# Learning Markov Games with Adversarial Opponents: Efficient Algorithms and Fundamental Limits

**Qinghua Liu** [*1] **Yuanhao Wang** [*1] **Chi Jin** [1]

## Abstract

An ideal strategy in zero-sum games should not only grant the player an average reward no less than the value of Nash equilibrium, but also exploit the (adaptive) opponents when they are suboptimal. While most existing works in Markov games focus exclusively on the former objective, it remains open whether we can achieve both objectives simultaneously. To address this problem, this work studies no-regret learning in Markov games with adversarial opponents when *competing against the best fixed policy in hindsight*. Along this direction, we present a new complete set of positive and negative results: When the policies of the opponents are revealed at the end of each episode, we propose new efficient algorithms achieving $\sqrt{K}$-regret bounds when either (1) the baseline policy class is small or (2) the opponent's policy class is small. This is complemented with an exponential lower bound when neither conditions are true. When the policies of the opponents are not revealed, we prove a statistical hardness result even in the most favorable scenario when both above conditions are true. Our hardness result is much stronger than the existing hardness results which either only involve computational hardness, or require further restrictions on the algorithms.

## 1. Introduction

Multi-agent reinforcement learning (MARL) studies how multiple players sequentially interact with each other and the environment to maximize the cumulative rewards. Recent years have witnessed inspiring breakthroughs in the application of multi-agent reinforcement learning to various challenging AI tasks, including, but not limited to, GO (Silver et al., 2016; 2017), Poker (Brown & Sandholm, 2019), real-time strategy games (e.g., StarCraft and Dota) (Vinyals et al., 2019; OpenAI, 2018), autonomous driving (Shalev-Shwartz et al., 2016), decentralized controls or multi-agent robotics systems (Brambilla et al., 2013), as well as complex social scenarios such as hide-and-seek (Baker et al., 2020).

Despite its great empirical success, MARL still suffers from limited theoretical understanding with many fundamental questions left open. Among them, one central and challenging question is how to exploit the (adaptive) suboptimal opponents while staying invulnerable to the optimal opponents. Achieving this objective requires a solution concept beyond Nash equilibria. As a motivating example, we consider the game of rock-paper-scissors with a suboptimal opponent who plays rock in the first $K/2$ games and then switches to paper in the next $K/2$ games. A strategic player in this case should be able to learn from the behavior of the opponent and exploit it to get a return of $\Omega(K)$. In contrast, playing a Nash equilibrium (which plays all actions uniformly) only yields an average return of zero.

In classical normal-form games (which can be viewed as special cases of MARL without transition and states), the question of exploiting adaptive opponents has been extensively studied under the framework of no-regret learning, where the agent is required to compete against the best fixed policy in hindsight even when facing adversarial opponents (see e.g., Cesa-Bianchi & Lugosi, 2006). On the other hand, addressing general MARL brings a number of new challenges such as unknown environment dynamics and sequential correlations between the player and the opponents. Consequently, all existing results (e.g., Brafman & Tennenholtz, 2002; Wei et al., 2017; Tian et al., 2021; Jin et al., 2021) have only focused on competing against Nash equilibria when facing adversarial opponents. This motivates us to ask the following question for MARL:

**Can we compete against the best fixed policy in hindsight and achieve no-regret learning in MARL?**

In this paper, we consider two-player zero-sum Markov games (Shapley, 1953; Littman, 1994) as a model for MARL, and address the above question by providing a complete set of positive and negative results as follows. We refer to *general policies* as policies which can depend on

[*]Equal contribution [1]Princeton University, New Jersey, USA. Correspondence to: Qinghua Liu <qinghual@princeton.edu>.

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

the entire history, in contrast to *Markov policies*, which can only depend on the state at the current step.

**Statistical efficiency (standard setting).** We first consider the most standard setting, in which only the actions of the opponents are observed, and prove an exponential lower bound for the regret. Importantly, the lower bound holds even if the baseline policy class only contains Markov policies and the opponent only alternates between a small number of Markov policies. Besides, this hardness result is much stronger than the existing ones which either only involve computational hardness (Bai et al., 2020), or require further restrictions on the algorithms (Tian et al., 2021). The proof of the lower bound builds upon the key observation that we can simulate any POMDP/latent MDP by a Markov game of similar size and an opponent playing general/Markov policies. This directly implies no-regret learning in MGs is no easier than learning POMDPs/latent MDPs which is statistically intractable in general (Jin et al., 2020; Kwon et al., 2021).

**Statistical efficiency (revealed-policy setting).** Given that only observing actions of opponents is insufficient for achieving sublinear regret, we then consider a setting more advantageous to the learner, in which the opponent reveals the policy she played at the end of each episode.

- When baseline policies—the set of policies we are competing against in the definition of regret (see Definition 3.1)—are Markov policies, we propose ***O**ptimistic **P**olicy **EXP3*** (OP-EXP3, Algorithm 1) that has $\tilde{\mathcal{O}}(\sqrt{H^4S^2AK})$-regret even when the opponent can play arbitrary general (history-dependent) policies, where $H$ is the length of each episode, $S$ is the number of states, $A$ is the number of actions, and $K$ is the number of episodes.

- When baseline policies are general policies, We further propose adaptive OP-EXP3 (Algorithm 2) that achieves regret $\tilde{\mathcal{O}}(\sqrt{H^4S^2AK} + \sqrt{|\Psi^\star|SAH^3K} + \sqrt{|\Psi^\star|^2H^2K})$ when the opponent only chooses policies from an unknown policy class $\Psi^\star$.

- Finally, we complement our upper bounds with an exponential lower bound for competing against general policies, which holds even when the opponent only plays deterministic Markov policies.

**Computational efficiency.** Finally, we prove that achieving sublinear regret is computationally hard even in the very favorable setting where (a) the learner only competes against the best fixed Markov policy in hindsight, (b) the opponent only chooses policies randomly from a known small set of Markov policies and reveals the policy she played at the end of each episode, (c) the MG model is known. We emphasize that this computational hardness holds under very weak conditions as stated above, and applies to all the settings studied in this paper.

To summarize, we provide a complete set of results including both efficient algorithms and fundamental limits for no-regret learning in Markov games with adversarial opponents. We refer the reader to Table 1 for a brief summary of our main results.

## 2. Related Work

**Learning Nash equilibria in Markov games.** There has been a long line of literature focusing on learning the Nash equilibrium of Markov games when either the dynamics are known, or the amount of collected data goes to infinity (Littman, 1994; Hu & Wellman, 2003; Hansen et al., 2013; Lee et al., 2020). Later works have considered self-play algorithms that incorporate exploration and can find Nash equilibrium in Markov games with unknown dynamics (Wei et al., 2017; Bai et al., 2020; Bai & Jin, 2020; Xie et al., 2020; Liu et al., 2021).

When the algorithm is only able to control one player and the other player is potentially adversarial, Brafman & Tennenholtz (2002) proposed the R-max algorithm, and showed that it is able to obtain average value close to the Nash value. Later works (Wei et al., 2017; Tian et al., 2021; Jin et al., 2021) obtain similar or improved results also for comparing to the Nash value.

**Learning latent MDPs.** In latent MDPs, sometimes also referred as multi-model MDPs, a latent variable is drawn from a fixed distribution at the start of each episode, and the dynamics of the MDP would be a function of this latent variable. Steimle et al. (2021) has shown that finding the optimal Markov policy in the latent MDP problem is computational hard; Kwon et al. (2021) considered reinforcement learning in latent MDPs, providing both statistical lower bounds for the general case and sample complexity upper bounds under further assumptions. Latent MDPs, and in fact POMDPs (Smallwood & Sondik, 1973; Azizzadenesheli et al., 2016; Jin et al., 2020) in general, can be simulated using Markov games with adversarial opponents as proved in this paper; thus learning latent MDPs can be viewed as a special case of the setting considered in this paper.

**Adversarial MDPs.** Another line of work focuses on the single-agent adversarial MDP setting where the transition or the reward function is adversarially chosen for each episode. When the adversary can arbitrarily alter the transition, Abbasi Yadkori et al. (2013) prove that no-regret learning is computationally at least as hard as learning parity with noise. Later work by Bai et al. (2020) adapt similar hard instance

| Baseline Policies | Opponent's Policies | Standard Setting | Revealed-policy Setting |
|---|---|---|---|
| Markov policies | General policies | $\Omega(\min\{K, 2^H\}/H)$ | $\tilde{\mathcal{O}}(\sqrt{H^4S^2AK})$ |
| General policies | Finite class $\Psi^\star$ | | $\tilde{\mathcal{O}}(\sqrt{H^4S^2AK} + \sqrt{\lvert\Psi^\star\rvert SAH^3K} + \sqrt{\lvert\Psi^\star\rvert^2H^2K})$ |
| | Markov policies | | $\Omega(\min\{K, 2^H\})$ |

*Table 1.* A summary of the main results. Baseline policies refer to the policies the algorithm competes against in the definition of regret (see Definition 3.1). General policies include both Markov and history-dependent policies.

for Markov games and prove that achieving sublinear regret in MGs against adversarial opponents is also computationally hard. On the other hand, if the transition is fixed and the adversary is only allowed to alter the reward function, sublinear regret can be achieved by various algorithms (Jin et al., 2019; Zimin & Neu, 2013; Rosenberg & Mansour, 2019; Shani et al., 2020) in competing against the best Markov policy in hindsight.

**Matrix games and extensive form games.** For matrix games, it is well known that playing EXP-style algorithms would allow one to compete with the best policy (action profile) in hindsight (see e.g., Cesa-Bianchi & Lugosi, 2006). For extensive form games (EFGs), similar no-regret guarantees can be achieved via counterfactual regret minimization (Zinkevich et al., 2007) or online convex optimization (Gordon, 2007; Farina et al., 2020; Farina & Sandholm, 2021; Kozuno et al., 2021). EFGs can be viewed a special subclass of MGs where the transition admits a strict tree structure. Therefore, results for EFGs do not directly apply to MGs.

## 3. Preliminaries

In this paper, we consider Markov Games (MGs, Shapley, 1953; Littman, 1994), which generalize the standard Markov Decision Processes (MDPs) into the multi-player setting, where each player seeks maximizing her own utility.

Formally, we study the tabular episodic version of two-player zero-sum Markov games, which is specified by a tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$. Here $\mathcal{S}$ denotes the state set with $|\mathcal{S}| \le S$. $\mathcal{A} = \mathcal{A}_{\max} \times \mathcal{A}_{\min}$ (with $|\mathcal{A}| \le A$) denotes the action-pair set that is equal to the Cartesian product of the action set of the max-player $\mathcal{A}_{\max}$ and the action set of the min-player $\mathcal{A}_{\min}$. $H$ denotes the length of each episode. $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ denotes a collection of transition matrices, so that $\mathbb{P}_h(\cdot|s, \mathbf{a})$ gives the distribution of the next state if action-pair $\mathbf{a} \in \mathcal{A}$ is taken at state $s$ at step $h$. $r = \{r_h\}_{h\in[H]}$ denotes a collection of expected reward functions, where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the expected reward function at step $h$. This reward represents both the gain of the max-player and the loss of the min-player, making the problem a zero-sum Markov game. For cleaner presentation, we assume the reward function is known in this work.[1]

In each episode, the environment starts from a *fixed initial state* $s_1$. At step $h \in [H]$, both players observe state $s_h \in \mathcal{S}$, and then pick their own actions $a_{h,\max} \in \mathcal{A}_{\max}$ and $a_{h,\min} \in \mathcal{A}_{\min}$ simultaneously. After that, both players observe the action of their opponent, receive reward $r_h(s_h, \mathbf{a}_h)$, and then the environment transitions to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, \mathbf{a}_h)$. The episode terminates immediately once $s_{H+1}$ is reached.

We use $\tau_h = (s_1, \mathbf{a}_1, \ldots, s_{h-1}, \mathbf{a}_{h-1}, s_h) \in (\mathcal{S}\times\mathcal{A})^{h-1} \times \mathcal{S}$ to denote a trajectory from step 1 to step $h$, which includes the state but excludes the action at step $h$. We use box brackets to denote the concatenation of trajectories, e.g., $[\tau_h, \mathbf{a}_h, s_{h+1}] \in (\mathcal{S} \times \mathcal{A})^h \times \mathcal{S}$ gives a trajectory from step 1 to step $h+1$ by concatenating $\tau_h$ with an action-state pair $(\mathbf{a}_h, s_{h+1})$.

**Policy.** We consider two classes of policies: Markov policies and general policies. A **Markov policy** $\mu = \{\mu_h : \mathcal{S} \to \Delta_{\mathcal{A}_{\max}}\}_{h\in[H]}$ of the max-player is a collection of $H$ functions, each mapping from a state to a distribution over actions. (Here $\Delta_{\mathcal{A}_{\max}}$ is the probability simplex over action set $\mathcal{A}_{\max}$.) Similarly, a Markov policy of the min-player is of form $\nu = \{\nu_h : \mathcal{S} \to \Delta_{\mathcal{A}_{\min}}\}_{h\in[H]}$. Different from Markov policies, a **general policy** can choose actions depending on the entire history of interactions. Formally, a general policy $\mu = \{\mu_h : (\mathcal{S}\times\mathcal{A})^{h-1}\times\mathcal{S} \to \Delta_{\mathcal{A}_{\max}}\}_{h\in[H]}$ of the max-player is a collection of $H$ functions where each function maps a trajectory to a distribution over actions. The definition of general policies of the min-player follows similarly. We remark that Markov policies are special cases of general policies, which pick actions only conditioning on the current state.

**Value function.** Given any pair of general policies $(\mu, \nu)$, we use $V_1^{\mu\times\nu}(s_1)$ to denote its value function, which is equal to the expected cumulative rewards received by the max-player, if the game starts at state $s_1$ at the $1^{\text{th}}$ step and the max-player and the min-player follow policy $\mu$ and $\nu$

[1] Our results immediately generalize to unknown reward functions effortlessly, since learning the transitions is more difficult than learning the rewards in tabular MGs.

respectively:

$$V_1^{\mu\times\nu}(s_1) := \mathbb{E}_{\mu\times\nu}\left[\sum_{h=1}^{H} r_h(s_h, \mathbf{a}_h)\middle| s_1\right], \quad (1)$$

where the expectation is taken with respect to the randomness of $\mathbf{a}_1, s_2, \mathbf{a}_2, \ldots, s_H, \mathbf{a}_H$.

**Best response and Nash equilibrium.** Given any general policy of the max-player $\mu$, there exists a *best response* of the min-player $\nu^\dagger(\mu)$ so that $V_1^{\mu\times\nu^\dagger(\mu)}(s_1) = \inf_\nu V_1^{\mu\times\nu}(s_1)$. For brevity of notations, we denote $V_1^{\mu,\dagger} := V_1^{\mu\times\nu^\dagger(\mu)}$. By symmetry, we can also define $\mu^\dagger(\nu)$ and $V_1^{\dagger,\nu}$. Moreover, previous works (e.g., Filar & Vrieze, 2012) prove that there exist policies $\mu^\star$, $\nu^\star$ that are optimal against the best responses of the opponents, in the sense that

$$V_1^{\mu^\star,\dagger}(s_1) = \sup_\mu V_1^{\mu,\dagger}(s_1), \quad V_1^{\dagger,\nu^\star}(s_1) = \inf_\nu V_1^{\dagger,\nu}(s_1).$$

We refer to such strategies $(\mu^\star, \nu^\star)$ as the Nash equilibria of the Markov game. Importantly, any Nash equilibrium satisfies the following minimax theorem [2]:

$$\sup_\mu \inf_\nu V_1^{\mu\times\nu}(s_1) = V_1^{\mu^\star\times\nu^\star}(s_1) = \inf_\nu \sup_\mu V_1^{\mu\times\nu}(s_1).$$

The minimax theorem above directly implies the value function of Nash equilibria is unique, which we denote as $V_1^\star(s_1)$. Furthermore, it is also known that there always exists a Markov Nash equilibrium in the sense that both $\mu^\star$ and $\nu^\star$ are Markov. Intuitively, a Nash equilibrium gives a solution in which no player can benefit from unilateral deviation.

**Learning objective.** In this work, we study no-regret learning of Markov games with adversarial opponents, and measure the performance of an algorithm by its regret against the best fixed policy in hindsight from a prespecified set of policies. From now on, we refer to this policy set as the *baseline policy class*, and denote it by $\Phi^\star$.

**Definition 3.1** (Regret)**.** Let $(\mu^k, \nu^k)$ denote the policies deployed by the algorithm and the opponent in the $k^{\text{th}}$ episode. After a total of $K$ episodes, the regret is defined as

$$\text{Regret}_{\Phi^\star}(K) = \max_{\mu\in\Phi^\star} \sum_{k=1}^{K} (V_1^{\mu\times\nu^k} - V_1^{\mu^k\times\nu^k})(s_1). \quad (2)$$

When the baseline policy class $\Phi^\star$ includes all the general policies, we will omit subscript $\Phi^\star$ and simply write $\text{Regret}(K)$.

[2] We remark that the minimax theorem for MGs is different from the one for matrix games, i.e. $\max_x \min_y x^\top A y = \min_y \max_x x^\top A y$ for any matrix $A$, because $V_1^{\mu\times\nu}(s_1)$ is in general not bilinear in $\mu, \nu$.

Compared to previous works (e.g., Brafman & Tennenholtz, 2002; Wei et al., 2017; Tian et al., 2021; Jin et al., 2021) that only pursue achieving the Nash value, i.e., considering the following version of regret

$$\sum_{k=1}^{K} (V_1^\star - V_1^{\mu^k\times\nu^k})(s_1), \quad (3)$$

our regret defined in (2) is a much stronger criterion because it forces the algorithm to exploit the opponents to achieve higher value than Nash equilibria whenever the opponent is exploitable. In stark contrast, the regret defined in (3) only requires the algorithm itself to be invulnerable. Moreover, if the baseline policy class includes all the Markov policies, then the regret defined in (2) is an upper bound for the latter one because there always exists a Markov Nash equilibrium as mentioned before.

Finally, observe that the regret defined in (2) does not depend on the payoff function of the min-player, so it is still well-defined in the general-sum setting. Actually, all the results derived in this work can be directly extended to the general-sum setting, although the current paper assumes zero MGs for cleaner presentation and more direct comparison to previous works.

## 4. Results for the Standard Setting

In this section, we consider the standard setting where the opponent only reveals her actions to the learner during their interaction. We show that achieving low regret in this setting is impossible in general even if (a) the baseline policy class consists of Markov policies, *and* (b) the opponent sticks to a fixed general policy or only alternates between $H$ different Markov policies. Our hardness results build on the generality of Markov games, i.e., the ability to simulate POMDPs and latent MDPs with specially designed opponents.

### 4.1. Against opponents playing a fixed general policy

To begin with, we show competing with the best Markov policy in hindsight is statistically hard when the opponent keeps playing a fixed general policy.

**Theorem 4.1.** *There exists a Markov game with $S, A = \mathcal{O}(1)$ and an opponent playing a fixed unknown general policy, such that the regret for competing with the best fixed Markov policy in hindsight is $\Omega(\min\{K, 2^H\})$.*

Theorem 4.1 claims that even in a Markov game of constant size, if the learner is only able to observe the opponent's actions instead of the opponent's policies, then there exists a regret lower bound exponential in the horizon length $H$ for competing with the best fixed Markov policy in hindsight when the opponent plays a fixed unknown general policy.

The proof relies on the fact that a POMDP can be simulated

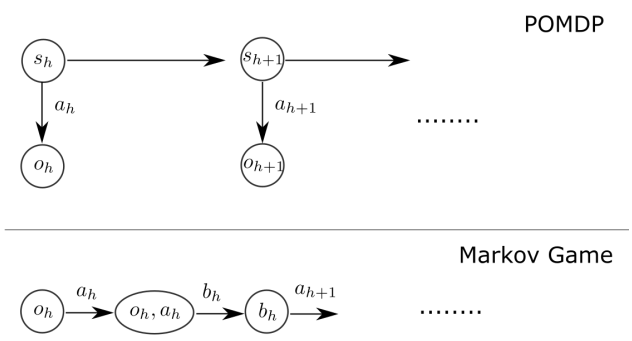


*Figure 1.* Simulating a POMDP using a Markov game with a history-dependent opponent. The player and the opponent dictates the transition dynamics in turn. The opponent, which has access to the full history $\{o_1, a_1, o_2, \cdots, o_h, a_h\}$, can always sample her action $b_h$ from $\mathbb{P}[o_{h+1} = \cdot | o_1, a_1, o_2, \cdots, o_h, a_h]$ in the POMDP above, and the next state is exactly equal to $b_h$.

by a Markov game of similar size and with an opponent who plays a fixed history-dependent policy.

**Proposition 4.2** (POMDP $\subseteq$ MG + opponent playing a general policy)**.** *A POMDP with $S$ hidden states, $A$ actions, $O$ observations, and episode length $H$ can be simulated by a Markov game with opponent playing **a fixed general policy**, where the Markov game has $OA + O$ states, $A$ actions for the learning agent, $O$ actions for the opponent, and episode length $2H$.*

The idea of simulating a POMDP is demonstrated in Figure 1: the opponent dictates the next state every two time steps, and since the opponent knows the full trajectory $\{o_1, a_1, o_2, \cdots, o_h, a_h\}$, she can choose $b_h$ according to the conditional distribution of $o_{h+1}$ given $\{o_1, a_1, o_2, \cdots, o_h, a_h\}$ in the POMDP. Thus we can simulate the POMDP with a Markov game whose number of states and actions are polynomially related to the original POMDP. We remark that in POMDPs the reward is typically included in the observation so here we do not need to handle it separately. A detailed proof of Proposition 4.2 is provided in Appendix A.2.

Given that there exists exponential regret lower bound for learning POMDPs (e.g., Jin et al., 2020), Proposition 4.2 immediately implies no-regret learning in Markov games is in general intractable if the opponent plays a fixed general policy. The proof is a straightforward combination of Proposition 4.2 and the hard instance constructed in Jin et al. (2020), which can be found in Appendix A.1.

### 4.2. Against opponents playing Markov policies

Theorem 4.1 shows that it is statistically hard to compete with the best Markov response to a non-Markov opponent, which is in stark contrast to the case where the opponent plays a fixed Markov policy and the Markov game can be reduced to a single-agent MDP. However, when the opponent is able to choose from a small set of Markov policies, the task of competing with the best Markov policy in hindsight becomes intractable again.

**Theorem 4.3.** *There exists a Markov game with $S, A = \mathcal{O}(H)$ and an opponent who chooses policy uniformly at random from an unknown set of $H$ Markov policies in each episode, such that the regret for competing with the best fixed Markov policy in hindsight is $\Omega(\min\{K, 2^H\}/H)$.*

Theorem 4.3 claims that even restricting the opponent to only play a finite number of Markov policies is insufficient to circumvent the exponential regret lower bound for competing with the best Markov policy in hindsight, as long as the opponent only reveals her actions to the learner.

The proof of Theorem 4.3 utilizes the following fact that we can simulate a latent MDP by a Markov game of similar size and an opponent who only plays a small set of Markov policies.

**Proposition 4.4** (Latent MDP $\subseteq$ MG + opponent playing multiple Markov policies)**.** *A latent MDP with $L$ latent variables, $S$ states, $A$ actions, and episode length $H$ and binary rewards can be simulated by a Markov game with opponent playing policies chosen from a set of $L$ Markov policies, where the Markov game has $SA + S$ states, $A$ actions for the learning agent, $2S$ actions for the opponent, and episode length $2H$.*

The proof of Proposition 4.4 is deferred to Appendix A.4, which is in a similar spirit to Proposition 4.2. Proposition 4.2 and 4.4 can be alternatively characterized by the Venn diagram in Figure 2.

Combining Proposition 4.4 with the hardness instance for learning latent MDPs (Kwon et al., 2021) immediately implies the exponential lower bound for playing against Markov opponents in Theorem 4.3. A detailed proof is provided in Appendix A.3.

## 5. Results for the Revealed-policy Setting

In this section, we study the setting where the opponent reveals the policy she just played to the learner at the end of each episode. Formally, in each round of interaction: first the learner and the opponent choose their policies $\mu$ and $\nu$ simultaneously, then an episode is played following $\mu \times \nu$, and after that the learner gets to observe the opponent policy $\nu$. For this setting, we propose two algorithms with $\sqrt{K}$-regret upper bounds, when either the log-cardinality

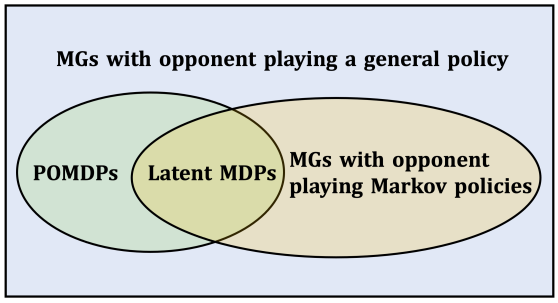


*Figure 2.* Relation between Markov games (reveal action only), latent MDPs and POMDPs.

of the baseline policy class or the cardinality of the opponent's policy class is small. This is complemented with an exponential lower bound when neither conditions are true.

### 5.1. Finite baseline policy class $\Phi^\star$

We first consider the case when the baseline policy class $\Phi^\star$ to compete with is finite but the opponent's policy class is arbitrary. Importantly, we allow both the opponent's polices and the baseline polices to be non-Markov (history-dependent).

**Algorithm.** We propose OP-EXP3 (Algorithm 1), which represents **O**ptimistic **P**olicy **EXP3**, for no-regret learning in this setting. At a high level, OP-EXP3 performs any-time EXP3 with optimistic gradient estimate in the baseline policy class $\Phi^\star$ by viewing each baseline policy as an"action". Specifically, OP-EXP3 maintains a distribution $\mathbf{p}$ over the baseline policy class, and in each episode $k$

- **Interaction** (Line 4-5). The learner samples a policy $\mu^k$ from $\Phi^\star$ according to $\mathbf{p}^k$ and the opponent chooses her policy $\nu^k$ simultaneously. Then a trajectory is sampled by following $\mu^k \times \nu^k$.
- **Optimistic EXP3** (Line 6-7). The opponent's policy $\nu^k$ is revealed to the learner, and for every baseline policy $\mu$ in $\Phi^\star$, the learner computes an optimistic estimate of the value function of $\mu \times \nu^k$ by using the **O**ptimistic **P**olicy **E**valuation (OPE) subroutine. Then the EXP3 update is incurred with the optimistic value estimates as the negative gradient.
- **Model estimate update** (Line 8). Using the newly collected data, we update the empirical estimate of the MG model.

In Subroutine 1, we formally describe the optimistic policy evaluation step. In brief, it utilizes the Bellman equation for general policies to perform dynamic programming from step $H$ to step 1, by using the empirical transition and additionally adding bonus to ensure optimism.

**Algorithm 1** ***O**ptimistic **P**olicy **EXP3***

1: **input**: bonus function $\beta : \mathbb{N} \to \mathbb{R}$, learning rate $(\eta_k)_{k=1}^K$, baseline policy class $\Phi^\star$
2: **initialize**: initial distribution $\mathbf{p}^1 \in \mathbb{R}^{|\Phi|}$ to be uniform over $\Phi$, visitation counters $N_h(s,\mathbf{a}) = N_h(s,\mathbf{a},s') = 0$ for all $(s,\mathbf{a},\mathbf{a}',h)$
3: **for** $k = 1,\dots,K$ **do**
4: the learner samples $\mu^k \sim \mathbf{p}^k$ and the adversary chooses $\nu^k$ *simultaneously*
5: follow $\pi^k = \mu^k \times \nu^k$ to sample $\{s_h^k, \mathbf{a}_h^k, r_h^k\}_{h=1}^H$
*# optimistic EXP3*
6: observe $\nu^k$, and for all $\mu \in \Phi^\star$ compute $\overline{V}_1^{\mu\times\nu^k}(s_1) = \mathrm{OPE}(N,\beta,\mu\times\nu^k)$
7: then update $\mathbf{p}^{k+1}(\mu) \propto \exp(\eta_k \cdot \sum_{t=1}^k \cdot \overline{V}_1^{\mu\times\nu^t}(s_1))$
*# update the counters*
8: for all $h \in [H]$: $N_h(s_h^k,\mathbf{a}_h^k) \leftarrow N_h(s_h^k,\mathbf{a}_h^k)+1$ and $N_h(s_h^k,\mathbf{a}_h^k,s_{h+1}^k) \leftarrow N_h(s_h^k,\mathbf{a}_h^k,s_{h+1}^k)+1$
9: **end for**

**Subroutine 1** ***O**ptimistic **P**olicy **E**valuation* $(N,\beta,\pi)$

initialize $V_{H+1}(\tau_{H+1}) = 0$ for all $\tau_{H+1}$
**for** $(s,\mathbf{a},h,s') \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$ **do**

$$\widehat{\mathbb{P}}_h(s' \mid s,\mathbf{a}) = \begin{cases} \frac{N_h(s,\mathbf{a},s')}{N_h(s,\mathbf{a})}, & \text{if } N_h(s,\mathbf{a}) \neq 0 \\ 1/S, & \text{otherwise} \end{cases}$$

**end for**
**for** $h = H,\dots,1$ **do**
**for** all $\tau_h = (s_1,\mathbf{a}_1,\dots,s_h) \in (\mathcal{S}\times\mathcal{A})^{h-1}\times\mathcal{S}$ **do**

$$Q_h(\tau_h,\mathbf{a}) = \mathbb{E}_{s'\sim\widehat{\mathbb{P}}_h(\cdot\mid s_h,\mathbf{a})}\left[V_{h+1}([\tau_h,\mathbf{a},s'])\right] + r_h(s_h,\mathbf{a}) + \beta(N_h(s_h,\mathbf{a}))$$

$$Q_h(\tau_h,\mathbf{a}) = \min\{Q_h(\tau_h,\mathbf{a}), H-h+1\}$$

$$V_h(\tau_h) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot\mid\tau_h)}[Q_h(\tau_h,\mathbf{a})]$$

**end for**
**end for**
**return** $V_1(s_1)$

**Theoretical guarantee.** Below we present the main theoretical guarantee for OP-EXP3.

**Theorem 5.1.** *Let $c$ be a large absolute constant. In Algorithm 1, choose $\eta_k = \sqrt{\log|\Phi^\star|/(kH^2)}$ and $\beta(n) = \sqrt{H^2S\iota/\max\{n,1\}}$ where $\iota = c\log(SAHK/\delta)$. Then with probability at least $1-\delta$, for all $k \in [K]$*

$$\mathrm{Regret}_{\Phi^\star}(k) \le \mathcal{O}\left(\sqrt{kH^2\log|\Phi^\star|} + \sqrt{kS^2AH^4\iota^2}\right).$$

Theorem 5.1 claims that OP-EXP3 with standard UCB-bonus achieves $\mathcal{O}(\sqrt{k})$-regret with high probability, when competing with the best policy in hindsight in the baseline class. Notably, the regret only depends logarithmically on the cardinality of the baseline class and is independent of the opponent's policy class. In particular, if we choose the baseline policy class to be the collections of all determin-

istic[3] Markov policies ($|\Phi^\star| = A^{SH}$), then Theorem 5.1 immediately implies $\mathcal{O}\left(\sqrt{kS^2AH^4\iota^2}\right)$ regret upper bound for competing with the best Markov policy in hindsight. Moreover, it further implies the same regret upper bound for competing against the value of Nash equilibria, i.e., the regret in equation (3), because there always exists a Markov Nash equilibrium. The proof of Theorem 5.1 can be found in Appendix B.1.

### 5.2. Finite unknown opponent policy class $\Psi^\star$

In Section 5.1, we study the problem of competing with a finite baseline policy while allowing arbitrary opponent polices. In this subsection, we turn to a complementary setting where the baseline policy class consists of *all* the general polices while the opponent policy class $\Psi^\star$ is finite but *unknown*.

**Algorithm.** Based on OP-EXP3, we propose Adaptive OP-EXP3 (Algorithm 2), which represents **adaptive** **O**ptimistic **P**olicy **EXP3**. Compared to its prototype, adaptive OP-EXP3 incorporates the following two key modifications

- **Lazy model update** (Line 9-10). Adaptive OP-EXP3 maintains two empirical model estimates: the latest version and a lazy version that are computed by using counter $N$ and $N^{\text{lazy}}$ respectively. Counter $N$ is promptly updated in each episode as in OP-EXP3, while counter $N^{\text{lazy}}$ copies the values in $N$ each time a state-action counter in $N$ is doubled or a new opponent policy is observed. Importantly, adaptive OP-EXP3 always uses the lazy model estimate for optimistic policy evaluation (Line 6).
- **Adaptive player policy class** (Line 9-12). Each time the opponent reveals a new policy (i.e., a policy not in the historical opponent policy set $\Psi^k$) or the lazy model is updated, the learner recomputes its policy class $\Phi$ to include the optimistic best responses to all the possible mixtures of historical opponent policies. After that, EXP3 is restarted from the uniform distribution over $\Phi$.

We formally describe how to recompute the player policy class in Subroutine 2 where we in fact only consider an $\epsilon$-cover of all the possible mixtures of historical opponent policies. And for each such mixture, we compute an optimistic best response, by invoking the optimistic policy evaluation subroutine with the lazy model estimate.

[3]Competing against all Markov policies is equivalent to competing with all deterministic Markov policies because for any general policy there always exists a Markov best-response that is also deterministic.

Intuitively, the reason for only including the best responses to policy mixtures in the player policy class is that the best general policy in hindsight is always a best response to a mixture of the historical opponent policies. Moreover, by doing so, we effectively shrink the log-cardinality of the baseline policy class to $\tilde{\mathcal{O}}(|\Psi^\star|)$ that is the size of the opponent policy class, while still remaining competitive with any general policy.

**Algorithm 2** ***A**daptive **O**ptimistic **P**olicy **EXP3***

1: **input**: bonus function $\beta : \mathbb{N} \to \mathbb{R}$, learning rate $(\eta_k)_{k=1}^K$, grid resolution $\epsilon$.
2: **initialize**: baseline policy class $\Phi$ and distribution $\mathbf{p}^1 \in \mathbb{R}^{|\Phi|}$ arbitrarily, visitation counters $N_h(s,\mathbf{a}) = N_h(s,\mathbf{a},s') = N_h^{\text{lazy}}(s,\mathbf{a}) = N_h^{\text{lazy}}(s,\mathbf{a},s') = 0$ for all $(s,\mathbf{a},\mathbf{a}',h)$, $\Psi^1 = \emptyset$, $m^1 = 0$
3: **for** $k = 1,\dots,K$ **do**
4: the learner samples $\mu^k \sim \mathbf{p}^k$ and the adversary chooses $\nu^k$ *simultaneously*
5: follow $\pi^k = \mu^k \times \nu^k$ to sample $\{s_h^k, \mathbf{a}_h^k, r_h^k\}_{h=1}^H$
*# optimistic EXP3*
6: observe $\nu^k$, and for all $\mu \in \Phi$ compute $\overline{V}_1^{\mu\times\nu^k}(s_1) = \text{OPE}(N^{\text{lazy}}, \beta, \mu\times\nu^k)$
7: then update $\mathbf{p}^{k+1}(\mu) \propto \exp(\eta_k \cdot \sum_{t=m^k+1}^k \cdot \overline{V}_1^{\mu\times\nu^t}(s_1))$
*# update the counters*
8: for all $h \in [H]$: $N_h(s_h^k,\mathbf{a}_h^k) \leftarrow N_h(s_h^k,\mathbf{a}_h^k)+1$ and $N_h(s_h^k,\mathbf{a}_h^k,s_{h+1}^k) \leftarrow N_h(s_h^k,\mathbf{a}_h^k,s_{h+1}^k)+1$
*# update the lazy model and policy class*
9: **if** $\nu^k \notin \Psi^k$ **or** $\exists h$ s.t. $N_h(s_h^k,\mathbf{a}_h^k) \geq 2N_h^{\text{lazy}}(s_h^k,\mathbf{a}_h^k)$ **then**
10: $N^{\text{lazy}} \leftarrow N$, $\Psi^{k+1} \leftarrow \Psi^k \cup \{\nu^k\}$, $m^{k+1} \leftarrow k$
11: $\Phi \leftarrow \text{OBR}(N^{\text{lazy}}, \beta, \Psi^{k+1}, \epsilon)$
12: reset $\mathbf{p}^{k+1}$ to be uniform over $\Phi$
13: **else**
14: $\Psi^{k+1} \leftarrow \Psi^k$ and $m^{k+1} \leftarrow m^k$
15: **end if**
16: **end for**

**Theoretical guarantee.** Now we present the theoretical guarantee for adaptive OP-EXP3, under the following adaptive learning rate schedule

$$\eta_k = \sqrt{\frac{|\Psi^k|\log(K)}{(k-m^k)H^2}}, \tag{4}$$

where $\Psi^k$ contains all the different policies the opponent has played before the $k^{\text{th}}$ episode, and $m^k$ denotes the index of the most recent episode when EXP3 is restarted (Line 11) before the $k^{\text{th}}$ episode.

**Theorem 5.2.** *Let $c$ be a large absolute constant. In Algorithm 2, choose the learning rate adaptively by (4), $\epsilon = 1/K$ and $\beta(n) = \sqrt{H^2S\iota/\max\{n,1\}}$ where $\iota =$*

**Subroutine 2** ***O**ptimistic **B**est **R**esponse* $(N^{\text{lazy}}, \beta, \Psi, \epsilon)$

**initialize:** $\text{BR} = \{\}$
denote the polices in $\Psi$ by $\nu^{(1)}, \dots, \nu^{(|\Psi|)}$
denote by $\Delta^{\epsilon}_{|\Psi|}$ an $\epsilon$-cover of $\Delta_{|\Psi|}$ w.r.t. $\ell_1$-norm
**for** $w \in \Delta^{\epsilon}_{|\Psi|}$ **do**
  Select an arbitrary
    $\mu \in \text{argmax}_{\hat{\mu}} \sum_{i=1}^{|\Psi|} w_i \times \text{OPE}(N^{\text{lazy}}, \beta, \hat{\mu} \times \nu^{(i)})$
  $\text{BR} \leftarrow \text{BR} \cup \{\mu\}$
**end for**
return BR

$c \log(SAHK/\delta)$. *Then with probability at least* $1-\delta$*, for all* $k \in [K]$

$$\text{Regret}(k) \le \mathcal{O}\left(\sqrt{k\left(S^2AH^2 + |\Psi^k|SAH + |\Psi^k|^2\right)H^2\iota^2}\right)$$

Given that the opponent only plays policies from a finite class $\Psi^\star$, Theorem 5.2 guarantees that adaptive OP-EXP3 suffers regret at most

$$\mathcal{O}\left(\sqrt{k\left(S^2AH^2 + |\Psi^\star|SAH + |\Psi^\star|^2\right)H^2\iota^2}\right)$$

in competing with the best *general* policy in hindsight. Moreover, note that the bound in Theorem 5.2 depends linearly on the number of different historical opponent policies. As a result, the regret of adaptive OP-EXP3 is still sublinear even if the opponent policy class keeps expanding as $k$ increases, as long as its cardinality is order $o(\sqrt{k})$. The proof of Theorem 5.2 can be found in Appendix B.2.

### 5.3. Statistical hardness with large $\Phi^\star$ and $\Psi^*$

Theorem 5.1 and 5.2 show that when either $\log|\Phi^\star|$ (the log-cardinality of the baseline policy class) or $|\Psi^*|$ (the cardinality of the opponent's policy class) is ploynomial, a sublinear regret bound is obtainable. We now complement these two results with a lower bound when both conditions are violated, i.e., when the size of $\Phi^\star$ is doubly exponential and the size of $\Psi^*$ is exponential.

**Theorem 5.3.** *There exists a Markov game with* $S = 1$, $|\mathcal{A}_{\max}| = |\mathcal{A}_{\min}| = 2$, $|\Psi^*| = 2^H$ *such that the regret for competing with the best general policy in hindsight is* $\Omega(\min\{K, 2^H\})$*, even if the adversary reveals her policy after each episode.*

The construction for this lower bound is quite simple. Consider a Markov game with horizon $H$ and only 1 state. The agent only receives non-zero reward if at the final time step, it plays the same action as the opponent, i.e., $r_H(s,(a,b)) = \mathbf{1}[a = b]$. Now, suppose that in each episode, the opponent samples randomly from the set of all deterministic Markov policies; any algorithm would have an expected value of $1/2$, as $b_H \sim \text{Ber}(1/2)$. However, the best history-dependent policy in hindsight would be able to predict $b_H$ by memorizing $b_1, \cdots, b_{H-1}$ when the number of episodes is not exponentially large. This gives the claimed $\Omega(\min\{K, 2^H\})$ lower bound. A formal proof can be found in Appendix B.3.

## 6. Computational Hardness

Finally, we provide a computational lower bound for this problem. We remark that this lower bound holds even if (a) the transitions of the Markov game are known, (b) the opponent reveals the policy she just played at the end of each episode, and (c) the opponent can only choose from a small known set of Markov policies ($|\Psi^*| = \mathcal{O}(H)$). Therefore, the lower bound applies to all the settings considered in this paper.

**Theorem 6.1.** *If an algorithm achieves* $\text{poly}(S, A, H) \cdot K^{1-c}$ *expected regret with a constant* $c > 0$ *in the setting that satisfies the above condition* $(a), (b), (c)$*, then its computational complexity cannot be* $\text{poly}(S, A, H, K)$ *unless* $\text{NP} \subseteq \text{BPP}$.[4]

This computational lower bound suggests that the best we can hope for is a statistically efficient but computationally intensive algorithm. It also renders statistically efficient value-iteration or Q-learning style algorithms for this problem unlikely, unless they employ NP-hard subroutines.

The proof of Theorem 6.1 depends on the construction in Proposition 6 of Steimle et al. (2021), which reduces solving 3-SAT to finding the best Markov policy in a latent MDP. We provide a full proof in Appendix C.1.

## 7. Conclusion

This paper studies no-regret learning of Markov games with adversarial opponents. We provide a complete set of positive and negative results for competing with the best fixed policy in hindsight. In the standard setting where only the actions of opponents are revealed, we prove it is statistically intractable to compete with the best fixed Markov policy in hindsight, even if the opponent only chooses from a limited number of Markov policies. In the revealed-policy setting, we propose new algorithms with $\sqrt{K}$-regret bound when either the log-cardinality of the baseline policy class or the cardinality of the opponent's policy class is small. Additionally, an exponential lower bound is derived when both quantities are large. Finally, we turn to the computational efficiency and prove achieving sublinear regret is in general computationally hard even in the very benign scenario.

[4] BPP is the probabilistic version of P, and $\text{NP} \subseteq \text{BPP}$ is believed to be highly unlikely in computational complexity literature.

## Acknowledgements

We thank Zhuoran Yang for valuable discussions. This work was partially supported by Office of Naval Research Grant N00014-22-1-2253.

## References

Abbasi Yadkori, Y., Bartlett, P. L., Kanade, V., Seldin, Y., and Szepesvári, C. Online learning in markov decision processes with adversarially chosen transition probability distributions. *Advances in neural information processing systems*, 26, 2013.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pp. 193–256. PMLR, 2016.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. *International Conference on Machine Learning*, 2020.

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

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxpxJBKwS.

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.

Brambilla, M., Ferrante, E., Birattari, M., and Dorigo, M. Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence*, 7(1):1–41, 2013.

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.

Farina, G. and Sandholm, T. Model-free online learning in unknown sequential decision making problems and games. *arXiv preprint arXiv:2103.04539*, 2021.

Farina, G., Kroer, C., and Sandholm, T. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. *arXiv preprint arXiv:2007.14358*, 2020.

Filar, J. and Vrieze, K. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Gordon, G. J. No-regret algorithms for online convex programs. In *Advances in Neural Information Processing Systems*, pp. 489–496. Citeseer, 2007.

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.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. *International Conference on Machine Learning*, 2019.

Jin, C., Kakade, S. M., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 2020.

Jin, C., Liu, Q., and Yu, T. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021.

Kozuno, T., Ménard, P., Munos, R., and Valko, M. Model-free learning for two-player zero-sum partially observable markov games with perfect recall. *arXiv preprint arXiv:2106.06279*, 2021.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 2021.

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

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. Linear last-iterate convergence for matrix games and stochastic games. *arXiv preprint arXiv:2006.09517*, 2020.

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

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

OpenAI. Openai five. https://blog.openai.com/openai-five/, 2018.

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.

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

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

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, 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.

Smallwood, R. D. and Sondik, E. J. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Steimle, L. N., Kaufman, D. L., and Denton, B. T. Multi-model markov decision processes. *IISE Transactions*, pp. 1–16, 2021.

Tian, Y., Wang, Y., Yu, T., and Sra, S. Online learning in unknown markov games. In *International conference on machine learning*, pp. 10279–10288. PMLR, 2021.

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.

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.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20:1729–1736, 2007.

# A. Proofs for Section 4

## A.1. Proof of Theorem 4.1

Because we can simulate any POMDP with a MG by using Proposition 4.2, it suffices to show there exists a hard POMDP instance with $\mathcal{O}(1)$ number of states, actions and observations so that any algorithm will suffer $\Omega(\min\{4^H, K\})$ regret when competing with the optimal Markov policy of this POMDP.

We use the hard instance constructed in Jin et al. (2020). There are two states $s_g, s_b$ and four actions. There is special action sequence $a_1^\star, \ldots, a_{H-1}^\star$ sampled independently and uniformly at random from the action set, which is unknown to the learner. The transition dynamics are constructed so that (a) the agent always starts in $s_g$ at step 1, (b) at each step $h$ the agent will transition to $s_g$ if and only if she is currently in $s_g$ and plays the special action $a_h^\star$, and otherwise will go to $s_b$. At the first $H-1$ steps, the two states emit the *same* observation that contains reward 0. At step $H$, $s_g$ emits reward 1 while $s_b$ still emits a zero-reward observation. It is straightforward to see the optimal policy is to play the special action sequence, which is *Markov*. However, because $s_g$ and $s_b$ are totally indistinguishable from observations at the first $H-1$ steps, finding this action sequence will cost at least $\Omega(4^H)$ episodes in general, which implies a $\Omega(\min\{4^H, K\})$ regret lower bound for competing with the optimal Markov policy.

## A.2. Proof of Proposition 4.2

We describe how to simulate a POMDP with a Markov game and an opponent playing a fixed general policy.

Each step in the POMDP is simulated by two consecutive steps in the Markov game, and the transition dynamics of the Markov game have the following special structures:

- At an even step, the transition only depends on the action of the opponent. Moreover, the next state is always equal to the opponent's action regardless of the current state.
- At an odd step, the transition only depends on the action of the learner, and the next state is simply an augmentation of the current state and the learner's action.

Specifically, suppose in the POMDP, at step $h$, the learner starts with history $o_1, a_1, \ldots, o_h$ and plays action $a_h$, then observes $o_{h+1}$ sampled from $\mathbb{P}(o_{h+1} = \cdot \mid o_1, a_1, \ldots, o_h, a_h)$. In this case, the corresponding two steps in the POMG will be: at step $2h-1$, the learner starts at state $o_h$ and takes action $a_h$, then the environment transitions to state $(o_h, a_h)$; at step $2h$, the opponent starts at state $(o_h, a_h)$ and takes action $o_{h+1}$ sampled from $\mathbb{P}(o_{h+1} = \cdot \mid o_1, a_1, \ldots, o_h, a_h)$, then the environment transitions to $o_{h+1}$ that is exactly equal to the action of the opponent. Note that here the opponent is playing a history-dependent policy.

It is direct to see there are $OA + O$ distinct states, $A$ actions for the learner and $O$ actions for the opponent in this Markov game. Besides, the episode length is $2H$.

## A.3. Proof of Theorem 4.3

By Proposition 4.4, we can simulate any latent MDP with a MG. As a result, it suffices to show there exists a hard latent MDP with $\mathcal{O}(1)$ states, $\mathcal{O}(H)$ actions and $H$ latent variables so that any algorithm will suffer $\Omega(\min\{4^H, K\}/H)$ regret when competing with the optimal Markov policy of this latent MDP.

We utilize the hard latent MDP instance constructed in Theorem 3.1 (Kwon et al., 2021).[5] In the latent MDP instance, there is a collection of $H$ unknown MDPs, each of which has $\mathcal{O}(1)$ states, $\mathcal{O}(H)$ actions and binary rewards. At the beginning of each episode the environment *secretly* draws an MDP uniformly at random from these $H$ MDPs, and then the algorithm interacts with this MDP without knowing which one it is. Kwon et al. (2021) prove that it takes $\Omega(4^H)$ episodes to learn a policy that is $\mathcal{O}(1/H)$-optimal compared to the best Markov policy, where the optimality is defined using the average value over the $H$ MDPs. By the standard online-to-batch conversion (e.g., Lattimore & Szepesvári, 2020), it immediately implies a $\Omega(\min\{4^H, K\}/H)$ regret lower bound for competing with the optimal Markov policy.

[5] Despite Kwon et al. (2021) study the stationary setting, their constructions can be trivially adapted to handle the nonstationary setting and gives a stronger lower bound which is the one we state here.

### A.4. Proof of Proposition 4.4

To begin with, we recall the definition of latent MDPs (Kwon et al., 2021). At the beginning of each episode the environment *secretly* draws an MDP uniformly at random from $L$ unknown MDPs, then the algorithm interacts with this MDP without knowing which one it is.

Denote by $q \in \Delta_L$ the latent distribution over these $L$ MDPs and $\mathbb{P}_h^i(s'|s,a)$ ($r_h^i(s,a)$) the transition (reward) function of the $i^{\text{th}}$ MDP. In each episode of the Markov game

- The opponent secretly samples $t \sim q$ before step 1, and keeps it hidden from the learner throughout.
- At step $2h-1$, the transitions are deterministic, and only depend on the current state and the learner's action. Specifically, the environment will transition to an augmenting state $(s,a)$ if the learner takes action $a$ at state $s$ regardless of what action the opponent picks. There is no reward at this step.
- At step $2h$, the transitions and rewards are still deterministic, but only depend on the opponent's action. Formally, the environment will transition to state $s'$ from an augmenting state $(s,a)$ and the learner will receive reward $r' \in \{0,1\}$, if the opponent takes action $(s',r')$, the probability of which is given by
$$\nu_{2h}\left((s',r')|(s,a),t\right) = \mathbb{P}_h^t(s'|s,a) \times \mathbf{1}\left(r_h^t(s,a) = r'\right).$$

It is direct to see interacting with this MG is exactly equivalent to interacting with the original latent MDP. In particular, there is *no additional information* revealed in the MG because the opponent's action is always equal to the next state and the reward.

## B. Proofs for Section 5

### B.1. Proof of Theorem 5.1

We first introduce several notations that will be frequently used in our proof. Let $\tau_h = [s_1, \mathbf{a}_1, \ldots, s_{h-1}, \mathbf{a}_{h-1}, s_h]$. Denote by $N^k$ the collection of counters at the *beginning* of episode $k$. Denote by $\hat{\mathbb{P}}^k$ the empirical transition computed by using $N^k$, i.e., for any $(s,\mathbf{a},h,s') \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$

$$\widehat{\mathbb{P}}_h^k(s' \mid s,\mathbf{a}) = \begin{cases} \frac{N_h^k(s,\mathbf{a},s')}{N_h^k(s,\mathbf{a})} & \text{if } N_h(s,\mathbf{a}) \neq 0, \\ 1/S & \text{otherwise.} \end{cases}$$

Given an arbitrary policy $\pi$, we define $\overline{V}^{\pi,k}$ ($\overline{Q}^{\pi,k}$) that is the optimistic estimate of $V^\pi$ ($Q^\pi$) as following: for any $h \in [H]$,

$$\begin{cases} \overline{V}_h^{k,\pi}(\tau_h) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\tau_h)}\left[\overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a})\right], \\ \overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) = \min\left\{\mathbb{E}_{s'\sim\hat{\mathbb{P}}_h^k(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] + r_h(s_h,\mathbf{a}) + \beta(N_h^k(s_h,\mathbf{a})), H-h+1\right\}, \end{cases} \tag{5}$$

and we define $\overline{V}_{H+1}^{k,\pi} \equiv 0$. We comment that by definition $\overline{V}_1^{k,\pi}(s_1) = \text{UCB-VI}(N^k,\beta,\pi)$ for all $k,\pi$.

For the purpose of proof, we further introduce the following auxiliary function for controlling the optimism of $\overline{V}^{\pi,k}$ ($\overline{Q}^{\pi,k}$) against the true value function $V^\pi$ ($Q^\pi$): for any $h \in [H]$

$$\begin{cases} \tilde{V}_h^{k,\pi}(\tau_h) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\tau_h)}\left[\tilde{Q}_h^{k,\pi}(\tau_h,\mathbf{a})\right], \\ \tilde{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) = \min\left\{\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\tilde{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] + 2\beta(N_h^k(s_h,\mathbf{a})), H-h+1\right\}, \end{cases} \tag{6}$$

and we define $\tilde{V}_{H+1}^{k,\pi} \equiv 0$. Compared to $\overline{V}^{k,\pi}$, $\tilde{V}_h^{k,\pi}$ is defined using the groundtruth transition function $\mathbb{P}$, it does not contain the reward function and the bonus function is doubled.

Finally, recall we choose the bonus function to be

$$\beta(t) = H\sqrt{\frac{S\iota}{\max\{t,1\}}},$$

where $\iota = \log(KHSA/\delta)$ with $c$ being some large absolute constant.

**Lemma B.1** (Optimism). *With probability at least $1-\delta$, for all $(k,h)\in[K]\times[H+1]$ and all general policy $\pi$,*

$$0 \le \overline{V}_h^{k,\pi}(\tau_h) - V_h^{\pi}(\tau_h) \le \tilde{V}_h^{k,\pi}(\tau_h) \quad \textit{for all } \tau_h.$$

*Proof of Lemma B.1.* To begin with, by the Azuma-Hoeffding inequality and standard union bound argument, we have that with probability at least $1-\delta$:

$$\|\hat{\mathbb{P}}_h^k(\cdot \mid s,\mathbf{a}) - \mathbb{P}_h(\cdot \mid s,\mathbf{a})\|_1 \le \frac{1}{H}\beta(N_h^k(s,\mathbf{a})) \quad \text{for all } (s,\mathbf{a},h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K].$$

Below, we prove the lemma conditioning on the event above being true. We prove the lemma by induction and start with the upper bound. The inequality holds for step $H+1$ trivially because $V_{H+1}^{k,\pi} = V_{H+1}^{\pi} = \tilde{V}_{H+1}^{k,\pi} = 0$. Assume the inequality holds for step $h+1$. At step $h$, notice that

$$\begin{aligned}
\tilde{V}_h^{k,\pi}(\tau_h) =& \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\tau_h)}\left[\tilde{Q}_h^{k,\pi}(\tau_h,\mathbf{a})\right],\\
\overline{V}_h^{k,\pi}(\tau_h) - V_h^{\pi}(\tau_h) =& \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\tau_h)}\left[\overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) - Q_h^{\pi}(\tau_h,\mathbf{a})\right].
\end{aligned}$$

Therefore, it suffices to prove

$$\overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) - Q_h^{\pi}(\tau_h,\mathbf{a}) \le \tilde{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) \quad \text{for all } \tau_h,\mathbf{a},$$

which follows from

$$\begin{aligned}
&\overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) - Q_h^{\pi}(\tau_h,\mathbf{a})\\
\le& \min\left\{\mathbb{E}_{s'\sim\hat{\mathbb{P}}_h^k(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[V_{h+1}^{\pi}([\tau_h,\mathbf{a},s'])\right] + \beta(N_h^k(s_h,\mathbf{a})), H-h+1\right\}\\
=& \min\Bigg\{\mathbb{E}_{s'\sim\hat{\mathbb{P}}_h^k(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right]\\
&\quad + \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s']) - V_{h+1}^{\pi}([\tau_h,\mathbf{a},s'])\right] + \beta(N_h^k(s_h,\mathbf{a})), H-h+1\Bigg\}\\
\le& \min\left\{2\beta(N_h^k(s_h,\mathbf{a})) + \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\tilde{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right], H-h+1\right\} = \tilde{Q}_h^{k,\pi}(\tau_h,\mathbf{a}),
\end{aligned}$$

where the last inequality follows from the induction hypothesis and $\|\hat{\mathbb{P}}_h^k(\cdot \mid s_h,\mathbf{a}) - \mathbb{P}_h(\cdot \mid s_h,\mathbf{a})\|_1 \le \beta(N_h^k(s_h,\mathbf{a}))/H$.

Similarly, for the lower bound, we only need to show

$$\overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) \ge Q_h^{\pi}(\tau_h,\mathbf{a}) \quad \text{for all } \tau_h,\mathbf{a},$$

which follows similarly from

$$\begin{aligned}
&\overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) - Q_h^{\pi}(\tau_h,\mathbf{a})\\
\ge& \min\left\{\mathbb{E}_{s'\sim\hat{\mathbb{P}}_h^k(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[V_{h+1}^{\pi}([\tau_h,\mathbf{a},s'])\right] + \beta(N_h^k(s_h,\mathbf{a})), 0\right\}\\
=& \min\Bigg\{\mathbb{E}_{s'\sim\hat{\mathbb{P}}_h^k(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right]\\
&\quad + \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s']) - V_{h+1}^{\pi}([\tau_h,\mathbf{a},s'])\right] + \beta(N_h^k(s_h,\mathbf{a})), 0\Bigg\}\\
\ge& \min\left\{-\beta(N_h^k(s_h,\mathbf{a})) + \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s']) - V_{h+1}^{\pi}([\tau_h,\mathbf{a},s'])\right] + \beta(N_h^k(s_h,\mathbf{a})), 0\right\}\\
\ge& 0,
\end{aligned}$$

where the last inequality follows from the induction hypothesis and the second last one uses $\|\hat{\mathbb{P}}_h^k(\cdot \mid s_h,\mathbf{a}) - \mathbb{P}_h(\cdot \mid s_h,\mathbf{a})\|_1 \le \beta(N_h^k(s_h,\mathbf{a}))/H$. □

*Proof of Theorem 5.1.* In the remainder of this section, we show how to control $\mathrm{Regret}(K)$. The upper bound for $\mathrm{Regret}(k)$ ($k \in [K]$) can be derived by repeating precisely the same arguments.

For simplicity of notations, denote $\pi^k = \mu^k \times \nu^k$. By the optimism of $\overline{V}$ (Lemma B.1), with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{\mu^\star} \sum_{k=1}^K V_1^{\mu^\star \times \nu^k}(s_1) - \sum_{k=1}^K V_1^{\pi^k}(s_1) \\
&\le \left( \max_{\mu^\star} \sum_{k=1}^K \overline{V}_1^{\mu^\star \times \nu^k, k}(s_1) - \sum_{k=1}^K \mathbb{E}_{\mu \sim \mathbf{p}^k}[\overline{V}_1^{\mu \times \nu^k, k}(s_1)] \right) + \left( \sum_{k=1}^K \mathbb{E}_{\mu \sim \mathbf{p}^k}[\overline{V}_1^{\mu \times \nu^k, k}(s_1)] - \sum_{k=1}^K V_1^{\pi^k}(s_1) \right).
\end{aligned}$$

The first term is upper bounded by the regret bound of anytime EXP3, which is of order $\mathcal{O}(H\sqrt{\log(|\Phi^\star|)K})$ (e.g., Lattimore & Szepesvári, 2020). Below, we focus on controlling the second term. Since $\mu^k \sim \mathbf{p}^k$, by the Azuma-Hoeffding inequality, with probability at least $1-\delta$,

$$\sum_{k=1}^K \mathbb{E}_{\mu \sim \mathbf{p}^k}[\overline{V}_1^{\mu \times \nu^k, k}(s_1)] - \sum_{k=1}^K V_1^{\pi^k}(s_1) \le \sum_{k=1}^K \overline{V}_1^{\mu^k \times \nu^k, k}(s_1) - \sum_{k=1}^K V_1^{\pi^k}(s_1) + \mathcal{O}(H\sqrt{K\log(1/\delta)}).$$

By Lemma B.1 and the Azuma-Hoeffding inequality, with probability at least $1-2\delta$,

$$\begin{aligned}
\sum_{k=1}^K \overline{V}_1^{\mu^k \times \nu^k, k}(s_1) - \sum_{k=1}^K V_1^{\pi^k}(s_1) \le& \sum_{k=1}^K \tilde{V}_1^{\mu^k \times \nu^k, k}(s_1) \\
\le& \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{(s_h,a_h)\sim\pi^t}\left[2\beta(N_h^k(s_h,a_h))\right] \\
\le& 2\sum_{h=1}^H \sum_{k=1}^K \beta(N_h^k(s_h^k,a_h^k)) + \mathcal{O}(H^2\sqrt{KS\iota^2}) \\
\le& \mathcal{O}\left(\sqrt{KS^2AH^4\iota^2}\right),
\end{aligned}$$

where the final inequality follows from the definition of $\beta$ and the standard pigeon-hole argument.

Combining all the relations above, taking a union bound and rescaling $\delta$ complete the proof. □

### B.2. Proof of Theorem 5.2

At the very beginning of the proof of Theorem 5.1, we define several useful quantities $\hat{\mathbb{P}}^k, \overline{V}^k, \tilde{V}^k$ using the regular counter $N^k$. In this section, with slight abuse of notations, we change their definitions by replacing $N^k$ with $N^{k,\mathrm{lazy}}$ that is the collection of the *lazy* counters at the *beginning* of episode $k$. Formally, denote by $\widehat{\mathbb{P}}^k$ the empirical transition computed by using $N^{k,\mathrm{lazy}}$, i.e., for any $(s,\mathbf{a},h,s') \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$

$$\widehat{\mathbb{P}}_h^k(s' \mid s,\mathbf{a}) = \begin{cases} \frac{N_h^{k,\mathrm{lazy}}(s,\mathbf{a},s')}{N_h^{k,\mathrm{lazy}}(s,\mathbf{a})} & \text{if } N_h(s,\mathbf{a}) \neq 0, \\ 1/S & \text{otherwise.} \end{cases}$$

Given an arbitrary policy $\pi$, we define $\overline{V}^{\pi,k}$ ($\overline{Q}^{\pi,k}$) that is the optimistic estimate of $V^\pi$ ($Q^\pi$) as following: for any $h \in [H]$,

$$\begin{cases} \overline{V}_h^{k,\pi}(\tau_h) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\tau_h)}\left[\overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a})\right], \\ \overline{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) = \min\left\{\mathbb{E}_{s'\sim\hat{\mathbb{P}}_h^k(\cdot|s_h,\mathbf{a})}\left[\overline{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] + r_h(s_h,\mathbf{a}) + \beta(N_h^k(s_h,\mathbf{a})), H-h+1\right\}, \end{cases} \tag{7}$$

and we define $\overline{V}_{H+1}^{k,\pi} \equiv 0$. We comment that by definition $\overline{V}_1^{k,\pi}(s_1) = \text{UCB-VI}(N^{k,\mathrm{lazy}},\beta,\pi)$ for all $k,\pi$.

For the purpose of proof, we further introduce the following auxiliary function for controlling the optimism of $\overline{V}^{\pi,k}$ ($\overline{Q}^{\pi,k}$) against the true value function $V^\pi$ ($Q^\pi$): for any $h \in [H]$

$$\begin{cases} \tilde{V}_h^{k,\pi}(\tau_h) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|\tau_h)}\left[\tilde{Q}_h^{k,\pi}(\tau_h,\mathbf{a})\right], \\ \tilde{Q}_h^{k,\pi}(\tau_h,\mathbf{a}) = \min\left\{\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s_h,\mathbf{a})}\left[\tilde{V}_{h+1}^{k,\pi}([\tau_h,\mathbf{a},s'])\right] + 2\beta(N_h^{k,\mathrm{lazy}}(s_h,\mathbf{a})), H-h+1\right\}, \end{cases} \tag{8}$$

and we define $\tilde{V}_{H+1}^{k,\pi} \equiv 0$. Compared to $\overline{V}^{k,\pi}$, $\tilde{V}_h^{k,\pi}$ is defined using the groundtruth transition function $\mathbb{P}$, it does not contain the reward function and the bonus function is doubled.

Finally, recall we choose

$$\beta(t) = H\sqrt{\frac{S\iota}{\max\{t,1\}}},$$

where $\iota = \log(KHSA/\delta)$ with $c$ being some large absolute constant.

**Lemma B.2** (Optimism). *With probability at least $1-\delta$, for all $(k,h) \in [K] \times [H+1]$ and all general policy $\pi$,*

$$0 \le \overline{V}_h^{k,\pi}(\tau_h) - V_h^{\pi}(\tau_h) \le \tilde{V}_h^{k,\pi}(\tau_h) \quad \textit{for all } \tau_h.$$

*Proof.* The proof of Lemma B.2 follows exactly the same as that of Lemma B.1 except that we replace $N^k$ with $N^{k,\mathrm{lazy}}$. □

*Proof of Theorem 5.2.* In the remainder of this section, we show how to control $\mathrm{Regret}(K)$. The upper bound for $\mathrm{Regret}(k)$ ($k \in [K]$) can be derived by repeating precisely the same arguments.

Denote by $\Phi^k$ ($\Psi^k$) the player (opponent) policy set at the *beginning* of episode $k$. Recall in Algorithm 2, each time we encounter a new opponent policy or one of the counters is doubled, we update the lazy counters to be the latest counters, recompute the player policy set, and restart EXP3 from the uniform distribution. We denote the indices of episodes where such restarting happens by $T_1, \ldots, T_L$. Observe that $L \le \mathcal{O}(SAH\log(K) + |\Psi^K|)$.

To begin with, we decompose the cumulative regret of $K$ episodes into the regret within $L+1$ segments divided by $T_1, \ldots, T_L$:

$$\begin{aligned}
&\max_{\mu} \sum_{k=1}^{K} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) \\
\le &\sum_{i=1}^{L-1} \max_{\mu} \sum_{k=T_i+1}^{T_{i+1}-1} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) + \max_{\mu} \sum_{k=T_L+1}^{K} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) \\
&\quad + \max_{\mu} \sum_{i=1}^{L} \left(V_1^{\mu\times\nu^{T_i}}(s_1) - V_1^{\mu^{T_i}\times\nu^{T_i}}(s_1)\right) \\
\le &\sum_{i=1}^{L-1} \max_{\mu} \sum_{k=T_i+1}^{T_{i+1}-1} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) + \max_{\mu} \sum_{k=T_L+1}^{K} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) + HL.
\end{aligned} \tag{9}$$

Below we show how to control $\sum_{k=T_i+1}^{T_{i+1}-1} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right)$ for any $i \in [L-1]$. The second term can be bounded in the same way. By Lemma B.2, with probability at least $1-2\delta$,

$$\begin{aligned}
&\max_{\mu} \sum_{k=T_i+1}^{T_{i+1}-1} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) \\
\le &\max_{\mu} \sum_{k=T_i+1}^{T_{i+1}-1} \left(\overline{V}_1^{k,\mu\times\nu^k}(s_1) - \overline{V}_1^{k,\mu^k\times\nu^k}(s_1)\right) + \sum_{k=T_i+1}^{T_{i+1}-1} \left(\overline{V}_1^{k,\mu^k\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) \\
\le &\max_{\mu} \sum_{k=T_i+1}^{T_{i+1}-1} \left(\overline{V}_1^{k,\mu\times\nu^k}(s_1) - \mathbb{E}_{\mu'\sim\mathbf{p}^k}\left[\overline{V}_1^{k,\mu'\times\nu^k}(s_1)\right]\right) + \mathcal{O}\left(H\sqrt{(T_{i+1}-T_i-1)\iota}\right) \\
&\qquad + \sum_{k=T_i+1}^{T_{i+1}-1} \left(\overline{V}_1^{k,\mu^k\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right),
\end{aligned} \tag{10}$$

where in the second inequality we use the Azuma-Hoeffding inequality and take a union bound for all the possible values of $T_i$ and $T_{i+1}$. Specifically, we use the fact that with probability at least $1-\delta$, for all $p, q \in [K]$,

$$\sum_{k=p+1}^{q-1} \mathbb{E}_{\mu' \sim \mathbf{p}^k}\left[\overline{V}_1^{k,\mu' \times \nu^k}(s_1)\right] - \overline{V}_1^{k,\mu^k \times \nu^k}(s_1) \le \mathcal{O}\left(H\sqrt{(q-p-1)\iota}\right).$$

The key to controlling the RHS of equation (10) is to show that the first term is approximately upper bounded by the regret of EXP3. Recall that for $k$ lying between $T_i$ and $T_{i+1}$, the opponent does not play any new policy and the lazy counter is never updated. As a result, for all $k$ satisfying $T_i < k < T_{i+1}$, we have

- $\overline{V}^{k,\pi} = \overline{V}^{T_i+1,\pi}$ for all $\pi$.
- $\Phi^k = \Phi^{T_i+1}$, $\Psi^k = \Psi^{T_i+1}$, and $\nu^k \in \Psi^{T_i+1}$.

Moreover, by the definition of the UCB-BestResponse subroutine, we have that for any policy $\tilde{\nu}$ that is a mixture of the policies in $\Psi^{T_i+1}$, there exists $\tilde{\mu} \in \Phi^{T_i+1}$ so that

$$\max_{\mu} \overline{V}_1^{T_i+1,\mu\times\tilde{\nu}}(s_1) - \overline{V}_1^{T_i+1,\tilde{\mu}\times\tilde{\nu}}(s_1) \le \epsilon H.$$

By utilizing the three relations above, we have

$$\begin{aligned}
&\max_{\mu} \sum_{k=T_i+1}^{T_{i+1}-1} \left(\overline{V}_1^{k,\mu\times\nu^k}(s_1) - \mathbb{E}_{\mu'\sim\mathbf{p}^k}\left[\overline{V}_1^{k,\mu'\times\nu^k}(s_1)\right]\right) \\
=&\max_{\mu} \sum_{k=T_i+1}^{T_{i+1}-1} \left(\overline{V}_1^{T_i+1,\mu\times\nu^k}(s_1) - \mathbb{E}_{\mu'\sim\mathbf{p}^k}\left[\overline{V}_1^{T_i+1,\mu'\times\nu^k}(s_1)\right]\right) \\
\le&\max_{\tilde{\mu}\in\Phi^{T_i+1}} \sum_{k=T_i+1}^{T_{i+1}-1} \left(\overline{V}_1^{T_i+1,\tilde{\mu}\times\nu^k}(s_1) - \mathbb{E}_{\mu'\sim\mathbf{p}^k}\left[\overline{V}_1^{T_i+1,\mu'\times\nu^k}(s_1)\right]\right) + (T_{i+1}-T_i-1)\,\epsilon H \\
\le&\mathcal{O}\left(\sqrt{\log|\Phi^{T_i+1}|\,(T_{i+1}-T_i-1)\,H^2} + (T_{i+1}-T_i-1)\,\epsilon H\right),
\end{aligned} \tag{11}$$

where the first inequality follows from $\frac{1}{T_{i+1}-T_i-1}\sum_{k=T_i+1}^{T_{i+1}-1}\nu^k$ being a mixture of policies in $\Psi^{T_i+1}$, and the second inequality follows from Algorithm 2 running anytime EXP on $\Phi^{T_i+1}$ and using $-\overline{V}^{T_i+1}$ as gradients for iterations between $T_i$ and $T_{i+1}$.

Finally, combining equations (9), (10), and (11) together, we obtain

$$\begin{aligned}
&\max_{\mu} \sum_{k=1}^{K} \left(V_1^{\mu\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) \\
\le&\min\left\{\mathcal{O}\left(HL + H\sqrt{KL\iota} + KH\epsilon + \sqrt{KLH^2\log|\Phi^K|}\right), KH\right\} \\
&\qquad + \sum_{k=1}^{k} \left(\overline{V}_1^{k,\mu^k\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right).
\end{aligned}$$

For the second term, note that $N_h^{k,\mathrm{lazy}}(s,\mathbf{a}) = \Theta(N_h^k(s,\mathbf{a}))$ for all $(s,\mathbf{a},h,k)$, so following the identical arguments in the proof of Theorem 5.1 gives: with probability at least $1-\delta$

$$\sum_{k=1}^{K} \left(\overline{V}_1^{k,\mu^k\times\nu^k}(s_1) - V_1^{\mu^k\times\nu^k}(s_1)\right) \le \mathcal{O}\left(\sqrt{KS^2AH^4\iota^2}\right).$$

For the first term, plug in $\epsilon = 1/K$ and notice that $\min\{HL, HK\} \le H\sqrt{KL\iota}$ as well as $\log\left|\Phi^K\right| \le \mathcal{O}(|\Psi^K|\log(K))$,

$$\begin{aligned}
&\min\left\{\mathcal{O}\left(HL + H\sqrt{KL\iota} + KH\epsilon + \sqrt{KLH^2\log|\Phi^K|}\right), KH\right\} \\
\le&\mathcal{O}\left(H\sqrt{KL\iota} + \sqrt{KL|\Psi^K|H^2\iota}\right) \\
\le&\mathcal{O}\left(H\sqrt{K|\Psi^K|SAH\iota^2} + H\sqrt{K|\Psi^K|^2\iota}\right),
\end{aligned}$$

where the second inequality uses $L \le \mathcal{O}(SAH\log(K) + |\Psi^K|)$. Combining all relations, we conlude the final upper bound is

$$\mathcal{O}\left(\sqrt{KS^2AH^4\iota^2} + \sqrt{K|\Psi^K|SAH^3\iota^2} + \sqrt{K|\Psi^K|^2H^2\iota}\right).$$

□

### B.3. Proof of Theorem 5.3

*Proof.* Consider the following Markov game with one state $s$ and horizon $H$. The action set for both the max-player and the min-player (adversary) is $\{0,1\}$. For $h = 1, \cdots, H-1$, $r_h(s,\cdot) = 0$. $r_H(s,(a,b)) = I[a=b]$.

Suppose that at episode $t$, the adversary chooses a policy $\nu_t$ which plays $b_h^t$ at time step $h$, where each $b_h^t$ is sampled independently from $\mathrm{Unif}(\{0,1\})$.[6] It can be easily seen that for each episode, the expected value for the max-player is always $\frac{1}{2}$. However, the best general policy in hindsight will be able to predict $b_H$ from $b_{1:h-1}$ to a large extent. Specifically, for each possible value of $b_{1:h-1}$, define $N(b_{1:h-1}) := \sum_{t=1}^T I[b_{1:h-1}^t = b_{1:h-1}]$. If $T < 2^{H-2}$,

$$\Pr\left[N(b_{1:h-1}) > 1\right] \le \frac{\mathbb{E}[N(b_{1:h-1})]}{2} \le \frac{1}{4}.$$

If $N(b_{1:h-1}) = 1$, denote the only episode in which it appears by $t$. The best general policy in hindsight could set $\mu(s, b_{1:h-1}) = b_H^t$ and achieve value 1 on episode $t$. In other words, there exists general policy $\mu$ such that

$$\mathbb{E}\left[\sum_{t=1}^T V_1^{\mu,\nu_t}(s)\right] \ge \mathbb{E}\left[\sum_{t=1}^T I[N(b_{1:h-1}^k) = 1]\right] \ge \frac{3}{4}T.$$

Therefore regret is at least $\frac{1}{4}T$, unless $T \ge 2^{H-2}$. □

## C. Proofs for Section 6

### C.1. Proof of Theorem 6.1

The proof of this theorem is essentially a reduction to Proposition 6 of Steimle et al. (2021). We present a full proof here for the sake of clarity and completeness.

We would prove the theorem via reduction from 3-SAT. Consider a 3-SAT instance with $m$ clauses and $n$ variables: $\wedge_{i=1}^m (y_{i1} \vee y_{i2} \vee y_{i3})$, where $y_{ij} \in \{x_1, \cdots, x_n, \bar{x}_1, \cdots, \bar{x}_n\}$. We would then construct a Markov game with $H = n$, $|\mathcal{A}_{\max}| = 2$ and $|\mathcal{A}_{\min}| = m$ as follows. The set of states are $\{s_1, \cdots, s_n, \mathrm{T}, \mathrm{F}\}$. The action set is $\{0,1\}$ for the max-player and $[m]$ for the min-player. The transitions are deterministic, independent of $h$, and are specified as follows:

$$\begin{aligned}
\mathbb{P}(\mathrm{T}|s_i,(a,j)) &= \begin{cases} 1 & (\text{setting } x_i = a \text{ sets clause } j \text{ to True}) \\ 0 & (\text{otherwise}) \end{cases} \\
\mathbb{P}(s_{i+1}|s_i,(a,j)) &= 1 - \mathbb{P}(\mathrm{T}|s_i,(a,j)), \\
\mathbb{P}(\mathrm{F}|s_n,(a,j)) &= 1 - \mathbb{P}(\mathrm{T}|s_n,(a,j)).
\end{aligned} \tag{i¡n}$$

[6] The policy itself is deterministic.

For $h < n$, $r_h(\cdot, \cdot) = 0$; for $h = n$, $r_n(\mathrm{T}, \cdot) = 1$, $r_n(\mathrm{F}, \cdot) = 0$. Every Markov policy $\mu$ of the max-player induces an assignment of the variables, *i.e.* $x_i = \mu(s_i)$. Moreover, denote the min-player's policy of playing action $j$ at all states by $\nu_j$. Notice that

$$V_1^{\mu,\nu_j}(s_1) = \begin{cases} 1 & (\text{assignment induced by } \mu \text{ satisfies clause } j) \\ 0 & (\text{assignment induced by } \mu \text{ violates clause } j) \end{cases}.$$

If an algorithm achieves $\mathbb{E}[\mathrm{Regret}(T)] = \mathrm{poly}(S, A, H) \cdot T^{1-c}$ regret, then there exists $T = \mathrm{poly}(n, m)$ such that $T \geq \max\left\{4m \cdot \mathbb{E}[\mathrm{Regret}(T)], 20m\sqrt{T}\right\}$. Now consider the following algorithm for 3-SAT:

1. Construct the aforementioned Markov game
2. Run algorithm $\mathcal{A}$, with the opponent playing $\nu_j$ with $j$ sampled from $\mathrm{Unif}([m])$ at the start of each episode independently
3. Calculate $R$, the total reward accumulated by the algorithm. Decide True (satisfiable) if $R > (1 - 1/2m)T$, and False otherwise.

We now claim that if the input instance is satisfiable, this algorithm returns True with probability at least $0.99$. This is because satisfiability implies $\exists \mu^* :, \sum_{t=1}^T V^{\mu^*,\nu_t}(s_1) = T$. By the definition of regret,

$$\mathbb{E}\left[\sum_{t=1}^T \left(V^{\mu^*,\nu_t} - V^{\mu_t,\nu_t}\right)(s_1)\right] \leq \frac{T}{4m}.$$

By Hoeffding's inequality, with probability at least $0.99$,

$$R \geq \mathbb{E}\left[\sum_{t=1}^T V^{\mu_t,\nu_t}\right] - 5\sqrt{T} > T - \frac{T}{4m} - \frac{T}{4m} = \left(1 - \frac{1}{2m}\right)T.$$

Meanwhile, if the input instance is insatisfiable, then with probability $0.99$, the algorithm returns False. This is because with probability $0.9$,

$$\mathbb{E}_{t\sim[T],j\sim[m]}\left[\sum_{t=1}^T V^{\pi_t,\nu_j}\right] \geq R - 5\sqrt{T} > R - \frac{T}{4m}.$$

Conditioned on this event, if the algorithm returns True, then

$$\mathbb{E}_{t\sim[T],j\sim[m]}\left[\sum_{t=1}^T V^{\pi_t,\nu_j}\right] \geq T - \frac{T}{2m} - \frac{T}{4m} > \left(1 - \frac{1}{m}\right)T.$$

This implies that $\exists t$:

$$\mathbb{E}_{j\sim[m]}\left[\sum_{t=1}^T V^{\pi_t,\nu_j}\right] > \left(1 - \frac{1}{m}\right)T,$$

which contradicts with the fact that the input is not satisfiable. Therefore the probability that the algorithm returns True when the input is not satisfiable is at most $0.01$.

The reduction above suggests that, if algorithm $\mathcal{A}$ runs in $\mathrm{poly}(S, A, H, T)$ time, we can obtain an algorithm that decides 3-SAT with high probability and runs in $\mathrm{poly}(m, n)$ time. In other words, this implies 3-SAT$\in$BPP, which further implies NP$\subseteq$BPP since 3-SAT is NP-complete.