We study online agnostic learning, 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, and present an algorithm that achieves after $K$ episodes a sublinear $\tilde{O}(K^{3/4})$ regret. This is the first sublinear regret bound (to our knowledge) in the online agnostic setting. 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) models 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 and Sandholm, 2018, 2019) and autonomous driving (Shalev-Shwartz et al., 2016).

When studying MARL, an often used computational model is that of Markov games (MGs) (Shapley, 1953). Compared with Markov decision processes (MDPs) (Puterman, 2014), MGs allow the players to influence the state transition and return, 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 learning 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 by 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.

*Equal contribution.
Key challenges in ensuring sample efficiency for MARL arise from the observation model. Here we distinguish between two online settings. In the informed setting, our player has access to the actions taken by the opponents. In the agnostic setting, such observations are unavailable; information flows to our player only through the revealed returns and state transitions. We emphasize that “agnostic learning” in this paper has nothing to do with the term “agnostic learning” in supervised learning.

The agnostic setting is harder, more general, and of greater practical relevance in some applications. It is thus highly desirable to achieve provably low regret. However, theoretical understanding for this setting is rather limited. Even the following fundamental question for analyzing agnostic learning in online MGs is open:

Q1. Can we achieve sublinear regret?

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

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

1.1 Contributions

We answer the above two questions affirmatively. In particular, we do so by developing Optimistic Nash V-learning for online learning (V-ol). This algorithm is significant in the following aspects:

- It achieves \( \tilde{O}(K^{3/4}) \) regret, the first sublinear regret bound in the online agnostic setting.
- Its regret does not depend on the size of the opponents’ action space (again the first regret bound of this kind in the online 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 \( m \).

The idea of Nash V-learning first appears in (Bai et al., 2020), where they design the V-sp (SP is an abbreviation for self-play) algorithm to achieve near-optimal sample complexity in the self-play setting. Compared with V-sp, our V-ol algorithm uses a learning rate that decays slower, which proves crucial to guarantee the above properties.

Furthermore, although the weaker notion of regret (see Section 2) that we use has appeared in prior works (Brafman and Tennenholtz, 2002; Xie et al., 2020), it is not clear why this choice is statistically reasonable. To justify this notion of regret, we show that competing with the best response in hindsight is statistically hard (Section 3). Specifically, the regret can be exponential in the horizon \( H \). This strengthens the computational lower bound in (Bai et al., 2020) in online agnostic setting. 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.

1.2 Related work

Learning in MGs without strategic exploration. A large body of literature on MGs focus on solving a known MG (Hansen et al., 2013; Littman, 1994) or learning with a generative model (Jia et al., 2019;
Sidford et al., 2020), using which we can sample transition and return from arbitrary state-action pair. Hu and Wellman (2003); Littman (2001); Wei et al. (2017) do not assume a generative model, but the theoretical guarantee only applies under the communicating assumption.

Online MGs. Brafman and Tennenholtz (2002) propose R-max, which does not provide a regret guarantee in general. Xie et al. (2020) study this setting for two-player zeros-sum games with linear function approximation using the same weaker definition of regret. They use a value iteration (VI) based algorithm and achieve $\tilde{O}(\sqrt{A^3 B^3 S^3 H^4 K})$ regret when translated into the tabular language, where $A$ and $B$ are number of actions for the two player, $S$ is the number of states and $H$ is horizon. Both algorithms require observing the opponents’ actions and thus cannot be applied to the agnostic setting.

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

MDPs with adversarial transitions. The online setting in MGs is 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 2). However, under additional structural assumptions, it is possible to achieve low regret (Cheung et al., 2019).

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

Single-agent RL. Finally, there are an abundance of works on sample-efficient learning in MDPs. Jakisch et al. (2010) first adopts optimism to achieve efficient exploration in MDPs and Jin et al. (2018) extends this idea to model-free methods. Azar et al. (2017) and Zhang et al. (2020) achieve the minimax regret bounds (up to log-factors) $\tilde{O}(\sqrt{H^3 S A K})$ 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 for multi-player general-sum MGs.
2.1 Markov games

Model. We consider episodic two-player zero-sum MGs, where the max-player (min-player) aims to maximize (minimize, resp.) its cumulative return. Such a MG is denoted by MG(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{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,
- \(\mathcal{A} = \bigcup_{h \in [H]} \mathcal{A}_h\) is the action space of the max-player (min player, resp.),
- \(\mathcal{B}\) is a collection of unknown transition functions \(\{\mathcal{P}_h : \mathcal{S}_h \times \mathcal{A}_h \times \mathcal{B}_h \rightarrow \Delta(\mathcal{S}_{h+1})\}\) for each step, and
- \(r\) is a collection of return functions \(\{r_h : \mathcal{S}_h \times \mathcal{A}_h \times \mathcal{B}_h \rightarrow [0, 1]\}\) for each step.

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

With a subscript \(h\) let \(\mathcal{S}_h, \mathcal{A}_h, \mathcal{B}_h, \mathcal{P}_h, r_h\) 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 \(s_{h+1} \sim \mathcal{P}_h(s_h, a_h, b_h)\) and outputs the return \(r_h(s_h, a_h, b_h)\). The max-player’s policy \(\mu\) specifies a distribution on \(\mathcal{A}\) at each step \(h\). Concretely, \(\mu = \{\mu_h\}_{h \in [H]}\) where \(\mu_h(\cdot) \in \Delta(\mathcal{A})\). Similarly we define the min-player’s policy \(\nu\).

Value functions. In an analogous manner to MDPs, for a policy pair \((\mu, \nu)\), 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 simplicity, let

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

Then we have the following Bellman equations by definition:

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

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 \(\nu\), there exists a best response \(\mu^*\) to it, such that \(V_h^{\mu^*, \nu}(s) = V_h^{\mu^*, \nu}(s) := \sup_{\mu} V_h^{\mu, \nu}(s)\) for any step \(h \in [H]\) and state \(s \in \mathcal{S}_h\). 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 $s \in S_h$,

$$\max_{\mu} \min_{v} V_{h}^{\mu, v}(s) = \min_{v} \max_{\mu} V_{h}^{\mu, v}(s).$$

A policy pair $(\mu^*, v^*)$ that achieves the equality is known as a Nash equilibrium. We use $V_h^*(s) := V_{h}^{\mu^*, v^*}(s)$ to denote the value at the Nash equilibrium, which is unique for the MG and we call the minimax value of the MG.

### 2.2 Problem setup

Now we formally define the online agnostic setting: we control the max-player and after each step, only the state $s_h$ and return $r_h$ are revealed, but not the action of the min-player $b_h$. 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:

$$\text{Regret}^t(K) := \sup_{\mu} \sum_{k=1}^{K} (V_{1}^{\mu, s_k^1} - V_{1}^{\mu, s_k^1}),$$

where the superscript $k$ denotes the corresponding objects in the $k$th episode. Although we use this compact notation, the regret depends on both $\mu$, and $v$.

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

Therefore, by noting

$$\max_{\mu \in M} V_{1}^{\mu, s_k^1} \geq V_{1}^{*, s_k^1} \geq V_{1}^{*}(s_k^1),$$

we consider a more modest goal. That is, to minimize the following regret against the minimax value of the game, which has appeared in prior works (Brafman and Tennenholtz, 2002; Xie et al., 2020):

$$\text{Regret}(K) := \sum_{k=1}^{K} (V_{1}^{*}(s_k^1) - V_{1}^{\mu, s_k^1}).$$

### 3 Statistical hardness for agnostic learning

As mentioned above, we use the minimax value of the game as the benchmark for online agnostic learning. By contrast, in adversarial MDPs (Jin et al., 2019), it is more common to compete against the best policy in hindsight, i.e., (2.1). In this section, we justify our usage of a weaker notion of regret (2.2) 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 agnostic learning). 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 (in the online agnostic setting) suffers the following worst-case one-sided regret:

$$\sup_{\mu} \sum_{k=1}^{K} \left(V_{1}^{\mu, s_k^1} - \mathbb{E}_{\mu} V_{1}^{\mu, s_k^1}\right) \geq \Omega \left(\min \left\{\sqrt{2HK}, K\right\}\right).$$
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 of each episode are adversarially determined, and the goal is to compare against the best (fixed) policy in hindsight. We show that this problem is statistically hard.

**Lemma 1** (Lower bound for adversarial MDPs (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 \( \Omega(\min\{\sqrt{2^H K}, K\}) \).

Note that Lemma 2 is different from a previous hardness result by 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 Fig. 1, \( M_{X,Y} \) has a layered structure, and reward is non-zero only at the final layer. The only way to achieve the high reward is to follow that path \( s_{0,0} \rightarrow s_{y_1,1} \rightarrow \cdots \rightarrow s_{y_H,H} \). Thus the corresponding optimal policy is \( \pi(s_{w,h}) = x_h \oplus w \), which is only a function of \( X \).

Now, in each episode, \( Y \) is chosen from a uniform distribution over \( \{0, 1\}^H \) while \( X \) is fixed. When the player interacts with \( M_{X,Y} \), because \( Y \) is uniformly random, it gets no effective feedback from the transitions observed, 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 2 to prove Theorem 1 by reducing the adversarial MDP problem to online agnostic learning in Markov games. The construction is straightforward. The state space and the action space for the max-player would be the same as that in the original MDP family. The min-player will have control over the transition function and reward at each player, and will execute 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 agnostic learning in this setting would then be the same as online learning in the adversarial MDP problem, and thus suffers from the same regret lower bound.

**Class of policies.** In Section 2, we define the policy \( \mu \) by a distribution on \( A \) at each step \( h \). This class of policies are known as Markov policies (Bai et al., 2020). All the policies induced by algorithms in the remaining part of the paper are Markov policies. However, our lower bound actually holds for a more general class of policies, which are known as general policies (Bai et al., 2020). For general policies, for an informed max-player the input of \( \mu_h \) can be the history \( (s_1, a_1, b_1, r_1, \cdots, s_h) \), while for an agnostic
max-player the input of $\mu_h$ can be the history $(s_1, a_1, r_1, \ldots, s_h)$. In words, the lower bound holds even of the policies depend on history.

4 The V-ol algorithm and its regret bound

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

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|
2: **for** episode $k = 1, \ldots, K** do
3:  
4:   **for** step $h = 1, \ldots, H$ do
5:     Take action $a_h \sim \mu_h(\cdot|s_h)$
6:     Observe return $r_h$ and next state $s_{h+1}$
7:     Increase counter $t = N_h(s_h) \leftarrow N_h(s_h) + 1$
8:     $V_h(s_h) \leftarrow (1 - \alpha_t) V_h(s_h) + \alpha_t (r_h + V_{h+1}(s_{h+1}) + \beta_t)$
9:     **for** all actions $a \in \mathcal{A}_h$ do
10:       $l_h(s_h, a) \leftarrow (H - r_h - V_{h+1}(s_{h+1})) \mathbb{I}(a_h = a) / (\mu_h(a_h|s_h) + \eta_t)$
11:       $L_h(s_h, a) \leftarrow (1 - \alpha_t) L_h(s_h, a) + \alpha_t l_h(s_h, a)$
12:     **end for**
13:     $\mu_h(\cdot|s_h) \leftarrow \exp \{-\eta_t L_h(s_h, \cdot) / \alpha_t \}$
14:     **end for**
15: **end for**

In this section, we introduce the V-ol algorithm and its regret guarantees for online agnostic learning in two-player zero-sum 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.

The V-ol algorithm. Bai et al. (2020) proposes V-sp as a near-optimal algorithm for the self-play setting of two-player zero-sum MGs.

Except for the parameters V-ol is the same as the max-player algorithm in V-sp, which we summarize in Algorithm 1. At each time step $h$, the player interacts with the environment, performs an incremental update to $V_h$, and updates its policy $\mu_h$. The optimism refers to the fact that $V_h$ is an upper confidence bound (UCB) on the minimax value $V^*_h$ of the MG.

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 MGs, algorithms based on it inevitably require observing the opponent’s actions, therefore inapplicable to the agnostic setting. 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 and Jin, 2020) and Optimistic Nash Q-learning (Bai et al., 2020), OMNI-VI (Xie et al., 2020), or in the online setting, such as Xie et al. (2020). By contrast, V-learning is promising in removing the dependence on $B$, as formalized in Theorem 2.
Slower decay in learning rate. Despite the aforementioned possible 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 (4.2) in Theorem 2 and discussed in Section 6 in more details. To resolve this problem, we adopt a learning rate that decays slower. Concretely, for the self-play setting, Bai et al. (2020) specify the following hyperparameters for V-sp:

$$
\alpha_t = \frac{H - 1}{G + t}, \quad \beta_t = c\sqrt{\frac{H^2 A_i}{t}}, \quad \eta_t = \sqrt{\frac{\log A_i}{A_i}},
$$

where $t = \log(\frac{HS AK}{p})$. For the online setting, we propose to keep the specification of $\eta_t$ and to parameterize $\alpha_t$ and $\beta_t$ as:

$$
\alpha_t = \frac{G H + 1}{G H + t}, \quad \beta_t = c G \sqrt{\frac{H^2 A_i}{t}}, \quad \eta_t = \sqrt{\frac{\log A_i}{A_i}},
$$

(4.1)

where $G \geq 1$ is a quantity for us to tune. We call this variant of V-learning V-ol, for which we have the following regret guarantees.

**Theorem 2 (Regret bound).** For any $p \in (0, 1)$, for large constant $c$ in (4.1) and $t = \log(\frac{HS AK}{p})$, if we run V-ol with our hyperparameter specification in a two-player zero-sum MG, then with probability at least $1 - p$, the regret in $K$ episodes satisfies

$$
\text{Regret}(K) \leq O \left( SH^2 + GH^3 \sqrt{SAK} + G^{-1} KH \right).
$$

(4.2)

In particular, if we take $G = K^{1/4}$ \footnote{Setting the parameter $G$ this way requires knowledge of $K$ beforehand. This assumption can be removed by using a standard doubling trick.} then

$$
\text{Regret}(K) \leq O \left( SH^2 + H^3 K^{3/4} \sqrt{SA} \right).
$$

Theorem 2 shows that a sublinear regret against the minimax value of our agent is achievable in the agnostic setting. And as expected, the regret bound does not depend on $B$, the size of the opponent’s action space. This independence of $B$ is particularly significant for large $B$, as is the case where our agent plays with multiple players, which we discuss in the following section.

## 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.

Consider the $m$-player general-sum MG

$$
\text{MG}_m(S, \{A_i\}_{i=1}^m, \{P_i\}_{i=1}^m, H),
$$

(5.1)

where $S, H$ follow from the same definition in two-player zero-sum Markov games, and

- for each $i \in [m]$, player $i$ has its own action space $A_i = \bigcup_{h \in [H]} A_i h$ and return function $r_i = \{r_{i,h} : S_h × \bigotimes_{i=1}^m A_i h → [0, 1]\}_{h=1}^m$, and aims to maximize its own cumulative return;

- $P$ is a collection of transition functions $\{P_h : S_h × \bigotimes_{i=1}^m A_i h → \Delta(S_{h+1})\}_{h=1}^H$. 


Online agnostic learning in a 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 agnostic learning in MG (5.1) is indistinguishable from that in the two-player zero-sum MG specified by \((S, A_1, B, P, r_1, H)\) where \(B = \otimes_{i=2}^{m} A_i\), since we only observe and care about player 1’s return. Define the value function using \(r_1\) as
\[
V^h_{\mu} (s) := \mathbb{E}_{\mu, v} [\sum_{h'=h}^{H} r_1 \mu'(s_{h'}, a_{h'}, b_{h'}) | s_h = s],
\]
and the regret against the minimax value of player 1 as
\[
\text{Regret}(K) := \sum_{k=1}^{K} (V^*_1(s_k^1) - V^h_{\mu, v}(s_k^1)),
\]
where we note that for all state \(s \in S_1\),
\[
V^*_1(s) := \max_{\mu} \min_{v} V^h_{\mu, v}(s) = \min_{v} \max_{\mu} V^h_{\mu, v}(s)
\]
is defined as the minimax value of player 1, which is no larger than the value at the Nash equilibrium of the multi-player general-sum MG.

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 at once we obtain the following corollary from Theorem 2.

**Corollary 3** (Regret bound in multi-player general-sum MG). For any \(p \in (0, 1)\), for large constant \(\iota\) in (4.1), \(i = \log(HSABK/p)\) and a given \(K \in \mathbb{N}_+\), if we run V-ol with our hyperparameter specification and \(G = K^{3/4}\) for player 1 in the multi-player general-sum MG (5.1), then with probability at least \(1 - p\), the regret in \(K\) episodes satisfies
\[
\text{Regret}(K) \leq O \left( SH^2 + H^3 K^{3/4} \sqrt{SA_1} \right).
\]

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 \(\hat{O}(\sqrt{H^3 S^3 \prod_{i=1}^{m} A_i^3 T})\), which depends exponentially on \(m\). On the other hand, since the online informed setting has weaker assumptions than the online agnostic setting, the regret bound of V-ol \(\hat{O}(H^{3/4} K^{3/4} / \sqrt{SA_1})\) 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 \(SH\) entries), rather than update the Q-table (which has \(SH\prod_{i=1}^{m} A_i\) entries) as in Xie et al. (2020), we can also improve the time and space complexity by an exponential factor.

### 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. Also, we expose the problem with V-sp in the online setting, which explains why we make the learning rate decay slower in V-ol.

In the analysis below, we use a superscript \(k\) to signify the corresponding quantities in the \(k\)-th episode. To express \(V^h_k\) in Algorithm 1 compactly, we introduce the following quantities.
\[
a^0_i := \prod_{j=1}^{l} (1 - a_j), \quad a_i^l := a_i \prod_{j=i+1}^{l} (1 - a_j).
\]
Let $t = N_h^k(s)$ and suppose $s$ is previously visited at episodes $k^1, \ldots, k^t \leq k$. Then $V_h^k(s)$ can be expressed as

$$a_0^t H + \sum_{i=1}^t a_i^t (r_h(s, a_h^i, t_h^i) + V_{h+1}^k(s_h^i + \beta_i)).$$

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

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

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 $\tilde{O}(\sqrt{T})$ regret. Concretely,

$$V_h^k(s) - V_h^*(s) \geq \max_{\mu \in \Delta_{A_h}} \sum_{i=1}^t a_i^t \mathbb{D}_{\mu^k, V_h^k} [r_h + \mathbb{P}_h V_{h+1}^k](s) - \mathbb{D}_{\mu^k, V_h^*} [r_h + \mathbb{P}_h V_{h+1}^*](s) \geq \sum_{i=1}^t a_i^t \mathbb{D}_{\mu^k, V_h^k} [r_h + \mathbb{P}_h V_{h+1}^k](s) - \mathbb{D}_{\mu^k, V_h^*} [r_h + \mathbb{P}_h V_{h+1}^*](s) = \sum_{i=1}^t a_i^t (\mathbb{D}_{\mu^k, V_h^k} - \mathbb{D}_{\mu^k, V_h^*}) [r_h + \mathbb{P}_h V_{h+1}^k](s) \geq \sum_{i=1}^t a_i^t (\mathbb{D}_{\mu^k, V_h^k} - \mathbb{D}_{\mu^k, V_h^*})(s) \geq C,$

where the first inequality is due to the V-learning lemma, the second inequality holds by taking a specific $\mu^*$, in the following equality we add and subtract the same term, and the last inequality follows from the above UCB. If the opponent is weak at some step $h \in [H]$ such that for all episodes $k \in [K]$,

$$(\mathbb{D}_{\mu^k, V_h^k} - \mathbb{D}_{\mu^k, V_h^*})(s) \geq C,$$

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

$$\sum_{k=1}^K (V_1^k - V_1^{\mu^k, V}) (s_h^k) = \tilde{O} \left( SH^2 + GH^3 \sqrt{SAK} + G^{-1}KH \right),$$

which is much stronger than merely a sublinear regret if the opponent is weak; in this case, V-ol performs much better than claimed in Theorem 2.

**Regret bound.** 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 and log factors. By definition, we have

$$\delta_h^k := (V_h^k - V_h^{\mu^k, V})(s_h^k) \lesssim GH^2 \sqrt{\sum_{i=1}^t a_i^t} [r_h + \mathbb{P}_h V_{h+1}^{\mu^k, V}](s_h^k) + \sum_{i=1}^t a_i^t \mathbb{D}_{\mu^k, V} [r_h + \mathbb{P}_h V_{h+1}^k](s_h^k).$$

By the same regrouping technique as that in Jin et al. (2018), for any quantity $f^j$ indexed by $i \in [K]$,

$$\sum_{k=1}^K \sum_{i=1}^t a_i^t f^k \leq \sum_{k=1}^K f^k \sum_{t=1}^\infty a_t^k = (1 + 1/GH) \sum_{k=1}^K f^k.$$
Taking $\mathbb{D}_{\mu^k, \nu^k} [r_h + \mathbb{P}_h V_{h+1}^k (s_h^k)]$ as $f^i$ yields
\[
\sum_{k=1}^{K} \sum_{i=1}^{l} \alpha^i_{k} \mathbb{D}_{\mu^k, \nu^k} [r_h + \mathbb{P}_h V_{h+1}^k (s_h^k)] \leq (1 + 1/GH) \sum_{k=1}^{K} \mathbb{D}_{\mu^k, \nu^k} [r_h + \mathbb{P}_h V_{h+1}^k (s_h^k)].
\]
And as a result,
\[
\sum_{k=1}^{K} \delta_{h+1}^k \leq \sum_{k=1}^{K} (GH^2 \sqrt{A_t} + (1 + 1/GH) \delta_{h+1}^k + 1/GH) \mathbb{D}_{\mu^k, \nu^k} [r_h + \mathbb{P}_h V_{h+1}^k (s_h^k)]
\]
\[
\leq \sum_{k=1}^{K} (GH^2 \sqrt{A_t} + (1 + 1/GH) \delta_{h+1}^k + 1/G).
\]
where the last inequality is due to
\[
\mathbb{D}_{\mu^k, \nu^k} [r_h + \mathbb{P}_h V_{h+1}^k (s_h^k)] \leq H.
\]
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
\[
\sum_{k=1}^{K} \delta_{1}^k \leq (1 + 1/GH)^H (GH^2 \sqrt{A_t} + 1/G).
\]
If $G \geq 1$ then the coefficient $(1 + 1/GH)^H \leq e$. By noting
\[
\sum_{k=1}^{K} \sqrt{1/t} = \sum_{k=1}^{K} \sqrt{1/n_i^k} \leq \sqrt{SK},
\]
we obtain
\[
\text{Regret}(K) \leq \sum_{k=1}^{K} \delta_{1}^k \leq GH^3 \sqrt{SAKt} + G^{-1}KH.
\]
For $V_{sp}$ where $G = 1$, the regret bound is linear in $K$ and therefore useless. To address this problem, we introduce the tunable parameter $G \geq 1$ that balances the $\sqrt{K}$ and $K$ terms in the above bound to produce a sublinear regret.

7 Conclusion

In this paper, we study online agnostic learning in Markov games using $V_{ol}$, which is based on the $V_{sp}$ algorithm ((Bai et al., 2020)). $V_{ol}$ achieves $\tilde{O}(K^{3/4})$ regret after $K$ episodes. Furthermore, the regret bound is independent of the size of opponents’ action space. It is still unclear whether we can achieve a $\tilde{O}(K^{1/2})$ regret in the agnostic setting, which is a question worthy of future study.

Another related question is: does $V_{ol}$ (or its variants) also achieve sublinear regret in adversarial MDPs? Given the many technical differences between adversarial MDP and Markov games, it is desirable to figure out how to solve these problems in a unified manner. In addition, the form of the model-free update in $V_{ol}$ should be of independent interest for adversarial MDPs.

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

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 263–272. JMLR. org, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. arXiv preprint arXiv:2002.04017, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. arXiv preprint arXiv:2006.12007, 2020.

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

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

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

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Non-stationary reinforcement learning: The blessing of (more) optimism. Available at SSRN 3397818, 2019.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with Rich Observations via Latent State Decoding. In International Conference on Machine Learning, pages 1665–1674, 2019.

Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. 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.

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

Thomas Jakobs, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11(Apr):1563–1600, 2010.

Zeyu Jia, Lin F Yang, and Mengdi Wang. Feature-based Q-learning for two-player stochastic games. arXiv preprint arXiv:1906.00423, 2019.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In Advances in Neural Information Processing Systems, pages 4863–4873, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. arXiv, pages arXiv–1912, 2019.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

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

Michael L Littman. Friend-or-foe Q-learning in general-sum games. In ICML, volume 1, pages 322–328, 2001.
Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

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

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

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

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

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *nature*, 550(7782):354–359, 2017.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning Zero-Sum Simultaneous-Move Markov Games Using Function Approximation and Correlated Equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Yasin Abbasi Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in neural information processing systems*, pages 2508–2516, 2013.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020.

Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pages 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 2, the optimal policies for $M_k$ are the same, so Lemma 2 indeed implies a lower bound on the regret defined against the best stationary policy in hindsight.

**Lemma 2** (Lower bound for adversarial MDPs). For any horizon $H \geq 2$ and $K \geq 1$, there exists a family of MDPs $M$ 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 $\mu^k$ in episode $k$, we have

$$\sup_{M_k, \ldots, M_{K} \in M} \sup_{\mu} \sum_{k=1}^{K} \left( V_{M_k}^\mu(s_0) - \mathbb{E}_{\mu^k} V_{M_k}^\mu(s_0) \right) \geq \Omega(\min \{\sqrt{2HK}, K\}),$$

where $V_{M_k}^\mu$ 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 $M_{X,Y,\epsilon}$). 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 $\epsilon \in (0, 1)$, the MDP $M_{X,Y,\epsilon}$ 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 $1/2$ each, regardless of the action taken.
   - For any $1 \leq h \leq H - 1$, $s_{0,h}$ transitions to $s_{1,h+1}$ deterministically if $a_h = x_h \oplus y_h$ (“correct state” in combination lock), and transitions to $s_{0,h+1}$ deterministically if $a_h = 1 - x_h \oplus y_h$.
   - For any $1 \leq h \leq H - 1$, $s_{1-h}$ transitions to $s_{1,1}$ 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
   - $r_H(s_{0,H}) \sim \text{Ber}(1/2 + \epsilon)$,
   - $r_H(s_{1-H}) \sim \text{Ber}(1/2 - \epsilon)$.

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

It is straightforward to see that the optimal value function of this MDP is $1/2(1/2 + \epsilon) + 1/2(1/2 - \epsilon) = 1/2$, and the only way to achieve higher reward than $1/2 - \epsilon$ is by following the path of “good states”: $(s_0, s_{y_1,1}, \ldots, s_{y_h,1}, \ldots, s_{y_H,1})$. 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 $Y_k \sim \text{Unif}(\{0, 1\}^H)$ for $1 \leq k \leq K$. We argue that if we provide $M_k := M_{X, Y_k, \epsilon}$ in episode $k$ (with some appropriate choice of $\epsilon$), then the problem is as hard as a $2^H$-armed bandit problem with (minimum) suboptimality gap $\epsilon$, 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 $s \in \{0, 1\}^H$ transitions to $s' \in \{0, 1\}^H$ with probability $1/2$ regardless of the actions taken; and the reward is $r_H \sim \text{Ber}(1/2)$ if $A = X \oplus Y$ and $r_H \sim \text{Ber}(1/2 - \epsilon)$ if $A \neq X \oplus Y$, where $A = \{a_1, \ldots, a_H\}$ are the actions taken in steps 1 through $H$. Indeed, consider the transition starting from $s_{y_h, h}$. Since $y_{h+1} \sim \text{Ber}(1/2)$, 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 $\text{Ber}(1/2)$, and otherwise the reward is $\text{Ber}(1/2 - \epsilon)$. 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 $k = 1, \ldots, K$ do
2:   Sample $Y \sim \text{Unif}(\{0, 1\}^H)$.
3:   Simulate the adversarial MDP algorithm by showing the trajectory $(s_0, s_{y_1, 1}, \ldots, s_{y_H, H})$.
4:   Denote the action sequence by $A = (a_1, \ldots, a_H)$.
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 $M_{X, Y_k, \epsilon}$. The trajectory is drawn from a uniform distribution, which is the same as that generated by $M_{X, Y_k, \epsilon}$. The reward is high, i.e. $\text{Ber}(1/2)$, 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 $\pi_\epsilon(s_{y_h, h}) = x_h \oplus y_h$.

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 and Szepesvári, 2020, Chapter 15) (which corresponds to taking $\epsilon = \epsilon_{H,K} := \min \{ \sqrt{2^H/K}, 1/4 \}$), we get

![Figure 2: $M(X, Y)$: “Combination lock” MDP specified by $X$ and $Y$. For $y \in \{0, 1\}$, $y'$ stands for $1 - y$.](image)
constructed in Lemma 2. Indeed, we augment the MDP prior distribution of MDPs, the above lower bound implies the following minimax lower bound for any adversarial MDP algorithm. Thus the previous lower bound can rewritten as a comparison with the best policy in hindsight:

\[
\sup_{X \in \{0,1\}^H} \sup_{\mu} \mathbb{E}_{Y_1,\ldots,Y_H \sim \text{Unif}\left(\{0,1\}^H\right)} \left[ \sum_{k=1}^{K} \left( V_{M,X,Y_i,[0,H]}^\mu (s_0) - \mathbb{E}_{\mu} V_{M,X,Y_i,[0,H]}^\mu (s_0) \right) \right] \geq \Omega(\min \left\{ \sqrt{2HK}, K \right\}).
\]

The adversarial MDP problem is as hard as the above random sequence of MDPs. Define \( \mathcal{M} := \{ M_{X,Y,[0,H]} : X,Y \in \{0,1\}^H \} \). 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 for any adversarial MDP algorithm.

\[
\sup_{M_1 \in \mathcal{M}} \sup_{\mu} \left[ \sum_{k=1}^{K} \left( V_{M_1}^\mu (s_0) - \mathbb{E}_{\mu} V_{M_1}^\mu (s_0) \right) \right] \geq \Omega(\min \left\{ \sqrt{2HK}, K \right\})
\]

Proof of Theorem 1. With Lemma 2 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 \( \{ \nu^k \} \) such that the transitions and rewards seen by the max-player are exactly equivalent to the MDP \( M_{X,Y,[0,H]} \) constructed in Lemma 2. Indeed, we augment the MDP \( M_{X,Y,[0,H]} \) with a set of min-player actions \( B_h = \{1,2,3,4\} \), and redefine the transition such that from any \( s_{i,h} \) where \( i \in \{0,1\} \) and \( 1 \leq h \leq H-1 \), 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 kernel of the state \( s_{i,h} \) for the hard instance of Markov games.

Such an action set \( B_h \) is powerful enough to reproduce all the possible transitions in the original single-player MDP. We then define \( \nu^k \) as the policy such that the transition follows exactly \( M_{X,Y_i} \). The reward function is determined only by states and thus remains the same. Therefore Lemma 2 implies the following one-sided regret bound for the max-player:

\[
\sup_{\nu} \sup_{\mu} \sum_{k=1}^{K} \left( V_{\nu^k}^\mu (s_0) - \mathbb{E}_{\mu} V_{\nu^k}^\mu (s_0) \right) \geq \Omega(\min \left\{ \sqrt{2HK}, K \right\}),
\]

which is the desired result.

B Proof for the V-ol algorithm

Throughout this section, let \( \epsilon = \log(HSAK/p) \). The following lemma summarizes the key properties of the choice of the learning rate \( a_t \), which are used in the proof below.
Proof. The proof is similar to that of (Bai et al., 2020, Lemma 15), except that we need to deal with an 

1. \( t^{-1/2} \leq \sum_{i=1}^{t} i^{-1/2} a_i^t \leq 2t^{-1/2} \) for all \( t \geq 1 \).

2. \( \sum_{i=1}^{t} (a_i^t)^2 \leq \max_{i \in [t]} a_i^t \leq 2t^{-1}GH \) for all \( t \geq 1 \).

3. \( \sum_{t=i}^{\infty} a_i^t = 1 + (GH)^{-1} \) for all \( i \geq 1 \).

B.1 Upper confidence bound on the minimax value function

Lemma 4 (V-learning lemma). Let \( t = N_h^k(s) \) and suppose state \( s \in S_h \) was previously visited at episodes \( k^1, \ldots, k^l < k \) at the \( h \)th step. Choosing \( \eta_t = \sqrt{\log A/\alpha_t} \), with probability \( 1 - p \), for any \( t \in [K], h \in [H] \) and \( s \in S_h \), there exists a constant \( c \) such that

\[
\max_{\mu \in \Delta_A} \sum_{i=1}^{t} x_i^t \mathbb{D}_{\mu, x_h^t}(r_h + \mathbb{P}_h V_{h+1}^{k}) (s) - \sum_{i=1}^{t} \alpha_i^t \left( r_h(s, a_h^t, b_h^t) + V_{h+1}^{k_i}(s_{h+1}) \right) \leq cGH^2\sqrt{\alpha_t}/t.
\]

Proof. The proof follows from that of (Bai et al., 2020, Lemma 14). The only difference is that here the

learning rate contains a parameter \( G \) such that

\[
R_t^* \leq 2Ha_t^0 \sqrt{\alpha_t} + \frac{3}{2} H \sqrt{\alpha_t} \sum_{i=1}^{t} i^{-1/2} a_i^t + \frac{1}{2} H a_t^0 + \frac{1}{2} H \sqrt{\alpha_t},
\]

\[
\leq 4GH^2\sqrt{\alpha_t}/t + 3H \sqrt{\alpha_t}/t + GH^2/t + 2H \sqrt{GH}/t
\]

\[
\leq cGH^2\sqrt{\alpha_t}/t
\]

for some large constant \( c \).

Lemma 5 (Upper confidence bound). For all \( k \in [K], h \in [H] \) and \( s \in S_h \), with probability \( 1 - p \), \( V_{h}^*(s) \leq V_{h}^k(s) \).

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 \( s \in S_h \) is observed at step \( h \) for the \( i \)th time. Where there

is no ambiguity, we use \( k_i^h \) 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 \( \beta_i \), we have \( \sum_{i=1}^{t} a_i^t \beta_i = \Theta(GH^2\sqrt{\alpha_t}/t) \) by Lemma 3.

Recall that

\[
V_h^k(s) := a_0^k H + \sum_{i=1}^{t} a_i^t \left( r_h(s, a_h^k, b_h^k) + V_{h+1}^{k_i}(s_{h+1}^k) + \beta_i \right),
\]

\[
V_h^*(s) := \mathbb{D}_{\mu_h^*, x_h^*}(r_h + \mathbb{P}_h V_{h+1}^*) (s).
\]

For \( h = H + 1 \) the UCB vacuously holds. To apply backward induction, assume that \( V_{h+1}^* \leq V_{h+1}^k \) holds
entrywise. Since \( \sum_{i=1}^{t} a_i = 1 \), we have for any \( s \in \mathcal{S}_h \), by definition,

\[
V_{h}^s(s) = \max_{\mu \in \Delta_{A_h}} \min_{\upsilon \in \Delta_{\mathcal{B}_h}} \mathbb{D}_{\mu, \nu} [r_h + P_h V_{h+1}^s](s) \\
= \max_{\mu \in \Delta_{A_h}} \sum_{t=1}^{T} a_t \min_{\upsilon \in \Delta_{\mathcal{B}_h}} \mathbb{D}_{\mu, \upsilon} [r_h + P_h V_{h+1}^s](s) \\
\leq \max_{\mu \in \Delta_{A_h}} \sum_{t=1}^{T} a_t \mathbb{D}_{\mu, \upsilon} [r_h + P_h V_{h+1}^s](s) \\
\leq \max_{\mu \in \Delta_{A_h}} \sum_{t=1}^{T} a_t \mathbb{D}_{\mu, \upsilon} [r_h + P_h V_{h+1}^s](s) \\
\leq V_{h}^k(s),
\]

where the last inequality holds with probability \( 1 - p \) by the V-learning lemma (Lemma 4) and our choice of \( \beta_i \). Inductively we have \( V_{h}^s(s) \leq V_{h}^k(s) \) for all \( k \in [K], h \in [H] \) and \( s \in \mathcal{S}_h \). \( \square \)

**B.2 Proof of Theorem 2**

**Proof.** In the proof below, we use ‘\( \leq \)’ to denote ‘\( \leq \)’ hiding some constants. Recall that

\[
V_{h}^{\mu, \nu}^k(s_h) = \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h).
\]

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

\[
\delta_h^k = a_0^k H + \sum_{i=1}^{t} a_i \left( r_h(s_h, a_h, b_h^k) + V_{h+1}^{k}(s_{h+1}^k) + \beta_i \right) - \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) \\
= a_0^k H + \sum_{i=1}^{t} a_i \left( r_h(s_h, a_h, b_h^k) + V_{h+1}^{k}(s_{h+1}^k) + \beta_i \right) - \sum_{i=1}^{t} a_i \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) \\
+ \sum_{i=1}^{t} a_i \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) - \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) \\
\leq a_0^k H + GH^2 / \sqrt{V/t} + \sum_{i=1}^{t} a_i \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) - \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h),
\]

where in the second equality we add and subtract the same term, and the last inequality follows from the property of \( \beta_i \) that \( \sum_{i=1}^{t} a_i / \beta_i = \Theta(GH^2 / \sqrt{V/t}) \) and the fact that by the Azuma-Hoeffding inequality,

\[
\sum_{i=1}^{t} a_i \left( r_h(s_h, a_h, b_h^k) + V_{h+1}^{k}(s_{h+1}^k) + \beta_i \right) - \sum_{i=1}^{t} a_i \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) \leq G / \sqrt{V/t}.
\]

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

\[
\sum_{k=1}^{K} \sum_{i=1}^{t} a_i \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) \leq \sum_{k'=1}^{K} \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) \sum_{t=b_h}^{\infty} a_t \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h) \\
\leq (1 + 1 / GH) \sum_{k=1}^{K} \mathbb{D}_{\mu, \upsilon}^k [r_h + P_h V_{h+1}^k](s_h),
\]

18
Substituting the above back into the bound on $\delta^k_h$ and taking sum over $k \in [K]$, we obtain

\[
\sum_{k=1}^K \delta^k_h \lesssim \sum_{k=1}^K \left( a^0_t H + GH^2 \sqrt{A_t l} + (1 + 1/GH) D_{\mu,h,t}^k |r_h + P_h V_{h+1}^k|(s^k_h) - D_{\mu,h,t}^k |r_h + P_h V_{h+1}^k|(s^k_h) \right) \\
= \sum_{k=1}^K \left( a^0_t H + GH^2 \sqrt{A_t l} + (1 + 1/GH)(\delta^k_{h+1} + \gamma^k_h) + 1/GH D_{\mu,h,t}^k |r_h + P_h V_{h+1}^k|(s^k_h) \right) \\
\leq \sum_{k=1}^K \left( a^0_t H + GH^2 \sqrt{A_t l} + (1 + 1/GH)(\delta^k_{h+1} + \gamma^k_h) + 1/G \right),
\]

where we define the martingale difference term $\gamma^k_h := D_{\mu,h,t}^k [P_h(V_{h+1}^k - V_{h+1}^h)](s^k_h) - (V_{h+1}^k - V_{h+1}^h)(s^k_{h+1})$ and the last inequality follows from that

\[
D_{\mu,h,t}^k |r_h + P_h V_{h+1}^k|(s^k_h) \leq H.
\]

Recursively,

\[
\sum_{k=1}^K \delta^k_t \lesssim (1 + 1/GH)^H \sum_{k=1}^K \sum_{h=1}^H \left( a^0_t H + GH^2 \sqrt{A_t l} + (1 + 1/GH)\gamma^k_h + 1/G \right).
\]

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

\[
\sum_{k=1}^K a^0_t H \leq \sum_{k=1}^K H \cdot 1(n^k_h = 0) \leq SH_t,
\]

\[
\sum_{k=1}^K GH^2 \sqrt{A_t l/n^k_h} = GH^2 \sqrt{A_t} \sum_{k=1}^K 1/n^k_h \leq GH^2 \sqrt{A_t} \sum_{k \in S_h} \sum_{n=1}^{n^k_h(s)} \sqrt{1/n} \lesssim GH^2 \sqrt{SAKi},
\]

\[
\sum_{k=1}^K \sum_{h=1}^H \gamma^k_h \lesssim H^3 Ki,
\]

where the second line follows from a pigeonhole argument and the third line follows from the Azuma-Hoeffding inequality. Combining the above bounds, we obtain

\[
\text{Regret}(K) \leq \sum_{k=1}^K \delta^k_t \lesssim \text{SH}^2 + GH^3 \sqrt{SAKi} + G^{-1}KH.
\]

\[\square\]