}{\sqrt{m}}}$	$\sqrt{H^2SK} + \sqrt{\frac{H^3SAK}{m}}$	non-fresh	stochastic	unknown
coop-nf-O-REPS	$\sqrt{H^2SK} + \sqrt{\frac{H^2SAK}{m}}$	$\sqrt{H^2SK} + \sqrt{\frac{H^2SAK}{m}}$	non-fresh	adversarial	known
coop-nf-UOB-REPS	$\sqrt{H^4S^2K}$ (*)	$\sqrt{H^2SK} + \sqrt{\frac{H^3SAK}{m}}$	non-fresh	adversarial	unknown

the failure of optimistic methods under non-fresh randomness and prove lower bounds that reveal a significant gap from the fresh randomness case. Our novel constructions for these lower bounds carefully take advantage of the agents' shared random seed to make sure that they cannot explore different areas of the environment simultaneously. Third, we develop a novel multi-agent action-elimination based algorithm for stochastic MDP with non-fresh randomness that forces the agents to scatter at a carefully chosen time so that exploration is maximized. Through novel analysis of the relations between the policies of different agents and the error propagation, we prove near-optimal regret for the algorithm. Finally, for adversarial MDP (where action-elimination is not possible) with non-fresh randomness, we design a novel exploration mechanism to replace optimism and show that it can achieve near-optimal regret for a large number of agents. Table 1 summarizes all our regret lower and upper bounds.

1.1. Related Work

Multi-agent multi-armed bandit. Cooperation was previously studied in both stochastic MAB (Dubey et al., 2020; Wang et al., 2020; Madhushani et al., 2021; Landgren et al., 2021) and adversarial MAB (Cesa-Bianchi et al., 2016; 2019; Bar-On & Mansour, 2019; Ito et al., 2020). While we extend some ideas from the MAB literature to RL, many of the challenges that this paper faces do not arise in MAB. Notably, fresh vs. non-fresh randomness, which is the main focus of the paper, is unique to MDPs as it involves dynamics (i.e., state transitions) that do not exist in MAB.

A different line of work studies *collisions* in cooperative MAB (Liu & Zhao, 2010; Bubeck et al., 2021), where agents are penalized by choosing the same action. In these works, the goal of the cooperation is to minimize the number of collisions rather than improve the agents' performance. Thus, this line of work which has different motivation is only marginally related to our work. See Bubeck et al. (2021) for a more thorough literature review on collisions in MAB.

MARL. There is a long line of research on the theoretical aspects of MARL, mainly focusing on MGs (Littman, 1994; Bai & Jin, 2020; Bai et al., 2020; Xie et al., 2020; Zhang et al., 2020; Liu et al., 2021; Jin et al., 2021). As mentioned before, this literature is only partially related to our model since it aims to converge to an equilibrium rather than minimize individual regret. More related is the literature on decentralized MARL that considers stochastic MDPs with fresh randomness. However, the theoretical guarantees provided by these works are either asymptotic or less tight than our bounds (Zhang et al., 2018a;b; 2021b; Lidard et al., 2021).

Single-agent RL. There is a rich literature on regret minimization in both stochastic (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018; 2020a;b; Yang & Wang, 2019; Zanette et al., 2020a;b) and adversarial (Zimin & Neu, 2013; Rosenberg & Mansour, 2019a;b; 2021; Jin & Luo, 2020; Cai et al., 2020; Luo et al., 2021) MDPs. Note that for a single agent, fresh and non-fresh randomness are identical.

2. Preliminaries

A finite-horizon episodic MDP $\mathcal{M}$ is defined by the tuple $(\mathcal{S}, \mathcal{A}, H, p, \{c^k\}_{k=1}^K)$. $\mathcal{S}$ (of size S) and $\mathcal{A}$ (of size A) are finite state and action spaces, H is the horizon and K is the number of episodes. p is a transition function such that the probability to move to state s' when taking action a in state s at time h is $p_h(s'|s,a)$. $c^k \in [0,1]^{HSA}$ is the cost function for episode k. In the *adversarial* setting the sequence of cost functions $\{c^k\}_{k=1}^K$ is chosen by an oblivious adversary before the interaction starts, while in the *stochastic* setting the costs are sampled i.i.d from a stationary distribution (that does not depend on k) with mean $c_h^k(s,a) = c_h(s,a)$. The adversarial MDP model generalizes stochastic MDPs.

A policy π is a function such that $\pi_h(a|s)$ gives the probability to take action a in state s at time h. If π is deterministic we often abuse notation and use $\pi_h(s)$ for the action chosen by the policy. Given a cost function c, the value $V_h^\pi(s)$ of π is the expected cost when starting from state s at time h, i.e., $V_h^\pi(s) = \mathbb{E}^{p,\pi}[\sum_{h'=h}^H c_{h'}(s_{h'}, a_{h'}) | s_h = s]$ where the notation $\mathbb{E}^{p,\pi}[\cdot]$ means that actions are chosen by π and transitions are determined by p. We also define the Q-function $Q_h^\pi(s,a) = \mathbb{E}^{p,\pi}[\sum_{h'=h}^H c_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a]$ that satisfies the Bellman equations (Sutton & Barto, 2018):

$$Q_h^\pi(s,a) = c_h(s,a) + \mathbb{E}_{p_h(\cdot|s,a)}\left[V_{h+1}^\pi\right]$$
$$V_h^\pi(s) = \langle \pi_h(\cdot \mid s), Q_h^\pi(s,\cdot)\rangle,$$

where $\mathbb{E}_{r(\cdot)}[f]$ denotes the expectation of $f(x)$ where x is sampled from the distribution r, and $\langle \cdot, \cdot \rangle$ is the dot product.

Multi-agent interaction. A team of m agents interacts with the MDP $\mathcal{M}$. At the beginning of episode k, every agent $v \in [m]$ picks a policy $\pi^{k,v}$ and starts in the initial state $s_1^{k,v} = s_{\text{init}}$. At time $h = 1, \ldots, H$, each agent observes its current state $s_h^{k,v}$ and samples an action $a_h^{k,v} \sim \pi_h^{k,v}(\cdot | s_h^{k,v})$. In the *fresh randomness* model, the next state is sampled independently for each agent, i.e., $s_{h+1}^{k,v} \sim p_h(\cdot | s_h^{k,v}, a_h^{k,v})$. For *non-fresh randomness*, the next state is sampled once for each state-action pair $S_h^k(s,a) \sim p_h(\cdot | s,a)$ ahead of the episode, and then every agent v that takes action a in s at time h transitions to the same state $S_h^k(s,a)$, i.e., $s_{h+1}^{k,v} = S_h^k(s_h^{k,v}, a_h^{k,v})$. Similarly, the cost $C_h^{k,v}$ suffered by the agent is either sampled independently when randomness is fresh, or sampled once for each state-action pair (s,a) ahead of the episode when randomness is non-fresh. Note that for adversarial MDPs the costs are not stochastic so it is always the case that $C_h^{k,v} = c_h^k(s_h^{k,v}, a_h^{k,v})$. At the end of the episode, the team observes the trajectories and costs of all agents $\{s_h^{k,v}, a_h^{k,v}, C_h^{k,v}\}_{h=1,v=1}^{H,m}$ (i.e., bandit feedback).

Regret. Let $V^{k,\pi}$ be the value function of π with respect to c^k. The pseudo-regret of an agent v is the cumulative difference between the values of its policies and the values of the best fixed policy in hindsight. The performance of the team is measured by the maximal individual pseudo-regret (note that this criterion is stronger than the average regret):

$$R_K = \max_{v \in [m]} \sum_{k=1}^K V_1^{k,\pi^{k,v}}(s_{\text{init}}) - \min_\pi \sum_{k=1}^K V_1^{k,\pi}(s_{\text{init}}).$$

For stochastic MDP, we use the more common definition of the regret in which, for $k \in [K]$, $V_h^{k,\pi}(s) = V_h^\pi(s)$ for the cost function c (the mean of the costs distribution).

Occupancy measures. For policy π, let q^π be its occupancy measure such that $q_h^\pi(s)$ is the probability to visit s at time h playing π and $q_h^\pi(s,a) = q_h^\pi(s)\pi_h(a \mid s)$. By definition, $V_1^{k,\pi}(s_{\text{init}}) = \langle q^\pi, c^k \rangle$, so we can write the regret in terms of occupancy measures as follows:

$$R_K = \max_{v \in [m]} \sum_{k=1}^K \langle q^{\pi^{k,v}}, c^k \rangle - \min_{q \in \Delta(\mathcal{M})} \sum_{k=1}^K \langle q, c^k \rangle,$$

where $\Delta(\mathcal{M})$ is the set of valid occupancy measures which corresponds to the set of stochastic policies and is a convex polytope in $\mathbb{R}^{HSA}$ defined by $O(HSA)$ linear inequalities.

Additional notations. The notation $\widetilde{O}(\cdot)$ hides constants, lower order terms and poly-logarithmic factors, including $\log(K/\delta)$ for some confidence parameter δ. For $n \in \mathbb{N}$ we denote $[n] = \{1, 2, \ldots, n\}$, the indicator of event E is denoted by $\mathbb{I}\{E\}$, and $x \vee y = \max\{x, y\}$. $\pi^\star$ denotes the optimal policy (best in hindsight for the adversarial case).

3. Fresh Randomness

The main principle that guides us in the design of algorithms for fresh randomness is the following: even if all agents play the same policy, the team still gathers "m times more data". Thus, we take a single-agent regret minimization algorithm `ALG` and let all the agents play the policy that it outputs. `ALG` is then updated based on the observations of all agents.

For the stochastic setting we propose an optimistic algorithm we call `coop-ULCVI` based on the single-agent algorithms of Azar et al. (2017); Dann et al. (2019). The algorithm maintains empirical estimates of the transition probabilities and costs, based on samples from all agents. At the beginning of episode k it constructs an optimistic estimate $\underline{Q}^k$ of the optimal Q-function Q^* so that with high probability (w.h.p) $\underline{Q}_h^k(s,a) \le Q_h^*(s,a)$. The agents all play the same deterministic policy which is greedy with respect to $\underline{Q}^k$: $\pi_h^{k,v}(s) = \arg\max_a \underline{Q}_h^k(s,a)$. Even if agents arrive together at the same state and take the same action, we still get multiple i.i.d samples. Hence, the empirical estimates are based on m times more samples compared to the non-cooperative (single-agent) setting. This is the key property that allows us to prove the following improved regret bound. Detailed description of the `coop-ULCVI` algorithm and the proof of Theorem 3.1 appear in Appendix B.

Theorem 3.1. *For stochastic MDP with fresh randomness,* `coop-ULCVI` *ensures w.h.p,* $R_K = \widetilde{O}\big(\sqrt{\frac{H^3 SAK}{m}}\big)$.

This bound improves upon Lidard et al. (2021) by a factor of $\sqrt{H}$, and is in fact optimal up to logarithmic factors as shown by our lower bound in Appendix A. The lower bound is built on a simple observation (Ito et al., 2020): minimizing the sum of regrets of m agents in K episodes is harder than minimizing the regret of a single agent in mK episodes. By single-agent lower bound (Domingues et al., 2021), the sum of regrets is $\Omega(\sqrt{H^3 SAKm})$, so that the lower bound on the average regret matches our regret bound in Theorem 3.1.

For the adversarial setting we propose `coop-O-REPS` which is based on the single-agent O-REPS algorithm (Zimin & Neu, 2013). Essentially, this is Online Mirror Descent (Shalev-Shwartz et al., 2011) on the set of occupancy measures with entropy regularization. More specifically, in episode k all agents play policy π^k computed as follows:

$$q^{\pi^k} = \arg\min_{q \in \Delta(\mathcal{M})} \eta \langle q, \hat{c}^{k-1} \rangle + \mathrm{KL}(q \parallel q^{\pi^{k-1}}), \qquad (1)$$

where $\mathrm{KL}(\cdot \parallel \cdot)$ is the KL-divergence, η is a learning rate and $\hat{c}^k$ is an importance sampling estimator. The main difference in our algorithm is the new estimator that incorporates the observations from all the different agents as follows:

$$\hat{c}_h^k(s,a) = \frac{c_h^k(s,a)\mathbb{I}\{\exists v : \ s_h^{k,v} = s, a_h^{k,v} = a\}}{W_h^k(s,a) + \gamma}, \qquad (2)$$

where γ is a bias added for high-probability regret (Neu, 2015), and $W_h^k(s,a)$ is the probability that *some* agent visits state s and takes action a at time h – this quantity will play a major role in the analysis of all our algorithms. Conveniently, $W_h^k(s,a) = 1 - (1 - q_h^{\pi^k}(s,a))^m$ as it is the complement of the event that all agents do not visit (s,a) at time h. Thus, the algorithm can be implemented efficiently similarly to the single-agent algorithm (Zimin & Neu, 2013).

For unknown transitions we propose `coop-UOB-REPS` based on single-agent UOB-REPS (Jin et al., 2020a), which is similar to `coop-O-REPS` but uses an estimate $\Delta(\mathcal{M}, k)$ of the set of occupancy measures which contains the true set $\Delta(\mathcal{M})$ with high probability. Note that without knowing p, we cannot compute W_h^k. Instead, we use an optimistic estimate U_h^k which bounds W_h^k from above with high probability. The full algorithms and analysis for the adversarial setting with fresh randomness appear in Appendices D and E.

Theorem 3.2. *For adversarial MDP with fresh randomness,* `coop-O-REPS` *ensures w.h.p,*

$$R_K = \widetilde{O}\left(H\sqrt{K} + \sqrt{\frac{H^2 SAK}{m}} \right),$$

for known dynamics, and `coop-UOB-REPS` *ensures w.h.p,*

$$R_K = \widetilde{O}\left(H\sqrt{K} + \sqrt{\frac{H^4 S^2 AK}{m}} \right),$$

for unknown dynamics.

In Appendix A we show these bounds are optimal, except for an extra $\sqrt{HS}$ factor in the second term of our unknown dynamics bound which we cannot hope to remove here since it is still an open problem even for single-agent. Notice the additional $H\sqrt{K}$ term that does not appear in the stochastic setting. It follows from the lower bound for single-agent adversarial MDP with full-information feedback (not bandit feedback), which is equivalent to our setting in the best case scenario where the agents manage to visit all state-actions.

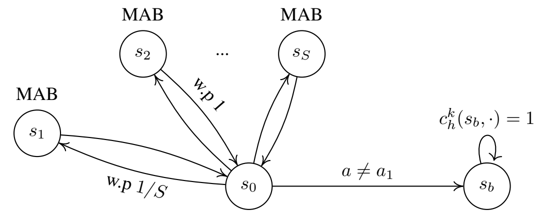

Figure 1. Lower bound construction for non-fresh randomness.

Proof sketch for Theorem 3.2. By standard analysis, the regret scales with two terms: *penalty* of order H/η, and *stability* of order $\eta \sum_{k,h,s,a} q_h^{\pi^k}(s,a)\hat{c}_h^k(s,a)^2$. We show that the stability (which accounts for the estimator's variance) decreases as the number of agents increases. In particular we prove that,

$$q_h^{\pi^k}(s,a) \le \Big(\frac{1}{m} + q_h^{\pi^k}(s,a) \Big) W_h^k(s,a). \qquad (3)$$

This implies that either the probability to observe cost $c_h^k(s,a)$ is m times the probability of a single agent to observe it , or that this probability is at least a constant. Hence, $q_h^{\pi^k}(s,a)\hat{c}_h^k(s,a)^2 \le (1/m + q_h^{\pi^k}(s,a))\hat{c}_h^k(s,a)$. Then, by concentration, the stability amounts to $\eta HK(1 + SA/m)$, and optimizing over η gives the desired bound.

With unknown dynamics, there is an additional error term that comes from the estimation of the occupancy measures set and the bias of the estimator (in particular U_h^k). This error is handled similarly to the stochastic case. □

4. The Challenges of Non-fresh Randomness

Unlike fresh randomness, in the non-fresh randomness setting the total amount of feedback is not necessarily m times the feedback of a single agent. In fact, any algorithm that uses deterministic policies (e.g., optimistic algorithms) simply fails in this setting. The reason is that all agents follow the exact same trajectory since the policy does not introduce any randomness and the transitions are fixed ahead of the episode. Thus, the total amount of feedback is exactly the same as a single-agent would have gathered, which means $\Omega(\sqrt{H^3 SAK})$ regret with no benefit from multiple agents.

The next theorem shows that the *non-fresh randomness* setting is significantly harder than fresh randomness even in terms of the statistical lower bound. While the regret for stochastic MDP with *fresh randomness* scales only logarithmically with K for large enough m, $H\sqrt{SK}$ regret is unavoidable under *non-fresh randomness* even if $m \to \infty$.

Theorem 4.1. *For any $S, A, H, m \in \mathbb{N}$ and $K \ge SAH$, and for any algorithm* `ALG`*, there exists a stochastic MDP with non-fresh randomness such that* `ALG` *suffers expected average regret of at least $\Omega\big(H\sqrt{SK} + \sqrt{\frac{H^3 SAK}{m}}\big)$.*

Proof sketch. We construct the following MDP illustrated in Figure 1. All agents start in s_0. Taking action a_1 transitions to one of the MAB states $s_1, \ldots, s_S$ with probability $1/S$ to each. Taking any other action $a \neq a_1$ transitions to a bad state s_b which is a sink with maximal cost 1. Each MAB state encodes a hard MAB instance: one action gives cost 0 with probability $1/2 + \epsilon$ and cost 1 otherwise, while the rest of the actions give cost 0 or 1 with probability $1/2$. From the MAB states all actions transition back to s_0 with probability 1.

Since the bad state has higher cost than every MAB state and does not contribute to exploration at all, we can assume that all agents choose action a_1 every time they arrive to s_0. Recall that transitions are non-fresh, so all agents visit exactly the same states. This is the critical point in our construction, as it means that exploration is limited to the A actions in the states that all agents visit (and cannot remove S from the regret).

Choosing ϵ as in standard MAB lower bounds, we get that the regret from each of MAB state is $\Omega(\sqrt{(1 + A/m)X})$, where X is the total number of visits to that state. For last, $X \approx K/S$ with high probability which implies that the regret from each MAB state is $\Omega(\sqrt{(1 + A/m)K/S})$. Summing over all states and time steps, we get the desired lower bound. We note that this is a simplified version, missing a factor of $\sqrt{H}$ in the second term. For the full construction and other lower bounds, see Appendix A. □

In Section 5 we face non-fresh randomness in stochastic MDP and present the `coop-ULCAE` algorithm based jointly on optimism and action-elimination. It is important to note that, much like optimistic algorithms, existing RL action elimination algorithms (e.g., Xu et al. (2021)) are deterministic and thus fail in the cooperative non-fresh randomness setting, even though they achieve optimal regret for single-agent. Moreover, naive ways to make these algorithms use stochastic policies that succeed in cooperative MAB, such as uniform exploration of non-eliminated arms, lead to sub-optimal regret in RL because exploration must be controlled more carefully to ensure important states are reached with large enough probability. Hence, we develop a novel exploration method for our algorithm which guarantees that agents can deviate from the optimistic policy and explore potentially optimal actions with minimal effect on the regret.

For adversarial MDP, non-fresh randomness introduces an additional challenge. Due to correlations between the trajectories of different agents, there is no clear and simple relation between $W_h^k(s,a)$ and $q_h^{\pi^k}(s,a)$ as in the fresh randomness setting. In fact, Equation (3) does not hold anymore. In Section 6.1 we present a sophisticated technique for bounding the ratio $q_h^{\pi^k}(s,a)/W_h^k(s,a)$ through a Linear Programming formulation. This allows us to prove optimal regret bounds for adversarial cost and known dynamics.

Existing algorithms for adversarial cost and *unknown* dynamics are optimistic in essence, and as mentioned before, such algorithms fail to utilize cooperation under non-fresh randomness. In Section 6.2 we overcome this challenge with a novel exploration mechanism and prove regret that does not depend on A (up to logarithmic factors) if there are at least $\sqrt{K}$ agents. However, finding the optimal regret for general m still remains an important open question.

5. Non-fresh Randomness - Stochastic MDP

As outlined in Section 4, under non-fresh randomness, we cannot let all agents play the optimistic policy as in the fresh randomness case. Hence, we want agents to occasionally deviate from the optimistic policy for the purpose of exploration. A naive approach would be to let agents explore a random action with probability ϵ and to follow the optimistic policy with probability $1 - \epsilon$. That way, we get $m\epsilon$ more feedback for $m \geq 1/\epsilon$ and the regret of playing the optimistic policy would scale as $\sqrt{SAK/(m\epsilon)}$ (ignoring dependency in H). On the other hand, deviating with an arbitrary action can lead to cost of order of H, which happens for approximately ϵK episodes, so one would have to set $\epsilon \leq \sqrt{SA/K}$ in order to obtain improvement over single agent regret. Thus, the number of agents must be at least $m \geq 1/\epsilon \geq \sqrt{K/(SA)}$ for an improvement. In this section we significantly reduce the number of agents required for a gain in the regret, and show that it can depend on A alone.

Another natural approach, which leads to optimal regret in cooperative MAB, is *action-elimination*, i.e., eliminate all actions that are clearly sub-optimal and explore uniformly at random over non-eliminated actions. However, this approach would also fail in RL because it does not explore efficiently enough. More precisely, agents deviate too much from the optimistic policy so we cannot guarantee that they visit important states. A closer look at action-elimination algorithms for RL (Xu et al., 2021) reveals that they use deterministic policies for this very reason.

Our algorithm, cooperative upper lower confidence action-elimination (`coop-ULCAE`), is presented in Algorithm 1 and in its full version (together with the full analysis) in Appendix C. It takes inspiration from the two previous approaches but utilizes multi-agent exploration in a nearly optimal way. It explores only over non-eliminated actions, but also makes sure that deviation from the optimistic policy is minimal, thus avoiding "non-important" states. This is achieved by playing a random non-eliminated action only at one step during the episode, selected uniformly at random.

Formally, the algorithm maintains a set of *active actions* in each state $\mathcal{A}_h^k(s)$, consisting of only potentially-optimal actions. In episode k, it computes optimistic and pessimistic

Algorithm 1 COOP-ULCAE

initialize: $\mathcal{A}_h^0(s) = \mathcal{A}$ for every $s \in \mathcal{S}$ and $h \in [H]$.
for $k = 1, \ldots, K$ **do**
  Compute $\underline{Q}^k, \overline{Q}^k$ based on empirical estimates.
  Set optimistic policy $\underline{\pi}_h^k(s) \in \arg\min_{a \in \mathcal{A}} \underline{Q}_h^k(s,a)$.
  Eliminate sub-optimal actions: remove a from $\mathcal{A}_h^k(s)$ if $\exists a' \in \mathcal{A}_h^k(s)$ s.t. $\underline{Q}_h^k(s,a) > \overline{Q}_h^k(s,a')$.
  Set policies for agents: for every $v \in [m]$ sample $h_v \in [H]$ uniformly at random and set:

$$\pi^{k,v} = \begin{cases} \underline{\pi}^k & \text{with probability } 1-\epsilon \\ \pi^{k,h_v} & \text{with probability } \epsilon, \end{cases}$$

  where $\pi_h^{k,h'} = \underline{\pi}_h^k$ for any $h \neq h'$ and uniform over $\mathcal{A}_h^k(s)$ at (h,s).
  Play episode k, observe feedback and update empirical estimates.
end for

estimates of Q^*, $\underline{Q}^k$ and $\overline{Q}^k$, respectively, such that w.h.p $\underline{Q}_h^k(s,a) \leq Q_h^*(s,a) \leq \overline{Q}_h^k(s,a)$. Hence, if for actions a and a', $\underline{Q}_h^k(s,a) > \overline{Q}_h^k(s,a')$, then a is clearly sub-optimal and we can eliminate it. The policies of the agents are determined as follows: agent v plays the optimistic policy (greedy with respect to $\underline{Q}^k$) with probability $1-\epsilon$, and with probability ϵ she plays the optimistic policy except for one random time step h_v where she takes a uniformly random active action. The key idea is that deviating on a single time step with an active arm would have only minor affect on the regret, so we can set ϵ much larger compared to the naive ϵ-exploration approach described in the beginning of this section.

Theorem 5.1. *For stochastic MDP with non-fresh randomness,* coop-ULCAE *ensures with high probability,*

$$R_K = \widetilde{O}\left(\sqrt{H^5 SK} + \sqrt{\frac{H^7 SAK}{\sqrt{m}}}\right).$$

If $m \geq H^4 A^2$, the first term is dominant, in which case the regret is nearly optimal and matches our lower bound (Theorem 4.1) up to $H^{3/2}$. Otherwise, we have optimal dependence in S, A, K but there is still a gap of H^2 and more importantly $1/\sqrt[4]{m}$. Determining the optimal dependency in m for this setting is an important open question.

Proof sketch for Theorem 5.1. To simplify presentation, we ignore $poly(H)$ factors in the proof sketch and use the notation $V^\pi = V_1^\pi(s_{\text{init}})$. For agent v, we first break the regret into episodes in which she plays the optimistic policy $\mathcal{K}_{\text{OP}}^v$ and episodes in which she plays an exploration policy $\mathcal{K}_{\text{EXP}}^v$:

$$\underbrace{\sum_{k \in \mathcal{K}_{\text{OP}}^v} V^{\underline{\pi}^k} - V^{\pi^\star}}_{R_{\text{OP}}^v} + \underbrace{\sum_{k \in \mathcal{K}_{\text{EXP}}^v} V^{\pi^{k,h_v}} - V^{\pi^\star}}_{R_{\text{EXP}}^v}.$$

Then, we show that the regret of playing $\underline{\pi}^k$ is bounded by the difference between the optimistic and pessimistic estimates of $Q^\star$ over the trajectory of $\underline{\pi}^k$. This difference shrinks with the confidence radius and mainly scales as,

$$R_{\text{OP}}^v \lesssim \sum_{k=1}^K \sum_{h=1}^H \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \frac{q_h^{\underline{\pi}^k}(s,a)}{\sqrt{n_h^k(s,a) \vee 1}}, \qquad (4)$$

where $n_h^k(s,a) = \sum_{j=1}^{k-1} \mathbb{I}\{\exists v : s_h^{j,v} = s, a_h^{j,v} = a\}$ is the number of times some agent visited (s,a) at time h before episode k. Now, one can show that $n_h^k(s,a)$ is approximately the sum of probabilities that *some* agent visits (h,s,a), i.e., $n_h^k(s,a) \approx \sum_{j=1}^{k-1} W_h^j(s,a)$. Trivially $W_h^j(s,a) \geq q_h^{\underline{\pi}^j}(s,a)$, but we can further utilize the exploration of all the agents to bound $W_h^j(s,a)$ in terms of $q_h^{\underline{\pi}^j}(s)$ and not $q_h^{\underline{\pi}^j}(s,a)$, as follows: With probability $1-(1-\epsilon/HA)^m$ some agent plays the policy $\pi^{j,h}$ and takes action a at time h. In that case, she would arrive to s in time h with probability $q_h^{\pi^{j,h}}(s)$. Recall that $\pi^{j,h}$ and $\underline{\pi}^j$ are identical up to time h, and so $q_h^{\pi^{j,h}}(s) = q_h^{\underline{\pi}^j}(s)$. Also, it is possible to show that $1-(1-\epsilon/A)^m \approx m\epsilon/A$ whenever $\epsilon \leq A/m$. Thus, we get the better bound:

$$n_h^k(s,a) \approx \sum_{j=1}^{k-1} W_h^j(s,a) \gtrsim \frac{m\epsilon}{A} \sum_{j=1}^{k-1} q_h^{\underline{\pi}^j}(s).$$

Combining this with Equation (4), we obtain:

$$R_{\text{OP}}^v \lesssim \sqrt{\frac{A}{m\epsilon}} \sum_{k,h,s} \frac{\sum_a q_h^{\underline{\pi}^k}(s,a)}{\sqrt{\sum_{j=1}^{k-1} q_h^{\underline{\pi}^j}(s)}} \lesssim \sqrt{\frac{SAK}{m\epsilon}}, \qquad (5)$$

where the last relation uses $\sum_{a \in \mathcal{A}} q_h^{\underline{\pi}^k}(s,a) = q_h^{\underline{\pi}^k}(s)$, the Cauchy–Schwarz inequality, and standard arguments (Rosenberg et al., 2020, Lemma B.18) to bound $\sum_k q_h^{\underline{\pi}^k}(s) / \sum_{j=1}^{k-1} q_h^{\underline{\pi}^j}(s) \lesssim \log K$.

For R_{EXP}^v we utilize the fact that when the agent plays an exploration policy, she deviates from the optimistic policy using an active action. Particularly, we show that similar to the regret of the optimistic policy, the regret of the exploration episodes scales with the difference between the optimistic and pessimistic estimates of $Q^\star$ over the trajectory of π^{k,h_v}, but with additional penalty due to the deviation which is overall bounded recursively by R_{OP}^v, i.e.,

$$R_{\text{EXP}}^v \lesssim R_{\text{OP}}^v + \sum_{k \in \mathcal{K}_{\text{EXP}}^v} \sum_{h,s,a} \frac{q_h^{\pi^{k,h_v}}(s,a)}{\sqrt{n_h^k(s,a) \vee 1}}. \qquad (6)$$

While we can bound $n_h^k(s,a)$ as before in terms of $q_h^{\pi^j}(s)$, it can be very different than $q_h^{\pi^{k,h_v}}(s)$. Intuitively, once the agent deviated from the optimal policy, we have a much weaker guarantee on the quality of our confidence sets in the states that she reaches since the cooperative exploration is done over the trajectory of the optimistic policy. Thus, we cannot use similar arguments to the ones we used to bound $R^v_{\texttt{OP}}$, and in particular Equation (4). Instead, we only utilize samples gathered by v in $\mathcal{K}^v_{\texttt{EXP}}$, and bound $n_h^k(s,a) \gtrsim \sum_{j\in\mathcal{K}^v_{\texttt{EXP}}, j<k} q_h^{\pi^{j,h_v}}(s,a)$. Using the fact that the number of exploration episodes for v is approximately $|\mathcal{K}^v_{\texttt{EXP}}| \approx \epsilon K$ and standard arguments, the second term in Equation (6) is bounded by $\sqrt{SAK\epsilon}$. To finish, combine the bounds and set $\epsilon = \min\{\frac{A}{m}, \frac{1}{\sqrt{m}}\}$. □

6. Non-fresh Randomness - Adversarial MDP

6.1. Known Transitions

Before tackling the most challenging model – adversarial MDP with non-fresh randomness and unknown transitions, we first study the case of known transitions. While some of the challenges that we tackled in Section 5 are alleviated when transitions are known, in this section we face additional challenges that stem from the fact that now an adversary is choosing the sequence of cost functions instead of them being sampled from a fixed distribution.

More precisely, under non-fresh randomness there are strong correlations between the trajectories of different agents. This is in stark contrast to the fresh randomness setting where, by playing the same policy for all agents, we obtained different i.i.d samples which enabled us to prove that our estimator has reduced variance and therefore get an improved regret bound (Theorem 3.2). The correlations between the agents' trajectories introduce two main challenges: a statistical challenge and a computational challenge.

The statistical challenge resembles the ones we faced in the stochastic case (Section 5) – whenever the policy is close to being deterministic, the trajectories of the agents are almost identical. However, since costs are adversarial, here we need different techniques that are compatible with adversarial online learning. More formally, for a near-deterministic policy π^k, we have $W_h^k(s,a) \approx q_h^{\pi^k}(s,a)$ which means that there is almost no benefit from the cooperation between the agents. Even though algorithms for adversarial environments inherently choose stochastic policies, it is still unclear a-priori how to bound the ratio $q_h^{\pi^k}(s,a)/W_h^k(s,a)$, except for the trivial bound of 1 which leads to single-agent regret guarantees. Recall that for fresh randomness we bounded this ratio in Equation (3) which relies on the independence of the agents' trajectories given the policy π^k. However, this bound no longer holds and we develop a new bound that is suitable to the non-fresh randomness setting and builds on a novel LP formulation.

The computational challenge follows because we no longer have a closed-form expression to compute $W_h^k(s,a)$ like we had under fresh randomness. We propose to solve this challenge by Monte Carlo estimation of $W_h^k(s,a)$ which we show does not damage the final regret guarantees.

For this setting we propose `coop-nf-O-REPS`, presented (together with its analysis) in Appendix F. It follows the same update rule (Equation (1)) of `coop-O-REPS`, but instead of $W_h^k(s,a)$ (which is now hard to compute) in the definition of the importance sampling estimator (Equation (2)), it uses a Monte Carlo estimate $\widetilde{W}_h^k(s,a)$. The estimate is computed by simulating the run of multiple agents playing policy π^k over the MDP for $\widetilde{O}(K)$ times and taking the fraction of times where some agent visited the state-action pair (s,a) at time h. The approximation error adds a small bias of order $O(1/\sqrt{K})$ which affects the total regret by only a constant factor. We note that the computational complexity of this algorithm is similar to standard O-REPS-based algorithms which are known to have $poly(K,H,S,A)$ per-episode computational complexity (Dick et al., 2014).

Theorem 6.1. *For adversarial MDP with non-fresh randomness and known dynamics,* `coop-nf-O-REPS` *ensures w.h.p,*

$$R_K = \widetilde{O}\left(H\sqrt{SK} + \sqrt{\frac{H^2SAK}{m}}\right).$$

The above regret bound is optimal up to logarithmic factors. The lower bound can be found in Appendix A, and features a similar construction to the one in the proof of Theorem 4.1. Note that in Theorem 4.1 there is an extra $\sqrt{H}$ factor in the second term, which only appears for unknown dynamics.

Proof sketch for Theorem 6.1. Similarly to the proof of Theorem 3.2, the regret scale with the penalty term H/η, and the stability term $\eta\sum_{k,h,s,a} q_h^{\pi^k}(s,a)\hat{c}_h^k(s,a)^2$. Bounding the approximation error of $\widetilde{W}_h^k(s,a)$ by $\gamma/2$ gives us:

$$(stability) \lesssim \eta \sum_{k,h,s,a} \frac{q_h^{\pi^k}(s,a)}{W_h^k(s,a)}\hat{c}_h^k(s,a) \qquad (7)$$

To further bound the right-hand-side we use a standard concentration bound of $\hat{c}_h^k(s,a)$ around $c_h^k(s,a) \le 1$. It remains to bound the ratio $q_h^{\pi^k}(s,a)/W_h^k(s,a)$ which is our main technical novelty in this proof. Let $M_h^k(s)$ be the random variable that represents the number of agents that arrive at state s in time h and denote $p_i = \Pr[M_h^k(s) = i]$. Let $\mathbb{E}^k[\cdot]$ denote an expectation conditioned on everything that occurred before the start of episode k. By definition,

$$W_h^k(s,a) = \mathbb{E}^k\left[1-(1-\pi_h^k(a \mid s))^{M_h^k(s)}\right]$$
$$\geq \mathbb{E}^k\left[\frac{M_h^k(s)\pi_h^k(a \mid s)}{1+M_h^k(s)\pi_h^k(a \mid s)}\right] = \sum_{i=0}^{m}\frac{p_i i\pi_h^k(a \mid s)}{1+i\pi_h^k(a \mid s)}, \quad (8)$$

where the inequality holds deterministically for every realization of $M_h^k(s)$ (see Lemma D.3). Note that the expected value of $M_h^k(s)$ is $mq_h^{\pi^k}(s)$, so the right-hand-side of Eq. (8) is bounded by the value of the following linear program:

$$\min_{p_0,\dots,p_m}\sum_{i=0}^{m} p_i \frac{i\pi_h^k(a \mid s)}{1+i\pi_h^k(a \mid s)}$$
$$s.t. \quad \sum_{i=0}^{m} p_i i = mq_h^k(s) \quad ; \quad \sum_{i=0}^{m} p_i = 1.$$

Now, we can solve the LP by considering the dual problem (see Lemma F.3), and get that $W_h^k(s,a) \geq \frac{mq_h^{\pi^k}(s)\pi_h^k(a|s)}{1+m\pi_h^k(a|s)}$. Hence, Eq. (7) is bounded by $\eta HSK(1+A/m)$, and we obtain the claim by optimizing over η. □

6.2. Unknown Transitions

When facing adversarial MDPs with unknown transitions and non-fresh randomness, we encounter all the challenges presented in previous sections. The combination of these challenges makes this model especially hard from an algorithmic perspective. Specifically, the only way (currently) to obtain regret bounds in adversarial MDPs with unknown dynamics is via optimism. Unfortunately, as discussed in Sections 4 and 5, optimistic methods fail under non-fresh randomness. Moreover, our solution for the stochastic case is based on action-elimination so it cannot be extended to adversarial costs. Instead, in this section we present a novel exploration mechanism which guarantees near-optimal regret for large enough number of agents. Importantly, if we used optimism, regret would not improve even for $m \to \infty$.

We propose the `coop-nf-UOB-REPS` algorithm, presented in Algorithm 2 and in full version (together with analysis) in Appendix G. Similarly to `coop-O-REPS` and `coop-nf-O-REPS`, it maintains a policy π^k through the O-REPS update rule (Equation (9)), however unlike the previous algorithms, some agents play a different policy than π^k for the purpose of exploration. We now present the two key features that allow our algorithm to perform efficient exploration in this challenging setting.

First, we equip the algorithm with a novel exploration mechanism: for every (h,a,k) we assign an agent $\sigma(h,a,k)$ to follow π^k up to time h, and then take action a. The rest of the agents follow the policy π^k. This exploration mechanism is motivated by `coop-ULCAE`, but since costs

Algorithm 2 COOP-NF-UOB-REPS

initialize: define a mapping $\sigma : [H] \times \mathcal{A} \times [K] \to [m]$.
for $k = 1,\dots,K$ **do**
  Compute $\pi_h^k(a \mid s) = q_h^k(s,a)/q_h^k(s)$ for:

$$q^k = \arg\min_{q\in\Delta(\mathcal{M},k)} \eta\langle q, \hat{c}^{k-1}\rangle + \mathrm{KL}(q \parallel q^{k-1}). \quad (9)$$

  Set policies for agents: For every (h,a,k) set the policy of agent $v = \sigma(h,a,k)$ to be:

$$\pi_{h'}^{k,v}(a' \mid s) = \begin{cases}\mathbb{I}\{a'=a\} & h'=h\\ \pi_{h'}^k(a' \mid s) & h' \neq h,\end{cases}$$

  and for the rest of the agents set $\pi^{k,v} = \pi^k$.
  Play episode k, observe feedback, update transition empirical estimates, and compute cost estimator $\hat{c}^k$.
end for

are adversarial, we cannot eliminate actions and thus have much weaker guarantees on the regret of the exploration policies. As a result, we require many agents so that each agent would explore less often. In particular, there are HAK targets to explore and whenever $m \geq \sqrt{K}$ we can choose σ so that each agent performs exploration for at most $HA\sqrt{K}$ episodes. Second, to avoid the complex dependencies between the agents' trajectories, we use a new importance sampling estimator that ignores all agents except for $\sigma(h,a,k)$, i.e.,

$$\hat{c}_h^k(s,a) = \frac{c_h^k(s,a)\mathbb{I}\{s_h^{k,\sigma(h,a,k)} = s\}}{u_h^k(s)+\gamma},$$

where $u_h^k(s) \approx q_h^{\pi^k}(s)$. Notice that this estimator is approximately unbiased (up to γ and approximation errors) since $\mathbb{I}\{a_h^{k,\sigma(h,a,k)} = a\} = 1$. Finally, since we do not know the set occupancy measures under unknown dynamics, we use an approximation $\Delta(\mathcal{M},k)$ based on empirical estimates.

Theorem 6.2. *Assume $S \geq A$ and $m \geq \sqrt{K}$. For adversarial MDP with non-fresh randomness and unknown dynamics,* `coop-nf-O-REPS` *ensures w.h.p,* $R_K = \widetilde{O}(H^2 S\sqrt{K})$.

The above result shows that optimal regret is attainable up to factor of $\sqrt{HS}$. Recall that the extra factor $\sqrt{HS}$ compared to the lower bound of Theorem 4.1 also appear in the state-of-the-art upper bound for single-agent. There is still a significant gap on the number of agents required for optimal regret, and finding the minimal number of agents to ensure such regret still remains an important open problem.

Proof sketch for Theorem 6.2. The proof focuses on bounding the regret of the O-REPS policies $\{\pi^k\}_{k=1}^K$, since the number of episodes that each agent does not play these policies is at most $HA\sqrt{K}$, resulting in extra regret of at most

$H^2 S\sqrt{K}$. Similarly to the proof of Theorem 6.1 we need to bound the stability term, but here the unknown dynamics also introduce additional approximation errors. The analysis of the stability term resembles the proof of Theorem 6.1 but utilizes the fact that $q_h^{\pi^k}(s,a)/u_h^k(s) \lesssim \pi_h^k(a \mid s)$. The analysis of the approximation errors takes inspiration from the proof of Theorem 5.1, and shows that it scales with the sum of confidence radius over the trajectory of π^k:

$$(*) = \sum_{h,s,a,k} \frac{H\sqrt{S} q_h^{\pi^k}(s,a)}{\sqrt{n_h^k(s,a)}} \approx \sum_{h,s,a,k} \frac{H\sqrt{S} q_h^{\pi^k}(s,a)}{\sqrt{\sum_{j=1}^{k-1} W_h^j(s,a)}}.$$

We then utilize agents' exploration to lower bound $W_h^k(s,a)$. Recall that agent $\sigma(h,s,k)$ follows π^k until time h, so she arrives at s with probability $q_h^{\pi^k}(s)$ and then takes action a deterministically. Hence, $W_h^k(s,a) \geq q_h^{\pi^k}(s)$, yielding:

$$(*) \lesssim \sum_{h,s,k} \frac{H\sqrt{S} \sum_a q_h^{\pi^k}(s,a)}{\sqrt{\sum_{j=1}^{k-1} q_h^{\pi^j}(s)}} \lesssim H^2 S\sqrt{K}. \qquad \square$$

7. Conclusions and Future Work

In this paper we studied cooperation in multi-agent RL. We introduced the non-fresh randomness model and characterized its challenges compared to standard fresh randomness. We provided nearly-matching regret lower and upper bounds in all relevant settings, and developed novel techniques for handling different types of randomness in various models.

Our work leaves two important directions for future work. First, our regret bounds for non-fresh randomness with unknown transitions are not tight for both stochastic and adversarial MDPs. Second, we assume agents communicate through a fully-connected graph. Extending our results to general communication graphs (as in MAB) is an interesting future direction that would also require analyzing delayed feedback (Lancewicki et al., 2020; Howson et al., 2021).

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

References

Adler, J. L. and Blue, V. J. A cooperative multi-agent transportation management and route guidance system. *Transportation Research Part C: Emerging Technologies*, 10(5-6):433–454, 2002.

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.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33, 2020.

Bar-On, Y. and Mansour, Y. Individual regret in cooperative nonstochastic multi-armed bandits. *Advances in Neural Information Processing Systems*, 32:3116–3126, 2019.

Bubeck, S., Budzinski, T., and Sellke, M. Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions. In *Conference on Learning Theory*, pp. 821–822. PMLR, 2021.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pp. 605–622. PMLR, 2016.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20(17):1–38, 2019.

Choi, J., Oh, S., and Horowitz, R. Distributed learning and cooperative control for multi-agent systems. *Automatica*, 45(12):2802–2814, 2009.

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34, 2021.

Cortes, J., Martinez, S., Karatas, T., and Bullo, F. Coverage control for mobile sensing networks. *IEEE Transactions on robotics and Automation*, 20(2):243–255, 2004.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: uniform pac bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5717–5727, 2017.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Dick, T., Gyorgy, A., and Szepesvari, C. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pp. 512–520. PMLR, 2014.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Dubey, A. et al. Cooperative multi-agent bandits with heavy tails. In *International Conference on Machine Learning*, pp. 2730–2739. PMLR, 2020.

Efroni, Y., Merlis, N., Saha, A., and Mannor, S. Confidence-budget matching for sequential budgeted learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2937–2947. PMLR, 2021. URL http://proceedings.mlr.press/v139/efroni21a.html.

Howson, B., Pike-Burke, C., and Filippi, S. Delayed feedback in episodic reinforcement learning. *arXiv preprint arXiv:2111.07615*, 2021.

Ito, S., Hatano, D., Sumita, H., Takemura, K., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-I. Delay and cooperation in nonstochastic linear bandits. *Advances in Neural Information Processing Systems*, 33:4872–4883, 2020.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020b.

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in Neural Information Processing Systems*, 33, 2020.

Lancewicki, T., Rosenberg, A., and Mansour, Y. Learning adversarial markov decision processes with delayed feedback. *arXiv preprint arXiv:2012.14843*, 2020.

Landgren, P., Srivastava, V., and Leonard, N. E. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125:109445, 2021.

Lee, J. W., Zhang, B.-T., et al. Stock trading system using reinforcement learning with cooperative agents. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 451–458, 2002.

Lee, J. W., Park, J., Jangmin, O., Lee, J., and Hong, E. A multiagent approach to q-learning for daily stock trading. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):864–877, 2007.

Lidard, J., Madhushani, U., and Leonard, N. E. Provably efficient multi-agent reinforcement learning with fully decentralized communication. *arXiv preprint arXiv:2110.07392*, 2021.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Liu, K. and Zhao, Q. Distributed learning in multi-armed bandit with multiple players. *IEEE transactions on signal processing*, 58(11):5667–5681, 2010.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Luo, H., Wei, C.-Y., and Lee, C.-W. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34, 2021.

Madhushani, U., Dubey, A., Leonard, N., and Pentland, A. One more step towards reality: Cooperative bandits with imperfect communication. *Advances in Neural Information Processing Systems*, 34, 2021.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28:3168–3176, 2015.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486. PMLR, 2019a.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32:2212–2221, 2019b.

Rosenberg, A. and Mansour, Y. Stochastic shortest path with adversarially changing costs. In Zhou, Z. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 2936–2942. ijcai.org, 2021. doi: 10.24963/ijcai.2021/404. URL https://doi.org/10.24963/ijcai.2021/404.

Rosenberg, A., Cohen, A., Mansour, Y., and Kaplan, H. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Seldin, Y., Bartlett, P., Crammer, K., and Abbasi-Yadkori, Y. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pp. 280–287. PMLR, 2014.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Streeter, M. and McMahan, H. B. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Wang, S., Wan, J., Zhang, D., Li, D., and Zhang, C. Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. *Computer networks*, 101:158–168, 2016.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pp. 3674–3682. PMLR, 2020.

Xu, H., Ma, T., and Du, S. S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In Belkin, M. and Kpotufe, S. (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4438–4472. PMLR, 2021. URL http://proceedings.mlr.press/v134/xu21a.html.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

Zhang, K., Yang, Z., and Basar, T. Networked multi-agent reinforcement learning in continuous spaces. In *2018 IEEE conference on decision and control (CDC)*, pp. 2771–2776. IEEE, 2018a.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Basar, T. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pp. 5872–5881. PMLR, 2018b.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021a.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Basar, T. Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 2021b.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Neural Information Processing Systems 26*, 2013.

Additional notations. While in the main paper the notation $\lesssim$ hides lower order terms and logarithmic factors, in the appendix it only hides constant factors, i.e., $x \lesssim y$ if and only if $x = O(y)$. We use the notation $\mathbb{E}^k[\cdot]$ to denote an expectation conditioned on everything that occurred before the beginning of episode k. Furthermore, $\mathbb{E}^k[\cdot \mid \pi]$ denotes an expectation conditioned on everything that occurred before the beginning of episode k, and when playing episode k using the policy π. $n_h^k(s,a)$ denotes the number of samples we have from (s,a,h) in the beginning of episode k. More precisely, for fresh randomness $n_h^k(s,a) = \sum_{j=1}^{k-1} \sum_{v=1}^{m} \mathbb{I}\{s_h^{j,v} = s, a_h^{j,v} = a\}$, and for non-fresh randomness $n_h^k(s,a) = \sum_{j=1}^{k-1} \mathbb{I}\{\exists v: \ s_h^{j,v} = s, a_h^{j,v} = a\}$. Finally, $\mathrm{Var}_{r(\cdot)}[f]$ denotes the variance of $f(x)$ where x is sampled from the distribution r. For transition function p' and policy π, $q_h^{p',\pi}(s,a,s') = \Pr[s_h = s, a_h = a, s_{h+1} = s' \mid p', \pi]$ denotes its occupancy measure, $q_h^{p',\pi}(s,a) = \sum_{s'} q_h^{p',\pi}(s,a,s')$ and $q_h^{p',\pi}(s) = \sum_a q_h^{p',\pi}(s,a)$. When the transition function is p, we often use the shorter notation $q^\pi = q^{p,\pi}$. $\pi^\star$ denotes the optimal policy (or best in hindsight), $V^\star$ denotes its values function and $q^\star$ its occupancy measure.

A. Lower bounds

In this section we provide proofs for the lower bounds that appear in Table 1.

A.1. Fresh randomness

Theorem A.1 (Lower bound for stochastic MDP with fresh randomness)**.** *Let $S, A, H, m \in \mathbb{N}$ and $K \geq SAH$. For any algorithm* `ALG` *there exists a stochastic MDP $\mathcal{M}$ with fresh randomness such that: (i) $\mathcal{M}$ has $\Theta(S)$ states, $\Theta(A)$ actions and horizon $\Theta(H)$; (ii) Running* `ALG` *with m agents for K episodes suffers expected average regret of at least $\Omega(\sqrt{\frac{H^3 SAK}{m}})$.*

Proof. The proof is similar to the proof of Ito et al. (2020, Theorem 4). Notice that cooperative regret minimization with m agents for K episodes is harder than single agent regret minimization for Km episodes, because we can solve the second problem using an algorithm for the first problem. Simply let the agents play the first m episodes one by one, feed the feedback to the algorithm and then again let the agents play sequentially. This implies that the cumulative expected regret of the agents is at least $\Omega(\sqrt{H^3 SAKm})$ by standard lower bounds for MDPs (Domingues et al., 2021). Thus, the average expected regret is at least $\Omega(\sqrt{\frac{H^3 SAK}{m}})$. □

Theorem A.2 (Lower bound for adversarial MDP with fresh randomness and known transition)**.** *Let $S, A, H, m \in \mathbb{N}$ and $K \geq SAH$. For any algorithm* `ALG` *there exists an adversarial MDP $\mathcal{M}$ with fresh randomness such that: (i) $\mathcal{M}$ has $\Theta(S)$ states, $\Theta(A)$ actions and horizon $\Theta(H)$; (ii) Running* `ALG` *with m agents for K episodes, when the transition function is known, suffers expected average regret of at least $\Omega(\sqrt{H^2 K} + \sqrt{\frac{H^2 SAK}{m}})$.*

Proof. The lower bound is obtained by a combination of the following two constructions, i.e., with probability $1/2$ the MDP has the structure of the first construction and with probability $1/2$ of the other one:

1. Similarly to the proof of Theorem A.1, cooperative regret minimization with m for K episodes is harder than single agent regret minimization for Km episodes. Invoking the $\Omega(\sqrt{H^2 SAKm})$ lower bound of Zimin & Neu (2013) for adversarial MDP with bandit feedback and known transition, this gives us the $\Omega(\sqrt{\frac{H^2 SAK}{m}})$ lower bound.
2. Cooperative regret minimization with bandit feedback is harder than single agent regret minimization with full-information feedback because the transition function is known so the agents share information only about the cost function (which is fully revealed under full-information feedback). Thus, for the second construction we can simply use the construction of Zimin & Neu (2013) for the $\Omega(\sqrt{H^2 K})$ lower bound of single agent adversarial MDP with known transition and full-information feedback. □

Theorem A.3 (Lower bound for adversarial MDP with fresh randomness and unknown transition)**.** *Let $S, A, H, m \in \mathbb{N}$ and $K \geq SAH$. For any algorithm* `ALG` *there exists an adversarial MDP $\mathcal{M}$ with fresh randomness such that: (i) $\mathcal{M}$ has $\Theta(S)$ states, $\Theta(A)$ actions and horizon $\Theta(H)$; (ii) Running* `ALG` *with m agents for K episodes, when the transition function is unknown, suffers expected average regret of at least $\Omega(\sqrt{H^2 K} + \sqrt{\frac{H^3 SAK}{m}})$.*

Proof. Similarly to the proof of Theorem A.2 we use a combination of two constructions. The first one is the construction from Theorem A.1 which gives the $\Omega(\sqrt{\frac{H^3SAK}{m}})$ lower bound, and the second one is the construction from Theorem A.2 which gives the $\Omega(\sqrt{H^2K})$ lower bound. □

A.2. Non-fresh randomness

Theorem A.4 (Lower bound for adversarial MDP with non-fresh randomness and known transition)**.** *Let $S, A, H, m \in \mathbb{N}$ and $K \geq SAH$. For any algorithm* `ALG` *there exists an adversarial MDP $\mathcal{M}$ with non-fresh randomness such that: (i) $\mathcal{M}$ has $\Theta(S)$ states, $\Theta(A)$ actions and horizon $\Theta(H)$; (ii) Running* `ALG` *with m agents for K episodes, when the transition function is known, suffers expected average regret of at least $\Omega(\sqrt{H^2SK} + \sqrt{\frac{H^2SAK}{m}})$.*

Proof. Consider the following MDP with horizon $2H$. There are A actions $a_1, a_2, \dots, a_A$ and $S+2$ states: the initial state s_0, a bad state s_b and the MAB states $s_1, s_2, \dots, s_S$. The agent starts in the initial state s_0 where action a_1 transitions to each of the MAB states $s_1, \dots, s_S$ with probability $1/S$, and all the other actions $a_2, \dots, a_A$ transition to the bad state s_b. In the bad state the cost is always 1 and all the actions just stay in it with probability 1. Each MAB state s_i encodes a hard multi-arm bandit problem for each horizon step h. That is, all the actions transition back to the initial state s_0, but one action (sampled uniformly at random) suffers cost 0 with probability $1/2 + \epsilon$ (and otherwise 1) while the other actions suffer cost 0 with probability $1/2$, where $\epsilon \approx \sqrt{SA/K}$ which is standard for MAB/RL lower bounds.

Without loss of generality we can assume that all of the agents always choose action a_1 in the initial state because otherwise they transition to the bad state and suffer maximal cost. Critically, this means that all of the agents visit the same state in every time step (because of non-fresh randomness).

Denote by $T_{i,h}$ the number of visits to MAB state s_i in step h. We utilize the lower bound for cooperation in multi-arm bandit (Seldin et al., 2014; Ito et al., 2020) in order to bound the average expected regret from below by

$$\Omega\left(\mathbb{E}\left[\sum_{i=1}^{S}\sum_{h=1}^{H}\sqrt{\left(1+\frac{A}{m}\right)T_{i,2h}}\right]\right) = \Omega\left(SH\sqrt{1+\frac{A}{m}}\mathbb{E}[\sqrt{X}]\right),$$

for $X \sim Bin(n = K, p = 1/S)$ because in each even step size $2h$ one of the MAB states is sampled uniformly at random. By Lemma A.7, we have $\mathbb{E}[\sqrt{X}] \geq \Omega(\sqrt{np})$ for $n \geq 1/p^2$ which proves the lower bound $\Omega(\sqrt{H^2SK} + \sqrt{\frac{H^2SAK}{m}})$ for $K \geq S^2$. We note that a more involved proof of the lower bound in each state reveals that the standard assumption $K \geq HSA$ is sufficient. For more details see the proof of Rosenberg et al. (2020, Theorem 2.7). □

Theorem A.5 (Lower bound for stochastic MDP with non-fresh randomness)**.** *Let $S, A, H, m \in \mathbb{N}$ and $K \geq SAH$. For any algorithm* `ALG` *there exists a stochastic MDP $\mathcal{M}$ with non-fresh randomness such that: (i) $\mathcal{M}$ has $\Theta(S)$ states, $\Theta(A)$ actions and horizon $\Theta(H)$; (ii) Running* `ALG` *with m agents for K episodes suffers expected average regret of at least $\Omega(\sqrt{H^2SK} + \sqrt{\frac{H^3SAK}{m}})$.*

Proof. Similarly to the proof of Theorem A.2 we use a combination of two constructions. The first one is presented in the rest of the proof and gives the $\Omega(\sqrt{\frac{H^3SAK}{m}})$ lower bound, and the second one is the construction from Theorem A.4 which gives the $\Omega(\sqrt{H^2SK})$ lower bound.

Consider the following MDP with horizon $2H+1$. There are A actions $a_1, a_2, \dots, a_A$ and $2S+3$ states: the initial state s_0, a bad state s_b, a good state s_g, the MAB states $s_1, s_2, \dots, s_S$, and the wait states $s_1^w, \dots, s_S^w$. The agent starts in the initial state s_0 where action a_1 transitions to each of the wait states $s_1^w, \dots, s_S^w$ with probability $1/S$, and all the other actions $a_2, \dots, a_A$ transition to the bad state s_b. In the bad state the cost is always 1 and all the actions just stay in it with probability 1, while in the good state the cost is always 0 and all the actions just stay in it with probability 1.

For each $i \in [S]$, the pair of states (s_i, s_i^w) encodes a hard multi-arm bandit problem with HA actions and costs either 0 or $\Omega(H)$. In the next paragraph we describe how the MAB problem is encoded, but first notice that this achieves the desired lower bound. In each episode all of the agents visit the same MAB problem and do not obtain any information about the other ones. Thus, similarly to the proof of Theorem A.4, we can utilize the lower bound for cooperation in MAB (Seldin et al., 2014) in order to prove the lower bound:

$$\Omega\left(\sum_{i=1}^{S} H\sqrt{\frac{HA}{m}\cdot\frac{K}{S}}\right) = \Omega\left(\sqrt{\frac{H^3SAK}{m}}\right).$$

Finally, we describe how to encode a hard MAB instance through the pair of state (s_i, s_i^w). In the wait state s_i^w the action a_1 transitions to s_i with probability 1, and the action a_2 stays in state s_i^w if the step h is at most $H+2$, otherwise it transitions to the bad state s_b. All the other actions $a_3, \dots, a_A$ always transition to the bad state. In state s_i all the actions transition to the good state s_g with probability $1/2$ and to the bad state s_b with probability $1/2$, except for one action in a specific time step (both sampled uniformly at random) that transition to the good state with probability $1/2+\epsilon$ (and to the bad state with probability $1/2-\epsilon$) for some $\epsilon \approx \sqrt{SA/K}$ which is standard for MAB/RL lower bounds.

Notice that this is in fact MAB with HA actions since the learner needs to pick both the right action and the right horizon step. Moreover, the cost is either 0 if the agents is successful in transitioning to the good state, or $\Theta(H)$ if the learner transitions to the bad state (in any case that the good state is not reached, the bad state will be reached before time step $H+2$). □

Theorem A.6 (Lower bound for adversarial MDP with non-fresh randomness and unknown transition)**.** *Let $S, A, H, m \in \mathbb{N}$ and $K \geq SAH$. For any algorithm* ALG *there exists an adversarial MDP $\mathcal{M}$ with non-fresh randomness such that: (i) $\mathcal{M}$ has $\Theta(S)$ states, $\Theta(A)$ actions and horizon $\Theta(H)$; (ii) Running* ALG *with m agents for K episodes, when the transition function is unknown, suffers expected average regret of at least $\Omega(\sqrt{H^2SK} + \sqrt{\frac{H^3SAK}{m}})$.*

Proof. Follows immediately from Theorem A.5 since adversarial MDPs generalize stochastic MDPs. □

A.3. Auxiliary lemmas

Lemma A.7. *Let $X \sim Bin(n,p)$ and assume that $n \geq 1/p^2$. Then, $\mathbb{E}[\sqrt{X}] \geq 0.01\sqrt{np}$.*

Proof. By Markov inequality we have:

$$\mathbb{E}[\sqrt{X}] \geq \frac{\sqrt{np}}{10}\Pr\left[\sqrt{X} \geq \frac{\sqrt{np}}{10}\right] = \frac{\sqrt{np}}{10}\Pr\left[X \geq \frac{np}{100}\right] = \frac{\sqrt{np}}{10}\left(1 - \Pr\left[X < \frac{np}{100}\right]\right).$$

Thus, it suffices to show that $\Pr\left[X < \frac{np}{100}\right] \leq 9/10$ which follows immediately from Hoeffding inequality and the assumption that $n \geq 1/p^2$. □

Algorithm 3 COOPERATIVE UPPER LOWER CONFIDENCE VALUE ITERATION (COOP-ULCVI)

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, confidence parameter δ, number of episodes K, number of agents m.
initialize: $n^1_h(s,a)=0, n^1_h(s,a,s')=0, C^1_h(s,a)=0\ \forall(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$.
for $k=1,\dots,K$ **do**
  set $\hat{p}^k_h(s'\mid s,a)\leftarrow\frac{n^k_h(s,a,s')}{n^k_h(s,a)\vee 1}, \hat{c}^k_h(s,a)\leftarrow\frac{C^k_h(s,a)}{n^k_h(s,a)\vee 1}\ \forall(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$.
  compute $\{\pi^k_h(s)\}_{s,h}$ via OPTIMISTIC-PESSIMISTIC VALUE ITERATION (Algorithm 4).
  set $n^{k+1}_h(s,a)\leftarrow n^k_h(s,a), n^{k+1}_h(s,a,s')\leftarrow n^k_h(s,a,s'), C^{k+1}_h(s,a)\leftarrow C^k_h(s,a)\ \forall(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$.
  for $v=1,\dots,m$ **do**
    observe initial state $s^{k,v}_1$.
    for $h=1,\dots,H$ **do**
      pick action $a^{k,v}_h=\pi^k_h(s^{k,v}_h)$, suffer cost $C^{k,v}_h\sim c_h(s^{k,v}_h,a^{k,v}_h)$ and observe next state $s^{k,v}_{h+1}\sim p_h(\cdot\mid s^{k,v}_h,a^{k,v}_h)$.
      update $n^{k+1}_h(s^{k,v}_h,a^{k,v}_h)\leftarrow n^{k+1}_h(s^{k,v}_h,a^{k,v}_h)+1, n^{k+1}_h(s^{k,v}_h,a^{k,v}_h,s^{k,v}_{h+1})\leftarrow n^{k+1}_h(s^{k,v}_h,a^{k,v}_h,s^{k,v}_{h+1})+1$.
      update $C^{k+1}_h(s^{k,v}_h,a^{k,v}_h)\leftarrow C^{k+1}_h(s^{k,v}_h,a^{k,v}_h)+C^{k,v}_h$.
    end for
  end for
end for

B. The coop-ULCVI algorithm for stochastic MDPs with fresh randomness

For the setting of stochastic MDPs with fresh randomness we propose the Cooperative Upper Lower Confidence Value Iteration algorithm (coop-ULCVI; see Algorithm 3). The idea is simple: all the agents run the same optimistic policy, but the estimated costs and transition models are updated based on the trajectories of all of them. Since the randomness is fresh in this setting, we expect the agents to observe m times more information. Next, we prove the following optimal regret bound for coop-ULCVI.

Theorem B.1. *With probability $1-\delta$, the individual regret of each agent of* coop-ULCVI *is*

$$R_K = O\left(\sqrt{\frac{H^3SAK}{m}}\log\frac{mKHSA}{\delta} + H^3S^2A\log^2\frac{mKHSA}{\delta}\right).$$

B.1. The good event, optimism and pessimism

Define the following events (for $\tau = 3\log\frac{6SAHKm}{\delta}$):

$$E^c(k)=\left\{\forall(s,a,h):\ |\hat{c}^k_h(s,a)-c_h(s,a)|\le\sqrt{\frac{2\tau}{n^k_h(s,a)\vee 1}}\right\}$$

$$E^p(k)=\left\{\forall(s,a,s',h):\ |p_h(s'|s,a)-\hat{p}^k_h(s'|s,a)|\le\sqrt{\frac{2p_h(s'|s,a)\tau}{n^k_h(s,a)\vee 1}}+\frac{2\tau}{n^k_h(s,a)\vee 1}\right\}$$

$$E^{pv1}(k)=\left\{\forall(s,a,h):\ |\big(\hat{p}^k_h(\cdot|s,a)-p_h(\cdot|s,a)\big)\cdot V^\star_{h+1}|\le\sqrt{\frac{2\mathrm{Var}_{p_h(\cdot|s,a)}(V^\star_{h+1})\tau}{n^k_h(s,a)\vee 1}}+\frac{5H\tau}{n^k_h(s,a)\vee 1}\right\}$$

$$E^{pv2}(k)=\left\{\forall(s,a,h):\ |\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V^\star_{h+1})}-\sqrt{\mathrm{Var}_{\hat{p}^k_h(\cdot|s,a)}(V^\star_{h+1})}|\le\sqrt{\frac{12H^2\tau}{n^k_h(s,a)\vee 1}}\right\}$$

The basic good event, which is the intersection of the above events, is the one used in Efroni et al. (2021). The following lemma establishes that the good event holds with high probability. The proof is supplied in Efroni et al. (2021, Lemma 13) by applying standard concentration results.

Lemma B.2 (The First Good Event). *Let $\mathbb{G}_1=\cap_{k=1}^K E^c(k)\cap_{k=1}^K E^p(k)\cap_{k=1}^K E^{pv1}(k)\cap_{k=1}^K E^{pv2}(k)$ be the basic good event. It holds that $\Pr(\mathbb{G}_1)\ge 1-\delta/2$.*

Under the first good event, we can prove that the value is optimistic using standard techniques (similar to Efroni et al. (2021, Lemma 14)).

Algorithm 4 OPTIMISTIC-PESSIMISTIC VALUE ITERATION

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, confidence parameter δ, number of episodes K, number of agents m, visit counters n^k, empirical transition function $\hat{p}^k$, empirical cost function $\hat{c}^k$.
initialize: $\underline{V}^k_{H+1}(s) = \overline{V}^k_{H+1}(s) = 0$ for all $s \in \mathcal{S}$.
for $h = H, H-1, \dots, 1$ **do**
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      set the bonus $b^k_h(s,a) = b^k_h(s,a;c) + b^k_h(s,a;p)$ defined as follows (for $\tau = 3\log\frac{6SAHKm}{\delta}$),

$$b^k_h(s,a;c) = \sqrt{\frac{2\tau}{n^k_h(s,a) \vee 1}}$$

$$b^k_h(s,a;p) = \sqrt{\frac{2\mathrm{Var}_{\hat{p}^k_h(\cdot|s,a)}(\underline{V}^k_{h+1})\tau}{n^k_h(s,a) \vee 1}} + \frac{44H^2 S\tau}{n^k_h(s,a) \vee 1} + \frac{1}{16H}\mathbb{E}_{\hat{p}^k_h(\cdot|s,a)}\left[\overline{V}^k_{h+1} - \underline{V}^k_{h+1}\right].$$

      compute optimistic and pessimistic Q-functions:

$$\underline{Q}^k_h(s,a) = \hat{c}^k_h(s,a) - b^k_h(s,a) + \mathbb{E}_{\hat{p}^k_h(\cdot|s,a)}[\underline{V}^k_{h+1}]$$
$$\overline{Q}^k_h(s,a) = \hat{c}^k_h(s,a) + b^k_h(s,a) + \mathbb{E}_{\hat{p}^k_h(\cdot|s,a)}[\overline{V}^k_{h+1}].$$

    end for
    set $\pi^k_h(s) \in \arg\min_{a\in\mathcal{A}} \underline{Q}^k_h(s,a)$.
    set $\underline{V}^k_h(s) = \max\{\underline{Q}^k_h(s,\pi^k_h(s)), 0\}$, $\overline{V}^k_h(s) = \min\{\overline{Q}^k_h(s,\pi^k_h(s)), H\}$.
  end for
end for

Lemma B.3 (Upper Value Function is Pessimistic, Lower Value Function is Optimistic)**.** *Conditioned on the first good event $\mathbb{G}_1$, it holds that $\underline{V}^k_h(s) \le V^\star_h(s) \le V^{\pi^k}_h(s) \le \overline{V}^k_h(s)$ for every $k = 1, \dots, K$, $s \in \mathcal{S}$ and $h = 1, \dots, H$.*

Finally, using similar techniques to Efroni et al. (2021, Lemma 21), we can prove an additional high probability event which hold alongside the basic good event $\mathbb{G}_1$. To that end, we define the filtration $\{\mathcal{F}^k\}_{k\ge1}$ as the σ-algebra that contains the information on all observed data until the beginning of episode k (including the initial state of episode k). In addition, we define the filtration $\{\mathcal{F}^k_h\}_{k\ge1,h\ge1}$ as the σ-algebra that contains the information on all observed data until step h of episode k (including the h-th state of episode k).

Lemma B.4 (The Good Event)**.** *Let $\mathbb{G}_1$ be the event defined in Lemma B.2. The second good event is the intersection of two events $\mathbb{G}_2 = E^{OP} \cap E^{\mathrm{Var}}$ defined as follows:*

$$E^{OP} = \left\{\forall h \in [H], v \in [m] :\ \sum_{k=1}^K \mathbb{E}[\overline{V}^k_h(s^{k,v}_h) - \underline{V}^k_h(s^{k,v}_h) \mid \mathcal{F}^k_h] \le 18H^2\tau + \left(1 + \frac{1}{2H}\right)\sum_{k=1}^K \overline{V}^k_h(s^{k,v}_h) - \underline{V}^k_h(s^{k,v}_h)\right\}$$
$$E^{\mathrm{Var}} = \left\{\forall v \in [m] :\ \sum_{k=1}^K\sum_{h=1}^H \mathrm{Var}_{p_h(\cdot|s^{k,v}_h,a^{k,v}_h)}(V^{\pi^k}_{h+1}) \le 4H^3\tau + 2\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}[\mathrm{Var}_{p_h(\cdot|s^{k,v}_h,a^{k,v}_h)}(V^{\pi^k}_{h+1}) \mid \mathcal{F}^k]\right\}.$$

Then, the good event $\mathbb{G} = \mathbb{G}_1 \cap \mathbb{G}_2$ holds with probability at least $1-\delta$.

B.2. Proof of Theorem B.1

Lemma B.5 (Key Recursion Bound)**.** *Conditioning on the good event $\mathbb{G}$, the following bound holds for all $h \in [H]$ and $v \in [m]$.*

$$\sum_{k=1}^{K} \overline{V}_h^k(s_h^{k,v}) - \underline{V}_h^k(s_h^{k,v}) \le 18H^2\tau + \sum_{k=1}^{K} \frac{226H^2 S\tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} + \sum_{k=1}^{K} \frac{2\sqrt{2\tau}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}}$$
$$+ \sum_{k=1}^{K} 2\sqrt{2\tau} \frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}} + \left(1 + \frac{1}{2H}\right)^2 \sum_{k=1}^{K} \overline{V}_{h+1}^k(s_{h+1}^{k,v}) - \underline{V}_{h+1}^k(s_{h+1}^{k,v}).$$

Proof. We bound each of the terms in the sum as follows:

$$\begin{aligned}
\overline{V}_h^k(s_h^{k,v}) - \underline{V}_h^k(s_h^{k,v}) &\le 2b_h^k(s_h^{k,v}, a_h^{k,v}; c) + 2b_h^k(s_h^{k,v}, a_h^{k,v}; p) + \mathbb{E}_{\hat{p}_h^k(\cdot|s_h^{k,v}, a_h^{k,v})}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k] \\
&= 2b_h^k(s_h^{k,v}, a_h^{k,v}; c) + 2b_h^k(s_h^{k,v}, a_h^{k,v}; p) \\
&\quad + \mathbb{E}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k] + (\hat{p}_h^k - p_h)(\cdot|s_h^{k,v}, a_h^{k,v}) \cdot (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k) \\
&\le 2b_h^k(s_h^{k,v}, a_h^{k,v}; c) + 2b_h^k(s_h^{k,v}, a_h^{k,v}; p) \\
&\quad + \frac{8H^2 S\tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} + \left(1 + \frac{1}{4H}\right) \mathbb{E}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k],
\end{aligned} \tag{10}$$

where the last relation holds by Cohen et al. (2021, Lemma B.13) which upper bounds

$$(\hat{p}_h^k - p_h)(\cdot|s_h^{k,v}, a_h^{k,v}) \cdot \left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right) \le \frac{8H^2 S\tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} + \frac{1}{4H} \mathbb{E}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k]$$

by setting $\alpha = 4H, C_1 = C_2 = 2$ and bounding $H\tau(2C_2 + \alpha S C_1/2) \le 8H^2 S\tau$ (the assumption of the lemma holds since the event $\cap_k E^p(k)$ holds). Taking the sum over $k \in [K]$ we get that

$$\begin{aligned}
\sum_{k=1}^{K} \overline{V}_h^k(s_h^{k,v}) - \underline{V}_h^k(s_h^{k,v}) &\le \sum_{k=1}^{K} 2b_h^k(s_h^{k,v}, a_h^{k,v}; c) + \sum_{k=1}^{K} 2b_h^k(s_h^{k,v}, a_h^{k,v}; p) + \sum_{k=1}^{K} \frac{8H^2 S\tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} \\
&\quad + \left(1 + \frac{1}{4H}\right) \sum_{k=1}^{K} \mathbb{E}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k].
\end{aligned} \tag{11}$$

The first sum is bounded by definition by

$$\sum_{k=1}^{K} b_h^k(s_h^{k,v}, a_h^{k,v}; c) \le \sum_{k=1}^{K} \sqrt{\frac{2\tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}},$$

and the second sum is bounded in Efroni et al. (2021, Lemma 24) by

$$\begin{aligned}
\sum_{k=1}^{K} b_h^k(s_h^{k,v}, a_h^{k,v}) &\le \sum_{k=1}^{K} \frac{109H^2 S\tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} + \sqrt{2\tau} \sum_{k=1}^{K} \frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}} \\
&\quad + \frac{1}{8H} \sum_{k=1}^{K} \mathbb{E}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k].
\end{aligned}$$

Plugging this into (11) and rearranging the terms we get

$$
\begin{aligned}
\sum_{k=1}^{K} \overline{V}_h^k(s_h^{k,v}) - \underline{V}_h^k(s_h^{k,v}) &\le \sum_{k=1}^{K} \frac{2\sqrt{2\tau}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}} + 2\sqrt{2\tau} \sum_{k=1}^{K} \frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}} \\
&\quad + \sum_{k=1}^{k} \frac{226 H^2 S \tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} + \left(1 + \frac{1}{2H}\right) \sum_{k=1}^{K} \mathbb{E}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k] \\
&\le 18 H^2 \tau + \sum_{k=1}^{K} \frac{2\sqrt{2\tau}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}} + \sum_{k=1}^{k} \frac{226 H^2 S \tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} \\
&\quad + \sum_{k=1}^{K} 2\sqrt{2\tau} \frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}} + \left(1 + \frac{1}{2H}\right)^2 \sum_{k=1}^{K} \overline{V}_{h+1}^k(s_{h+1}^{k,v}) - \underline{V}_{h+1}^k(s_{h+1}^{k,v}),
\end{aligned}
$$

where the last inequality follows since the second good event holds. □

Proof of Theorem B.1. Start by conditioning on the good event which holds with probability greater than $1 - \delta$. Applying the optimism-pessimism of the upper and lower value function we get

$$
\sum_{k=1}^{K} V_1^{\pi^k}(s_1^{k,v}) - V_1^\star(s_1^{k,v}) \le \frac{1}{m} \sum_{v=1}^{m} \sum_{k=1}^{K} \overline{V}_1^k(s_1^{k,v}) - \underline{V}_1^k(s_1^{k,v}). \tag{12}
$$

Iteratively applying Lemma B.5 and bounding the exponential growth by $(1 + \frac{1}{2H})^{2H} \le e \le 3$, the following upper bound on the cumulative regret is obtained.

$$
\begin{aligned}
(12) &\le 54 H^2 \tau + \frac{1}{m} \sum_{v=1}^{m} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{678 H^2 S \tau}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} \\
&\quad + \frac{1}{m} \sum_{v=1}^{m} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{6\sqrt{2\tau}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1}} + \frac{1}{m} \sum_{v=1}^{m} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{6\sqrt{2\tau \mathrm{Var}_{p_h(\cdot|s_h^{k,v}, a_h^{k,v})}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^{k,v}, a_h^{k,v})}}.
\end{aligned} \tag{13}
$$

We now bound each of the three sums in Equation (13). We bound the first sum in Equation (13) via standard analysis as follows:

$$
\begin{aligned}
\sum_{v=1}^{m} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{n_h^k(s_h^{k,v}, a_h^{k,v}) \vee 1} &= \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a) \vee 1} \\
&= \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \mathbb{I}\{n_h^k(s,a) \ge m\} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a) \vee 1} \\
&\quad + \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \mathbb{I}\{n_h^k(s,a) < m\} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a) \vee 1} \\
&\le 2HSA\tau + \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \mathbb{I}\{n_h^k(s,a) < m\} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a) \vee 1} \\
&\le 2HSA\tau + 2mHSA \le 4mHSA\tau,
\end{aligned} \tag{14}
$$

where the first inequality is by Lemma B.6.

The second sum in Equation (13) is bounded as follows,

$$\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{n_h^k(s_h^{k,v},a_h^{k,v})\vee 1}} \leq \sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\mathbb{I}\{n_h^k(s_h^{k,v},a_h^{k,v})\geq m\}}{\sqrt{n_h^k(s_h^{k,v},a_h^{k,v})\vee 1}} + 2mHSA$$
$$\leq \sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}1}\sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\mathbb{I}\{n_h^k(s_h^{k,v},a_h^{k,v})\geq m\}}{n_h^k(s_h^{k,v},a_h^{k,v})\vee 1}} + 2mHSA$$
$$\leq \sqrt{KHm}\sqrt{2HSA\tau} + 24HSA\tau = \sqrt{2mH^2SAK\tau} + 2mHSA,$$

where the first inequality is similar to Equation (14), the second is by Cauchy–Schwarz, and the third is by Lemma B.6.

The third sum in Equation (13) is bounded by applying the Cauchy–Schwarz inequality as follows,

$$\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^{k,v},a_h^{k,v})}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^{k,v},a_h^{k,v})}} \leq \sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\{n_h^k(s_h^{k,v},a_h^{k,v})\geq m\}\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^{k,v},a_h^{k,v})}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^{k,v},a_h^{k,v})}} + 2mH^2SA$$
$$\leq \sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_{p_h(\cdot|s_h^{k,v},a_h^{k,v})}(V_{h+1}^{\pi^k})}\sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{\mathbb{I}\{n_h^k(s_h^{k,v},a_h^{k,v})\geq m\}}{n_h^k(s_h^{k,v},a_h^{k,v})}} + 2mH^2SA$$
$$\leq \sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_{p_h(\cdot|s_h^{k,v},a_h^{k,v})}(V_{h+1}^{\pi^k})}\sqrt{2HSA\tau} + 2mH^2SA$$
$$\leq \sqrt{2HSA\tau}\sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\mathrm{Var}_{p_h(\cdot|s_h^{k,v},a_h^{k,v})}(V_{h+1}^{\pi^k}) \mid \mathcal{F}^k\right] + 4H^3\tau} + 2mH^2SA$$
$$\leq \sqrt{2HSA\tau}\sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\mathrm{Var}_{p_h(\cdot|s_h^{k,v},a_h^{k,v})}(V_{h+1}^{\pi^k}) \mid \mathcal{F}^k\right]} + 5mH^2SA\tau$$
$$\overset{(*)}{=} \sqrt{2HSA\tau}\sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}\mathbb{E}\left[\left(V_1^{\pi^k}(s_1^{k,v}) - \sum_{h=1}^{H}c_h(s_h^{k,v},a_h^{k,v})\right)^2 \mid \mathcal{F}^k\right]} + 5mH^2SA\tau$$
$$\leq \sqrt{2HSA\tau}\sqrt{\sum_{v=1}^{m}\sum_{k=1}^{K}H^2} + 5mH^2SA\tau \leq \sqrt{2mH^3SAK\tau} + 5mH^2SA\tau.$$

where the first inequality is similar to Equation (14), the third inequality is by Lemma B.6, the forth is by event E^{Var}, and $(*)$ is by the law of total variance (Cohen et al., 2021, Lemma B.14). □

B.3. Auxiliary lemmas

Lemma B.6. *It holds that*

$$\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{n_h^k(s,a)\geq m\}\frac{\sum_{v=1}^{m}\mathbb{I}\{s_h^{k,v}=s,a_h^{k,v}=a\}}{n_h^k(s,a)\vee 1} \leq 2HSA\log(Km).$$

Proof. By Rosenberg et al. (2020, Lemma B.18), we have that

$$\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{n_h^k(s,a)\geq m\}\frac{\sum_{v=1}^{m}\mathbb{I}\{s_h^{k,v}=s,a_h^{k,v}=a\}}{n_h^k(s,a)\vee 1} \leq \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}2\log\left(\sum_{k=1}^{K}\sum_{v=1}^{m}\mathbb{I}\{s_h^{k,v}=s,a_h^{k,v}=a\}\right)$$
$$\leq 2HSA\log(Km).$$
□

Algorithm 5 COOPERATIVE UPPER LOWER CONFIDENCE ACTION ELIMINATION (COOP-ULCAE)

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, confidence parameter δ, number of episodes K, number of agents m, exploration parameter $\epsilon > 0$.
initialize: $n_h^1(s,a) = 0, n_h^1(s,a,s') = 0, C_h^1(s,a) = 0, \mathcal{A}_h^0(s) = \mathcal{A} \; \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.
for $k = 1, \dots, K$ **do**
    set $\hat{p}_h^k(s' \mid s,a) \leftarrow \frac{n_h^k(s,a,s')}{n_h^k(s,a) \vee 1}, \hat{c}_h^k(s,a) \leftarrow \frac{C_h^k(s,a)}{n_h^k(s,a) \vee 1} \; \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.
    compute $\{\underline{\pi}_h^k(s)\}_{s,h}$ via OPTIMISTIC-PESSIMISTIC VALUE ITERATION (Algorithm 4).
    set $\mathcal{A}_h^k(s) \leftarrow \mathcal{A}_h^{k-1}(s)$ for every s, h.
    remove sub-optimal actions for every s, h: if $\exists a, a' \in \mathcal{A}_h^k(s)$ s.t. $\underline{Q}_h^k(s,a) > \overline{Q}_h^k(s,a')$, then $\mathcal{A}_h^k(s) \leftarrow \mathcal{A}_h^k(s) \backslash \{a\}$.
    set $I_h^k(s,a,s') = 0, I_h^k(s,a) = 0, IC_h^k(s,a) = 0 \; \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.
    for $v = 1, \dots, m$ **do**
        sample $h_v \in [H]$ uniformly at random.
        set $\pi^{k,v} = \begin{cases} \underline{\pi}^k & \text{with probability } 1-\epsilon \\ \pi^{k,h_v} & \text{with probability } \epsilon \end{cases}$, where: $\pi_h^{k,h'}(a \mid s) = \begin{cases} \underline{\pi}_h^k(a \mid s) & h \neq h' \\ \frac{1}{\mathcal{A}_h^k(s)} & h = h'. \end{cases}$
        observe initial state $s_1^{k,v}$.
        for $h = 1, \dots, H$ **do**
            pick action $a_h^{k,v} \sim \pi_h^{k,v}(\cdot \mid s_h^{k,v})$, suffer cost $C_h^{k,v}$ and observe next state $s_{h+1}^{k,v}$.
            update $I_h^k(s_h^{k,v}, a_h^{k,v}) \leftarrow 1, I_h^k(s_h^{k,v}, a_h^{k,v}, s_{h+1}^{k,v}) \leftarrow 1, IC_h^k(s_h^{k,v}, a_h^{k,v}) \leftarrow C_h^{k,v}$.
        end for
    end for
    set $n_h^{k+1}(s,a) \leftarrow n_h^k(s,a) + I_h^k(s,a), n_h^{k+1}(s,a,s') \leftarrow n_h^k(s,a,s') + I_h^k(s,a,s') \; \forall (s,a,s',h)$.
    set $C_h^{k+1}(s,a) \leftarrow C_h^k(s,a) + IC_h^k(s,a) \; \forall (s,a,h)$.
end for

C. The `coop-ULCAE` algorithm for stochastic MDPs with non-fresh randomness

For the setting of stochastic MDPs with non-fresh randomness we propose the Cooperative Upper Lower Confidence Action Elimination algorithm (`coop-ULCAE`; see Algorithm 5). Recall that if all the agents play the optimistic policy (like `coop-ULCVI`), the regret will not improve since the randomness is non-fresh. Thus, we want each agent to diverge from the trajectory of the optimistic policy at some point. To that end, at some step each agent takes a random action. At the other steps it follows the optimistic policy to make sure that its regret does not increase. Finally, since all actions have probability to be explored, we eliminate sub-optimal actions to avoid unnecessary over exploration.

Theorem C.1. *With probability $1-\delta$, setting $\epsilon = \min\{\frac{HA}{m}, \frac{1}{\sqrt{m}}\}$, the individual regret of each agent of* `coop-ULCAE` *is*

$$R_K = O\Bigg(\sqrt{H^5 SK} \log \frac{mHSAK}{\delta} + \sqrt{\frac{H^7 SAK}{\sqrt{m}}} \log \frac{mHSAK}{\delta} + \sqrt{\frac{H^8 SAK}{m}} \log \frac{mHSAK}{\delta} + H^5 S^2 A \log^2 \frac{mHSAK}{\delta} + \frac{H^6 S^2 A^2}{\sqrt{m}} \log^2 \frac{mHSAK}{\delta} \Bigg).$$

C.1. The good event, optimism and pessimism

Define the following events (for $\tau = 3\log\frac{6SAHKm}{\delta}$):

$$E^c(k) = \left\{ \forall(s,a,h) : \ |\hat{c}_h^k(s,a) - c_h(s,a)| \le \sqrt{\frac{2\tau}{n_h^k(s,a) \vee 1}} \right\}$$

$$E^p(k) = \left\{ \forall(s,a,s',h) : \ |p_h(s'|s,a) - \hat{p}_h^k(s'|s,a)| \le \sqrt{\frac{2p_h(s'|s,a)\tau}{n_h^k(s,a) \vee 1}} + \frac{2\tau}{n_h^k(s,a) \vee 1} \right\}$$

$$E^{pv1}(k) = \left\{ \forall(s,a,h) : \ | \big((\hat{p}_h^k(\cdot|s,a) - p_h(\cdot|s,a) \big) \cdot V_{h+1}^\star | \le \sqrt{\frac{2\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^\star)\tau}{n_h^k(s,a) \vee 1}} + \frac{5H\tau}{n_h^k(s,a) \vee 1} \right\}$$

$$E^{pv2}(k) = \left\{ \forall(s,a,h) : \ | \sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^\star)} - \sqrt{\mathrm{Var}_{\hat{p}_h^k(\cdot|s,a)}(V_{h+1}^\star)} | \le \sqrt{\frac{12H^2\tau}{n_h^k(s,a) \vee 1}} \right\}$$

The basic good event, which is the intersection of the above events, is the one used in Efroni et al. (2021). The following lemma establishes that the good event holds with high probability. The proof is supplied in Efroni et al. (2021, Lemma 13) by applying standard concentration results.

Lemma C.2 (The First Good Event)**.** *Let $\mathbb{G}_1 = \cap_{k=1}^K E^c(k) \cap_{k=1}^K E^p(k) \cap_{k=1}^K E^{pv1}(k) \cap_{k=1}^K E^{pv2}(k)$ be the basic good event. It holds that $\Pr(\mathbb{G}_1) \ge 1 - \delta/2$.*

Under the first good event, we can prove that the value is optimistic using standard techniques (similar to Efroni et al. (2021, Lemma 14)).

Lemma C.3 (Upper Value Function is Pessimistic, Lower Value Function is Optimistic)**.** *Conditioned on the first good event $\mathbb{G}_1$, it holds that $\underline{V}_h^k(s) \le V_h^\star(s) \le V_h^{\pi^k}(s) \le \overline{V}_h^k(s)$ and that $\underline{Q}_h^k(s,a) \le Q_h^\star(s,a) \le Q_h^{\pi^k}(s,a) \le \overline{Q}_h^k(s,a)$ for every $k = 1,\dots,K$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $h = 1,\dots,H$. Moreover, $\pi_h^\star(s) \in \mathcal{A}_h^k(s)$ for every $k = 1,\dots,K$, $s \in \mathcal{S}$ and $h = 1,\dots,H$.*

Finally, we define the following events that are more specific to our algorithmic action elimination framework:

$$E^{n1} = \left\{ \forall(k,h,s) \in [K]\times[H]\times\mathcal{S} \ \forall a \in \mathcal{A}_h^k(s) : n_h^k(s,a) \ge \frac{m\epsilon}{4HA}\sum_{j=1}^{k-1} q_h^{\pi^j}(s) - \log\frac{6HSA}{\delta} \right\}$$

$$E^{n2} = \left\{ \forall(k,h,s,a,v) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[m] : n_h^k(s,a) \ge \frac{1}{2}\sum_{j=1}^{k-1} q_h^{\pi^{j,v}}(s,a) - \log\frac{6mHSA}{\delta} \right\}$$

$$E^{\epsilon} = \left\{ \forall(h',v) \in [H]\times[m] : \sum_{k=1}^{K} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \le \frac{\epsilon}{H}K + \sqrt{K\log\frac{6mH}{\delta}} \right\}$$

Lemma C.4 (The Second Good Event)**.** *Let $\mathbb{G}_2 = E^{n1} \cap E^{n2} \cap E^\epsilon$ be the second good event. It holds that $\Pr(\mathbb{G}_2) \ge 1 - \delta/2$.*

As a direct consequence, we get that the good event $\mathbb{G}$ which is the intersection of $\mathbb{G}_1$ and $\mathbb{G}_2$ holds with probability $1 - \delta$.

Lemma C.5 (The Good Event)**.** *Let $\mathbb{G}_1$ be the first good event defined in Lemma C.2, and $\mathbb{G}_2$ be the second good event defined in Lemma C.4. Then, the good event $\mathbb{G} = \mathbb{G}_1 \cap \mathbb{G}_2$ holds with probability $1 - \delta$.*

Proof of Lemma C.4. We show that each of the events $\neg E^{n1}, \neg E^{n2}, \neg E^\epsilon$ occur with probability at most $\delta/6$. Then, by a union bound we obtain the statement.

$\Pr[\neg E^{n1}] \le \delta/6$**:** Without loss of generality, assume that in each episode, each agent uniformly randomizes a permutation over all actions, $\sigma^{k,v}$, and in case of exploration takes the first active arm in the permutation $\sigma^{k,v}$: $\arg\min_{a \in \mathcal{A}_h^k(s)} \sigma^{k,v}(a)$.

For any $a \in \mathcal{A}_h^k(s)$,

$$\begin{aligned}
n_h^k(s,a) &= \sum_{j=1}^{k-1} \mathbb{I}\{\exists v : s_h^{j,v} = s, a_h^{j,v} = a\} \\
&\geq \sum_{j=1}^{k-1} \mathbb{I}\{\exists v : s_h^{j,v} = s, a_h^{j,v} = a, h_v = h, \pi^{j,v} = \pi^{j,h_v}, \sigma^{j,v}(a) = 1\} \\
&\underset{(*)}{\geq} \sum_{j=1}^{k-1} \mathbb{I}\{\exists v : s_h^{j,v} = s, h_v = h, \pi^{j,v} = \pi^{j,h_v}, \sigma_s^{j,v}(a) = 1\} \\
&\underset{(**)}{=} \sum_{j=1}^{k-1} \mathbb{I}\{s_h^{j,\pi^j} = s\}\mathbb{I}\{\exists v : h_v = h, \pi^{j,v} = \pi^{j,h_v}, \sigma_s^{j,v}(a) = 1\}, \qquad (15)
\end{aligned}$$

For $(*)$, recall that $a \in \mathcal{A}_h^k(s)$, which implies that $a \in \mathcal{A}_h^j(s)$. Therefore, if $h_v = h, \pi^{j,v} = \pi^{j,h_v}, \sigma^{j,v}(a) = 1$ (that is, the agent explores h, and a is the first action in the permutation), then $a_h^k = a$. $(**)$ is because each agent that randomize $h_v = h$ follows the (deterministic) optimistic until horizon h. Since h_v, $\sigma^{j,v}$ and the event $\{\pi^{j,v} = \pi^{j,h_v}\}$ are randomized independently,

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{I}\{s_h^{j,\pi^j} = s\}\mathbb{I}\{\exists v : h_v = h, \pi^{j,v} = \pi^{j,h_v}, \sigma_s^{j,v}(a) = 1\} \mid \mathcal{F}^j\Big] &= \\
&= q_h^{\pi^j}(s) \Pr\left[\exists v : h_v = h, \pi^{j,v} = \pi^{j,h_v}, \sigma_s^{j,v}(a) = 1\right] \\
&= q_h^{\pi^j}(s) \left[1 - \Pr\left[\forall v : h_v \neq h \vee \pi^{j,v} \neq \pi^{j,h_v} \vee \sigma_s^{j,v}(a) \neq 1\right]\right] \\
&= q_h^{\pi^j}(s) \left[1 - \left(\Pr[h_1 \neq h \vee \pi^{j,1} \neq \pi^{j,h_1} \vee \sigma_s^{j,1}(a) \neq 1]\right)^m\right] \\
&= q_h^{\pi^j}(s) \left[1 - \left(1 - \Pr[h_1 = h, \pi^{j,1} = \pi^{j,h_1}, \sigma_s^{j,1}(a) = 1]\right)^m\right] \\
&= q_h^{\pi^j}(s) \left[1 - \left(1 - \frac{\epsilon}{HA}\right)^m\right] \\
&= q_h^{\pi^j}(s) \left[1 - \left(\left(1 - \frac{\epsilon}{HA}\right)^{\frac{HA}{\epsilon}}\right)^{m\frac{\epsilon}{HA}}\right] \\
&\geq q_h^{\pi^j}(s) \left[1 - e^{-\frac{m\epsilon}{HA}}\right] && ((1-x^{-1})^x \leq e) \\
&\geq q_h^{\pi^j}(s) \left[\frac{m\epsilon}{HA} - \frac{1}{2}\left(\frac{m\epsilon}{HA}\right)^2\right] && (e^{-x} \leq 1 - x + \tfrac{x^2}{2}) \\
&\geq q_h^{\pi^j}(s) \frac{m\epsilon}{2HA}. && (\epsilon \leq \tfrac{HA}{m})
\end{aligned}$$

By (Dann et al., 2017)[Lemma F.4] and Equation (15) we have, $\Pr\left[\exists k : n_h^k(s,a) < \frac{m\epsilon}{4HA} \sum_{j=1}^{k-1} q_h^{\pi^j}(s) - \log \frac{6HSA}{\delta}\right] \leq \frac{\delta}{6HSA}$. By taking the union bound over all $h \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we get $\Pr(\neg E^{n1}) \leq \delta/6$.

$\Pr[\neg E^{n2}] \leq \delta/6$: For any $v \in [m]$,

$$n_h^k(s,a) = \sum_{j=1}^{k-1} \mathbb{I}\{\exists v' \in [m] : s_h^{j,v'} = s, a_h^{j,v'} = a\} \geq \sum_{j=1}^{k-1} \mathbb{I}\{s_h^{j,v} = s, a_h^{j,v} = a\}.$$

Again, by (Dann et al., 2017)[Lemma F.4], we get $\Pr\left[\exists k \in [K] : n_h^k(s,a) < \frac{1}{2} \sum_{j=1}^{k-1} q_h^{\pi^{j,v}}(s,a) - \log \frac{6mHSA}{\delta}\right] \leq \frac{\delta}{6mHSA}$. Taking the union bound we get $\Pr[\neg E^{n2}] \leq \delta/6$.

$\Pr[\neg E^{\epsilon}] \leq \delta/6$: Directly from Hoeffding's inequality and a union bound. □

C.2. Proof of Theorem C.1

Proof of Theorem C.1. By Lemma C.5, the good event holds with probability $1-\delta$. We now analyze the regret under the assumption that the good event holds. We start by decomposing the regret according to the policy played by agent v:

$$
\begin{aligned}
R_K &= \sum_{k=1}^{K} V_1^{\pi^{k,v}}(s_1^{k,v}) - V_1^{\star}(s_1^{k,v}) \\
&= \sum_{k=1}^{K} \mathbb{I}\{\pi^{k,v} = \underline{\pi}^k\} \left(V_1^{\underline{\pi}^k}(s_1^{k,v}) - V_1^{\star}(s_1^{k,v}) \right) + \sum_{k=1}^{K}\sum_{h'=1}^{H} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \left(V_1^{\pi^{k,h'}}(s_1^{k,v}) - V_1^{\star}(s_1^{k,v}) \right) \\
&\le \sum_{k=1}^{K} V_1^{\underline{\pi}^k}(s_1^{k,v}) - V_1^{\star}(s_1^{k,v}) + \sum_{k=1}^{K}\sum_{h'=1}^{H} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \left(V_1^{\pi^{k,h'}}(s_1^{k,v}) - V_1^{\star}(s_1^{k,v}) \right).
\end{aligned}
$$

For the first term we use Lemma C.12, then Lemma C.7 and then Lemma C.6:

$$
\begin{aligned}
\sum_{k=1}^{K} V_1^{\underline{\pi}^k}(s_1^{k,v}) - V_1^{\star}(s_1^{k,v}) &\lesssim H \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}^k \left[\frac{\sqrt{\tau \mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} + \frac{H^2 S\tau}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \underline{\pi}^k \right] \\
&\lesssim H\sqrt{\tau} \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}^k \left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \underline{\pi}^k \right] + \frac{H^5 S^2 A^2 \tau^2}{m\epsilon} \\
&\lesssim \sqrt{\frac{H^6 SAK\tau^2}{m\epsilon}} + \frac{H^5 S^2 A^2 \tau^2}{m\epsilon}.
\end{aligned}
$$

For the second term we use Lemma C.11, then Lemmas C.7 and C.8 and then Lemmas C.9 and C.10:

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h'=1}^{H} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \left(V_1^{\pi^{k,h'}}(s_1^{k,v}) - V_1^{\star}(s_1^{k,v}) \right) \lesssim \\
&\quad \lesssim H \sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \mathbb{E}^k \left[\frac{\sqrt{\tau \mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} + \frac{H^2 S\tau}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \pi^{k,h'} \right] \\
&\qquad + H^2 \sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \mathbb{E}^k \left[\frac{\sqrt{\tau \mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} + \frac{H^2 S\tau}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \underline{\pi}^k \right] \\
&\quad \lesssim H\sqrt{\tau} \sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \mathbb{E}^k \left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \pi^{k,h'} \right] + H^5 S^2 A \tau^2 \\
&\qquad + H^2\sqrt{\tau} \sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H} \mathbb{I}\{\pi^{k,v} = \pi^{k,h'}\} \mathbb{E}^k \left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \underline{\pi}^k \right] + \frac{H^6 S^2 A^2 \tau^2}{m\epsilon} \\
&\quad \lesssim \sqrt{H^7 SAK\epsilon\tau^2} + \tau\sqrt{H^7 SA\tau^3} K^{1/4} + H^5 S^2 A \tau^2 + \sqrt{\frac{H^8 SAK\tau}{m}} + \sqrt{\frac{H^9 SA\tau^2}{m\epsilon}} K^{1/4} + \frac{H^6 S^2 A^2 \tau^2}{m\epsilon}.
\end{aligned}
$$

Setting $\epsilon = \min\left\{\frac{HA}{m}, \frac{1}{\sqrt{m}}\right\}$, we get:

$$\begin{aligned}
R_K &\lesssim \sqrt{\frac{H^6SAK\tau^2}{m\epsilon}} + \frac{H^6S^2A^2\tau^2}{m\epsilon} + \sqrt{H^7SAK\epsilon\tau^2} + \tau\sqrt{H^7SA\tau^3}K^{1/4} \\
&\quad + H^5S^2A\tau^2 + \sqrt{\frac{H^8SAK\tau}{m}} + \sqrt{\frac{H^9SA\tau^2}{m\epsilon}}K^{1/4} \\
&\lesssim \sqrt{H^5SK\tau^2} + \sqrt{\frac{H^7SAK\tau^2}{\sqrt{m}}} + \sqrt{\frac{H^8SAK\tau}{m}} + H^5S^2A\tau^2 + \frac{H^6S^2A^2\tau^2}{\sqrt{m}} \\
&\quad + \sqrt{H^7SA\tau^3}K^{1/4} + \sqrt{H^8S\tau^2}K^{1/4} + \sqrt{\frac{H^9SA\tau^2}{\sqrt{m}}}K^{1/4} \\
&\lesssim \sqrt{H^5SK\tau^2} + \sqrt{\frac{H^7SAK\tau^2}{\sqrt{m}}} + \sqrt{\frac{H^8SAK\tau}{m}} + H^5S^2A\tau^2 + \frac{H^6S^2A^2\tau^2}{\sqrt{m}},
\end{aligned}$$

where the last inequality follows because the $K^{1/4}$ terms are dominant only when K is small, and in these cases the constant terms are larger. □

C.3. Bounds on the cumulative bonuses

Lemma C.6. *Under the good event, if* $\frac{m\epsilon}{HA} \le 1$*,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \underline{\pi}^k\right] \lesssim \sqrt{\frac{H^4SAK\tau}{m\epsilon}} + \frac{H^3SA^2\tau}{m\epsilon}.$$

Proof. By the event E^{n1}, we have:

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \underline{\pi}^k\right] &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{q_h^{\pi^k}(s,a)\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s,a)\vee 1}} \\
&\le \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s,a)\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k})}}{\sqrt{(\frac{m\epsilon}{4HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)-\tau)\vee 1}} \\
&\le \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k:\frac{m\epsilon}{8HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)\le\tau}^{K} q_h^{\pi^k}(s,a)\underbrace{\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k})}}_{\le H} \\
&\quad + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k:\frac{m\epsilon}{8HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)>\tau}^{K}\frac{q_h^{\pi^k}(s,a)\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k})}}{\sqrt{(\frac{m\epsilon}{4HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)-\tau)\vee 1}} \\
&\lesssim \frac{H^3A^2S\tau}{m\epsilon} + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s,a)\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k})}}{\sqrt{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s))\vee 1}} \\
&\le \frac{H^3A^2S\tau}{m\epsilon} + \sqrt{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k=1}^{K}q_h^{\pi^k}(s,a)\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k})}\sqrt{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s,a)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s))\vee 1}} \\
&= \frac{H^3A^2S\tau}{m\epsilon} + \sqrt{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\sum_{k=1}^{K}q_h^{\pi^k}(s,a)\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k})}\sqrt{\sum_{s\in\mathcal{S}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s))\vee 1}},
\end{aligned}$$

where the last inequality is Cauchy–Schwarz inequality. Using the law of total variance (Cohen et al., 2021, Lemma B.14),

$$\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^{\pi^k}(s,a)\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k}) = \mathbb{E}\left[\sum_{h=1}^{H}\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\pi^k}) \mid \underline{\pi}^k\right]$$
$$= \mathbb{E}\left[\left(V_1^{\pi^k}(s_1^k) - \sum_h c_h(s_h^k,a_h^k)\right)^2 \mid \underline{\pi}^k\right] \le H^2. \tag{16}$$

Using Rosenberg et al. (2020)[Lemma B.18], since $\frac{m\epsilon}{HA} \le 1$,

$$\sum_{s\in\mathcal{S}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s))\vee 1} = \frac{HA}{m\epsilon}\sum_{s\in\mathcal{S}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\frac{m\epsilon}{HA}q_h^{\pi^k}(s)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s))\vee 1} \lesssim \frac{H^2SA\tau}{m\epsilon}. \tag{17}$$

Combining the last three inequalities completes the proof. □

Lemma C.7. *Under the good event,* $\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{1}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \underline{\pi}^k\right] \lesssim \frac{H^2SA^2\tau}{m\epsilon}$.

Proof. By the event E^{n1}, we have:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{1}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \underline{\pi}^k\right] = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a)\vee 1} \le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s,a)}{\left(\frac{m\epsilon}{4HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)-\tau\right)\vee 1}$$
$$\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k:\frac{m\epsilon}{8HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)\le\tau} q_h^{\pi^k}(s,a) + \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k:\frac{m\epsilon}{8HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)>\tau}\frac{q_h^{\pi^k}(s,a)}{\left(\frac{m\epsilon}{4HA}\sum_{j=1}^{k-1}q_h^{\pi^j}(s)-\tau\right)\vee 1}$$
$$\lesssim \frac{H^2SA^2\tau}{m\epsilon} + \frac{HA}{m\epsilon}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{q_h^{\pi^k}(s,a)}{\left(\sum_{j=1}^{k-1}q_h^{\pi^j}(s)\right)\vee 1} \lesssim \frac{H^2SA^2\tau}{m\epsilon},$$

where the last inequality follows from Rosenberg et al. (2020)[Lemma B.18]. □

Lemma C.8. *Let $h' \in [H]$. Under the good event,* $\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\mathbb{E}^k\left[\frac{1}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \pi^{k,h'}\right] \lesssim HSA\tau$.

Proof. By the event E^{n2}, we have:

$$\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{1}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \pi^{k,h'}\right] = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\frac{q_h^{\pi^{k,h'}}(s,a)}{n_h^k(s,a)\vee 1}$$
$$\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\frac{q_h^{\pi^{k,h'}}(s,a)}{\left(\frac{1}{2}\sum_{j=1}^{k-1}q_h^{\pi^{j,v}}(s,a)-\tau\right)\vee 1}$$
$$\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}{\left(\frac{1}{2}\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)-\tau\right)\vee 1}$$
$$\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k:\frac{1}{4}\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)\le\tau}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)$$
$$+\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k:\frac{1}{4}\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)>\tau}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}{\left(\frac{1}{2}\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)-\tau\right)\vee 1}$$
$$\lesssim HSA\tau + \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}{\left(\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)\right)\vee 1} \lesssim HSA\tau,$$

where the last inequality follows (Rosenberg et al., 2020)[Lemma B.18]. □

Lemma C.9. *Let $h' \in [H]$. Under the good event,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\mathbb{E}^k\left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \pi^{k,h'}\right] \lesssim \sqrt{H^3SAK\epsilon\tau}+\tau\sqrt{H^3SAK}K^{1/4}+HSA\tau.$$

Proof. First we bound $\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\pi^k})}$ by H. Now, under the good event,

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{1}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \pi^{k,h'}\right] &= \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\frac{q_h^{\pi^{k,h'}}(s,a)}{\sqrt{n_h^k(s,a)\vee 1}}\\
&\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\frac{q_h^{\pi^{k,h'}}(s,a)}{\sqrt{\left(\frac{1}{2}\sum_{j=1}^{k-1}q_h^{\pi^{j,v}}(s,a)-\tau\right)\vee 1}}\\
&\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}{\sqrt{\left(\frac{1}{2}\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)-\tau\right)\vee 1}}\\
&\lesssim HSA\tau+\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}{\sqrt{\left(\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)\right)\vee 1}}\\
&\le \sqrt{\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}{\left(\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)\right)\vee 1}}\sqrt{\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}\\
&\quad+HSA\tau,
\end{aligned}$$

where the first inequality is by E^{n2}, the second inequality is done by breaking the sum to ks such that $\frac{1}{4}\sum_{j=1}^{k-1}q_h^{\pi^{j,v}}(s,a)\le\tau$ and the rest of the ks, as done in the proof of Lemma C.8 for example, and the last is Cauchy–Schwarz inequality. By Rosenberg et al. (2020)[Lemma B.18],

$$\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)}{\left(\sum_{j=1}^{k-1}\mathbb{I}\{\pi^{j,v}=\pi^{j,h'}\}q_h^{\pi^{j,h'}}(s,a)\right)\vee 1}\lesssim HSA\tau.$$

By the good event E^{ϵ},

$$\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\pi^{k,h'}}(s,a)=H\sum_{k=1}^{K}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\le K\epsilon+\sqrt{K\tau}.$$ □

Lemma C.10. *Under the good event,*

$$\sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\mathbb{E}^k\left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} \mid \underline{\pi}^k\right]\lesssim\sqrt{\frac{H^4SAK\tau}{m}}+\sqrt{\frac{H^5SA\tau^2}{m\epsilon}}K^{1/4}+\frac{H^3A^2S\tau}{m\epsilon}.$$

Proof. By the event E^{n1}, we have:

$$\sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\mathbb{E}^k\left[\frac{\sqrt{\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}}\mid\underline{\pi}^k\right]=$$

$$=\sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\frac{q_h^{\underline{\pi}^k}(s,a)\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}}$$

$$\le\sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\frac{q_h^{\underline{\pi}^k}(s,a)\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{(\frac{m\epsilon}{4HA}\sum_{j=1}^{k-1}q_h^{\underline{\pi}^j}(s)-\tau)\vee 1}}$$

$$\lesssim\frac{H^3A^2S\tau}{m\epsilon}+\sum_{k=1}^{K}\sum_{h'=1}^{H}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\frac{q_h^{\underline{\pi}^k}(s,a)\sqrt{\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\underline{\pi}^j}(s))\vee 1}}$$

$$\le\sqrt{\sum_{k,h',h,s,a}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\underline{\pi}^k}(s,a)\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\underline{\pi}^k})}\sqrt{\sum_{k,h',h,s,a}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\underline{\pi}^k}(s,a)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\underline{\pi}^j}(s))\vee 1}}+\frac{H^3A^2S\tau}{m\epsilon}$$

$$=\sqrt{\sum_{k,h'}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\sum_{h,s,a}q_h^{\underline{\pi}^k}(s,a)\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\underline{\pi}^k})}\sqrt{\sum_{k,h',h,s}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\underline{\pi}^k}(s)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\underline{\pi}^j}(s))\vee 1}}+\frac{H^3A^2S\tau}{m\epsilon},$$

where the last inequality is by Cauchy-Schwarz. By Equation (16) and the good event E^{ϵ},

$$\sum_{k=1}^{K}\sum_{h'=1}^{H}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}q_h^{\underline{\pi}^k}(s,a)\mathrm{Var}_{p_h(\cdot|s,a)}(V_{h+1}^{\underline{\pi}^k})\le H^2\sum_{k=1}^{K}\sum_{h'=1}^{H}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}$$
$$\le H^2K\epsilon+H^3\sqrt{K\log\frac{mH}{\delta'}}.$$

For last,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{h'=1}^{H}\frac{\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}q_h^{\underline{\pi}^k}(s)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\underline{\pi}^j}(s))\vee 1}\le\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\frac{q_h^{\underline{\pi}^k}(s)}{(\frac{m\epsilon}{HA}\sum_{j=1}^{k-1}q_h^{\underline{\pi}^j}(s))\vee 1}\lesssim\frac{H^2SA\tau}{m\epsilon},$$

where the first inequality is since $\sum_{h'=1}^{H}\mathbb{I}\{\pi^{k,v}=\pi^{k,h'}\}\le 1$ and the last is as in Equation (17). Combining the last three inequalities completes the proof. ☐

C.4. Bounding the regret of $\pi^{k,h'}$ and the optimistic policy

Lemma C.11. *Under the good event, for every $k\in[K]$ it holds that:*

$$V_1^{\pi^{k,h'}}(s_1^k)-V_1^{\star}(s_1^k)\lesssim H\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{\sqrt{\tau\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}}+\frac{H^2S\tau}{n_h^k(s_h^k,a_h^k)\vee 1}\mid\pi^{k,h'}\right]$$
$$+H^2\sum_{h=1}^{H}\mathbb{E}^k\left[\frac{\sqrt{\tau\mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}}+\frac{H^2S\tau}{n_h^k(s_h^k,a_h^k)\vee 1}\mid\underline{\pi}^k\right].$$

Proof. Apply Lemma C.13 for every $k \in [K]$ and then apply Lemmas C.16 to C.19 iteratively. □

Lemma C.12. *Under the good event, for every $k \in [K]$ it holds that:*

$$V_1^{\underline{\pi}^k}(s_1^k) - V_1^\star(s_1^k) \lesssim H \sum_{h=1}^{H} \mathbb{E}^k \left[\frac{\sqrt{\tau \mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\underline{\pi}^k})}}{\sqrt{n_h^k(s_h^k,a_h^k) \vee 1}} + \frac{H^2 S \tau}{n_h^k(s_h^k,a_h^k) \vee 1} \mid \underline{\pi}^k \right].$$

Proof. Similar to the proof of Lemma C.11. □

Lemma C.13. *Let $h' \in [H]$. Under the good event, for every $k \in [K]$ it holds that:*

$$\begin{aligned} V_1^{\pi^{k,h'}}(s_1^k) - V_1^\star(s_1^k) \le \sum_{h=1}^{H} \mathbb{E}^k \left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h^k(s_h^k,a_h^k) \mid \pi^{k,h'} \right] \\ + \mathbb{E}^k \left[\overline{Q}_{h'}^k(s_{h'}^k, \underline{\pi}_{h'}^k(s_{h'}^k)) - \underline{Q}_{h'}^k(s_{h'}^k, \underline{\pi}_{h'}^k(s_{h'}^k)) \mid \pi^{k,h'} \right]. \end{aligned}$$

Proof. It holds that:

$$\begin{aligned} V_1^{\pi^{k,h'}}(s_1^k) - V_1^\star(s_1^k) = V_1^{\pi^{k,h'}}(s_1^k) - V_1^\star(s_1^k) &= \sum_h \mathbb{E}^k \left[\langle Q_h^\star(s_h^k,\cdot), \pi^{k,h'}(\cdot \mid s_h^k) - \pi_h^\star(\cdot \mid s_h^k) \rangle \mid \pi^{k,h'} \right] \\ &= \sum_h \mathbb{E}^k \left[Q_h^\star(s_h^k,a_h^k) - V_h^\star(s_h^k) \mid \pi^{k,h'} \right] \le \sum_h \mathbb{E}^k \left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{V}_h^k(s_h^k) \mid \pi^{k,h'} \right] \\ &= \sum_h \mathbb{E}^k \left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h^k(s_h^k,\underline{\pi}_h^k(s_h^k)) \mid \pi^{k,h'} \right] \\ &= \sum_{h \ne h'} \mathbb{E}^k \left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h^k(s_h^k,\underline{\pi}_h^k(s_h^k)) \mid \pi^{k,h'} \right] \\ &\quad + \mathbb{E}^k \left[\overline{Q}_{h'}^k(s_{h'}^k,a_{h'}^k) - \underline{Q}_{h'}^k(s_{h'}^k,\underline{\pi}_{h'}^k(s_{h'}^k)) \mid \pi^{k,h'} \right] \\ &= \sum_{h \ne h'} \mathbb{E}^k \left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h^k(s_h^k,a_h^k) \mid \pi^{k,h'} \right] \\ &\quad + \mathbb{E}^k \left[\overline{Q}_{h'}^k(s_{h'}^k,a_{h'}^k) - \underline{Q}_{h'}^k(s_{h'}^k,\underline{\pi}_{h'}^k(s_{h'}^k)) \mid \pi^{k,h'} \right] \\ &\le \sum_h \mathbb{E}^k \left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h^k(s_h^k,a_h^k) \mid \pi^{k,h'} \right] \\ &\quad + \mathbb{E}^k \left[\overline{Q}_{h'}^k(s_{h'}^k,\underline{\pi}_{h'}^k(s_{h'}^k)) - \underline{Q}_{h'}^k(s_{h'}^k,\underline{\pi}_{h'}^k(s_{h'}^k)) \mid \pi^{k,h'} \right]. \end{aligned}$$

The first inequality is by Lemma C.3, the last equality is since $\pi^{k,h'} = \underline{\pi}^k$ for any $h \ne h'$ and the last inequality is by Lemma C.15. □

Lemma C.14. *Under the good event, for every $k \in [K]$ it holds that:*

$$V_1^{\underline{\pi}^k}(s_1^k) - V_1^\star(s_1^k) \le \sum_{h=1}^{H} \mathbb{E}^k \left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h^k(s_h^k,a_h^k) \mid \underline{\pi}^k \right].$$

Proof. Similar to the proof of Lemma C.13. □

C.5. Auxiliary lemmas

Lemma C.15. *If $a, a' \in \mathcal{A}_h^k(s)$ then,*

$$\overline{Q}_h^k(s,a) - \underline{Q}_h^k(s_h^k,a') \le \overline{Q}_h^k(s,a) - \underline{Q}_h^k(s,a) + \overline{Q}_h^k(s,a') - \underline{Q}_h^k(s,a').$$

Proof. Since $a, a' \in \mathcal{A}_h^k(s)$, we have that $\underline{Q}_h^k(s,a) \le \overline{Q}_h^k(s,a')$. Thus:

$$\overline{Q}_h^k(s,a) - \underline{Q}_h^k(s,a') = \overline{Q}_h^k(s,a) - \underline{Q}_h^k(s,a) + \overline{Q}_h^k(s,a') - \underline{Q}_h^k(s,a') + \underbrace{\underline{Q}_h^k(s,a) - \overline{Q}_h^k(s,a')}_{\le 0}. \qquad \square$$

Lemma C.16 (Recursion with Optimistic Next-Action)**.** *Let $h \neq h' - 1$. Under the good event, for every $k \in [K]$ it holds that:*

$$\mathbb{E}^k\left[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h(s_h^k,a_h^k) \mid \pi^{k,h'}\right] \le \mathbb{E}^k\left[\frac{8\sqrt{\tau \mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^k,a_h^k) \vee 1}} + \frac{118H^2S\tau}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \pi^{k,h'}\right] + \left(1+\frac{1}{4H}\right)\mathbb{E}^k\left[\overline{Q}_{h+1}^k(s_{h+1}^k,a_{h+1}^k) - \underline{Q}_{h+1}^k(s_{h+1}^k,a_{h+1}^k) \mid \pi^{k,h'}\right].$$

Proof. By definition of the optimistic and pessimistic Q-functions, we have:

$$\begin{aligned}
&\mathbb{E}^k[\overline{Q}_h^k(s_h^k,a_h^k) - \underline{Q}_h(s_h^k,a_h^k) \mid \pi^{k,h'}] = \\
&\quad = \mathbb{E}^k[2b_h^k(s_h^k,a_h^k;c) + 2b_h^k(s_h^k,a_h^k;p) \mid \pi^{k,h'}] \\
&\qquad + \mathbb{E}^k\left[\mathbb{E}_{\hat{p}_h^k(\cdot|s_h^k,a_h^k)}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k] \mid \pi^{k,h'}\right] \\
&\quad \le \mathbb{E}^k[2b_h^k(s_h^k,a_h^k;c) + 2b_h^k(s_h^k,a_h^k;p) \mid \pi^{k,h'}] \\
&\qquad + \mathbb{E}^k\left[\frac{18H^2S\tau}{n_h^k(s_h^k,a_h^k)\vee 1} + \left(1+\frac{1}{16H}\right)\mathbb{E}_{p_h(\cdot|s_h^k,a_h^k)}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k] \mid \pi^{k,h'}\right] \\
&\quad = \mathbb{E}^k\left[2b_h^k(s_h^k,a_h^k;c) + 2b_h^k(s_h^k,a_h^k;p) + \frac{18H^2S\tau}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \pi^{k,h'}\right] \\
&\qquad + \left(1+\frac{1}{16H}\right)\mathbb{E}^k\left[\mathbb{E}_{p_h(\cdot|s_h^k,a_h^k)}[\overline{V}_{h+1}^k(s_{h+1}^k) - \underline{V}_{h+1}^k(s_{h+1}^k)] \mid \pi^{k,h'}\right] \\
&\quad \le \mathbb{E}^k\left[\frac{8\sqrt{\tau \mathrm{Var}_{p_h(\cdot|s_h^k,a_h^k)}(V_{h+1}^{\pi^k})}}{\sqrt{n_h^k(s_h^k,a_h^k)\vee 1}} + \frac{118H^2S\tau}{n_h^k(s_h^k,a_h^k)\vee 1} \mid \pi^{k,h'}\right] \\
&\qquad + \left(1+\frac{1}{4H}\right)\mathbb{E}^k\left[\mathbb{E}_{p_h(\cdot|s_h^k,a_h^k)}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k] \mid \pi^{k,h'}\right]
\end{aligned}$$

where the first inequality is by Cohen et al. (2021, Lemma B.13), and the second one is by Cohen et al. (2021, Lemma B.6). Let $\overline{\pi}_h^k(s) = \arg\min_a \overline{Q}_h^k(s,a)$. For the last term we have:

$$\begin{aligned}
&\mathbb{E}^k\left[\mathbb{E}_{p_h(\cdot|s_h^k,a_h^k)}[\overline{V}_{h+1}^k - \underline{V}_{h+1}^k] \mid \pi^{k,h'}\right] = \\
&\quad = \sum_{h,s,a,s'} q_h^{\pi^{k,h'}}(s,a) p_h(s' \mid s,a)\left(\overline{V}_{h+1}^k(s') - \underline{V}_{h+1}^k(s')\right) \\
&\quad = \sum_{h,s} q_{h+1}^{\pi^{k,h'}}(s)\left(\overline{V}_{h+1}^k(s) - \underline{V}_{h+1}^k(s)\right) \\
&\quad = \mathbb{E}^k\left[\overline{V}_{h+1}^k(s_{h+1}^k) - \underline{V}_{h+1}^k(s_{h+1}^k) \mid \pi^{k,h'}\right] \\
&\quad = \mathbb{E}^k\left[\overline{Q}_{h+1}^k(s_{h+1}^k,\overline{\pi}_{h+1}^k(s_{h+1}^k)) - \underline{Q}_{h+1}^k(s_{h+1}^k,\underline{\pi}_{h+1}^k(s_{h+1}^k)) \mid \pi^{k,h'}\right] \\
&\quad \le \mathbb{E}^k\left[\overline{Q}_{h+1}^k(s_{h+1}^k,a_{h+1}^k) - \underline{Q}_{h+1}^k(s_{h+1}^k,\underline{\pi}_{h+1}^k(s_{h+1}^k)) \mid \pi^{k,h'}\right] \\
&\quad = \mathbb{E}^k\left[\overline{Q}_{h+1}^k(s_{h+1}^k,a_{h+1}^k) - \underline{Q}_{h+1}^k(s_{h+1}^k,a_{h+1}^k) \mid \pi^{k,h'}\right],
\end{aligned}$$

where the last equality follows because $h \neq h' - 1$. $\square$

Lemma C.17 (Recursion with Non-Optimistic Next-Action)**.** *Let $h' \in [H]$. Under the good event, for every $k \in [K]$ it holds that:*

$$\mathbb{E}^k\Big[\overline{Q}^k_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) - \underline{Q}_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) \mid \pi^{k,h'}\Big] \leq$$
$$\leq \mathbb{E}^k\left[\frac{8\sqrt{\tau \mathrm{Var}_{p_{h'-1}(\cdot|s^k_{h'-1},a^k_{h'-1})}(V^{\pi^k}_{h'})}}{\sqrt{n^k_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) \vee 1}} + \frac{118H^2S\tau}{n^k_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) \vee 1} \mid \pi^{k,h'}\right]$$
$$+ \left(1 + \frac{1}{4H}\right)\mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, a^k_{h'}) - \underline{Q}^k_{h'}(s^k_{h'}, a^k_{h'}) \mid \pi^{k,h'}\Big]$$
$$+ \left(1 + \frac{1}{4H}\right)\mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) - \underline{Q}^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \mid \pi^{k,h'}\Big].$$

Proof. Similarly to Lemma C.16, we have,

$$\mathbb{E}^k\left[\overline{Q}^k_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) - \underline{Q}_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) \mid \pi^{k,h'}\right] \leq$$
$$\leq \mathbb{E}^k\left[\frac{8\sqrt{\tau \mathrm{Var}_{p_{h'-1}(\cdot|s^k_{h'-1},a^k_{h'-1})}(V^{\pi^k}_{h'})}}{\sqrt{n^k_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) \vee 1}} + \frac{118H^2S\tau}{n^k_{h'-1}(s^k_{h'-1}, a^k_{h'-1}) \vee 1} \mid \pi^{k,h'}\right]$$
$$+ \left(1 + \frac{1}{4H}\right)\mathbb{E}^k\left[\overline{Q}^k_{h'}(s^k_{h'}, a^k_{h'}) - \underline{V}^k_{h'}(s^k_{h'}) \mid \pi^{k,h'}\right].$$

Now, by Lemma C.15,

$$\mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, a^k_{h'}) - \underline{V}^k_{h'}(s^k_{h'}) \mid \pi^{k,h'}\Big] = \mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, a^k_{h'}) - \underline{Q}^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \mid \pi^{k,h'}\Big]$$
$$\leq \mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, a^k_{h'}) - \underline{Q}^k_{h'}(s^k_{h'}, a^k_{h'}) \mid \pi^{k,h'}\Big]$$
$$+ \mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) - \underline{Q}^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \mid \pi^{k,h'}\Big]. \qquad \square$$

Lemma C.18 (Exploration Penalty Term Recursion)**.** *Let $h' \in [H]$. Under the good event, for every $k \in [K]$ it holds that:*

$$\mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) - \underline{Q}_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \mid \pi^{k,h'}\Big] \leq$$
$$\leq \mathbb{E}^k\left[\frac{8\sqrt{\tau \mathrm{Var}_{p_{h'}(\cdot|s^k_{h'},a^k_{h'})}(V^{\pi^k}_{h'+1})}}{\sqrt{n^k_{h'}(s^k_{h'}, a^k_{h'}) \vee 1}} + \frac{118H^2S\tau}{n^k_{h'}(s^k_{h'}, a^k_{h'}) \vee 1} \mid \underline{\pi}^k\right]$$
$$+ \left(1 + \frac{1}{4H}\right)\mathbb{E}^k\left[\overline{Q}^k_{h'+1}(s^k_{h'+1}, a^k_{h'+1}) - \underline{Q}^k_{h'+1}(s^k_{h'+1}, a^k_{h'+1}) \mid \underline{\pi}^k\right].$$

Proof. Again, similar to Lemma C.16,

$$\mathbb{E}^k\Big[\overline{Q}^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) - \underline{Q}_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \mid \pi^{k,h'}\Big] \leq$$
$$\leq \mathbb{E}^k\left[\frac{8\sqrt{\tau \mathrm{Var}_{p_{h'}(\cdot|s^k_{h'},\underline{\pi}^k_{h'}(s^k_{h'}))}(V^{\pi^k}_{h'+1})}}{\sqrt{n^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \vee 1}} + \frac{118H^2S\tau}{n^k_{h'}(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \vee 1} \mid \pi^{k,h'}\right]$$
$$+ \left(1 + \frac{1}{4H}\right)\mathbb{E}^k\left[\mathbb{E}_{p_{h'}(\cdot|s^k_{h'},\underline{\pi}^k_{h'}(s^k_{h'}))}[\overline{V}^k_{h+1} - \underline{V}^k_{h+1}] \mid \pi^{k,h'}\right]$$

Note that $q^{\pi^{k,h'}}_{h'}(s) = q^{\underline{\pi}^k}_{h'}(s)$ since until step h' the policies are the same, i.e., $\pi^{h',k}_h = \underline{\pi}^k_h$ for all $h < h'$. Hence, denoting the first term above by $\mathbb{E}^k[x(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \mid \pi^{k,h'}]$, we can write

$$\mathbb{E}^k[x(s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'})) \mid \pi^{k,h'}] = \sum_s q^{\pi^{k,h'}}_{h'}(s) x(s, \underline{\pi}^k_{h'}(s)) = \sum_s q^{\underline{\pi}^k}_{h'}(s) x(s, \underline{\pi}^k_{h'}(s))$$
$$= \sum_{s,a} q^{\underline{\pi}^k}_{h'}(s,a) x(s,a) = \mathbb{E}^k[x(s^k_{h'}, a^k_{h'}) \mid \underline{\pi}^k],$$

where the third equality is since $\underline{\pi}^k$ is deterministic. In a similar way, the second term can bounded by,

$$\mathbb{E}^k\left[\mathbb{E}_{p_{h'}(\cdot \mid s^k_{h'}, \underline{\pi}^k_{h'}(s^k_{h'}))}[\overline{V}^k_{h+1} - \underline{V}^k_{h+1}] \mid \pi^{k,h'}\right] =$$
$$= \sum_{s,s'} q^{\pi^{k,h'}}_{h'}(s) p_{h'}(\cdot \mid s^k_h, \underline{\pi}^k_{h'}(s^k_{h'})) \left(\overline{V}^k_{h'+1}(s') - \underline{V}^k_{h'+1}(s')\right)$$
$$= \sum_{s,s'} q^{\underline{\pi}^k}_{h'}(s) p_{h'}(\cdot \mid s^k_h, \underline{\pi}^k_{h'}(s^k_{h'})) \left(\overline{V}^k_{h'+1}(s') - \underline{V}^k_{h'+1}(s')\right)$$
$$= \sum_{s} q^{\underline{\pi}^k}_{h'+1}(s) \left(\overline{V}^k_{h'+1}(s') - \underline{V}^k_{h'+1}(s')\right)$$
$$= \mathbb{E}^k\left[\overline{V}^k_{h'+1}(s^k_{h'+1}) - \underline{V}^k_{h'+1}(s^k_{h'+1}) \mid \underline{\pi}^k\right]$$
$$\le \mathbb{E}^k\left[\overline{Q}^k_{h'+1}(s^k_{h'+1}, a^k_{h'+1}) - \underline{V}^k_{h'+1}(s^k_{h'+1}) \mid \underline{\pi}^k\right]$$
$$= \mathbb{E}^k\left[\overline{Q}^k_{h'+1}(s^k_{h'+1}, a^k_{h'+1}) - \underline{Q}^k_{h'+1}(s^k_{h'+1}, a^k_{h'+1}) \mid \underline{\pi}^k\right].$$

□

Lemma C.19 (Recursion Optimistic Policy). *Let $h \in [H]$. Under the good event, for every $k \in [K]$ it holds that:*

$$\mathbb{E}^k\left[\overline{Q}^k_h(s^k_h, a^k_h) - \underline{Q}_h(s^k_h, a^k_h) \mid \underline{\pi}^k\right] \le \mathbb{E}^k\left[\frac{8\sqrt{\tau \mathrm{Var}_{p_h(\cdot \mid s^k_h, a^k_h)}(V^{\underline{\pi}^k}_{h+1})}}{\sqrt{n^k_h(s^k_h, a^k_h) \vee 1}} + \frac{118 H^2 S \tau}{n^k_h(s^k_h, a^k_h) \vee 1} \mid \underline{\pi}^k\right]$$
$$+ \left(1 + \frac{1}{4H}\right) \mathbb{E}^k\left[\overline{Q}^k_{h+1}(s^k_{h+1}, a^k_{h+1}) - \underline{Q}^k_{h+1}(s^k_{h+1}, a^k_{h+1}) \mid \underline{\pi}^k\right].$$

Proof. Similar to the proof of Lemma C.18. □

Algorithm 6 COOPERATIVE O-REPS (COOP-O-REPS)

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, transition function p, number of episodes K, number of agents m, exploration parameter γ, learning rate η.
initialize: $\pi^1_h(a \mid s) = 1/A, q^1_h(s,a) = q^{\pi^1}_h(s,a)\ \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.
for $k = 1,\dots,K$ **do**
  for $v = 1,\dots,m$ **do**
    observe initial state $s^{k,v}_1$.
    for $h = 1,\dots,H$ **do**
      pick action $a^{k,v}_h \sim \pi^k_h(\cdot \mid s^{k,v}_h)$.
      suffer and observe cost $c^k_h(s^{k,v}_h, a^{k,v}_h)$.
      observe next state $s^{k,v}_{h+1} \sim p_h(\cdot \mid s^{k,v}_h, a^{k,v}_h)$.
    end for
  end for
  compute $W^k_h(s,a) = \Pr[\exists v:\ s^{k,v}_h = s, a^{k,v}_h = a \mid \pi^k] = 1 - \big(1 - q^k_h(s,a)\big)^m\ \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.
  compute $\hat{c}^k_h(s,a) = \frac{c^k_h(s,a)\mathbb{I}\{\exists v:\ s^{k,v}_h = s, a^{k,v}_h = a\}}{W^k_h(s,a)+\gamma}\ \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.
  compute $q^{k+1} = \arg\min_{q \in \Delta(\mathcal{M})} \eta\langle q, \hat{c}^k\rangle + \mathrm{KL}(q \,\|\, q^k)$.
  compute $\pi^{k+1}_h(a \mid s) = \frac{q^k_h(s,a)}{\sum_{a'\in\mathcal{A}} q^k_h(s,a')}\ \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.
end for

D. The coop-O-REPS algorithm for adversarial MDPs with fresh randomness and known p

For the setting of adversarial MDPs with fresh randomness and known transitions we propose the Cooperative O-REPS algorithm (coop-O-REPS; see Algorithm 6). The idea is simple: all the agents run the same O-REPS algorithm, but the estimated costs are updated based on the trajectories of all of them. Since the randomness is fresh in this setting, we expect the agents to observe m times more information. Next, we prove the following optimal regret bound for coop-O-REPS.

Similarly to Zimin & Neu (2013), We use the notations $\Delta(\mathcal{M})$ and $\mathrm{KL}(\cdot \,\|\, \cdot)$ for the set of occupancy measures of the MDP $\mathcal{M}$ and the KL-divergence between occupancy measures, respectively.

Theorem D.1. *With probability $1-\delta$, setting $\eta = \gamma = \sqrt{\frac{\log\frac{HSA}{\delta}}{(1+\frac{SA}{m})K}}$, the individual regret of each agent of* coop-O-REPS *is*

$$R_K = O\left(H\sqrt{K\log\frac{HSA}{\delta}} + \sqrt{\frac{H^2SAK}{m}\log\frac{HSA}{\delta}} + \frac{HSA}{m}\log\frac{HSA}{\delta}\right).$$

D.1. The good event

Define the following events:

$$E^c = \left\{\sum_{k=1}^K \langle \mathbb{E}^k[\hat{c}^k] - \hat{c}^k, q^k\rangle \le 4H\sqrt{K\log\frac{3}{\delta}}\right\}$$

$$E^{\hat{c}} = \left\{\sum_{k=1}^K\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\frac{1}{m} + q^k_h(s,a)\right)\left(\hat{c}^k_h(s,a) - 2c^k_h(s,a)\right) \le \frac{10HSA\log\frac{3HSA}{\delta}}{m\gamma} + \frac{10H\log\frac{3HSA}{\delta}}{\gamma}\right\}$$

$$E^\star = \left\{\sum_{k=1}^K \langle \hat{c}^k - c^k, q^{\pi^\star}\rangle \le \frac{H\log\frac{3HSA}{\delta}}{\gamma}\right\}$$

The good event is the intersection of the above events. The following lemma establishes that the good event holds with high probability.

Lemma D.2 (The Good Event)**.** *Let $\mathbb{G} = E^c \cap E^{\hat{c}} \cap E^\star$ be the good event. It holds that $\Pr[\mathbb{G}] \ge 1-\delta$.*

Proof. We show that each of the events $\neg E^c, \neg E^{\hat{c}}, \neg E^\star$ occur with probability at most $\delta/3$. Then, by a union bound we obtain the statement. Notice that:

$$\begin{aligned}\langle \hat{c}^k, q^k \rangle &= \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a)\hat{c}_h^k(s,a) \le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a)\frac{\mathbb{I}\{\exists v: s_h^{k,v}=s, a_h^{k,v}=a\}}{W_h^k(s,a)+\gamma} \\ &\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{q_h^k(s,a)}{1-\left(1-q_h^k(s,a)\right)^m}\mathbb{I}\{\exists v: s_h^{k,v}=s, a_h^{k,v}=a\} \\ &\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left(\frac{1}{m}+q_h^k(s,a)\right)\mathbb{I}\{\exists v: s_h^{k,v}=s, a_h^{k,v}=a\} \\ &\le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a) + \frac{1}{m}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathbb{I}\{\exists v: s_h^{k,v}=s, a_h^{k,v}=a\} \le H + \frac{1}{m}\cdot Hm = 2H,\end{aligned}$$

where the third inequality is by Lemma D.3, and the last inequality follows because for each step h the agents visit at most m state-action pairs. Thus, event E^c holds by Azuma inequality.

Event $E^{\hat{c}}$ holds by Cohen et al. (2021, Lemma E.2) since $\sum_{h,s,a}\left(\frac{1}{m}+q_h^k(s,a)\right)\hat{c}_h^k(s,a) \le \frac{1}{\gamma}\left(\frac{HSA}{m}+H\right)$ and $\mathbb{E}^k[\hat{c}_h^k(s,a)] \le c_h^k(s,a)$. Event $E^\star$ holds by Jin et al. (2020a, Lemma 14). □

D.2. Proof of Theorem D.1

Proof of Theorem D.1. By Lemma D.2, the good event holds with probability $1-\delta$. We now analyze the regret under the assumption that the good event holds. We start by decomposing the regret as follows:

$$\begin{aligned}R_K &= \sum_{k=1}^{K} V_1^{k,\pi^k}(s_1^{k,v}) - V_1^{k,\pi^\star}(s_1^{k,v}) = \sum_{k=1}^{K}\langle c^k, q^k - q^{\pi^\star}\rangle \\ &= \underbrace{\sum_{k=1}^{K}\langle c^k-\hat{c}^k, q^k\rangle}_{(A)} + \underbrace{\sum_{k=1}^{K}\langle \hat{c}^k, q^k - q^{\pi^\star}\rangle}_{(B)} + \underbrace{\sum_{k=1}^{K}\langle \hat{c}^k - c^k, q^{\pi^\star}\rangle}_{(C)}.\end{aligned}$$

Term (A) can be further decomposed as:

$$(A) = \sum_{k=1}^{K}\langle c^k-\hat{c}^k, q^k\rangle = \sum_{k=1}^{K}\langle c^k - \mathbb{E}^k[\hat{c}^k], q^k\rangle + \sum_{k=1}^{K}\langle \mathbb{E}^k[\hat{c}^k]-\hat{c}^k, q^k\rangle.$$

The second term is bounded by $4H\sqrt{K\log\frac{3}{\delta}}$ by the good event E^c, and for the first term:

$$\begin{aligned}\sum_{k=1}^{K}\langle c^k - \mathbb{E}^k[\hat{c}^k], q^k\rangle &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a)c_h^k(s,a)\left(1-\frac{\mathbb{E}^k[\mathbb{I}\{\exists v:\ s_h^{k,v}=s, a_h^{k,v}=a\}]}{W_h^k(s,a)+\gamma}\right) \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a)c_h^k(s,a)\left(1-\frac{W_h^k(s,a)}{W_h^k(s,a)+\gamma}\right) \le \gamma\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{q_h^k(s,a)}{W_h^k(s,a)+\gamma} \\ &\le \gamma\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{q_h^k(s,a)}{1-\left(1-q_h^k(s,a)\right)^m} \le \gamma\sum_{k,h,s,a}\left(\frac{1}{m}+q_h^k(s,a)\right) = \gamma HK\left(1+\frac{SA}{m}\right),\end{aligned}$$

where the last inequality is by Lemma D.3.

Term (B) is bounded by OMD (see, e.g., Zimin & Neu (2013)) as follows:

$$
\begin{aligned}
(B) = \sum_{k=1}^{K} \langle \hat{c}^k, q^k - q^{\pi^\star} \rangle &\leq \frac{H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^k(s,a) \hat{c}_h^k(s,a)^2 \\
&\leq \frac{H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^k(s,a) \frac{\hat{c}_h^k(s,a)}{W_h^k(s,a) + \gamma} \\
&\leq \frac{H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \hat{c}_h^k(s,a) \frac{q_h^k(s,a)}{1 - \left(1 - q_h^k(s,a)\right)^m} \\
&\leq \frac{H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left(\frac{1}{m} + q_h^k(s,a) \right) \hat{c}_h^k(s,a) \\
&\leq \frac{H \log(HSA)}{\eta} + 2\eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left(\frac{1}{m} + q_h^k(s,a) \right) c_h^k(s,a) + \frac{10\eta HSA \log \frac{3HSA}{\delta}}{m\gamma} + \frac{10\eta H \log \frac{3HSA}{\delta}}{\gamma} \\
&\lesssim \frac{H \log \frac{HSA}{\delta}}{\eta} + \frac{\eta HSAK}{m} + \eta HK + \frac{\eta HSA \log \frac{3HSA}{\delta}}{m\gamma} + \frac{\eta H \log \frac{3HSA}{\delta}}{\gamma},
\end{aligned}
$$

where the forth inequality is by Lemma D.3, and the fifth inequality is by the good event $E^{\hat{c}}$.

Term (C) is bounded by $\frac{H \log \frac{3H}{\delta}}{\gamma}$ by the good event $E^\star$. Putting the three terms together gives the final regret bound when setting $\eta = \gamma = \sqrt{\frac{\log \frac{HSA}{\delta}}{(1 + \frac{SA}{m})K}}$. □

D.3. Auxiliary lemmas

Lemma D.3. *Let* $x \in (0,1)$. *Then,* $\frac{x}{1-(1-x)^m} \leq \frac{1}{m} + x$.

Proof. Using AM-GM inequaility,

$$
\begin{aligned}
&\left((1-x)^m (1+xm)\right)^{\frac{1}{m+1}} \leq \frac{m(1-x) + 1 + xm}{m+1} = 1 \\
\Longrightarrow &(1-x)^m \leq \frac{1}{1+xm} \\
\Longrightarrow &1 - (1-x)^m \geq \frac{xm}{1+xm} \\
\Longrightarrow &\frac{x}{1-(1-x)^m} \leq \frac{1}{m} + x.
\end{aligned}
$$

□

Algorithm 7 COOPERATIVE UOB-REPS (COOP-UOB-REPS)

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, confidence parameter δ, number of episodes K, number of agents m, exploration parameter γ, learning rate η.
initialize: $n^1_h(s,a) = 0, n^1_h(s,a,s') = 0, \pi^1_h(a \mid s) = 1/A, q^1_h(s,a,s') = 1/S^2A \; \forall(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$.
for $k = 1, \ldots, K$ **do**
  set $n^{k+1}_h(s,a) \leftarrow n^k_h(s,a), n^{k+1}_h(s,a,s') \leftarrow n^k_h(s,a,s') \; \forall(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$.
  for $v = 1, \ldots, m$ **do**
    observe initial state $s^{k,v}_1$.
    for $h = 1, \ldots, H$ **do**
      pick action $a^{k,v}_h \sim \pi^k_h(\cdot \mid s^{k,v}_h)$.
      suffer and observe cost $c^k_h(s^{k,v}_h, a^{k,v}_h)$.
      observe next state $s^{k,v}_{h+1} \sim p_h(\cdot \mid s^{k,v}_h, a^{k,v}_h)$.
      update $n^{k+1}_h(s^{k,v}_h, a^{k,v}_h) \leftarrow n^{k+1}_h(s^{k,v}_h, a^{k,v}_h) + 1, n^{k+1}_h(s^{k,v}_h, a^{k,v}_h, s^{k,v}_{h+1}) \leftarrow n^{k+1}_h(s^{k,v}_h, a^{k,v}_h, s^{k,v}_{h+1}) + 1$.
    end for
  end for
  set $\hat{p}^{k+1}_h(s' \mid s,a) \leftarrow \frac{n^{k+1}_h(s,a,s')}{n^{k+1}_h(s,a)\vee 1} \; \forall(s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$.
  compute confidence set for $\epsilon^{k+1}_h(s' \mid s,a) = 4\sqrt{\frac{\hat{p}^{k+1}_h(s'|s,a)\ln\frac{HSAK}{4\delta}}{n^{k+1}_h(s,a)\vee 1}} + 10\frac{\ln\frac{HSAK}{4\delta}}{n^{k+1}_h(s,a)\vee 1}$:

$$\mathcal{P}^{k+1} = \left\{p' \mid \forall(s,a,s',h) : |\hat{p}^{k+1}_h(s' \mid s,a) - p'_h(s' \mid s,a)| \le \epsilon^{k+1}_h(s' \mid s,a)\right\}.$$

  compute $u^k_h(s) = \max_{p'\in\mathcal{P}^k} q^{p',\pi^k}_h(s) = \max_{p'\in\mathcal{P}^k} \Pr[s_h = s \mid \pi^k, p'] \; \forall s \in \mathcal{S}$.
  compute $U^k_h(s,a) = 1 - \left(1 - u^k_h(s,a)\right)^m \; \forall(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.
  compute $\hat{c}^k_h(s,a) = \frac{c^k_h(s,a)\mathbb{I}\{\exists v:\, s^{k,v}_h = s, a^{k,v}_h = a\}}{U^k_h(s,a)+\gamma} \; \forall(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.
  compute $q^{k+1} = \arg\min_{q\in\Delta(\mathcal{M},k+1)} \eta\langle q, \hat{c}^k\rangle + \text{KL}(q \,\|\, q^k)$.
  compute $\pi^{k+1}_h(a \mid s) = \frac{q^{k+1}_h(s,a)}{\sum_{a'\in\mathcal{A}} q^{k+1}_h(s,a')} \; \forall(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, where $q^{k+1}_h(s,a) = \sum_{s'\in\mathcal{S}} q^{k+1}_h(s,a,s')$.
end for

E. The `coop-UOB-REPS` algorithm for adversarial MDPs with fresh randomness and unknown p

For the setting of adversarial MDPs with fresh randomness and unknown transitions we propose the Cooperative UOB-REPS algorithm (`coop-UOB-REPS`; see Algorithm 7). The idea is simple: all the agents run the same UOB-REPS algorithm, but the estimated costs and transitions are updated based on the trajectories of all of them. Since the randomness is fresh in this setting, we expect the agents to observe m times more information. Next, we prove the following regret bound for `coop-UOB-REPS`. Note that this bound is optimal up to a $\sqrt{HS}$ factor. Removing this extra $\sqrt{HS}$ is an open-problem even for adversarial MDPs with a single agent.

Similarly to Rosenberg & Mansour (2019a), We use the notation $\Delta(\mathcal{M}, k)$ for the set of occupancy measures whose induced transition function is within the confidence set $\mathcal{P}^k$.

Theorem E.1. *With probability $1-\delta$, setting $\eta = \gamma = \sqrt{\frac{\log\frac{mKHSA}{\delta}}{\left(1+\frac{SA}{m}\right)K}}$, the individual regret of each agent of* `coop-UOB-REPS` *is*

$$R_K = O\left(H\sqrt{K\log\frac{mKHSA}{\delta}} + \sqrt{\frac{H^4S^2AK}{m}\log^3\frac{mKHSA}{\delta}} + H^3S^3A\log^3\frac{mKHSA}{\delta}\right).$$

E.1. The good event

Denote $\epsilon_h^k(s' \mid s,a) = \sqrt{\frac{2\hat{p}_h^k(s'|s,a)\log\frac{30KHSA}{\delta}}{n_h^k(s,a)\vee 1}} + \frac{2\log\frac{30KHSA}{\delta}}{n_h^k(s,a)\vee 1}$, $\epsilon_h^k(s,a) = \sum_{s'\in\mathcal{S}} \epsilon_h^k(s' \mid s,a)$, $\tilde{\epsilon}_h^k(s' \mid s,a) = 8\sqrt{\frac{p_h(s'|s,a)\log\frac{30KHSA}{\delta}}{n_h^k(s,a)\vee 1}} + \frac{100\log\frac{30KHSA}{\delta}}{n_h^k(s,a)\vee 1}$, and $\tilde{\epsilon}_h^k(s,a) = \sum_{s'} \tilde{\epsilon}_h^k(s' \mid s,a)$. Define the following events:

$$E^p = \left\{\forall (k,s,a,s',h) : \ |p_h(s'|s,a) - \hat{p}_h^k(s'|s,a)| \le \epsilon_h^k(s' \mid s,a)\right\}$$

$$E^{on1} = \left\{\forall v \in [m] : \sum_{k,h,s,a} \left(q_h^{\pi^k}(s,a) - \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}\right)\min\{2, \epsilon_h^k(s,a)\} \le 10\sqrt{K\log\frac{30KHSAm}{\delta}}\right\}$$

$$E^{on2} = \left\{\forall v \in [m] : \sum_{k,h,s,a} q_h^{\pi^k}(s,a)\tilde{\epsilon}_h^k(s,a) \le 2\sum_{k,h,s,a} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}\tilde{\epsilon}_h^k(s,a) + 100HS\log^2\frac{30KHSAm}{\delta}\right\}$$

$$E^{on3} = \left\{\forall v \in [m] : \sum_{k,s,a,h} \frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a)} \le 2\sum_{k,s,a,h} \frac{\mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a)} + H\log\frac{m}{\delta}\right\}$$

$$E^c = \left\{\sum_{k=1}^K \langle \mathbb{E}^k[\hat{c}^k] - \hat{c}^k, q^k\rangle \le 4H\sqrt{K\log\frac{6}{\delta}}\right\}$$

$$E^{\hat{c}} = \left\{\sum_{k=1}^K\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left(\frac{1}{m} + q_h^k(s,a)\right)\left(\hat{c}_h^k(s,a) - c_h^k(s,a)\right) \le \frac{HSA\log\frac{3HSA}{\delta}}{\gamma}\right\}$$

$$E^\star = \left\{\sum_{k=1}^K \langle \hat{c}^k - c^k, q^{\pi^\star}\rangle \le \frac{H\log\frac{3HSA}{\delta}}{\gamma}\right\}$$

The good event is the intersection of the above events. The following lemma establishes that the good event holds with high probability.

Lemma E.2 (The Good Event)**.** *Let $\mathbb{G} = E^p \cap E^{on1} \cap E^{on2} \cap E^{on3} \cap E^c \cap E^{\hat{c}} \cap E^\star$ be the good event. It holds that $\Pr[\mathbb{G}] \ge 1-\delta$.*

Proof. E^{on2} and E^{on3} follows Cohen et al. (2021, Lemma E.2). The rest are similar to the proofs of Lemmas B.4 and D.2 and to proofs in Jin et al. (2020a). □

E.2. Proof of Theorem E.1

Proof of Theorem E.1. By Lemma E.2, the good event holds with probability $1-\delta$. We now analyze the regret under the assumption that the good event holds. We start by decomposing the regret as follows:

$$R_K = \sum_{k=1}^K V_1^{k,\pi^k}(s_1^{k,v}) - V_1^{k,\pi^\star}(s_1^{k,v}) = \sum_{k=1}^K \langle c^k, q^{\pi^k} - q^{\pi^\star}\rangle$$
$$= \underbrace{\sum_{k=1}^K \langle c^k, q^{\pi^k} - q^k\rangle}_{(A)} + \underbrace{\sum_{k=1}^K \langle c^k - \hat{c}^k, q^k\rangle}_{(B)} + \underbrace{\sum_{k=1}^K \langle \hat{c}^k, q^k - q^{\pi^\star}\rangle}_{(C)} + \underbrace{\sum_{k=1}^K \langle \hat{c}^k - c^k, q^{\pi^\star}\rangle}_{(D)}.$$

Let $\tau = \log\frac{KHSAm}{\delta}$ be a logarithmic term. Term (A) can be decomposed using the value difference lemma (see, e.g., Shani et al. (2020)):

$$
\begin{aligned}
(A) = \sum_{k=1}^{K} \langle c^k, q^{\pi^k} - q^k \rangle &\leq 2H \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^{\pi^k}(s,a) \| p_h(\cdot \mid s,a) - \hat{p}_h^k(\cdot \mid s,a) \|_1 \\
&\leq 2H \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^{\pi^k}(s,a) \min\{2, \epsilon_h^k(s,a)\} \\
&\lesssim \frac{H}{m} \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\} \epsilon_h^k(s,a) + H\sqrt{K\tau} \\
&\lesssim \frac{H\sqrt{S\tau}}{m} \sum_{k,h,s,a} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sqrt{n_h^k(s,a) \vee 1}} + \frac{HS\tau}{m} \sum_{k,h,s,a} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a) \vee 1} + H\sqrt{K\tau},
\end{aligned}
$$

where the second inequality is by event E^p, and the third inequality uses event E^{on1} and Cauchy–Schwarz inequality. We now bound each of the two sums separately. For the second sum recall that $n_h^k(s,a) = \sum_{j=1}^{k-1} \sum_{v=1}^{m} \mathbb{I}\{s_h^{j,v} = s, a_h^{j,v} = a\}$, thus we have:

$$
\begin{aligned}
\sum_{k,h,s,a} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a) \vee 1} &\leq 2HSAm + \sum_{h,s,a} \sum_{k: n_h^k(s,a) \geq m} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{n_h^k(s,a)} \\
&= 2HSAm + \sum_{h,s,a} \sum_{k: n_h^k(s,a) \geq m} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sum_{j=1}^{k-1} \sum_{v=1}^{m} \mathbb{I}\{s_h^{j,v} = s, a_h^{j,v} = a\}} \\
&\leq 2HSAm + HSA \log(Km), \qquad (18)
\end{aligned}
$$

where the last inequality is by Rosenberg et al. (2020, Lemma B.18). For the first term:

$$
\begin{aligned}
\sum_{k,h,s,a} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sqrt{n_h^k(s,a) \vee 1}} &\leq 2HSAm + \sum_{h,s,a} \sum_{k: n_h^k(s,a) \geq m} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sqrt{n_h^k(s,a)}} \\
&= 2HSAm + \sum_{h,s,a} \sum_{k: n_h^k(s,a) \geq m} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sqrt{n_h^{k+1}(s,a)}} \sqrt{\frac{n_h^{k+1}(s,a)}{n_h^k(s,a)}} \\
&\leq 2HSAm + \sum_{h,s,a} \sum_{k: n_h^k(s,a) \geq m} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sqrt{n_h^{k+1}(s,a)}} \sqrt{\frac{n_h^k(s,a) + m}{n_h^k(s,a)}} \\
&\leq 2HSAm + 2 \sum_{h,s,a} \sum_{k: n_h^k(s,a) \geq m} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sqrt{n_h^{k+1}(s,a)}} \\
&= 2HSAm + 2 \sum_{h,s,a} \sum_{k: n_h^k(s,a) \geq m} \frac{\sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}}{\sqrt{\sum_{j=1}^{k} \sum_{v=1}^{m} \mathbb{I}\{s_h^{j,v} = s, a_h^{j,v} = a\}}} \\
&\leq 2HSAm + 4 \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sqrt{\sum_{k=1}^{K} \sum_{v=1}^{m} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}} \\
&\leq 2HSAm + 4 \sum_{h=1}^{H} \sqrt{SA \sum_{k=1}^{K} \sum_{v=1}^{m} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{I}\{s_h^{k,v} = s, a_h^{k,v} = a\}} \\
&= 2HSAm + 4H\sqrt{SAKm}, \qquad (19)
\end{aligned}
$$

where the forth inequality is by Streeter & McMahan (2010, Lemma 1), and the last inequality is by Jensen's inequality. Putting these together we get that: $(A) \lesssim \sqrt{\frac{H^4 S^2 A K \tau}{m}} + H^2 S^2 A \tau^2 + H\sqrt{K\tau}$.

Term (B) can be further decomposed as:

$$(B) = \sum_{k=1}^{K} \langle c^k - \hat{c}^k, q^k \rangle = \sum_{k=1}^{K} \langle c^k - \mathbb{E}^k[\hat{c}^k], q^k \rangle + \sum_{k=1}^{K} \langle \mathbb{E}^k[\hat{c}^k] - \hat{c}^k, q^k \rangle.$$

The second term is bounded by $4H\sqrt{K \log \frac{6}{\delta}}$ by the good event E^c. For the first term, let $W_h^{\pi^k}(s,a) = \Pr[\exists v : s_h^{k,v} = s, a_h^{k,v} = a \mid \pi^k] = 1 - (1 - q_h^{\pi^k}(s,a))^m$:

$$\begin{aligned}
\sum_{k=1}^{K} \langle c^k - \mathbb{E}^k[\hat{c}^k], q^k \rangle &= \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^k(s,a) c_h^k(s,a) \left(1 - \frac{\mathbb{E}^k[\mathbb{I}\{\exists v :\ s_h^{k,v} = s, a_h^{k,v} = a\}]}{U_h^k(s,a) + \gamma}\right) \\
&= \sum_{k,h,s,a} q_h^k(s,a) c_h^k(s,a) \left(1 - \frac{W_h^{\pi^k}(s,a)}{U_h^k(s,a) + \gamma}\right) \le \sum_{k,h,s,a} \frac{q_h^k(s,a)}{W_h^{\pi^k}(s,a)} (U_h^k(s,a) - W_h^{\pi^k}(s,a) + \gamma) \\
&\le \sum_{k,h,s,a} \left(\frac{1}{m} + q_h^k(s,a)\right) (U_h^k(s,a) - W_h^{\pi^k}(s,a) + \gamma) \\
&= \sum_{k,h,s,a} \left(\frac{1}{m} + q_h^k(s,a)\right) \Big((1 - q_h^{\pi^k}(s,a))^m - (1 - u_h^k(s,a))^m\Big) + \frac{\gamma HSAK}{m} + \gamma HK \\
&\le \sum_{k,h,s,a} \left(\frac{1}{m} + q_h^k(s,a)\right) m(1 - q_h^{\pi^k}(s,a))^m (u_h^k(s,a) - q_h^{\pi^k}(s,a)) + \frac{\gamma HSAK}{m} + \gamma HK \\
&\le \sum_{k,h,s,a} u_h^k(s,a) - q_h^{\pi^k}(s,a) + \frac{\gamma HSAK}{m} + \gamma HK \\
&\quad + \sum_{k,h,s,a} m q_h^k(s,a)(1 - q_h^{\pi^k}(s,a))^m (u_h^k(s,a) - q_h^{\pi^k}(s,a)) \\
&\le 3 \sum_{k,h,s,a} (u_h^k(s,a) - q_h^{\pi^k}(s,a)) \log m + \frac{\gamma HSAK}{m} + \gamma HK,
\end{aligned}$$

where the second inequality is by Lemma D.3, the third inequality is by convexity of the function $f(x) = (1-x)^m$ for $x \in [0,1]$, and the last inequality follows because $mx(1-x)^m \le \log m$ for every $x \in [0,1]$ since if $1 - x \le 1 - \frac{2 \log m}{m}$ then $(1-x)^m \le \frac{1}{m^2}$; otherwise $x \le \frac{2 \log m}{m}$. Finally, $\sum_{k,h,s,a} (u_h^k(s,a) - q_h^{\pi^k}(s,a))$ is bounded by Lemma E.5.

Term (C) is bounded by OMD (see, e.g., Rosenberg & Mansour (2019a)) as follows:

$$\begin{aligned}
(C) = \sum_{k=1}^{K} \langle \hat{c}^k, q^k - q^{\pi^\star} \rangle &\le \frac{2H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^k(s,a) \hat{c}_h^k(s,a)^2 \\
&\le \frac{2H \log(HSA)}{\eta} + \eta \sum_{k,h,s,a} q_h^k(s,a) \frac{\hat{c}_h^k(s,a)}{U_h^k(s,a) + \gamma} \le \frac{2H \log(HSA)}{\eta} + \eta \sum_{k,h,s,a} \hat{c}_h^k(s,a) \frac{q_h^k(s,a)}{1 - \big(1 - q_h^k(s,a)\big)^m} \\
&\le \frac{2H \log(HSA)}{\eta} + \eta \sum_{k,h,s,a} \left(\frac{1}{m} + q_h^k(s,a)\right) \hat{c}_h^k(s,a) \\
&\le \frac{2H \log(HSA)}{\eta} + \eta \sum_{k,h,s,a} \left(\frac{1}{m} + q_h^k(s,a)\right) c_h^k(s,a) + \frac{\eta HSA \log \frac{3HSA}{\delta}}{\gamma} \\
&\le \frac{2H \log \frac{HSA}{\delta}}{\eta} + \frac{\eta HSAK}{m} + \eta HK + \frac{\eta HSA \log \frac{3HSA}{\delta}}{\gamma},
\end{aligned}$$

where the forth inequality is by Lemma D.3, and the fifth inequality is by the good event $E^{\hat{c}}$.

Term (D) is bounded by $\frac{H \log \frac{3H}{\delta}}{\gamma}$ by the good event $E^\star$. Putting the three terms together gives the final regret bound when setting $\eta = \gamma = 1/\sqrt{\left(1 + \frac{SA}{m}\right) K}$. □

E.3. Auxiliary Lemmas

The following Lemma is by Jin et al. (2020a, Lemma 8), Cohen et al. (2021, Lemma B.13).

Lemma E.3. *Under the good event we have,*

$$\forall(k,s,a,s',h): \quad |p_h(s'|s,a) - \hat{p}_h^k(s'|s,a)| \leq \tilde{\epsilon}_h^k(s' \mid s,a).$$

The following Lemma is part of the proof of Jin et al. (2020a, Lemma 4). We provide the proof here for completeness.

Lemma E.4. *Let $q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid \tilde{s}';h)$ be the probability to visit $(\tilde{s},\tilde{a})$ in time $\tilde{h}$ given that we visited $\tilde{s}'$ in time h. Under the good event,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} |u_h^k(s,a) - q_h^{\pi^k}(s,a)| \lesssim H\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \tilde{\epsilon}_h^k(s,a) q_h^{\pi^k}(s,a) \tag{20}$$
$$+ HS\sum_{k=1}^{K}\sum_{1\leq h<\tilde{h}\leq H}\sum_{s\in\mathcal{S},a\in\mathcal{A},s'\in\mathcal{S}}\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \tilde{\epsilon}_h^k(s' \mid s,a) q_h^{\pi^k}(s,a) \min\left\{2, \sum_{\tilde{s}'\in\mathcal{S}} \tilde{\epsilon}_{\tilde{h}}^k(\tilde{s}' \mid \tilde{s},\tilde{a})\right\} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1).$$

Proof. Let $q^{k,s,h}$ be the occupancy measure such that $q_h^{k,s,h}(s) = u_h^k(s)$, and let $p^{k,s,h}$ be the transition that corresponds to $q^{k,s,h}$. Let $\sigma_h(s)$ be the set of all trajectories that end in s in time h, i.e., $\sigma_h(s) = \{s_1,a_1,\ldots,s_{h-1},a_{h-1},s_h\}$ where $s_h = s$. We have:

$$u_h^k(s,a) = q_h^{k,s,h}(s,a) = \pi_h^k(a \mid s)\sum_{\sigma_h(s)}\prod_{h'=1}^{h-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'},a_{h'})$$
$$q_h^{\pi^k}(s,a) = \pi_h^k(a \mid s)\sum_{\sigma_h(s)}\prod_{h'=1}^{h-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) p_{h'}(s_{h'+1} \mid s_{h'},a_{h'}).$$

Then,

$$|u_h^k(s,a) - q_h^{\pi^k}(s,a)| = \pi_h^k(a \mid s)\sum_{\sigma_h(s)}\prod_{h'=1}^{h-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) \left|\prod_{h'=1}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'},a_{h'}) - \prod_{h'=1}^{h-1} p_{h'}(s_{h'+1} \mid s_{h'},a_{h'})\right|.$$

We can rewrite the following term as,

$$
\begin{aligned}
&\left| \prod_{h'=1}^{h-1} p^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) - \prod_{h'=1}^{h-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \right| \\
&\qquad = \left| \sum_{l=2}^{h-1} \prod_{h'=1}^{l-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) + \prod_{h'=1}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right. \\
&\qquad \left. - \prod_{h'=1}^{h-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) - \sum_{l=2}^{h-1} \prod_{h'=1}^{l-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right| \\
&\qquad = \left| \sum_{l=1}^{h-1} \prod_{h'=1}^{l-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right. \\
&\qquad \left. - \sum_{l=2}^{h} \prod_{h'=1}^{l-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right| \\
&\qquad = \left| \sum_{l=1}^{h-1} \prod_{h'=1}^{l-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right. \\
&\qquad \left. - \sum_{l=1}^{h-1} \prod_{h'=1}^{l} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l+1}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right| \\
&\qquad = \sum_{l=1}^{h-1} \left| p_l^{k,s,h}(s_{l+1} \mid s_l, a_l) - p_l(s_{l+1} \mid s_l, a_l) \right| \prod_{h'=1}^{l-1} p_h(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l+1}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&|u_h^k(s,a) - q_h^{\pi^k}(s,a)| \\
&\quad \le \pi_h^k(a \mid s) \sum_{\sigma_h(s)} \prod_{h'=1}^{h-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) \sum_{l=1}^{h-1} \left| p_l^{k,s,h}(s_{l+1} \mid s_l, a_l) - p_l(s_{l+1} \mid s_l, a_l) \right| \\
&\qquad\qquad \cdot \prod_{h'=1}^{l-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \prod_{h'=l+1}^{h-1} p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \\
&\quad \le \sum_{l=1}^{h-1} \sum_{\sigma_h(s)} \left| p_l^{k,s,h}(s_{l+1} \mid s_l, a_l) - p_l(s_{l+1} \mid s_l, a_l) \right| \left(\pi_l^k(a_l \mid s_l) \prod_{h'=1}^{l-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \right) \\
&\qquad\qquad \cdot \left(\pi_h^k(a \mid s) \prod_{h'=l+1}^{h-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right) \\
&\quad = \sum_{l=1}^{h-1} \sum_{s_l \in \mathcal{S}, a_l \in \mathcal{A}, s_{l+1} \in \mathcal{S}} \left| p_l^{k,s,h}(s_{l+1} \mid s_l, a_l) - p_l(s_{l+1} \mid s_l, a_l) \right| \\
&\qquad\qquad \cdot \left(\sum_{\sigma_l(s_l)} \pi_l^k(a_l \mid s_l) \prod_{h'=1}^{l-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \right) \\
&\qquad\qquad \cdot \left(\sum_{a_{l+1} \in \mathcal{A}} \sum_{\{s_{h''} \in \mathcal{S}, a_{h''} \in \mathcal{A}\}_{h''=l+2}^{h-1}} \pi_h^k(a \mid s) \prod_{h'=l+1}^{h-1} \pi_{h'}^k(a_{h'} \mid s_{h'}) p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) \right) \\
&\quad = \sum_{l=1}^{h-1} \sum_{s_l \in \mathcal{S}, a_l \in \mathcal{A}, s_{l+1} \in \mathcal{S}} \left| p_l^{k,s,h}(s_{l+1} \mid s_l, a_l) - p_l(s_{l+1} \mid s_l, a_l) \right| q_l^{\pi^k}(s_l, a_l) \cdot q_h^{k,s,h}(s, a \mid s_{l+1}),
\end{aligned}
$$

where we ease notation and denote $q_h^{k,s,h}(s,a \mid s_{l+1}) = q_h^{k,s,h}(s,a \mid s_{l+1}; l+1)$. Similarly, we can show that,

$$
\begin{aligned}
&|q_h^{k,s,h}(s,a \mid s_{l+1}) - q_h^{\pi^k}(s,a \mid s_{l+1})| \\
&\quad \lesssim \sum_{h'=l+1}^{h-1} \sum_{s_{h'}\in\mathcal{S}, a_{h'}\in\mathcal{A}, s_{h'+1}\in\mathcal{S}} \left| p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) - p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \right| q_{h'}^{\pi^k}(s_{h'}, a_{h'} \mid s_{l+1}) q_{h'}^{k,s,h}(s,a \mid s_{h'+1}) \\
&\quad \le \pi_h^k(a \mid s) \sum_{h'=l+1}^{h-1} \sum_{s_{h'}\in\mathcal{S}, a_{h'}\in\mathcal{A}, s_{h'+1}\in\mathcal{S}} \left| p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) - p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \right| q_{h'}^{\pi^k}(s_{h'}, a_{h'} \mid s_{l+1}),
\end{aligned}
$$

where the last is since $q_{h'}^{k,s,h}(s,a \mid s_{h'+1}) \le \pi_h^k(a \mid s)$. Combining the last two,

$$
\begin{aligned}
&\sum_{h,s,a,k} |u_h^k(s,a) - q_h^{\pi^k}(s,a)| \\
&\quad \lesssim \sum_{h,s,a,k} \sum_{l=1}^{h-1} \sum_{s_l\in\mathcal{S}, a_l\in\mathcal{A}, s_{l+1}\in\mathcal{S}} \left| p_l^{k,s,h}(s_{l+1} \mid s_l, a_l) - p_l(s_{l+1} \mid s_l, a_l) \right| q_l^{\pi^k}(s_l, a_l) \cdot q_h^{k,s,h}(s,a \mid s_{l+1}) \\
&\quad \le \sum_{h,s,a,k} \sum_{l=1}^{h-1} \sum_{s_l\in\mathcal{S}, a_l\in\mathcal{A}, s_{l+1}\in\mathcal{S}} \tilde{\epsilon}_l^k(s_{l+1} \mid s_l, a_l) q_l^{\pi^k}(s_l, a_l) \cdot q_h^{\pi^k}(s,a \mid s_{l+1}) \\
&\qquad + \sum_{h,s,a,k} \sum_{l=1}^{h-1} \sum_{s_l\in\mathcal{S}, a_l\in\mathcal{A}, s_{l+1}\in\mathcal{S}} \tilde{\epsilon}_l^k(s_{l+1} \mid s_l, a_l) q_l^{\pi^k}(s_l, a_l) \pi_h^k(a \mid s) \\
&\qquad\quad \cdot \left(\sum_{h'=l+1}^{h-1} \sum_{s_{h'}\in\mathcal{S}, a_{h'}\in\mathcal{A}, s_{h'+1}\in\mathcal{S}} \left| p_{h'}^{k,s,h}(s_{h'+1} \mid s_{h'}, a_{h'}) - p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}) \right| q_{h'}^{\pi^k}(s_{h'}, a_{h'} \mid s_{l+1}) \right) \\
&\quad \le \sum_{k,h} \sum_{l=1}^{h-1} \sum_{s_l\in\mathcal{S}, a_l\in\mathcal{A}, s_{l+1}\in\mathcal{S}} \tilde{\epsilon}_l^k(s_{l+1} \mid s_l, a_l) q_l^{\pi^k}(s_l, a_l) \cdot \left(\sum_{s,a} q_h^{\pi^k}(s,a \mid s_{l+1}) \right) \\
&\qquad + \sum_{h,s,k} \sum_{l=1}^{h-1} \sum_{s_l\in\mathcal{S}, a_l\in\mathcal{A}, s_{l+1}\in\mathcal{S}} \tilde{\epsilon}_l^k(s_{l+1} \mid s_l, a_l) q_l^{\pi^k}(s_l, a_l) \sum_a \pi_h^k(a \mid s) \\
&\qquad\quad \cdot \left(\sum_{h'=l+1}^{h-1} \sum_{s_{h'}\in\mathcal{S}, a_{h'}\in\mathcal{A}} \min\left\{ 2, \sum_{s_{h'+1}\in\mathcal{S}} \tilde{\epsilon}_{h'}^k(s_{h'+1} \mid s_{h'}, a_{h'}) \right\} q_{h'}^{\pi^k}(s_{h'}, a_{h'} \mid s_{l+1}) \right) \\
&\quad \le H \sum_{k=1}^{K} \sum_{1\le l\le H} \sum_{s_l\in\mathcal{S}, a_l\in\mathcal{A}, s_{l+1}\in\mathcal{S}} \tilde{\epsilon}_l^k(s_{l+1} \mid s_l, a_l) q_l^{\pi^k}(s_l, a_l) \\
&\qquad + HS \sum_{k=1}^{K} \sum_{1\le l<h'\le H} \sum_{s_l\in\mathcal{S}, a_l\in\mathcal{A}, s_{l+1}\in\mathcal{S}} \sum_{s_{h'}\in\mathcal{S}, a_{h'}\in\mathcal{A}} \tilde{\epsilon}_l^k(s_{l+1} \mid s_l, a_l) q_l^{\pi^k}(s_l, a_l) \\
&\qquad\qquad \cdot \min\left\{ 2, \sum_{s_{h'+1}\in\mathcal{S}} \tilde{\epsilon}_{h'}^k(s_{h'+1} \mid s_{h'}, a_{h'}) \right\} q_{h'}^{\pi^k}(s_{h'}, a_{h'} \mid s_{l+1}) \\
&\quad = H \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}} \tilde{\epsilon}_h^k(s' \mid s,a) q_h^{\pi^k}(s,a) \\
&\qquad + HS \sum_{k=1}^{K} \sum_{1\le h<\tilde{h}\le H} \sum_{s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}} \sum_{\tilde{s}\in\mathcal{S}, \tilde{a}\in\mathcal{A}} \tilde{\epsilon}_h^k(s' \mid s,a) q_h^{\pi^k}(s,a) \min\left\{ 2, \sum_{\tilde{s}'\in\mathcal{S}} \tilde{\epsilon}_{\tilde{h}}^k(\tilde{s}' \mid \tilde{s}, \tilde{a}) \right\} q_{\tilde{h}}^{\pi^k}(\tilde{s}, \tilde{a} \mid s'; h+1),
\end{aligned}
$$

where the last inequality is by Lemma E.3 and since $p_l^{k,s,h}(\cdot|s,a), p_l(\cdot|s,a)$ are probability distributions $\forall (s,a,l)$. □

Lemma E.5. *Under the good event,*

$$\sum_{h,s,a,k} |u_h^k(s,a) - q_h^{\pi^k}(s,a)| \lesssim \sqrt{\frac{H^4 S^2 A K \tau}{m}} + H^3 S^3 A \tau^2.$$

Proof. We first bound $\sum_{h,s,a,k} |u_h^k(s,a) - q_h^{\pi^k}(s,a)|$ using Lemma E.4. Now, for the first term in Equation (20):

$$\begin{aligned}
\sum_{k,h,s,a} \tilde{\epsilon}_h^k(s,a) q_h^{\pi^k}(s,a) &= \frac{1}{m}\sum_{v=1}^m \sum_{k,h,s,a} \tilde{\epsilon}_h^k(s,a) q_h^{\pi^k}(s,a) \le \frac{2}{m}\sum_{v=1}^m \sum_{k,h,s,a} \mathbb{I}\{s_h^{k,v}=s, a_h^{k,v}=a\}\tilde{\epsilon}_h^k(s,a) + 18HS\tau^2 \\
&\lesssim \frac{\sqrt{S\tau}}{m}\sum_{k,h,s,a} \frac{\sum_{v=1}^m \mathbb{I}\{s_h^{k,v}=s, a_h^{k,v}=a\}}{\sqrt{n_h^k(s,a)\vee 1}} \\
&\quad + \frac{S\tau}{m}\sum_{k,h,s,a} \frac{\sum_{v=1}^m \mathbb{I}\{s_h^{k,v}=s, a_h^{k,v}=a\}}{n_h^k(s,a)\vee 1} + HS\tau^2 \\
&\lesssim \sqrt{\frac{H^2 S^2 A K\tau}{m}} + HS^2 A\tau^2,
\end{aligned}$$

where the first inequality is by event E^{on2} and the last inequality is by Equations (18) and (19). Plugging the definition of $\tilde{\epsilon}$, we break the second sum in Eq. (20) as follows:

$$\begin{aligned}
&\sum_{k=1}^K \sum_{1\le h<\tilde{h}\le H} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a}} \tilde{\epsilon}_h^k(s' \mid s,a) q_h^{\pi^k}(s,a) \cdot \min\left\{2, \sum_{\tilde{s}'} \tilde{\epsilon}_{\tilde{h}}^k(\tilde{s}' \mid \tilde{s},\tilde{a})\right\} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) \\
&\quad \lesssim \sum_{k=1}^K \sum_{1\le h<\tilde{h}\le H} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a},\tilde{s}'} \sqrt{\frac{p_h^k(s'|s,a)\tau}{n_h^k(s,a)\vee 1}} q_h^{\pi^k}(s,a) \cdot \sqrt{\frac{p_{\tilde{h}}^k(\tilde{s}'|\tilde{s},\tilde{a})\tau}{n_{\tilde{h}}^k(\tilde{s},\tilde{a})\vee 1}} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) \\
&\qquad + \sum_{k=1}^K \sum_{1\le h<\tilde{h}\le H} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a}} \sqrt{\frac{p_h^k(s'|s,a)\tau}{n_h^k(s,a)\vee 1}} q_h^{\pi^k}(s,a) \cdot \min\left\{2, \frac{S\tau}{n_{\tilde{h}}^k(\tilde{s},\tilde{a})\vee 1}\right\} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) \\
&\qquad + \sum_{k=1}^K \sum_{1\le h<\tilde{h}\le H} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a}} \frac{\tau}{n_h^k(s,a)\vee 1} q_h^{\pi^k}(s,a) \cdot \min\left\{2, \sum_{\tilde{s}'} \tilde{\epsilon}_{\tilde{h}}^k(\tilde{s}' \mid \tilde{s},\tilde{a})\right\} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) \\
&\quad \lesssim \underbrace{\sum_{k=1}^K \sum_{1\le h<\tilde{h}\le H} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a},\tilde{s}'} \sqrt{\frac{p_h^k(s'|s,a)\tau}{n_h^k(s,a)\vee 1}} q_h^{\pi^k}(s,a) \cdot \sqrt{\frac{p_{\tilde{h}}^k(\tilde{s}'|\tilde{s},\tilde{a})\tau}{n_{\tilde{h}}^k(\tilde{s},\tilde{a})\vee 1}} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1)}_{(i)} \\
&\qquad + \underbrace{\sum_{k=1}^K \sum_{1\le h<\tilde{h}\le H} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a},\tilde{s}'} q_h^{\pi^k}(s,a) p_h^k(s'|s,a) \cdot \frac{\tau}{n_{\tilde{h}}^k(\tilde{s},\tilde{a})\vee 1} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1)}_{(ii)} \\
&\qquad + \underbrace{\sum_{k=1}^K \sum_{1\le h<\tilde{h}\le H} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a}} \frac{\tau}{n_h^k(s,a)\vee 1} q_h^{\pi^k}(s,a) \cdot q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1)}_{(iii)},
\end{aligned}$$

where the last inequality follows because $\sqrt{xy} \le x + y$ for every $x, y \ge 0$. Term (i) is bounded as follows:

$$
\begin{aligned}
(i) &= \tau \sum_{1 \le h < \tilde{h} \le H} \sum_{k=1}^{K} \sum_{s,a,s'} \sum_{\tilde{s},\tilde{a},\tilde{s}'} \sqrt{\frac{q_h^{\pi^k}(s,a) q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) p_h^k(\tilde{s}'|\tilde{s},\tilde{a})}{n_h^k(s,a) \vee 1}} \cdot \sqrt{\frac{q_h^{\pi^k}(s,a) q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) p_h^k(s'|s,a)}{n_{\tilde{h}}^k(\tilde{s},\tilde{a}) \vee 1}} \\
&\le \tau \sum_{1 \le h < \tilde{h} \le H} \sqrt{\sum_{k,s,a,s',\tilde{s},\tilde{a},\tilde{s}'} \frac{q_h^{\pi^k}(s,a) q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) p_h^k(\tilde{s}'|\tilde{s},\tilde{a})}{n_h^k(s,a) \vee 1}} \\
&\qquad \cdot \sqrt{\sum_{k,s,a,s',\tilde{s},\tilde{a},\tilde{s}'} \frac{q_h^{\pi^k}(s,a) q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) p_h^k(s'|s,a)}{n_{\tilde{h}}^k(\tilde{s},\tilde{a}) \vee 1}} \\
&= \tau \sum_{1 \le h < \tilde{h} \le H} \sqrt{S \sum_{k=1}^{K} \sum_{s,a} \frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a) \vee 1}} \cdot \sqrt{\sum_{k=1}^{K} \sum_{s',\tilde{s},\tilde{a},\tilde{s}'} \frac{q_{h+1}^{\pi^k}(s') q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1)}{n_{\tilde{h}}^k(\tilde{s},\tilde{a}) \vee 1}} \\
&\le HS\tau \sqrt{\sum_{k=1}^{K} \sum_{s,a,h} \frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a) \vee 1}} \cdot \sqrt{\sum_{k=1}^{K} \sum_{\tilde{s},\tilde{a},\tilde{h}} \frac{q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a})}{n_{\tilde{h}}^k(\tilde{s},\tilde{a}) \vee 1}} \le HS\tau \sum_{k=1}^{K} \sum_{s,a,h} \frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a) \vee 1} \lesssim H^2 S^2 A \tau^2,
\end{aligned}
$$

where the last inequality is by event E^{on3} and Equation (18). Term (ii) is bounded as follows:

$$
\begin{aligned}
(ii) &= S \sum_{k=1}^{K} \sum_{1 \le h < \tilde{h} \le H} \sum_{\tilde{s},\tilde{a},s'} \frac{\tau}{n_{\tilde{h}}^k(\tilde{s},\tilde{a}) \vee 1} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) \sum_{s,a} q_h^{\pi^k}(s,a) p_h^k(s'|s,a) \\
&= S \sum_{k=1}^{K} \sum_{1 \le h < \tilde{h} \le H} \sum_{\tilde{s},\tilde{a}} \frac{\tau}{n_{\tilde{h}}^k(\tilde{s},\tilde{a}) \vee 1} \sum_{s'} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) q_{h+1}^{\pi^k}(s') \\
&= HS\tau \sum_{k=1}^{K} \sum_{\tilde{h}} \sum_{\tilde{s},\tilde{a}} \frac{q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a})}{n_{\tilde{h}}^k(\tilde{s},\tilde{a}) \vee 1} \lesssim H^2 S^2 A \tau^2.
\end{aligned}
$$

Term (iii) is bounded as follows:

$$
\begin{aligned}
(iii) &\le \sum_{k=1}^{K} \sum_{1 \le h < \tilde{h} \le H} \sum_{s,a,s'} \frac{\tau}{n_h^k(s,a) \vee 1} q_h^{\pi^k}(s,a) \sum_{\tilde{s},\tilde{a}} q_{\tilde{h}}^{\pi^k}(\tilde{s},\tilde{a} \mid s'; h+1) \\
&\le HS\tau^2 \sum_{k=1}^{K} \sum_{s,a,h} \frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a) \vee 1} \lesssim H^2 S^2 A \tau^2.
\end{aligned}
$$

□

Algorithm 8 COOPERATIVE O-REPS WITH NON-FRESH RANDOMNESS (COOP-NF-O-REPS)

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, transition function p, number of episodes K, number of agents m, exploration parameter γ, learning rate η, confidence parameter δ.
initialize: $\pi_h^1(a \mid s) = 1/A, q_h^1(s,a) = q_h^{\pi^1}(s,a) \; \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
for $k = 1, \dots, K$ **do**
  for $v = 1, \dots, m$ **do**
    observe initial state $s_1^{k,v}$.
    for $h = 1, \dots, H$ **do**
      pick action $a_h^{k,v} \sim \pi_h^k(\cdot \mid s_h^{k,v})$, suffer cost $c_h^k(s_h^{k,v}, a_h^{k,v})$ and observe next state $s_{h+1}^{k,v}$.
    end for
  end for
  For every (s,a,h) compute $\widetilde{W}_h^k(s,a)$ – the estimate of $W_h^k(s,a) = \Pr[\exists v : \; s_h^{k,v} = s, a^{k,v} = a \mid \pi^k]$ using $N = 10\gamma^{-2} \log \frac{KHSAm}{\delta}$ samples (Algorithm 9).
  compute $\hat{c}_h^k(s,a) = \frac{c_h^k(s,a)\mathbb{I}\{\exists v: \, s_h^{k,v}=s, a_h^{k,v}=a\}}{\widetilde{W}_h^k(s,a)+\gamma} \; \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
  compute $q^{k+1} = \arg\min_{q \in \Delta(\mathcal{M})} \eta \langle q, \hat{c}^k \rangle + \mathrm{KL}(q \,\|\, q^k)$.
  compute $\pi_h^{k+1}(a \mid s) = \frac{q_h^k(s,a)}{\sum_{a' \in \mathcal{A}} q_h^k(s,a')} \; \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
end for

F. The `coop-nf-O-REPS` algorithm for adversarial MDPs with non-fresh randomness and known p

For the setting of adversarial MDPs with non-fresh randomness and known transitions we propose the Cooperative O-REPS with non-fresh randomness algorithm (`coop-nf-O-REPS`; see Algorithm 8). The idea is similar to the `coop-O-REPS` algorithm for fresh randomness, but the key difference is that the probability to reach some state-action pair that the algorithm uses (i.e., $W_h^k(s,a)$) must be computed differently in order to suit non-fresh randomness. In fact, computing $W_h^k(s,a)$ becomes a difficult challenge once the randomness is non-fresh, and a naive computation takes exponential time. Instead we propose to estimate $W_h^k(s,a)$ from samples. That is, we simulate $N = 10\gamma^{-2} \log \frac{KHSAm}{\delta}$ i.i.d episodes in which all agents use policy π^k and then estimate $W_h^k(s,a)$ by the fraction of episodes in which (s,a) was reached in step h by at least one of the m agents. This way our algorithm keeps polynomial running time.

Theorem F.1. *With probability $1 - \delta$, setting $\eta = \gamma = 1/\sqrt{\left(1 + \frac{SA}{m}\right) K}$, the individual regret of each agent of* `coop-nf-O-REPS` *is*

$$R_K = O\left(H\sqrt{SK \log \frac{HSA}{\delta}} + \sqrt{\frac{H^2SAK}{m} \log \frac{HSA}{\delta}} + \frac{HSA}{m} \log \frac{HSA}{\delta} + HS \log \frac{HSA}{\delta} \right).$$

F.1. The good event

Define the following events:

$$E^{app} = \left\{ \forall (s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K] : \; |\widetilde{W}_h^k(s,a) - W_h^k(s,a)| \le \gamma/2 \right\}$$

$$E^{c} = \left\{ \sum_{k=1}^{K} \langle \mathbb{E}^k[\hat{c}^k] - \hat{c}^k, q^k \rangle \le 4H\sqrt{K \log \frac{6}{\delta}} \right\}$$

$$E^{\hat{c}} = \left\{ \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \left(\frac{1}{m} + \pi_h^k(a \mid s) \right) \left(\hat{c}_h^k(s,a) - 2c_h^k(s,a) \right) \le \frac{10HSA \log \frac{6HSA}{\delta}}{m\gamma} + \frac{10HS \log \frac{6HSA}{\delta}}{\gamma} \right\}$$

$$E^{\star} = \left\{ \sum_{k=1}^{K} \langle \hat{c}^k - c^k, q^{\pi^\star} \rangle \le \frac{2H \log \frac{6HSA}{\delta}}{\gamma} \right\}$$

The good event is the intersection of the above events. The following lemma establishes that the good event holds with high probability.

Algorithm 9 ESTIMATE REACHABILITY PROBABILITY FOR NON-FRESH RANDOMNESS

input: state space $\mathcal{S}$, action space $\mathcal{A}$, transition function p, number of agents m, policy π, number of samples N, state-action-step triplet to estimate $(\bar{s}, \bar{a}, \bar{h})$.
initialize indicator for reaching $I(n) \leftarrow 0$ for $n \in [N]$.
for $n = 1, \ldots, N$ **do**
  initialize realized transitions $p^r_h(s' \mid s, a) = 0 \; \forall (s, a, s', h)$.
  for $h = 1, \ldots, \bar{h}$ **do**
    for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
      sample $s' \sim p_h(\cdot \mid s, a)$ and set $p^r_h(s' \mid s, a) = 1$.
    end for
  end for
  for $v = 1, \ldots, m$ **do**
    observe initial state $s^v_1 = s_{\text{init}}$.
    for $h = 1, \ldots, \bar{h}$ **do**
      pick action $a^v_h \sim \pi_h(\cdot \mid s^v_h)$ and observe next state $s^v_{h+1} \sim p^r_h(\cdot \mid s, a)$.
    end for
    if $s^v_{\bar{h}} = \bar{s}, a^v_{\bar{h}} = \bar{a}$ **then**
      set $I(n) \leftarrow 1$.
      break
    end if
  end for
end for
return $\frac{1}{N} \sum_{n=1}^{N} I(n)$.

Lemma F.2 (The Good Event). *Let $\mathbb{G} = E^{app} \cap E^{c} \cap E^{\hat{c}} \cap E^{\star}$ be the good event. It holds that $\Pr[\mathbb{G}] \geq 1 - \delta$.*

Proof. By Hoeffding inequality we have that $\Pr[\neg E^{app}] \leq \delta/6$, and the other events are similar to Lemma D.2. ☐

F.2. Proof of Theorem F.1

Proof of Theorem F.1. By Lemma F.2, the good event holds with probability $1 - \delta$. We now analyze the regret under the assumption that the good event holds. We start by decomposing the regret as follows:

$$
\begin{aligned}
R_K &= \sum_{k=1}^{K} V_1^{k,\pi^k}(s_1^{k,v}) - V_1^{k,\pi^\star}(s_1^{k,v}) = \sum_{k=1}^{K} \langle c^k, q^k - q^{\pi^\star} \rangle \\
&= \underbrace{\sum_{k=1}^{K} \langle c^k - \hat{c}^k, q^k \rangle}_{(A)} + \underbrace{\sum_{k=1}^{K} \langle \hat{c}^k, q^k - q^{\pi^\star} \rangle}_{(B)} + \underbrace{\sum_{k=1}^{K} \langle \hat{c}^k - c^k, q^{\pi^\star} \rangle}_{(C)} .
\end{aligned}
$$

Term (A) can be further decomposed as:

$$
(A) = \sum_{k=1}^{K} \langle c^k - \hat{c}^k, q^k \rangle = \sum_{k=1}^{K} \langle c^k - \mathbb{E}^k[\hat{c}^k], q^k \rangle + \sum_{k=1}^{K} \langle \mathbb{E}^k[\hat{c}^k] - \hat{c}^k, q^k \rangle.
$$

The second term is bounded by $4H\sqrt{K\log\frac{6}{\delta}}$ by the good event E^c, and for the first term:

$$\begin{aligned}\sum_{k=1}^{K}\langle c^k-\mathbb{E}^k[\hat{c}^k],q^k\rangle &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a)c_h^k(s,a)\left(1-\frac{\mathbb{E}^k[\mathbb{I}\{\exists v:\ s_h^{k,v}=s,a_h^{k,v}=a\}]}{\widetilde{W}_h^k(s,a)+\gamma}\right)\\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a)c_h^k(s,a)\left(1-\frac{\mathbb{E}^k[\mathbb{I}\{\exists v:\ s_h^{k,v}=s,a_h^{k,v}=a\}]}{W_h^k(s,a)+\gamma/2}\right)\\ &= \sum_{k,h,s,a} q_h^k(s,a)c_h^k(s,a)\left(1-\frac{W_h^k(s,a)}{W_h^k(s,a)+\gamma/2}\right)\le\gamma\sum_{k,h,s,a}\frac{q_h^k(s,a)}{W_h^k(s,a)+\gamma/2}\\ &\le\gamma\sum_{k,h,s,a}\frac{q_h^k(s)\pi_h^k(a\mid s)}{W_h^k(s,a)}\le\gamma\sum_{kh,s,a}\sum_{a\in\mathcal{A}}\left(\frac{1}{m}+\pi_h^k(a\mid s)\right)=\frac{\gamma HSAK}{m}+\gamma HSK,\end{aligned}$$

where the first inequality is by the event E^{app}, and the last inequality is by Lemma F.3.

Term (B) is bounded by OMD (see, e.g., Zimin & Neu (2013)) as follows:

$$\begin{aligned}(B)=\sum_{k=1}^{K}\langle\hat{c}^k,q^k-q^{\pi^\star}\rangle &\le\frac{H\log(HSA)}{\eta}+\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}q_h^k(s,a)\hat{c}_h^k(s,a)^2\\ &\le\frac{H\log(HSA)}{\eta}+\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}q_h^k(s,a)\frac{\hat{c}_h^k(s,a)}{\widetilde{W}_h^k(s,a)+\gamma}\\ &\le\frac{H\log(HSA)}{\eta}+\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}q_h^k(s,a)\frac{\hat{c}_h^k(s,a)}{W_h^k(s,a)+\gamma/2}\\ &\le\frac{H\log(HSA)}{\eta}+\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{q_h^k(s)\pi_h^k(a\mid s)}{W_h^k(s,a)}\hat{c}_h^k(s,a)\\ &\le\frac{H\log(HSA)}{\eta}+\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\frac{1}{m}+\pi_h^k(a\mid s)\right)\hat{c}_h^k(s,a)\\ &\le\frac{H\log(HSA)}{\eta}+2\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\frac{1}{m}+\pi_h^k(a\mid s)\right)c_h^k(s,a)+\frac{10\eta HSA\log\frac{6HSA}{\delta}}{m\gamma}+\frac{10\eta HS\log\frac{6HSA}{\delta}}{\gamma}\\ &\lesssim\frac{H\log\frac{HSA}{\delta}}{\eta}+\frac{\eta HSAK}{m}+\eta HSK+\frac{\eta HSA\log\frac{6HSA}{\delta}}{m\gamma}+\frac{\eta HS\log\frac{6HSA}{\delta}}{\gamma},\end{aligned}$$

where the forth inequality is by Lemma F.3, and the fifth inequality is by the good event $E^{\hat{c}}$.

Term (C) is bounded by $\frac{2H\log\frac{6H}{\delta}}{\gamma}$ by the good event $E^\star$. Putting the three terms together gives the final regret bound when setting $\eta=\gamma=\sqrt{\frac{\log\frac{HSA}{\delta}}{(1+\frac{A}{m})SK}}$. $\square$

F.3. Auxiliary lemmas

Lemma F.3. *Let π be a policy and denote by $q_h^\pi(s)$ the probability to reach state s in time h when playing policy π. Assume that m agents use the same policy π in an MDP $\mathcal{M}$ with non-fresh randomness, and denote by $W_h(s,a)$ the probability that at least one agent to reaches (s,a) in time h. Then, for every $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, it holds that:*

$$\frac{q_h^\pi(s)\pi_h(a\mid s)}{W_h(s,a)}\le\frac{1}{m}+\pi_h(a\mid s).$$

Proof. Let $M_h(s)$ be the number of agents that arrive at state s in time h. We have that,

$$W_h(s,a) = \Pr[\exists v \in [m] : \; s_h^v = s, a_h^v = a \mid \pi] = \mathbb{E}\left[1 - (1-\pi_h(a \mid s))^{M_h(s)} \mid \pi\right]$$
$$\geq \mathbb{E}\left[\frac{\pi_h(a \mid s)}{\frac{1}{M_h(s)} + \pi_h(a \mid s)} \mid \pi\right] = \mathbb{E}\left[\frac{M_h(s)\pi_h(a \mid s)}{1 + M_h(s)\pi_h(a \mid s)} \mid \pi\right], \tag{21}$$

where the inequality is by Lemma D.3.

Notice that $\mathbb{E}[M_h(s) \mid \pi] = mq_h^\pi(s)$ by linearity of expectation. Therefore, Equation (21) is bounded from below by the value of the following optimization problem:

$$\begin{aligned}
\min_{p_0,\dots,p_m} & \sum_{i=0}^m p_i \frac{i\pi_h(a \mid s)}{1 + i\pi_h(a \mid s)}, \\
s.t. \quad & \sum_{i=0}^m p_i i = mq_h^\pi(s), \\
& \sum_{i=0}^m p_i = 1, \\
& p_i \geq 0 \quad \forall i \in [m],
\end{aligned}$$

where p_i represents $\Pr[M_h(s) = i]$. Since the coefficient of p_i in the constrains and the objective are non-negative, we can substitute the equality constrains with " $\geq$ " constrains. We get the following standard form Linear Programming:

$$\begin{aligned}
\min_{p \in \mathbb{R}^{m+1}} & b^T p, \\
s.t \quad & A^T p \geq c, \\
& p \geq 0,
\end{aligned}$$

where,

$$b = \begin{pmatrix} 0 \\ \frac{\pi_h(a|s)}{1+\pi_h(a|s)} \\ \vdots \\ \frac{m\pi_h(a|s)}{1+m\pi_h(a|s)} \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & m \end{pmatrix}, \quad c = (1, mq_h^\pi(s)).$$

The dual problem is,

$$\begin{aligned}
\max_{x_1,x_2} & (x_1 + x_2 mq_h^\pi(s)) \\
s.t \quad & x_1 \leq 0, \\
& x_1 + x_2 \leq \frac{\pi_h(a \mid s)}{1 + \pi_h(a \mid s)}, \\
& x_1 + 2x_2 \leq \frac{2\pi_h(a \mid s)}{1 + 2\pi_h(a \mid s)}, \\
& \vdots \\
& x_1 + mx_2 \leq \frac{m\pi_h(a \mid s)}{1 + m\pi_h(a \mid s)}, \\
& x_1, x_2 \geq 0.
\end{aligned}$$

From the first and the last constrains we have $x_1 = 0$ and the rest of the constrains are equivalent to $x_2 \leq \frac{\pi_h(a|s)}{1+m\pi_h(a|s)}$. Hence the maximum value is $\frac{m\pi_h(a|s)}{1+m\pi_h(a|s)} q_h^\pi(s)$, which completes the proof. $\square$

Algorithm 10 COOPERATIVE UOB-REPS WITH NON-FRESH RANDOMNESS (COOP-NF-UOB-REPS)

input: state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, number of episodes K, number of agents $m = \sqrt{K}$, exploration parameter γ, learning rate η, confidence parameter δ.
initialize: $n_h^1(s,a) = 0, n_h^1(s,a,s') = 0, \pi_h^1(a \mid s) = 1/A, q_h^1(s,a,s') = 1/S^2A \; \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.
initialize: define a mapping $\sigma : [H] \times \mathcal{A} \times [K] \to [m]$ such that $\sigma(h,a,k) \neq \sigma(h',a',k)$ whenever $h \neq h'$ or $a \neq a'$, and such that each agent is assigned by σ exactly $HA\sqrt{K}$ times (i.e., $|\sigma^{-1}(v)| = HA\sqrt{K}$ for every $v \in [m]$).
for $k = 1, \ldots, K$ **do**
  set $I_h^k(s,a,s') = 0, I_h^k(s,a) = 0 \; \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.
  for $v = 1, \ldots, m$ **do**
    observe initial state $s_1^{k,v}$.
    for $h = 1, \ldots, H$ **do**
      if $\exists \tilde{a} \in \mathcal{A} : \; \sigma(h, \tilde{a}, k) = v$ **then**
        pick action $a_h^{k,v} = \tilde{a}$.
      else
        pick action $a_h^{k,v} \sim \pi_h^k(\cdot \mid s_h^{k,v})$.
      end if
      suffer cost $c_h^k(s_h^{k,v}, a_h^{k,v})$ and observe next state $s_{h+1}^{k,v}$.
      update $I_h^k(s_h^{k,v}, a_h^{k,v}) \leftarrow 1, I_h^k(s_h^{k,v}, a_h^{k,v}, s_{h+1}^{k,v}) \leftarrow 1$.
    end for
    set $n_h^{k+1}(s,a) \leftarrow n_h^k(s,a) + I_h^k(s,a), n_h^{k+1}(s,a,s') \leftarrow n_h^k(s,a,s') + I_h^k(s,a,s') \; \forall (s,a,s',h)$.
    set $\hat{p}_h^{k+1}(s' \mid s,a) \leftarrow \frac{n_h^{k+1}(s,a,s')}{n_h^{k+1}(s,a) \vee 1} \; \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$.
    compute confidence set for $\epsilon_h^{k+1}(s' \mid s,a) = 4\sqrt{\frac{\hat{p}_h^{k+1}(s'|s,a) \ln \frac{HSAK}{4\delta}}{n_h^{k+1}(s,a) \vee 1}} + 10 \frac{\ln \frac{HSAK}{4\delta}}{n_h^{k+1}(s,a) \vee 1}$:

$$\mathcal{P}^{k+1} = \left\{ p' \mid \forall (s,a,s',h) : |\hat{p}_h^{k+1}(s' \mid s,a) - p'_h(s' \mid s,a)| \leq \epsilon_h^{k+1}(s' \mid s,a) \right\}.$$

    compute $u_h^k(s) = \max_{p' \in \mathcal{P}^k} q_h^{p',\pi^k}(s) = \max_{p' \in \mathcal{P}^k} \Pr[s_h = s \mid \pi^k, p'] \; \forall s \in \mathcal{S}$.
    compute $\hat{c}_h^k(s,a) = \frac{c_h^k(s,a) \mathbb{I}\{s_h^{k,\sigma(h,a,k)} = s\}}{u_h^k(s) + \gamma} \; \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
    compute $q^{k+1} = \arg\min_{q \in \Delta(\mathcal{M}, k+1)} \eta \langle q, \hat{c}^k \rangle + \mathrm{KL}(q \parallel q^k)$.
    compute $\pi_h^{k+1}(a \mid s) = \frac{q_h^{k+1}(s,a)}{\sum_{a' \in \mathcal{A}} q_h^{k+1}(s,a')} \; \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, where $q_h^{k+1}(s,a) = \sum_{s' \in \mathcal{S}} q_h^{k+1}(s,a,s')$.
  end for
end for

G. The `coop-nf-UOB-REPS` algorithm for adversarial MDPs with non-fresh randomness and unknown p

For the setting of adversarial MDPs with non-fresh randomness and unknown transitions we propose the Cooperative UOB-REPS with non-fresh randomness algorithm (`coop-nf-UOB-REPS`; see Algorithm 10). The idea is to combine the `coop-nf-O-REPS` algorithm for known transitions with ideas from the `coop-ULCAE` algorithm in order to handle unknown transitions under non-fresh randomness. The main challenge is that, unlike the stochastic case, we cannot eliminate sub-optimal actions. Thus, our method requires $\sqrt{K}$ agents to attain near-optimal regret as opposed to the stochastic case where only H^2A^2 agents are required.

Theorem G.1. *Assume that* `coop-nf-UOB-REPS` *is run with* $m = \sqrt{K}$ *agents. With probability* $1 - \delta$*, setting* $\eta = \gamma = \sqrt{\frac{\log \frac{KHSA}{\delta}}{SK}}$*, the individual regret of each agent of* `coop-nf-O-REPS` *is*

$$R_K = O\left(H^2 S \sqrt{K \log \frac{KHSA}{\delta}} + H^3 S^3 \log^2 \frac{KHSA}{\delta} \right).$$

G.1. The good event

Denote $\epsilon_h^k(s' \mid s,a) = \sqrt{\frac{2\hat{p}_h^k(s'|s,a)\log\frac{30KHSA}{\delta}}{n_h^k(s,a)\vee 1}} + \frac{2\log\frac{30KHSA}{\delta}}{n_h^k(s,a)\vee 1}$ and $\epsilon_h^k(s,a) = \sum_{s'\in\mathcal{S}}\epsilon_h^k(s' \mid s,a)$. Define the following events:

$$E^p = \left\{\forall(k,s,a,s',h): \ |p_h(s'|s,a) - \hat{p}_h^k(s'|s,a)| \le \epsilon_h^k(s' \mid s,a)\right\}$$
$$E^{on} = \left\{\forall(k,h,s,a,v)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}: n_h^k(s,a) \ge \frac{1}{2}\sum_{j=1}^{k-1} q_h^{\pi^j}(s) - \log\frac{6mHSA}{\delta}\right\}$$
$$E^c = \left\{\sum_{k=1}^K \langle \mathbb{E}^k[\hat{c}^k \mid \pi^k] - \hat{c}^k, q^k\rangle \le 4HS\sqrt{K\log\frac{6}{\delta}}\right\}$$
$$E^{\hat{c}} = \left\{\sum_{k=1}^K\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_h^k(a \mid s)\left(\hat{c}_h^k(s,a) - 2c_h^k(s,a)\right) \le \frac{10HS\log\frac{3HSA}{\delta}}{\gamma}\right\}$$
$$E^\star = \left\{\sum_{k=1}^K\langle\hat{c}^k - c^k, q^{\pi^\star}\rangle \le \frac{H\log\frac{3HSA}{\delta}}{\gamma}\right\}$$

The good event is the intersection of the above events. The following lemma establishes that the good event holds with high probability.

Lemma G.2 (The Good Event)**.** *Let $\mathbb{G} = E^p\cap E^{on}\cap E^c\cap E^{\hat{c}}\cap E^\star$ be the good event. It holds that* $\Pr[\mathbb{G}] \ge 1-\delta$.

Proof. Similar to the proofs of Lemmas C.5 and F.2 and to proofs in Jin et al. (2020a). □

G.2. Proof of Theorem G.1

Proof of Theorem G.1. By Lemma G.2, the good event holds with probability $1-\delta$. We now analyze the regret under the assumption that the good event holds. Note that each agent plays the OMD policy π^k in all except for $HA\sqrt{K}$ episodes. Thus, the regret is bounded by the regret of the policies $\{\pi^k\}_{k=1}^K$ plus a $H^2A\sqrt{K}$ term which is at most $H^2S\sqrt{K}$. Next, we focus on bounding the regret of $\{\pi^k\}_{k=1}^K$, starting with the following decomposition:

$$\sum_{k=1}^K V_1^{k,\pi^k}(s_1^{k,v}) - V_1^{k,\pi^\star}(s_1^{k,v}) = \sum_{k=1}^K\langle c^k, q^{\pi^k} - q^{\pi^\star}\rangle$$
$$= \underbrace{\sum_{k=1}^K\langle c^k, q^{\pi^k} - q^k\rangle}_{(A)} + \underbrace{\sum_{k=1}^K\langle c^k - \hat{c}^k, q^k\rangle}_{(B)} + \underbrace{\sum_{k=1}^K\langle \hat{c}^k, q^k - q^{\pi^\star}\rangle}_{(C)} + \underbrace{\sum_{k=1}^K\langle \hat{c}^k - c^k, q^{\pi^\star}\rangle}_{(D)}.$$

Let $\tau = \log\frac{KHSAm}{\delta}$ be a logarithmic term. Term (A) can be decomposed using the value difference lemma (see, e.g., Shani et al. (2020)):

$$(A) = \sum_{k=1}^K\langle c^k, q^{\pi^k} - q^k\rangle \le 2H\sum_{k=1}^K\sum_{h=1}^H\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^{\pi^k}(s,a)\|p_h(\cdot \mid s,a) - \hat{p}_h^k(\cdot \mid s,a)\|_1$$
$$\lesssim H\sqrt{S\tau}\sum_{k,h,s,a}\frac{q_h^{\pi^k}(s,a)}{\sqrt{n_h^k(s,a)\vee 1}} + HS\tau\sum_{k,h,s,a}\frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a)\vee 1},$$

where the second inequality is by event E^p. We now bound each of the two sums separately using the event E^{on}. For the second sum we have:

$$\sum_{k,h,s,a} \frac{q_h^{\pi^k}(s,a)}{n_h^k(s,a) \vee 1} \leq \sum_{k,h,s,a} \frac{q_h^{\pi^k}(s,a)}{(\frac{1}{2}\sum_{j=1}^{k-1} q_h^{\pi^j}(s) - \log\frac{6mHSA}{\delta}) \vee 1} = \sum_{k,h,s} \frac{q_h^{\pi^k}(s)\sum_a \pi_h^k(a \mid s)}{(\frac{1}{2}\sum_{j=1}^{k-1} q_h^{\pi^j}(s) - \log\frac{6mHSA}{\delta}) \vee 1}$$
$$\leq 2HS\tau + 2\sum_{h,s} \sum_{k:\sum_{j=1}^{k-1} q_h^{\pi^j}(s) \geq 2\log\frac{6mHSA}{\delta}} \frac{q_h^{\pi^k}(s)}{\sum_{j=1}^{k-1} q_h^{\pi^j}(s)} \lesssim HS\tau,$$

where the last inequality is by Rosenberg et al. (2020, Lemma B.18). For the first term:

$$\sum_{k,h,s,a} \frac{q_h^{\pi^k}(s,a)}{\sqrt{n_h^k(s,a) \vee 1}} \leq \sum_{k,h,s,a} \frac{q_h^{\pi^k}(s,a)}{\sqrt{(\frac{1}{2}\sum_{j=1}^{k-1} q_h^{\pi^j}(s) - \log\frac{6mHSA}{\delta}) \vee 1}} = \sum_{k,h,s} \frac{q_h^{\pi^k}(s)\sum_a \pi_h^k(a \mid s)}{\sqrt{(\frac{1}{2}\sum_{j=1}^{k-1} q_h^{\pi^j}(s) - \log\frac{6mHSA}{\delta}) \vee 1}}$$
$$\leq 2HS\tau + 2\sum_{h,s} \sum_{k:\sum_{j=1}^{k-1} q_h^{\pi^j}(s) \geq 2\log\frac{6mHSA}{\delta}} \frac{q_h^{\pi^k}(s)}{\sqrt{\sum_{j=1}^{k-1} q_h^{\pi^j}(s)}}$$
$$= 2HS\tau + 2\sum_{h,s} \sum_{k:\sum_{j=1}^{k-1} q_h^{\pi^j}(s) \geq 2\log\frac{6mHSA}{\delta}} \frac{q_h^{\pi^k}(s)}{\sqrt{\sum_{j=1}^{k} q_h^{\pi^j}(s)}} \sqrt{\frac{\sum_{j=1}^{k} q_h^{\pi^j}(s)}{\sum_{j=1}^{k-1} q_h^{\pi^j}(s)}}$$
$$= 2HS\tau + 4\sum_{h,s} \sum_{k:\sum_{j=1}^{k-1} q_h^{\pi^j}(s) \geq 2\log\frac{6mHSA}{\delta}} \frac{q_h^{\pi^k}(s)}{\sqrt{\sum_{j=1}^{k} q_h^{\pi^j}(s)}}$$
$$\leq 2HS\tau + 8\sum_{h,s} \sqrt{\sum_{k=1}^{K} q_h^{\pi^k}(s)} \leq 2HS\tau + 8\sqrt{HS\sum_{k=1}^{K}\sum_{h,s} q_h^{\pi^k}(s)} = 2HS\tau + 8H\sqrt{SK},$$

where the third inequality is by Streeter & McMahan (2010, Lemma 1), and the last inequality is by Jensen's inequality. Putting these together we get that: $(A) \lesssim H^2 S\sqrt{K\tau} + H^2S^2\tau^2$.

Term (B) can be further decomposed as:

$$(B) = \sum_{k=1}^{K} \langle c^k - \hat{c}^k, q^k \rangle = \sum_{k=1}^{K} \langle c^k - \mathbb{E}^k[\hat{c}^k \mid \pi^k], q^k \rangle + \sum_{k=1}^{K} \langle \mathbb{E}^k[\hat{c}^k \mid \pi^k] - \hat{c}^k, q^k \rangle.$$

The second term is bounded by $4HS\sqrt{K\log\frac{6}{\delta}}$ by the good event E^c, and for the first term:

$$\sum_{k=1}^{K} \langle c^k - \mathbb{E}^k[\hat{c}^k \mid \pi^k], q^k \rangle = \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^k(s,a)c_h^k(s,a)\left(1 - \frac{\mathbb{E}^k[\mathbb{I}\{s_h^{k,\sigma(h,a,k)} = s\} \mid \pi^k]}{u_h^k(s) + \gamma}\right)$$
$$= \sum_{k,h,s,a} q_h^k(s,a)c_h^k(s,a)\left(1 - \frac{q_h^{\pi^k}(s)}{u_h^k(s) + \gamma}\right) \leq ` \sum_{k,h,s} q_h^k(s)\left(1 - \frac{q_h^{\pi^k}(s)}{u_h^k(s) + \gamma}\right)$$
$$\leq 2\sum_{k,h,s} \frac{q_h^k(s)}{u_h^k(s)}(u_h^k(s) - q_h^{\pi^k}(s) + \gamma) \leq 2\sum_{k,h,s}(u_h^k(s) - q_h^{\pi^k}(s)) + \gamma HSK,$$

where the second equality is because agent $\sigma(h,a,k)$ plays policy π^k until step h. Finally, $\sum_{k,h,s}(u_h^k(s) - q_h^k(s))$ is bounded by similarly to Lemma E.5.

Term (C) is bounded by OMD (see, e.g., Rosenberg & Mansour (2019a)) as follows:

$$
\begin{aligned}
(B) = \sum_{k=1}^{K} \langle \hat{c}^k, q^k - q^{\pi^\star} \rangle &\leq \frac{2H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^k(s,a) \hat{c}_h^k(s,a)^2 \\
&\leq \frac{2H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q_h^k(s) \pi_h^k(a \mid s) \frac{\hat{c}_h^k(s,a)}{u_h^k(s) + \gamma} \\
&\leq \frac{2H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_h^k(a \mid s) \hat{c}_h^k(s,a) \\
&\leq \frac{2H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_h^k(a \mid s) c_h^k(s,a) + \frac{\eta HS \log \frac{3HSA}{\delta}}{\gamma} \\
&\leq \frac{2H \log(HSA)}{\eta} + \eta \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi_h^k(a \mid s) + \frac{\eta HS \log \frac{3HSA}{\delta}}{\gamma} \\
&= \frac{2H \log(HSA)}{\eta} + \eta HSK + \frac{\eta HS \log \frac{3HSA}{\delta}}{\gamma},
\end{aligned}
$$

where the forth inequality is by the good event $E^{\hat{c}}$.

Term (D) is bounded by $\frac{H \log \frac{3H}{\delta}}{\gamma}$ by the good event $E^\star$. Putting the three terms together gives the final regret bound when setting $\eta = \gamma = \sqrt{\frac{\log \frac{KHSA}{\delta}}{SK}}$. □