Online Learning in Unknown Markov Games

Yi Tian [1 *] **Yuanhao Wang** [2 *] **Tiancheng Yu** [1 *] **Suvrit Sra** [1]

Abstract

We study online learning in unknown Markov games, a problem that arises in episodic multi-agent reinforcement learning where the actions of the opponents are unobservable. We show that in this challenging setting, achieving sublinear regret against the best response in hindsight is statistically hard. We then consider a weaker notion of regret by competing with the *minimax value* of the game, and present an algorithm that achieves a sublinear $\tilde{\mathcal{O}}(K^{2/3})$ regret after K episodes. This is the first sublinear regret bound (to our knowledge) for online learning in unknown Markov games. Importantly, our regret bound is independent of the size of the opponents' action spaces. As a result, even when the opponents' actions are fully observable, our regret bound improves upon existing analysis (e.g., (Xie et al., 2020)) by an exponential factor in the number of opponents.

1. Introduction

Multi-agent reinforcement learning (MARL) helps us model strategic decision making problems in an interactive environment with multiple players. It has witnessed notable recent success (with two or more agents), e.g., in Go (Silver et al., 2016; 2017), video games (Vinyals et al., 2019), Poker (Brown & Sandholm, 2018; 2019), and autonomous driving (Shalev-Shwartz et al., 2016).

When studying MARL, often Markov games (MGs) (Shapley, 1953) are used as the computational model. Compared with Markov decision processes (MDPs) (Puterman, 2014), Markov games allow the players to influence the state transition and returns, and are thus capable of modeling competitive and collaborative behaviors that arise in MARL.

A fundamental problem in MGs is sample efficiency. Unlike MDPs, there are at least two key ways to measure performance in MGs: (1) the offline (self-play) setting, where we control both/all players and aim to minimize the number of episodes required to find a good policy; and (2) the online setting, where we can only control one player (which we refer to as *our player*), treat other players as opponents, and judge how our player performs in the whole process using regret. The offline setting is more useful when training players in a controllable environment (e.g., a simulator) and the online setting is more favorable for life-long learning.

When ensuring sample efficiency for MARL, key challenges arise from the observation model. We distinguish between two online settings. When learning in *informed* MGs, our player can observe the actions taken by the opponents. For learning in *unknown* MGs (Cesa-Bianchi & Lugosi, 2006), such observations are unavailable; information flows to our player only through the revealed returns and state transitions. We emphasize that both *informed games* and *unknown games* are describing the observation process instead of our prior knowledge of the parameters: We always assume zero knowledge of the transition function of the MG.

Learning in unknown MGs is harder, more general, and potentially of greater practical relevance than informed MGs. It is thus important to discover algorithms that can guarantee low regret. However, theoretical understanding for unknown MGs is rather limited. Even the following *fundamental question* for analyzing online learning in unknown MGs is open:

Q1. *Is sublinear regret achievable?*

To see why learning in unknown MGs is challenging, notice that without observing an opponents' actions, we cannot learn the transition function of the MG, even with infinitely many episodes to collect data. Therefore, explore-then-commit type of algorithms cannot achieve sublinear regret.

Another concern arises when the number of players involved increases, as then the effective size of the opponents' action space grows exponentially in it. Therefore, the following question is also crucial, even in (easier) informed MGs:

Q2. *Can the regret be independent of the size of the opponents' action space?*

Contributions. We answer both questions Q1 and Q2 affirmatively in this paper. At the heart of our answers lies an Optimistic Nash V-learning algorithm for online learning

[*]Equal contribution [1]Department of EECS, MIT [2]Department of Computer Science, Princeton University. Correspondence to: Suvrit Sra <suvrit@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

(V-OL) that we develop. This algorithm is significant in the following aspects:

- It achieves $\tilde{\mathcal{O}}(K^{2/3})$ regret, the first sublinear regret bound for online learning in unknown MGs. This bound is nontrivial because without observing opponents' actions, we cannot learn the transition function of the MG, even with infinitely many episodes to collect data.
- Its regret does not depend on the size of the opponents' action space. This regret bound is also the first of this kind in the online setting, even for the (easier) *informed* MG setting. For m-player MGs, the effective size of the opponents' action space is A^{m-1} with A the size of each player's action space. Therefore, compared with existing algorithms (Xie et al., 2020) even in the informed setting, we save an exponential factor.
- It is computationally efficient. The computational complexity does not scale up as the number of players m increases; existing algorithms such as (Xie et al., 2020) suffer space and time complexities exponential in m. Also, in existing algorithms, a subprocedure to find a Nash equilibrium in two-player zero-sum games is called in each step, which becomes the computational bottleneck. In sharp contrast, our algorithm does not require calling any such subprocedures.

The idea of Nash V-learning first appears in (Bai et al., 2020). We denote their original Nash V-learning algorithm by V-SP (SP is an acronym for self-play) to distinguish it from our algorithm V-OL. See the discussion at the end of Section 4 for a detailed comparison of the two algorithms.

Furthermore, although the weaker notion of regret (see Section 2) that we use has appeared in prior works (Brafman & Tennenholtz, 2002; Xie et al., 2020), it is not clear why this choice is statistically reasonable. We justify this notion of regret by showing that competing with the best response in hindsight is statistically hard (Section 3). Specifically, the regret can be exponential in the horizon H. This result also strengthens the computational lower bound in (Bai et al., 2020) for online learning in unknown MGs. As an intermediate step, we prove that competing with the optimal policy in hindsight is also statistically hard in MDPs with adversarial transitions under bandit feedback, which strengthens the computational lower bound in (Yadkori et al., 2013) under bandit feedback and is a result of independent interest.

1.1. Related work

Learning in MGs without strategic exploration. A large body of literature focuses on solving known MGs (Littman, 1994; Hansen et al., 2013) or learning with a generative model (Jia et al., 2019; Sidford et al., 2020; Zhang et al., 2020a), using which we can sample transitions and returns for arbitrary state-action pairs. Littman (2001); Hu & Wellman (2003); Wei et al. (2017) do not assume a generative model, but their results only apply to communicating MGs.

Online MGs. Brafman & Tennenholtz (2002) propose R-max, which does not provide a regret guarantee in general. Xie et al. (2020) study this setting for two-player zero-sum games with linear function approximation using the same weaker definition of regret. They use a value iteration (VI) based algorithm and achieve $\tilde{\mathcal{O}}(\sqrt{H^4A^3B^3S^3K})$ regret when translated into the tabular language, where A and B are number of actions for the two players, S is the number of states and H is the horizon. In Appendix C, we adapt the Optimistic Nash Q-learning algorithm (Q-SP) (Bai et al., 2020) to the online setting (Q-OL, Algorithm 3) and prove for Q-OL a $\tilde{\mathcal{O}}(\sqrt{H^5ABSK})$ regret (Theorem 4). All the three algorithms require observing the opponents' actions and thus cannot be applied to learning in unknown MGs.

Self-play. There is a recent line of work focusing on achieving near-optimal sample complexity in offline two-player zero-sum MGs (Bai & Jin, 2020; Xie et al., 2020; Bai et al., 2020; Liu et al., 2020). The goal is to find an ϵ-approximate Nash equilibrium within K episodes. VI-based methods (Bai & Jin, 2020; Xie et al., 2020) achieve $K = \tilde{\mathcal{O}}(S^2AB/\epsilon^2)$. Q-SP (Bai et al., 2020) achieves $K = \tilde{\mathcal{O}}(SAB/\epsilon^2)$, and the V-SP algorithm (Bai et al., 2020) achieves the best existing result $K = \tilde{\mathcal{O}}(S(A+B)/\epsilon^2)$, matching the lower bound w.r.t. the dependence on S, A, B and ϵ. Note that in the self-play setting, we need to find good policies for both players, so the dependence on B is inevitable. Extensions to multi-player general-sum games are discussed in (Liu et al., 2020) but the dependence on the number of players is exponential.

MDPs with adversarial transitions. Online MGs are closely related to adversarial MDPs. In general, competing with the optimal policy in hindsight in MDPs with adversarial transitions is intractable. With full-information feedback, the problem is computationally hard (Yadkori et al., 2013). With bandit feedback, the problem is statistically hard (Lemma 1). However, under additional structural assumptions, one can achieve low regret (Cheung et al., 2019).

MDPs with adversarial rewards. We can ensure sublinear regret if the transition is fixed (but unknown) and only the reward is chosen adversarially (Zimin & Neu, 2013; Rosenberg & Mansour, 2019; Jin et al., 2019). This yields another useful model for adversarial MDPs. The best existing result in adversarial episodic MDPs with bandit feedback and unknown transition is achieved in (Jin et al., 2019) with $\tilde{\mathcal{O}}(\sqrt{H^3S^2AK})$ regret, where H is the horizon.

Single-agent RL. Finally, there is an abundance of works on sample efficient learning in MDPs. Jaksch et al. (2010) first adopt optimism to achieve efficient exploration in MDPs and Jin et al. (2018) extend this idea to model-free

methods. Azar et al. (2017) and Zhang et al. (2020b) achieve minimax regret bounds (up to log-factors) $\tilde{\mathcal{O}}(\sqrt{H^3SAK})$ for model-based and model-free methods, respectively.

2. Background and problem setup

For simplicity, we formulate the problem of two-player zero-sum MGs in this section and provide our algorithmic solution in Section 4. Please see Section 5 for extensions to multi-player general-sum MGs.

2.1. Markov games: setup and notation

Model. We consider episodic two-player zero-sum MGs, where the max-player (min-player) aims to maximize (minimize) its cumulative return. Let $[H] := \{1, 2, \ldots, H\}$ for positive integer H, and let $\Delta(\mathcal{X})$ be the set of probability distributions on set $\mathcal{X}$. Then such an MG is denoted by $\mathrm{MG}(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r, H)$, where

- $H \in \mathbb{N}_+$ is the number of steps in each episode,
- $\mathcal{S} = \bigcup_{h\in[H+1]} \mathcal{S}_h$ is the state space,
- $\mathcal{A} = \bigcup_{h\in[H]} \mathcal{A}_h$ ($\mathcal{B} = \bigcup_{h\in[H]} \mathcal{B}_h$) is the action space of the max-player (min-player, resp.).
- $\mathbb{P}$ is a collection of *unknown* transition functions $\{\mathbb{P}_h : \mathcal{S}_h \times \mathcal{A}_h \times \mathcal{B}_h \to \Delta(\mathcal{S}_{h+1})\}_{h\in[H]}$, and
- r is a collection of return functions $\{r_h : \mathcal{S}_h \times \mathcal{A}_h \times \mathcal{B}_h \to [0, 1]\}_{h\in[H]}$.

The return r is usually called reward in MDPs, which a player aims to maximize. We will use the term "return" for MGs and reserve the term "reward" for (adversarial) MDPs.

With a subscript h let $\mathcal{S}_h, \mathcal{A}_h, \mathcal{B}_h, \mathbb{P}_h, r_h$ denote the corresponding objects at step h. Let $|\cdot|$ denote cardinality of a set; then define the following terms:

$$S := \sup_{h\in[H]} |\mathcal{S}_h|, \quad A := \sup_{h\in[H]} |\mathcal{A}_h|, \quad B := \sup_{h\in[H]} |\mathcal{B}_h|.$$

Interaction protocol. In each episode, the MG starts at an adversarially chosen initial state $s_1 \in \mathcal{S}_1$. At each step $h \in [H]$, the two players observe the state $s_h \in \mathcal{S}_h$ and simultaneously take actions $a_h \in \mathcal{A}_h$, $b_h \in \mathcal{B}_h$; then the environment transitions to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h, b_h)$ and outputs the return $r_h(s_h, a_h, b_h)$. The max-player's policy μ specifies a distribution on $\mathcal{A}_h$ at each step h. Concretely, $\mu = \{\mu_h\}_{h\in[H]}$ where $\mu_h : \mathcal{S}_h \to \Delta(\mathcal{A}_h)$. Similarly we define the min-player's policy ν.

Value functions. Analogously to MDPs, for a policy pair (μ, ν), step $h \in [H]$, state $s \in \mathcal{S}_h$, and actions $a \in \mathcal{A}_h, b \in \mathcal{B}_h$, define the state value function and Q-value function as:

$$V_h^{\mu,\nu}(s) := \mathbb{E}_{\mu,\nu}[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})|s_h = s],$$

$$Q_h^{\mu,\nu}(s, a, b) := \mathbb{E}_{\mu,\nu}[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})|s_h = s, a_h = a, b_h = b].$$

For compactness of notation, define the operators:

$$\mathbb{P}_h V(s, a, b) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a,b)}[V(s')],$$
$$\mathbb{D}_{\mu,\nu}[Q](s) := \mathbb{E}_{a\sim\mu(\cdot|s),b\sim\nu(\cdot|s)}[Q(s, a, b)].$$

Then we have the following Bellman equations:

$$V_h^{\mu,\nu}(s) = \mathbb{D}_{\mu_h,\nu_h}[Q_h^{\mu,\nu}](s),$$
$$Q_h^{\mu,\nu}(s, a, b) = (r_h + \mathbb{P}_h V_{h+1}^{\mu,\nu})(s, a, b).$$

For convenience define $V_{H+1}^{\mu,\nu}(s) := 0$ for $s \in \mathcal{S}_{H+1}$.

Optimality. For a given min-player's policy ν, there exists a *best response* $\mu^\dagger(\nu)$ to it, such that for any step $h \in [H]$ and state $s \in \mathcal{S}_h$,

$$V_h^{\dagger,\nu}(s) \equiv V_h^{\mu^\dagger(\nu),\nu}(s) := \sup_\mu V_h^{\mu,\nu}(s).$$

Again, a symmetric discussion applies to the best response to a max-player's policy. The following minimax theorem holds for two-player zero-sum MGs: for any step $h \in [H]$ and state $s \in \mathcal{S}_h$,

$$\max_\mu \min_\nu V_h^{\mu,\nu}(s) = \min_\nu \max_\mu V_h^{\mu,\nu}(s).$$

Moreover, the best policies against the best responses

$$\mu^* \in \operatorname*{argmax}_{\mu\in M} V_1^{\mu,\dagger}, \quad \nu^* \in \operatorname*{argmin}_{\nu\in N} V_1^{\dagger,\nu}$$

attain the minimax value. Such a policy pair is known as a Nash equilibrium (NE). We use $V_h^*(s) := V_h^{\mu^*,\nu^*}(s)$ to denote the value at the NE, which is unique for the MG and we call the *minimax value* of the MG.

2.2. Problem setup

We are now ready to formally define the problem of online learning in an *unknown* MG: we control the max-player and in each step, only the state s_h and return r_h are revealed, but not the action of the min-player b_h. Recall that if b_h is also accessible, we call it the *informed* setting.

Our goal is to maximize the expected cumulative return, or equivalently, to minimize the regret. The conventional definition of regret is to compete against the best fixed policy in hindsight:

$$\mathrm{Regret}'(K) := \sup_\mu \sum_{k=1}^{K} \big(V_1^{\mu,\nu^k}(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\big), \quad (1)$$

where the superscript k denotes the corresponding objects in the kth episode. Although we use this compact notation, the regret depends on both μ^k and ν^k.

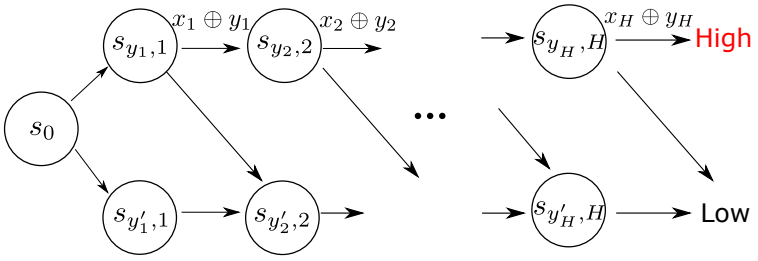

Figure 1. Illustration of the MDP $M_{X,Y}$. For $y \in \{0,1\}$, y' stands for $1-y$.

However, even in the informed setting, achieving sublinear regret in this form is computationally hard (Bai et al., 2020). For online learning in unknown MGs, the problem is statistically hard (Section 3), thus is still intractable even if we have infinite computational power.

Therefore, by noting

$$\max_{\mu \in M} V_1^{\mu,\nu^k}(s_1^k) \geq V_1^{\mu^*,\nu^k}(s_1^k) \geq V_1^*(s_1^k),$$

we consider a more modest goal. That is, to compete against the minimax value of the game, which has appeared in (Brafman & Tennenholtz, 2002). Specifically, we define the following regret (as used by Xie et al. (2020)):

$$\text{Regret}(K) := \sum_{k=1}^{K} \big(V_1^*(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\big). \qquad (2)$$

As a special case, if the opponent is omniscient and plays the best response $\nu^k = \nu^\dagger(\mu^k)$, then a sublinear regret guarantee for (2) implies a sample complexity guarantee to approximate a Nash equilibrium policy.

3. Statistical hardness of online learning in unknown MGs

As mentioned above, we use the minimax value of the game as the benchmark for online learning in unknown MGs. In contrast, in adversarial MDPs (Jin et al., 2019), it is more common to compete against the best policy in hindsight (using regret (1)). In this section, we justify our usage of the weaker notion of regret (2) by showing that, in general, competing against the best policy in hindsight is statistically intractable. In particular, we show that in this case, the regret has to be either linear in K or exponential in H.

Theorem 1 (Statistical hardness for online learning in unknown MGs)**.** *For any $H \geq 2$ and $K \geq 1$, there exists a two-player zero-sum MG with horizon H, $|S_h| \leq 2$, $|A_h| \leq 2$, $|B_h| \leq 4$ such that any algorithm for unknown MGs suffers the following worst-case one-sided regret:*

$$\sup_{\mu} \sum_{k=1}^{K} \Big(V_1^{\mu,\nu^k}(s_1) - \mathbb{E}_{\mu^k} V_1^{\mu^k,\nu^k}(s_1)\Big)$$
$$\geq \Omega\big(\min\{\sqrt{2^H K}, K\}\big).$$

In particular, any algorithm has to suffer linear regret unless $K \geq \Omega(2^H)$.

Here we give a sketch of our proof, while the full proof is deferred to Appendix A.

We start by considering online learning in (single-agent) MDPs, where the reward and transition function in each episode are adversarially determined, and the goal is to compete against the best (fixed) policy in hindsight. In the following lemma we show that this problem is statistically hard; see Lemma 1 in the appendix for its formal statement.

Lemma (informal). *For any algorithm, there exists a sequence of single agent MDPs with horizon H, $S = O(H)$ states and $A = O(1)$ actions, such that the regret defined against the best policy in hindsight is $\Omega(\min\{\sqrt{2^H K}, K\})$.*

Remark 1. *The above lemma is different from a previous hardness result in (Yadkori et al., 2013), which states that this problem is* computationally *hard.*

We now briefly explain how this family of hard MDPs is constructed, which is inspired by the "combination lock" MDP (Du et al., 2019). Every MDP $M_{X,Y}$ is specified by two H-bit strings: $X, Y \in \{0,1\}^H$. The states are $\{s_{0,0}, s_{0,1}, s_{1,1}, \cdots, s_{0,H}, s_{1,H}\}$. As shown in Figure 1, $M_{X,Y}$ has a layered structure, and the reward is nonzero only at the final layer. The only way to achieve the high reward is to follow the path $s_{0,0} \to s_{y_1,1} \to \cdots s_{y_H,H}$. Thus, the corresponding optimal policy is $\pi(s_{w,h}) = x_h \oplus w$, which is only a function of X. Here, $\oplus$ denotes the bitwise exclusive or operator.

Now, in each episode, Y is chosen from a uniform distribution over $\{0,1\}^H$ while X is fixed. When the player interacts with $M_{X,Y}$, since Y is uniformly random, it gets no effective feedback from the observed transitions, and the only informative feedback is the reward at the end. However, achieving the high reward requires guessing every bit of X correctly. This "needle in a haystack" situation makes the problem as hard as a multi-armed bandit problem with 2^H arms. The regret lower bound immediately follows.

Next, we use the hard family of MDPs in Lemma 1 to prove Theorem 1 by reducing the adversarial MDP problem to online learning in unknown MGs. The construction is

straightforward. The state space and the action space for the max-player are the same as that in the original MDP family. The min-player has control over the transition function and reward at each step, and executes a policy such that the induced MDP for the max-player is the same as $M_{X,Y}$. This is possible using only $B = O(1)$ actions as $M_{X,Y}$ has a layered structure. Online learning in unknown MGs then simulates the online learning in the adversarial MDP problem, and thus has the same regret lower bound.

Classes of policies. In Section 2, we define the policy μ by mappings from $\mathcal{S}_h$ to a distribution on $\mathcal{A}_h$ at each step h. Such policies are called *Markov policies* (Bai et al., 2020). The policies induced by the algorithms in the remaining part of this paper are always Markov policies. However, our lower bound also holds for *general policies* (Bai et al., 2020). Here, for an informed max-player the input of μ_h can be the history $(s_1, a_1, b_1, r_1, \cdots, s_h)$, while for a max-player in an unknown MG the input of μ_h can be the history $(s_1, a_1, r_1, \cdots, s_h)$. In words, the lower bound holds even for policies that depend on histories.

Regret minimization in self-play. We emphasize that our lower bound applies to online learning in unknown MGs. For the self-play setting, people indeed minimize the strong regret (1) as an intermediate step toward PAC guarantees (Bai & Jin, 2020; Bai et al., 2020; Xie et al., 2020). This is possible because in self-play *both* players are running the policies specified by the algorithm designer. Therefore, they do not need to worry about the adversarial scenario described in the lower bound here.

We emphasize that our lower bound applies to *online learning* in unknown MGs. In self-play, as an intermediate step toward PAC guarantees, people indeed minimize an even stronger notion called duality gap (Bai & Jin, 2020; Bai et al., 2020; Xie et al., 2020), which is defined as

$$\begin{aligned}\mathrm{Gap}(K) &:= \sum_{k=1}^{K}\left(V_1^{\dagger,\nu^k}(s_1^k) - V_1^{\mu^k,\dagger}(s_1^k)\right)\\ &= \sum_{k=1}^{K}\left(V_1^{\dagger,\nu^k}(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\right)\\ &\quad + \sum_{k=1}^{K}\left(V_1^{\mu^k,\nu^k}(s_1^k) - V_1^{\mu^k,\dagger}(s_1^k)\right),\end{aligned}$$

where the two terms in the last equality are no smaller than the stronger regrets (1) of the two players respectively. This is possible because in self-play *both* players are running the policies specified by the algorithm. Therefore, they do not need to worry about the adversarial scenario described in the lower bound here.

4. The V-OL algorithm

In this section, we introduce the V-OL algorithm and its regret guarantees for online learning in two-player zero-sum *unknown* Markov games. We show that not only can we achieve a sublinear regret in this challenging setting, but the regret bound can be independent of the size of the opponent's action space as well.

Algorithm 1 Optimistic Nash V-learning for Online Learning (V-OL)

Require: Learning rate $\{\alpha_t\}_{t\geq 1}$, exploration bonus $\{\beta_t\}_{t\geq 1}$, policy update parameter $\{\eta_t\}_{t\geq 1}$
Initialize: for any $h \in [H], s \in \mathcal{S}_h, a \in \mathcal{A}_h$, $V_h(s) \leftarrow H$, $L_h(s,a) \leftarrow 0$, $N_h(s) \leftarrow 0$, $\mu_h(a|s) \leftarrow 1/|\mathcal{A}_h|$.
for episode $k = 1, \ldots, K$ **do**
  Receive s_1
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h \sim \mu_h(\cdot|s_h)$
    Observe return r_h and next state s_{h+1}
    Increase counter $t = N_h(s_h) \leftarrow N_h(s_h) + 1$
    $V_h(s_h) \leftarrow (1-\alpha_t)V_h(s_h) + \alpha_t(r_h + V_{h+1}(s_{h+1}) + \beta_t)$
    for all actions $a \in \mathcal{A}_h$ **do**
      $l_h(s_h, a) \leftarrow (H - r_h - V_{h+1}(s_{h+1}))\mathbb{I}(a_h = a)/(\mu_h(a_h|s_h) + \eta_t)$
      $L_h(s_h, a) \leftarrow (1-\alpha_t)L_h(s_h, a) + \alpha_t l_h(s_h, a)$
    end for
    Update policy μ by

$$\mu_h(\cdot|s_h) \leftarrow \frac{\exp\{-\eta_t L_h(s_h, \cdot)/\alpha_t\}}{\sum_a \exp\{-\eta_t L_h(s_h, a)/\alpha_t\}}$$

  end for
end for

The V-OL algorithm. V-OL is a variant of V-learning algorithms. Bai et al. (2020) first propose V-SP as a near-optimal algorithm for the self-play setting of two-player zero-sum MGs. See the discussion at the end of this section for a detailed comparison between V-OL and V-SP.

In V-OL (Algorithm 1), at each time step h, the player interacts with the environment, performs an incremental update to V_h, and updates its policy μ_h. Note that the estimated value function V_h is only used for the intermediate loss $l_h(s_h, \cdot)$ in this time step, but not used in decision making. To encourage exploration in less visited states, we add a bonus term β_t. As we will see in Section 6, this update rule is optimistic, i.e., V_h is an upper confidence bound (UCB) on the minimax value V_h^* of the MG. Then the player samples the action according to the exponentially weighted averaged loss $L_h(s_h, \cdot)$, which is a popular decision rule in adversarial environments (Auer et al., 1995).

Intuition behind V-learning. Most existing provably efficient tabular RL algorithms learn a Q-table (table consisting of Q-values). However, since state-action pairs are necessary for updating the Q-table, for online learning in MGs,

algorithms based on it inevitably require observing the opponent's actions and are thus inapplicable to unknown MGs. By contrast, V-OL does not need to maintain the Q-table at all and bypasses this challenge naturally.

Moreover, learning a Q-value function in two-player Markov games usually results in a regret or sample complexity that depends on its size SAB, whether in the self-play setting, such as VI-ULCB (Bai & Jin, 2020) and Q-SP (Bai et al., 2020), OMNI-VI-offline (Xie et al., 2020), or in the online setting, such as OMNI-VI-online (Xie et al., 2020) and Q-OL (Appendix C). By contrast, V-learning removes the dependence on B, as formalized in Theorem 2.

Note that we analyze Q-OL in Appendix C to more clearly demonstrate V-OL's advantage of avoiding learning a Q-table. Q-OL is a Q-learning-type algorithm for online MGs adapted from Q-SP. It updates the Q-values by a termporal difference method like V-OL but makes decisions based on the Q-values instead. Therefore, Q-OL applies only to the informed setting and its regret depends on AB (Theorem 4).

Favoring more recent samples. Despite the above noted advantages of V-learning, the V-SP algorithm (Bai et al., 2020) may have a regret bound that is linear in K, as indicated by (4) in Theorem 2 and discussed in Section 6 in more detail. To resolve this problem, we adopt a different set of hyperparameters to learn more aggressively by giving more weight to more recent samples. Concretely, for the self-play setting, Bai et al. (2020) specify the following hyperparameters for V-SP:

$$\alpha_t = \tfrac{H+1}{H+t},\ \beta_t = c\sqrt{\tfrac{H^4 A\iota}{t}},\ \eta_t = \sqrt{\tfrac{\log A}{At}},$$

where ι is a log factor defined later and $c > 0$ is a constant. For the online setting, we set these hyperparameters as:

$$\alpha_t = \tfrac{GH+1}{GH+t},\ \beta_t = c(\sqrt{\tfrac{GH^3 A\iota}{t}} + \tfrac{GH^2\iota}{t}),\ \eta_t = \sqrt{\tfrac{GH\iota}{At}}, \tag{3}$$

where $G \geq 1$ is a quantity that we tune and $c > 0$ is a constant. Ostensibly, these changes may appear small, but they are essential to attaining a sublinear regret.

Remark 2. *Compared with $\alpha_t = 1/t$, the learning rate $\alpha_t = {H+1}/{H+t}$ first proposed in (Jin et al., 2018) already favors more recent samples. Here we go one step further: our algorithm learns even more aggressively by taking $\alpha_t = {GH+1}/{GH+t}$ with $G \geq 1$. Moreover, we choose a larger η_t to make our algorithm care more about more recently incurred loss. β_t is set accordingly to achieve optimism.*

We call this variant of V-learning V-OL, for which we prove the following regret guarantees.

Theorem 2 (Regret bounds)**.** *For any $p \in (0,1)$, let $\iota = \log({HSAK}/{p})$. If we run* V-OL *with our hyperparameter specification (3) for some large constant $c > 0$ and $G \geq 1$ in an online two-player zero-sum MG, then with probability at least $1-p$, the regret in K episodes satisfies*

$$\text{Regret}(K) = \mathcal{O}\big(GH^3S\iota^2 + \sqrt{GH^5SAK\iota} + G^{-1}KH\big). \tag{4}$$

In particular, by taking $G = \frac{1}{H}(\frac{K}{SA})^{1/3}$ if $K \geq H^3SA$ and $G = K^{1/3}$ otherwise, with probability at least $1-p$, the regret satisfies

$$\text{Regret}(K) = \begin{cases} \tilde{\mathcal{O}}\big(H^2 S^{\frac{1}{3}} A^{\frac{1}{3}} K^{\frac{2}{3}}\big), & \text{if } K \geq H^3SA, \\ \tilde{\mathcal{O}}\big(\sqrt{H^5SA}K^{\frac{2}{3}} + H^3SK^{\frac{1}{3}}\big), & \text{otherwise.} \end{cases}$$

Theorem 2 shows that a sublinear regret against the minimax value of the MG is achievable for online learning in unknown MGs. As expected, the regret bound does not depend on the size of the opponent's action space B. This independence of B is particularly significant for large B, as is the case where our player plays with multiple opponents. Note that although in Theorem 2 setting the parameter G requires knowledge of K beforehand, we can use a standard doubling trick to bypass this requirement.

Remark 3. *In* V-SP *the parameter G is set to be 1. Then our choice of η_t and β_t become $\sqrt{\frac{H\iota}{At}}$ and $c(\sqrt{\frac{H^3A\iota}{t}} + \frac{H^2\iota}{t})$. If the other player also adopts the corresponding new policy update parameter and exploration bonus, then the sample complexity of* V-SP *can actually be improved upon (Bai et al., 2020) by an H factor to $\tilde{\mathcal{O}}(H^5S(A+B)/\epsilon^2)$.*

Comparison between V-OL and V-SP. Apart from the difference in parameter choices, we now point out other differences between V-OL and V-SP.

1. To achieve near-optimal sample complexity in the self-play setting, V-SP needs to construct upper and lower confidence bounds not only for the *minimax value* of the game, but also for the *best response* values. As a result, it uses a complicated certified policy technique, and must store the whole history of states and policies in the past K episodes for resampling. By comparing with the *minimax value* directly, we can make V-OL provably efficient without extracting a certified policy. Therefore, V-OL only needs $\mathcal{O}(HSA)$ space instead of $\mathcal{O}(HSAK)$, and the resampling procedure is no more necessary.
2. A key feature of the proof in (Bai et al., 2020) is to make full use of a symmetric structure, which naturally arises because in the self-play setting we can control both players to follow the same learning algorithm. However, this property no longer holds for the online setting, and we must take a different proof route. Algorithmically, V-OL learns more aggressively to be provably efficient.

3. V-OL also works in multi-player general-sum MGs; see Section 5.

5. Multi-player general-sum games

In this section, we extend the regret guarantees of V-OL to multi-player general-sum MGs, demonstrating the generality of our algorithm. Notably, the result in multi-player MGs highlights the significance of removing the dependence on B in the regret bound, which is now an exponential factor in the number of opponents.

Formally, consider the m-player general-sum MG

$$\mathrm{MG}_m(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, \mathbb{P}, \{r_i\}_{i=1}^m, H), \tag{5}$$

where $\mathcal{S}$, H follow from the same definition in two-player zero-sum MGs, and

- for each $i \in [m]$, player i has its own action space $\mathcal{A}_i = \bigcup_{h\in[H]} \mathcal{A}_{i,h}$ and return function $r_i = \{r_{i,h} : \mathcal{S}_h \times \bigotimes_{i=1}^m \mathcal{A}_{i,h} \to [0,1]\}_{i=1}^m$, and aims to maximize its own cumulative return (here $\bigotimes$ denotes the Cartesian product of sets);
- $\mathbb{P}$ is a collection of transition functions $\{\mathbb{P}_h : \mathcal{S}_h \times \bigotimes_{i=1}^m \mathcal{A}_{i,h} \to \Delta(\mathcal{S}_{h+1})\}_{h\in[H]}$.

Like in two-player MGs, let

$$S := \sup_{h\in[H]} |\mathcal{S}_h|, \quad A_i := \sup_{h\in[H]} |\mathcal{A}_{i,h}| \text{ for all } i \in [m].$$

Online learning in an unknown multi-player general-sum MG can be reduced to that in a two-player zero-sum MG. Concretely, suppose we are player 1, then online learning in unknown MGs (5) is indistinguishable from that in the two-player zero-sum MG specified by $(\mathcal{S}, \mathcal{A}_1, \mathcal{B}, \mathbb{P}, r_1, H)$ where $\mathcal{B} = \bigotimes_{i=2}^m \mathcal{A}_i$, since we only observe and care about player 1's return. For all states $s \in \mathcal{S}_1$, define the value function using r_1 as

$$V_h^{\mu,\nu}(s) := \mathbb{E}_{\mu,\nu}[\sum_{h'=h}^H r_{1,h'}(s_{h'}, a_{h'}, b_{h'})|s_h = s],$$

and define the minimax value of player 1 as

$$V_1^*(s) := \max_\mu \min_\nu V_1^{\mu,\nu}(s) = \min_\nu \max_\mu V_1^{\mu,\nu}(s),$$

which is no larger than the value at any Nash equilibrium of the multi-player general-sum MG. Then we define the regret against the minimax value of player 1 as

$$\mathrm{Regret}(K) := \sum_{k=1}^K \big(V_1^*(s_1^k) - V_1^{\mu^k,\nu^k}(s_1^k)\big).$$

We argue that this notion of regret is reasonable since we have control of only player 1 and all opponents may collude to compromise our performance. Then immediately we obtain the following corollary from Theorem 2.

Corollary 3 (Regret bound in multi-player MGs)**.** *For any $p \in (0,1)$, let $A = A_1$ and $\iota = \log(HSAK/p)$. If we run* V-OL *with our hyperparameter specification (3) for some large constant $c > 0$ and the choice of G in Theorem 2 for player 1 in the online multi-player general-sum MG (5), then with probability at least $1-p$, the regret in K episodes satisfies*

$$\mathrm{Regret}(K) = \begin{cases} \tilde{\mathcal{O}}\big(H^2 S^{\frac{1}{3}} A_1^{\frac{1}{3}} K^{\frac{2}{3}}\big), & \text{if } K \geq H^3 S A_1, \\ \tilde{\mathcal{O}}\big(\sqrt{H^5 S A_1} K^{\frac{2}{3}} + H^3 S K^{\frac{1}{3}}\big), & \text{otherwise.} \end{cases}$$

In a multi-player MG, the size of the opponents' joint action space B grows exponentially in the number of opponents. Corollary 3 shows that the regret of V-OL only depends on the size of our player's action space A_1. The savings arise because V-OL bypasses the need to learn Q-tables, and the multi-player setting makes no real difference in our analysis.

In the online informed setting, the same equivalence to a two-player zero-sum MG holds, since the other players' actions we observe can be seen as a single action $(a_i)_{i=2}^m$, and whether we observe the other players' returns does not help us decide our policies to maximize our own cumulative return. In this setting, the regret bound in (Xie et al., 2020) becomes $\tilde{\mathcal{O}}(\sqrt{H^4 S^3 \prod_{i=1}^m A_i^3 K})$, which depends exponentially on m. On the other hand, since the online informed setting has stronger assumptions than online learning in unknown MGs, the $\tilde{\mathcal{O}}(H^2 S^{1/3} A_1^{1/3} K^{2/3})$ regret bound of V-OL carries over, which has no dependence on m. This sharp contrast highlights the importance of achieving a regret independent of the size of the opponent's action space.

Furthermore, since in V-OL we only need to update the value function (which has HS entries), rather than update the Q-table (which has $HS\prod_{i=1}^m A_i$ entries) as in (Xie et al., 2020), we can also improve the time and space complexity by an exponential factor in m.

6. Proof sketch of Theorem 2

In this section, we sketch the proof of Theorem 2. We also highlight an observation that V-OL can perform much better than claimed in Theorem 2. Moreover, we expose the problem with V-SP in the online setting, which explains why we favor more recent samples in V-OL.

In the analysis below, we use a superscript k to signify the corresponding quantities at the beginning of the kth episode. To express V_h^k in Algorithm 1 compactly, we introduce the following quantities.

$$\alpha_t^0 := \prod_{j=1}^t (1-\alpha_j), \quad \alpha_t^i := \alpha_i \prod_{j=i+1}^t (1-\alpha_j).$$

Let $t := N_h^k(s)$ and suppose s is previously visited at episodes $k^1, \ldots, k^t \leq k$. Then we can express $V_h^k(s)$ as

$$\alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \big(r_h(s, a_h^{k^i}, b_h^{k^i}) + V_{h+1}^{k^i}(s_{h+1}^{k^i}) + \beta_i\big).$$

It is easy to verify that $\{\alpha_t^i\}_{i=1}^t$ satisfies the normalization property that $\sum_{i=1}^t \alpha_t^i = 1$ for any sequence $\{\alpha_t\}_{t\geq 1}$ and any $t \geq 1$. Moreover, for $\{\alpha_t\}_{t\geq 1}$ specified in (3), $\{\alpha_t^i\}$ has several other desirable properties (Lemma 2), resembling (Jin et al., 2018, Lemma 4.1).

Upper confidence bound (UCB). In Algorithm 1, by bonus β_t we ensure that V_h^k is an entrywise UCB on V_h^* using standard techniques (Bai et al., 2020), building on the normalization property of $\{\alpha_t^i\}_{i=1}^t$ and the key V-learning lemma (Lemma 3) based on the regret bound of the adversarial bandit problem we solve to derive the policy update.

Remark 4. *A main difference from the previous UCB framework (e.g., Azar et al. (2017)) is that here the gap between V_h^k and V_h^* is not necessarily diminishing, which partially explains why we do not achieve the conventional $\tilde{\mathcal{O}}(\sqrt{T})$ regret. Concretely, by taking $\mu = \mu^*$ in the V-learning lemma (Lemma 3), we have*

$$\begin{aligned}
&V_h^k(s) - V_h^*(s)\\
\geq&\sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu_h^*,\nu_h^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s)\\
&\qquad - \mathbb{D}_{\mu_h^*,\nu_h^*}[r_h + \mathbb{P}_h V_{h+1}^*](s)\\
=&\sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu_h^*,\nu_h^{k^i}}[\mathbb{P}_h(V_{h+1}^{k^i} - V_{h+1}^*)](s)\\
&+\sum_{i=1}^t \alpha_t^i (\mathbb{D}_{\mu_h^*,\nu_h^{k^i}} - \mathbb{D}_{\mu_h^*,\nu_h^*})[r_h + \mathbb{P}_h V_{h+1}^*](s)\\
\overset{(i)}{\geq}&\sum_{i=1}^t \alpha_t^i (\mathbb{D}_{\mu_h^*,\nu_h^{k^i}} - \mathbb{D}_{\mu_h^*,\nu_h^*})[r_h + \mathbb{P}_h V_{h+1}^*](s),
\end{aligned}$$

where (i) follows from the above UCB. If the opponent is weak at some step $h \in [H]$ such that for all episodes $k \in [K]$,

$$(\mathbb{D}_{\mu_h^*,\nu_h^k} - \mathbb{D}_{\mu_h^*,\nu_h^*})[r_h + \mathbb{P}_h V_{h+1}^*](s) \geq C,$$

then $\sum_{k=1}^K (V_h^k(s) - V_h^*(s)) \geq CK$. This indicates that the gap between the sum of the UCBs and that of the minimax values can be linear in K. As proved below, we actually show that $\sum_{k=1}^K (V_1^k - V_1^{\mu^k,\nu^k})(s_h^k)$ is sublinear in K, which is much stronger than that merely the regret is sublinear if the opponent is weak. In words, V-OL *performs much better than claimed in Theorem 2 against a weak opponent.*

Regret bounds. Note that the above proof of the UCB holds for any $G > 0$. We now illustrate what problem appears if $G = 1$ and where the constraint $G \geq 1$ comes from. Let "$\lesssim$" denote "$\leq$" up to multiplicative constants. Define $\delta_h^k := (V_h^k - V_h^{\mu^k,\nu^k})(s_h^k)$. Then by the UCB, $\text{Regret}(K) \leq \sum_{k=1}^K \delta_1^k$. It then suffices to bound $\sum_{k=1}^K \delta_1^k$.

By the decomposition of V_h^k, the standard concentration inequality and our choice of β_t, we have (with some lower-order terms hidden)

$$\begin{aligned}
\delta_h^k \lesssim& \sqrt{\tfrac{GH^3A\iota}{t}} + \tfrac{GH^2\iota}{t} - \mathbb{D}_{\mu_h^k,\nu_h^k}[r_h + \mathbb{P}_h V_{h+1}^{\mu^k,\nu^k}](s_h^k)\\
&+\sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu^{k^i},\nu^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s_h^k).
\end{aligned}$$

To treat the last term, we need the regrouping technique (see, e.g., (Jin et al., 2018)): for any quantity f^k indexed by $k \in [K]$,

$$\begin{aligned}
\sum_{k=1}^K \sum_{i=1}^t \alpha_t^i f^{k^i} &\leq \sum_{k'=1}^K f^{k'} \sum_{t=n_h^{k'}}^\infty \alpha_t^{n_h^{k'}}\\
&\leq (1 + \tfrac{1}{GH}) \sum_{k=1}^K f^k.
\end{aligned}$$

Taking $\mathbb{D}_{\mu^{k^i},\nu^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s_h^k)$ as f^i yields (with some lower-order terms hidden)

$$\begin{aligned}
\sum_{k=1}^K \delta_h^k \lesssim \sum_{k=1}^K \big((1 + \tfrac{1}{GH})\delta_{h+1}^k\\
+ \sqrt{\tfrac{GH^3A\iota}{t}} + \tfrac{GH^2\iota}{t} + \tfrac{1}{G}\big),
\end{aligned}$$

where $\frac{1}{G}$ (not arising in the proof of V-SP) results from

$$\tfrac{1}{GH}\mathbb{D}_{\mu_h^k,\nu_h^k}[r_h + \mathbb{P}_h V_{h+1}^{\mu^k,\nu^k}](s_h^k) \leq \tfrac{1}{GH} \cdot H = \tfrac{1}{G}.$$

Since $\sum_{k=1}^K \delta_{H+1}^k = 0$, a recursion over $h \in [H]$ for $\sum_{k=1}^K \delta_h^k$ yields

$$\sum_{k=1}^K \delta_1^k \lesssim (1 + \tfrac{1}{GH})^H \sum_{k=1}^K \sum_{h=1}^H \big(\sqrt{\tfrac{GH^3A\iota}{t}} + \tfrac{GH^2\iota}{t} + \tfrac{1}{G}\big).$$

To bound the coefficient $(1 + \frac{1}{GH})^H \leq e$, we need $G \geq 1$. By standard pigeonhole arguments,

$$\begin{aligned}
\sum_{k=1}^K \sqrt{\tfrac{1}{t}} &= \sum_{s\in\mathcal{S}_h} \sum_{n=1}^{n_h^K(s)} \sqrt{\tfrac{1}{n}} \lesssim \sqrt{SK},\\
\sum_{k=1}^K \tfrac{1}{t} &= \sum_{s\in\mathcal{S}_h} \sum_{n=1}^{n_h^K(s)} \tfrac{1}{n} \lesssim S\log K \leq S\iota.
\end{aligned}$$

Hence, we obtain

$$\sum_{k=1}^K \delta_1^k \lesssim GH^3S\iota^2 + \sqrt{GH^5SAK\iota} + G^{-1}KH.$$

If we take $G = 1$ as in V-SP, the regret is linear in K and therefore useless. To address this problem, we introduced the tunable parameter $G \geq 1$ that balances the $\sqrt{K}$ and K terms in the above bound to yield a sublinear regret.

7. Conclusion and Future Work

In this paper, we study online learning in unknown Markov games using V-OL, which is based on the V-SP algorithm of Bai et al. (2020). V-OL achieves $\tilde{\mathcal{O}}(K^{2/3})$ regret after K episodes. Furthermore, the regret bound is independent of the size of opponents' action space. It is still unclear whether one can achieve a sharper regret bound, which is

a question worthy of future study. We briefly comment on two other future directions.

Toward $\tilde{\mathcal{O}}(K^{1/2})$ regret in MDPs. A key reason why we need to learn more aggressively in online learning is that a symmetric structure (like that in the proof of V-SP) is absent. However, it exists if the opponent plays a *fixed* policy, in which case the Markov game becomes an MDP. To see why, we can imagine the opponent is also executing V-OL, which makes no difference since $B = 1$. However, even in that case, a gap remains: we can only upper and lower bound V_h^* but not $V_h^{\mu^k,\nu^k}$. Figuring out how to fill this gap will make V-OL become the first policy-based algorithm without an estimation of Q-value functions that achieves $\tilde{\mathcal{O}}(K^{1/2})$ regret for tabular RL.

Strong regret for MDPs with adversarial rewards. Another special case is MDPs with adversarial rewards, where the transitions are fixed across episodes. In this case, achieving sublinear regret using strong regret (1) is possible (Jin et al., 2019). A question is then: does V-OL (or its variants) achieve sublinear regret using the strong regret? Given the many technical differences between MDPs with adversarial rewards and online Markov games, it is desirable to resolve these problems in a unified manner. In addition, the form of the model-free update in V-OL should be of independent interest for MDPs with adversarial rewards.

Acknowledgement

YT, TY, SS acknowledge partial support from the NSF BIGDATA grant (number 1741341). We thank Yu Bai, Kefan Dong and Chi Jin for useful discussions.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331. IEEE, 1995.

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. *arXiv preprint arXiv:2002.04017*, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

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

Brown, N. and Sandholm, T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Brown, N. and Sandholm, T. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Non-stationary reinforcement learning: The blessing of (more) optimism. *Available at SSRN 3397818*, 2019.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient RL with Rich Observations via Latent State Decoding. In *International Conference on Machine Learning*, pp. 1665–1674, 2019.

Hansen, T. D., Miltersen, P. B., and Zwick, U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

Hu, J. and Wellman, M. P. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

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

Jia, Z., Yang, L. F., and Wang, M. Feature-based Q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

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 mdps with bandit feedback and unknown transition. *arXiv*, pp. arXiv–1912, 2019.

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

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

Littman, M. L. Friend-or-foe Q-learning in general-sum games. In *ICML*, volume 1, pp. 322–328, 2001.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002, 2020.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *nature*, 550(7676):354–359, 2017.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pp. 4987–4997, 2017.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning Zero-Sum Simultaneous-Move Markov Games Using Function Approximation and Correlated Equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Yadkori, Y. A., Bartlett, P. L., Kanade, V., Seldin, Y., and Szepesvári, C. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in neural information processing systems*, pp. 2508–2516, 2013.

Zhang, K., Kakade, S. M., Başar, T., and Yang, L. F. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020b.

Zimin, A. and Neu, G. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pp. 1583–1591, 2013.

A. Proof of the lower bound

The lower bound builds on the following lower bound for adversarial MDPs where both the transition and the reward function of each episode are chosen adversarially. Note that in our proof of Lemma 1, the optimal policies for M_k are the same, so Lemma 1 indeed implies a lower bound on the regret defined against the best stationary policy in hindsight.

Lemma 1 (Lower bound for adversarial MDPs)**.** *For any horizon $H \geq 2$ and $K \geq 1$, there exists a family of MDPs $\mathcal{M}$ with horizon H, state space $\{S_h\}_{h\leq H}$ with $|S_h| \leq 2$, action space $\{A_h\}_{h\leq H}$ with $|A_h| \leq 2$, and reward $r_h \in [0,1]$ such that the following is true: for any algorithm that deploys policy μ^k in episode k, we have*

$$\sup_{M_1,\cdots,M_K\in\mathcal{M}} \sup_{\mu} \sum_{k=1}^{K} \left(V_{M_k}^{\mu}(s_0) - \mathbb{E}_{\mu_k} V_{M_k}^{\mu^k}(s_0)\right) \geq \Omega(\min\left\{\sqrt{2^H K}, K\right\}),$$

where $V_{M_k}^*$ refers to the optimal value function of MDP M_k.

Proof. Our construction is inspired by the "combination lock" MDP (Du et al., 2019). Let us redefine the horizon length as $H+1$ (so that $H \geq 1$) and let h start from 0. We now define our family of MDPs.

Definition 1 (MDP $M_{X,Y,\varepsilon}$)**.** *For any pair of bit strings $X = (x_1,\ldots,x_H) \in \{0,1\}^H$, $Y = (y_1,\ldots,y_H) \in \{0,1\}^H$ and any $\varepsilon \in (0,1)$, the MDP $M_{X,Y,\varepsilon}$ is defined as follows.*

1. *The state space is $S_0 = \{s_0\}$ and $S_h = \{s_{0,h}, s_{1,h}\}$ for all $1 \leq h \leq H$. The MDP starts at s_0 deterministically and terminates at $s_{0,H}$ or $s_{1,H}$.*
2. *The action space is $A_h = \{0,1\}$ for all $0 \leq h \leq H$.*
3. *The transition is defined as follows:*
   - *s_0 transitions to $s_{0,1}$ or $s_{1,1}$ with probability at least $1/2$ each, regardless of the action taken.*
   - *For any $1 \leq h \leq H-1$, $s_{y_h,h}$ transitions to $s_{y_{h+1},h+1}$ deterministically if $a_h = x_h \oplus y_h$ ("correct state" in combination lock), and transitions to $s_{1-y_{h+1},h+1}$ deterministically if $a_h = 1 - x_h \oplus y_h$.*
   - *For any $1 \leq h \leq H-1$, $s_{1-y_h,h}$ transitions to $s_{1-y_{h+1},h+1}$ deterministically regardless of the action taken ("wrong state" in combination lock).*
4. *The reward is $r_h \equiv 0$ for all $0 \leq h \leq H-1$. At step H, we have*
   - *$r_H(s_{y_H,H}) \sim \mathsf{Ber}(1/2+\varepsilon)$,*
   - *$r_H(s_{1-y_H,H}) \sim \mathsf{Ber}(1/2-\varepsilon)$.*

A visualization for the MDP specified by X, Y and ε is shown in Figure 2.

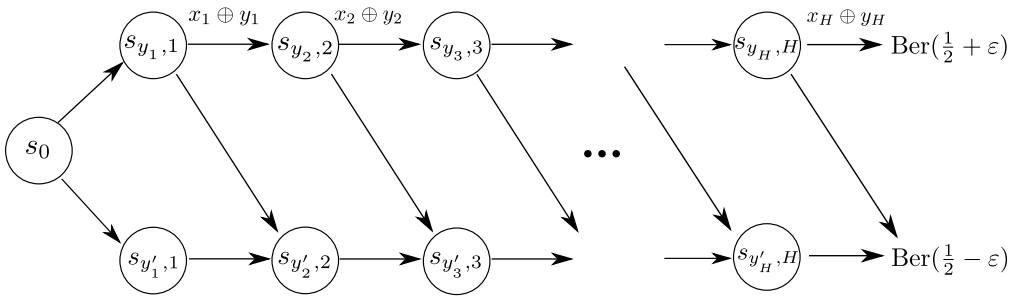

Figure 2. $M(X,Y)$: "Combination lock" MDP specified by X and Y. For $y \in \{0,1\}$, y' stands for $1-y$.

It is straightforward to see that the optimal value function of this MDP is $1/2(1/2+\varepsilon) + 1/2(1/2-\varepsilon) = 1/2$, and the only way to achieve higher reward than $1/2-\varepsilon$ is by following the path of "good states": $(s_0, s_{y_1,1}, \cdots, s_{y_h,h}, \cdots, s_{y_H,H})$. The corresponding optimal policy is $\pi^*(s_{w,h}) = w \oplus x_h$, which is independent of Y.

Random sequence of MDPs is as hard as a 2^H-armed bandit. We now consider any fixed (but unknown) $X \in \{0,1\}^H$ and draw K independent samples $Y_k \sim \mathsf{Unif}(\{0,1\}^H)$ for $1 \le k \le K$. We argue that if we provide $M_k := M_{X,Y_k,\varepsilon}$ in episode k (with some appropriate choice of ε), then the problem is as hard as a 2^H-armed bandit problem with (minimum) suboptimality gap ε, and thus must have the desired regret lower bound.

Our first claim is that, on average over Y_k, the trajectory seen by the algorithm is equivalent (equal in distribution) to the following "completely random" MDP: each state $s_{\{0,1\},h}$ transitions to $s_{\{0,1\},h+1}$ with probability at least $1/2$ regardless of the actions taken; and the reward is $r_H \sim \mathsf{Ber}(1/2)$ if $A = X \oplus Y$ and $r_H \sim \mathsf{Ber}(1/2-\varepsilon)$ if $A \neq X \oplus Y$, where $A = \{a_1, \dots, a_h\}$ are the actions taken in steps 1 through H. Indeed, consider the transition starting from $s_{y_h,h}$. Since $y_{h+1} \sim \mathsf{Ber}(1/2)$, the transition probability to $s_{0,h+1}$ and $s_{1,h+1}$ must be $1/2$ each, regardless of the action taken. The claim about the reward follows from the definition of the MDP.

We now construct a bandit instance, and show that solving this bandit problem can be reduced to online learning in the sequence of MDPs above. The bandit instance has 2^H arms indexed by $\{0,1\}^H$. The arm indexed by X gives reward $\mathsf{Ber}(1/2)$, and otherwise the reward is $\mathsf{Ber}(1/2-\varepsilon)$. Now, for any algorithm solving the adversarial MDP problem, consider the following induced algorithm for the bandit problem.

Algorithm 2 Reducing bandits to adversarial MDPs

1: **for** $k = 1, \dots, K$ **do**
2: Sample $Y \sim \mathsf{Unif}(\{0,1\}^H)$.
3: Simulate the adversarial MDP algorithm by showing the trajectory $(s_0, s_{y_1,1}, \dots, s_{y_H,H})$.
4: Denote the action sequence by $A = (a_1, \dots, a_H)$.
5: Play $A \oplus Y$ in the bandit environment.
6: Show the received bandit reward to the adversarial MDP algorithm as the last step reward.
7: **end for**

We now argue that the interaction seen by the adversarial MDP algorithm is identical in distribution to the sequence $M_{X,Y_k,\varepsilon}$. The trajectory is drawn from a uniform distribution, which is the same as that generated by $M_{X,Y_k,\varepsilon}$. The reward is high, i.e. $\mathsf{Ber}(1/2)$, if and only if $A \oplus Y = X$, which is equivalent to $A = X \oplus Y$. This is also the case in the adversarial MDP problem, since playing the action sequence $X \oplus Y$ corresponds to playing the optimal policy $\pi^*(s_{y_h,h}) = x_h \oplus y_h$.

Therefore, the regret achieved by the induced algorithm in the bandit environment would be equal (in distribution) to the regret achieved by this algorithm in the adversarial MDP environment. Applying classical lower bounds on stochastic bandits (Lattimore & Szepesvári, 2020, Chapter 15) (which corresponds to taking $\varepsilon = \varepsilon_{H,K} := \min\left\{\sqrt{2^H/K}, 1/4\right\}$), we obtain

$$\sup_{X\in\{0,1\}^H} \mathbb{E}_{Y_1,\dots,Y_k\sim\mathsf{Unif}(\{0,1\}^H)}\left[\sum_{k=1}^K \left(V^*_{M_{X,Y_k,\varepsilon_{H,K}}}(s_0) - \mathbb{E}_{\mu^k} V^{\mu^k}_{M_{X,Y_k,\varepsilon_{H,K}}}(s_0)\right)\right] \ge \Omega(\min\left\{\sqrt{2^H K}, K\right\}),$$

where $\mathbb{E}_{\mu^k}$ denotes the randomness in the algorithm execution (which includes the randomness of the realized transitions and rewards that were used by the algorithm to determine μ^k). Note that for the MDP $M_{X,Y_k,\epsilon_H,T}$, the optimal policy is dictated by X and independent of Y_k (hence independent of k). Thus, the previous lower bound can rewritten as a comparison with the best policy in hindsight:

$$\sup_{X\in\{0,1\}^H} \sup_{\mu} \mathbb{E}_{Y_1,\dots,Y_k\sim\mathsf{Unif}(\{0,1\}^H)}\left[\sum_{k=1}^K \left(V^{\mu}_{M_{X,Y_k,\varepsilon_{H,K}}}(s_0) - \mathbb{E}_{\mu^k} V^{\mu^k}_{M_{X,Y_k,\varepsilon_{H,K}}}(s_0)\right)\right] \ge \Omega(\min\left\{\sqrt{2^H K}, K\right\}).$$

The adversarial MDP problem is as hard as the above random sequence of MDPs. Define $\mathcal{M} := \left\{M_{X,Y,\varepsilon_{H,K}} : X, Y \in \{0,1\}^H\right\}$. As the minimax regret is lower bounded by the average regret over any prior distribution of MDPs, the above lower bound implies the following minimax lower bound

$$\sup_{M_k\in\mathcal{M}} \sup_{\mu} \left[\sum_{k=1}^K \left(V^{\mu}_{M_k}(s_0) - \mathbb{E}_{\mu^k} V^{\mu^k}_{M_k}(s_0)\right)\right] \ge \Omega(\min\left\{\sqrt{2^H K}, K\right\})$$

for any adversarial MDP algorithm. □

Proof of Theorem 1. With Lemma 1 in hand, we are in a position to prove the main theorem.

Our proof follows by defining a two-player Markov game and a set of min-player policies $\left\{\nu^k\right\}$ such that the transitions and rewards seen by the max-player are exactly equivalent to the MDP $M_{X,Y_k,\varepsilon_{H,K}}$ constructed in Lemma 1. Indeed, we augment the MDP $M_{X,Y_k,\varepsilon_{H,K}}$ with a set of min-player actions $\mathcal{B}_h = \{1,2,3,4\}$, and redefine the transition such that from any $s_{i,h}$ where $i \in \{0,1\}$ and $1 \le h \le H-1$, the Markov game transitions according to Table 1.

a/b	1	2	3	4
0	$s_{i,h+1}$	$s_{1-i,h+1}$	$s_{i,h+1}$	$s_{1-i,h+1}$
1	$s_{i,h+1}$	$s_{1-i,h+1}$	$s_{1-i,h+1}$	$s_{i,h+1}$

Table 1. transition function of the state $s_{i,h}$ for the hard instance of Markov games.

Such an action set $\mathcal{B}_h$ is powerful enough to reproduce all the possible transitions in the original single-player MDP. We then define ν^k as the policy such that the transition follows exactly M_{X,Y_k}. The reward function is determined only by states and thus remains the same. Therefore, Lemma 1 implies the following one-sided regret bound for the max-player:

$$\sup_{\nu^k}\sup_{\mu}\sum_{k=1}^{K}\left(V^{\mu,\nu^k}(s_0) - \mathbb{E}_{\mu^k}V^{\mu^k,\nu^k}(s_0)\right) \ge \Omega(\min\left\{\sqrt{2^H K}, K\right\}),$$

which is the desired result. □

B. Proof for the V-OL algorithm

Throughout this section, let $\iota = \log(HSAK/p)$, and we use '$\lesssim$' to denote '$\le$' hiding positive universal constants.

The following lemma summarizes the key properties of the choice of the learning rate α_t, which are used in the proof below.

Lemma 2. *The following properties hold for α_t^i.*

1. $1/\sqrt{t} \le \sum_{i=1}^t \alpha_t^i/\sqrt{i} \le 2/\sqrt{t}$ *and* $1/t \le \sum_{i=1}^t \alpha_t^i/i \le 2/t$ *for all* $t \ge 1$.
2. $\sum_{i=1}^t (\alpha_t^i)^2 \le \max_{i\in[t]} \alpha_t^i \le 2GH/t$ *for all* $t \ge 1$.
3. $\sum_{t=i}^\infty \alpha_t^i = 1 + 1/GH$ *for all* $i \ge 1$.

Proof. The properties are copied from (Jin et al., 2018, Lemma 4.1) up to an additional parameter G, except that $1/t \le \sum_{i=1}^t \alpha_t^i/i \le 2/t$ for all $t \ge 1$, which we prove below by induction.

Recall that

$$\alpha_t^0 := \prod_{j=1}^t (1-\alpha_j), \quad \alpha_t^i := \alpha_i \prod_{j=i+1}^t (1-\alpha_j).$$

For the base case $t = 1$, $\sum_{i=1}^t \alpha_t^i/i = \alpha_1^1 = 1$; hence the statement holds. For $t \ge 2$, by noticing $\alpha_t^i = (1-\alpha_t)\alpha_{t-1}^i$, we have

$$\sum_{i=1}^t \frac{\alpha_t^i}{i} = \frac{\alpha_t}{t} + (1-\alpha_t)\sum_{i=1}^{t-1}\frac{\alpha_{t-1}^i}{i}.$$

Then by induction, on the one hand,

$$\sum_{i=1}^t \frac{\alpha_t^i}{i} = \frac{\alpha_t}{t} + (1-\alpha_t)\sum_{i=1}^{t-1}\frac{\alpha_{t-1}^i}{i} \ge \frac{\alpha}{t} + \frac{1-\alpha}{t-1} \ge \frac{\alpha}{t} + \frac{1-\alpha}{t} = \frac{1}{t};$$

on the other hand,

$$\sum_{i=1}^t \frac{\alpha_t^i}{i} = \frac{\alpha_t}{t} + (1-\alpha_t)\sum_{i=1}^{t-1}\frac{\alpha_{t-1}^i}{i} \le \frac{\alpha_t}{t} + \frac{2(1-\alpha_t)}{t} = \frac{H+1}{t(H+t)} + \frac{2}{H+t} = \frac{H+1+2t}{t(H+t)} \overset{(i)}{\le} \frac{2(H+t)}{t(H+t)} = \frac{2}{t},$$

where (i) holds since $H \ge 1$. □

B.1. Upper confidence bound on the minimax value function

Lemma 3 (V-learning lemma)**.** *In Algorithm 1, let $t = N_h^k(s)$ and suppose state $s \in \mathcal{S}_h$ was previously visited at episodes $k^1, \ldots, k^t < k$ at the hth step. For any $p \in (0,1)$, let $\iota = \log(HSAK/p)$. Choose $\eta_t = \sqrt{GH\iota/At}$. Then with probability at least $1-p$, for any $t \in [K]$, $h \in [H]$ and $s \in \mathcal{S}_h$, there exists a constant c such that*

$$\max_{\mu \in \Delta_{\mathcal{A}_h}} \sum_{i=1}^{t} \alpha_t^i \mathbb{D}_{\mu, \nu_h^{k^i}} \left[r_h + \mathbb{P}_h V_{h+1}^{k^i} \right](s) - \sum_{i=1}^{t} \alpha_t^i \left(r_h(s, a_h^{k^i}, b_h^{k^i}) + V_{h+1}^{k^i}(s_{h+1}^{k^i}) \right) \le c(\sqrt{GH^3A\iota/t} + GH^2\iota/t). \quad (6)$$

Proof. By the Azuma-Hoeffding inequality and Lemma 2,

$$\sum_{i=1}^{t} \alpha_t^i \mathbb{D}_{\mu_h^{k^i} \times \nu_h^{k^i}} \left(r_h + \mathbb{P}_h V_{h+1}^{k^i} \right)(s) - \sum_{i=1}^{t} \alpha_t^i \left[r_h\left(s, a_h^{k^i}, b_h^{k^i}\right) + V_{h+1}^{k^i}\left(s_{h+1}^{k^i}\right) \right] \le 2\sqrt{GH^3\iota/t}.$$

Hence, we only need to bound

$$R_t^* := \max_{\mu \in \Delta_{\mathcal{A}_h}} \sum_{i=1}^{t} \alpha_t^i \mathbb{D}_{\mu \times \nu_h^{k^i}} \left(r_h + \mathbb{P}_h V_{h+1}^{k^i} \right)(s) - \sum_{i=1}^{t} \alpha_t^i \mathbb{D}_{\mu_h^{k^i} \times \nu_h^{k^i}} \left(r_h + \mathbb{P}_h V_{h+1}^{k^i} \right)(s), \quad (7)$$

which is the weighted regret of an adversarial bandit problem where the loss is bounded in $[0, H]$. From an intermediate result (the last but one inequality) in the proof of (Bai et al., 2020, Lemma 17), V-OL guarantees that with probability at least $1-p$, the scaled regret R_t^*/H is bounded by

$$\frac{R_t^*}{H} \le \frac{\alpha_t^t \log A}{\eta_t} + \frac{A}{2} \sum_{i=1}^{t} \eta_i \alpha_t^i + \frac{1}{2} \max_{i \le t} \alpha_t^i \iota + A \sum_{i=1}^{t} \eta_i \alpha_t^i + \sqrt{2\iota \sum_{i=1}^{t} (\alpha_t^i)^2} + \max_{i \le t} \frac{\alpha_t^i \iota}{\eta_t},$$

where for the parameters w_i and γ_t in (Bai et al., 2020, Lemma 17), we take $w_i = \alpha_t^i$ and $\gamma_t = \eta_t = \sqrt{\frac{GH\iota}{At}}$. This choice of γ_t satisfies the requirements in the proof of (Bai et al., 2020, Lemma 17) that $\eta_i \le 2\gamma_i$ for all $i \le t$ (Bai et al., 2020, Lemma 19) and that γ_t is nondecreasing in t (Bai et al., 2020, Lemma 21). Therefore, by noticing that $\max_{i \le t} \alpha_t^i = \alpha_t^t = \frac{GH+1}{GH+t} \le 2GH/t$, we have

$$\begin{aligned}
R_t^* &\le H\left(\frac{\alpha_t^t \log A}{\eta_t} + \frac{3A}{2} \sum\nolimits_{i=1}^{t} \eta_i \alpha_t^i + \frac{1}{2} \alpha_t^t \iota + \sqrt{2\iota \sum\nolimits_{i=1}^{t} \left(\alpha_t^i\right)^2} + \frac{\alpha_t^t \iota}{\eta_t} \right) \\
&\overset{(i)}{\lesssim} H\left(\alpha_t^t \sqrt{\frac{At\iota}{GH}} + \sqrt{GHA\iota} \sum\nolimits_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} + GH\iota/t + \sqrt{\iota \sum\nolimits_{i=1}^{t} \left(\alpha_t^i\right)^2} \right) \\
&\overset{(ii)}{\le} H\left(\frac{2GH}{t} \sqrt{\frac{At\iota}{GH}} + 2\sqrt{GHA\iota/t} + GH\iota/t + \sqrt{2GH\iota/t} \right) \\
&\lesssim \sqrt{GH^3A\iota/t} + GH^2\iota/t,
\end{aligned}$$

where (i) is by setting $\eta_t = \sqrt{\frac{GH\iota}{At}}$ and (ii) is by Lemma 2. Taking the union bound for all $(t, h, s) \in [K] \times [H] \times \mathcal{S}$ completes the proof. □

Lemma 4 (Upper confidence bound)**.** *In Algorithm 1, for any $p \in (0,1)$, let $\iota = \log(HSAK/p)$ and choose $\beta_t = c(\sqrt{GH^3A\iota/t} + GH^2\iota/t)$ for some large constant c. Then with probability at least $1-p$, $V_h^*(s) \le V_h^k(s)$ for all $k \in [K]$, $h \in [H]$ and $s \in \mathcal{S}_h$.*

Proof. The proof is similar to that of (Bai et al., 2020, Lemma 15), except that we need to deal with an extra parameter G here.

Let $k_h^i(s)$ denote the index of the episode where $s \in \mathcal{S}_h$ is observed at step h for the ith time. Where there is no ambiguity, we use k^i as a shorthand for $k_h^i(s)$. Let s_h^k be the state actually observed in the algorithm at step h in episode k. For our choice of β_i, we have $\sum_{i=1}^{t} \alpha_t^i \beta_i = \Theta(\sqrt{GH^3A\iota/t} + GH^2\iota/t)$ by Lemma 2.

Recall that

$$V_h^k(s) := \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \left(r_h(s, a_h^{k^i}, b_h^{k^i}) + V_{h+1}^{k^i}(s_{h+1}^{k^i}) + \beta_i \right), \tag{8}$$
$$V_h^*(s) := \mathbb{D}_{\mu_h^*, \nu_h^*}[r_h + \mathbb{P}_h V_{h+1}^*](s).$$

For $h = H+1$ the UCB vacuously holds. To apply backward induction, assume that $V_{h+1}^* \le V_{h+1}^k$ holds entrywise. Then by definition, for any $s \in \mathcal{S}_h$,

$$\begin{aligned} V_h^*(s) &= \max_{\mu \in \Delta_{\mathcal{A}_h}} \min_{\nu \in \Delta_{\mathcal{B}_h}} \mathbb{D}_{\mu,\nu}[r_h + \mathbb{P}_h V_{h+1}^*](s) \\ &\overset{(i)}{=} \max_{\mu \in \Delta_{\mathcal{A}_h}} \sum\nolimits_{i=1}^t \alpha_t^i \min_{\nu \in \Delta_{\mathcal{B}_h}} \mathbb{D}_{\mu,\nu}[r_h + \mathbb{P}_h V_{h+1}^*](s) \\ &\le \max_{\mu \in \Delta_{\mathcal{A}_h}} \sum\nolimits_{i=1}^t \alpha_t^i \mathbb{D}_{\mu, \nu_h^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^*](s) \\ &\overset{(ii)}{\le} \max_{\mu \in \Delta_{\mathcal{A}_h}} \sum\nolimits_{i=1}^t \alpha_t^i \mathbb{D}_{\mu, \nu_h^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s) \overset{(iii)}{\le} V_h^k(s), \end{aligned}$$

where (i) follows from $\sum_{i=1}^t \alpha_t^i = 1$, in (ii) we apply the induction assumption, and (iii) holds with probability at least $1-p$ by the V-learning lemma (Lemma 3) and the decomposition (8) with $\sum_{i=1}^t \alpha_t^i \beta_i = \Theta(\sqrt{GH^3 A\iota/t} + GH^2\iota/t)$. Inductively we have $V_h^*(s) \le V_h^k(s)$ for all $k \in [K]$, $h \in [H]$ and $s \in \mathcal{S}_h$. □

B.2. Proof of Theorem 2

Proof. Recall that

$$V_h^{\mu^k, \nu^k}(s_h^k) = \mathbb{D}_{\mu_h^k, \nu_h^k}[r_h + \mathbb{P}_h V_{h+1}^{\mu^k, \nu^k}](s_h^k).$$

Then define $\delta_h^k := (V_h^k - V_h^{\mu^k, \nu^k})(s_h^k)$. By definition,

$$\begin{aligned} \delta_h^k &= \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \left(r_h(s_h^k, a_h^{k^i}, b_h^{k^i}) + V_{h+1}^{k^i}(s_{h+1}^{k^i}) + \beta_i \right) - \mathbb{D}_{\mu_h^k, \nu_h^k}[r_h + \mathbb{P}_h V_{h+1}^{\mu^k, \nu^k}](s_h^k) \\ &\overset{(i)}{=} \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \left(r_h(s_h^k, a_h^{k^i}, b_h^{k^i}) + V_{h+1}^{k^i}(s_{h+1}^{k^i}) + \beta_i \right) - \sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu^{k^i}, \nu^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s_h^k) \\ &\quad + \sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu^{k^i}, \nu^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s_h^k) - \mathbb{D}_{\mu_h^k, \nu_h^k}[r_h + \mathbb{P}_h V_{h+1}^{\mu^k, \nu^k}](s_h^k) \\ &\overset{(ii)}{\lesssim} \alpha_t^0 H + \sqrt{\tfrac{GH^3 A\iota}{t}} + \tfrac{GH^2\iota}{t} + \sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu^{k^i}, \nu^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s_h^k) - \mathbb{D}_{\mu_h^k, \nu_h^k}[r_h + \mathbb{P}_h V_{h+1}^{\mu^k, \nu^k}](s_h^k), \end{aligned}$$

where in (i) we add and subtract the same term, and (ii) follows from the property of β_i that $\sum_{i=1}^t \alpha_t^i \beta_i = \Theta(\sqrt{GH^3 A\iota/t} + GH^2\iota/t)$ and the fact that by the Azuma-Hoeffding inequality and Property 2 of Lemma 2,

$$\sum_{i=1}^t \alpha_t^i \left(r_h(s_h^k, a_h^{k^i}, b_h^{k^i}) + V_{h+1}^{k^i}(s_{h+1}^{k^i}) \right) - \sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu^{k^i}, \nu^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s_h^k) \lesssim \sqrt{\tfrac{GH^3\iota}{t}}.$$

By the same regrouping technique as that in (Jin et al., 2018),

$$\sum_{k=1}^K \sum_{i=1}^t \alpha_t^i \mathbb{D}_{\mu^{k^i}, \nu^{k^i}}[r_h + \mathbb{P}_h V_{h+1}^{k^i}](s_h^k) \le \sum_{k'=1}^K \mathbb{D}_{\mu^{k'}, \nu^{k'}}[r_h + \mathbb{P}_h V_{h+1}^{k'}](s_h^k) \sum_{t=n_h^{k'}}^{\infty} \alpha_t^{n_h^{k'}}$$

$$\leq (1+\tfrac{1}{GH})\sum_{k=1}^{K}\mathbb{D}_{\mu^k,\nu^k}[r_h+\mathbb{P}_hV_{h+1}^k](s_h^k).$$

Substituting the above back into the bound on δ_h^k and taking sum over $k\in[K]$, we obtain

$$\begin{aligned}
\sum_{k=1}^{K}\delta_h^k &\lesssim \sum_{k=1}^{K}\left(\alpha_t^0 H+\sqrt{\tfrac{GH^3A\iota}{t}}+\tfrac{GH^2\iota}{t}+(1+\tfrac{1}{GH})\mathbb{D}_{\mu^k,\nu^k}[r_h+\mathbb{P}_hV_{h+1}^k](s_h^k)-\mathbb{D}_{\mu_h^k,\nu_h^k}[r_h+\mathbb{P}_hV_{h+1}^{\mu^k,\nu^k}](s_h^k)\right)\\
&\overset{(i)}{=}\sum_{k=1}^{K}\left(\alpha_t^0 H+\sqrt{\tfrac{GH^3A\iota}{t}}+\tfrac{GH^2\iota}{t}+(1+\tfrac{1}{GH})(\delta_{h+1}^k+\gamma_h^k)+\tfrac{1}{GH}\mathbb{D}_{\mu_h^k,\nu_h^k}[r_h+\mathbb{P}_hV_{h+1}^{\mu^k,\nu^k}](s_h^k)\right)\\
&\overset{(ii)}{\leq}\sum_{k=1}^{K}\left(\alpha_t^0 H+\sqrt{\tfrac{GH^3A\iota}{t}}+\tfrac{GH^2\iota}{t}+(1+\tfrac{1}{GH})(\delta_{h+1}^k+\gamma_h^k)+\tfrac{1}{G}\right),
\end{aligned}$$

where in (i) we define the martingale difference term $\gamma_h^k:=\mathbb{D}_{\mu_h^k,\nu_h^k}[\mathbb{P}_h(V_{h+1}^k-V_{h+1}^{\mu^k,\nu^k})](s_h^k)-(V_{h+1}^k-V_{h+1}^{\mu^k,\nu^k})(s_{h+1}^k)$ and (ii) follows from that

$$\mathbb{D}_{\mu_h^k,\nu_h^k}[r_h+\mathbb{P}_hV_{h+1}^{\mu^k,\nu^k}](s_h^k)\leq H.$$

Recursively,

$$\sum_{k=1}^{K}\delta_1^k\lesssim(1+\tfrac{1}{GH})^H\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\alpha_t^0H+\sqrt{\tfrac{GH^3A\iota}{t}}+\tfrac{GH^2\iota}{t}+(1+\tfrac{1}{GH})\gamma_h^k+\tfrac{1}{G}\right).$$

Now we bound each term in $\sum_{k=1}^{K}\delta_1^k$ separately by standard techniques in (Jin et al., 2018; Xie et al., 2020):

$$\begin{aligned}
&\sum_{k=1}^{K}\alpha_{n_h^k}^0H\leq\sum_{k=1}^{K}H\cdot\mathbb{I}(n_h^k=0)\leq HS,\\
&\sum_{k=1}^{K}\sqrt{\tfrac{GH^3A\iota}{n_h^k}}=\sqrt{GH^3A\iota}\sum_{k=1}^{K}\sqrt{\tfrac{1}{n_h^k}}\overset{(i)}{=}\sqrt{GH^3A\iota}\sum_{s\in\mathcal{S}_h}\sum_{n=1}^{n_h^K(s)}\sqrt{\tfrac{1}{n}}\lesssim\sqrt{GH^3SAK\iota},\\
&\sum_{k=1}^{K}\tfrac{GH^2\iota}{t}=GH^2\iota\sum_{k=1}^{K}\tfrac{1}{n_h^k}\overset{(ii)}{=}GH^2\sum_{s\in\mathcal{S}_h}\sum_{n=1}^{n_h^K(s)}\tfrac{1}{n}\lesssim GH^2S\iota^2,\\
&\sum_{k=1}^{K}\sum_{h=1}^{H}\gamma_h^k\overset{(iii)}{\lesssim}\sqrt{H^3K\iota},
\end{aligned}$$

where (i) and (ii) follow from a pigeonhole argument and (iii) follows from the Azuma-Hoeffding inequality. Combining the above bounds, we obtain

$$\text{Regret}(K)\leq\sum_{k=1}^{K}\delta_1^k\lesssim GH^3S\iota^2+\sqrt{GH^5SAK\iota}+G^{-1}KH.$$

If $K\geq H^3SA$ then we take we take $G=\frac{1}{H}(\frac{K}{SA})^{1/3}$; otherwise we take $G=K^{\frac{1}{3}}$. Then the following regret bounds holds:

$$\text{Regret}(K)=\begin{cases}\tilde{\mathcal{O}}\big(H^2S^{\frac{1}{3}}A^{\frac{1}{3}}K^{\frac{2}{3}}\big), & \text{if } K\geq H^3SA,\\ \tilde{\mathcal{O}}\big(\sqrt{H^5SA}K^{\frac{2}{3}}+H^3SK^{\frac{1}{3}}\big), & \text{otherwise.}\end{cases}$$

□

C. The Q-OL Algorithm

When explaining the intuition behind the V-OL in Section 4, we mentioned that learning a Q-table will result in a regret bound depending on AB. This is clear for the algorithms we mentioned in the literature. However, the regret bounds of Q-learning-type algorithms have not been studied to our best knowledge. In this section, we study a Q-learning-type algorithm for online MGs. We formalize Q-OL in Algorithm 3, which is similar to the Optimistic Nash Q-learning (Q-SP) algorithm in (Bai et al., 2020). We emphasize that since learning a Q-table requires knowing the opponents' actions, Q-OL only works for informed MGs, but not for unknown MGs.

Algorithm 3 Optimistic Nash Q-learning for Online Learning (Q-OL)

Require: Learning rate $\{\alpha_t\}_{t\geq 1}$, exploration bonus $\{\beta_t\}_{t\geq 1}$
Initialize: for any (s,a,b,h), $Q_h(s,a,b) \leftarrow H$, $N_h(s,a,b) \leftarrow 0$, $\mu_h(a|s) \leftarrow 1/A$
for episode $k = 1,\ldots,K$ **do**
  Receive s_1
  for step $h = 1,\ldots,H$ **do**
    Take action $a_h \sim \mu_h(\cdot|s_h)$
    Observe action b_h, reward $r_h(s_h,a_h,b_h)$ and next state s_{h+1}
    $t = N_h(s_h,a_h,b_h) \leftarrow N_h(s_h,a_h,b_h)+1$
    $Q_h(s_h,a_h,b_h) \leftarrow (1-\alpha_t)Q_h(s_h,a_h,b_h) + \alpha_t(r_h(s_h,a_h,b_h) + V_{h+1}(s_{h+1}) + \beta_t)$
    Solve the NE $(\mu_h(\cdot,|s_h),\nu_h(\cdot,|s_h))$ of the matrix game with payoff matrix $Q_h^k(s_h,\cdot,\cdot)$
    $V_h(s_h) \leftarrow (\mathbb{D}_{\mu_h\times\nu_h}Q_h)(s_h)$
  end for
end for

In Algorithm 3, we set $\alpha_t := {}^{H+1}/_{H+t}$. As in the analysis of V-OL, below we use a superscript k to signify the corresponding quantities at the beginning of the kth episode. The following lemma claims that Q_h^k and V_h^k are the entrywise upper confidence bounds of Q_h^* and V_h^* for all $k\in[K]$ and $h\in[H]$; see the proof of (Bai et al., 2020, Lemma 3) for its proof.

Lemma 5 (Upper confidence bounds). *In Algorithm 3, for any $p\in(0,1)$, $\iota = \log({}^{HSAK}/_{p})$ and choose $\beta_t = c\sqrt{H^3\iota/t}$ for some large constant c. Then with probability at least $1-p$, $Q_h^*(s,a,b)\leq Q_h^k(s,a,b)$ and $V_h^*(s)\leq V_h^k(s)$ for all $k\in[K]$, $h\in[H]$ and $(s,a,b)\in\mathcal{S}_h\times\mathcal{A}_h\times\mathcal{B}_h$.*

Then for Q-OL, we have the following regret guarantees.

Theorem 4 (Regret bound of Q-OL). *For any $p\in(0,1)$, let $\iota = \log({}^{HSAK}/_{p})$ and choose $\beta_t = c\sqrt{H^3\iota/t}$ for some large constant $c>0$. If we run Q-OL in a two-player zero-sum MG, then with probability at least $1-p$, the regret in K episodes satisfies*

$$\text{Regret}(K) = \mathcal{O}\left(SABH^2 + \sqrt{H^5SABK\iota}\right). \tag{9}$$

Proof. Let $k_h^i(s,a,b)$ denote the index of the episode where (s,a,b) is observed at step h for the ith time. Where there is no ambiguity, we use k^i as a shorthand for $k_h^i(s,a,b)$. Let s_h^k be the state actually observed in the algorithm at step h in episode k.

By defining

$$\overline{\gamma}_h^k := \mathbb{E}_{a\sim\mu_h^k(s_h^k)}[Q_h^k(s_h^k,a,b_h^k)] - Q_h^k(s_h^k,a_h^k,b_h^k),$$
$$\hat{\gamma}_h^k := \mathbb{E}_{a\sim\mu_h^k(s_h^k),b\sim\omega_h^k}[Q_h^{\mu^k,\omega^k}(s_h^k,a,b)] - Q_h^{\mu^k,\omega^k}(s_h^k,a_h^k,b_h^k),$$

we have

$$V_h^k(s_h^k) = \min_{\nu\in\Delta_{\mathcal{B}_h}} \mathbb{E}_{a\sim\mu_h^k(s_h^k),b\sim\nu}[Q_h^k(s_h^k,a,b)] \leq \mathbb{E}_{a\sim\mu_h^k(s_h^k)}[Q_h^k(s_h^k,a,b_h^k)] = Q_h^k(s_h^k,a_h^k,b_h^k) + \overline{\gamma}_h^k,$$
$$V_h^{\mu^k,\omega^k}(s_h^k) = \mathbb{E}_{a\sim\mu_h^k(s_h^k),b\sim\omega_h^k}[Q_h^{\mu^k,\omega^k}(s_h^k,a,b)] = Q_h^{\mu^k,\omega^k}(s_h^k,a_h^k,b_h^k) + \hat{\gamma}_h^k.$$

Define $\delta_h^k := V_h^k(s_h^k) - V_h^{\mu^k,\omega^k}(s_h^k)$ and $\phi_h^k := V_h^k(s_h^k) - V_h^*(s_h^k)$. Then

$$\delta_h^k \le Q_h^k(s_h^k, a_h^k, b_h^k) + \overline{\gamma}_h^k - Q_h^{\mu^k,\omega^k}(s_h^k, a_h^k, b_h^k) - \hat{\gamma}_h^k.$$

In Algorithm 3, for any $k \in [K]$, $h \in [H]$ and $(s,a,b) \in \mathcal{S}_h \times \mathcal{A}_h \times \mathcal{B}_h$, let $t := N_h^k(s,a,b)$ and suppose (s,a,b) is previously visited at episodes $k^1, \cdots, k^t \le k$. Then we can rewrite $Q_h^k(s,a,b)$ as

$$Q_h^k(s,a,b) = \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \left(r_h(s,a,b) + V_{h+1}^{k^i}(s_{h+1}^{k^i}) + \beta_i \right),$$

and recall that

$$Q_h^*(s,a,b) = r_h(s,a,b) + \mathbb{P}_h V_{h+1}^*(s,a,b).$$

Then the difference between Q_h^k and $Q_h^{\mu^k,\omega^k}$ at (s_h^k, a_h^k, b_h^k) satisfies

$$\begin{aligned}
(Q_h^k - Q_h^{\mu^k,\omega^k})(s_h^k, a_h^k, b_h^k) &\overset{(i)}{=} (Q_h^k - Q_h^* + Q_h^* - Q_h^{\mu^k,\omega^k})(s_h^k, a_h^k, b_h^k) \\
&\overset{(ii)}{\le} \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k^i} + 2\tilde{\beta}_t + \mathbb{P}_h(V_{h+1}^* - V_{h+1}^{\mu^k,\omega^k})(s_h^k, a_h^k, b_h^k) \\
&\overset{(iii)}{=} \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k^i} + 2\tilde{\beta}_t + \delta_{h+1}^k - \phi_{h+1}^k + \zeta_h^k,
\end{aligned}$$

where in (i) we add and subtract the same term, in (ii) we define $\tilde{\beta}_t := \sum_{i=1}^t \alpha_t^i \beta_i = \mathcal{O}(\sqrt{H^3\iota/t})$ and by the Azuma-Hoeffding inequality we have

$$\left| \sum_{i=1}^t \alpha_t^i \left(\mathbb{P}_h V_{h+1}^*(s,a,b) - V_{h+1}^{k^i}(s_{h+1}^{k^i}) \right) \right| \le 2H\sqrt{\iota \sum_{i=1}^t (\alpha_t^i)^2} = \mathcal{O}\left(\sqrt{\frac{H^3\iota}{t}} \right) \overset{\text{choice of } \beta_i}{=\!=} \tilde{\beta}_t,$$

and in (iii) we define

$$\zeta_h^k := \mathbb{P}_h(V_{h+1}^* - V_{h+1}^{\mu^k,\omega^k})(s_h^k, a_h^k, b_h^k) - (V_{h+1}^* - V_{h+1}^{\mu^k,\omega^k})(s_{h+1}^k).$$

Therefore,

$$\delta_h^k \le \delta_{h+1}^k + \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \phi_{h+1}^k + 2\tilde{\beta}_t - \phi_{h+1}^k + \zeta_h^k + \overline{\gamma}_h^k - \hat{\gamma}_h^k.$$

Recursively,

$$\delta_1^k \le \sum_{h=1}^H \left(\alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \phi_{h+1}^k + 2\tilde{\beta}_t - \phi_{h+1}^k + \zeta_h^k + \overline{\gamma}_h^k - \hat{\gamma}_h^k \right). \tag{10}$$

By Lemma 5, the regret that we aim to bound is upper bounded by $\sum_{k=1}^K \delta_1^k$. Let $n_h^k := N_k(s_h^k, a_h^k, b_h^k)$. By the regrouping technique in (Jin et al., 2018),

$$\sum_{k=1}^K \sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k^i} \le \sum_{k'=1}^K \phi_{h+1}^{k'} \sum_{t=n_h^{k'}}^\infty \alpha_t^{n_h^{k'}} \le (1 + \frac{1}{H}) \sum_{k=1}^K \phi_{h+1}^k.$$

Substituting the above into (10) yields

$$\sum_{k=1}^K \delta_1^k \le \sum_{k=1}^K \sum_{h=1}^H \left(\alpha_{n_h^k}^0 H + \frac{1}{H} \phi_{h+1}^k + 2\tilde{\beta}_t + \zeta_h^k + \overline{\gamma}_h^k - \hat{\gamma}_h^k \right).$$

Now we bound each term in $\sum_{k=1}^K \delta_1^k$ separately by standard techniques in (Jin et al., 2018; Xie et al., 2020):

$$\begin{aligned}
&\sum_{k=1}^K \alpha_{n_h^k}^0 H \le \sum_{k=1}^K H \cdot \mathbb{I}(n_h^k = 0) \le SABH, \\
&\sum_{k=1}^K \tilde{\beta}_{n_h^k} \le \mathcal{O}(1) \sum_{k=1}^K \sqrt{\frac{H^3\iota}{n_h^k}} \le \mathcal{O}(\sqrt{H^3SABK\iota}), \\
&\sum_{k=1}^K \sum_{h=1}^H (\zeta_h^k + \overline{\gamma}_h^k - \hat{\gamma}_h^k) = \mathcal{O}(\sqrt{H^3K\iota}) = \mathcal{O}(\sqrt{H^3K\iota}).
\end{aligned} \tag{11}$$

Bounding $\frac{1}{H}\sum_{k=1}^K \sum_{h=1}^H \phi_{h+1}^k$ requires additional efforts, since here the relationship $\phi_{h+1}^k \le \delta_{h+1}^k$ in (Jin et al., 2018) does not necessarily hold. Define the martingale difference sequence

$$\gamma_h^k = \mathbb{E}_{a\sim\mu_h^*(s_h^k), b\sim\nu_h^*}[Q_h^*(s_h^k, a, b)] - Q_h^*(s_h^k, a_h^k, b_h^k).$$

Then by noting

$$\phi_h^k = Q_h^k(s_h^k, a_h^k, b_h^k) + \overline{\gamma}_h^k - Q_h^*(s_h^k, a_h^k, b_h^k) - \gamma_h^k \le \alpha_t^0 H + \sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k^i} + 2\tilde{\beta}_t + \overline{\gamma}_h^k - \gamma_h^k,$$

we obtain

$$\sum_{k=1}^K \phi_h^k \le (1 + \frac{1}{H}) \sum_{k=1}^K \phi_{h+1}^k + \sum_{k=1}^K \left(\alpha_t^0 H + 2\tilde{\beta}_t + \overline{\gamma}_h^k - \gamma_h^k\right).$$

Recursively, for all $h' \in [H]$,

$$\sum_{k=1}^K \phi_{h'}^k \le (1 + \frac{1}{H})^{H+1-h'} \sum_{k=1}^K \sum_{h=h'}^H \left(\alpha_t^0 H + 2\tilde{\beta}_t + \overline{\gamma}_h^k - \gamma_h^k\right).$$

Then by similar arguments to those in (11),

$$\frac{1}{H}\sum_{k=1}^K \sum_{h=1}^H \phi_h^k \lesssim SABH^2 + \sqrt{H^5SABK\iota}. \tag{12}$$

Finally, combining the above separate bounds in (11) and (12) yields

$$\text{Regret}(K) \lesssim SABH^2 + \sqrt{H^5SABK\iota}.$$

□