# Minimax-Optimal Multi-Agent RL in Markov Games With a Generative Model

**Gen Li** UPenn
**Yuejie Chi** CMU
**Yuting Wei** UPenn
**Yuxin Chen** UPenn

## Abstract

This paper studies multi-agent reinforcement learning in Markov games, with the goal of learning Nash equilibria or coarse correlated equilibria (CCE) sample-optimally. All prior results suffer from at least one of the two obstacles: the curse of multiple agents and the barrier of long horizon, regardless of the sampling protocol in use. We take a step towards settling this problem, assuming access to a flexible sampling mechanism: the generative model. Focusing on non-stationary finite-horizon Markov games, we develop a fast learning algorithm called Q-FTRL and an adaptive sampling scheme that leverage the optimism principle in online adversarial learning (particularly the Follow-the-Regularized-Leader (FTRL) method). Our algorithm learns an $\varepsilon$-approximate CCE in a general-sum Markov game using

$$\widetilde{O}\left(\frac{H^4 S \sum_{i=1}^{m} A_i}{\varepsilon^2}\right)$$

samples, where $m$ is the number of players, $S$ indicates the number of states, $H$ is the horizon, and $A_i$ denotes the number of actions for the $i$-th player. This is minimax-optimal (up to log factor) when $m$ is fixed. When applied to two-player zero-sum Markov games, our algorithm provably finds an $\varepsilon$-approximate Nash equilibrium with a minimal number of samples. Along the way, we derive a refined regret bound for FTRL that makes explicit the role of variance-type quantities, which might be of independent interest.

## 1 Introduction

The thriving field of multi-agent reinforcement learning (MARL) studies how a group of interacting agents make decisions autonomously in a shared dynamic environment [81]. The recent developments in game playing [67, 9], self-driving vehicles [59], and multi-robot control [46] are prime examples of MARL in action. In practice, there is no shortage of situations where the agents involved have conflict of interest, and they have to act competitively in order to promote their own benefits (possibly at the expense of one another). Scenarios of this kind are frequently modeled via Markov games (MGs) [60, 42], a framework that has been a fruitful playground to formalize and stimulate the studies of competitive MARL.

In view of the irreconcilable competition between individual players, solutions of competitive MARL normally take the form of certain equilibrium strategy profiles, which are perhaps best epitomized by the concept of Nash equilibrium (NE) [50]. In a Nash equilibrium, no gain can be realized through a unilateral change — assuming no coordination between players — and hence no player has incentives to deviate from her current strategy/policy. A myriad of research has been conducted surrounding NE, which spans various aspects like existence, learnability, computational hardness, and algorithm design, among others [60, 20, 12, 54, 53, 22, 42, 28, 51, 33]. Given that finding NE is notoriously expensive in general (except for special cases like two-player zero-sum MGs) [20, 21], several more tractable solution concepts have emerged in the studies of game theory and MARL, a

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

prominent example being the coarse correlated equilibirum (CCE) [48]. A key compromise made in the CCE is that it permits the players to act in an coordinated fashion, which contrasts sharply with the absence of coordination in the definition of NE.

One critical challenge impacting modern MARL applications is data efficiency. The players involved often have minimal knowledge about how the environment responds to their actions, and have to learn the dynamics and preferable actions by probing the unknown environment. For MARL to expand into applications with enormous dimensionality and long planning horizon, the learning algorithms must manage to make efficient use of the collected data. Nevertheless, how to learn NE and/or CCE with optimal sample complexity remains by and large unsettled even when it comes to the most basic setting: two-player zero-sum Markov games, as we shall discuss below.

**Example: inadequacy in learning two-player zero-sum Markov games.** To facilitate concrete comparisons, let us review two representative algorithms aimed at learning NE in two-player zero-sum MGs. These algorithms have been studied under two drastically different sampling protocols, and we shall discuss the shortfalls of the cutting-edge sample complexity results. In a two-player zero-sum MG, we denote by $S$ the number of states and $H$ the horizon or effective horizon, whereas $A_1$ and $A_2$ denote respectively the number of actions for the max-player and the min-player.

- *Model-based methods under either a generative model or online exploration.* Assuming access to a generative model (so that one can sample arbitrary state-action tuples), Zhang et al. [80] investigated a natural model-based algorithm, which performs planning (e.g., value iteration) on an empirical MG derived from samples produced non-adaptively by the generative model. Focusing on *stationary* discounted infinite-horizon MGs, their algorithm finds an $\varepsilon$-approximate NE with
$$\widetilde{O}\left(\frac{H^3 S A_1 A_2}{\varepsilon^2}\right) \text{ samples.} \tag{1}$$
In parallel, Liu et al. [43] studied *non-stationary* finite-horizon MGs with online exploration, and obtained similar sample complexity bounds, i.e.,
$$\widetilde{O}\left(\frac{H^4 S A_1 A_2}{\varepsilon^2}\right) \text{ samples} \qquad \text{or} \qquad \widetilde{O}\left(\frac{H^3 S A_1 A_2}{\varepsilon^2}\right) \text{ episodes} \tag{2}$$
for learning an $\varepsilon$-approximate NE. While these bounds achieve minimax-optimal dependency on the horizon $H$, a major drawback emerges — commonly referred to as the curse of multiple agents; namely, these results scale proportionally with the total number of *joint actions* (i.e., $\prod_{1\le i\le 2} A_i$), a quantity that blows up exponentially with the number of players.
- *V-learning for online exploration settings.* Focusing on online exploration settings, Bai et al. [5], Jin et al. [31] proposed an algorithm called V-learning that leverages the advances in online adversarial learning (e.g., adversarial bandits) to circumvent the curse of multiple agents. This algorithm provably yields an $\varepsilon$-approximate NE in non-stationary finite-horizon MGs using
$$\widetilde{O}\left(\frac{H^6 S(A_1+A_2)}{\varepsilon^2}\right) \text{ samples} \qquad \text{or} \qquad \widetilde{O}\left(\frac{H^5 S(A_1+A_2)}{\varepsilon^2}\right) \text{ episodes,} \tag{3}$$
which effectively brings down the sample size scaling (2) from $A_1 A_2$ (i.e., the number of joint actions) to $A_1 + A_2$ (i.e., the sum of individual actions). It is worth pointing out, however, that this theory appears sub-optimal in terms of the horizon dependency, as it is a factor of $H^2$ above the minimax lower bound.

**Key issues and our main contributions.** While the above summary focuses on two-player zero-sum MGs, it unveils a fundamental issue surrounding the sample efficiency of learning equilibria; that is, all existing results in this front — irrespective of the sampling mechanism in use — fall short of overcoming at least one of the two major hurdles: (i) the *curse of multiple agents*, and (ii) the *barrier of long horizon*. A natural question to pose is:

**Question:** *can we learn a Nash equilibrium in a two-player zero-sum Markov game in a sample-optimal and computation-efficient fashion?*

To settle this favorably, both of the above hurdles need to be crossed simultaneously. Moving beyond two-player zero-sum MGs, it is not surprising to see that general-sum multi-player MGs have to grapple with the aforementioned two hurdles as well. Thus, the following question also comes into mind when learning CCE (a compromise due to the general intractability of learning NE):

**Question:** *can we learn a coarse correlated equilibrium in a multi-player general-sum Markov game in a sample-optimal and computation-efficient fashion?*

Note that these questions remain open regardless of the sampling scheme in use.

This paper takes a first step towards solving the problem by assuming access to the most flexible sampling protocol: the generative model (or simulator). In stark contrast to the single-agent case where uniform sampling of all state-action pairs suffices [3, 38], the multi-agent scenario requires one to take samples intelligently and adaptively, a crucial step to avoid inefficient use of data (otherwise one cannot hope to break the curse of multiple agents). With the aim of computing an $\varepsilon$-approximate equilibrium in a *non-stationary* finite-horizon MG, we come up with a computationally efficient learning algorithm (accompanied by an adaptive sampling strategy) that accomplishes this goal with no more than

$$\begin{cases} \widetilde{O}\Big(\frac{H^4 S(A_1+A_2)}{\varepsilon^2}\Big) \text{ samples} & \text{(learning } \varepsilon\text{-NE in two-player zero-sum MGs)} \\ \widetilde{O}\Big(\frac{H^4 S\left(\sum_{i=1}^m A_i\right)}{\varepsilon^2}\Big) \text{ samples} & \text{(learning } \varepsilon\text{-CCE in multi-player general-sum MGs)} \end{cases} \tag{4}$$

drawn from the generative model. Encouragingly, this sample complexity bound matches the minimax lower limit (up to a logarithmic factor) as long as the number of players $m \geq 2$ is a fixed constant or grows only logarithmically in problem parameters. Our sample complexity theory is valid for the full $\varepsilon$-range (i.e., any $\varepsilon \in (0, H]$); this unveils that no burn-in cost is needed for our algorithm to achieve sample optimality, which lends itself well to sample-hungry applications.

The proposed algorithm is inspired by two key algorithmic ideas in RL and bandit literature: (i) optimism in the face of uncertainty (by leveraging upper confidence bounds (UCBs) in value estimation), and (ii) online and adversarial learning (particularly the Follow-the-Regularized-Leader (FTRL) algorithm). Note that the optimal design of bonus terms — typically based on certain data-driven variance estimates — is substantially more challenging than the single-agent case, as it requires intricate adaptation in response to the policy changes of one another as well as compatibility with the FTRL dynamics. Two points are worth emphasizing (which will be made precise later on):

- The efficacy of FTRL in breaking the curse of multiple agents has been illustrated in Jin et al. [31], Song et al. [64], Mao and Başar [45]. To improve horizon dependency, one needs to exploit connections between the performance of FTRL and certain variances. Towards this, we develop a refined regret bound for FTRL that unveils the role of variance-style quantities, which was previously unavailable.
- The bonus terms entail Bernstein-style variance estimates that mimic the variance-style quantities appearing in our refined FTRL regret bounds, and are carefully chosen so as to ensure certain decomposability over steps. This is crucial in optimizing the horizon dependency.

Additionally, the policy returned by our algorithm is Markovian (i.e., the action selection probability depends only on the current state $s$ and step $h$), and the algorithm can be carried out in a decentralized manner without the need of directly observing the opponents' actions.

**Notation.** Let us also gather several convenient notation that shall be used multiple times. For any positive integer $n$, we write $[n] \coloneqq \{1, \cdots, n\}$. We shall abuse notation and let $1$ and $0$ denote the all-one vector and the all-zero vector, respectively. For a sequence $\{\alpha_k\}_{k\geq 1} \subseteq (0, 1]$, we define

$$\alpha_i^k \coloneqq \begin{cases} \alpha_i \prod_{j=i+1}^k (1-\alpha_j), & \text{if } 0 < i < k \\ \alpha_k, & \text{if } i = k \end{cases} \tag{5}$$

for any $1 \leq i \leq k$. For a given vector $x \in \mathbb{R}^{SA}$ (resp. $y \in \mathbb{R}^{SAB}$), we denote by $x(s,a)$ (resp. $y(s,a,b)$) the entry of $x$ (resp. $y$) associated with the state-action combination $(s,a)$ (resp. $(s,a,b)$), as long as it is clear from the context. Next, consider any two vectors $a = [a_i]_{1\leq i\leq n}$ and $b = [b_i]_{1\leq i\leq n}$. We use $a \leq b$ (resp. $a \geq b$) to indicate that $a_i \geq b_i$ (resp. $a_i \leq b_i$) holds for all $i$; we allow scalar functions to take vector-valued arguments in order to denote entrywise operations (e.g., $a^2 = [a_i^2]_{1\leq i\leq n}$ and $a^4 = [a_i^4]_{1\leq i\leq n}$); and we denote by $a \circ b = [a_i b_i]_{1\leq i\leq n}$ the Hadamard product. For a finite set $\mathcal{A} = \{1, \cdots, A\}$, we denote by $\Delta(\mathcal{A}) = \{x \in \mathbb{R}^A \mid \sum_i x_i = 1; x \geq 0\}$ the probability simplex over $\mathcal{A}$. For any function $f$ with domain $\mathcal{A}$ (or $\mathcal{B}$), we adopt the notation

$$\mathbb{E}_\pi[f] \coloneqq \sum\nolimits_a \pi(a) f(a) \qquad \text{and} \qquad \mathsf{Var}_\pi(f) \coloneqq \sum\nolimits_a \pi(a)\big(f(a) - \mathbb{E}_\pi[f]\big)^2. \tag{6}$$

## 2 Background and models

In this section, we introduce the basics for Markov games, as well as the solution concepts of Nash equilibrium and coarse correlated equilibrium.

**Markov games.** A non-stationary finite-horizon *multi-player general-sum Markov game*, denoted by $\mathcal{MG} = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le m}, H, P, r\big\}$, involves $m$ players competing against each other, and consists of several key elements to be formalized below. Recall that $\Delta(\mathcal{S})$ represents the probability simplex over the set $\mathcal{S}$.

- $\mathcal{S} = \{1, \cdots, S\}$ is the state space of the shared environment, which comprises $S$ different states.
- For each $1 \le i \le m$, let $\mathcal{A}_i = \{1, \cdots, A_i\}$ represent the action space of the $i$-th player, which contains $A_i$ different actions. Here and below, we denote

$$\mathcal{A} \coloneqq \mathcal{A}_1 \times \cdots \times \mathcal{A}_m \qquad \text{and} \qquad \mathcal{A}_{-i} \coloneqq \prod_{j:j\neq i} \mathcal{A}_j \quad (1 \le i \le m). \tag{7}$$

  Throughout the paper, we shall often use the boldface letter $\boldsymbol{a} \in \mathcal{A}$ (resp. $\boldsymbol{a}_{-i} \in \mathcal{A}_{-i}$) to denote a joint action profile of all players (resp. a joint action profile excluding the $i$-th player's action).
- $H$ stands for the horizon length of the Markov game.
- $P = \{P_h\}_{1\le h\le H}$ — with $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ — denotes the probability transition kernel of $\mathcal{MG}$. Namely, for any $(s, \boldsymbol{a}, h, s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$, we let $P_h(s' \mid s, \boldsymbol{a})$ indicate the probability of $\mathcal{MG}$ transitioning from state $s$ to state $s'$ at step $h$ when the joint action profile taken by the players is $\boldsymbol{a}$.
- $r = \{r_{i,h}\}_{1\le h\le H, 1\le i\le m}$ — with $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ — represents the (deterministic) reward function. Namely, for any $(s, \boldsymbol{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $r_{i,h}(s, \boldsymbol{a})$ stands for the immediate reward the $i$-th player gains in state $s$ at step $h$, if the joint action profile is $\boldsymbol{a}$. Here and throughout, we assume normalized rewards in the sense that $r_{i,h}(s, \boldsymbol{a}) \in [0, 1]$ for any $(s, \boldsymbol{a}, h, i) \in \mathcal{S} \times \mathcal{A} \times [H] \times [m]$.

As an important special case, a *two-player zero-sum Markov game* — denoted by $\mathcal{MG} = \big\{\mathcal{S}, \{\mathcal{A}_1, \mathcal{A}_2\}, H, P, r\big\}$ — satisfies $r_{2,h} = -r_{1,h}$ for all $h \in [H]$. Following the convention, we assume that $r_{1,h} \ge 0$ for all $h \in [H]$,[1] and refer to the first (resp. second) player as the max-player (resp. the min-player).

**Markov policies.** This paper focuses on the class of Markov policies, such that the action selection strategies of the players are determined by the current state $s$ and the step number $h$, without depending on previously visited states. To begin with, let $\pi_i = \{\pi_{i,h}\}_{1\le h\le H}$ represent the policy of the $i$-th player. Here, $\pi_{i,h}(\cdot \mid s) \in \Delta(\mathcal{A}_i)$ for any $(s, h) \in \mathcal{S} \times [H]$, where $\pi_{i,h}(a \mid s)$ indicates the probability of the $i$-th player selecting action $a$ in state $s$ at step $h$. The joint Markov policy can be defined analogously: we let $\pi = (\pi_1, \ldots, \pi_m) : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$ represent a joint Markov policy of all players, where the joint actions of all players in state $s$ and step $h$ are chosen according to the distribution specified by $\pi_h(\cdot \mid s) = (\pi_{1,h}, \ldots, \pi_{m,h})(\cdot \mid s) \in \Delta(\mathcal{A})$. For any given joint policy $\pi$, we employ $\pi_{-i}$ to represent the policies of all but the $i$-th player, and let $\pi_{-i,h}$ denote the policies of all but the $i$-th player at step $h$. All policies are assumed throughout to be Markovian, except our brief remarks on non-Markovian policies in Section 3.2.

Additionally, a joint policy $\pi$ is said to be a *product policy* if $\pi_1, \ldots, \pi_m$ are executed in a statistically independent manner (namely, under policy $\pi$ the players take actions independently), and we shall adopt the notation $\pi = \pi_1 \times \cdots \times \pi_m$ to indicate that $\pi$ is a product policy.

**Value functions.** Consider a Markovian trajectory $\{(s_h, \boldsymbol{a}_h)\}_{1\le h\le H}$, where $s_h \in \mathcal{S}$ is the state at step $h$ and $\boldsymbol{a}_h \in \mathcal{A}$ is the joint action profile at step $h$. For any given joint policy $\pi$ and any step $h \in [H]$, we define the value function $V_{i,h}^{\pi} : \mathcal{S} \to \mathbb{R}$ of the $i$-th player under policy $\pi$ as follows:

$$V_{i,h}^{\pi}(s) \coloneqq \mathbb{E}\left[\sum_{t=h}^{H} r_{i,t}\big(s_t, \boldsymbol{a}_t\big) \mid s_h = s\right], \qquad \forall s \in \mathcal{S}, \tag{8}$$

[1] The careful reader might immediately note that $r_{2,h} \le 0$, thus falling outside our assumed range for the reward function. This, however, can be easily addressed by enforcing a positive global shift to $r_{2,h}$ without changing the learning process.

where the expectation is taken over the Markovian trajectory $\{(s_h, \boldsymbol{a}_h)\}$ with the $m$ players jointly executing policy $\pi$; that is, conditional on $s_h$, we draw $\boldsymbol{a}_h \sim \pi_h(\cdot \mid s_h)$ and then $s_{h+1} \sim P_h(\cdot \mid s_h, \boldsymbol{a}_h)$.

In addition, consider the case where (i) all but the $i$-th player executes the joint policy $\pi_{-i}$ and (ii) the $i$-th player executes policy $\pi_i'$ *independently* from the other players; we shall denote by $V_{i,h}^{\pi_i' \times \pi_{-i}}$ the resulting value function under this joint policy $\pi_i' \times \pi_{-i}$. By optimizing over all $\pi_i'$, we can further define

$$V_{i,h}^{\star,\pi_{-i}}(s) := \max_{\pi_i' : \mathcal{S} \times [H] \to \Delta(\mathcal{A}_i)} V_{i,h}^{\pi_i' \times \pi_{-i}}(s), \qquad \forall (s,h,i) \in \mathcal{S} \times [H] \times [m]. \tag{9}$$

It is known that there exists at least one policy, denoted by $\pi_i^\star(\pi_{-i}) : \mathcal{S} \times [H] \to \Delta(\mathcal{A}_i)$ and commonly referred to as the *best-response policy*, that can simultaneously attain $V_{i,h}^{\star,\pi_{-i}}(s)$ for all $h \in [H]$ and all $s \in \mathcal{S}$. It is worth emphasizing that the best-response policy $\pi_i^\star(\pi_{-i})$ is the best among all policies of the $i$-th player executed independently of $\pi_{-i}$. Furthermore, if we freeze $\pi_{-i}$, then the Bellman optimality condition for the $i$-th player can be expressed as [8]

$$V_{i,h}^{\star,\pi_{-i}}(s) = \max_{a_i \in \mathcal{A}_i} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{a}_{-i} \sim \pi_{-i,h}(\cdot \mid s)} \left[ r_{i,h}(s, \boldsymbol{a}) + \left\langle P_h(\cdot \mid s, \boldsymbol{a}),\, V_{i,h+1}^{\star,\pi_{-i}} \right\rangle \right] \right\} \tag{10}$$

for all $(s,h,i) \in \mathcal{S} \times [H] \times [m]$, where the joint action profile $\boldsymbol{a}$ is composed of $a_i$ for the $i$-th player and $\boldsymbol{a}_{-i}$ for the remaining ones.

**Equilibria of Markov games.** In a multi-agent Markov game, each player wishes to maximize its own value function. Due to the competing objectives, finding some sorts of equilibria — e.g., the Nash equilibrium [49] and the coarse correlated equilibrium [48, 2] — becomes a central topic in the studies of Markov games. Let us introduce these solution concepts below.

- *Nash equilibrium.* A product policy $\pi = \pi_1 \times \cdots \times \pi_m$ is said to be a *(mixed-strategy) Nash equilibrium* of $\mathcal{MG}$ if the following holds:

$$V_{i,1}^{\pi}(s) = V_{i,1}^{\star,\pi_{-i}}(s), \qquad \text{for all } (s,i) \in \mathcal{S} \times [m]. \tag{11}$$

  In other words, conditional on the opponents' current policy and the assumption that all players take actions *independently*, no player can harvest any gain by unilaterally deviating from its current policy.

- *Coarse correlated equilibrium.* A joint policy $\pi$ is said to be a coarse correlated equilibrium of $\mathcal{MG}$ if

$$V_{i,1}^{\pi}(s) \geq V_{i,1}^{\star,\pi_{-i}}(s), \qquad \text{for all } (s,i) \in \mathcal{S} \times [m]. \tag{12}$$

  While a CCE also ensures that no unilateral deviation (performed independently from others) is beneficial, its key distinction from the definition of NE lies in the fact that it allows the policy to be correlated across the players. Any NE of $\mathcal{MG}$ is, self-evidently, also a CCE.

In practice, it might be challenging to compute an "exact" equilibrium, and instead one would seek to find approximate solutions. Towards this end, we find it helpful to define the sub-optimality gap of a policy $\pi$ as follows (measured in an $\ell_\infty$-based manner)

$$\mathsf{gap}(\pi) := \max_{s \in \mathcal{S}} \mathsf{gap}(\pi; s), \tag{13a}$$

where

$$\mathsf{gap}(\pi; s) := \max_{1 \leq i \leq m} \left\{ V_{i,1}^{\star,\pi_{-i}}(s) - V_{i,1}^{\pi}(s) \right\}. \tag{13b}$$

With this sub-optimality measure in place, a *product* policy $\pi = \pi_1 \times \cdots \times \pi_m$ is said to be an $\varepsilon$-approximate NE — or more concisely, $\varepsilon$-Nash — if the resultant sub-optimality gap obeys $\mathsf{gap}(\pi) \leq \varepsilon$. Similarly, a joint (and possibly correlated) policy $\pi$ is said to be an $\varepsilon$-approximate CCE — or more concisely, $\varepsilon$-CCE — if $\mathsf{gap}(\pi) \leq \varepsilon$.

**Generative model / simulator.** In reality, we oftentimes do not have access to perfect descriptions (e.g., accurate knowledge of the transition kernel $P$) of the Markov game under consideration; instead, one has to learn the true model on the basis of data samples. When it comes to the data generating mechanism, this paper assumes access to a generative model (also called a simulator) [35, 34]:

in each call to the generative model, the learner can choose an arbitrary $(s, \boldsymbol{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and obtain an independent sample generated based on the true transition kernel:

$$s' \sim P_h(\cdot \mid s, \boldsymbol{a}).$$

In words, a generative model facilitates query of arbitrary state-action-step tuples, which helps alleviate the sampling constraints arising in online episodic settings for exploration. The goal of the current paper is to compute an $\varepsilon$-approximate equilibrium (either NE or CCE) of $\mathcal{MG}$ with as few samples as possible, i.e., using a minimal number of calls to the generative model.

## 3 Sample-efficient learning with a generative model

In this section, we put forward an efficient algorithm aimed at learning an $\varepsilon$-approximate equilibrium with the assistance of a generative model, and demonstrate its sample optimality for the full $\varepsilon$-range.

### 3.1 Algorithm description

We now describe the proposed algorithm, which is inspired by the optimism principle and the FTRL algorithm for online/adversarial learning. Following the dynamic programming approach [8], our algorithm employs backward recursion from step $h = H$ back to $h = 1$; in fact, we shall finish the sampling and learning processes for step $h$ before moving backward to step $h - 1$. For each $h$, the $i$-th player calls the generative model for $K$ rounds, with each round drawing $SA_i$ independent samples; as a result, the total sample size is given by $KSH \sum_{i=1}^{m} A_i$. In what follows, let us first introduce some convenient notation that facilitates our exposition of the algorithm.

**Notation.** Consider any step $h \in [H]$, any player $i \in [m]$, and any data collection round $k \in [K]$. The algorithm maintains the following iterates, whose notation is gathered here with their formal definitions introduced later.

- $\widehat{V}_{i,h} \in \mathbb{R}^S$ represents the final estimate of the value function at step $h$ by the $i$-th player; in particular, we set $\widehat{V}_{i,H+1} = 0$.
- $Q_{i,h}^k \in \mathbb{R}^{SA_i}$ represents the Q-function estimate of the $i$-th player at step $h$ after the $k$-th round of data collection.
- $q_{i,h}^k \in \mathbb{R}^{SA_i}$ stands for a certain "one-step-look-ahead" Q-function estimate of the $i$-th player at step $h$ using samples collected in the $k$-th round.
- $r_{i,h}^k \in \mathbb{R}^{SA_i}$ denotes the sample reward vector for step $h$ received by the $i$-th player in the $k$-th round.
- $P_{i,h}^k \in \mathbb{R}^{SA_i \times S}$ denotes the empirical probability transition matrix for step $h$ constructed using the samples collected by the $i$-th player in the $k$-th round.
- $\beta_{i,h} \in \mathbb{R}^S$ denotes the bonus vector chosen by the $i$-th player at step $h$ during final value estimation.
- $\pi_{i,h}^k : \mathcal{S} \to \Delta(\mathcal{A}_i)$ denotes the policy iterate of the $i$-th player at step $h$ before the beginning of the $k$-th round of data collection; in particular, we set $\pi_{i,h}^1$ to be uniform, namely, $\pi_{i,h}^1(a_i \mid s) = 1/A_i$ for any $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$.

Crucially, the above objects are all constructed from the perspective of a single player, and hence resemble those needed to operate a "single-agent" MDP (as opposed to MARL). As such, the complexity of storing/updating the above objects only scales with the aggregate size of the individual action space, rather than the size of the product action space.

**Main steps of the proposed algorithm.** As mentioned above, our algorithm collects multiple rounds of independent samples for each $h$. In what follows, let us describe the proposed procedure for the $i$-th player in the $k$-th round for step $h$.

1. *Sampling and model estimation.* For each $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, draw an *independent* sample as follows

$$s'_{k,h,s,a_i} \sim P_h\big(\,\cdot \mid s, \boldsymbol{a}(k,h,s,a_i)\big) \qquad \text{and} \qquad r_{k,i,h,s,a_i} = r_{i,h}\big(s, \boldsymbol{a}(k,h,s,a_i)\big), \tag{14a}$$

where $\boldsymbol{a}(k,h,s,a_i) = [a_j(k,h,s,a_i)]_{1\le j\le m} \in \mathcal{A}$ consists of independent individual actions drawn from

$$a_j(k,h,s,a_i) \overset{\text{ind.}}{\sim} \pi_{j,h}^k(\cdot\,|\,s) \quad (j\neq i) \qquad \text{and} \qquad a_i(k,h,s,a_i) = a_i. \tag{14b}$$

These samples are then employed to construct the sample reward vector $r_{i,h}^k \in \mathbb{R}^{SA_i}$ and empirical probability transition kernel $P_{i,h}^k \in \mathbb{R}^{SA_i\times S}$ such that

$$r_{i,h}^k(s,a_i) = r_{k,i,h,s,a_i} \qquad \text{and} \qquad P_{i,h}^k(s'\,|\,s,a_i) = \begin{cases} 1, & \text{if } s' = s'_{k,h,s,a_i} \\ 0, & \text{else} \end{cases} \tag{14c}$$

for all $(s,a_i,s') \in \mathcal{S}\times\mathcal{A}_i\times\mathcal{S}$. Note that the $i$-th player only needs to compute (14c), without the need of directly observing the other players' actions.

2. *Q-function estimation.* Following the dynamic programming approach, we first compute the "one-step-look-ahead" Q-function estimate as follows

$$q_{i,h}^k = r_{i,h}^k + P_{i,h}^k \widehat{V}_{i,h+1}. \tag{15}$$

We then adopt the update rule of Q-learning:

$$Q_{i,h}^k = (1-\alpha_k)Q_{i,h}^{k-1} + \alpha_k q_{i,h}^k, \tag{16}$$

where $0<\alpha_k<1$ is the learning rate. Applying (16) recursively and using the quantities defined in (5), we easily arrive at the following expansion:

$$Q_{i,h}^k = \sum_{j=1}^k \alpha_j^k q_{i,h}^j. \tag{17}$$

3. *Policy updates.* Once the Q-estimates are updated, we adopt the exponential weights strategy to update the policy iterate of the $i$-th player as follows

$$\pi_{i,h}^{k+1}(a_i\,|\,s) = \frac{\exp\big(\eta_{k+1}Q_{i,h}^k(s,a_i)\big)}{\sum_{a'\in\mathcal{A}_i}\exp\big(\eta_{k+1}Q_{i,h}^k(s,a')\big)}, \qquad \forall (s,a_i)\in\mathcal{S}\times\mathcal{A}_i, \tag{18}$$

where $\eta_{k+1}>0$ is another learning rate associated with policy updates (to be specified shortly). In fact, this subroutine implements the Follow-the-Regularized-Leader strategy [57]:

$$\pi_{i,h}^{k+1}(\cdot\,|\,s) = \arg\min_{\mu\in\Delta(\mathcal{A}_i)} \left\{ -\big\langle \mu, Q_{i,h}^k(s,\cdot)\big\rangle + \frac{1}{\eta_{k+1}}F(\mu)\right\}, \tag{19}$$

where the regularizer $F(\cdot)$ is chosen to be the negative entropy function $F(\mu) := \sum_{a\in\mathcal{A}_i}\mu(a)\log\big(\mu(a)\big)$.

After carrying out $K$ rounds of the above procedure, our final policy estimate $\widehat{\pi}:\mathcal{S}\times[H]\to\Delta(\mathcal{A})$ and the value estimate $\widehat{V}_{i,h}:\mathcal{S}\to\mathbb{R}$ for step $h$ are taken respectively to be

$$\widehat{V}_{i,h}(s) = \min\left\{\sum_{k=1}^K \alpha_k^K \Big\langle \pi_{i,h}^k(\cdot\,|\,s),\, q_{i,h}^k(s,\cdot)\Big\rangle + \beta_{i,h}(s),\ H-h+1\right\} \qquad \text{and} \tag{20a}$$

$$\widehat{\pi}_h(\boldsymbol{a}\,|\,s) = \sum_{k=1}^K \alpha_k^K \prod_{i=1}^m \pi_{i,h}^k(a_i\,|\,s) \tag{20b}$$

for any $\big(s,\boldsymbol{a}=[a_1,\ldots,a_m]\big)\in\mathcal{S}\times\mathcal{A}$, where $\{\alpha_k^K\}$ is defined in (5) and $\beta_{i,h}(s)\ge 0$ is some bonus term (taking the form of some data-driven upper confidence bound) to be specified momentarily. It is worth pointing out that the final policy (20b) takes the form of *a mixture of product policies*. In the special case of two-player zero-sum MGs, we can alternatively output a product policy

$$\text{(two-player zero-sum MGs)} \qquad \widehat{\pi} = \widehat{\pi}_1\times\widehat{\pi}_2, \tag{21}$$

where for each $i=1,2$, we take $\widehat{\pi}_i = \{\widehat{\pi}_{i,h}\}_{1\le h\le H}$ with $\widehat{\pi}_{i,h} = \sum_{k=1}^K \alpha_k^K \pi_{i,h}^k$.

**Algorithm 1:** Q-FTRL.

**Input:** number of rounds $K$ for each step, learning rates $\{\alpha_k\}$ (cf. (22)) and $\{\eta_{k+1}\}$ (cf. (23)).

`// set initial value estimates to 0, and initial policies to uniform distributions.`

**Initialize:** for any $i \in [m]$ and any $(s, a_i, h) \in \mathcal{S} \times \mathcal{A}_i \times [H]$, set $\widehat{V}_{i,H+1}(s) = Q^0_{i,h}(s, a_i) = 0$ and $\pi^1_{i,h}(a_i \mid s) = 1/A_i$.

**for** $h = H$ **to** $1$ **do**

  **for** $k = 1$ **to** $K$ **do**

    **for** $i = 1$ **to** $m$ **do**

      `// draw independent samples, and construct empirical models.`

      $\big(r^k_{i,h}, P^k_{i,h}\big) \leftarrow \texttt{sampling}\big(i, h, \pi^k_h = \{\pi^k_{j,h}\}_{j\in[m]}\big)$. `/* see Algorithm 2. */`

      `// update Q-estimates with upper confidence bounds.`

      Compute $q^k_{i,h} = r^k_{i,h} + P^k_{i,h}\widehat{V}_{i,h+1}$, and update $Q^k_{i,h}$ according to (16).

      `// update policy estimates using FTRL.`

      Update $\pi^{k+1}_{i,h}$ according to (18).

  `// output the final value estimate for step` $h$.

  **for** $i = 1$ **to** $m$ **do**

    Update $\widehat{V}_{i,h}$ according to (20a), where $\beta^k_{i,h}$ is given in (24).

**if** *$\mathcal{MG}$ is a two-player zero-sum Markov game* **then**

  **output:** $\widehat{\pi}_1 \times \widehat{\pi}_2$, where for any $i = 1, 2$, $\widehat{\pi}_i = \{\widehat{\pi}_{i,h}\}_{1\le h\le H}$ with $\widehat{\pi}_{i,h} = \sum_{k=1}^K \alpha^K_k \pi^k_{i,h}$.

**if** *$\mathcal{MG}$ is a multi-player general-sum Markov game* **then**

  **output:** $\widehat{\pi} = \{\widehat{\pi}_h\}_{1\le h\le H}$, where $\widehat{\pi}_h = \sum_{k=1}^K \alpha^K_k \big(\pi^k_{1,h} \times \cdots \times \pi^k_{m,h}\big)$.

The whole procedure is summarized in Algorithm 1.

**Choices of learning rates.** Thus far, we have not yet specified the two sequences of learning rates, which we describe now. The learning rates associated with Q-function updates are set to be rescaled linear, namely,

$$\alpha_k = \frac{c_\alpha \log K}{k - 1 + c_\alpha \log K}, \qquad k = 1, 2, \ldots \tag{22}$$

for some constant $c_\alpha \ge 24$. In addition, the learning rates associated with policy updates are chosen to be:

$$\eta_{k+1} = \sqrt{\frac{\log K}{\alpha_k H}}, \qquad k = 1, 2, \ldots \tag{23}$$

**Choices of bonus terms.** It remains to specify the bonus terms, which are selected based on fairly intricate upper confidence bounds. This constitutes a key — and perhaps the most challenging — component of our algorithm design. Specifically, we take

$$\beta_{i,h}(s) = c_{\mathsf{b}} \sqrt{\frac{\log^3\big(\frac{KS\sum_i A_i}{\delta}\big)}{KH}} \sum_{k=1}^K \alpha^K_k \left\{ \mathsf{Var}_{\pi^k_{i,h}(\cdot|s)}\Big(q^k_{i,h}(s,\cdot)\Big) + H \right\} \tag{24}$$

for any $(i, s, h) \in [m] \times \mathcal{S} \times [H]$, where $c_{\mathsf{b}} > 0$ is some sufficiently large constant; see also (6) for the definition of the variance-style quantity. As in previous works, the bonus terms, which are chosen carefully in a data-driven fashion, need to compensate for the uncertainty incurred during the estimation process.

### 3.2 Main results

As it turns out, the proposed algorithm is tractable and provably sample-efficient. We begin by characterizing its sample complexity when learning Nash equilibria in two-player zero-sum MGs,

**Algorithm 2:** Auxiliary function $\texttt{sampling}\big(i, h, \pi_h = \{\pi_{j,h}\}_{j\in[m]}\big)$.

**Initialize:** $\overline{r} = 0 \in \mathbb{R}^{SA_i}$, and $\overline{P} = 0 \in \mathbb{R}^{SA_i \times S}$.

**for** $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ **do**

Draw an independent sample from the generative model: $s'_{s,a_i} \sim P_h\big(\,\cdot \mid s, \boldsymbol{a}(s, a_i)\big)$, where $\boldsymbol{a}(s, a_i) = [a_j(s, a_i)]_{1\le j\le m}$ is composed of independent individual actions drawn from

$$a_j(s, a_i) \overset{\text{ind.}}{\sim} \pi_{j,h}(\cdot \mid s) \quad (j \neq i) \qquad \text{and} \qquad a_i(s, a_i) = a_i. \tag{25}$$

Set $\overline{r}(s, a_i) = r_{i,h}\big(s, \boldsymbol{a}(s, a_i)\big)$ and $\overline{P}\big(s'_{s,a_i} \mid s, a_i\big) = 1$.

**Return:** $\big(\overline{r}, \overline{P}\big)$.

and then shift attention to learning CCE in multi-player general-sum MGs (given the intractability of learning NEs in general).

**Theorem 1** (NE for two-player zero-sum MGs)**.** *Consider a two-player zero-sum Markov game, and consider any $\varepsilon \in (0, H]$ and any $0 < \delta < 1$. Suppose that*

$$K \ge \frac{c_{\mathsf{k}} H^3 \log^4\big(\frac{KS(A_1+A_2)}{\delta}\big)}{\varepsilon^2} \tag{26}$$

*for some large enough universal constant $c_{\mathsf{k}} > 0$. With probability at least $1-\delta$, the product policy $\widehat{\pi}_1 \times \widehat{\pi}_2$ computed by Algorithm 1 is an $\varepsilon$-approximate Nash equilibrium, i.e., its sub-optimality gap (cf. (13)) obeys $\mathsf{gap}\big(\widehat{\pi}_1 \times \widehat{\pi}_2\big) \le \varepsilon$.*

**Theorem 2** (CCE for multi-player general-sum MGs)**.** *Consider an $m$-player general-sum Markov game, and consider any $\varepsilon \in (0, H]$ and any $0 < \delta < 1$. Suppose that*

$$K \ge \frac{c_{\mathsf{k}} H^3 \log^4\big(\frac{KS\sum_{i=1}^m A_i}{\delta}\big)}{\varepsilon^2} \tag{27}$$

*for some large enough universal constant $c_{\mathsf{k}} > 0$. With probability at least $1-\delta$, the joint policy $\widehat{\pi}$ returned by Algorithm 1 is an $\varepsilon$-approximate CCE, i.e., its sub-optimality gap (cf. (13)) obeys $\mathsf{gap}\big(\widehat{\pi}\big) \le \varepsilon$.*

Theorems 1-2 establish sample complexity upper bounds for the proposed algorithm, which we take a moment to interpret as follows. The proofs of these two theorems are postponed to Appendix B in the supplementary material.

**Sample complexity.** When a generative model is available, Theorems 1-2 assert that the total number of samples (i.e., $KSH\sum_i A_i$) needed for Algorithm 1 to work is

$$\begin{cases} \widetilde{O}\big(\frac{H^4 S(A_1+A_2)}{\varepsilon^2}\big), & \text{for learning an } \varepsilon\text{-NE in two-player zero-sum MGs;} \\ \widetilde{O}\big(\frac{H^4 S\sum_{i=1}^m A_i}{\varepsilon^2}\big), & \text{for learning an } \varepsilon\text{-CCE in multi-player general-sum MGs.} \end{cases} \tag{28}$$

As far as we know, our theorems deliver the first results that uncover the plausibility of simultaneously overcoming the long-horizon barrier and the curse of multi-agents. Let us compare (28) with prior art.

- *NE in two-player zero-sum MGs.* First, consider learning $\varepsilon$-NE policies in two-player zero-sum MGs. In comparison to Zhang et al. [80] (cf. (1)), our result reveals that what ultimately matters is the total number of individual actions (i.e., $A_1 + A_2$) as opposed to the total number $A_1 A_2$ of possible joint actions; additionally, our results exhibit improved horizon dependency (by a factor of $H^2$) compared to Bai et al. [5], Jin et al. [31] (see (3)), although we remark that the online sampling protocol therein is clearly more restrictive than a generative model.
- *CCE in multi-player general-sum MGs (for a fixed $m$).* Similar messages carry over to the task of learning multi-player general-sum MGs when the number of players $m$ is a fixed constant. Liu et al. [43] provided the first non-asymptotic result on learning CCE in the exploration setting; the model-based algorithm studied therein learns an $\varepsilon$-CCE using

$$\widetilde{O}\bigg(\frac{H^5 S^2 \prod_{i=1}^m A_i}{\varepsilon^2}\bigg) \text{ samples} \qquad \text{or} \qquad \widetilde{O}\bigg(\frac{H^4 S^2 \prod_{i=1}^m A_i}{\varepsilon^2}\bigg) \text{ episodes} \tag{29}$$

which is sub-optimal in terms of the dependency on both $H$ and $S$ and suffers from the curse of multiple agents. A more recent strand of works focused on a type of online RL algorithms called V-learning, which exploited the effectiveness of adversarial learning subroutines in overcoming the curse of multi-agents [45, 64, 31]; along this line, the state-of-the-art sample complexity bound is [31]:

$$\widetilde{O}\left(\frac{H^6 S \max_{1\le i\le m} A_i}{\varepsilon^2}\right) \text{ samples} \qquad \text{or} \qquad \widetilde{O}\left(\frac{H^5 S \max_{1\le i\le m} A_i}{\varepsilon^2}\right) \text{ episodes,} \tag{30}$$

which remains suboptimal in terms of the horizon dependency. As a drawback of these works, the policy returned by V-learning is non-Markovian, an issue that has been recently addressed by Daskalakis et al. [23] at the price of a much higher sample complexity. It is worth emphasizing that all these works assume the online exploration setting as opposed to the scenario with a generative model.

**Minimax optimality.** To assess the tightness of our result (28), it is helpful to look at the information-theoretic limit. Following the minimax lower bound for single-agent MDPs [3, 41], one can develop a minimax sample complexity lower bound for Markov games (w.r.t. finding either an $\varepsilon$-NE or an $\varepsilon$-CCE) that scales as

$$\text{(minimax lower bound)} \qquad \frac{H^4 S\big(\sum_i A_i\big)}{\varepsilon^2} \tag{31}$$

modulo some logarithmic factor, provided that $m$ is a fixed constant or grows only logarithmically in problem parameters; see Appendix D.3 in the supplementary material for a formal statement and its proof. Taking this together with (28) confirms the minimax optimality of our algorithm (up to logarithmic terms) when the number $m$ of players is fixed.

**No burn-in sample size and full $\varepsilon$-range.** It is noteworthy that the validity of our sample complexity bound (28) is guaranteed for the entire range of $\varepsilon$-levels (i.e., any $\varepsilon \in (0, H]$). This feature is particularly appealing in the data-starved applications, as it implies that there is no burn-in sample size needed for our algorithm to work optimally.

**Miscellaneous properties of our algorithm.** Finally, we would like to remark in passing that our learning algorithm enjoys several properties that might be practically appealing. For instance, the output policies are Markovian in nature, which depend only on the current state $s$ and step number $h$. This is enabled thanks to the availability of the generative model, which allows us to settle the sampling and learning process for step $h+1$ completely before moving backward to step $h$; in contrast, the online sampling protocol studied in Bai et al. [5], Jin et al. [31] cannot be implemented in this way without incurring information loss. In addition, our algorithm can be carried out in a decentralized fashion (except that the final estimate $\widehat{\pi}$ needs to aggregate policy iterates from all players), with each player acting in a symmetric yet independent manner (without the need of knowing each other's individual action). Our algorithm is also "rational" in the sense that it converges to the best-response policy of a player if all other players freeze their policies. All this is achieved under minimal sample complexity with the aid of the generative model.

## 4 Discussion

The primary contribution of this paper has been to develop a sample-optimal paradigm that simultaneously overcomes the curse of multiple agents and optimizes the horizon dependency when solving multi-player Markov games. This goal was not accomplished in any of the previous works, regardless of the sampling mechanism in use. The adoption of the adversarial learning subroutine helps break the curse of multiple agents compared to the prior model-based approach [80, 43], whereas the availability of the generative model in conjunction with the variance-aware bonus design improves horizon dependency compared to Bai et al. [5], Jin et al. [31].

**Acknowledgements:** Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, IIS-2218713 and IIS-2218773. Y. Wei is supported in part by the the NSF grants CCF-2106778, DMS-2147546/2015447 and CAREER award DMS-2143215. Y. Chi are supported in part by the grants ONR N00014-19-1-2404, NSF CCF-2106778 and DMS-2134080, and CAREER award ECCS-1818571. Part of this work was done while G. Li, Y. Wei and Y. Chen were visiting the Simons Institute for the Theory of Computing.

## References


[1] A. Agarwal, S. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[2] R. J. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18, 1987.

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

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

[5] Y. Bai, C. Jin, and T. Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

[6] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[7] C. L. Beck and R. Srikant. Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208, 2012.

[8] D. P. Bertsekas. *Dynamic programming and optimal control (4th edition)*. Athena Scientific, 2017.

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

[10] S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] S. Cen, Y. Chi, S. Du, and L. Xiao. Faster last-iterate convergence of policy optimization in zero-sum markov games. *arXiv preprint arXiv:2210.01050*, 2022.

[12] X. Chen, Y. Cheng, and B. Tang. Well-supported versus approximate Nash equilibria: Query complexity of large games. *arXiv preprint arXiv:1511.00785*, 2015.

[13] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam. Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*, 2020.

[14] Z. Chen, S. Ma, and Y. Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum Markov game. In *International Conference on Learning Representations*, 2021.

[15] Z. Chen, D. Zhou, and Q. Gu. Almost optimal algorithms for two-player Markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021.

[16] Z. Chen, D. Zhou, and Q. Gu. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022.

[17] Q. Cui and S. S. Du. When is offline two-player zero-sum Markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022.

[18] Q. Cui and S. S. Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *arXiv preprint arXiv:2206.00159*, 2022.

[19] Q. Cui and L. F. Yang. Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504. PMLR, 2021.

[20] C. Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

[21] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[22] C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5527–5540, 2020.

[23] C. Daskalakis, N. Golowich, and K. Zhang. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

[24] Z. Dou, Z. Yang, Z. Wang, and S. Du. Gap-dependent bounds for two-player markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 432–455, 2022.

[25] S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

[26] E. Even-Dar and Y. Mansour. Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25, 2003.

[27] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[28] T. D. Hansen, P. B. Miltersen, and U. 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.

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

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

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

[32] Y. Jin and A. Sidford. Towards tight bounds on the sample complexity of average-reward MDPs. In *International Conference on Machine Learning*, pages 5055–5064. PMLR, 2021.

[33] Y. Jin, V. Muthukumar, and A. Sidford. The complexity of infinite-horizon general-sum stochastic games. *arXiv preprint arXiv:2204.04186*, 2022.

[34] S. Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London, 2003.

[35] M. Kearns, Y. Mansour, and A. Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49(2-3):193–208, 2002.

[36] K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040, 2021.

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

[38] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[39] G. Li, C. Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

[40] G. Li, L. Shi, Y. Chen, Y. Gu, and Y. Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[41] G. Li, L. Shi, Y. Chen, Y. Chi, and Y. Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

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

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

[44] Q. Liu, C. Szepesvári, and C. Jin. Sample-efficient reinforcement learning of partially observable markov games. *arXiv preprint arXiv:2206.01315*, 2022.

[45] W. Mao and T. Başar. Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games and Applications*, pages 1–22, 2022.

[46] L. Matignon, L. Jeanpierre, and A.-I. Mouaddib. Coordinated multi-robot exploration under communication constraints using decentralized markov decision processes. In *Twenty-sixth AAAI conference on artificial intelligence*, 2012.

[47] W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. *arXiv preprint arXiv:2004.04719*, 2020.

[48] H. Moulin and J.-P. Vial. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3):201–221, 1978.

[49] J. Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[50] J. F. Nash Jr. Equilibrium points in $n$-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[51] A. Ozdaglar, M. O. Sayin, and K. Zhang. Independent learning in stochastic games. *arXiv preprint arXiv:2111.11743*, 2021.

[52] A. Pananjady and M. J. Wainwright. Instance-dependent $\ell_\infty$-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585, 2020.

[53] J. Perolat, B. Scherrer, B. Piot, and O. Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329. PMLR, 2015.

[54] A. Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. In *Annual Symposium on Foundations of Computer Science (FOCS)*, pages 258–265, 2016.

[55] M. Sayin, K. Zhang, D. Leslie, T. Basar, and A. Ozdaglar. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34, 2021.

[56] S. Shalev-Shwartz. Online learning: Theory, algorithms, and applications. 2007.

[57] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[58] S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2):115–142, 2007.

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

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

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

[62] A. Sidford, M. Wang, X. Wu, and Y. Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Symposium on Discrete Algorithms*, pages 770–787. SIAM, 2018.

[63] A. Sidford, M. Wang, L. Yang, and Y. 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. PMLR, 2020.

[64] Z. Song, S. Mei, and Y. Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

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

[66] S. Vaswani, L. F. Yang, and C. Szepesvári. Near-optimal sample complexity bounds for constrained MDPs. *arXiv preprint arXiv:2206.06270*, 2022.

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

[68] M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp $\ell_\infty$-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

[69] M. J. Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

[70] B. Wang, Y. Yan, and J. Fan. Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in Neural Information Processing Systems*, 34, 2021.

[71] C.-Y. Wei, Y.-T. Hong, and C.-J. Lu. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

[72] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on Learning Theory*, pages 4259–4299. PMLR, 2021.

[73] G. Weisz, P. Amortila, and C. Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

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

[75] Y. Yan, G. Li, Y. Chen, and J. Fan. Model-based reinforcement learning is minimax-optimal for offline zero-sum markov games. *arXiv preprint arXiv:2206.04044*, 2022.

[76] L. Yang and M. Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.

[77] Y. Yang and C. Ma. $o(t^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. *arXiv preprint arXiv:2209.12430*, 2022.

[78] A. Zanette, M. J. Kochenderfer, and E. Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. *Advances in Neural Information Processing Systems*, 32, 2019.

[79] A. Zanette, A. Lazaric, M. J. Kochenderfer, and E. Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 33:11756–11766, 2020.

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

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

[82] R. Zhang, Q. Liu, H. Wang, C. Xiong, N. Li, and Y. Bai. Policy optimization for Markov games: Unified framework and faster convergence. *arXiv preprint arXiv:2206.02640*, 2022.

[83] Y. Zhao, Y. Tian, J. D. Lee, and S. S. Du. Provably efficient policy gradient methods for two-player zero-sum Markov games. *arXiv preprint arXiv:2102.08903*, 2021.

[84] H. Zhong, W. Xiong, J. Tan, L. Wang, T. Zhang, Z. Wang, and Z. Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. *arXiv preprint arXiv:2202.07511*, 2022.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is a theoretical work that we do not foresee any potential negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

**Other related works.** Let us discuss in passing additional prior works on learning equilibrium solutions in MARL, which have attracted an explosion of interest in recent years. While the Nash equilibrium is arguably the most compelling solution concept in Markov games, the finite-sample/finite-time studies of NE learning concentrate primarily on two-player zero-sum MGs (e.g., Bai and Jin [4], Chen et al. [16], Mao and Başar [45], Wei et al. [71], Tian et al. [65], Cui and Du [17, 18], Zhong et al. [84], Jia et al. [29], Yang and Ma [77], Yan et al. [75], Dou et al. [24]), mainly because computing NEs becomes, for the most part, computationally infeasible (i.e., PPAD-complete) when going beyond two-player zero-sum MGs [20, 21]. Roughly speaking, previous NE-finding algorithms for two-player zero-sum Markov games can be categorized into model-based algorithms [53, 80, 43], value-based algorithms [4, 5, 74, 55, 31, 15], and policy-based algorithms [10, 22, 72, 83, 14, 82, 11]. In particular, Bai et al. [5], Jin et al. [31] developed the first algorithms to beat the curse of multiple agents in two-player zero-sum MGs, while Jin et al. [31], Daskalakis et al. [23], Mao and Başar [45], Song et al. [64] further demonstrated how to accomplish the same goal when learning other computationally tractable solution concepts (e.g., coarse correlated equilibria) in general-sum multi-player Markov games. The recent works Cui and Du [17, 18], Yan et al. [75] studied how to alleviate the sample size scaling with the number of agents in the presence of offline data, with Cui and Du [18] providing a sample-efficient algorithm that also learns NEs in multi-agent Markov games (despite computational intractability). The studies of Markov games have recently been extended to partially observable settings as well [44], which are beyond the scope of the present work.

We shall also briefly remark on the prior works that concern RL with a generative model. While there are multiple sampling mechanisms (e.g., online exploratory sampling, offline data) that bear practical relevance, the generative model (or simulator) serves as an idealistic sampling protocol that has received much recent attention, covering the design of various model-based, model-free and policy-based algorithms [35, 70, 1, 3, 38, 32, 61, 68, 69, 39, 34, 62, 52, 36, 13, 26, 7, 72, 66, 78, 73, 47, 76, 79, 25]. In single-agent RL, the model-based approach has been shown to be minimax-optimal for the entire $\varepsilon$-range [38, 1, 3]. When it comes to multi-agent RL, sample-efficient solutions with a generative model have been proposed in the recent works [63, 19, 80], although a provably sample-optimal strategy was previously unavailable.

# A Regret bounds for FTRL via variance-type quantities

Before embarking on our analysis for Markov games, we take a detour to study the celebrated Follow-the-Regularized-Leader algorithm for online weighted linear optimization, which plays a central role in the analysis of Markov games.

## A.1 Setting: online learning for weighted linear optimization

Let $\ell_1, \ldots, \ell_n \in \mathbb{R}^A$ represent an arbitrary sequence of *non-negative* loss vectors. We focus on the following setting of online learning or adversarial learning [37]: in each round $k$,

1. the learner makes a randomized prediction by choosing a distribution $\pi_k \in \Delta(\mathcal{A})$ over the actions in $\mathcal{A} = \{1, \cdots, A\}$;
2. subsequently, the learner observes the loss vector $\ell_k$, which is permitted to be adversarially chosen.

To evaluate the performance of the learner, we resort to a regret metric w.r.t. a certain weighted linear objective function. To be precise, consider a non-negative sequence $\{\alpha_k\}_{1\le k\le n}$ with $0 \le \alpha_k \le 1$; for each $1 \le k \le n$, we define recursively the following weighted average of the loss vectors:

$$L_0 = 0 \qquad \text{and} \qquad L_k = (1-\alpha_k)L_{k-1} + \alpha_k \ell_k, \quad k \ge 1,$$

which can be easily shown to enjoy the following expression

$$L_k = \sum_{i=1}^{k} \alpha_i^k \ell_k$$

with $\alpha_i^k$ defined in (5). When the sequential predictions made by the learner are $\{\pi_k\}_{k\ge1}$, we define the associated regret w.r.t. the above weighted sum of loss vectors as follows:

$$R_n \coloneqq \max_{a\in\mathcal{A}} R_n(a) \qquad \text{with } R_n(a) \coloneqq \sum_{k=1}^{n} \alpha_k^n \langle \pi_k, \ell_k\rangle - \sum_{k=1}^{n} \alpha_k^n \ell_k(a), \tag{32}$$

which compares the learner's performance (i.e., the expected loss of the learner over time if it draws actions based on $\pi_k$ in round $k$) against that of the best *fixed* action in hindsight.

### A.2 Refined regret bounds for FTRL

**Follow-the-Regularized-Leader.** The FTRL algorithm [58, 56] tailored to the above online optimization setting adopts the following update rule:

$$\pi_{k+1} = \arg\min_{\pi\in\Delta(\mathcal{A})} \left\{ \langle \pi, L_k \rangle + \frac{1}{\eta_{k+1}} F(\pi) \right\}, \qquad k = 1, 2, \ldots \tag{33}$$

where $\eta_{k+1} > 0$ denotes the learning rate, and $F(\cdot)$ is some convex regularization function employed to stabilize the learning process [57]. Throughout this section, we restrict our attention to negative-entropy regularization, namely,

$$F(\pi) = \sum_{a\in\mathcal{A}} \pi(a) \log\big(\pi(a)\big),$$

which allows one to express the FTRL update rule as the following exponential weights strategy (see, e.g., Lattimore and Szepesvári [37, Section 28.1])

$$\pi_{k+1}(a) = \frac{\exp\big(-\eta_{k+1} L_k(a)\big)}{\sum_{a'\in\mathcal{A}} \exp\big(-\eta_{k+1} L_k(a')\big)} \qquad \text{for all } a \in \mathcal{A}. \tag{34}$$

This update rule is also intimately connected to online mirror descent [37].

**Refined regret bounds via variance-style quantities.** As it turns out, the regret of FTRL can be upper bounded by certain (weighted) variance-type quantities, as asserted by the following theorem.

**Theorem 3.** *Suppose that* $0 < \alpha_1 \le 1$ *and* $\eta_1 = \eta_2(1-\alpha_1)$. *Also, assume that* $0 < \alpha_k < 1$ *and* $0 < \eta_{k+1}(1-\alpha_k) \le \eta_k$ *for all* $k \ge 2$. *In addition, define*

$$\widehat{\eta}_k := \begin{cases} \eta_2, & \text{if } k = 1, \\ \frac{\eta_k}{1-\alpha_k}, & \text{if } k > 1. \end{cases} \tag{35}$$

*Then the regret (cf.* (32)*) of the FTRL algorithm satisfies*

$$R_n \le \frac{5}{3} \sum_{k=1}^{n} \alpha_k^n \widehat{\eta}_k \alpha_k \mathsf{Var}_{\pi_k}(\ell_k) + \frac{\log A}{\eta_{n+1}} + 3 \sum_{k=1}^{n} \alpha_k^n \widehat{\eta}_k^2 \alpha_k^2 \big\|\ell_k\big\|_\infty^3 \mathbb{1}\left( \widehat{\eta}_k \alpha_k \big\|\ell_k\big\|_\infty > \frac{1}{3} \right), \tag{36}$$

*where for any* $\ell \in \mathbb{R}^A$ *and any* $\pi \in \Delta(\mathcal{A})$ *we define*

$$\mathsf{Var}_\pi(\ell) := \sum_a \pi(a) \Big( \ell(a) - \sum_{a'} \pi(a')\ell(a') \Big)^2.$$

***Remark*** 1. Note that the FTRL algorithm and the data generating process in this section are both described in a completely deterministic manner; no randomness is involved in the above theorem even though we introduce the variance-style quantities.

The proof of Theorem 3 is postponed to Appendix C. Let us take a moment to discuss the key distinction between Theorem 3 and prior theory.

- A key term in the regret bound (36) is a weighted sum of the "variance-style" quantities $\{\mathsf{Var}_{\pi_k}(\ell_k)\}$. In comparison, prior regret bounds typically involve the norm-type quantities (e.g., the infinity norms $\{\|\ell_k\|_\infty^2\}$) as opposed to the "variances"; see, for instance, Lattimore and Szepesvári [37, Corollary 28.8] for a representative existing regret bound that takes the form of the sum of $\{\|\ell_k\|_\infty^2\}$ that takes the form of the sum of $\{\|\ell_k\|_\infty^2\}$.[2] While $\mathsf{Var}(\ell_k) \le \|\ell_k\|_\infty^2$ is orderwise tight in the worst-case scenario for a given iteration $k$, exploiting the problem-specific variance-type structure across time is crucial in sharpening the horizon dependence in many RL problems (e.g., Azar et al. [3], Jin et al. [30], Li et al. [41, 40]).

[2]Note that the Bregman divergence generated by the negative entropy function is the (generalized) KL divergence [6], which is strongly convex w.r.t. $\|\cdot\|_1$ due to Pinsker's inequality. Additionally, the dual norm of $\|\cdot\|_1$ is the infinity norm.

- The careful reader would remark that the final term of (36) relies on the infinity norm $\|\ell_k\|_\infty$ as well. Fortunately, when the products of the learning rates $\widehat{\eta}_k\alpha_k$ are chosen to be diminishing (which is the case in our analysis for Markov games), the number of iterations obeying $\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > 1/3$ is reasonably small, thus ensuring that this term does not exert too much of an influence on the regret bound.

# B Proof of Theorems 1-2

To begin with, we claim that Theorem 1 is a direct consequence of Theorem 2. Towards this, note that in a two-player zero-sum Markov game, it is self-evident that $\widehat{\pi}_{-1} = \widehat{\pi}_2$ and $\widehat{\pi}_{-2} = \widehat{\pi}_1$ (see line 12 of Algorithm 1). Consequently, Theorem 2 (if it is valid) reveals that

$$\begin{aligned}\varepsilon \geq \mathsf{gap}(\widehat{\pi}; s) &= \max\left\{V_{1,1}^{\star,\widehat{\pi}_{-1}}(s) - V_{1,1}^{\widehat{\pi}}(s),\, V_{2,1}^{\star,\widehat{\pi}_{-2}}(s) - V_{2,1}^{\widehat{\pi}}(s)\right\} \\ &= \max\left\{V_{1,1}^{\star,\widehat{\pi}_2}(s) - V_{1,1}^{\widehat{\pi}}(s),\, V_{2,1}^{\widehat{\pi}_1,\star}(s) - V_{2,1}^{\widehat{\pi}}(s)\right\}, \qquad \text{for all } s \in \mathcal{S}. \end{aligned} \tag{37}$$

Moreover, recalling that $r_{1,h} = -r_{2,h}$ for all $h \in [H]$, one has $V_{1,1}^{\pi}(s) = -V_{2,1}^{\pi}(s)$ for any joint policy profile $\pi$, which taken collectively with (37) results in

$$\begin{aligned}V_{1,1}^{\star,\widehat{\pi}_2}(s) - V_{1,1}^{\widehat{\pi}_1\times\widehat{\pi}_2}(s) &= V_{1,1}^{\star,\widehat{\pi}_2}(s) + V_{2,1}^{\widehat{\pi}_1\times\widehat{\pi}_2}(s) \leq V_{1,1}^{\star,\widehat{\pi}_2}(s) + V_{2,1}^{\widehat{\pi}_1,\star}(s) \\ &= V_{1,1}^{\star,\widehat{\pi}_2}(s) - V_{1,1}^{\widehat{\pi}}(s) + V_{2,1}^{\widehat{\pi}_1,\star}(s) - V_{2,1}^{\widehat{\pi}}(s) \leq 2\varepsilon.\end{aligned}$$

Analogously, one has $V_{2,1}^{\widehat{\pi}_1,\star}(s) - V_{2,1}^{\widehat{\pi}_1\times\widehat{\pi}_2}(s) \leq 2\varepsilon$. Replacing $\varepsilon$ with $\varepsilon/2$ immediately establishes Theorem 1.

With the above argument in mind, the remainder of this section is devoted to proving Theorem 2.

## B.1 Preliminaries and notation

Let us start with some preliminary facts and notation. Given that $\varepsilon \leq H$, the assumption (27) requires

$$K \geq c_{\mathsf{k}} H \log^4\left(\frac{KS\sum_i A_i}{\delta}\right) \tag{38}$$

for some large enough constant $c_{\mathsf{k}} > 0$, which will be a condition assumed throughout the proof. We also gather below several basic facts about our choices of learning rates $\{\alpha_i\}$ (cf. (22)) and the corresponding quantities $\{\alpha_i^k\}$ (cf. (5)).

**Lemma 1.** *For any $k \geq 1$, one has*

$$\alpha_1 = 1, \qquad \sum_{i=1}^{k}\alpha_i^k = 1, \qquad \max_{1\leq i\leq k}\alpha_i^k \leq \frac{2c_\alpha \log K}{k}. \tag{39a}$$

*In addition, if $k \geq c_\alpha \log K + 1$ and $c_\alpha \geq 24$, then one has*

$$\max_{1\leq i\leq k/2}\alpha_i^k \leq 1/K^6. \tag{39b}$$

*Proof.* The result (39a) is standard and has been recorded in previous works (e.g., Jin et al. [30, Appendix B]). Regarding (39b), we note that for any $i \leq k/2$ and $k \geq c_\alpha \log K + 1$,

$$\alpha_i^k \leq \prod_{j=i+1}^{k}(1-\alpha_j) \leq \prod_{j=k/2+1}^{k}(1-\alpha_j) \leq (1-\alpha_k)^{k/2} \leq \left(1-\frac{c_\alpha \log K}{2k}\right)^{k/2} \leq \exp\left(-\frac{c_\alpha \log K}{4}\right) \leq \frac{1}{K^6},$$

where we have used the fact that $\alpha_k = \frac{c_\alpha \log K}{k-1+c_\alpha \log K} \geq \frac{c_\alpha \log K}{2k}$ and the assumption $c_\alpha \geq 24$. □

Additionally, recognizing the definition in (15) and the upper bound $\widehat{V}_{i,h+1}(s) \leq H - h + 1$ (cf. (20a)), we make note of the range of the iterates $\left\{q_{i,h}^k\right\}$ as follows.

**Lemma 2.** *For any $i \in [m]$ and any $(h, k, s, a_i) \in [H] \times [K] \times \mathcal{S} \times \mathcal{A}_i$, it holds that*

$$0 \le q_{i,h}^k(s, a_i) \le H - h + 1. \tag{40}$$

Next, we introduce several additional notation that helps simplify our presentation of the proof. For any policy $\mu : \mathcal{S} \times [H] \to \Delta(\mathcal{A}_i)$, we adopt the convenient notation

$$\mu_h(s) \coloneqq \mu_h(\cdot \mid s) \in \Delta(\mathcal{A}_i).$$

We shall also employ the expectation operator $\mathbb{E}_{h,k-1}[\cdot]$ (resp. variance operator $\mathsf{Var}_{h,k-1}[\cdot]$) to denote the expectation (resp. variance) conditional on what happens before the beginning of the $k$-th round of data collection for step $h$ (see Section 3.1 about the data collection process).

### B.2 Proof outline

With the above preliminaries in place, we are in a position to present our analysis. Recall that the joint policy $\widehat{\pi}$ computed by Algorithm 1 takes the form of a mixture of product policies

$$\sum_{k=1}^{K} \alpha_k^K \underbrace{\left(\pi_{1,h}^k \times \cdots \times \pi_{m,h}^k\right)}_{=:\, \pi_h^k} \tag{41}$$

at step $h$. Consequently, the value function under policy $\widehat{\pi}$ satisfies the following Bellman equation:

$$V_{i,H+1}^{\widehat{\pi}}(s) \coloneqq 0 \tag{42a}$$

$$V_{i,h}^{\widehat{\pi}}(s) \coloneqq \sum_{k=1}^{K} \sum_{\boldsymbol{a} \in \mathcal{A}} \alpha_k^K \pi_h^k(\boldsymbol{a} \mid s) \Big[ r_{i,h}(s, \boldsymbol{a}) + \big\langle P_h(\cdot \mid s, \boldsymbol{a}), V_{i,h+1}^{\widehat{\pi}} \big\rangle \Big] \tag{42b}$$

for all $(i, s, h) \in [m] \times \mathcal{S} \times [H]$. To establish Theorem 2, we seek to prove the following inequality:

$$V_{i,1}^{\star, \widehat{\pi}_{-i}}(s) - V_{i,1}^{\widehat{\pi}}(s) \le \varepsilon, \qquad 1 \le i \le m, \tag{43}$$

where we remind the reader of the definition of $V_{i,1}^{\star, \widehat{\pi}_{-i}}$ in (9).

Towards this, let us introduce the following best-response policy of the $i$-th player:

$$\widetilde{\pi}_i^\star = \left[\widetilde{\pi}_{i,h}^\star\right]_{h \in [H]} \coloneqq \arg \max_{\pi_i' : \mathcal{S} \times [H] \to \Delta(\mathcal{A}_i)} V_{i,1}^{\pi_i' \times \widehat{\pi}_{-i}}.$$

We make note of the following key decomposition:

$$V_{i,h}^{\star, \widehat{\pi}_{-i}} - V_{i,h}^{\widehat{\pi}} \le \big(V_{i,h}^{\star, \widehat{\pi}_{-i}} - \overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\big) + \big(\overline{V}_{i,h}^{\star, \widehat{\pi}_{-i}} - \overline{V}_{i,h}^{\widehat{\pi}}\big) + \big(\overline{V}_{i,h}^{\widehat{\pi}} - V_{i,h}^{\widehat{\pi}}\big), \tag{44}$$

where we define the following auxiliary value functions:

$$\overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}(s) \coloneqq \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{a_i \sim \widetilde{\pi}_{i,h}^\star(s)} \Big[ r_{i,h}^k(s, a_i) + \big\langle P_{i,h}^k(\cdot \mid s, a_i), \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \big\rangle \Big], \qquad \text{with } \overline{V}_{i,H+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} = 0, \tag{45a}$$

$$\overline{V}_{i,h}^{\star, \widehat{\pi}_{-i}}(s) \coloneqq \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^{K} \alpha_k^K \Big[ r_{i,h}^k(s, a_i) + \big\langle P_{i,h}^k(\cdot \mid s, a_i), \overline{V}_{i,h+1}^{\star, \widehat{\pi}_{-i}} \big\rangle \Big], \qquad \text{with } \overline{V}_{i,H+1}^{\star, \widehat{\pi}_{-i}} = 0, \tag{45b}$$

$$\overline{V}_{i,h}^{\widehat{\pi}}(s) \coloneqq \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{a_i \sim \pi_{i,h}^k(s)} \Big[ r_{i,h}^k(s, a_i) + \big\langle P_{i,h}^k(\cdot \mid s, a_i), \overline{V}_{i,h+1}^{\widehat{\pi}} \big\rangle \Big], \qquad \text{with } \overline{V}_{i,H+1}^{\widehat{\pi}} = 0. \tag{45c}$$

Here, we have used the elementary fact $\overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \le \overline{V}_{i,h}^{\star, \widehat{\pi}_{-i}}$. We shall establish bounds for the above terms in (44), which consists of three steps as outlined below.

**Step 1: showing that $\widehat{V}_{i,h}$ is an entrywise upper bound on $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i}}$.** The following lemma ascertains that the value estimate $\widehat{V}_{i,h}$ of the $i$-th player returned by Algorithm 1 is an optimistic estimate of the auxiliary value $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i}}$ defined in (45b). Evidently, this result cannot happen unless the bonus terms are suitably chosen.

**Lemma 3.** *With probability at least $1-\delta$, it holds that*

$$\widehat{V}_{i,h} \geq \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i}}, \qquad \textit{for all } (i,h) \in [m] \times [H]. \tag{46}$$

The proof of this lemma is postponed to Appendix D.1. Armed with Lemma 3, we can further bound (44) as follows

$$V_{i,h}^{\star,\widehat{\pi}_{-i}} - V_{i,h}^{\widehat{\pi}} \leq \big(V_{i,h}^{\star,\widehat{\pi}_{-i}} - \overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\big) + \big(\widehat{V}_{i,h} - \overline{V}_{i,h}^{\widehat{\pi}}\big) + \big(\overline{V}_{i,h}^{\widehat{\pi}} - V_{i,h}^{\widehat{\pi}}\big). \tag{47}$$

**Step 2: establishing a key recursion.** Recall the definition of $\pi_h^k$ in (41). Let us define the following auxiliary reward vectors $r_{i,h}^{\widehat{\pi}}, r_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}, \overline{r}_{i,h} \in \mathbb{R}^S$ as well as the auxiliary probability transition matrices $P_{i,h}^{\widehat{\pi}}, P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}, \overline{P}_{i,h} \in \mathbb{R}^{S \times S}$ such that: for any $s, s' \in \mathcal{S}$,

$$r_{i,h}^{\widehat{\pi}}(s) := \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_h^k(s)} \big[r_{i,h}(s, \boldsymbol{a})\big], \tag{48a}$$

$$P_{i,h}^{\widehat{\pi}}(s, s') := \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_h^k(s)} \big[P_h(s' \mid s, \boldsymbol{a})\big], \tag{48b}$$

$$r_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}(s) := \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{(a_i, \boldsymbol{a}_{-i}) \sim \widetilde{\pi}_{i,h}^\star(s) \times \pi_{-i,h}^k(s)} \big[r_{i,h}(s, \boldsymbol{a})\big], \tag{48c}$$

$$P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}(s, s') := \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{(a_i, \boldsymbol{a}_{-i}) \sim \widetilde{\pi}_{i,h}^\star(s) \times \pi_{-i,h}^k(s)} \big[P_h(s' \mid s, \boldsymbol{a})\big], \tag{48d}$$

$$\overline{r}_{i,h}(s) := \sum_{k=1}^{K} \alpha_k^K \sum_{a_i \in \mathcal{A}_i} \pi_{i,h}^k(a_i \mid s) r_{i,h}^k(s, a_i), \tag{48e}$$

$$\overline{P}_{i,h}(s, s') := \sum_{k=1}^{K} \alpha_k^K \sum_{a_i \in \mathcal{A}_i} \pi_{i,h}^k(a_i \mid s) P_{i,h}^k(s' \mid s, a_i). \tag{48f}$$

As it turns out, $\overline{V}_{i,h}^{\widehat{\pi}}$ (resp. $\overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}$, $\widehat{V}_{i,h}$) stays reasonably close to the "one-step-look-ahead" expression $r_{i,h}^{\widehat{\pi}} + P_{i,h}^{\widehat{\pi}} \overline{V}_{i,h+1}^{\widehat{\pi}}$ (resp. $r_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} + P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}$, $\overline{r}_{i,h} + \overline{P}_{i,h} \widehat{V}_{i,h+1}$), as revealed by the recursive relations stated in the following lemma; the proof of this lemma is deferred to Appendix D.2.

**Lemma 4.** *There exists some universal constant $c_3 > 0$ such that with probability exceeding $1-\delta$,*

$$\begin{aligned}\Big|\overline{V}_{i,h}^{\widehat{\pi}} - \big(r_{i,h}^{\widehat{\pi}} + P_{i,h}^{\widehat{\pi}} \overline{V}_{i,h+1}^{\widehat{\pi}}\big)\Big| &\leq c_3 \sqrt{\frac{H \log^3\big(\frac{KS\sum_i A_i}{\delta}\big)}{K}} 1 \\ &\quad + c_3 \sqrt{\frac{\log^3\big(\frac{KS\sum_i A_i}{\delta}\big)}{KH}} \Big[P_{i,h}^{\widehat{\pi}}\big(\overline{V}_{i,h+1}^{\widehat{\pi}} \circ \overline{V}_{i,h+1}^{\widehat{\pi}}\big) - \big(P_{i,h}^{\widehat{\pi}} \overline{V}_{i,h+1}^{\widehat{\pi}}\big) \circ \big(P_{i,h}^{\widehat{\pi}} \overline{V}_{i,h+1}^{\widehat{\pi}}\big)\Big],\end{aligned} \tag{49a}$$

$$\begin{aligned}\Big|\overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} - \big(r_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} + P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\big)\Big| &\leq c_3 \sqrt{\frac{H \log^3\big(\frac{KS\sum_i A_i}{\delta}\big)}{K}} 1 \\ &\quad + c_3 \sqrt{\frac{\log^3\big(\frac{KS\sum_i A_i}{\delta}\big)}{KH}} \Big[P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\big(\overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \circ \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\big) - \big(P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\big) \circ \big(P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\big)\Big],\end{aligned} \tag{49b}$$

$$\left|\widehat{V}_{i,h} - \left(\overline{r}_{i,h} + \overline{P}_{i,h}\widehat{V}_{i,h+1}\right)\right| \le c_3\sqrt{\frac{H\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}}1 + c_3\sqrt{\frac{\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{KH}}\left[\overline{P}_{i,h}\left(\widehat{V}_{i,h+1}\circ\widehat{V}_{i,h+1}\right) - \left(\overline{P}_{i,h}\widehat{V}_{i,h+1}\right)\circ\left(\overline{P}_{i,h}\widehat{V}_{i,h+1}\right)\right] \tag{49c}$$

*hold for all* $h \in [H]$.

***Remark*** 2. The right-hand side of each of the bounds in (49) contains a variance-style term (e.g., those terms taking the form of $P_{i,h}(V_{i,h+1}\circ V_{i,h+1}) - (P_{i,h}V_{i,h+1})\circ(P_{i,h}V_{i,h+1})$ for some probability transition matrix $P_{i,h}$ and value vector $V_{i,h+1}$). Such variance-style terms are direct consequences of our Bernstein-style bonus terms, and are crucial in optimizing the horizon dependency.

With the above lemma in place, one can readily show that

$$\begin{aligned}\left|\overline{V}^{\widehat{\pi}}_{i,h} - P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right| &\le r^{\widehat{\pi}}_{i,h} + c_3\sqrt{\frac{H\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}}1 \\ &\quad + \frac{c_3}{H}\sqrt{\frac{H\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}}\left[P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\right] \\ &\le \frac{c_4}{4}1 + \frac{1}{4H}\left[P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\right] =: \zeta_0\end{aligned} \tag{50}$$

for some large enough constant $c_4 > 0$, where the last line holds due to Condition (38), the basic fact $P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) \ge \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)$, and the following fact (for large enough $c_4$)

$$c_3\sqrt{\frac{H\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}}1 + r^{\widehat{\pi}}_{i,h} \le c_3\sqrt{\frac{H\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}}1 + 1 \le \frac{c_4}{4}1.$$

In addition, recalling that $\|\overline{V}^{\widehat{\pi}}_{i,h}\|_\infty, \|\overline{V}^{\widehat{\pi}}_{i,h+1}\|_\infty \le H$ (cf. (20a)) and recognizing that $\zeta_0 \ge 0$ (see (50)), we can demonstrate that

$$\begin{aligned}\left|\overline{V}^{\widehat{\pi}}_{i,h}\circ\overline{V}^{\widehat{\pi}}_{i,h} - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\right| &= \left|\left(\overline{V}^{\widehat{\pi}}_{i,h} + P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(\overline{V}^{\widehat{\pi}}_{i,h} - P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\right| \\ &\le \left(\overline{V}^{\widehat{\pi}}_{i,h} + P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\zeta_0 \le 2H\zeta_0 \\ &= \frac{c_4}{2}H1 + \frac{1}{2}\left[P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\right].\end{aligned} \tag{51}$$

This further leads to

$$\begin{aligned}&P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right) \\ &\quad = P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \overline{V}^{\widehat{\pi}}_{i,h}\circ\overline{V}^{\widehat{\pi}}_{i,h} + \overline{V}^{\widehat{\pi}}_{i,h}\circ\overline{V}^{\widehat{\pi}}_{i,h} - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right) \\ &\quad \le P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \overline{V}^{\widehat{\pi}}_{i,h}\circ\overline{V}^{\widehat{\pi}}_{i,h} + \frac{c_4}{2}H1 + \frac{1}{2}\left[P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\right],\end{aligned}$$

which can be rearranged to yield

$$P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\circ\left(P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right) \le 2\left[P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \overline{V}^{\widehat{\pi}}_{i,h}\circ\overline{V}^{\widehat{\pi}}_{i,h}\right] + c_4H1.$$

Substituting it into (49a) and combining terms give

$$\begin{aligned}\left|\overline{V}^{\widehat{\pi}}_{i,h} - \left(r^{\widehat{\pi}}_{i,h} + P^{\widehat{\pi}}_{i,h}\overline{V}^{\widehat{\pi}}_{i,h+1}\right)\right| &\le c_5\sqrt{\frac{H\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}}1 \\ &\quad + 2c_3\sqrt{\frac{\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{KH}}\left[P^{\widehat{\pi}}_{i,h}\left(\overline{V}^{\widehat{\pi}}_{i,h+1}\circ\overline{V}^{\widehat{\pi}}_{i,h+1}\right) - \overline{V}^{\widehat{\pi}}_{i,h}\circ\overline{V}^{\widehat{\pi}}_{i,h}\right],\end{aligned} \tag{52}$$

where we take $c_5 = c_3 + c_3 c_4$.

An analogous argument (which is omitted here for brevity) also reveals that

$$\left|\overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} - \left(r_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} + P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\right)\right|$$
$$\leq c_5 \sqrt{\frac{H \log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}} 1 + 2c_3 \sqrt{\frac{\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{KH}} \left[P_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\left(\overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \circ \overline{V}_{i,h+1}^{\widetilde{\pi}_i^\star, \widehat{\pi}_{-i}}\right) - \overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \circ \overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}}\right], \tag{53}$$

$$\left|\widehat{V}_{i,h} - \left(\overline{r}_{i,h} + \overline{P}_{i,h}\widehat{V}_{i,h+1}\right)\right|$$
$$\leq c_5 \sqrt{\frac{H \log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}} 1 + 2c_3 \sqrt{\frac{\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{KH}} \left[\overline{P}_{i,h}\left(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\right) - \widehat{V}_{i,h} \circ \widehat{V}_{i,h}\right]. \tag{54}$$

**Step 3: invoking the key recursion to establish the desired bound.** We find it helpful to introduce the following notation (please note the order of the matrix product)

$$\prod_{j:j<h} P_{i,j}^{\widehat{\pi}} := \begin{cases} P_{i,1}^{\widehat{\pi}} \cdots P_{i,h-1}^{\widehat{\pi}}, & \text{if } h > 1, \\ I, & \text{if } h = 1. \end{cases}$$

Armed with this notation, we can invoke the relation (52) recursively and use $\overline{V}_{i,h+1}^{\widehat{\pi}} = V_{i,h+1}^{\widehat{\pi}} = 0$ to obtain

$$\begin{aligned}
\overline{V}_{i,h}^{\widehat{\pi}} - V_{i,h}^{\widehat{\pi}} &\overset{\text{(i)}}{=} r_{i,h}^{\widehat{\pi}} + P_{i,h}^{\widehat{\pi}} \overline{V}_{i,h+1}^{\widehat{\pi}} + \left(\overline{V}_{i,h}^{\widehat{\pi}} - \left(r_{i,h}^{\widehat{\pi}} + P_{i,h}^{\widehat{\pi}} \overline{V}_{i,h+1}^{\widehat{\pi}}\right)\right) - \left(r_{i,h}^{\widehat{\pi}} + P_{i,h}^{\widehat{\pi}} V_{i,h+1}^{\widehat{\pi}}\right) \\
&\leq P_{i,h}^{\widehat{\pi}}\left(\overline{V}_{i,h+1}^{\widehat{\pi}} - V_{i,h+1}^{\widehat{\pi}}\right) + \left|\overline{V}_{i,h}^{\widehat{\pi}} - \left(r_{i,h}^{\widehat{\pi}} + P_{i,h}^{\widehat{\pi}} \overline{V}_{i,h+1}^{\widehat{\pi}}\right)\right| \qquad (55) \\
&\overset{\text{(ii)}}{\leq} c_5 \sqrt{\frac{H \log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}} \left(\sum_{h=1}^{H} \prod_{j:j<h} P_{i,j}^{\widehat{\pi}}\right) 1 \\
&\quad + 2c_3 \sqrt{\frac{\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{KH}} \sum_{h=1}^{H} \prod_{j:j<h} P_{i,j}^{\widehat{\pi}} \left[P_{i,h}^{\widehat{\pi}}\left(\overline{V}_{i,h+1}^{\widehat{\pi}} \circ \overline{V}_{i,h+1}^{\widehat{\pi}}\right) - \overline{V}_{i,h}^{\widehat{\pi}} \circ \overline{V}_{i,h}^{\widehat{\pi}}\right] \\
&\overset{\text{(iii)}}{\leq} c_5 \sqrt{\frac{H \log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}} \left(\sum_{h=1}^{H} \prod_{j:j<h} P_{i,j}^{\widehat{\pi}}\right) 1 = c_5 \sqrt{\frac{H^3 \log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{K}} 1 \leq \frac{\varepsilon}{3} 1. \qquad (56)
\end{aligned}$$

Here, (i) uses the Bellman equation; (ii) applies the bound (52) recursively; (iii) holds since for any transition matrices $\{P_{i,h}\}$ and any sequence $\{V_{i,h}\}$ obeying $V_{i,h+1} = 0$, one can use the telescoping sum to obtain

$$\begin{aligned}
\sum_{h=1}^{H} \prod_{j:j<h} P_{i,j}\left[P_{i,h}\left(V_{i,h+1} \circ V_{i,h+1}\right) - V_{i,h} \circ V_{i,h}\right] &= \sum_{h=1}^{H} \prod_{j:j\leq h} P_{i,j}\left(V_{i,h+1} \circ V_{i,h+1}\right) - \sum_{h=1}^{H} \prod_{j:j<h} P_{i,j}\left(V_{i,h} \circ V_{i,h}\right) \\
&= \prod_{j:j\leq H} P_{i,j}\left(V_{i,h+1} \circ V_{i,h+1}\right) - V_{i,1} \circ V_{i,1} \\
&= -V_{i,1} \circ V_{i,1} \leq 0,
\end{aligned}$$

whereas the last inequality in (56) arises from the assumption (27) when $c_{\mathsf{k}}$ is large enough. Similarly, replacing $\widehat{\pi}_i$ with $\widetilde{\pi}_i^\star$ in the above argument and recalling (53) directly lead to

$$V_{i,h}^{\star, \widehat{\pi}_{-i}} - \overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} = V_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} - \overline{V}_{i,h}^{\widetilde{\pi}_i^\star \times \widehat{\pi}_{-i}} \leq \frac{\varepsilon}{3} 1. \tag{57}$$

In addition, recalling the definition of $\overline{V}_{i,h}^{\widehat{\pi}}$ (cf. (45c)), $\overline{r}_{i,h}$ and $\overline{P}_{i,h}$ (see (48)), we can deduce that

$$\widehat{V}_{i,h} - \overline{V}_{i,h}^{\widehat{\pi}} = \overline{r}_{i,h} + \overline{P}_{i,h}\widehat{V}_{i,h+1} + \left\{\widehat{V}_{i,h} - \left(\overline{r}_{i,h} + \overline{P}_{i,h}\widehat{V}_{i,h+1}\right)\right\} - \overline{r}_{i,h} - \overline{P}_{i,h}\overline{V}_{i,h+1}^{\widehat{\pi}}$$

$$\leq \overline{P}_{i,h}\big(\widehat{V}_{i,h+1} - \overline{V}^{\widehat{\pi}}_{i,h+1}\big) + \Big|\widehat{V}_{i,h} - \big(\overline{r}_{i,h} + \overline{P}_{i,h}\widehat{V}_{i,h+1}\big)\Big|,$$

which resembles (55). Thus, repeating the above argument for (56) and applying (54) recursively, we reach

$$\widehat{V}_{i,h} - \overline{V}^{\widehat{\pi}}_{i,h} \leq \frac{\varepsilon}{3}1. \tag{58}$$

To finish up, combining (56), (57) and (58) with (47), we arrive at

$$V^{\star,\widehat{\pi}_{-i}}_{i,h} - V^{\widehat{\pi}}_{i,h} \leq \big(V^{\star,\widehat{\pi}_{-i}}_{i,h} - \overline{V}^{\widetilde{\pi}^{\star}_i \times \widehat{\pi}_{-i}}_{i,h}\big) + \big(\widehat{V}_{i,h} - \overline{V}^{\widehat{\pi}}_{i,h}\big) + \big(\overline{V}^{\widehat{\pi}}_{i,h} - V^{\widehat{\pi}}_{i,h}\big) \leq \varepsilon 1.$$

This establishes the first inequality in (43), while the second inequality in (43) can be validated via the same argument. We have thus completed the proof of Theorem 2.

# C Proof of Theorem 3

This section is devoted to presenting the proof of Theorem 3. Before embarking on the analysis, let us introduce a convenient auxiliary iterate

$$\pi^{-}_{k+1} = \arg\min_{\pi\in\Delta(\mathcal{A})} \left\{\langle \pi, L_k\rangle + \frac{1}{\widehat{\eta}_k}F(\pi)\right\}, \tag{59}$$

or equivalently,

$$\pi^{-}_{k+1}(a) = \frac{\exp\big(-\widehat{\eta}_k L_k(a)\big)}{\sum_{a'\in\mathcal{A}}\exp\big(-\widehat{\eta}_k L_k(a')\big)} \qquad \text{for all } a\in\mathcal{A}, \tag{60}$$

which differs from (34) only in the learning rates being used (namely, $\pi_{k+1}$ uses $\eta_{k+1}$ while $\pi^{-}_{k+1}$ adopts $\widehat{\eta}_k$).

## C.1 Main steps of the proof

The key steps of the proof lie in justifying the following two claims:

$$R_n \leq \sum_{k=1}^{n} \alpha_k^n \big\langle \pi_k - \pi^{-}_{k+1}, \ell_k\big\rangle + \frac{\log A}{\eta_{n+1}}; \tag{61}$$

and for all $a\in\mathcal{A}$ and all $k\geq 1$,

$$\pi^{-}_{k+1}(a) \geq \begin{cases} \big[1-\widehat{\eta}_k\alpha_k\ell_k(a)\big]\pi_k(a), & \text{if } \widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > \frac{1}{3}, \\ \Big\{1-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big) - 2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}\big(\ell_k\big)\Big\}\pi_k(a), & \text{if } \widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \leq \frac{1}{3}, \end{cases} \tag{62}$$

where for any vector $\ell\in\mathbb{R}^A$ we define

$$\mathbb{E}_{\pi_k}[\ell] \coloneqq \sum_{a\in\mathcal{A}}\pi_k(a)\ell(a).$$

In words, the first claim (61) allows us to replace the action that appears best in hindsight (cf. (32)) by the time-varying predictions $\{\pi^{-}_{k+1}\}$ without incurring much cost, whereas the second claim (62) controls the proximity of $\pi^{-}_{k+1}$ and $\pi_k$ in each round. Let us assume the validity of these two claims for the moment, and return to prove them shortly.

In view of the upper bound (61), we are in need of controlling $\big\langle \pi_k - \pi^{-}_{k+1}, \ell_k\big\rangle$. We divide into two cases.

- For any $k$ obeying $\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > 1/3$, invoke (62) and the non-negativity of $\ell_k$ to reach

$$\big\langle \pi_k - \pi^{-}_{k+1}, \ell_k\big\rangle \leq \sum_{a\in\mathcal{A}}\widehat{\eta}_k\alpha_k\pi_k(a)\big[\ell_k(a)\big]^2 \leq \sum_{a\in\mathcal{A}}\widehat{\eta}_k\alpha_k\pi_k(a)\big\|\ell_k\big\|_\infty^2 = \widehat{\eta}_k\alpha_k\big\|\ell_k\big\|_\infty^2. \tag{63}$$

- In contrast, if $\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \le 1/3$, then it follows from (62) that

$$\begin{aligned}
\langle \pi_k - \pi_{k+1}^-, \ell_k\rangle &\le \sum_{a\in\mathcal{A}} \left\{ \widehat{\eta}_k\alpha_k\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big) + 2\widehat{\eta}_k^2\alpha_k^2 \mathsf{Var}_{\pi_k}(\ell_k)\right\}\pi_k(a)\ell_k(a) \\
&= \widehat{\eta}_k\alpha_k \sum_{a\in\mathcal{A}} \pi_k(a)\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big)\mathbb{E}_{\pi_k}[\ell_k] + \widehat{\eta}_k\alpha_k \sum_{a\in\mathcal{A}} \pi_k(a)\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big)^2 \\
&\quad + 2\widehat{\eta}_k^2\alpha_k^2 \mathsf{Var}_{\pi_k}(\ell_k) \sum_{a\in\mathcal{A}} \pi_k(a)\ell_k(a) \\
&= \widehat{\eta}_k\alpha_k \sum_{a\in\mathcal{A}} \pi_k(a)\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big)^2 + 2\widehat{\eta}_k^2\alpha_k^2 \mathsf{Var}_{\pi_k}(\ell_k) \sum_{a\in\mathcal{A}} \pi_k(a)\ell_k(a) \\
&\le \widehat{\eta}_k\alpha_k \mathsf{Var}_{\pi_k}(\ell_k) + 2\widehat{\eta}_k^2\alpha_k^2 \mathsf{Var}_{\pi_k}(\ell_k)\|\ell_k\|_\infty, \qquad (64)
\end{aligned}$$

  where we invoke the elementary facts that $\sum_a \pi_k(a)\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big) = 0$ and $\sum_a \pi_k(a)\ell_k(a) \le \|\ell_k\|_\infty$.

Putting the above two cases together yields

$$\begin{aligned}
&\sum_{k=1}^n \alpha_k^n \langle \pi_k - \pi_{k+1}^-, \ell_k\rangle \\
&\quad\le \sum_{k=1}^n \alpha_k^n \widehat{\eta}_k\alpha_k \|\ell_k\|_\infty^2 \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > \frac{1}{3}\right) + \sum_{k=1}^n \alpha_k^n \widehat{\eta}_k\alpha_k \mathsf{Var}_{\pi_k}(\ell_k)\, \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \le \frac{1}{3}\right) \\
&\qquad + 2\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k^2\alpha_k^2 \mathsf{Var}_{\pi_k}(\ell_k)\|\ell_k\|_\infty\, \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \le \frac{1}{3}\right) \\
&\quad\le \frac{5}{3}\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k\alpha_k \mathsf{Var}_{\pi_k}(\ell_k) + 3\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k^2\alpha_k^2 \|\ell_k\|_\infty^3\, \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > \frac{1}{3}\right), \qquad (65)
\end{aligned}$$

where the last inequality holds true since

$$\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k\alpha_k \|\ell_k\|_\infty^2 \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > \frac{1}{3}\right) \le 3\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k^2\alpha_k^2 \|\ell_k\|_\infty^3 \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > \frac{1}{3}\right),$$

$$\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k^2\alpha_k^2 \|\ell_k\|_\infty \mathsf{Var}_{\pi_k}(\ell_k)\, \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \le \frac{1}{3}\right) \le \frac{1}{3}\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k\alpha_k \mathsf{Var}_{\pi_k}(\ell_k).$$

Substituting (65) into (61), we can readily arrive at

$$R_n \le \frac{5}{3}\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k\alpha_k \mathsf{Var}_{\pi_k}(\ell_k) + \frac{\log A}{\eta_{n+1}} + 3\sum_{k=1}^n \alpha_k^n \widehat{\eta}_k^2\alpha_k^2 \|\ell_k\|_\infty^3\, \mathbb{1}\left(\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > \frac{1}{3}\right).$$

It thus remains to establish the claims (61) and (62), which we shall accomplish next.

### C.2 Proof of claim (61)

We claim that it suffices to prove that

$$\begin{aligned}
&\alpha_1^n\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^n}{\eta_2\alpha_1}F(\pi_2) + \sum_{k=2}^n \left\{ \alpha_k^n\langle \pi_{k+1}^-, \ell_k\rangle + \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1})\right\} \\
&\qquad\le \min_{\pi\in\Delta(\mathcal{A})} \left\{ \left\langle \pi, \sum_{k=1}^n \alpha_k^n \ell_k\right\rangle + \frac{1}{\eta_{n+1}}F(\pi)\right\}. \qquad (66)
\end{aligned}$$

In fact, suppose that this inequality (66) is valid, then one can easily obtain

$$\alpha_1^n\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^n}{\eta_2\alpha_1}F(\pi_2) + \sum_{k=2}^n \left\{ \alpha_k^n\langle \pi_{k+1}^-, \ell_k\rangle + \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1})\right\}$$

$$\leq \min_{\pi\in\Delta(\mathcal{A})}\left\{\left\langle \pi, \sum_{k=1}^n \alpha_k^n \ell_k\right\rangle + \frac{1}{\eta_{n+1}}F(\pi)\right\} \leq \min_{\pi\in\{e_a \,|\, a\in\mathcal{A}\}}\left\{\left\langle \pi, \sum_{k=1}^n \alpha_k^n \ell_k\right\rangle + \frac{1}{\eta_{n+1}}F(\pi)\right\}$$
$$= \min_{\pi\in\{e_a \,|\, a\in\mathcal{A}\}}\left\langle \pi, \sum_{k=1}^n \alpha_k^n \ell_k\right\rangle = \min_{a\in\mathcal{A}}\sum_{k=1}^n \alpha_k^n \ell_k(a)$$

with $e_a$ the $a$-th standard basis vector in $\mathbb{R}^A$, where the last line holds true since the negative entropy obeys $F(e_a) = 0$ for any $a \in \mathcal{A}$. In turn, this implies that

$$\begin{aligned} R_n &= \sum_{k=1}^n \alpha_k^n \langle \pi_k, \ell_k\rangle - \min_{a\in\mathcal{A}}\sum_{k=1}^n \alpha_k^n \ell_k(a) \\ &\leq \sum_{k=1}^n \alpha_k^n \langle \pi_k - \pi_{k+1}^-, \ell_k\rangle - \sum_{k=2}^n \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1}) + \frac{\alpha_1^n}{\eta_2\alpha_1}\log A, \end{aligned} \tag{67}$$

where the last inequality invokes the elementary fact $-F(\pi) \leq \log A$ for any $\pi \in \Delta(\mathcal{A})$. Additionally, under the assumptions that $\eta_{k+1}(1-\alpha_k) \leq \eta_k$ ($k \geq 1$), we can use the definition (5) to obtain

$$\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} = \frac{\prod_{j=k+1}^n (1-\alpha_j)}{\eta_{k+1}} \geq \frac{\prod_{j=k}^n (1-\alpha_j)}{\eta_k} = \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}},$$

for any $k \geq 2$, which together with the basic fact $0 \leq -F(\pi) \leq \log A$ yields

$$\begin{aligned} -\sum_{k=2}^n \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1}) + \frac{\alpha_1^n}{\eta_2\alpha_1}\log A &\leq \sum_{k=2}^n \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]\log A + \frac{\alpha_1^n}{\eta_2\alpha_1}\log A \\ &= \frac{\alpha_n^n}{\eta_{n+1}\alpha_n}\log A = \frac{\log A}{\eta_{n+1}}. \end{aligned} \tag{68}$$

Substitution into (67) leads to

$$R_n \leq \sum_{k=1}^n \alpha_k^n \langle \pi_k - \pi_{k+1}^-, \ell_k\rangle + \frac{\log A}{\eta_{n+1}} \tag{69}$$

as advertised. As a consequence, everything boils down to establishing (66).

Towards this end, we would like to proceed with an induction argument, with the induction hypothesis w.r.t. $n$ given by (66). Firstly, the base case with $n = 1$ simplifies to

$$\alpha_1^1\langle \pi_2^-, \ell_1\rangle + \frac{1}{\eta_2}F(\pi_2) \leq \min_{\pi\in\Delta(\mathcal{A})}\left\{\langle \pi, \alpha_1^1\ell_1\rangle + \frac{1}{\eta_2}F(\pi)\right\}$$

given that $\alpha_1 = \alpha_1^1$; this inequality clearly holds since, according to (33) and (59),

$$\pi_2^- = \pi_2 = \arg\min_{\pi\in\Delta(\mathcal{A})}\left\{\langle \pi, L_1\rangle + \frac{1}{\eta_2}F(\pi)\right\} = \arg\min_{\pi\in\Delta(\mathcal{A})}\left\{\langle \pi, \alpha_1\ell_1\rangle + \frac{1}{\eta_2}F(\pi)\right\}.$$

Secondly, suppose that (66) holds w.r.t. $n$, and we intend to justify it w.r.t. $n+1$. To do so, we observe that

$$\begin{aligned} &\alpha_1^{n+1}\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^{n+1}}{\eta_2\alpha_1}F(\pi_2) + \sum_{k=2}^n \left\{\alpha_k^{n+1}\langle \pi_{k+1}^-, \ell_k\rangle + \Big(\frac{\alpha_k^{n+1}}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^{n+1}}{\eta_k\alpha_{k-1}}\Big)F(\pi_{k+1})\right\} + \alpha_{n+1}\langle \pi_{n+2}^-, \ell_{n+1}\rangle \\ &\overset{\text{(i)}}{=} (1-\alpha_{n+1})\left\{\alpha_1^n\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^n}{\eta_2\alpha_1}F(\pi_2) + \sum_{k=2}^n \left\{\alpha_k^n\langle \pi_{k+1}^-, \ell_k\rangle + \Big(\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big)F(\pi_{k+1})\right\}\right\} \\ &\qquad + \alpha_{n+1}\langle \pi_{n+2}^-, \ell_{n+1}\rangle \\ &\overset{\text{(ii)}}{\leq} (1-\alpha_{n+1})\left\{\left\langle \pi_{n+2}^-, \sum_{k=1}^n \alpha_k^n\ell_k\right\rangle + \frac{1}{\eta_{n+1}}F(\pi_{n+2}^-)\right\} + \alpha_{n+1}\langle \pi_{n+2}^-, \ell_{n+1}\rangle \\ &\overset{\text{(iii)}}{=} \left\langle \pi_{n+2}^-, \sum_{k=1}^{n+1} \alpha_k^{n+1}\ell_k\right\rangle + \frac{1-\alpha_{n+1}}{\eta_{n+1}}F(\pi_{n+2}^-) = \min_{\pi\in\Delta(\mathcal{A})}\left\{\left\langle \pi, \sum_{k=1}^{n+1} \alpha_k^{n+1}\ell_k\right\rangle + \frac{1}{\widehat{\eta}_{n+1}}F(\pi)\right\}. \end{aligned} \tag{70}$$

Here, (i) and (iii) invoke the fact $\alpha_k^{n+1} = (1-\alpha_{n+1})\alpha_k^n$ and $\alpha_{n+1}^{n+1} = \alpha_{n+1}$ (according to (5)), (ii) relies on the induction hypothesis (66) w.r.t. $n$. To finish up, invoke (70) and the definition (5) to arrive at

$$\begin{aligned}
&\alpha_1^{n+1}\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^{n+1}}{\eta_2\alpha_1}F(\pi_2) + \sum_{k=2}^{n+1}\left\{\alpha_k^{n+1}\langle \pi_{k+1}^-, \ell_k\rangle + \Big[\frac{\alpha_k^{n+1}}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^{n+1}}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1})\right\} \\
&= \left\{\alpha_1^{n+1}\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^{n+1}}{\eta_2\alpha_1}F(\pi_2) + \sum_{k=2}^{n}\left\{\alpha_k^{n+1}\langle \pi_{k+1}^-, \ell_k\rangle + \Big[\frac{\alpha_k^{n+1}}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^{n+1}}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1})\right\} + \alpha_{n+1}\langle \pi_{n+2}^-, \ell_{n+1}\rangle\right\} \\
&\qquad + \Big[\frac{1}{\eta_{n+2}} - \frac{1-\alpha_{n+1}}{\eta_{n+1}}\Big]F(\pi_{n+2}) \\
&\le \left\{\Big\langle \pi_{n+2}, \sum_{k=1}^{n+1}\alpha_k^{n+1}\ell_k\Big\rangle + \frac{1-\alpha_{n+1}}{\eta_{n+1}}F(\pi_{n+2})\right\} + \Big[\frac{1}{\eta_{n+2}} - \frac{1-\alpha_{n+1}}{\eta_{n+1}}\Big]F(\pi_{n+2}) \\
&= \Big\langle \pi_{n+2}, \sum_{k=1}^{n+1}\alpha_k^{n+1}\ell_k\Big\rangle + \frac{1}{\eta_{n+2}}F(\pi_{n+2}) = \min_{\pi\in\Delta(\mathcal{A})}\left\{\Big\langle \pi, \sum_{k=1}^{n+1}\alpha_k^{n+1}\ell_k\Big\rangle + \frac{1}{\eta_{n+2}}F(\pi)\right\},
\end{aligned}$$

where the inequality above makes use of (70), and the last identity comes from (33). This justifies the induction hypothesis w.r.t. $n+1$. Applying the induction argument in turn establishes (66) for all $n$, thereby concluding the proof.

### C.3 Proof of claim (62)

We first make the observation that

$$\begin{aligned}
\sum_a \exp\big(-\widehat{\eta}_k L_k(a)\big) &= \sum_a \exp\big(-\eta_k L_{k-1}(a)\big)\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big) \\
&= \sum_a \left\{\pi_k(a)\sum_{a'}\exp\big(-\eta_k L_{k-1}(a')\big)\right\}\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big) \\
&= \sum_{a'}\exp\big(-\eta_k L_{k-1}(a')\big)\sum_a\big\{\pi_k(a)\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big)\big\},
\end{aligned}$$

where the second equality follows from (34). This in turn allows us to demonstrate that

$$\begin{aligned}
\pi_{k+1}^-(a) &= \frac{\exp\big(-\widehat{\eta}_k L_k(a)\big)}{\sum_{a'}\exp\big(-\widehat{\eta}_k L_k(a')\big)} = \frac{\exp\big(-\eta_k L_{k-1}(a)\big)}{\sum_{a'}\exp\big(-\eta_k L_{k-1}(a')\big)}\cdot\frac{\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a')\big)} \\
&= \pi_k(a)\frac{\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a')\big)} \ge \big[1-\widehat{\eta}_k\alpha_k\ell_k(a)\big]\pi_k(a),
\end{aligned}$$

where the last inequality holds since $\exp(-x) \ge 1-x$ and $\sum_a \pi_k(a)\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big) \le \sum_a \pi_k(a) = 1$.

Next, suppose that $\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \le 1/3$. In this case, it is self-evident that $\widehat{\eta}_k\alpha_k|\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]| \le 2/3$ for all $a\in\mathcal{A}$. Recalling that $\mathbb{E}_{\pi_k}[\ell_k] = \sum_a \pi_k(a)\ell_k(a)$, one can derive

$$\begin{aligned}
\pi_{k+1}^-(a) &= \pi_k(a)\frac{\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a')\big)} = \frac{\exp\big(-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\big(\ell_k(a')-\mathbb{E}_{\pi_k}[\ell_k]\big)\big)}\pi_k(a) \\
&\ge \frac{1-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\big(\ell_k(a')-\mathbb{E}_{\pi_k}[\ell_k]\big)\big)}\pi_k(a) \\
&\ge \frac{1-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)}{1+\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}(\ell_k)}\pi_k(a); \qquad (71)
\end{aligned}$$

here, the first inequality arises since $\exp(-x) \ge 1-x$, while the second inequality can be shown via the elementary inequality $\exp(-x) \le 1-x+x^2$ for any $x \ge -1.5$ and therefore

$$\sum_a \pi_k(a)\exp\Big(-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)\Big)$$

$$
\begin{aligned}
&\le \sum_a \pi_k(a)\Big\{1 - \widehat{\eta}_k\alpha_k\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big) + \widehat{\eta}_k^2\alpha_k^2\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big)^2\Big\} \\
&= \sum_a \pi_k(a)\Big\{1 + \widehat{\eta}_k^2\alpha_k^2\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big)^2\Big\} \\
&= 1 + \widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}(\ell_k).
\end{aligned}
$$

Applying the elementary inequality $\frac{1-a}{1+b} \ge (1-a)(1-b) = 1 - a - b + ab \ge 1 - a - 2b$ for any $a \in [-1, 1]$ and $b > 0$, we can continue to lower bound (71) as follows

$$
(71) \ge \Big\{1 - \widehat{\eta}_k\alpha_k\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big) - 2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}\big(\ell_k\big)\Big\}\pi_k(a),
$$

thereby completing the proof.

# D Proofs of auxiliary lemmas and details

## D.1 Proof of Lemma 3

This section aims to prove Lemma 3, which establishes the inequality $\widehat{V}_{i,h} \ge \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i}}$. In what follows, we shall proceed with an induction argument. The base case with step $H+1$ is trivially true, given that

$$
\widehat{V}_{i,H+1} = \overline{V}_{i,H+1}^{\star,\widehat{\pi}_{-i}} = 0
$$

holds for any joint policy. Next, let us assume that the claim (46) is valid for step $h+1$, namely,

$$
\widehat{V}_{i,h+1} \ge \overline{V}_{i,h+1}^{\star,\widehat{\pi}_{-i}}, \tag{72}
$$

and attempt to justify the validity of this result when $h+1$ is replaced with $h$.

This step is mainly accomplished by applying our refined theory (cf. Theorem 3) for FTRL (see (19)). More precisely, we claim that

$$
\begin{aligned}
\max_{a_i} Q_{i,h}^K(s, a_i) \le &\sum_{k=1}^K \alpha_k^K \Big\langle \pi_{i,h}^k(s),\, q_{i,h}^k(s,\cdot)\Big\rangle \\
&+ 10\sqrt{\frac{c_\alpha \log^3(KA_i)}{KH}} \sum_{k=1}^K \alpha_k^K \mathsf{Var}_{\pi_{i,h}^k(s)}\Big(q_{i,h}^k(s,\cdot)\Big) + 2\sqrt{\frac{c_\alpha H \log^3(KA_i)}{K}}
\end{aligned} \tag{73}
$$

for any $s \in \mathcal{S}$, whose proof is deferred to Appendix D.1.1. Recall the construction (20a) of $\widehat{V}_{i,h}$. If $\widehat{V}_{i,h} = H - h + 1$, then the claimed result $\widehat{V}_{i,h} \ge \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i}}$ holds trivially. It thus suffices to focus on the case where

$$
\widehat{V}_{i,h}(s) = \sum_{k=1}^K \alpha_k^K \Big\langle \pi_{i,h}^k(s),\, q_{i,h}^k(s,\cdot)\Big\rangle + \beta_{i,h}(s). \tag{74}
$$

In this case, recalling the definition of $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i}}(s)$ in (45b) gives

$$
\begin{aligned}
\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i}}(s) &= \max_{a_i} \sum_{k=1}^K \alpha_k^K \Big[r_{i,h}^k(s,a_i) + \big\langle P_{i,h}^k(\cdot \mid s, a_i), \overline{V}_{i,h+1}^{\star,\widehat{\pi}_{-i}}\big\rangle\Big] \\
&\le \max_{a_i} \sum_{k=1}^K \alpha_k^K \Big[r_{i,h}^k(s,a_i) + \big\langle P_{i,h}^k(\cdot \mid s, a_i), \widehat{V}_{i,h+1}\big\rangle\Big] = \max_{a_i} Q_{i,h}^K(s,a_i) \\
&\le \sum_{k=1}^K \alpha_k^K \Big\langle \pi_{i,h}^k(s),\, q_{i,h}^k(s,\cdot)\Big\rangle + 10\sqrt{\frac{c_\alpha \log^3(KA_i)}{KH}} \sum_{k=1}^K \alpha_k^K \mathsf{Var}_{\pi_{i,h}^k(s)}\Big(q_{i,h}^k(s,\cdot)\Big) + 2\sqrt{\frac{c_\alpha H \log^3(KA_i)}{K}}
\end{aligned}
$$

$$\leq \sum_{k=1}^{K} \alpha_k^K \left\langle \pi_{i,h}^k(s),\, q_{i,h}^k(s,\cdot) \right\rangle + \beta_{i,h}(s) = \widehat{V}_{i,h}(s)$$

simultaneously for all $(s,h) \in \mathcal{S} \times [H]$. Here, the second line follows from the induction hypothesis (72) and the definition (17) of $Q_{i,h}^K$, the third line invokes the claim (73), whereas the last line comes from our choice (24) of $\beta_{i,h}$ (provided $c_{\mathsf{b}}$ is large enough) and (74). This concludes the proof, as long as (73) can be justified.

### D.1.1 Proof of claim (73)

Consider any state $s \in \mathcal{S}$. By virtue of the identity $Q_{i,h}^k = \sum_{j=1}^k \alpha_j^k q_{i,h}^j$ (see (17)), the policy update rule (18) (or (19)) for $\pi_{i,h}^k(s)$ can essentially be viewed as the FTRL algorithm applied to the loss vectors

$$\ell_k = -q_{i,h}^k(s,\cdot), \qquad k \geq 1.$$

Moreover, recalling the definition (23) of $\eta_{k+1}$ and the definition (22) of $\alpha_k$ (with $c_\alpha \geq 24$), we have

$$\left(\frac{\eta_k}{\eta_{k+1}}\right)^2 = \frac{\alpha_k}{\alpha_{k-1}} = \frac{k-2+c_\alpha \log K}{k-1+c_\alpha \log K} \geq \frac{k-1}{k-1+c_\alpha \log K} = 1-\alpha_k > (1-\alpha_k)^2. \tag{75}$$

This property (75) permits us to invoke Theorem 3 to obtain

$$\begin{aligned}
&\max_{a_i \in \mathcal{A}_i} Q_{i,h}^K(s,a_i) - \sum_{k=1}^K \alpha_k^K \left\langle \pi_{i,h}^k(s),\, q_{i,h}^k(s,\cdot) \right\rangle = \max_{a_i \in \mathcal{A}_i} \left\{ \sum_{k=1}^K \alpha_k^K \left\langle \pi_{i,h}^k(s),\, \ell_k \right\rangle - \sum_{k=1}^K \alpha_k^K \ell_k(a_i) \right\} \\
&\leq \frac{5}{3} \sum_{k=2}^K \alpha_k^K \frac{\eta_k \alpha_k}{1-\alpha_k} \mathsf{Var}_{\pi_{i,h}^k(s)}\left(q_{i,h}^k(s,\cdot)\right) + \frac{\log A_i}{\eta_{K+1}} + \xi_{i,h} \\
&\overset{\text{(i)}}{\leq} \frac{5}{3} \sum_{k=2}^{K/2} \frac{(2c_\alpha)^{1.5} \log^2 K}{\sqrt{kH}} \alpha_k^K \mathsf{Var}_{\pi_{i,h}^k(s)}\left(q_{i,h}^k(s,\cdot)\right) \\
&\qquad + \frac{20}{3} \sum_{k=K/2+1}^K \alpha_k^K \sqrt{\frac{c_\alpha \log^2 K}{KH}} \mathsf{Var}_{\pi_{i,h}^k(s)}\left(q_{i,h}^k(s,\cdot)\right) + \frac{\log A_i}{\eta_{K+1}} + \xi_{i,h},
\end{aligned} \tag{76}$$

where $\xi_{i,h}$ is defined as

$$\xi_{i,h} := \frac{5}{3}\alpha_1^K \eta_2 \left\|q_{i,h}^1\right\|_\infty^2 + \left\{ 3\sum_{k=2}^K \alpha_k^K \frac{\eta_k^2 \alpha_k^2}{(1-\alpha_k)^2} \left\|q_{i,h}^k\right\|_\infty^3 \mathbb{1}\left( \frac{\eta_k \alpha_k}{1-\alpha_k} \left\|q_{i,h}^k\right\|_\infty > \frac{1}{3} \right) \right\} + 3\alpha_1^K \eta_2^2 \left\|q_{i,h}^1\right\|_\infty^3. \tag{77}$$

Here, to see why (i) holds, we make use of the facts that

$$1-\alpha_k = 1 - \frac{c_\alpha \log K}{k-1+c_\alpha \log K} \geq \begin{cases} 1 - \frac{c_\alpha \log K}{1+c_\alpha \log K} = \frac{1}{1+c_\alpha \log K} \geq \frac{1}{2c_\alpha \log K}, & \text{if } k \geq 2, \\ 1 - \frac{c_\alpha \log K}{K/2+c_\alpha \log K} = \frac{K}{K+2c_\alpha \log K} \geq \frac{1}{2}, & \text{if } k \geq K/2+1, \end{cases} \tag{78a}$$

$$\eta_k \alpha_k = \sqrt{\frac{\log K}{\alpha_{k-1} H}} \cdot \alpha_k \leq \sqrt{\frac{\log K}{\alpha_k H}} \cdot \alpha_k = \sqrt{\frac{\alpha_k \log K}{H}} \leq \sqrt{\frac{2c_\alpha \log^2 K}{kH}}, \tag{78b}$$

where the first line makes use of (38) for large enough $c_{\mathsf{k}}$, and the second line relies on (39a) in Lemma 1.

To proceed, let us control the terms in (76) separately.

- We start with the first term in (76). The elementary bound $\left\|q_{i,h}^k\right\|_\infty \leq H$ in Lemma 2 taken together with (39b) in Lemma 1 helps us derive

$$\sum_{k=2}^{K/2} \frac{\alpha_k^K \log^2 K}{\sqrt{kH}} \mathsf{Var}_{\pi_{i,h}^k(s)}\left(q_{i,h}^k(s,\cdot)\right) \leq \sum_{k=2}^{K/2} \frac{\log^2 K}{K^6 \sqrt{kH}} \mathsf{Var}_{\pi_{i,h}^k(s)}\left(q_{i,h}^k(s,\cdot)\right)$$

$$\leq \sum_{k=2}^{K/2} \frac{\log^2 K}{K^6\sqrt{kH}} \left\|q_{i,h}^k(s,\cdot)\right\|_\infty^2 \leq \frac{H^{3/2}\log^2 K}{K^6} \sum_{k=2}^{K/2} \frac{1}{\sqrt{k}}$$
$$\leq \frac{2H^{3/2}\log^2 K}{K^6} \cdot \sqrt{K/2} \leq \frac{2H^{3/2}\log^2 K}{K^5}. \tag{79}$$

- Turning to the third term in (76), we recall the definition of $\eta_{K+1}$ (cf. (23)) to obtain

$$\frac{\log A_i}{\eta_{K+1}} = \log A_i \sqrt{\frac{\alpha_K H}{\log K}} \leq \sqrt{\frac{2c_\alpha H \log^2 A_i}{K}}, \tag{80}$$

  where the inequality comes from Lemma 1.

- Finally, we move on to the last term in (76). For any $k \geq 2$, combine Lemma 2 with (78) to obtain

$$\frac{\eta_k \alpha_k}{1-\alpha_k} \left\|q_{i,h}^k\right\|_\infty \leq \frac{\sqrt{\frac{2c_\alpha \log^2 K}{kH}}}{\frac{1}{2c_\alpha \log K}} \cdot H = \sqrt{\frac{8c_\alpha^3 H \log^4 K}{k}}. \tag{81}$$

  Clearly, the right-hand side of (81) is upper bounded by $1/3$ for all $k$ obeying $k \geq c_9 H \log^4 \frac{K}{\delta}$ for some large enough constant $c_9 > 0$ (see also (38)). Consequently, one can derive

$$\begin{aligned}
\xi_{i,h} &= \frac{5}{3}\alpha_1^K \eta_2 \left\|q_{i,h}^1\right\|_\infty^2 + \left\{3\sum_{k=2}^K \alpha_k^K \frac{\eta_k^2 \alpha_k^2}{(1-\alpha_k)^2} \left\|q_{i,h}^k\right\|_\infty^3 \mathbb{1}\left(\frac{\eta_k\alpha_k}{1-\alpha_k}\left\|q_{i,h}^k\right\|_\infty > \frac{1}{3}\right)\right\} + 3\alpha_1^K \eta_2^2 \left\|q_{i,h}^1\right\|_\infty^3 \\
&\leq \frac{5}{3K^6}\sqrt{\frac{\log K}{H}} \left\|q_{i,h}^1\right\|_\infty^2 + \frac{(2c_\alpha \log K)^2}{K^6}\left\{3\sum_{k=2}^{c_9 H \log^4 \frac{K}{\delta}} \eta_k^2 \alpha_k^2 \left\|q_{i,h}^k\right\|_\infty^3\right\} + \frac{3}{K^6}\frac{\log K}{H}\left\|q_{i,h}^1\right\|_\infty^3 \\
&\leq \frac{24c_\alpha^3 \log^4 K}{K^6 H}\left\{\sum_{k=1}^K \frac{1}{k} H^3\right\} \\
&\leq \frac{24c_\alpha^3 H^2 \log^5 K}{K^6} \leq \frac{1}{K^4},
\end{aligned} \tag{82}$$

  where the second line comes from (78) and the fact that $K/2 > c_9 H \log^4 \frac{K}{\delta}$ (as a consequence of (38)), and the third line holds due to Lemma 2.

Putting the preceding bounds together and substituting them into (76), we arrive at

$$\begin{aligned}
&\max_{a_i} Q_{i,h}^K(s,a_i) - \sum_{k=1}^K \alpha_k^K \left\langle \pi_{i,h}^k(s),\, q_{i,h}^k(s,\cdot)\right\rangle \\
&\leq \frac{5(2c_\alpha)^{1.5}}{3} \cdot \frac{2H^{3/2}\log^2 K}{K^5} + \frac{20}{3}\sqrt{\frac{c_\alpha \log^2 K}{KH}} \sum_{k=K/2+1}^K \alpha_k^K \mathsf{Var}_{\pi_{i,h}^k(s)}\left(q_{i,h}^k(s,\cdot)\right) + \sqrt{\frac{2c_\alpha H \log^2 A_i}{K}} + \frac{1}{K^4} \\
&\leq 10\sqrt{\frac{c_\alpha \log^3(KA_i)}{KH}} \sum_{k=1}^K \alpha_k^K \mathsf{Var}_{\pi_{i,h}^k(s)}\left(q_{i,h}^k(s,\cdot)\right) + 2\sqrt{\frac{c_\alpha H \log^3(KA_i)}{K}},
\end{aligned} \tag{83}$$

where the last line is valid under Condition (38). This completes the proof of Claim (73).

### D.2 Proof of Lemma 4

In this section, we present the proof of Lemma 4. To begin with, we introduce the auxiliary quantities

$$\widetilde{q}_{i,h}^k(s,a_i) := r_{i,h}^k(s,a_i) + P_{i,h}^k(\cdot \mid s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi}}, \qquad \forall (s,a_i) \in \mathcal{S}\times\mathcal{A}_i.$$

It is also helpful to introduce an auxiliary random action $a_{k,s} \in \mathcal{A}_i$ generated in a way that

$$a_{k,s} \sim \pi_{i,h}^k(s),$$

which is independent from $\widetilde{q}_{i,h}^k$ conditional on $\pi_{i,h}^k$. This allows us to define another set of random variables

$$\widehat{q}_{i,h}^k(s) \coloneqq \widetilde{q}_{i,h}^k\big(s, a_{k,s}\big), \qquad \forall s \in \mathcal{S}, \tag{84}$$

which plays a central role in our analysis. It is readily seen from the facts $\overline{V}_{i,h+1}(s) \le H - h$ (cf. (20a)) and $r_{i,h}^k(s, a_i) \in [0, 1]$ that

$$0 \le \widehat{q}_{i,h}^k(s), \widetilde{q}_{i,h}^k(s, a_i) \le H - h + 1, \qquad \forall (s, a_i, h, k) \in \mathcal{S} \times \mathcal{A}_i \times [H] \times [K]. \tag{85}$$

Letting $e(i) \in \mathbb{R}^{A_i}$ denote the $i$-th standard basis vector, we learn from the law of total variance that

$$\begin{aligned}
\mathsf{Var}_{h,k-1}\Big(\widehat{q}_{i,h}^k(s)\Big) &= \mathsf{Var}_{h,k-1}\Big(\big\langle e(a_{k,s}),\, \widetilde{q}_{i,h}^k(s,\cdot)\big\rangle\Big) \\
&\ge \mathsf{Var}_{h,k-1}\Big(\mathbb{E}_{h,k-1}\Big[\big\langle e(a_{k,s}),\, \widetilde{q}_{i,h}^k(s,\cdot)\big\rangle \mid \widetilde{q}_{i,h}^k\Big]\Big) \\
&= \mathsf{Var}_{h,k-1}\Big(\big\langle \pi_{i,h}^k(s),\, \widetilde{q}_{i,h}^k(s,\cdot)\big\rangle\Big).
\end{aligned} \tag{86}$$

With these preparations in place, we are ready to embark on the proof.

### D.2.1 Proof of inequalities (49a) and (49b)

Recall the definition of $\overline{V}_{i,h}^{\widehat{\pi}}(s)$ in (45c) that

$$\overline{V}_{i,h}^{\widehat{\pi}}(s) = \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{a_i \sim \pi_{i,h}^k(s)} \Big[r_{i,h}^k(s, a_i) + P_{i,h}^k(\cdot \mid s, a_i)\overline{V}_{i,h+1}^{\widehat{\pi}}\Big] = \sum_{k=1}^K \alpha_k^K \Big\langle \pi_{i,h}^k(s),\, \widetilde{q}_{i,h}^k(s,\cdot)\Big\rangle. \tag{87}$$

It is first observed that

$$\begin{aligned}
\sum_{k=1}^K \mathbb{E}_{h,k-1}\Big[\alpha_k^K \big\langle \pi_{i,h}^k(s), \widetilde{q}_{i,h}^k(s,\cdot)\big\rangle\Big] &= \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_h^k(s)} \Big[r_{i,h}(s, \boldsymbol{a}) + \big\langle P_{i,h}(\cdot \mid s, \boldsymbol{a}), \overline{V}_{i,h+1}^{\widehat{\pi}}\big\rangle \mid \overline{V}_{i,h+1}^{\widehat{\pi}}, \pi_{i,h}^k\Big] \\
&= r_{i,h}^{\widehat{\pi}}(s) + \big\langle P_{i,h}^{\widehat{\pi}}(s,\cdot), \overline{V}_{i,h+1}^{\widehat{\pi}}\big\rangle,
\end{aligned} \tag{88}$$

where the second identity arises from the definitions (48) of $r_{i,h}^{\widehat{\pi}}$ and $P_{i,h}^{\widehat{\pi}}$. It is also seen that

$$R_1 \coloneqq \max_k \Big|\alpha_k^K \big\langle \pi_{i,h}^k(s), \widetilde{q}_{i,h}^k(s,\cdot)\big\rangle\Big| \le \Big\{\max_k \alpha_k^K\Big\}\Big\{\max_k \big\|\pi_{i,h}^k(s)\big\|_1 \big\|\widetilde{q}_{i,h}^k\big\|_\infty\Big\} \le \frac{2c_\alpha H \log K}{K},$$

where the first line invokes Lemma 1, (85) and the fact $\|\pi_{i,h}^k(s)\|_1 = 1$. Another observation is that

$$\begin{aligned}
W_1 = \sum_{k=1}^K \big(\alpha_k^K\big)^2 \mathsf{Var}_{h,k-1}\Big(\big\langle \pi_{i,h}^k(s), \widetilde{q}_{i,h}^k(s,\cdot)\big\rangle\Big) &\le \Big\{\max_k \alpha_k^K\Big\}\Big\{\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\big\langle \pi_{i,h}^k(s), \widetilde{q}_{i,h}^k(s,\cdot)\big\rangle\Big)\Big\} \\
&\le \frac{2c_\alpha \log K}{K} \sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\widehat{q}_{i,h}^k(s)\Big),
\end{aligned} \tag{89}$$

where the second line makes use of Lemma 1 and the inequality (86). With the definitions (87) and (88) in mind, invoking Freedman's inequality (i.e., Theorem 5) with $\kappa_1 = \sqrt{\frac{K \log \frac{K}{\delta}}{H}}$ then leads to

$$\begin{aligned}
&\Big|\overline{V}_{i,h}^{\widehat{\pi}}(s) - \Big(r_{i,h}^{\widehat{\pi}}(s) + \big\langle P_{i,h}^{\widehat{\pi}}(s,\cdot),\, \overline{V}_{i,h+1}^{\widehat{\pi}}\big\rangle\Big)\Big| \\
&\quad = \Big|\sum_{k=1}^K \alpha_k^K \Big\langle \pi_{i,h}^k(s), \widetilde{q}_{i,h}^k(s,\cdot)\Big\rangle - \sum_{k=1}^K \mathbb{E}_{h,k-1}\Big[\alpha_k^K \Big\langle \pi_{i,h}^k(s), \widetilde{q}_{i,h}^k(s,\cdot)\Big\rangle\Big]\Big| \\
&\quad \le \kappa_1 W_1 + \Big(\frac{2}{\kappa_1} + 5R_1\Big) \log \frac{3K}{\delta}
\end{aligned}$$

$$\leq 2c_\alpha \sqrt{\frac{\log^3 \frac{K}{\delta}}{KH}} \sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{h,k-1}\left(\widehat{q}_{i,h}^k(s)\right) + \left(2\sqrt{\frac{H}{K \log \frac{K}{\delta}}} + \frac{10 c_\alpha H \log K}{K}\right) \log \frac{3K}{\delta}$$

$$\leq 2c_\alpha \sqrt{\frac{\log^3 \frac{K}{\delta}}{KH}} \sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{h,k-1}\left(\widehat{q}_{i,h}^k(s)\right) + 4\sqrt{\frac{H \log \frac{3K}{\delta}}{K}} \tag{90}$$

with probability at least $1-\delta$, where the last relation holds true under Condition (38).

To continue, we note the first term in (90) can be bounded by Cauchy-Schwarz as follows:

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{h,k-1}\left(\widehat{q}_{i,h}^k(s)\right) &= \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\left(\widehat{q}_{i,h}^k(s)\right)^2\right] - \sum_{k=1}^{K} \alpha_k^K \left(\mathbb{E}_{h,k-1}\left[\widehat{q}_{i,h}^k(s)\right]\right)^2 \\ &\leq \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\left(\widehat{q}_{i,h}^k(s)\right)^2\right] - \left(\sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\widehat{q}_{i,h}^k(s)\right]\right)^2. \end{aligned} \tag{91}$$

Further, we make note of two additional facts:

- The weighted mean of $\widehat{q}_{i,h}^k(s)$ obeys

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\widehat{q}_{i,h}^k(s)\right] &= \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi_h^k(s)}\left[r_{i,h}(s,\boldsymbol{a})\right] + \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi_h^k(s)}\left[\left\langle P_{i,h}(\cdot \mid s,\boldsymbol{a}),\, \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right] \\ &= r_{i,h}^{\widehat{\pi}}(s) + \left\langle P_{i,h}^{\widehat{\pi}}(s,\cdot), \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle \geq \left\langle P_{i,h}^{\widehat{\pi}}(s,\cdot), \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle. \end{aligned} \tag{92}$$

- Regarding the square of $\widehat{q}_{i,h}^k(s)$, one has (see (84))

$$\begin{aligned}\left(\widehat{q}_{i,h}^k(s)\right)^2 &= \left(r_{i,h}^k(s,a_{k,s}) + \left\langle P_{i,h}^k(\cdot \mid s,a_{k,s}),\, \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right)^2 \\ &= \left(\left\langle P_{i,h}^k(\cdot \mid s,a_{k,s}),\, \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right)^2 + \left(r_{i,h}^k(s,a_{k,s})\right)^2 + 2r_{i,h}^k(s,a_{k,s})\left\langle P_{i,h}^k(\cdot \mid s,a_{k,s}),\, \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle \\ &\leq \left(\left\langle P_{i,h}^k(\cdot \mid s,a_{k,s}),\, \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right)^2 + 3H \\ &\leq \left\langle P_{i,h}^k(\cdot \mid s,a_{k,s}),\, \overline{V}_{i,h+1}^{\widehat{\pi}} \circ \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle + 3H, \end{aligned}$$

  where we have used the fact that $\|\overline{V}_{i,h+1}^{\widehat{\pi}}\|_\infty \leq H$ and $\|r_{i,h}^k\|_\infty \leq 1$; consequently,

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\left(\widehat{q}_{i,h}^k(s)\right)^2\right] &\leq \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\left\langle P_{i,h}^k(\cdot \mid s,a_{k,s}),\, \overline{V}_{i,h+1}^{\widehat{\pi}} \circ \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right] + 3H \\ &= \sum_{k=1}^{K} \alpha_k^K \sum_{a_i\in\mathcal{A}_i} \pi_{i,h}^k(a_i \mid s) \mathbb{E}_{h,k-1}\left[\left\langle P_{i,h}^k(\cdot \mid s,a_i),\, \overline{V}_{i,h+1}^{\widehat{\pi}} \circ \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right] + 3H \\ &= \left\langle P_{i,h}^{\widehat{\pi}}(s,\cdot), \overline{V}_{i,h+1}^{\widehat{\pi}} \circ \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle + 3H. \end{aligned} \tag{93}$$

Taking (92) and (93) together with (91) yields

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{h,k-1}\left(\widehat{q}_{i,h}^k(s)\right) &\leq \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\left(\widehat{q}_{i,h}^k(s)\right)^2\right] - \left(\sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{h,k-1}\left[\widehat{q}_{i,h}^k(s)\right]\right)^2 \\ &\leq \left\langle P_{i,h}^{\widehat{\pi}}(s,\cdot), \overline{V}_{i,h+1}^{\widehat{\pi}} \circ \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle - \left(\left\langle P_{i,h}^{\widehat{\pi}}(s,\cdot), \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right)^2 + 3H. \end{aligned}$$

To finish up, substituting these into (90) and making use of the assumption (38) give

$$\left|\overline{V}_{i,h}^{\widehat{\pi}}(s) - \left(r_{i,h}^{\widehat{\pi}}(s) + \left\langle P_{i,h}^{\widehat{\pi}}(s,\cdot),\, \overline{V}_{i,h+1}^{\widehat{\pi}}\right\rangle\right)\right|$$

$$\leq 2c_\alpha \sqrt{\frac{\log^3 \frac{K}{\delta}}{KH}} \left[ \left\langle P^{\widehat{\pi}}_{i,h}(s,\cdot), \overline{V}^{\widehat{\pi}}_{i,h+1} \circ \overline{V}^{\widehat{\pi}}_{i,h+1} \right\rangle - \left( \left\langle P^{\widehat{\pi}}_{i,h}(s,\cdot), \overline{V}^{\widehat{\pi}}_{i,h+1} \right\rangle \right)^2 \right] + (6c_\alpha + 4) \sqrt{\frac{H \log^3 \frac{K}{\delta}}{K}}$$

for any $s \in \mathcal{S}$, thus concluding the proof of the first claim (49a) of Lemma 4.

The second claim (49b) of Lemma 4 can be established using exactly the same argument, and hence we omit the proof here for the sake of brevity.

### D.2.2 Proof of inequality (49c)

We then turn to the last advertised inequality (49c). Given that $\overline{r}_{i,h}(s) + \overline{P}_{i,h}(s,\cdot)\widehat{V}_{i,h+1} \in [0, H - h + 1]$ for all $s \in \mathcal{S}$, we can recall the definition (20a) of $\widehat{V}_{i,h}$ to obtain

$$\left| \widehat{V}_{i,h}(s) - \left(\overline{r}_{i,h}(s) + \overline{P}_{i,h}(s,\cdot)\widehat{V}_{i,h+1}\right) \right| \leq \left| \sum_{k=1}^{K} \alpha_k^K \left\langle \pi^k_{i,h}(\cdot \mid s),\, q^k_{i,h}(s,\cdot) \right\rangle + \beta_{i,h}(s) - \left(\overline{r}_{i,h}(s) + \overline{P}_{i,h}(s,\cdot)\widehat{V}_{i,h+1}\right) \right| \tag{94}$$

for all $s \in \mathcal{S}$. The remaining analysis is dedicated to bounding the right-hand side of (94).

Let us begin with the following identity:

$$\begin{aligned} \sum_{k=1}^{K} \alpha_k^K \left\langle \pi^k_{i,h}(\cdot \mid s),\, q^k_{i,h}(s,\cdot) \right\rangle + \beta_{i,h}(s) &= \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{a_i \sim \pi^k_{i,h}(s)} \left[ r^k_{i,h}(s,a_i) + P^k_{i,h}(\cdot \mid s, a_i)\widehat{V}_{i,h+1} \right] + \beta_{i,h}(s) \\ &= \overline{r}_{i,h}(s) + \left\langle \overline{P}_{i,h}(s,\cdot), \widehat{V}_{i,h+1} \right\rangle + \beta_{i,h}(s), \end{aligned} \tag{95}$$

where we recall the definitions of $\overline{r}_{i,h} \in \mathbb{R}^S$ and $\overline{P}_{i,h} \in \mathbb{R}^{S \times S}$ in (48). The key step boils down to bounding the bonus term defined in (24), towards which we first claim that

$$\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{\pi^k_{i,h}(s)}\left(q^k_{i,h}(s,\cdot)\right) \leq 2 + 2\left[ \overline{P}_{i,h}(s,\cdot)\left(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\right) - \left(\overline{P}_{i,h}(s,\cdot)\widehat{V}_{i,h+1}\right)^2 \right] \tag{96}$$

holds for all $s \in \mathcal{S}$. Assuming the validity of this claim, we can then demonstrate that

$$\begin{aligned} \beta_{i,h}(s) &= c_{\mathsf{b}} \sqrt{\frac{\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{KH}} \sum_{k=1}^{K} \alpha_k^K \left\{ \mathsf{Var}_{\pi^k_{i,h}(s)}\left(q^k_{i,h}(s,\cdot)\right) + H \right\} \\ &\leq 2c_{\mathsf{b}} \sqrt{\frac{\log^3\left(\frac{KS\sum_i A_i}{\delta}\right)}{KH}} \left\{ \overline{P}_{i,h}(s,\cdot)\left(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\right) - \left(\overline{P}_{i,h}(s,\cdot)\widehat{V}_{i,h+1}\right)^2 + H \right\}, \end{aligned} \tag{97}$$

where we have used the identity $\sum_{k=1}^{K} \alpha_k^K = 1$. Hence, we can readily establish the desired result (49c) by combining (97) with (95) and (94), provided that $c_3 > 0$ is sufficiently large.

It remains to justify the claim (96). Towards this end, we make the observation that

$$\begin{aligned} \mathsf{Var}_{\pi^k_{i,h}(s)}\left(q^k_{i,h}(s,\cdot)\right) &\leq 2\mathsf{Var}_{\pi^k_{i,h}(s)}\left(r^k_{i,h}(s,\cdot)\right) + 2\mathsf{Var}_{\pi^k_{i,h}(s)}\left(\sum_{s'} P^k_{i,h}(s' \mid s,\cdot)\widehat{V}_{i,h+1}(s')\right) \\ &\leq 2 + 2\left[ \sum_{a_i} \pi^k_{i,h}(a_i \mid s) P^k_{i,h}(\cdot \mid s, a_i)\left(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\right) - \left( \sum_{a_i} \pi^k_{i,h}(a_i \mid s) P^k_{i,h}(\cdot \mid s, a_i)\widehat{V}_{i,h+1} \right)^2 \right], \end{aligned}$$

which results from $\|r^k_{i,h}\|_\infty \leq 1$ and the following relation:

$$\begin{aligned} &\mathsf{Var}_{\pi^k_{i,h}(s)}\left(\sum_{s'} P^k_{i,h}(s' \mid s,\cdot)\widehat{V}_{i,h+1}(s')\right) \\ &\quad= \sum_{a_i} \pi^k_{i,h}(a_i \mid s)\left(P^k_{i,h}(\cdot \mid s, a_i)\widehat{V}_{i,h+1}\right)^2 - \left( \sum_{a_i} \pi^k_{i,h}(a_i \mid s) P^k_{i,h}(\cdot \mid s, a_i)\widehat{V}_{i,h+1} \right)^2 \\ &\quad\leq \sum_{a_i} \pi^k_{i,h}(a_i \mid s) P^k_{i,h}(\cdot \mid s, a_i)\left(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\right) - \left( \sum_{a_i} \pi^k_{i,h}(a_i \mid s) P^k_{i,h}(\cdot \mid s, a_i)\widehat{V}_{i,h+1} \right)^2. \end{aligned}$$

This taken together with the fact $\sum_{k=1}^{K} \alpha_k^K = 1$ and Jensen's inequality yields

$$\begin{aligned}
&\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{\pi_{i,h}^k(s)}\big(q_{i,h}^k(s,\cdot)\big) \\
&\quad \le \sum_{k=1}^{K} \alpha_k^K \left\{ 2 + 2\left[ \sum_{a_i} \pi_{i,h}^k(a_i \mid s) P_{i,h}^k(\cdot \mid s, a_i)\big(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\big) - \Big( \sum_{a_i} \pi_{i,h}^k(a_i \mid s) P_{i,h}^k(\cdot \mid s, a_i) \widehat{V}_{i,h+1} \Big)^2 \right] \right\} \\
&\quad \le 2 + 2\overline{P}_{i,h}(s,\cdot)\big(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\big) - 2\Big( \sum_{k=1}^{K} \alpha_k^K \sum_{a_i} \pi_{i,h}^k(a_i \mid s) P_{i,h}^k(\cdot \mid s, a_i) \widehat{V}_{i,h+1} \Big)^2 \\
&\quad = 2 + 2\Big[ \overline{P}_{i,h}(s,\cdot)\big(\widehat{V}_{i,h+1} \circ \widehat{V}_{i,h+1}\big) - \big(\overline{P}_{i,h}(s,\cdot)\widehat{V}_{i,h+1}\big)^2 \Big]
\end{aligned}$$

as claimed.

### D.3 Minimax lower bound

In this section, we formalize the minimax lower bound claimed in (31).

**Theorem 4** (Minimax lower bound)**.** *Consider any $m \ge 2$ and any $0 < \varepsilon \le c_1 H$ for some small enough constant $c_1 > 0$. Then one can construct a collection of $m$-player zero-sum Markov games $\{\mathcal{MG}_\theta \mid \theta \in \Theta\}$ with $S$ states, horizon $H$, and $A_i$ actions for the $i$-th player ($1 \le i \le m$) such that*

$$\inf_{\widehat{\pi}} \max_{\theta \in \Theta} \mathbb{P}^{\mathcal{MG}_\theta}\left\{ \mathsf{gap}\big(\widehat{\pi}\big) > \varepsilon \right\} \ge \frac{1}{4}, \tag{98}$$

*provided that the total sample size obeys*

$$N \le \frac{c_2 H^4 S \max_{1 \le i \le m} A_i}{\varepsilon^2} \tag{99}$$

*for some sufficiently small constant $c_2 > 0$. Here, the infimum is over all (joint) policy estimator $\widehat{\pi}$, and $\mathbb{P}^{\mathcal{MG}_\theta}$ denotes the probability when the Markov game is $\mathcal{MG}_\theta$.*

*Proof.* Suppose without loss of generality that $A_1 \ge \max\{A_2, \ldots, A_m\}$. Let us begin by considering the special scenario with $A_2 = \ldots = A_m = 1$; in this case, computing either the NE or the CCE reduces to finding the optimal policy of a single-agent MDP with $S$ states and $A_1$ actions. It is well-known that for any given accuracy level $\varepsilon \in (0, H]$, there exists a non-stationary MDP with $S$ states and $A_1$ actions such that no algorithm can learn an $\varepsilon$-optimal policy with $o\big(\frac{H^4 S A_1}{\varepsilon^2}\big)$ samples [3, 41]. More precisely, for any given $0 < \varepsilon \le c_1 H$ for some small enough constant $c_1 > 0$, one can construct a collection of MDPs $\{\mathcal{M}_\theta \mid \theta \in \Theta\}$ such that

$$\inf_{\widehat{\mu}} \max_{\theta \in \Theta} \mathbb{P}^{\mathcal{M}_\theta}\left\{ \max_{s \in \mathcal{S}} \big(V_1^\star(s) - V_1^{\widehat{\mu}}(s)\big) > \varepsilon \right\} \ge \frac{1}{4}, \tag{100}$$

with the proviso that the total sample size obeys

$$N \le \frac{c_2 H^4 S A_1}{\varepsilon^2} \tag{101}$$

for some small enough constant $c_2 > 0$. Here, the infimum is over all policy estimate $\widehat{\mu}$ in this single-agent scenario, and $\mathbb{P}^{\mathcal{M}_\theta}$ denotes the probability when the MDP is $\mathcal{M}_\theta$.

Next, let us construct a collection of Markov games by augmenting each of the single-agent MDPs $\mathcal{M}_\theta$ with $A_i$ completely identical actions for the $i$-th player ($2 \le i \le m$); that is, to construct $\mathcal{MG}_\theta$, we take its reward function and probability transition kernel to be

$$r_{i,h}^{\mathcal{MG}_\theta}(s, \boldsymbol{a}) = \begin{cases} r_h^{\mathcal{M}_\theta}(s, a_1) & \text{if } i = 1 \\ -r_h^{\mathcal{M}_\theta}(s, a_1) & \text{if } i = m \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad P_h^{\mathcal{MG}_\theta}(\cdot \mid s, \boldsymbol{a}) = P_h^{\mathcal{M}_\theta}(\cdot \mid s, a_1) \tag{102}$$

for all $(s, h, \boldsymbol{a} = [a_1, \ldots, a_m]) \in \mathcal{S} \times [H] \times \mathcal{A}$. Evidently, finding either an NE or a CCE of $\mathcal{MG}_\theta$ is equivalent to computing the optimal policy of $\mathcal{M}_\theta$, given the non-distinguishability of the actions of all but the first player in $\mathcal{MG}_\theta$. This in turn immediately establishes the advertised lower bound. $\square$

### D.4 Freedman's inequality

In this section, we record the Freedman inequality for martingales [27] with slight modification, which is a crucial concentration bound for our analysis.

**Theorem 5.** *Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \le R \qquad \text{and} \qquad \mathbb{E}\left[X_k \mid \{X_j\}_{j:j<k}\right] = 0 \qquad \text{for all } k \ge 1$$

*for some quantity $R > 0$. Define*

$$W_n := \sum_{k=1}^n \mathbb{E}_{k-1}\left[X_k^2\right],$$

*where $\mathbb{E}_{k-1}$ stands for the expectation conditional on $\{X_j\}_{j:j<k}$. Consider any arbitrary quantity $\kappa > 0$. With probability at least $1-\delta$, one has*

$$|Y_n| \le \sqrt{8W_n \log\frac{3n}{\delta}} + 5R\log\frac{3n}{\delta} \le \kappa W_n + \left(\frac{2}{\kappa} + 5R\right)\log\frac{3n}{\delta}. \tag{103}$$

*Proof.* Suppose that $W_n \le \sigma^2$ holds deterministically for some quantity $\sigma^2$. As has been demonstrated in Li et al. [39, Theorem 5], with probability at least $1-\delta$ we have

$$|Y_n| \le \sqrt{8\max\left\{W_n, \frac{\sigma^2}{2^K}\right\}\log\frac{2K}{\delta}} + \frac{4}{3}R\log\frac{2K}{\delta} \tag{104}$$

for any positive integer $K \ge 1$. Recognizing the trivial bound $W_n \le nR^2$, one can take $\sigma^2 = nR^2$ and $K = \log_2 n$ to obtain

$$\begin{aligned}
|Y_n| &\le \sqrt{8\max\left\{W_n, R^2\right\}\log\frac{4\log_2 n}{\delta}} + \frac{4}{3}R\log\frac{4\log_2 n}{\delta} \\
&\le \sqrt{8W_n\log\frac{3n}{\delta}} + \sqrt{8R^2\log\frac{3n}{\delta}} + \frac{4}{3}R\log\frac{3n}{\delta} \\
&\le \sqrt{8W_n\log\frac{3n}{\delta}} + 5R\log\frac{3n}{\delta},
\end{aligned}$$

where we have used $4\log_2 n \le 3n$ for any integer $n \ge 1$. This establishes the first inequality in (103). The second inequality in (103) is then a direct consequence of the elementary inequality $2ab \le a^2 + b^2$. □