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

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

August 20, 2022

Abstract

This paper is concerned with two-player zero-sum Markov games — arguably the most basic setting in multi-agent reinforcement learning — with the goal of learning a Nash equilibrium (NE) 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 learning algorithm called Nash-Q-FTRL and an adaptive sampling scheme that leverage the optimism principle in adversarial learning (particularly the Follow-the-Regularized-Leader (FTRL) method), with a delicate design of bonus terms that ensure certain decomposability under the FTRL dynamics. Our algorithm learns an ε-approximate Markov NE policy using

$$\widetilde{O}\Big(\frac{H^4S(A+B)}{\varepsilon^2}\Big)$$

samples, where S is the number of states, H is the horizon, and A (resp. B) denotes the number of actions for the max-player (resp. min-player). This is nearly un-improvable in a minimax sense. 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.

Keywords: Markov games, sample complexity, Nash equilibrium, adversarial learning, Follow-the-Regularized-Leader

Contents

*Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.
†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

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 (Zhang et al., 2021). The recent developments in game playing (Brown and Sandholm, 2019; Vinyals et al., 2019), self-driving vehicles (Shalev-Shwartz et al., 2016), and multi-robot control (Matignon et al., 2012) 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) (Littman, 1994; Shapley, 1953), 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). In a Nash equilibrium, no gain can be realized through a unilateral change, and hence no player has incentives to deviate from her current strategy/policy (assuming no coordination between players). A myriad of research has been conducted surrounding NEs, which spans various aspects like existence, learnability, computational hardness, and algorithm design, among others (Chen et al., 2015; Daskalakis, 2013; Daskalakis et al., 2020; Hansen et al., 2013; Jin et al., 2022; Littman, 1994; Ozdaglar et al., 2021; Perolat et al., 2015; Rubinstein, 2016; Shapley, 1953).

Sample efficiency in zero-sum Markov games. 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 NEs with optimal sample complexity remains by and large unsettled, even when it comes to the most basic setting — two-player zero-sum Markov games. In what follows, let us review two representative algorithms developed on this front under two drastically different sampling protocols, and discuss the shortfalls of these cutting-edge results. To facilitate precise comparisons, our discussion concentrates on zero-sum Markov games involving two players, where S is the number of states, H indicates the horizon or effective horizon, and A and B 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 combinations), Zhang et al. (2020) investigated a natural model-based algorithm, which performs planning (e.g., value iteration) on an empirical MG derived from samples produced by the generative model. Focusing on *stationary* discounted infinite-horizon MGs, their algorithm finds an ε-approximate NE with no more than

$$\widetilde{O}\left(\frac{H^3SAB}{\varepsilon^2}\right) \quad \text{samples.} \tag{1}$$

In parallel, Liu et al. (2021) studied *non-stationary* finite-horizon MGs with online exploration, and obtained similar sample complexity bounds, i.e.,

$$\widetilde{O}\left(\frac{H^4SAB}{\varepsilon^2}\right) \text{ samples} \qquad \text{or} \qquad \widetilde{O}\left(\frac{H^3SAB}{\varepsilon^2}\right) \text{ episodes} \tag{2}$$

for learning ε-approximate NEs. 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., AB), 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. (2020); Jin et al. (2021) 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 ε-approximate NE in non-stationary finite-horizon MGs using

$$\widetilde{O}\left(\frac{H^6S(A+B)}{\varepsilon^2}\right) \text{ samples} \qquad \text{or} \qquad \widetilde{O}\left(\frac{H^5S(A+B)}{\varepsilon^2}\right) \text{ episodes,} \tag{3}$$

  which effectively brings down the sample size scaling (2) from AB (i.e., the number of joint actions) to $A+B$ (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.

The above results represent the state-of-the-art sample complexity theory thus far for learning NEs in two-player zero-sum MGs. In summary, these existing results — 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 Nash equilibria in a two-player zero-sum Markov game in a sample-optimal and computation-efficient fashion?*

To settle this question favorably, both of the above-mentioned hurdles need to be crossed simultaneously.

Main contributions. Recognizing that the above question remains open regardless of the sampling scheme in use, this paper takes a first step towards solving it assuming access to the most flexible sampling protocol: the generative model (or simulator). In sharp contrast to the single-agent case where uniform sampling of all state-action pairs suffices (Azar et al., 2013; Li et al., 2020), 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 ε-approximate NE in a *non-stationary* finite-horizon two-player zero-sum MG, we come up with a learning algorithm (accompanied by an adaptive sampling strategy) that accomplishes this goal with no more than

$$\widetilde{O}\left(\frac{H^4S(A+B)}{\varepsilon^2}\right) \text{ samples} \tag{4}$$

drawn from the generative model. Encouragingly, this sample complexity bound matches the minimax lower limit (up to a logarithmic factor). Our sample complexity theory is valid for the full ε-range (i.e., any $\varepsilon \in (0, H]$); this unveils that no burn-in cost whatsoever 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: (1) optimism in the face of uncertainty (by leveraging upper confidence bounds (UCBs) in value estimation), and (2) 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 proved in Jin et al. (2021); Song et al. (2021). 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 opponent's actions.

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 (2020); Chen et al. (2022); Cui and Du (2022a,b); Dou et al. (2022); Jia et al. (2019); Mao and Başar (2022); Tian et al. (2021); Wei et al. (2017); Yan et al. (2022); Zhong et al. (2022)), mainly because computing NEs becomes, for the most part, computationally infeasible (i.e., PPAD-complete) when going beyond two-player zero-sum MGs (Daskalakis, 2013; Daskalakis et al., 2009). Roughly speaking, previous NE-finding algorithms for two-player zero-sum Markov games can be categorized into model-based algorithms (Liu et al., 2021; Perolat et al., 2015; Zhang et al., 2020), value-based algorithms (Bai and Jin, 2020; Bai et al., 2020; Chen et al., 2021b; Jin et al., 2021; Sayin et al., 2021; Xie et al., 2020), and policy-based algorithms (Cen et al., 2021; Chen et al., 2021a; Daskalakis et al., 2020; Wei et al., 2021; Zhang et al., 2022; Zhao et al., 2021). In particular, Bai et al. (2020); Jin et al. (2021) developed the first algorithms to beat the curse of multiple agents in two-player zero-sum MGs, while Daskalakis et al. (2022); Jin et al. (2021); Mao and Başar (2022); Song et al. (2021) 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 (2022a,b); Yan et al. (2022) studied how to alleviate the sample size scaling with the number of agents in the presence of offline data, with Cui and Du (2022a) 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 (Liu et al., 2022), 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 (Agarwal et al., 2020; Azar et al., 2013; Beck and Srikant, 2012; Chen et al., 2020; Du et al., 2020; Even-Dar and Mansour, 2003; Jin and Sidford, 2021; Kakade, 2003; Kearns et al., 2002; Khamaru et al., 2021; Li et al., 2021a, 2020; Mou et al., 2020; Pananjady and Wainwright, 2020; Sidford et al., 2018a,b; Vaswani et al., 2022; Wainwright, 2019a,b; Wang et al., 2021; Wei et al., 2021; Weisz et al., 2021; Yang and Wang, 2019; Zanette et al., 2019, 2020). In single-agent RL, the model-based approach has been shown to be minimax-optimal for the entire ε-range (Agarwal et al., 2020; Azar et al., 2013; Li et al., 2020). When it comes to multi-agent RL, sample-efficient solutions with a generative model have been proposed in the recent works (Cui and Yang, 2021; Sidford et al., 2020; Zhang et al., 2020), although a provably sample-optimal strategy was previously unavailable.

Paper organization and notation. The rest of the paper is organized as follows. Section 2 introduces the background of Markov games, the preliminaries of the solution concepts of Nash equilibrium, and formulates the sampling protocol. The proposed learning algorithm and the sampling strategy are described in Section 3.1, with the theoretical guarantees provided in Section 3.2. Section 4 takes a detour to develop our refined regret bound for FTRL, which plays a crucial role in our main sample complexity analysis in Section 5. Proof details (particularly those for auxiliary lemmas) are postponed to the appendix.

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 convenient 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 two-player zero-sum Markov games and the solution concept of Nash equilibrum.

Two-player Markov games. A non-stationary finite-horizon two-player Markov game, denoted by $\mathcal{MG} = \big\{\mathcal{S}, \mathcal{A}, \mathcal{B}, H, P, r\big\}$, involves two players competing against each other — with the first one called a "max-player" and the second one called a "min-player" — and consists of several key elements to be formalized below. To begin with, we denote by $\mathcal{S} = \{1, \cdots, S\}$ a shared state space comprising S different states, and let $\mathcal{A} = \{1, \cdots, A\}$ (resp. $\mathcal{B} = \{1, \cdots, B\}$) represent the action space of the max-player (resp. min-player) containing A (resp. B) different actions. The horizon length of this finite-horizon Markov game is denoted by H. The probability transition kernel of $\mathcal{MG}$ is denoted by $P = \{P_h\}_{1\leq h\leq H}$ with $P_h \in \mathbb{R}^{SAB\times S}$; namely, for any $(s,a,b,h,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]\times\mathcal{S}$, we let $P_h(s' \,|\, s,a,b)$ indicate the probability of $\mathcal{MG}$ transitioning from state s to state s' at step h when the max-player takes action a and the min-player takes action b. Additionally, $r = \{r_h\}_{1\leq h\leq H}$ with $r_h \in \mathbb{R}^{SAB}$ represents the reward function; namely, for any $(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$, $r_h(s,a,b)$ stands for the immediate reward the max-player gains (or the min-player loses) in state s at step h, if the max-player (resp. min-player) executes action a (resp. b). Given that our focal point is the family of zero-sum Markov games, it suffices to introduce a single reward function r. We shall also assume normalized rewards throughout this paper in the sense that $r_h(s,a,b) \in [0,1]$ for any $(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$.

Value functions under product policies. Let $\mu = \{\mu_h\}_{1\leq h\leq H}$ (resp. $\nu = \{\nu_h\}_{1\leq h\leq H}$) denote the policy of the max-player (resp. min-player), where $\mu_h(\cdot \,|\, s) \in \Delta(\mathcal{A})$ and $\nu_h(\cdot \,|\, s) \in \Delta(\mathcal{B})$ for any $s \in \mathcal{S}$. More specifically, $\mu_h(a \,|\, s)$ indicates the probability of the max-player selecting action a in state s at step h; and $\nu_h(b \,|\, s)$ is defined analogously. Consider a Markovian trajectory $\{(s_t, a_t, b_t, r_t)\}_{1\leq t\leq H}$, where s_t is the state at time t, a_t (resp. b_t) is the action of the max-player (resp. min-player) at time t, and r_t is the immediate reward observed at time t. For any given policy μ (resp. ν) of the max-player (resp. min-player) and any step h, we define the value function $V_h^{\mu,\nu} : \mathcal{S} \to \mathbb{R}$ under the product policy $\mu\times\nu$ as follows:

$$V_h^{\mu,\nu}(s) \coloneqq \mathbb{E}\left[\sum_{t=h}^{H} r_t\big(s_t, a_t, b_t\big) \mid s_h = s\right], \qquad \forall s \in \mathcal{S}, \tag{7}$$

where the expectation is taken over the Markovian trajectory $\{(s_t, a_t, b_t, r_t)\}$ with the max-player and the min-player executing policies μ and ν, respectively, in an *independent* fashion; that is, conditional on s_t, we draw $a_t \sim \mu_t(\cdot \,|\, s_t)$ and $b_t \sim \nu_t(\cdot \,|\, s_t)$ independently, and then $s_{t+1} \sim P_t(\cdot \,|\, s_t, a_t, b_t)$. In addition, conditional

on the min-player executing policy ν, the optimal value function $V^{\star,\nu} = \{V_h^{\star,\nu}\}_{1\le h\le H}$ (with $V_h^{\star,\nu} : \mathcal{S} \to \mathbb{R}$) of the max-player is defined as

$$V_h^{\star,\nu}(s) \coloneqq \max_{\mu:\, \mathcal{S}\times[H]\to\Delta(\mathcal{A})} V_h^{\mu,\nu}(s), \qquad \forall (s,h) \in \mathcal{S}\times[H]; \tag{8a}$$

similarly, when the max-player chooses to execute policy μ, the optimal value function $V^{\mu,\star} = \{V_h^{\mu,\star}\}_{1\le h\le H}$ (with $V_h^{\mu,\star} : \mathcal{S} \to \mathbb{R}$) of the min-player is defined as

$$V_h^{\mu,\star}(s) \coloneqq \min_{\nu:\, \mathcal{S}\times[H]\to\Delta(\mathcal{B})} V_h^{\mu,\nu}(s), \qquad \forall (s,h) \in \mathcal{S}\times[H]. \tag{8b}$$

Furthermore, if we freeze the policy of the min-player to ν, then the Bellman optimality condition for the max-player can be expressed as (Bertsekas, 2017)

$$V_h^{\star,\nu}(s) = \max_{a\in\mathcal{A}} \left\{ \mathop{\mathbb{E}}_{b\sim\nu_h(\cdot|s)} \left[r_h(s,a,b) + \left\langle P_h(\cdot \,|\, s,a,b),\, V_{h+1}^{\star,\nu} \right\rangle \right] \right\}, \qquad \forall (s,h) \in \mathcal{S}\times[H]. \tag{9}$$

Solution concept: Nash equilibrium. In the zero-sum two-player Markov game, the max-player wishes to maximize the value function, while the min-player aims to minimize the value function instead. Due to the competing objectives, finding some sort of equilibrium — particularly the Nash equilibrium (Nash, 1951) — becomes a central topic in the studies of Markov games. More precisely, a policy pair $(\mu^\star, \nu^\star) \in \Delta(\mathcal{A})^{SH} \times \Delta(\mathcal{B})^{SH}$ is said to be a *(mixed-strategy) Nash equilibrium* of $\mathcal{MG}$ if the resulting product policy $\mu^\star \times \nu^\star$ obeys

$$V_1^{\mu^\star,\nu^\star}(s) = V_1^{\star,\nu^\star}(s) \qquad \text{and} \qquad V_1^{\mu^\star,\nu^\star}(s) = V_1^{\mu^\star,\star}(s), \qquad \text{for all } s \in \mathcal{S}. \tag{10}$$

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

In practice, it might be challenging to compute an "exact" Nash 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 pair (μ, ν) as follows (measured in an ℓ_∞-based manner)

$$\mathsf{NE\text{-}gap}(\mu,\nu) \coloneqq \max_{s\in\mathcal{S}} \mathsf{NE\text{-}gap}(\mu,\nu;s), \tag{11a}$$

where

$$\mathsf{NE\text{-}gap}(\mu,\nu;s) \coloneqq \max\left\{ V_1^{\star,\nu}(s) - V_1^{\mu,\nu}(s),\, V_1^{\mu,\nu}(s) - V_1^{\mu,\star}(s) \right\}. \tag{11b}$$

With this sub-optimality measure in place, a policy pair (μ, ν) is said to be an ε-approximate Nash-equilibrium — or more concisely, ε-Nash — if the resultant sub-optimality gap of the product policy $\mu \times \nu$ obeys $\mathsf{NE\text{-}gap}(\mu,\nu) \le \varepsilon$.

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 a set of data samples. When it comes to the data generating mechanism, this paper assumes access to a generative model (also called a simulator) (Kakade, 2003; Kearns et al., 2002): in each call to the generative model, the learner can choose an arbitrary quadruple $(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$ and obtain an independent sample generated based on the true transition kernel:

$$s' \sim P_h(\cdot \,|\, s,a,b).$$

In words, a generative model facilitates query of arbitrary state-action-step combinations, which helps alleviate the sampling constraints arising in online episodic settings for exploration. The goal of the current paper is to compute an ε-approximate Nash equilibrium 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 for NE

In this section, we put forward an algorithm aimed at finding an ε-Nash policy pair with the assistance of a generative model, and demonstrate its sample optimality for the full ε-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 (Bertsekas, 2017), 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 max-player (resp. min-player) calls the generative model for K rounds, with each round drawing SA (resp. SB) independent samples; as a result, the total sample size is given by $KS(A+B)H$. In what follows, let us first introduce some convenient notation that facilitates our exposition of the algorithm.

Notation. Consider any step $h \in [H]$, 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.

- $\overline{V}_h \in \mathbb{R}^S$ (resp. $\underline{V}_h \in \mathbb{R}^S$) represents the final estimate of the value function at step h by the max-player (resp. min-player); in particular, we set $\overline{V}_{H+1} = \underline{V}_{H+1} = 0$.
- $\overline{Q}_h^k \in \mathbb{R}^{SA}$ (resp. $\underline{Q}_h^k \in \mathbb{R}^{SB}$) represents the Q-function estimate of the max-player (resp. min-player) at step h after the k-th round of data collection.
- $\overline{q}_h^k \in \mathbb{R}^{SA}$ (resp. $\underline{q}_h^k \in \mathbb{R}^{SB}$) represents a certain "one-step-look-ahead" Q-function estimate of the max-player (resp. min-player) at step h using samples collected in the k-th round.
- $\overline{r}_h^k \in \mathbb{R}^{SA}$ (resp. $\underline{r}_h^k \in \mathbb{R}^{SB}$) denotes the sample reward vector for step h received by the max-player (resp. min-player) in the k-th round.
- $\overline{P}_h^k \in \mathbb{R}^{SA\times S}$ (resp. $\underline{P}_h^k \in \mathbb{R}^{SB\times S}$) denotes the empirical probability transition matrix for step h constructed using the samples collected by the max-player (resp. min-player) in the k-th round.
- $\overline{\beta}_{h,V} \in \mathbb{R}^S$ (resp. $\underline{\beta}_{h,V} \in \mathbb{R}^S$) denotes the bonus (resp. penalty) vector chosen by the max-player (resp. min-player) during final value estimation.
- $\mu_h^k : \mathcal{S} \to \Delta(\mathcal{A})$ (resp. $\nu_h^k : \mathcal{S} \to \Delta(\mathcal{B})$) denotes the policy iterate of the max-player (resp. min-player) at step h before the beginning of the k-th round of data collection; in particular, we set both μ_h^1 and ν_h^1 to be uniform, namely, $\mu_h^1(a \,|\, s) = 1/A$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $\nu_h^1(b \,|\, s) = 1/B$ for any $(s,b) \in \mathcal{S}\times\mathcal{B}$.

Crucially, the above objects are all constructed from the perspective of a single player (either the max-player or the min-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 max-player in the k-th round for step h; the procedure for the min-player proceeds analogously.

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

$$b_{k,h,s,a} \sim \nu_h^k(\cdot \,|\, s), \qquad s'_{k,h,s,a} \sim P_h(\cdot \,|\, s, a, b_{k,h,s,a}) \tag{12a}$$

and receive the reward $r_{k,h,s,a} = r_h\big(s,a,b_{k,h,s,a}\big)$. These samples are then employed to construct the sample reward vector $\overline{r}_h^k \in \mathbb{R}^{SA}$ and empirical probability transition kernel $\overline{P}_h^k \in \mathbb{R}^{SA\times S}$ such that

$$\overline{r}_h^k(s,a) = r_{k,h,s,a} \qquad \text{and} \qquad \overline{P}_h^k(s' \,|\, s,a) = \begin{cases} 1, & \text{if } s' = s'_{k,h,s,a} \\ 0, & \text{else} \end{cases} \tag{12b}$$

for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Note that the max-player only needs to compute (12b), without the need of directly observing the opponent's actions (i.e., $\{b_{k,h,s,a}\}$).

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

$$\overline{q}_h^k = \overline{r}_h^k + \overline{P}_h^k \overline{V}_{h+1}. \tag{13}$$

We then adopt the update rule of Q-learning:

$$\overline{Q}_h^k = (1 - \alpha_k)\overline{Q}_h^{k-1} + \alpha_k \overline{q}_h^k, \tag{14}$$

where $0 < \alpha_k < 1$ is the learning rate. Applying (14) recursively and using the quantities defined in (5), we easily arrive at the following expansion:

$$\overline{Q}_h^k = \sum_{i=1}^{k} \alpha_i^k \overline{q}_h^i. \tag{15}$$

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

$$\mu_h^{k+1}(a \,|\, s) = \frac{\exp\big(\eta_{k+1}\overline{Q}_h^k(s,a)\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\eta_{k+1}\overline{Q}_h^k(s,a')\big)}, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{16}$$

where $\eta_{k+1} > 0$ denotes another learning rate associated with policy updates for the max-player (to be specified shortly). In fact, this subroutine implements the Follow-the-Regularized-Leader update rule (Shalev-Shwartz, 2012) with

$$\mu_h^{k+1}(\cdot \,|\, s) = \arg\min_{\pi \in \Delta(\mathcal{A})} \left\{ -\big\langle \pi, \overline{Q}_h^k(s,\cdot)\big\rangle + \frac{1}{\eta_{k+1}} F(\pi) \right\}, \tag{17}$$

where the regularizer $F(\cdot)$ is chosen to be the negative entropy function $F(\pi) \coloneqq \sum_a \pi(a) \log\big(\pi(a)\big)$.

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

$$\widehat{\mu}_h = \sum_{k=1}^{K} \alpha_k^K \mu_h^k \qquad \text{and} \tag{18a}$$

$$\overline{V}_h(s) = \min\left\{ \sum_{k=1}^{K} \alpha_k^K \Big\langle \mu_h^k(\cdot \,|\, s),\, \overline{q}_h^k(s,\cdot)\Big\rangle + \overline{\beta}_{h,V}(s),\ H - h + 1 \right\} \tag{18b}$$

with $\{\alpha_k^K\}$ defined in (5), where $\overline{\beta}_{h,V}(s) \geq 0$ is some bonus term (taking the form of some data-driven upper confidence bound) to be specified momentarily. In particular, the final policy (18a) is a mixture of the policy iterates $\{\mu_h^k\}$ and is Markovian in nature (i.e., the action selection rule depends only on the current state s and step h).

The whole procedure, including the algorithm for the min-player, 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,

$$\alpha_k = \frac{c_\alpha \log K}{k - 1 + c_\alpha \log K}, \qquad k = 1, 2, \ldots \tag{19}$$

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

$$\eta_{k+1} = \sqrt{\frac{\log K}{\alpha_k H}}, \qquad k = 1, 2, \ldots \tag{20}$$

Algorithm 1: Nash-Q-FTRL.

Input: nmber of rounds K for each step, learning rates $\{\alpha_k\}$ (see (19)) and $\{\eta_{k+1}\}$ (see (20)).

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

Initialize: for all $(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$, set $\overline{V}_{H+1}(s) = \underline{V}_{H+1}(s) = \overline{Q}_h^0(s,a) = \underline{Q}_h^0(s,b) = 0$, $\mu_h^1(a\,|\,s) = 1/A$ and $\nu_h^1(b\,|\,s) = 1/B$.

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

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

    // draw independent samples, and construct empirical models.

    $\big(\overline{r}_h^k, \underline{r}_h^k, \overline{P}_h^k, \underline{P}_h^k\big) \leftarrow$ `sampling`$\big(h, \mu_h^k, \nu_h^k\big)$. /* see Algorithm 2. */

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

    Compute $\overline{q}_h^k = \overline{r}_h^k + \overline{P}_h^k \overline{V}_{h+1}$, $\underline{q}_h^k = \underline{r}_h^k + \underline{P}_h^k \underline{V}_{h+1}$, and update

$$\overline{Q}_h^k = (1-\alpha_k)\overline{Q}_h^{k-1} + \alpha_k \overline{q}_h^k, \quad \underline{Q}_h^k = (1-\alpha_k)\underline{Q}_h^{k-1} + \alpha_k \underline{q}_h^k.$$

    // update policy estimates using FTRL.

    for $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ **do**

$$\mu_h^{k+1}(a\,|\,s) = \frac{\exp\big(\eta_{k+1}\overline{Q}_h^k(s,a)\big)}{\sum_{a'}\exp\big(\eta_{k+1}\overline{Q}_h^k(s,a')\big)}, \quad \nu_h^{k+1}(b\,|\,s) = \frac{\exp\big(-\eta_{k+1}\underline{Q}_h^k(s,b)\big)}{\sum_{b'}\exp\big(-\eta_{k+1}\underline{Q}_h^k(s,b')\big)}.$$

  // output the final policy estimate and value estimate for step h.

  Update

$$\widehat{\mu}_h = \sum_{k=1}^K \alpha_k^K \mu_h^k, \qquad \overline{V}_h(s) = \min\left\{\sum_{k=1}^K \alpha_k^K \big\langle \mu_h^k(\cdot\,|\,s), \overline{q}_h^k(s,\cdot)\big\rangle + \overline{\beta}_{h,V}(s),\ H-h+1\right\},\ \forall s\in\mathcal{S},$$

$$\widehat{\nu}_h = \sum_{k=1}^K \alpha_k^K \nu_h^k, \qquad \underline{V}_h(s) = \max\left\{\sum_{k=1}^K \alpha_k^K \big\langle \mu_h^k(\cdot\,|\,s), \underline{q}_h^k(s,\cdot)\big\rangle - \underline{\beta}_{h,V}(s),\ 0\right\},\ \forall s\in\mathcal{S},$$

  where $\overline{\beta}_{h,V}^k$ is given in (21), and $\underline{\beta}_{h,V}^k$ is obtained by replacing $(\mu_h^k, \overline{q}_h^k)$ in (21) with $(\nu_h^k, \underline{q}_h^k)$.

Output: $\widehat{\mu} = \{\widehat{\mu}_h\}_{1\le h\le H}$ and $\widehat{\nu} = \{\widehat{\nu}_h\}_{1\le h\le H}$.

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

$$\overline{\beta}_{h,V}(s) = c_{\mathsf{b}}\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\sum_{k=1}^K \alpha_k^K\left\{\mathsf{Var}_{\mu_h^k(\cdot|s)}\big(\overline{q}_h^k(s,\cdot)\big) + H\right\} \tag{21}$$

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

Algorithm 2: Auxiliary function $\texttt{sampling}\big(h, \mu_h, \nu_h\big)$.

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

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

  Draw an independent sample from the generative model:

$$b_{h,s,a} \sim \nu_h(\cdot \,|\, s), \qquad s'_{h,s,a} \sim P_h(\cdot \,|\, s, a, b_{h,s,a}). \tag{22}$$

  set $\overline{r}(s, a) = r_h\big(s, a, b_{h,s,a}\big)$ and $\overline{P}\big(s'_{h,s,a} \,|\, s, a\big) = 1$.

for $(s, b) \in \mathcal{S} \times \mathcal{B}$ **do**

  Draw an independent sample from the generative model:

$$a_{h,s,b} \sim \mu_h(\cdot \,|\, s), \qquad s'_{h,s,a} \sim P_h(\cdot \,|\, s, a_{h,s,b}, b). \tag{23}$$

  set $\underline{r}(s, b) = r_h\big(s, a_{h,s,b}, b\big)$ and $\underline{P}\big(s'_{h,s,a} \,|\, s, b\big) = 1$.

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

3.2 Main results

As it turns out, the proposed algorithm is provably sample-efficient, whose sample complexity is characterized by the following theorem.

Theorem 1. *Consider any $\varepsilon \in (0, H]$ and any $0 < \delta < 1$. Suppose that*

$$K \geq \frac{c_{\mathsf{k}} H^3 \log^4 \frac{KS(A+B)}{\delta}}{\varepsilon^2} \tag{24}$$

for some large enough universal constant $c_{\mathsf{k}} > 0$. With probability at least $1 - \delta$, the sub-optimality gap (cf. (11)) of the policies $(\widehat{\mu}, \widehat{\nu})$ returned by Algorithm 1 obeys

$$\mathsf{NE\text{-}gap}\big(\widehat{\mu}, \widehat{\nu}\big) \leq \varepsilon.$$

Theorem 1 establishes a sample complexity upper bound for the proposed algorithm, which we take a moment to interpret as follows. The proof of this theorem is postponed to Section 5.

Sample complexity. When a generative model is available, Theorem 1 asserts that the total number of samples (i.e., $KS(A + B)H$) needed for Algorithm 1 to yield ε-Nash policies is at most

$$\text{(sample complexity)} \qquad \widetilde{O}\bigg(\frac{H^4 S(A + B)}{\varepsilon^2}\bigg). \tag{25}$$

As far as we know, this delivers the first result — for any sampling protocol — that simultaneously overcomes the long-horizon barrier and the curse of multiple agents. In comparison to Zhang et al. (2020) (cf. (1)), our result reveals that what ultimately matters is the total number of individual actions (i.e., $A + B$) as opposed to the total number AB of possible joint actions. Additionally, our result exhibits improved horizon dependency (by a factor of H^2) compared to Bai et al. (2020); Jin et al. (2021) (see (3)), although we remark that the online sampling protocol therein is more restrictive than a generative model.

Minimax optimality. In order to assess the tightness of our sample complexity bound (25), it is helpful to take a look at the information theoretical limit. Suppose without loss of generality that $A \geq B$. Consider the extreme scenario where $B = 1$, in which case the Nash-equilibrium reduces to the optimal policy of a single-agent MDP with S states and A actions. It is well-known that for any given accuracy level $\varepsilon \in (0, H]$, one can construct a non-stationary MDP with S states and A actions such that no algorithm can learn an

ε-optimal policy with fewer than the order of $\frac{H^4SA}{\varepsilon^2}$ samples (Azar et al., 2013; Li et al., 2022). This means that in general, the minimax sample complexity lower bound (w.r.t. finding an ε-Nash policy pair) scales as

$$\text{(minimax lower bound)} \qquad \frac{H^4S(A+B)}{\varepsilon^2} \tag{26}$$

modulo some logarithmic factor; see Appendix B.3 for a formal statement and its proof. Taking this together with (25) confirms the minimax optimality of our algorithm (up to logarithmic terms).

No burn-in sample size and full ε-range. It is noteworthy that the validity of our sample complexity bound (25) is guaranteed for the entire range of ε-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. (2020); Jin et al. (2021) cannot be implemented in this way without incurring information loss. In addition, our algorithm can be carried out in a decentralized fashion, with two players acting in a symmetric and independent manner (without the need of knowing each other's individual action); and our algorithm is "rational" in the sense that it converges to the best-response policy if one of the players freezes its policy. All this is achieved under minimal sample complexity with the aid of the generative model.

4 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.

4.1 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 (Lattimore and Szepesvári, 2020): 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 ℓ_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 α_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:

$$R_n := \max_{a\in\mathcal{A}} R_n(a) \qquad \text{with } R_n(a) := \sum_{k=1}^{n} \alpha_k^n \langle \pi_k, \ell_k\rangle - \sum_{k=1}^{n} \alpha_k^n \ell_k(a), \tag{27}$$

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

4.2 Refined regret bounds for FTRL

Follow-the-Regularized-Leader. The FTRL algorithm (Shalev-Shwartz, 2007; Shalev-Shwartz and Singer, 2007) tailored to the above online optimization setting adopts the following update rule:

$$\pi_{k+1} = \arg\min_{\pi \in \Delta(\mathcal{A})} \left\{ \langle \pi, L_k \rangle + \frac{1}{\eta_{k+1}} F(\pi) \right\}, \qquad k = 1, 2, \dots \tag{28}$$

where $\eta_{k+1} > 0$ denotes the learning rate, and $F(\cdot)$ is some convex regularization function employed to stabilize the learning process (Shalev-Shwartz, 2012). 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 (2020, Section 28.1))

$$\pi_{k+1}(a) = \frac{\exp\big(-\eta_{k+1} L_k(a)\big)}{\sum_{a' \in \mathcal{A}} \exp\big(-\eta_{k+1} L_k(a')\big)} \qquad \text{for all } a \in \mathcal{A}. \tag{29}$$

This update rule is also intimately connected to online mirror descent (Lattimore and Szepesvári, 2020).

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 2. *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*

$$\widehat{\eta}_k := \begin{cases} \eta_2, & \text{if } k = 1, \\ \frac{\eta_k}{1 - \alpha_k}, & \text{if } k > 1. \end{cases} \tag{30}$$

Then the regret (cf. (27)) of the FTRL algorithm satisfies

$$R_n \le \frac{5}{3} \sum_{k=1}^{n} \alpha_k^n \widehat{\eta}_k \alpha_k \mathsf{Var}_{\pi_k}(\ell_k) + \frac{\log A}{\eta_{n+1}} + 3 \sum_{k=1}^{n} \alpha_k^n \widehat{\eta}_k^2 \alpha_k^2 \big\|\ell_k\big\|_\infty^3 \mathbb{1}\left(\widehat{\eta}_k \alpha_k \big\|\ell_k\big\|_\infty > \frac{1}{3} \right), \tag{31}$$

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 2 is postponed to Appendix A. Let us take a moment to discuss the key distinction between Theorem 2 and prior theory.

- A key term in the regret bound (31) 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 (2020, 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 $\{\|\ell_k\|_\infty^2\}$.[1] 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. (2013); Jin et al. (2018); Li et al. (2022, 2021c)).

- The careful reader would remark that the final term of (31) 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.

5 Proof of Theorem 1

In this section, we present the proof of Theorem 1, following some preliminary facts and notation.

5.1 Preliminaries and notation

Given that $\varepsilon \le H$, the assumption (24) necessarily requires that

$$K \ge c_{\mathsf{k}} H \log^4 \frac{KS(A+B)}{\delta} \tag{32}$$

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. (19)) and the corresponding quantities $\{\alpha_i^k\}$ (cf. (5)).

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

$$\alpha_1 = 1, \qquad \sum_{i=1}^{k} \alpha_i^k = 1, \qquad \max_{1\le i\le k} \alpha_i^k \le \frac{2c_\alpha \log K}{k}. \tag{33a}$$

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

$$\max_{1\le i\le k/2} \alpha_i^k \le 1/K^6. \tag{33b}$$

Proof. The result (33a) is standard and has been recorded in previous works (e.g., Jin et al. (2018, Appendix B)). Regarding (33b), we note that for any $i \le k/2$ and $k \ge c_\alpha \log K + 1$,

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

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

Additionally, recognizing the definition in (13) and the upper bound $\overline{V}_{h+1}(s) \le H-h+1$ (cf. (18b)), we make note of the range of the iterates $\{\overline{q}_h^k\}$ as follows.

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

$$0 \le \overline{q}_h^k(s,a) \le H-h+1. \tag{34}$$

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

$$\mu(s) \coloneqq \mu(\cdot \mid s) \in \Delta(\mathcal{A}) \qquad \text{and} \qquad \nu(s) \coloneqq \nu(\cdot \mid s) \in \Delta(\mathcal{B}), \qquad \forall s \in \mathcal{S}.$$

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).

[1] Note that the Bregman divergence generated by the negative entropy function is the (generalized) KL divergence (Beck, 2017), 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.

5.2 Proof outline

With the above preliminaries in place, we are in a position to present our analysis. Let us single out an intermediate value function that is intimately connected to both $\widehat{\mu}$ and $\widehat{\nu}$ as follows:

$$V_{H+1}^{\widehat{\mu}\cdot\widehat{\nu}}(s) \coloneqq 0 \tag{35a}$$

$$V_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) \coloneqq \sum_{k=1}^{K} \sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} \alpha_k^K \mu_h^k(a\,|\,s)\nu_h^k(b\,|\,s)\Big[r_h(s,a,b) + \big\langle P_h(\cdot\,|\,s,a,b), V_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle\Big] \tag{35b}$$

for all $(s,h)\in\mathcal{S}\times[H]$. In other words, in each step h, the joint policy pair generating the above value function is chosen to be the mixture of product policies $\sum_{k=1}^K \alpha_k^K\big(\mu_h^k\times\nu_h^k\big)$. We remark that here and below, when we write the joint policy pair $\widehat{\mu}\cdot\widehat{\nu}$, we allow them to be dependent of each other, which should not be understood differently as the product measure $\widehat{\mu}\times\widehat{\nu}$ when policy $\widehat{\mu}$ and $\widehat{\nu}$ are executed independently.

It is easily seen that to establish Theorem 1, it suffices to prove the following two inequalities:

$$V_1^{\star,\widehat{\nu}}(s) - V_1^{\widehat{\mu}\cdot\widehat{\nu}}(s) \le \varepsilon/2 \qquad \text{and} \qquad V_1^{\widehat{\mu}\cdot\widehat{\nu}}(s) - V_1^{\widehat{\mu},\star}(s) \le \varepsilon/2, \tag{36}$$

In this subsection, we shall only prove the first inequality in (36); the second one in (36) can be established in the same way and hence we omit the proof for brevity.

Towards this, let us introduce the following policy:

$$\widetilde{\mu}^\star = \arg\max_{\mu:\mathcal{S}\times[H]\to\Delta(\mathcal{A})} V_1^{\mu,\widehat{\nu}}.$$

We observe the following key decomposition

$$V_h^{\star,\widehat{\nu}} - V_h^{\widehat{\mu}\cdot\widehat{\nu}} \le \big(V_h^{\star,\widehat{\nu}} - \overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}\big) + \big(\overline{V}_h^{\star,\widehat{\nu}} - \overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\big) + \big(\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - V_h^{\widehat{\mu}\cdot\widehat{\nu}}\big), \tag{37}$$

where we define

$$\overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}(s) \coloneqq \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{a\sim\widetilde{\mu}_h^\star(s)}\Big[\overline{r}_h^k(s,a) + \big\langle \overline{P}_h^k(\cdot\,|\,s,a), \overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\big\rangle\Big], \qquad \text{with } \overline{V}_{H+1}^{\widetilde{\mu}^\star,\widehat{\nu}} = 0, \tag{38a}$$

$$\overline{V}_h^{\star,\widehat{\nu}}(s) \coloneqq \max_{a\in\mathcal{A}} \sum_{k=1}^K \alpha_k^K\Big[\overline{r}_h^k(s,a) + \big\langle \overline{P}_h^k(\cdot\,|\,s,a), \overline{V}_{h+1}^{\star,\widehat{\nu}}\big\rangle\Big], \qquad \text{with } \overline{V}_{H+1}^{\star,\widehat{\nu}} = 0, \tag{38b}$$

$$\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) \coloneqq \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{a\sim\mu_h^k(s)}\Big[\overline{r}_h^k(s,a) + \big\langle \overline{P}_h^k(\cdot\,|\,s,a), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle\Big], \qquad \text{with } \overline{V}_{H+1}^{\widehat{\mu}\cdot\widehat{\nu}} = 0. \tag{38c}$$

Here, we have used $\overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}} \le \overline{V}_h^{\star,\widehat{\nu}}$ given that $\widetilde{\mu}^\star$ is a policy that does not change with the index k. We shall establish bounds for the above terms, which is composed of three steps as outlined below.

Step 1: showing that $\overline{V}_h$ is an entrywise upper bound on $\overline{V}_h^{\star,\widehat{\nu}}$. The following lemma ascertains that the value estimate $\overline{V}_h$ of the max-player returned by Algorithm 1 is an optimistic estimate of the auxiliary value $\overline{V}_h^{\star,\widehat{\nu}}$ defined in (38b). Evidently, this result cannot happen unless the bonus terms are suitably chosen.

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

$$\overline{V}_h \ge \overline{V}_h^{\star,\widehat{\nu}}, \qquad \text{for all } 1\le h\le H. \tag{39}$$

The proof of this lemma is postponed to Appendix B.1. Armed with Lemma 3, we can further bound (37) as follows

$$V_h^{\star,\widehat{\nu}} - V_h^{\widehat{\mu}\cdot\widehat{\nu}} \le \big(V_h^{\star,\widehat{\nu}} - \overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}\big) + \big(\overline{V}_h - \overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\big) + \big(\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - V_h^{\widehat{\mu}\cdot\widehat{\nu}}\big). \tag{40}$$

Step 2: establishing a key recurrence. Let us define the auxiliary reward vectors $r_h^{\widehat{\mu}\cdot\widehat{\nu}}, r_h^{\widetilde{\mu}^\star,\widehat{\nu}}, \overline{r}_h \in \mathbb{R}^S$ and auxiliary probability transition matrices $P_h^{\widehat{\mu}\cdot\widehat{\nu}}, P_h^{\widetilde{\mu}^\star,\widehat{\nu}}, \overline{P}_h \in \mathbb{R}^{S\times S}$ such that: for any $s, s' \in \mathcal{S}$,

$$r_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) \coloneqq \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{(a,b)\sim\mu_h^k(s)\times\nu_h^k(s)} \big[r_h(s,a,b)\big], \tag{41a}$$

$$P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,s') \coloneqq \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{(a,b)\sim\mu_h^k(s)\times\nu_h^k(s)} \big[P_h(s' \,|\, s,a,b)\big], \tag{41b}$$

$$r_h^{\widetilde{\mu}^\star,\widehat{\nu}}(s) \coloneqq \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{(a,b)\sim\widetilde{\mu}_h^\star(s)\times\nu_h^k(s)} \big[r_h(s,a,b)\big], \tag{41c}$$

$$P_h^{\widetilde{\mu}^\star,\widehat{\nu}}(s,s') \coloneqq \sum_{k=1}^{K} \alpha_k^K \mathop{\mathbb{E}}_{(a,b)\sim\widetilde{\mu}_h^\star(s)\times\nu_h^k(s)} \big[P_h(s' \,|\, s,a,b)\big], \tag{41d}$$

$$\overline{r}_h(s) \coloneqq \sum_{k=1}^{K} \alpha_k^K \sum_{a\in\mathcal{A}} \mu_h^k(a\,|\,s)\overline{r}_h^k(s,a), \tag{41e}$$

$$\overline{P}_h(s,s') \coloneqq \sum_{k=1}^{K} \alpha_k^K \sum_{a\in\mathcal{A}} \mu_h^k(a\,|\,s)\overline{P}_h^k(s'\,|\,s,a). \tag{41f}$$

As it turns out, $\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}$ (resp. $\overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}$, $\overline{V}_h$) stays reasonably close to the "one-step-look-ahead" expression $r_h^{\widehat{\mu}\cdot\widehat{\nu}} + P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}$ (resp. $r_h^{\widetilde{\mu}^\star,\widehat{\nu}} + P_h^{\widetilde{\mu}^\star,\widehat{\nu}}\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}$, $\overline{r}_h + \overline{P}_h\overline{V}_{h+1}$), as revealed by the recursive relations stated in the following lemma; the proof of this lemma is deferred to Appendix B.2.

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

$$\begin{aligned}
\Big|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - \big(r_h^{\widehat{\mu}\cdot\widehat{\nu}} + P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big| &\le c_3\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1 \\
&\quad + c_3\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big) - \big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big],
\end{aligned} \tag{42a}$$

$$\begin{aligned}
\Big|\overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}} - \big(r_h^{\widetilde{\mu}^\star,\widehat{\nu}} + P_h^{\widetilde{\mu}^\star,\widehat{\nu}}\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\big)\Big| &\le c_3\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1 \\
&\quad + c_3\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\Big[P_h^{\widetilde{\mu}^\star,\widehat{\nu}}\big(\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\circ\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\big) - \big(P_h^{\widetilde{\mu}^\star,\widehat{\nu}}\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\big)\circ\big(P_h^{\widetilde{\mu}^\star,\widehat{\nu}}\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\big)\Big],
\end{aligned} \tag{42b}$$

$$\begin{aligned}
\Big|\overline{V}_h - \big(\overline{r}_h + \overline{P}_h\overline{V}_{h+1}\big)\Big| &\le c_3\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1 \\
&\quad + c_3\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\Big[\overline{P}_h\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big) - \big(\overline{P}_h\overline{V}_{h+1}\big)\circ\big(\overline{P}_h\overline{V}_{h+1}\big)\Big]
\end{aligned} \tag{42c}$$

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

Remark 2. The right-hand side of each of the bounds in (42) contains a variance-style term (e.g., those terms taking the form of $P_h(V_{h+1}\circ V_{h+1}) - (P_hV_{h+1})\circ(P_hV_{h+1})$ for some probability transition matrix P_h and value vector V_{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

$$\Big|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\Big| \le r_h^{\widehat{\mu}\cdot\widehat{\nu}} + c_3\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1$$

$$+\frac{c_3}{H}\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big]$$
$$\le\frac{c_4}{4}1+\frac{1}{4H}\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big]=:\zeta_0 \tag{43}$$

for some large enough constant $c_4>0$, where the last line holds due to the condition (32), the basic fact $P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\ge\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)$, and the following fact (for large enough c_4)

$$c_3\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1+r_h^{\widehat{\mu}\cdot\widehat{\nu}}\le c_3\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1+1\le\frac{c_4}{4}1.$$

In addition, recalling that $\|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\|_\infty,\|\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\|_\infty\le H$ (cf. (18b)) and recognizing that $\zeta_0\ge0$ (see (43)), we can demonstrate that

$$\begin{aligned}\Big|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big|&=\Big|\big(\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}+P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}-P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big|\\&\le\big(\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}+P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\zeta_0\le2H\zeta_0\\&=\frac{c_4}{2}H1+\frac{1}{2}\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big].\end{aligned} \tag{44}$$

This further leads to

$$\begin{aligned}&P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\\&\quad=P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}+\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\\&\quad\le P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}+\frac{c_4}{2}H1+\frac{1}{2}\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big],\end{aligned}$$

which can be rearranged to yield

$$P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\circ\big(P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\le2\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\Big]+c_4H1.$$

Substituting it into (42a) and combining terms give

$$\begin{aligned}\Big|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}-\big(r_h^{\widehat{\mu}\cdot\widehat{\nu}}+P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big|&\le c_5\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1\\&\quad+2c_3\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)-\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\Big],\end{aligned} \tag{45}$$

where we take $c_5=c_3+c_3c_4$.

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

$$\begin{aligned}\Big|\overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}-\big(r_h^{\widetilde{\mu}^\star,\widehat{\nu}}+P_h^{\widetilde{\mu}^\star,\widehat{\nu}}\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\big)\Big|&\le c_5\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1\\&\quad+2c_3\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\Big[P_h^{\widetilde{\mu}^\star,\widehat{\nu}}\big(\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\circ\overline{V}_{h+1}^{\widetilde{\mu}^\star,\widehat{\nu}}\big)-\overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}\circ\overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}\Big],\end{aligned} \tag{46}$$

$$\begin{aligned}\Big|\overline{V}_h-\big(\overline{r}_h+\overline{P}_h\overline{V}_{h+1}\big)\Big|&\le c_5\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}1\\&\quad+2c_3\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\Big[\overline{P}_h\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big)-\overline{V}_h\circ\overline{V}_h\Big].\end{aligned} \tag{47}$$

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)

$$\prod_{j:j<h} P_j^{\widehat{\mu}\cdot\widehat{\nu}} := \begin{cases} P_1^{\widehat{\mu}\cdot\widehat{\nu}}\cdots P_{h-1}^{\widehat{\mu}\cdot\widehat{\nu}}, & \text{if } h>1, \\ I, & \text{if } h=1. \end{cases}$$

Armed with this notation, we can invoke the relation (45) recursively and use $\overline{V}_{H+1}^{\widehat{\mu}\cdot\widehat{\nu}} = V_{H+1}^{\widehat{\mu}\cdot\widehat{\nu}} = 0$ to obtain

$$\begin{aligned}
\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - V_h^{\widehat{\mu}\cdot\widehat{\nu}} &= r_h^{\widehat{\mu}\cdot\widehat{\nu}} + P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} + \Big(\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - \big(r_h^{\widehat{\mu}\cdot\widehat{\nu}} + P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big) - \big(r_h^{\widehat{\mu}\cdot\widehat{\nu}} + P_h^{\widehat{\mu}\cdot\widehat{\nu}}V_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big) \\
&\le P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} - V_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big) + \Big|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - \big(r_h^{\widehat{\mu}\cdot\widehat{\nu}} + P_h^{\widehat{\mu}\cdot\widehat{\nu}}\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big)\Big| \qquad (48) \\
&\le c_5\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}\left(\sum_{h=1}^{H}\prod_{j:j<h} P_j^{\widehat{\mu}\cdot\widehat{\nu}}\right)1 \\
&\quad + 2c_3\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\sum_{h=1}^{H}\prod_{j:j<h} P_j^{\widehat{\mu}\cdot\widehat{\nu}}\Big[P_h^{\widehat{\mu}\cdot\widehat{\nu}}\big(\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big) - \overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\circ\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\Big] \\
&\le c_5\sqrt{\frac{H\log^3\frac{KS(A+B)}{\delta}}{K}}\left(\sum_{h=1}^{H}\prod_{j:j<h} P_j^{\widehat{\mu}\cdot\widehat{\nu}}\right)1 = c_5\sqrt{\frac{H^3\log^3\frac{KS(A+B)}{\delta}}{K}}1 \le \frac{\varepsilon}{6}1. \qquad (49)
\end{aligned}$$

Here, the first line uses the Bellman equation, the third inequality holds since for any transition matrices $\{P_h\}$ and any sequence $\{V_h\}$ obeying $V_{H+1}=0$, one can use the telescoping sum to obtain

$$\begin{aligned}
\sum_{h=1}^{H}\prod_{j:j<h} P_j\Big[P_h\big(V_{h+1}\circ V_{h+1}\big) - V_h\circ V_h\Big] &= \sum_{h=1}^{H}\prod_{j:j\le h} P_j\big(V_{h+1}\circ V_{h+1}\big) - \sum_{h=1}^{H}\prod_{j:j<h} P_j\big(V_h\circ V_h\big) \\
&= \prod_{j:j\le H} P_j\big(V_{H+1}\circ V_{H+1}\big) - V_1\circ V_1 \\
&= -V_1\circ V_1 \le 0,
\end{aligned}$$

whereas the last inequality in (49) arises from the assumption (24) when c_{k} is large enough. Similarly, replacing $\widehat{\mu}$ with $\widetilde{\mu}^\star$ in the above argument and recalling (46) directly lead to

$$V_h^{\star,\widehat{\nu}} - \overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}} = V_h^{\widetilde{\mu}^\star,\widehat{\nu}} - \overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}} \le \frac{\varepsilon}{6}1. \qquad (50)$$

In addition, recalling the definition of $\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}$ (cf. (38c)), $\overline{r}_h$ and $\overline{P}_h$ (see (41)), we can deduce that

$$\begin{aligned}
\overline{V}_h - \overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} &= \overline{r}_h + \overline{P}_h\overline{V}_{h+1} + \Big\{\overline{V}_h - \big(\overline{r}_h + \overline{P}_h\overline{V}_{h+1}\big)\Big\} - \overline{r}_h - \overline{P}_h\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \\
&\le \overline{P}_h\big(\overline{V}_{h+1} - \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big) + \Big|\overline{V}_h - \big(\overline{r}_h + \overline{P}_h\overline{V}_{h+1}\big)\Big|,
\end{aligned}$$

which resembles (48). Thus, repeating the above argument for (49) and applying (47) recursively, we reach

$$\overline{V}_h - \overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} \le \frac{\varepsilon}{6}1. \qquad (51)$$

Combining (49), (50), and (51) with (40), we arrive at

$$V_h^{\star,\widehat{\nu}} - V_h^{\widehat{\mu}\cdot\widehat{\nu}} \le \big(V_h^{\star,\widehat{\nu}} - \overline{V}_h^{\widetilde{\mu}^\star,\widehat{\nu}}\big) + \big(\overline{V}_h - \overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}\big) + \big(\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}} - V_h^{\widehat{\mu}\cdot\widehat{\nu}}\big) \le \frac{\varepsilon}{2}1.$$

This establishes the first inequality in (36), while the second inequality in (36) can be validated via the same argument. We have thus completed the proof of Theorem 1.

6 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 two-player zero-sum 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 (Liu et al., 2021; Zhang et al., 2020), whereas the availability of the generative model in conjunction with the variance-aware bonus design enables sharpened horizon dependency compared to Bai et al. (2020); Jin et al. (2021). Our work opens further questions surrounding sample efficiency in solving Markov games. For instance, how to attain minimax-optimal sample complexity if we only have access to less idealistic sampling protocol (e.g., local access models (Li et al., 2021b; Yin et al., 2022), and online sampling protocols (Azar et al., 2017; Jin et al., 2018)) as opposed to the flexible generative model? How can we optimize the horizon dependency when computing (coarse) correlated equilibria in multi-agent general-sum scenarios (Daskalakis et al., 2022; Jin et al., 2021; Song et al., 2021) without compromising the dependency on the size of the action spaces. In addition, our refined regret bound for FTRL (based on variance-type quantities) only covers the full-information case; it would be of interest to generalize it to the bandit-feedback setting (where only partial entries of the loss vectors are observable each time).

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.

A Proof of Theorem 2

This section is devoted to presenting the proof of Theorem 2. Before embarking on the analysis, let us introduce a convenient auxiliary iterate

$$\pi_{k+1}^{-} = \arg\min_{\pi\in\Delta(\mathcal{A})} \left\{ \langle \pi, L_k\rangle + \frac{1}{\widehat{\eta}_k} F(\pi) \right\}, \tag{52}$$

or equivalently,

$$\pi_{k+1}^{-}(a) = \frac{\exp\big(-\widehat{\eta}_k L_k(a)\big)}{\sum_{a'\in\mathcal{A}} \exp\big(-\widehat{\eta}_k L_k(a')\big)} \qquad \text{for all } a\in\mathcal{A}, \tag{53}$$

which differs from (29) only in the learning rates being used (namely, π_{k+1} uses η_{k+1} while π_{k+1}^{-} adopts $\widehat{\eta}_k$).

A.1 Main steps of the proof

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

$$R_n \le \sum_{k=1}^{n} \alpha_k^n \big\langle \pi_k - \pi_{k+1}^{-}, \ell_k\big\rangle + \frac{\log A}{\eta_{n+1}}; \tag{54}$$

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

$$\pi_{k+1}^{-}(a) \ge \begin{cases} \big[1-\widehat{\eta}_k\alpha_k\ell_k(a)\big]\pi_k(a), & \text{if } \widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > \frac{1}{3}, \\ \Big\{1-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big) - 2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}\big(\ell_k\big)\Big\}\pi_k(a), & \text{if } \widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \le \frac{1}{3}, \end{cases} \tag{55}$$

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 (54) allows us to replace the action that appears best in hindsight (cf. (27)) by the time-varying predictions $\{\pi^-_{k+1}\}$ without incurring much cost, whereas the second claim (55) controls the proximity of π^-_{k+1} and π_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 (54), we are in need of controlling $\langle \pi_k - \pi^-_{k+1}, \ell_k\rangle$. We divide into two cases.

- For any k obeying $\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty > 1/3$, invoke (55) and the non-negativity of ℓ_k to reach

$$\langle \pi_k - \pi^-_{k+1}, \ell_k\rangle \le \sum_{a\in\mathcal{A}} \widehat{\eta}_k\alpha_k\pi_k(a)\big[\ell_k(a)\big]^2 \le \sum_{a\in\mathcal{A}} \widehat{\eta}_k\alpha_k\pi_k(a)\big\|\ell_k\big\|_\infty^2 = \widehat{\eta}_k\alpha_k\big\|\ell_k\big\|_\infty^2. \tag{56}$$

- In contrast, if $\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \le 1/3$, then it follows from (55) that

$$\begin{aligned}
\langle \pi_k - \pi^-_{k+1}, \ell_k\rangle &\le \sum_{a\in\mathcal{A}} \left\{\widehat{\eta}_k\alpha_k\big(\ell_k(a) - \mathbb{E}_{\pi_k}[\ell_k]\big) + 2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}(\ell_k)\right\}\pi_k(a)\ell_k(a) \\
&= \widehat{\eta}_k\alpha_k \sum_{a\in\mathcal{A}} \pi_k(a)\big(\ell_k(a) - \mathbb{E}_{\pi_k}\big[\ell_k\big]\big)\mathbb{E}_{\pi_k}\big[\ell_k\big] + \widehat{\eta}_k\alpha_k \sum_{a\in\mathcal{A}} \pi_k(a)\big(\ell_k(a) - \mathbb{E}_{\pi_k}\big[\ell_k\big]\big)^2 \\
&\quad + 2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}(\ell_k)\sum_{a\in\mathcal{A}} \pi_k(a)\ell_k(a) \\
&= \widehat{\eta}_k\alpha_k \sum_{a\in\mathcal{A}} \pi_k(a)\big(\ell_k(a) - \mathbb{E}_{\pi_k}\big[\ell_k\big]\big)^2 + 2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}(\ell_k)\sum_{a\in\mathcal{A}} \pi_k(a)\ell_k(a) \\
&\le \widehat{\eta}_k\alpha_k\mathsf{Var}_{\pi_k}\big(\ell_k\big) + 2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}(\ell_k)\big\|\ell_k\big\|_\infty,
\end{aligned} \tag{57}$$

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

Putting the above two cases together yields

$$\begin{aligned}
&\sum_{k=1}^n \alpha_k^n\langle \pi_k - \pi^-_{k+1}, \ell_k\rangle \\
&\le \sum_{k=1}^n \alpha_k^n\widehat{\eta}_k\alpha_k\big\|\ell_k\big\|_\infty^2 \mathbb{1}\left(\widehat{\eta}_k\alpha_k\big\|\ell_k\big\|_\infty > \frac{1}{3}\right) + \sum_{k=1}^n \alpha_k^n\widehat{\eta}_k\alpha_k\mathsf{Var}_{\pi_k}\big(\ell_k\big)\mathbb{1}\left(\widehat{\eta}_k\alpha_k\big\|\ell_k\big\|_\infty \le \frac{1}{3}\right) \\
&\quad + 2\sum_{k=1}^n \alpha_k^n\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}\big(\ell_k\big)\big\|\ell_k\big\|_\infty\mathbb{1}\left(\widehat{\eta}_k\alpha_k\big\|\ell_k\big\|_\infty \le \frac{1}{3}\right) \\
&\le \frac{5}{3}\sum_{k=1}^n \alpha_k^n\widehat{\eta}_k\alpha_k\mathsf{Var}_{\pi_k}\big(\ell_k\big) + 3\sum_{k=1}^n \alpha_k^n\widehat{\eta}_k^2\alpha_k^2\big\|\ell_k\big\|_\infty^3\mathbb{1}\left(\widehat{\eta}_k\alpha_k\big\|\ell_k\big\|_\infty > \frac{1}{3}\right),
\end{aligned} \tag{58}$$

where the last inequality holds true since

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

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

Substituting (58) into (54), we can readily arrive at

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

It thus remains to establish the claims (54) and (55), which we shall accomplish next.

A.2 Proof of claim (54)

We claim that it suffices to prove that

$$\alpha_1^n\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^n}{\eta_2\alpha_1}F(\pi_2) + \sum_{k=2}^n \left\{\alpha_k^n\langle \pi_{k+1}^-, \ell_k\rangle + \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1})\right\}$$
$$\leq \min_{\pi\in\Delta(\mathcal{A})}\left\{\Big\langle \pi, \sum_{k=1}^n \alpha_k^n\ell_k\Big\rangle + \frac{1}{\eta_{n+1}}F(\pi)\right\}. \tag{59}$$

In fact, suppose that this inequality (59) is valid, then one can easily obtain

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

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

$$\begin{aligned}
R_n &= \sum_{k=1}^n \alpha_k^n\langle \pi_k, \ell_k\rangle - \min_{a\in\mathcal{A}}\sum_{k=1}^n \alpha_k^n\ell_k(a)\\
&\leq \sum_{k=1}^n \alpha_k^n\langle \pi_k - \pi_{k+1}^-, \ell_k\rangle - \sum_{k=2}^n \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1}) + \frac{\alpha_1^n}{\eta_2\alpha_1}\log A,
\end{aligned} \tag{60}$$

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

$$\begin{aligned}
-\sum_{k=2}^n \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]F(\pi_{k+1}) + \frac{\alpha_1^n}{\eta_2\alpha_1}\log A &\leq \sum_{k=2}^n \Big[\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}}\Big]\log A + \frac{\alpha_1^n}{\eta_2\alpha_1}\log A\\
&= \frac{\alpha_n^n}{\eta_{n+1}\alpha_n}\log A = \frac{\log A}{\eta_{n+1}}.
\end{aligned} \tag{61}$$

Substitution into (60) leads to

$$R_n \leq \sum_{k=1}^n \alpha_k^n\langle \pi_k - \pi_{k+1}^-, \ell_k\rangle + \frac{\log A}{\eta_{n+1}} \tag{62}$$

as advertised. As a consequence, everything boils down to establishing (59).

Towards this end, we would like to proceed with an induction argument, with the induction hypothesis w.r.t. n given by (59). 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 (28) and (52),

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

Secondly, suppose that (59) holds w.r.t. n, and we intend to justify it w.r.t. $n+1$. To do so, we observe that

$$\begin{aligned}
&\alpha_1^{n+1}\langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^{n+1}}{\eta_2\alpha_1} F(\pi_2) + \sum_{k=2}^{n} \left\{ \alpha_k^{n+1}\langle \pi_{k+1}^-, \ell_k\rangle + \Big(\frac{\alpha_k^{n+1}}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^{n+1}}{\eta_k\alpha_{k-1}} \Big) F(\pi_{k+1}) \right\} + \alpha_{n+1}\langle \pi_{n+2}^-, \ell_{n+1}\rangle \\
&\overset{\text{(i)}}{=} (1-\alpha_{n+1}) \left\{ \alpha_1^n \langle \pi_2^-, \ell_1\rangle + \frac{\alpha_1^n}{\eta_2\alpha_1} F(\pi_2) + \sum_{k=2}^{n} \left\{ \alpha_k^n \langle \pi_{k+1}^-, \ell_k\rangle + \Big(\frac{\alpha_k^n}{\eta_{k+1}\alpha_k} - \frac{\alpha_{k-1}^n}{\eta_k\alpha_{k-1}} \Big) F(\pi_{k+1}) \right\} \right\} \\
&\qquad + \alpha_{n+1}\langle \pi_{n+2}^-, \ell_{n+1}\rangle \\
&\overset{\text{(ii)}}{\le} (1-\alpha_{n+1}) \left\{ \Big\langle \pi_{n+2}^-, \sum_{k=1}^{n} \alpha_k^n \ell_k \Big\rangle + \frac{1}{\eta_{n+1}} F(\pi_{n+2}^-) \right\} + \alpha_{n+1}\langle \pi_{n+2}^-, \ell_{n+1}\rangle \\
&\overset{\text{(iii)}}{=} \Big\langle \pi_{n+2}^-, \sum_{k=1}^{n+1} \alpha_k^{n+1} \ell_k \Big\rangle + \frac{1-\alpha_{n+1}}{\eta_{n+1}} F(\pi_{n+2}^-) = \min_{\pi\in\Delta(\mathcal{A})} \left\{ \Big\langle \pi, \sum_{k=1}^{n+1} \alpha_k^{n+1} \ell_k \Big\rangle + \frac{1}{\widehat{\eta}_{n+1}} F(\pi) \right\}.
\end{aligned} \tag{63}$$

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 (59) w.r.t. n. To finish up, invoke (63) 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 (63), and the last identity comes from (28). This justifies the induction hypothesis w.r.t. $n+1$. Applying the induction argument in turn establishes (59) for all n, thereby concluding the proof.

A.3 Proof of claim (55)

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 (29). This in turn allows us to demonstrate that

$$\pi_{k+1}^-(a) = \frac{\exp\big(-\widehat{\eta}_k L_k(a)\big)}{\sum_{a'} \exp\big(-\widehat{\eta}_k L_k(a')\big)} = \frac{\exp\big(-\eta_k L_{k-1}(a)\big)}{\sum_{a'} \exp\big(-\eta_k L_{k-1}(a')\big)} \cdot \frac{\exp\big(-\widehat{\eta}_k \alpha_k \ell_k(a)\big)}{\sum_{a'} \pi_k(a') \exp\big(-\widehat{\eta}_k \alpha_k \ell_k(a')\big)}$$

$$= \pi_k(a)\frac{\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a')\big)} \geq \big[1-\widehat{\eta}_k\alpha_k\ell_k(a)\big]\pi_k(a),$$

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

Next, suppose that $\widehat{\eta}_k\alpha_k\|\ell_k\|_\infty \leq 1/3$. In this case, it is self-evident that $\widehat{\eta}_k\alpha_k|\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]|\leq 2/3$ 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

$$\begin{aligned}
\pi^-_{k+1}(a) = \pi_k(a)\frac{\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a)\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\ell_k(a')\big)} &= \frac{\exp\big(-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\big(\ell_k(a')-\mathbb{E}_{\pi_k}[\ell_k]\big)\big)}\pi_k(a) \\
&\geq \frac{1-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)}{\sum_{a'}\pi_k(a')\exp\big(-\widehat{\eta}_k\alpha_k\big(\ell_k(a')-\mathbb{E}_{\pi_k}[\ell_k]\big)\big)}\pi_k(a) \\
&\geq \frac{1-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)}{1+\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}(\ell_k)}\pi_k(a); \qquad (64)
\end{aligned}$$

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

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

Applying the elementary inequality $\frac{1-a}{1+b}\geq (1-a)(1-b)=1-a-b+ab\geq 1-a-2b$ for any $a\in[-1,1]$ and $b>0$, we can continue to lower bound (64) as follows

$$(64)\geq \Big\{1-\widehat{\eta}_k\alpha_k\big(\ell_k(a)-\mathbb{E}_{\pi_k}[\ell_k]\big)-2\widehat{\eta}_k^2\alpha_k^2\mathsf{Var}_{\pi_k}\big(\ell_k\big)\Big\}\pi_k(a),$$

thereby completing the proof.

B Proofs of auxiliary lemmas and details

B.1 Proof of Lemma 3

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

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

holds for any ν. Next, let us assume that the claim (39) is valid for step $h+1$, namely,

$$\overline{V}_{h+1} \geq \overline{V}_{h+1}^{\star,\widehat{\nu}}, \qquad (65)$$

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 2) for FTRL (see (17)). More precisely, we claim that

$$\max_a \overline{Q}_h^K(s,a) \leq \sum_{k=1}^K \alpha_k^K\Big\langle \mu_h^k(s),\, \overline{q}_h^k(s,\cdot)\Big\rangle$$

$$+10\sqrt{\frac{c_\alpha \log^3(KA)}{KH}}\sum_{k=1}^{K}\alpha_k^K \mathsf{Var}_{\mu_h^k(s)}\Big(\overline{q}_h^k(s,\cdot)\Big)+2\sqrt{\frac{c_\alpha H\log^3(KA)}{K}} \tag{66}$$

for any $s\in\mathcal{S}$, whose proof is deferred to Appendix B.1.1. Recall the construction (18b) of $\overline{V}_h$. If $\overline{V}_h = H-h+1$, then the claimed result $\overline{V}_h \geq \overline{V}_h^{\star,\widehat{\nu}}$ holds trivially. Therefore, it suffices to focus on the case where

$$\overline{V}_h(s)=\sum_{k=1}^{K}\alpha_k^K\Big\langle \mu_h^k(s),\,\overline{q}_h^k(s,\cdot)\Big\rangle+\overline{\beta}_{h,V}(s).$$

In this case, recalling the definition of $\overline{V}_h^{\star,\widehat{\nu}}(s)$ in (38b) gives

$$\begin{aligned}
\overline{V}_h^{\star,\widehat{\nu}}(s) &= \max_a \sum_{k=1}^{K}\alpha_k^K\big(\overline{r}_h^k(s,a)+\overline{P}_h^k(\cdot\,|\,s,a)\overline{V}_{h+1}^{\star,\widehat{\nu}}\big)\\
&\leq \max_a \sum_{k=1}^{K}\alpha_k^K\big(\overline{r}_h^k(s,a)+\overline{P}_h^k(\cdot\,|\,s,a)\overline{V}_{h+1}\big)=\max_a \overline{Q}_h^K(s,a)\\
&\leq \sum_{k=1}^{K}\alpha_k^K\Big\langle \mu_h^k(s),\,\overline{q}_h^k(s,\cdot)\Big\rangle+10\sqrt{\frac{c_\alpha\log^3(KA)}{KH}}\sum_{k=1}^{K}\alpha_k^K\mathsf{Var}_{\mu_h^k(s)}\Big(\overline{q}_h^k(s,\cdot)\Big)+2\sqrt{\frac{c_\alpha H\log^3(KA)}{K}}\\
&\leq \sum_{k=1}^{K}\alpha_k^K\Big\langle \mu_h^k(s),\,\overline{q}_h^k(s,\cdot)\Big\rangle+\overline{\beta}_{h,V}(s)=\overline{V}_h(s)
\end{aligned}$$

holds simultaneously for all $(s,h)\in\mathcal{S}\times[H]$. Here, the second line follows from the induction hypothesis (65) and the definition of $\overline{Q}_h^K$ in (15), the third line invokes the claim (66), whereas the last line comes from our choice (21) of $\overline{\beta}_{h,V}$ (provided that c_{b} is sufficiently large).

B.1.1 Proof of claim (66)

Consider any state $s\in\mathcal{S}$. By virtue of the identity $\overline{Q}_h^k=\sum_{i=1}^{k}\alpha_i^k\overline{q}_h^i$ (see (15)), the policy update rule (16) (or (17)) for $\mu_h^k(s)$ can essentially be viewed as the FTRL algorithm applied to the sequence of loss vectors

$$\ell_k=-\overline{q}_h^k(s,\cdot),\qquad k\geq 1.$$

Moreover, recalling the definition (20) of η_{k+1} and the definition (19) of α_k (with $c_\alpha\geq 24$), we have

$$\left(\frac{\eta_k}{\eta_{k+1}}\right)^2=\frac{\alpha_k}{\alpha_{k-1}}=\frac{k-2+c_\alpha\log K}{k-1+c_\alpha\log K}\geq\frac{k-1}{k-1+c_\alpha\log K}=1-\alpha_k>(1-\alpha_k)^2. \tag{67}$$

This property (67) permits us to invoke Theorem 2 to obtain

$$\begin{aligned}
\max_{a\in\mathcal{A}}\overline{Q}_h^K(s,a)-\sum_{k=1}^{K}\alpha_k^K\Big\langle\mu_h^k(s),\,\overline{q}_h^k(s,\cdot)\Big\rangle &= \max_{a\in\mathcal{A}}\left\{\sum_{k=1}^{K}\alpha_k^K\big\langle\mu_h^k(s),\,\ell_k\big\rangle-\sum_{k=1}^{K}\alpha_k^K\ell_k(a)\right\}\\
&\leq \frac{5}{3}\sum_{k=2}^{K}\alpha_k^K\frac{\eta_k\alpha_k}{1-\alpha_k}\mathsf{Var}_{\mu_h^k(s)}\Big(\overline{q}_h^k(s,\cdot)\Big)+\frac{\log A}{\eta_{K+1}}+\xi_h\\
&\overset{\text{(i)}}{\leq} \frac{5}{3}\sum_{k=2}^{K/2}\frac{(2c_\alpha)^{1.5}\log^2 K}{\sqrt{kH}}\alpha_k^K\mathsf{Var}_{\mu_h^k(s)}\Big(\overline{q}_h^k(s,\cdot)\Big)\\
&\quad+\frac{20}{3}\sum_{k=K/2+1}^{K}\alpha_k^K\sqrt{\frac{c_\alpha\log^2 K}{KH}}\mathsf{Var}_{\mu_h^k(s)}\Big(\overline{q}_h^k(s,\cdot)\Big)+\frac{\log A}{\eta_{K+1}}+\xi_h,
\end{aligned} \tag{68}$$

where ξ_h is defined as

$$\xi_h \coloneqq \frac{5}{3}\alpha_1^K \eta_2 \big\|\overline{q}_h^1\big\|_\infty^2 + \left\{3\sum_{k=2}^{K} \alpha_k^K \frac{\eta_k^2 \alpha_k^2}{(1-\alpha_k)^2} \big\|\overline{q}_h^k\big\|_\infty^3 \mathbb{1}\left(\frac{\eta_k \alpha_k}{1-\alpha_k}\big\|\overline{q}_h^k\big\|_\infty > \frac{1}{3}\right)\right\} + 3\alpha_1^K \eta_2^2 \big\|\overline{q}_h^1\big\|_\infty^3. \tag{69}$$

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

$$1-\alpha_k = 1 - \frac{c_\alpha \log K}{k-1+c_\alpha \log K} \geq \begin{cases} 1 - \frac{c_\alpha \log K}{1+c_\alpha \log K} = \frac{1}{1+c_\alpha \log K} \geq \frac{1}{2c_\alpha \log K}, & \text{if } k \geq 2, \\ 1 - \frac{c_\alpha \log K}{K/2+c_\alpha \log K} = \frac{K}{K+2c_\alpha \log K} \geq \frac{1}{2}, & \text{if } k \geq K/2+1, \end{cases} \tag{70a}$$

$$\eta_k \alpha_k = \sqrt{\frac{\log K}{\alpha_{k-1} H}} \cdot \alpha_k \leq \sqrt{\frac{\log K}{\alpha_k H}} \cdot \alpha_k = \sqrt{\frac{\alpha_k \log K}{H}} \leq \sqrt{\frac{2c_\alpha \log^2 K}{kH}}, \tag{70b}$$

where the first line makes use of (32) for large enough c_{k}, and the second line relies on (33a) in Lemma 1.

To proceed, let us control the terms in (68) separately.

- We start with the first term in (68). The elementary bound $\big\|\overline{q}_h^k\big\|_\infty \leq H$ in Lemma 2 taken together with (33b) in Lemma 1 helps us derive

$$\begin{aligned} \sum_{k=2}^{K/2} \frac{\alpha_k^K \log^2 K}{\sqrt{kH}} \mathsf{Var}_{\mu_h^k(s)}\left(\overline{q}_h^k(s,\cdot)\right) &\leq \sum_{k=2}^{K/2} \frac{\log^2 K}{K^6\sqrt{kH}} \mathsf{Var}_{\mu_h^k(s)}\left(\overline{q}_h^k(s,\cdot)\right) \\ &\leq \sum_{k=2}^{K/2} \frac{\log^2 K}{K^6\sqrt{kH}} \big\|\overline{q}_h^k(s,\cdot)\big\|_\infty^2 \leq \frac{H^{3/2}\log^2 K}{K^6} \sum_{k=2}^{K/2} \frac{1}{\sqrt{k}} \\ &\leq \frac{2H^{3/2}\log^2 K}{K^6} \cdot \sqrt{K/2} \leq \frac{2H^{3/2}\log^2 K}{K^5}. \end{aligned} \tag{71}$$

- Turning to the third term in (68), we recall the definition of η_{K+1} (cf. (20)) to obtain

$$\frac{\log A}{\eta_{K+1}} = \log A \sqrt{\frac{\alpha_K H}{\log K}} \leq \sqrt{\frac{2c_\alpha H \log^2 A}{K}}, \tag{72}$$

  where the inequality comes from Lemma 1.

- Finally, we move on to the last term in (68). For any $k \geq 2$, one can combine Lemma 2 with (70) to deduce that

$$\frac{\eta_k \alpha_k}{1-\alpha_k}\big\|\overline{q}_h^k\big\|_\infty \leq \frac{\sqrt{\frac{2c_\alpha \log^2 K}{kH}}}{\frac{1}{2c_\alpha \log K}} \cdot H = \sqrt{\frac{8c_\alpha^3 H \log^4 K}{k}}. \tag{73}$$

  Clearly, the right-hand side of (73) 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 (32)). Consequently, one can derive

$$\begin{aligned} \xi_h &= \frac{5}{3}\alpha_1^K \eta_2 \big\|\overline{q}_h^1\big\|_\infty^2 + \left\{3\sum_{k=2}^{K} \alpha_k^K \frac{\eta_k^2 \alpha_k^2}{(1-\alpha_k)^2} \big\|\overline{q}_h^k\big\|_\infty^3 \mathbb{1}\left(\frac{\eta_k \alpha_k}{1-\alpha_k}\big\|\overline{q}_h^k\big\|_\infty > \frac{1}{3}\right)\right\} + 3\alpha_1^K \eta_2^2 \big\|\overline{q}_h^1\big\|_\infty^3 \\ &\leq \frac{5}{3K^6}\sqrt{\frac{\log K}{H}}\big\|\overline{q}_h^1\big\|_\infty^2 + \frac{\left(2c_\alpha \log K\right)^2}{K^6}\left\{3\sum_{k=2}^{c_9 H \log^4 \frac{K}{\delta}} \eta_k^2 \alpha_k^2 \big\|\overline{q}_h^k\big\|_\infty^3\right\} + \frac{3}{K^6}\frac{\log K}{H}\big\|\overline{q}_h^1\big\|_\infty^3 \\ &\leq \frac{24c_\alpha^3 \log^4 K}{K^6 H}\left\{\sum_{k=1}^{K} \frac{1}{k} H^3\right\} \\ &\leq \frac{24c_\alpha^3 H^2 \log^5 K}{K^6} \leq \frac{1}{K^4}, \end{aligned} \tag{74}$$

  where the second line comes from (70) and the fact that $K/2 > c_9 H \log^4 \frac{K}{\delta}$ (as a consequence of (32)), and the third line holds due to Lemma 2.

Putting the preceding bounds together and substituting them into (68), we arrive at

$$\begin{aligned}
&\max_a \overline{Q}_h^K(s,a) - \sum_{k=1}^K \alpha_k^K \Big\langle \mu_h^k(s),\, \overline{q}_h^k(s,\cdot)\Big\rangle \\
&\quad\le \frac{5(2c_\alpha)^{1.5}}{3}\cdot\frac{2H^{3/2}\log^2 K}{K^5} + \frac{20}{3}\sqrt{\frac{c_\alpha \log^2 K}{KH}}\sum_{k=K/2+1}^K \alpha_k^K \mathsf{Var}_{\mu_h^k(s)}\Big(\overline{q}_h^k(s,\cdot)\Big) + \sqrt{\frac{2c_\alpha H\log^2 A}{K}} + \frac{1}{K^4} \\
&\quad\le 10\sqrt{\frac{c_\alpha \log^3(KA)}{KH}}\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{\mu_h^k(s)}\Big(\overline{q}_h^k(s,\cdot)\Big) + 2\sqrt{\frac{c_\alpha H\log^3(KA)}{K}},
\end{aligned} \tag{75}$$

where the last line is valid under Condition (32). This completes the proof of Claim (66).

B.2 Proof of Lemma 4

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

$$\widetilde{q}_h^k(s,a) \coloneqq \overline{r}_h^k(s,a) + \overline{P}_h^k(\cdot\,|\,s,a)\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}, \qquad \forall (s,a)\in\mathcal{S}\times\mathcal{A}.$$

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

$$a_{k,s}\sim\mu_h^k(s),$$

which is independent from $\widetilde{q}_h^k$ conditional on μ_h^k. This allows us to define another set of random variables

$$\widehat{q}_h^k(s) \coloneqq \widetilde{q}_h^k\big(s,a_{k,s}\big), \qquad \forall s\in\mathcal{S}, \tag{76}$$

which plays a central role in our analysis. It is readily seen from the facts $\overline{V}_{h+1}(s)\le H-h$ (cf. (18b)) and $\overline{r}_h^k(s,a)\in[0,1]$ that

$$0\le \widehat{q}_h^k(s), \widetilde{q}_h^k(s,a)\le H-h+1, \qquad \forall (s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[H]\times[K]. \tag{77}$$

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

$$\begin{aligned}
\mathsf{Var}_{h,k-1}\Big(\widehat{q}_h^k(s)\Big) &= \mathsf{Var}_{h,k-1}\Big(\big\langle e(a_{k,s}),\, \widetilde{q}_h^k(s,\cdot)\big\rangle\Big) \\
&\ge \mathsf{Var}_{h,k-1}\Big(\mathbb{E}_{h,k-1}\Big[\big\langle e(a_{k,s}),\, \widetilde{q}_h^k(s,\cdot)\big\rangle \mid \widetilde{q}_h^k\Big]\Big) \\
&= \mathsf{Var}_{h,k-1}\Big(\big\langle \mu_h^k(s),\, \widetilde{q}_h^k(s,\cdot)\big\rangle\Big).
\end{aligned} \tag{78}$$

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

B.2.1 Proof of inequalities (42a) and (42b)

Recall the definition of $\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}(s)$ in (38c) that

$$\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) = \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{a\sim\mu_h^k(s)}\Big[\overline{r}_h^k(s,a) + \overline{P}_h^k(\cdot\,|\,s,a)\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\Big] = \sum_{k=1}^K \alpha_k^K \Big\langle \mu_h^k(s),\, \widetilde{q}_h^k(s,\cdot)\Big\rangle. \tag{79}$$

It is first observed that

$$\begin{aligned}
\sum_{k=1}^K \mathbb{E}_{h,k-1}\Big[\alpha_k^K\big\langle \mu_h^k(s), \widetilde{q}_h^k(s,\cdot)\big\rangle\Big] &= \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{(a,b)\sim\mu_h^k(s)\times\nu_h^k(s)}\Big[r_h(s,a,b) + \big\langle P_h(\cdot\,|\,s,a,b), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle \mid \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}, \mu_h^k, \nu_h^k\Big] \\
&= r_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) + \big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle,
\end{aligned} \tag{80}$$

where the second identity arises from the definitions (41) of $r_h^{\widehat{\mu}\cdot\widehat{\nu}}$ and $P_h^{\widehat{\mu}\cdot\widehat{\nu}}$. It is also seen that

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

where the first line invokes Lemma 1, (77) and the fact $\|\mu_h^k(s)\|_1 = 1$. Another observation is that

$$\begin{aligned} W_1 = \sum_{k=1}^K \big(\alpha_k^K\big)^2 \mathsf{Var}_{h,k-1}\Big(\langle \mu_h^k(s), \widetilde{q}_h^k(s,\cdot)\rangle\Big) &\le \Big\{\max_k \alpha_k^K\Big\}\Big\{\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\langle \mu_h^k(s), \widetilde{q}_h^k(s,\cdot)\rangle\Big)\Big\} \\ &\le \frac{2c_\alpha \log K}{K}\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\widehat{q}_h^k(s)\Big), \end{aligned} \tag{81}$$

where the second line makes use of Lemma 1 and the inequality (78). With the definitions (79) and (80) in mind, invoking Freedman's inequality (i.e., Theorem 4) with $\kappa_1 = \sqrt{\frac{K\log\frac{K}{\delta}}{H}}$ then leads to

$$\begin{aligned} &\left|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) - \Big(r_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) + \big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle\Big)\right| \\ &\quad = \left|\sum_{k=1}^K \alpha_k^K \Big\langle \mu_h^k(s), \widetilde{q}_h^k(s,\cdot)\Big\rangle - \sum_{k=1}^K \mathbb{E}_{h,k-1}\Big[\alpha_k^K \Big\langle \mu_h^k(s), \widetilde{q}_h^k(s,\cdot)\Big\rangle\Big]\right| \\ &\quad \le \kappa_1 W_1 + \left(\frac{2}{\kappa_1} + 5R_1\right)\log\frac{3K}{\delta} \\ &\quad \le 2c_\alpha\sqrt{\frac{\log^3\frac{K}{\delta}}{KH}}\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\widehat{q}_h^k(s)\Big) + \left(2\sqrt{\frac{H}{K\log\frac{K}{\delta}}} + \frac{10c_\alpha H\log K}{K}\right)\log\frac{3K}{\delta} \\ &\quad \le 2c_\alpha\sqrt{\frac{\log^3\frac{K}{\delta}}{KH}}\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\widehat{q}_h^k(s)\Big) + 4\sqrt{\frac{H\log\frac{3K}{\delta}}{K}} \end{aligned} \tag{82}$$

with probability at least $1-\delta$, where the last relation holds true under Condition (32).

To continue, we note the first term in (82) can be bounded by Cauchy-Schwarz as follows:

$$\begin{aligned} \sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\widehat{q}_h^k(s)\Big) &= \sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\Big[\big(\widehat{q}_h^k(s)\big)^2\Big] - \sum_{k=1}^K \alpha_k^K \Big(\mathbb{E}_{h,k-1}\big[\widehat{q}_h^k(s)\big]\Big)^2 \\ &\le \sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\Big[\big(\widehat{q}_h^k(s)\big)^2\Big] - \Big(\sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\big[\widehat{q}_h^k(s)\big]\Big)^2. \end{aligned} \tag{83}$$

Further, we make note of two additional facts:

- The weighted mean of $\widehat{q}_h^k(s)$ obeys

$$\begin{aligned} \sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\big[\widehat{q}_h^k(s)\big] &= \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{(a,b)\sim\mu_h^k(s)\times\nu_h^k(s)}\big[r_h(s,a,b)\big] + \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{(a,b)\sim\mu_h^k(s)\times\nu_h^k(s)}\Big[\big\langle P_h(\cdot\,|\,s,a,b), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle\Big] \\ &= r_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) + \big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle \ge \big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle; \end{aligned} \tag{84}$$

- Regarding the square of $\widehat{q}_h^k(s)$, one has (see (76))

$$\begin{aligned} \big(\widehat{q}_h^k(s)\big)^2 &= \Big(\overline{r}_h^k(s,a_{k,s}) + \big\langle \overline{P}_h^k(\cdot\,|\,s,a_{k,s}), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle\Big)^2 \\ &= \Big(\big\langle \overline{P}_h^k(\cdot\,|\,s,a_{k,s}), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle\Big)^2 + \Big(\overline{r}_h^k(s,a_{k,s})\Big)^2 + 2\overline{r}_h^k(s,a_{k,s})\big\langle \overline{P}_h^k(\cdot\,|\,s,a_{k,s}), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\big\rangle \end{aligned}$$

$$\leq \Big(\big\langle \overline{P}_h^k(\cdot \,|\, s, a_{k,s}),\, \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \big\rangle\Big)^2 + 3H$$
$$\leq \Big\langle \overline{P}_h^k(\cdot \,|\, s, a_{k,s}),\, \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \circ \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \Big\rangle + 3H,$$

where we have used the fact that $\|\overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}}\|_\infty \leq H$ and $\|\overline{r}_h^k\|_\infty \leq 1$; consequently,

$$\begin{aligned}
\sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\Big[\big(\widehat{q}_h^k(s)\big)^2\Big] &\leq \sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\Big[\big\langle \overline{P}_h^k(\cdot \,|\, s, a_{k,s}),\, \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \circ \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \big\rangle\Big] + 3H \\
&= \sum_{k=1}^K \alpha_k^K \sum_a \mu_k^h(a \,|\, s) \mathbb{E}_{h,k-1}\Big[\big\langle \overline{P}_h^k(\cdot \,|\, s, a),\, \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \circ \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \big\rangle\Big] + 3H \\
&= \Big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \circ \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \Big\rangle + 3H. \qquad (85)
\end{aligned}$$

Taking (84) and (85) together with (83) yields

$$\begin{aligned}
\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{h,k-1}\Big(\widehat{q}_h^k(s)\Big) &\leq \sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\Big[\big(\widehat{q}_h^k(s)\big)^2\Big] - \Big(\sum_{k=1}^K \alpha_k^K \mathbb{E}_{h,k-1}\big[\widehat{q}_h^k(s)\big]\Big)^2 \\
&\leq \Big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \circ \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \Big\rangle - \Big(\big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \big\rangle\Big)^2 + 3H.
\end{aligned}$$

To finish up, substituting these into (82) and making use of the assumption (32) give

$$\begin{aligned}
&\Big|\overline{V}_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) - \Big(r_h^{\widehat{\mu}\cdot\widehat{\nu}}(s) + \big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot),\, \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \big\rangle\Big)\Big| \\
&\quad\leq 2c_\alpha \sqrt{\frac{\log^3 \frac{K}{\delta}}{KH}} \left[\Big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \circ \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \Big\rangle - \Big(\big\langle P_h^{\widehat{\mu}\cdot\widehat{\nu}}(s,\cdot), \overline{V}_{h+1}^{\widehat{\mu}\cdot\widehat{\nu}} \big\rangle\Big)^2\right] + (6c_\alpha + 4)\sqrt{\frac{H\log^3 \frac{K}{\delta}}{K}}
\end{aligned}$$

for any $s \in \mathcal{S}$, thus concluding the proof of the first claim (42a) of Lemma 4.

The second claim (42b) of Lemma 4 can be established using exactly the same argument, and hence we omit the proof here for the sake of brevity.

B.2.2 Proof of inequality (42c)

We then turn to the last advertised inequality (42c). Given that $\overline{r}_h(s) + \overline{P}_h(s,\cdot)\overline{V}_{h+1} \in [0, H-h+1]$ for all $s \in \mathcal{S}$, we can recall the definition (18b) of $\overline{V}_h$ to obtain

$$\Big|\overline{V}_h(s) - \big(\overline{r}_h(s) + \overline{P}_h(s,\cdot)\overline{V}_{h+1}\big)\Big| \leq \left|\sum_{k=1}^K \alpha_k^K \Big\langle \mu_h^k(\cdot \,|\, s),\, \overline{q}_h^k(s,\cdot)\Big\rangle + \overline{\beta}_{h,V}(s) - \big(\overline{r}_h(s) + \overline{P}_h(s,\cdot)\overline{V}_{h+1}\big)\right| \qquad (86)$$

for all $s \in \mathcal{S}$. The remaining analysis is dedicated to bounding the right-hand side of (86).

Let us begin with the following identity:

$$\begin{aligned}
\sum_{k=1}^K \alpha_k^K \Big\langle \mu_h^k(\cdot \,|\, s),\, \overline{q}_h^k(s,\cdot)\Big\rangle + \overline{\beta}_{h,V}(s) &= \sum_{k=1}^K \alpha_k^K \mathop{\mathbb{E}}_{a\sim\mu_h^k(s)}\Big[\overline{r}_h^k(s,a) + \overline{P}_h^k(\cdot \,|\, s,a)\overline{V}_{h+1}\Big] + \overline{\beta}_{h,V}(s) \\
&= \overline{r}_h(s) + \big\langle \overline{P}_h(s,\cdot), \overline{V}_{h+1}\big\rangle + \overline{\beta}_{h,V}(s), \qquad (87)
\end{aligned}$$

where we recall the definitions of $\overline{r}_h \in \mathbb{R}^S$ and $\overline{P}_h \in \mathbb{R}^{S\times S}$ in (41). The key step boils down to bounding the bonus term defined in (21), towards which first we claim that

$$\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{\mu_h^k(s)}\big(\overline{q}_h^k(s,\cdot)\big) \leq 2 + 2\Big[\overline{P}_h(s,\cdot)\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big) - \big(\overline{P}_h(s,\cdot)\overline{V}_{h+1}\big)^2\Big] \qquad (88)$$

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

$$\begin{aligned}\overline{\beta}_{h,V}(s) &= c_{\mathsf{b}}\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\sum_{k=1}^{K}\alpha_k^K\left\{\mathsf{Var}_{\mu_h^k(s)}\left(\overline{q}_h^k(s,\cdot)\right)+H\right\}\\ &\le 2c_{\mathsf{b}}\sqrt{\frac{\log^3\frac{KS(A+B)}{\delta}}{KH}}\left\{\overline{P}_h(s,\cdot)\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big)-\big(\overline{P}_h(s,\cdot)\overline{V}_{h+1}\big)^2+H\right\},\end{aligned} \tag{89}$$

where we have used the identity $\sum_{k=1}^K \alpha_k^K = 1$. Hence, we can readily establish the desired result (42c) by combining (89) with (87) and (86), provided that $c_3 > 0$ is sufficiently large.

It remains to justify the claim (88). Towards this end, we make the observation that

$$\begin{aligned}\mathsf{Var}_{\mu_h^k(s)}\big(\overline{q}_h^k(s,\cdot)\big) &\le 2\mathsf{Var}_{\mu_h^k(s)}\big(\overline{r}_h^k(s,\cdot)\big)+2\mathsf{Var}_{\mu_h^k(s)}\Big(\sum_{s'}\overline{P}_h^k(s'\mid s,\cdot)\overline{V}_{h+1}(s')\Big)\\ &\le 2+2\left[\sum_a \mu_h^k(a\mid s)\overline{P}_h^k(s,a)\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big)-\Big(\sum_a\mu_h^k(a\mid s)\overline{P}_h^k(s,a)\overline{V}_{h+1}\Big)^2\right],\end{aligned}$$

which results from $\|\overline{r}_h^k\|_\infty \le 1$ and the following relation:

$$\begin{aligned}\mathsf{Var}_{\mu_h^k(s)}\Big(\sum_{s'}\overline{P}_h^k(s'\mid s,\cdot)\overline{V}_{h+1}(s')\Big) &= \sum_a \mu_h^k(a\mid s)\Big(\overline{P}_h^k(\cdot\mid s,a)\overline{V}_{h+1}\Big)^2-\Big(\sum_a\mu_h^k(a\mid s)\overline{P}_h^k(\cdot\mid s,a)\overline{V}_{h+1}\Big)^2\\ &\le \sum_a \mu_h^k(a\mid s)\overline{P}_h^k(\cdot\mid s,a)\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big)-\Big(\sum_a\mu_h^k(a\mid s)\overline{P}_h^k(\cdot\mid s,a)\overline{V}_{h+1}\Big)^2.\end{aligned}$$

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

$$\begin{aligned}\sum_{k=1}^K\alpha_k^K\mathsf{Var}_{\mu_h^k(s)}\big(\overline{q}_h^k(s,\cdot)\big) &\le \sum_{k=1}^K\alpha_k^K\left\{2+2\left[\sum_a\mu_h^k(a\mid s)\overline{P}_h^k(s,a)\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big)-\Big(\sum_a\mu_h^k(a\mid s)\overline{P}_h^k(s,a)\overline{V}_{h+1}\Big)^2\right]\right\}\\ &\le 2+2\overline{P}_h(s,\cdot)\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big)-2\Big(\sum_{k=1}^K\alpha_k^K\sum_a\mu_h^k(a\mid s)\overline{P}_h^k(s,a)\overline{V}_{h+1}\Big)^2\\ &= 2+2\Big[\overline{P}_h(s,\cdot)\big(\overline{V}_{h+1}\circ\overline{V}_{h+1}\big)-\big(\overline{P}_h(s,\cdot)\overline{V}_{h+1}\big)^2\Big]\end{aligned}$$

as claimed.

B.3 Minimax lower bound

In this section, we formalize the minimax lower bound claimed in (26).

Theorem 3 (Minimax lower bound)**.** *Consider any $0 < \varepsilon \le c_1 H$ for some small enough constant $c_1 > 0$. Then one can construct a collection of Markov games $\{\mathcal{MG}_\theta \mid \theta \in \Theta\}$ such that*

$$\inf_{\widehat{\mu},\widehat{\nu}}\max_{\theta\in\Theta}\mathbb{P}^{\mathcal{MG}_\theta}\left\{\mathsf{NE\text{-}gap}\big(\widehat{\mu},\widehat{\nu}\big)>\varepsilon\right\}\ge\frac{1}{4}, \tag{90}$$

provided that the total sample size obeys

$$N \le \frac{c_2 H^4 S(A+B)}{\varepsilon^2} \tag{91}$$

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

Proof. Suppose without loss of generality that $A \geq B$. Let us begin by considering the special scenario with $B = 1$; in this case, computing the Nash-equilibrium reduces to finding the optimal policy of a single-agent MDP with S states and A 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 actions such that no algorithm can learn an ε-optimal policy with $o\big(\frac{H^4SA}{\varepsilon^2}\big)$ samples (Azar et al., 2013; Li et al., 2022). More precisely, for any given $0 < \varepsilon \leq c_1 H$ for some small enough constant $c_1 > 0$, one can construct a collection of MDPs $\{\mathcal{M}_\theta \mid \theta \in \Theta\}$ such that

$$\inf_{\widehat{\mu}} \max_{\theta \in \Theta} \mathbb{P}^{\mathcal{M}_\theta} \Big\{ \max_s \big(V_1^\star(s) - V_1^{\widehat{\mu}}(s)\big) > \varepsilon \Big\} \geq \frac{1}{4}, \tag{92}$$

with the proviso that the total sample size

$$N \leq \frac{c_2 H^4 SA}{\varepsilon^2} \tag{93}$$

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 B completely identical actions for the min-player; that is, to construct $\mathcal{MG}_\theta$, we take its reward function and probability transition kernel to be

$$r_h^{\mathcal{MG}_\theta}(s, a, b) = r_h^{\mathcal{M}_\theta}(s, a) \qquad \text{and} \qquad P_h^{\mathcal{MG}_\theta}(\cdot \mid s, a, b) = P_h^{\mathcal{M}_\theta}(\cdot \mid s, a) \tag{94}$$

for all $(s, a, b, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$. Evidently, finding the Nash-equilibrium of $\mathcal{MG}_\theta$ is equivalent to computing the optimal policy of $\mathcal{M}_\theta$, given the non-distinguishability of the actions of the min-player in $\mathcal{MG}_\theta$. This in turn immediately establishes the advertised lower bound. □

B.4 Freedman's inequality

In this section, we record the Freedman inequality for martingales (Freedman, 1975) with slight modification, which is a crucial concentration bound for our analysis.

Theorem 4. *Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \leq R \qquad \text{and} \qquad \mathbb{E}\Big[X_k \mid \{X_j\}_{j:j<k}\Big] = 0 \qquad \text{for all } k \geq 1$$

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

$$|Y_n| \leq \sqrt{8 W_n \log\frac{3n}{\delta}} + 5R \log\frac{3n}{\delta} \leq \kappa W_n + \Big(\frac{2}{\kappa} + 5R\Big)\log\frac{3n}{\delta}. \tag{95}$$

Proof. Suppose that $W_n \leq \sigma^2$ holds deterministically for some quantity σ^2. As has been demonstrated in Li et al. (2021a, Theorem 5), with probability at least $1 - \delta$ we have

$$|Y_n| \leq \sqrt{8 \max\Big\{W_n, \frac{\sigma^2}{2^K}\Big\} \log\frac{2K}{\delta}} + \frac{4}{3} R \log\frac{2K}{\delta} \tag{96}$$

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

$$|Y_n| \leq \sqrt{8 \max\big\{W_n, R^2\big\} \log\frac{4\log_2 n}{\delta}} + \frac{4}{3} R \log\frac{4\log_2 n}{\delta}$$

$$\leq \sqrt{8W_n \log \frac{3n}{\delta}} + \sqrt{8R^2 \log \frac{3n}{\delta}} + \frac{4}{3} R \log \frac{3n}{\delta}$$
$$\leq \sqrt{8W_n \log \frac{3n}{\delta}} + 5R \log \frac{3n}{\delta},$$

where we have used $4 \log_2 n \leq 3n$ for any integer $n \geq 1$. This establishes the first inequality in (95). The second inequality in (95) is then a direct consequence of the elementary inequality $2ab \leq a^2 + b^2$. □

References

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

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

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272. JMLR. org.

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

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

Beck, A. (2017). *First-order methods in optimization*. SIAM.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hansen, T. D., Miltersen, P. B., and Zwick, U. (2013). 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.

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

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

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

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

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

Kakade, S. (2003). *On the sample complexity of reinforcement learning*. PhD thesis, University of London.

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

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

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

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

Li, G., Chen, Y., Chi, Y., Gu, Y., and Wei, Y. (2021b). Sample-efficient reinforcement learning is feasible for linearly realizable MDPs with limited revisiting. *Advances in Neural Information Processing Systems*, 34:16671–16685.

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

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

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

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

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

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

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

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

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

Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, pages 286–295.

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

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

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

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

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

Shalev-Shwartz, S. (2007). Online learning: Theory, algorithms, and applications.

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

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

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

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

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018a). 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.

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

Sidford, A., Wang, M., Yang, L., and Ye, Y. (2020). 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.

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

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

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

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

Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*.

Wainwright, M. J. (2019b). Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*.

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

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

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

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

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

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

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

Yin, D., Hao, B., Abbasi-Yadkori, Y., Lazić, N., and Szepesvári, C. (2022). Efficient local planning with linear function approximation. In *International Conference on Algorithmic Learning Theory*, pages 1165–1192. PMLR.

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

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

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

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

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

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

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