Improved Sample Complexity for Markov Games via Variance-Aware Bandit Learning

Hanbin Zhou[1*] **Canzhe Zhao**[2*,†] **Shuai Li**[3,4‡]
[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] National University of Singapore
[3] The University of Hong Kong
[4] Shanghai Innovation Institute
hanbin.zhou@connect.hkust-gz.edu.cn canzhezhao@nus.edu.sg shuaili@hku.hk

Abstract

We study the problem of learning in multi-player general-sum Markov games. In the simulator setting, where the learner can sample the next state conditioned on an arbitrary state action pair, the previous best-known upper bound of the sample complexity for learning an ε-approximate coarse correlated equilibrium (CCE) is $\widetilde{O}(H^4 S \sum_{i=1}^m A_i/\varepsilon^2)$ simulator queries (Li et al., 2022), where H is the horizon, S is the number of states, and A_i denotes the number of actions for the i-th player. This work improves the upper bound to $\widetilde{O}(H^4 S \max_{i\in[m]} A_i/\varepsilon^2)$, matching the lower bound of $\Omega(H^4 S \max_{i\in[m]} A_i/\varepsilon^2)$ (Li et al., 2022). In the online setting, in which the learner can only sample a trajectory, the previous best-known sample complexity upper bound for learning CCE in multi-player general-sum Markov games is $\widetilde{O}(H^6 S \max_{i\in[m]} A_i/\varepsilon^2)$ transition observations (Song et al., 2022; Jin et al., 2024; Mao et al., 2022). In this work, we improve this to $\widetilde{O}(H^5 S \max_{i\in[m]} A_i/\varepsilon^2)$. To our knowledge, this is the tightest upper bound that breaks the curse of multi-agency. The core of our algorithmic design and analysis is a new gap decomposition and, in particular, a bandit algorithm with a high-probability empirical variance regret bound, which might be of independent interest.

1 Introduction

The thriving field of multi-agent reinforcement learning (MARL) studies how interacting agents make autonomous decisions in a shared environment. Recently, MARL has achieved remarkable empirical success in complex sequential tasks, including the game of Go (Silver et al., 2016), Poker (Brown and Sandholm, 2018), real-time strategy games (Vinyals et al., 2019), autonomous driving (Shalev-Shwartz et al., 2016), and robotics (Kober et al., 2013). To formally model competitive scenarios where agents have conflicting interests, Markov games (or stochastic games) (Shapley, 1953) have emerged as the standard mathematical framework.

In competitive MARL, solutions typically take the form of equilibrium strategy profiles. The most prominent concept is the Nash equilibrium (NE) (Nash, 1950), where no player can gain by unilaterally deviating from their strategy. However, finding an exact NE in general-sum multi-player games suffers from Polynomial Parity Arguments on Directed graphs (PPAD)-hardness and extreme computational expense (Daskalakis et al., 2009). Consequently, more computationally tractable

*Equal contribution.
†Project Leader.
‡Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

solution concepts like the coarse correlated equilibrium (CCE) (Moulin and Vial, 1978) are widely adopted. The CCE relaxes the strict independence requirements of NE by allowing players to coordinate their actions through a shared source of randomness.

In two-player zero-sum Markov games, a recent line of work (Bai and Jin, 2020; Xie et al., 2020; Liu et al., 2021) developed provably efficient algorithms based on optimistic value iteration for finding NE in the online exploration setting. In particular, Liu et al. (2021) achieve a sample complexity of $\widetilde{O}(H^3SAB/\varepsilon^2)$ episodes, where H is the horizon, S is the number of states, A and B denote the number of actions of the two players respectively. However, such centralized model-based algorithms usually suffer from the well-known "curse of multiagents" (Jin et al., 2024), a phenomenon where the joint action space is the Cartesian product of individual action spaces, often causing the required sample complexity to scale exponentially with the number of agents. A recent line of works (Song et al., 2022; Jin et al., 2024; Mao et al., 2022) provably addressed this issue by a decentralized model-free algorithm called V-learning (Bai et al., 2020). In particular, Song et al. (2022), Jin et al. (2024), and Mao et al. (2022) obtain a result of $\widetilde{O}(H^6S\max_{i\in[m]} A_i/\varepsilon^2)$ for learning CCE in multi-player general-sum Markov Games. Nevertheless, their algorithms output a non-Markov policy and are not truly optimal since there exists an $O(H^2)$ gap between their sample complexity guarantee and the lower bound $\Omega(H^4S\max_i A_i/\varepsilon^2)$. Assuming access to a generative model, by which one can sample from arbitrary state-action pairs, Li et al. (2022) attained minimax optimality with the sample complexity of $\widetilde{O}(H^4S\sum_{i=1}^m A_i/\varepsilon^2)$. However, compared to the lower bound $\Omega(H^4S\max_i A_i/\varepsilon^2)$ for the simulator setting (Li et al., 2022), their algorithm is likely suboptimal by a factor of $\frac{\sum_i A_i}{\max_i A_i}$.

Contributions. To address the limitations of existing studies for learning in Markov games, in this work, we make a further step towards tightening the sample complexity, which obtains the truly optimal result for learning CCE in general-sum multi-player Markov games in the simulator setting and improves upon the currently best-known result for the same problem in the online setting. The main technical ingredients in this work are a new gap decomposition technique for learning in Markov games, and a bandit algorithm with a high-probability regret bound that depends on the empirical variance of the realized losses, which we believe might be of independent interest. Our detailed contributions are as follows:

- **High-Probability Second-Order Regret Bound for Adversarial Bandits:** To achieve a sharper horizon dependence in our MARL algorithms, we require an adversarial bandit subroutine with a high-probability regret bound that scales with the empirical variance of the realized losses. To this end, we propose an online mirror descent algorithm featuring a novel correction bonus. While prior variance-style bounds (Bubeck et al., 2018; Wei and Luo, 2018; Lee et al., 2020a; Ito et al., 2020; Ito, 2021) either depend on unrealized losses or only have expected regret bounds, our specifically designed bonus structurally cancels out any dependence on unrealized losses in the regret bound. Consequently, our algorithm achieves a high-probability regret bound of $\widetilde{O}(\sqrt{AQ_n} + A^{3/4}n^{1/4})$ strictly in terms of the realized-loss variance Q_n. This provides the exact variance-aware mechanism required for our MARL algorithms—a second-order bound that, to our knowledge, is also new to the bandit learning literature.
- **Improved Sample Complexity under the simulator setting:** For the simulator setting, we propose a significantly simple V-Value-Iteration algorithm that learns a ε-approximate CCE in $\widetilde{O}(H^4S\max_i A_i/\varepsilon^2)$ samples in multi-player general-sum Markov games. Our approach successfully removes the complicated full-information weighted policy/value update structure and refines the sample complexity that scales with $\sum_{i=1}^m A_i$ by prior methods (Li et al., 2022), while maintaining optimal horizon dependency, effectively breaking the curse of multi-agency.
- **Near-Optimal Sample Complexity under the online setting:** For the online exploration setting, we propose a decentralized V-learning algorithm equipped with a variance-aware bonus. By integrating our novel second-order adversarial bandit subroutine—to our knowledge, the first of its kind used in decentralized online Markov games—into a stage-based framework, our algorithm learns an ε-approximate CCE within $\widetilde{O}(H^5S\max_i A_i/\varepsilon^2)$ samples. This reduces the sample complexity by a factor of H compared to the previous state-of-the-art results of $\widetilde{O}(H^6S\max_i A_i/\varepsilon^2)$ established by Song et al. (2022), Jin et al.

(2024), and Mao et al. (2022). Analytically, we also introduce an alternative gap decomposition technique different from the existing analysis for V-learning algorithms that eliminates the need to maintain a *noisy* adversarial bandit problem.

1.1 Related Works

Table 1: Comparison of MARL algorithms and lower bounds in different sampling protocols. Sample complexity counts one transition observation per simulator query and H transition observations per online episode. For comparison, the episode bounds in Jin et al. (2024) and Liu et al. (2021) are multiplied by H. Also, we remark that algorithm by Li et al. (2022) is not truly decentralized as it requires the agents to cooperate with each other for their sampling subroutine. In contrast, our algorithm for simulator setting could be carried out without even knowing the existence of other players (except that the output policy needs correlated action).

Algorithm	Sample Complexity	Decentralized	Markov	Sampling Protocol
Li et al. (2022)	$\widetilde{O}(H^4 S \sum_{i=1}^m A_i/\varepsilon^2)$	△	✓	Simulator
Ours (Theorem 4.2)	$\widetilde{O}(H^4 S \max_i A_i/\varepsilon^2)$	✓	✓	
Lower Bound	$\Omega(H^4 S \max_i A_i/\varepsilon^2)$	–	–	
Jin et al. (2024)	$\widetilde{O}(H^6 S \max_i A_i/\varepsilon^2)$	✓	✗	Online
Liu et al. (2021)	$\widetilde{O}(H^4 SAB/\varepsilon^2)$	✗	✓	
Ours (Theorem 5.1)	$\widetilde{O}(H^5 S \max_i A_i/\varepsilon^2)$	✓	✗	
Lower Bound	$\Omega(H^4 S \max_i A_i/\varepsilon^2)$	–	–	

An extensive line of research has focused on sample-efficient learning in Markov games with online exploration. Early works predominantly developed centralized learning methods that model the joint transition or Q-function values. In two-player zero-sum tabular Markov games, Xie et al. (2020) and Bai and Jin (2020) developed the first sample efficient algorithms for learning NE based on optimistic value-iteration. Liu et al. (2021) further builds on their works and achieves a sample complexity of $\widetilde{O}(H^4 SAB/\varepsilon^2)$, which is optimal in terms of the horizon H, size of state space S and the error tolerance ε.

However, for multi-player general-sum tabular Markov games, existing centralized model-based approaches based on value iteration or Q-learning inevitably suffer from the curse of multi-agency (Jin et al., 2024). To circumvent this barrier, decentralized algorithms based on the V-learning (Bai et al., 2020) were proposed. Initially introduced by Bai et al. (2020) for two-player zero-sum games, V-learning has been successfully extended to multi-player general-sum games (Song et al., 2022; Jin et al., 2024; Mao et al., 2022). However, the output policy of V-learning algorithms is non-Markov. Daskalakis et al. (2023) further make the output policy Markov, although at the cost of worse sample complexity that scales with ε^{-3}. This is later improved by Cui et al. (2023), Wang et al. (2023), and Dai et al. (2024) to ε^{-2}, but their dependencies on H,S and A remain suboptimal when specializing to tabular Markov games. Assuming access to a generative model, Li et al. (2022) achieved near-optimal sample complexity for learning NE in both two-player zero-sum and learning CCE in multi-player general-sum Markov games while returning a Markov policy.

In scenarios with enormous state spaces, as in the single-agent setting, function approximation is widely used to model the state space in MARL. Some early works apply centralized function approximation in Markov games (Xie et al., 2020; Chen et al., 2022; Jin et al., 2022; Huang et al., 2022; Ni et al., 2022), which will suffer from the curse of multi-agency. A recent line of works consider learning Markov games with decentralized linear function approximation (Cui et al., 2023; Wang et al., 2023; Fan et al., 2024; Dai et al., 2024) to break the curse of multi-agency.

2 Preliminaries

Notation. Throughout this paper, we establish the following standard notations. For any positive integer n, we denote the set $\{1, \ldots, n\}$ by $[n]$. For any finite set $\mathcal{X}$, we let $|\mathcal{X}|$ denote its cardinality, $\Delta(\mathcal{X})$ denote the probability simplex over $\mathcal{X}$. For any positive integer K, Δ_K represents the $(K-1)$-dimensional probability simplex over K available actions. For any two vectors $x, y \in \mathbb{R}^K$, $\langle x, y \rangle$ denotes their standard inner product. We use $\mathbb{I}\{\cdot\}$ to denote the indicator function, which evaluates to

1 if the condition inside is true and 0 otherwise. Finally, we use standard asymptotic notations $O(\cdot)$ and $\Omega(\cdot)$, and we use $\widetilde{O}(\cdot)$ to hide poly-logarithmic factors in the problem parameters and the failure probability p.

Markov Games. Formally, for some $m \in \mathbb{N}$, we define an m-player tabular episodic Markov game by a tuple $\left(H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[m]}, \mathbb{P}, \{r_i\}_{i\in[m]}\right)$ where $H \in \mathbb{N}_+$ denotes the length of each episode; $\mathcal{S}$ denotes the state space with $|\mathcal{S}| = S$; $\mathcal{A}_i$ denotes the action space of player i with $|\mathcal{A}_i| = A_i$ and $A = \max_{i\in[m]} A_i$. Let $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ be the joint action space of m players, $\mathbb{P} = \{\mathbb{P}_h : \mathcal{S}\times\boldsymbol{\mathcal{A}} \mapsto \Delta(\mathcal{S})\}_{h\in[H]}$ denotes the collection of transition kernels; $r_i = \{r_{i,h} : \mathcal{S}\times\boldsymbol{\mathcal{A}} \mapsto [0,1]\}_{h\in[H]}$ denotes the collection of reward functions of player i.

Policy and Value Function. A (Markov) policy $\pi_i = \{\pi_{i,h}\}_{h=1}^H$ for player $i \in [m]$ specifies a random strategy where at each step $h \in [H]$ and state $s \in \mathcal{S}$, the player selects an action according to the probability distribution $\pi_{i,h}(\cdot|s) \in \Delta(\mathcal{A}_i)$. A joint policy $\pi = (\pi_1, \ldots, \pi_m)$ is a tuple of policies of all m players. We use the subscript $-i$ to denote the collection of all players except player i, allowing us to write the joint policy as $\pi = (\pi_i, \pi_{-i})$. A joint policy π is called a product policy if all players choose their actions independently, denoted as $\pi = \pi_1 \times \cdots \times \pi_m$.

For any Markov joint policy π, we define the state-value function (V-function) as follows:

$$V_{i,h}^{\pi}(s) := \mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, a_{h'}) \;\middle|\; s_h = s\right],$$

where the expectation is taken over the randomness of the transition kernel P and the joint policy π. Similarly, the action-value function (Q-function) could be defined as:

$$Q_{i,h}^{\pi}(s, \boldsymbol{a}) := \mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, a_{h'}) \;\middle|\; (s_h, \boldsymbol{a}_h) = s, \boldsymbol{a}\right].$$

In the online setting, we also allow policies and deviations to depend on the observed history. A CCE deviation uses randomization independent of the opponents' correlation device. Initial values are defined by the same expected cumulative reward; values at later steps will refer to a specified initialization of the certified policy.

Best response and equilibrium. In competitive environments, we are interested in the highest possible reward a player can obtain if the policies of all other players are fixed. We define the best response value function for player i against the opponents' joint policy π_{-i} as:

$$V_{i,1}^{\dagger,\pi_{-i}}(s) := \max_{\pi_i'} V_{i,1}^{\pi_i'\times\pi_{-i}}(s). \tag{1}$$

This represents the optimal expected return player i can achieve by unilaterally adopting the best possible policy π_i' in response to π_{-i}.

Given the competing objectives of the players, we evaluate our learning algorithms based on their ability to find approximate equilibria. The two standard solution concepts are the Nash Equilibrium (NE) and the Coarse Correlated Equilibrium (CCE). Given the PPAD-hardness of computing NE in multi-player general-sum games, we only consider the CCE in this work. Formally, we define

Definition 2.1 (ε-Coarse Correlated Equilibrium)**.** For any $\varepsilon > 0$, given initial state s_1, an *joint* policy π is a ε-approximate CCE, if

$$\mathrm{CCEGap}(\pi) := \max_{i\in[m]}(V_{i,1}^{\dagger,\pi_{-i}} - V_{i,1}^{\pi})(s_1) \le \varepsilon.$$

While an NE requires the players to act independently, a CCE relaxes this requirement by permitting players to correlate their actions, often through a shared source of randomness.

Online and Simulator Setting. We now introduce and compare the two sampling protocol considered in our work. In the simulator setting, we assume access to a generative model, which is formally stated below. Note that this assumption is also made by Li et al. (2022) and Zhang et al. (2023).

Assumption 1 (Simulator)**.** The algorithm has access to a generative model (simulator) that takes a state-action-step tuple $(s, \boldsymbol{a}, h) \in \mathcal{S} \times \boldsymbol{\mathcal{A}} \times [H]$ as input, then returns rewards $r_{i,h}(s, \boldsymbol{a})$ for $i = 1, \ldots, m$, and an independent random sample of the next state s' from the true transition probability $\mathbb{P}_h(\cdot \mid s, \boldsymbol{a})$.

In the online setting, which makes the least stringent requirement for the environment, the players can only play a whole episode each time. That is, at each time, the transition probability of next state is conditioned on the current state and chosen action. And in both the simulator and online settings, our goal is to learn an ε-approximate equilibrium of the Markov game using as few total samples as possible.

3 Adversarial Bandit with Empirical Variance Bound

In this section we introduce an adversarial bandit algorithm with high-probability second-order regret bound, which is essential for a variance-aware bonus design in MARL algorithms in later sections.

Consider a multi-armed bandit problem with adversarial loss. At each round $j \in [n]$, the learner picks a strategy w^j over A available actions $a^j \in \mathcal{A}$ while the adversary decides a loss vector $\ell^j \in [0,1]^A$. Then the learner takes an action a^j that is sampled from distribution w^j and observes only the loss of the chosen action $\ell^j(a^j)$. We also define regret as

$$\mathrm{Regret}_\star^n := \sum_{j=1}^n \ell^j(a^j) - \min_{a \in \mathcal{A}} \sum_{j=1}^n \ell^j(a)\,.$$

Algorithm 1 uses online mirror descent (OMD) with the weighted log-barrier regularizer $\psi^j(w) = -\sum_{a \in \mathcal{A}} \frac{1}{\eta^j(a)} \ln w(a)$, where $\eta^j(a) > 0$ are coordinate learning rates. For any $w, w' \in \Delta_{\mathcal{A}}$, their Bregman divergence is $D_{\psi^j}(w, w') := \psi^j(w) - \psi^j(w') - \langle \nabla \psi^j(w'), w - w' \rangle$. This measures the excess above the first-order approximation at w', so OMD penalizes changes according to the regularizer's geometry. Its curvature grows as an action's probability approaches zero, making changes in small-probability coordinates more costly.

Algorithm 1 Adversarial Bandit with Empirical Variance Bound (Simplified)

Input: Initial learning rate η, confidence level δ
Define truncated simplex $\Omega = \{w \in \Delta(\mathcal{A}) : w(a) \geq 1/(An),\ \forall a \in \mathcal{A}\}$.
Initialize: $w^1(a) = 1/A$, $\rho^1(a) = 2A$, and $\eta^1(a) = \eta$ for all $a \in \mathcal{A}$.
Set $\kappa = e^{1/\ln(eAn)}$.
for $j = 1, \ldots, n$ **do**
  $\psi^j(w) \leftarrow -\sum_{a \in \mathcal{A}} \frac{1}{\eta^j(a)} \ln w(a)$.
  Choose the measurable scalar $m^j \in [0,1]$ from past observations.
  Sample $a^j \sim w^j$ and receive bandit feedback $\ell^j(a^j)$.
  Construct loss estimator $\hat{\ell}^j(a)$ and bonus $\hat{b}^j(a)$ as in Eq. (2), Eq. (3) respectively.
  Compute $w^{j+1} \leftarrow \arg\min_{w \in \Omega} \langle w, \hat{\ell}^j + \hat{b}^j \rangle + D_{\psi^j}(w, w^j)$.
  Update $\rho^{j+1}(a)$ and $\eta^{j+1}(a)$ for all $a \in \mathcal{A}$ as in Algorithm 5.
end for

Algorithm Design. Due to space constraints, we present the core mechanism as Algorithm 1, and defer the full pseudocode to Appendix E as Algorithm 5. Algorithm 5 follows the log-barrier OMD framework with a predictive loss estimator, commonly used to obtain small-loss regret bounds for adversarial bandits (Bubeck et al., 2018; Wei and Luo, 2018; Ito et al., 2020; Ito, 2021). At round j, a measurable scalar predictor $m^j \in [0,1]$ is used to construct the loss estimator $\hat{\ell}^j$

$$\hat{\ell}^j(a) := \frac{\mathbb{I}\{a = a^j\}}{w^j(a)}(\ell^j(a) - m^j) + m^j\,, \tag{2}$$

and correction term estimator $\hat{b}^j(a)$

$$\hat{b}^j(a) := \frac{\mathbb{I}\{a = a^j\}}{w^j(a)} b^j(a), \qquad b^j(a) := \eta\big(1 - (\ell^j(a) - m^j)^2\big)\iota_0. \tag{3}$$

Together with the learning-rate update rule of Lee et al. (2020a), the log-barrier regularizer enables a high-probability second-order regret bound.

Further we choose the correction term to be $b^j(a) := \eta\big(1-(\ell^j(a)-m^j)^2\big)\iota_0$ where $\iota_0 := 60C^2\ln(2CA/\delta)\ln(eAn)$, where $C = \lceil\log_2(16An)\rceil\lceil\log_2(256A^2n^3)\rceil$. Briefly, the negative quadratic term $-\eta(\ell^j(a)-m^j)^2\iota_0$ cancels out the fixed action related variance bound from the importance weighted estimator, which motivates the term *correction*. The design of the correction bonus is inspired by Wei and Luo (2018), where they also use it for achieving a certain variance regret bound, though defined differently. More details about the correction bonus could be found in Appendix E.

With the above configuration, we have the following high-probability guarantee of Algorithm 5.

Theorem 3.1. *Assume that $m^j \in [0,1]$ is a measurable scalar for every $j \in [n]$. For any $\delta \in (0,1)$, with a suitable choice of η, Algorithm 5 ensures that with probability at least $1-\delta$, $\mathrm{Regret}_\star^n \le \sqrt{A\sum_{j=1}^n(\ell^j(a^j)-m^j)^2}\iota_1 + A^{3/4}n^{1/4}\iota_1$, for $\iota_1 = 40\iota_0^4$, which is polylogarithmic in A, n, and $1/\delta$.*

Further, setting prediction term $m^1 = 0$ and $m^j = \frac{1}{j-1}\sum_{\tau<j}\ell^\tau(a^\tau)$ for $j \ge 2$, the predictor lies in $[0,1]$. We have the following result for Algorithm 5.

Corollary 3.2 (Main guarantee for Algorithm 5)**.** *With $m^1 = 0$ and $m^j = \frac{1}{j-1}\sum_{\tau<j}\ell^\tau(a^\tau)$ for $j \ge 2$ and a suitable choice of η, for any $\delta \in (0,1)$, Algorithm 5 ensures that with probability at least $1-\delta$, $\mathrm{Regret}_\star^n \le \sqrt{AQ_n}\iota_1 + 4A^{3/4}n^{1/4}\iota_1$, for $Q_n := \sum_{j=1}^n\big(\ell^j(a^j) - \frac{1}{n}\sum_{j'=1}^n\ell^{j'}(a^{j'})\big)^2$.*

The leading $\sqrt{AQ_n}$ term in Corollary 3.2 adapts to the empirical fluctuations of the realized losses. In our Markov-game algorithms, these losses are normalized reward-plus-value observations: if their empirical variance before normalization is W, then $Q_n = nW/H^2$ and

$$\frac{H}{n}\sqrt{AQ_n} = \sqrt{\frac{AW}{n}}.$$

This gives a variance-based bonus for the policy-update error, which can be controlled together with variance-dependent sampling errors using total-variance bounds, as in Bernstein-style RL analyses (Azar et al., 2017; Li et al., 2022). The additional $A^{3/4}n^{1/4}$ term is retained in the analysis and absorbed under the sample budgets and accuracy regimes in Theorems 4.2 and 5.1. The proofs of Theorem 3.1 and Corollary 3.2 are deferred to Appendix E due to limited space.

4 Learning CCE in Multi-Player General-Sum Markov Games in the Simulator Setting

In this section, we introduce the V-Value-Iteration algorithm (Algorithm 2), which could convert any adversarial bandit algorithm with suitable regret into a RL algorithm with the help of a generative model. We also establish theoretical guarantees for learning ε-approximate equilibrium. In particular, with the adversarial bandit subroutine instantiated as Algorithm 5, Algorithm 2 could attain minimax optimal sample complexity.

4.1 Algorithm Description

We now describe the proposed algorithm. Following the design of dynamic programming and value iteration, our algorithm calculates the V-value and output policy state-wise and recursively from step $h = H$ down to $h = 1$.

Specifically, for each $h \in [H]$, $s_h \in \mathcal{S}$, the algorithm instantiates a no-regret learning subroutine with total round T for all m players. Within each round t, one joint action $\boldsymbol{a}$ is sampled from the product policy $\pi_h^t(\cdot|s_h) = \prod_{i=1}^m \pi_{i,h}^t(\cdot|s_h)$, and then used for querying the generative model. After obtained the sampling result, we then construct a loss and feed it to the adversarial bandit subroutine to compute the next policy for this state-step pair. The clipping in Line 16 ensures that $0 \le \hat{V}_{i,h}(s) \le H-h+1$ for every (i,h,s): the terminal values are zero, and the rewards and bonuses are nonnegative. Consequently, $0 \le r_{i,h}(s,\boldsymbol{a}) + \hat{V}_{i,h+1}(s') \le H-h+1 \le H$. Thus the normalized

Algorithm 2 V-Value-Iteration

1: **Initialize:** $\hat{V}_{i,H+1}(s) = 0, \forall (i,s) \in [m] \times \mathcal{S}$
2: **for** $h = H, \ldots, 1$ **do**
3: **for** $s \in \mathcal{S}$ **do**
4: Initialize $\overline{v}_i(s), \overline{w}_i(s) \leftarrow 0$ for all $i \in [m]$.
5: **for** $t = 1, \ldots, T$ **do**
6: Compute $\pi_h^t(\cdot \mid s) \leftarrow \prod_{i=1}^m \pi_{i,h}^t(\cdot|s)$ and sample $\boldsymbol{a} \sim \pi_h^t(\cdot \mid s)$.
 `// Draw an independent sample from the generative model.`
7: Query the generative model with $(s, \boldsymbol{a}, h)$ and observe $\{r_{i,h}(s,\boldsymbol{a})\}_{i=1}^m$, $s' \sim \mathbb{P}_h(\cdot \mid s, \boldsymbol{a})$.
8: **for** $i \in [m]$ **do**
9: Update $\overline{v}_i(s) \leftarrow \overline{v}_i(s) + r_{i,h}(s,\boldsymbol{a}) + \hat{V}_{i,h+1}(s')$.
10: Update $\overline{w}_i(s) \leftarrow \overline{w}_i(s) + \big(r_{i,h}(s,\boldsymbol{a}) + \hat{V}_{i,h+1}(s')\big)^2$.
 `// Update policy with adversarial bandit subroutine.`
11: Update $\pi_{i,h}^{t+1}(\cdot \mid s) \leftarrow \text{ADV_BANDIT}\Big(\frac{H - r_{i,h}(s,\boldsymbol{a}) - \hat{V}_{i,h+1}(s')}{H}\Big)$.
12: **end for**
13: **end for**
14: **for** $i \in [m]$ **do**
15: Calculate $W_{i,h}(s) \leftarrow \overline{w}_i(s)/T - (\overline{v}_i(s)/T)^2$.
 `// Calculate the final value estimate for step-state pair (h,s).`
16: Calculate $\hat{V}_{i,h}(s) \leftarrow \min\{\overline{v}_i(s)/T + b_{i,h}(s), H - h + 1\}$.
17: Reset $\overline{v}_i(s), \overline{w}_i(s) \leftarrow 0$.
18: **end for**
19: **end for**
20: **end for**
21: **Return** $\hat{\pi} = \{\hat{\pi}_h\}_{h \in [H]}$, where $\hat{\pi}_h := \text{Unif}(\pi_h^1, \ldots, \pi_h^T)$.

loss in Line 11 lies in $[0,1]$, as required by the bandit subroutine. This is consistent with recent work (Li et al., 2022) but with a few significant improvements: In Li et al. (2022), the policy is updated using a full-information weighted FTRL subroutine. The requirement for full-information feedback makes the number of samples they need scale with $\sum_{i=1}^m A_i$ for each state, and thus worsens the sample complexity. In contrast, our algorithm removes the complicated weighted update structure and can adopt adversarial bandit algorithms as the subroutine for policy update (Line 11), which is the key for our algorithm to have a better sample complexity that only scales with $\max_{i \in [m]} A_i$. A more detailed comparison with Li et al. (2022) is provdided in Appendix A.

At the end of each T rounds, we compute $\hat{V}_{i,h}(s)$ by clipping the empirical average of $r_{i,h}(s,\boldsymbol{a}) + \hat{V}_{i,h+1}(s')$ plus the bonus $b_{i,h}(s)$ at $H - h + 1$ (Line 16).

For any $(i,h,s) \in [m] \times [H] \times \mathcal{S}$, we define the bonus value with two different options as follows:

$$b_{i,h}(s) = \begin{cases} 0\,, & \text{(Option I)} \\ \sqrt{A W_{i,h}(s)/T}\,\iota_1 + 4H(A/T)^{3/4}\iota_1 & \text{(Option II)}\,, \end{cases} \tag{4}$$

where for Option II,

$$W_{i,h}(s) = \frac{\overline{w}_i(s)}{T} - \left(\frac{\overline{v}_i(s)}{T}\right)^2 \tag{5}$$

is the empirical variance of the observations $r_{i,h}(s,\boldsymbol{a}) + \hat{V}_{i,h+1}(s')$. With Algorithm 5 as the subroutine, Option II matches the bonus to the empirical variance of the reward-plus-value observations through Corollary 3.2. The benefit comes from controlling the accumulation of these variances across steps. Lemma C.5 bounds the expected sum of $W_{i,h}(s_h)$ along the empirical transitions by $\widetilde{O}(H^2)$. Cauchy–Schwarz therefore gives a leading accumulated bonus of order $\widetilde{O}(\sqrt{H^3 A/T})$, improving the $\widetilde{O}(H^2\sqrt{A/T})$ scale of the variance-unaware gap bound by $\sqrt{H}$. This translates into a factor-H improvement in sample complexity. The guarantees for both bonus options are presented in Section 4.2.

After the training process, based on the policy trajectories from Algorithm 2, we construct a Markov policy $\hat{\pi}_h := \mathrm{Unif}(\pi_h^1, \ldots, \pi_h^T)$ which is a uniform mixture of past policies $\{\pi_h^t\}_{t=1}^T$. More specifically, at each step h, the output policy draws a fresh index $t_h \sim \mathrm{Unif}([T])$, independently of previous draws and shared by all players. Conditional on t_h and the observed state s_h, each player i independently samples $a_{i,h} \sim \pi_{i,h}^{t_h}(\cdot \mid s_h)$. Our construction of the policy follows the design in Li et al. (2022) but removes the need for playing a weighted mixture of $\{\pi^t\}_{t=1}^T$ and only need to uniformly sample one of them at each step.

4.2 Main Results

We now introduce the theoretical guarantee of Algorithm 2 for finding an ε-approximate CCE policy in multi-player general-sum games. The following theorems present the sample complexity bounds for Algorithm 2 with different bonus designs and adversarial bandit subroutines. Here T counts sampling rounds at each state-step pair, so the total number of simulator queries is HST. Our sample complexity bound improves over Li et al. (2022) and matches the minimax lower bound also established in Li et al. (2022), while significantly simplifying their analysis by dedicated gap decomposition techniques. The proof of Theorem 4.1 and Theorem 4.2 are deferred to Appendix C.1 and Appendix C.2 respectively due to limited space.

Theorem 4.1. *With $b_{i,h}(s) = 0$ and* ADV_BANDIT *instantiated by Exp3-IX (Neu, 2015), for any $\varepsilon \in (0, H]$, $p \in (0,1)$, set $T = \widetilde{\Theta}\left(H^4 A/\varepsilon^2\right)$. With probability at least $1-p$, the output policy $\hat{\pi}$ of Algorithm 2 is an ε-approximate CCE.*

Theorem 4.1 therefore gives a sample-complexity upper bound of $\widetilde{O}(H^5 SA/\varepsilon^2)$ simulator queries for learning an ε-approximate CCE, where $A = \max_{i\in[m]} A_i$. This result surprisingly shows that we could actually break the curse of multi-agency without complicated weighted structure and even a bonus when we have access to a generative model.

Next, we present that, with a variance-aware bonus and the adversarial bandit subroutine instantiated by Algorithm 5, we could further improve the horizon dependency of sample complexity.

Theorem 4.2. *With $b_{i,h}(s) = \sqrt{AW_{i,h}(s)/T}\iota_1 + 4H(A/T)^{3/4}\iota_1$ and* ADV_BANDIT *instantiated by Algorithm 5, for any $\varepsilon \in (0, \sqrt{H}]$, $p \in (0,1)$, set $T = \widetilde{\Theta}\left(H^3 A/\varepsilon^2\right)$. With probability at least $1-p$, the output policy $\hat{\pi}$ of Algorithm 2 is an ε-approximate CCE.*

Theorem 4.2 therefore gives a sample-complexity upper bound of $\widetilde{O}(H^4 SA/\varepsilon^2)$ simulator queries. This improves over the previous results of $\widetilde{O}(H^4 S\sum_{i=1}^m A_i/\varepsilon^2)$ for finding CCE in multi-player general-sum games (Li et al., 2022) by a factor of $\frac{\sum_{i=1}^m A_i}{\max_{i\in[m]} A_i}$ and matches the sample complexity lower bound of $\Omega(H^4 S \max_{i\in[m]} A_i/\varepsilon^2)$ (Li et al., 2022).

5 Learning CCE in Multi-Player General-Sum Markov Games in the Online Setting

In this section, we introduce our stage-based V-learning algorithm for learning CCE in multi-player general-sum Markov games, and establish its sample complexity guarantee.

5.1 Algorithm Description

We now describe the proposed algorithm. Generally, the policy and value update rule of our algorithm is based on specially defined stages. For each step-state pair (h, s), we divide the visitations into multiple stages. The length of these stages increases exponentially at a rate of roughly $(1 + 1/H)$ (Line 17). This helps to circumvent the complicated no-weighted-regret adversarial bandit subroutine entailed by learning rate $\alpha_t = \frac{H+1}{H+t}$ in previous V-learning algorithms (Song et al., 2022; Jin et al., 2024), allowing us to plug in the adversarial bandit algorithm established in Section 3 for a variance-aware bonus design and better sample complexity. The idea of stage-based update is first introduced in single-agent RL (Zhang et al., 2020) and later adopted to multi-agent setting by Mao et al. (2022).

Specifically, we update our optimistic value estimates $\overline{V}_{i,h}(s)$ only at the end of each stage (Line 16), using exclusively the samples collected during the most recent stage. For the bonus term $b_{i,h}^t(s)$,

Algorithm 3 Stage-Based V-Learning with Variance-Aware Bonus (i-th player's version)

1: **Initialize** $\lambda \leftarrow \min\{1/16H, \sqrt{SA/T}\}$.
2: Set $\check{v}_{i,h}(s), \check{w}_{i,h}(s), \check{N}_h(s) \leftarrow 0$, $e_h(s) \leftarrow H$, $\overline{V}_{i,h}(s) \leftarrow H-h+1$, for all (i,h,s).
3: Instantiate an adversarial bandit with horizon $e_h(s) = H$ and set $\pi_{i,h}(\cdot|s) = \mathbf{1}/A_i$ to its initial distribution, for all (i,h,s).
4: **for** $t = 1, \ldots, T$ **do**
5: Start with initial state s_1.
6: **for** $h = 1, 2, \ldots, H$ **do**
7: Take action $a_{i,h} \sim \pi^t_{i,h}(\cdot \mid s_h)$.
8: Receive reward $r_{i,h}(s_h, \boldsymbol{a}_h)$ and observe next state $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, \boldsymbol{a}_h)$.
// Update policy using adversarial bandit subroutine.
9: $\pi_{i,h}(\cdot \mid s_h) \leftarrow \text{ADV_BANDIT}\Big(a_{i,h}, \frac{H - r_{i,h}(s_h, \boldsymbol{a}_h) - \overline{V}_{i,h+1}(s_{h+1})}{H}\Big)$ on the (h, s_h)-th adversarial bandit.
// Update accumulators.
10: $\check{v}_{i,h}(s_h) \leftarrow \check{v}_{i,h}(s_h) + r_{i,h}(s_h, \boldsymbol{a}_h) + \overline{V}_{i,h+1}(s_{h+1})$.
11: $\check{w}_{i,h}(s_h) \leftarrow \check{w}_{i,h}(s_h) + \big(r_{i,h}(s_h, \boldsymbol{a}_h) + \overline{V}_{i,h+1}(s_{h+1})\big)^2$.
12: $\check{n} := \check{N}_h(s_h) \leftarrow \check{N}_h(s_h) + 1$.
// Reaching the end of a stage.
13: **if** $\check{n} = e_h(s_h)$ **then**
14: $W^t_{i,h}(s_h) \leftarrow \check{w}_{i,h}(s_h)/\check{n} - (\check{v}_{i,h}(s_h)/\check{n})^2$.
15: $b_{i,h}(s_h) \leftarrow 4\iota/(\lambda\check{n}) + \sqrt{AW^t_{i,h}(s_h)/\check{n}\iota_1} + 4H(A/\check{n})^{3/4}\iota_1$.
// Update value estimates.
16: $\overline{V}_{i,h}(s_h) \leftarrow \min\{\check{v}_{i,h}(s_h)/\check{n} + b_{i,h}(s_h), H-h+1\}$.
// Increase the length of next stage by roughly (1+1/H).
17: $e_h(s_h) \leftarrow \lfloor e_h(s_h)(1+1/H) \rfloor$.
// Reset the bandit subroutine and accumulators.
18: Restart the (h, s_h)-th adversarial bandit with new horizon $e_h(s_h)$.
19: $\check{v}_{i,h}(s_h), \check{w}_{i,h}(s_h), \check{N}_h(s_h) \leftarrow 0$.
20: **end if**
21: **end for**
22: **end for**
23: Output policy $\hat{\pi} := \text{Unif}(\hat{\pi}^1_1, \ldots, \hat{\pi}^T_1)$, for $\hat{\pi}^t_1$ defined in Algorithm 4.

we include a regret bound of $\sqrt{AW^t_{i,h}(s_h)/\check{n}\iota_1} + 4H(A/\check{n})^{3/4}\iota_1$ and a concentration bound of $4\iota/(\lambda\check{n})$ (Line 15). It also worth noting that, the latter part of the bonus adopts the Quasi-Optimism principle to remove the needs of maintaining more variance terms and simplify the analysis. The Quasi-Optimism principle is first proposed by Lee and Oh (2025) to remove the need to maintain a variance-type bonus while attaining a minimax optimal regret bound simultaneously in model-based single-agent algorithms. At each stage h, We do also clip the value estimator by $H-h+1$ after adding the bonus to ensure that every normalized loss in Line 9 belongs to $[0,1]$.

For the policy, the algorithm will instantiate an adversarial bandit algorithm at the start of each stage for each state s_h. And every time this state is visited within this stage, the algorithm feeds $\frac{H - r_{i,h}(s_h, \boldsymbol{a}_h) - \overline{V}_{i,h+1}(s_{h+1})}{H}$ to the adversarial bandit subroutine and use it to update policy $\pi_h(\cdot|s_h)$ (Line 9). This is consistent with the policy update rule in Song et al. (2022), Jin et al. (2024), and Mao et al. (2022). We instantiate the adversarial bandit subroutine with Algorithm 5. Within each stage, $Q_{\check{n}} = \check{n}W^t_{i,h}(s_h)/H^2$, so its leading average regret is proportional to $\sqrt{AW^t_{i,h}(s_h)/\check{n}}$, matching the variance term in the bonus. The analysis combines this control with squared-value telescoping and stage-counting bounds to limit the total bonus cost (Lemmas D.5 and D.3). Together with quasi-optimism, this gives leading CCE-gap contributions of order $\widetilde{O}(\lambda H^2 + H^2SA/(\lambda T))$. Balancing these terms yields the $\widetilde{O}(H^2\sqrt{SA/T})$ leading gap bound underlying Theorem 5.1, and hence the factor-H improvement in sample complexity over previous V-learning guarantees.

Algorithm 4 Certified policy $\hat{\pi}_h^t$

Input: The distribution trajectory specified by Algorithm 3 $\{\pi_{i,h}^t : i \in [m], h \in [H], t \in [T]\}$.
Input: $h \in [H]$, $t \in [T]$.
for $\ell = h, \ldots, H$ **do**
  Observe s_ℓ.
  Uniformly sample $j \in [\check{N}_\ell^t(s_\ell)]$.
  Set $t \leftarrow \check{t}_j$, where $\check{t}_j$ is the index of the episode such that s_ℓ is visited the j-th time in the previous stage.
  Take action $\boldsymbol{a}_\ell \sim \pi_\ell^t$.
end for

After the training process, we construct our output policy $\hat{\pi}$ as a uniform mixture of the certified policies $\{\hat{\pi}_1^t\}_{t=1}^T$ defined by Algorithm 4, i.e. $\hat{\pi} := \mathrm{Unif}(\hat{\pi}_1^1, \ldots, \hat{\pi}_1^T)$. At the start of each episode, draw one common index $J \sim \mathrm{Unif}([T])$ and execute the certified policy $\hat{\pi}_1^J$ throughout that episode. The design of certified policy $\hat{\pi}_1^t$ follows the one in Mao et al. (2022). At each step ℓ, the episode index is sampled using a common random seed from the last completed stage of (ℓ, s_ℓ) before the current episode index t. If no such stage exists, the players just play uniformly random actions independently for the remaining steps. The recursively updated episode index t carries information from earlier steps, so $\hat{\pi}$ is generally non-Markov. The improved online sample complexity in Theorem 5.1 retains this limitation of prior V-learning methods; achieving the same rate with a Markov output policy remains open.

5.2 Main Results

The sample complexity guarantee of Algorithm 3 for learning CCE in general-sum Markov games is presented in the following theorem. The proof of Theorem 5.1 could be found in Appendix D.1.

Theorem 5.1. *For any $p \in (0,1)$ and $\varepsilon \in (0, \sqrt{H}]$, set $T = \widetilde{\Theta}(H^4SA/\varepsilon^2)$. Then, with probability at least $1-p$, the output policy $\hat{\pi}$ of Algorithm 3 is an ε-approximate CCE.*

Here T counts episodes, so the total number of transition observations is $HT = \widetilde{\Theta}(H^5SA/\varepsilon^2)$.

The sample complexity bound in Theorem 5.1 improves over existing works (Song et al., 2022; Jin et al., 2024; Mao et al., 2022) by a factor of $O(H)$ for learning CCE in multi-player general-sum Markov games. To our knowledge, this is the best existing guarantee in the online setting, leaving only an $O(H)$ gap to the lower bound of $\Omega(H^3SA/\varepsilon^2)$ episodes.

6 Conclusion

In this paper, we have developed provably efficient MARL algorithms for learning CCE in general-sum Markov games in both the simulator and online settings. We first proposed an adversarial bandit algorithm with a variance-style regret bound that enables us to adopt a variance-aware bonus design and improve the horizon dependency when applied to our MARL algorithms, which might be of independent interest. In the simulator setting, our V-Value Iteration algorithm improves over existing works through both a simplified algorithmic design and a sharper sample complexity bound. In the online exploration setting, our variance-aware, stage-based V-learning algorithm reduce the sample complexity by a factor of H to improve upon the previous state-of-the-art results. An interesting future direction would be to further tighten the gap between the sample complexity lower bound and the upper bound established in our work in the online setting. Additionally, how to extend our variance-aware framework to learn correlated equilibrium in general-sum Markov games still remains open to future work.

Acknowledgments

The corresponding author Shuai Li is supported by National Natural Science Foundation of China (62376154).

References

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272.

Bai, Y. and Jin, C. (2020). Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 551–560.

Bai, Y., Jin, C., and Yu, T. (2020). Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, volume 33, pages 2159–2170.

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

Bubeck, S., Cohen, M. B., and Li, Y. (2018). Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 111–127.

Chen, Z., Zhou, D., and Gu, Q. (2022). Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 227–261.

Cui, Q., Zhang, K., and Du, S. (2023). Breaking the curse of multiagents in a large state space: RL in Markov games with independent linear function approximation. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2651–2652.

Dai, Y., Cui, Q., and Du, S. S. (2024). Refined sample complexity for Markov games with independent linear function approximation (extended abstract). In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1260–1261.

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. (2023). The complexity of Markov equilibrium in stochastic games. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4180–4234.

Fan, J., Han, Y., Zeng, J., Cai, J.-F., Wang, Y., Xiang, Y., and Zhang, J. (2024). RL in Markov games with independent function approximation: Improved sample complexity bound under the local access model. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2035–2043.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. (2022). Towards general function approximation in zero-sum Markov games. In *International Conference on Learning Representations*.

Ito, S. (2021). Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2552–2583.

Ito, S., Hirahara, S., Soma, T., and Yoshida, Y. (2020). Tight first- and second-order regret bounds for adversarial linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 2028–2038.

Jin, C., Liu, Q., Wang, Y., and Yu, T. (2024). V-Learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*, 49(4):2295–2322.

Jin, C., Liu, Q., and Yu, T. (2022). The power of exploiter: Provable multi-agent RL in large state spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10251–10279.

Kober, J., Bagnell, J., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274.

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. (2020a). Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 15522–15533.

Lee, C.-W., Luo, H., and Zhang, M. (2020b). A closer look at small-loss bounds for bandits with graph feedback. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2516–2564.

Lee, H. and Oh, M.-h. (2025). Minimax optimal reinforcement learning with quasi-optimism. In *International Conference on Learning Representations*.

Li, G., Chi, Y., Wei, Y., and Chen, Y. (2022). Minimax-optimal multi-agent RL in Markov games with a generative model. In *Advances in Neural Information Processing Systems*, volume 35, pages 15353–15367.

Liu, Q., Yu, T., Bai, Y., and Jin, C. (2021). A sharp analysis of model-based reinforcement learning with self-play. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7001–7010.

Mao, W., Yang, L., Zhang, K., and Basar, T. (2022). On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15007–15049.

Moulin, H. and Vial, J. P. (1978). Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3–4):201–221.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.

Neu, G. (2015). Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 28, pages 3168–3176.

Ni, C., Song, Y., Zhang, X., Jin, C., and Wang, M. (2022). Representation learning for general-sum low-rank Markov games. arXiv preprint arXiv:2210.16976.

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.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. (2020). Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8604–8613.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Song, Z., Mei, S., and Bai, Y. (2022). When can we learn general-sum Markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Wang, Y., Liu, Q., Bai, Y., and Jin, C. (2023). Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2793–2848.

Wei, C.-Y. and Luo, H. (2018). More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. (2020). Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3674–3682.

Zhang, K., Kakade, S. M., Basar, T., and Yang, L. F. (2023). Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24(175):1–53.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 15198–15207.

A Comparison with Li et al. (2022)

Our work (Algorithm 2) and Li et al. (2022) both use backward induction and return a Markov mixture of product policies. The main difference is the adversarial subroutine used in both algorithms.

Consider the subroutine instantiated for a fixed step-state pair, our work uses adversarial bandit algorithms that requires bandit loss feedback, while Li et al. (2022) uses weighted FTRL, which requires the information of the full loss vector. As a result, for the simulator querying process, our Algorithm 2 draws one joint action $\boldsymbol{a}_h^t \sim \prod_i \pi_{i,h}^t(\cdot \mid s_h)$ and makes one shared query, returning all players' rewards and one next state per round. Along with the proposed analysis of counterfactual transition samples, our approach thus gives a dependence of $\max_i A_i$ in the sample complexity. In comparison, in each round, Q-FTRL queries every action of each player against independently sampled opponents' actions, constructing a full sampled action-value vector for each player's weighted FTRL update (Li et al., 2022, Algorithm 2). This requires $\sum_{i=1}^m A_i$ simulator queries per round, which leads to the dependence of $\sum_{i=1}^m A_i$ in their guarantee.

Another difference is in the design we adopted for obtaining variance-dependent regret guarantees for the adversarial subroutine. Both methods use such guarantees to control the accumulation of errors across steps. The FTRL guarantee of Li et al. (2022) involves weighted within-round variances of the full sampled action-value vectors under the current policies. Corollary 3.2 provides a high-probability bandit guarantee governed by Q_n, which measures the empirical variation of realized scalar losses across rounds. Our analysis combines this guarantee with variance-dependent sampling-error bounds and the total-variance argument in Lemma C.5 to establish Theorem 4.2.

B Auxiliary Lemmas

Theorem B.1 (Azuma–Hoeffding inequality)**.** *Let $C > 0$ be a constant, and let $\{X_t\}_{t=1}^\infty$ be a martingale difference sequence with respect to filtration $\{\mathcal{F}_t\}_{t=1}^\infty$ such that $|X_t| \le C$ almost surely for all $t \in \mathbb{N}$. Then, for any fixed $n \in \mathbb{N}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$\sum_{t=1}^n X_t \le C\sqrt{2n \ln\left(\frac{1}{\delta}\right)}\,.$$

Proof. Conditional Hoeffding's lemma gives $\mathbb{E}[e^{\theta X_t} \mid \mathcal{F}_{t-1}] \le e^{\theta^2 C^2/2}$. Iteration and Markov's inequality give $\mathbb{P}(\sum_{t=1}^n X_t \ge z) \le \exp(-\theta z + n\theta^2 C^2/2)$. Minimize over $\theta > 0$ to obtain $\exp(-z^2/(2nC^2))$. □

Theorem B.2 (Freedman's inequality)**.** *Let $C > 0$ be a constant, and let $\{X_t\}_{t=1}^\infty$ be a martingale difference sequence with respect to filtration $\{\mathcal{F}_t\}_{t=1}^\infty$ such that $|X_t| \le C$ almost surely for all $t \in \mathbb{N}$. Then, for any fixed $n \in \mathbb{N}$ and $\lambda' \in (0, 1/C)$, and any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$\sum_{t=1}^n X_t \le \lambda' \sum_{t=1}^n \mathbb{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right] + \frac{4}{3\lambda'} \ln\left(\frac{1}{\delta}\right)\,.$$

Proof. For $|z| \le 1$, $e^z \le 1 + z + z^2$. Conditional expectation therefore gives $\mathbb{E}[e^{\theta X_t} \mid \mathcal{F}_{t-1}] \le \exp(\theta^2 \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}])$. Multiplication, iteration up to n, and Markov's inequality prove the statement. □

Theorem B.3 (Performance Difference Lemma)**.** *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \mathbb{P}, H)$ be an MDP, and let $\pi, \hat{\pi} \in \Pi_{\mathcal{M}}$ be two Markov policies. Let $\hat{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $h \in [H]$, be arbitrary functions. Set $\hat{V}_{H+1}(s) = 0$ for all $s \in \mathcal{S}$, and define, for $h \in [H]$,*

$$\hat{V}_h(s) := \sum_a \hat{\pi}(a \mid s)\hat{Q}_h(s,a)\,.$$

Then

$$V_1^\pi(s_1) - \hat{V}_1(s_1) = \mathbb{E}_{\mathbb{P},\pi}\left[\sum_{h=1}^H \left\langle \hat{Q}_h(s_h,\cdot),\, (\pi - \hat{\pi})(\cdot \mid s_h)\right\rangle\right]$$

$$+ \mathbb{E}_{\mathbb{P},\pi}\left[\sum_{h=1}^{H}\big(r_h + \mathbb{P}\hat{V}_{h+1} - \hat{Q}_h\big)(s_h, a_h)\right].$$

C Missing Proofs for Simulator Setting

Notation. Before we start, we first introduce some notation here. We denote by $\boldsymbol{a}_h^t$ the actual action taken by the players, in the t-th round of sampling for state $s_h \in \mathcal{S}$. Note that we omit the dependence of $s_h \in \mathcal{S}$ in the notation of $\boldsymbol{a}_h^t$ for brevity. We denote by $\hat{V}_{i,h}^t, b_{i,h}^t$ the $\hat{V}_{i,h}, b_h$ function at episode t. We also denote by $\hat{\mathbb{P}}_h^t$ the empirical probability transition kernel such that

$$\hat{\mathbb{P}}_h^t(s_{h+1}|s_h, \boldsymbol{a}_h) = \begin{cases} 1 & \text{if } s_{h+1} = s_{h+1,s_h,\boldsymbol{a}_h}^t \\ 0 & \text{else} \end{cases}$$

For the analysis, we work on an enlarged probability space with independent samples $s_{h+1,s,\boldsymbol{a}}^t \sim \mathbb{P}_h(\cdot \mid s, \boldsymbol{a})$ for all $(h, s, \boldsymbol{a}, t) \in [H] \times \mathcal{S} \times \boldsymbol{\mathcal{A}} \times [T]$, independent of the randomization of the bandit subroutines. Only the sample corresponding to the actual joint action $\boldsymbol{a}_h^t$ is observed by the players; the counterfactual samples define $\hat{\mathbb{P}}_h^t$ for unplayed actions without changing the joint law of the observed data.

Let $\mathcal{F}_h$ be the σ-algebra generated by all algorithmic randomness and transition samples, including the counterfactual samples, at levels $h' > h$. For $t = 0, \ldots, T$, let $\mathcal{F}_{h,s,t}$ be $\mathcal{F}_h$ augmented by all local randomness and samples at (h, s) through round t, where round 0 contains the initialization. Using independent randomization for the statewise bandit instances, their local histories are conditionally independent across $s \in \mathcal{S}$ given $\mathcal{F}_h$. These histories satisfy

$$\mathcal{F}_{h-1} = \bigvee_{s\in\mathcal{S}} \mathcal{F}_{h,s,T},$$

with $\mathcal{F}_H$ trivial. In particular, $\hat{V}_{i,h+1}$ is $\mathcal{F}_h$-measurable and $\pi_h^t(\cdot \mid s)$ is $\mathcal{F}_{h,s,t-1}$-measurable. Conditional on $\mathcal{F}_{h,s,t-1}$, the joint action has law $\pi_h^t(\cdot \mid s)$ and is independent of the next-state samples $\{s_{h+1,s,\boldsymbol{a}}^t\}_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}$. Concentration bounds are established uniformly over the relevant state and action indices before averaging under the possibly data-dependent evaluation policy; we do not condition the sampling process on that policy or on the good event.

We denote by $\mu_i^\dagger$ the best response policy of the i-th player against $\hat{\pi}_{-i}$,

$$\mu_i^\dagger = \arg\max_{\pi_i'} V_{i,1}^{\pi_i' \times \hat{\pi}_{-i}}(s_1)\,, \tag{6}$$

where $\mu_i^\dagger$ is selected by backward Bellman maximization with a fixed deterministic tie-breaking rule at every state and step. This gives a Markov best response optimal from every state-step pair. Its restriction to levels $h' > h$ depends only on $\hat{\pi}_{-i,h+1:H}$ and is therefore $\mathcal{F}_h$-measurable.

Also, for any $Q : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, $V : \mathcal{S} \mapsto \mathbb{R}$ and $\pi : \mathcal{S} \mapsto \Delta(\boldsymbol{\mathcal{A}})$, we define for simplicity of notation that

$$\begin{aligned} Q(s, \pi) :=& \mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s, a)] \\ \mathbb{V}^\pi V(s) :=& \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)}\left[V(s')^2\right] - \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim P(\cdot|s,a)}[V(s')]^2\,. \end{aligned}$$

With slight abuse of notation, we also use this operator for the variance defined below:

$$\begin{aligned} \mathbb{V}_h^\pi(r + V)(s) :=& \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim\mathbb{P}_h(\cdot|s,a)}\left[(r(s, a) + V(s'))^2\right] \\ & - \left(\mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim\mathbb{P}_h(\cdot|s,a)}[r(s, a) + V(s')]\right)^2. \end{aligned}$$

The variance is over both the action and the next state.

With above notation, the calculation rule of $\hat{V}_{i,h}$ at the end of loop for state $s_h \in \mathcal{S}$ can be written as following, for any $i \in [m], h \in [H]$,

$$\hat{V}_{i,h}(s_h) = \min\left\{\frac{1}{T}\sum_{t=1}^{T}(r_{i,h} + \hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h^t) + b_{i,h}(s_h),\ H - h + 1\right\}. \tag{7}$$

For each step h, let $\pi^t_{-i,h} = \prod_{j\neq i} \pi^t_{j,h}$ and $\hat{\pi}_{-i,h} := \mathrm{Unif}(\pi^1_{-i,h}, \ldots, \pi^T_{-i,h})$ denote the opponents' joint marginal of the component and output policies, respectively. Since the output policy draws a fresh mixture index at each step, for every $(i,h,s) \in [m] \times [H] \times \mathcal{S}$, we have

$$V^{\hat{\pi}}_{i,h}(s) = \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \mathbb{P}_h V^{\hat{\pi}}_{i,h+1}\big)(s, \pi^t_h),$$
$$V^{\mu_i^\dagger \times \hat{\pi}_{-i}}_{i,h}(s) = \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \mathbb{P}_h V^{\mu_i^\dagger \times \hat{\pi}_{-i}}_{i,h+1}\big)(s, \mu^\dagger_{i,h} \times \pi^t_{-i,h}),$$

with terminal values $V^{\hat{\pi}}_{i,H+1}(s) = V^{\mu_i^\dagger \times \hat{\pi}_{-i}}_{i,H+1}(s) = 0$.

CCEGap Decomposition We notice that the CCEGap of output policy $\hat{\pi}$ with respect to player i could be decomposed in the following way

$$V^{\dagger,\hat{\pi}_{-i}}_{i,1}(s_1) - V^{\hat{\pi}}_{i,1}(s_1) = \underbrace{V^{\dagger,\hat{\pi}_{-i}}_{i,1}(s_1) - \hat{V}_{i,1}(s_1)}_{\text{term-P}} + \underbrace{\hat{V}_{i,1}(s_1) - V^{\hat{\pi}}_{i,1}(s_1)}_{\text{term-Q}}. \tag{8}$$

In the following sections, we will use the above decomposition to bound the CCEGap.

C.1 Proof of Theorem 4.1

In this section, we prove the Theorem 4.1. And the key step is Lemma C.1, which establishes the desired CCE-Gap bound.

Lemma C.1. *With $b_{i,h}(s) = 0$ and* ADV_BANDIT *instantiated by Exp3-IX,, for any $p \in (0,1)$, $s_1 \in \mathcal{S}_1$, it holds with probability at least $1-p$ that the output policy $\hat{\pi}$ of Algorithm 2 satisfies*

$$\max_{i\in[m]} V^{\dagger,\hat{\pi}_{-i}}_{i,1}(s_1) - V^{\hat{\pi}}_{i,1}(s_1) \le \widetilde{O}\Big(\sqrt{H^4 A/T}\Big).$$

Before we start the proof, we claim that we use Exp3-IX(Algorithm 1 of Neu (2015)) as one choice of adversarial bandit subroutine in Algorithm 2. Since Exp3-IX is a standard procedure and we use it without modification, we only recall the regret guarantee needed in our proof. More precisely, for an adversarial bandit problem with losses $\ell^j \in [0,1]^A$ at round t, Exp3-IX outputs actions $a^t \in \mathcal{A}$ and satisfies the following high-probability regret bound.

Theorem C.2 (Theorem 1 of Neu (2015))**.** *Fix an arbitrary $\delta > 0$. With $\eta = 2\gamma = \sqrt{\frac{2\log A}{An}}$, Exp3-IX guarantees it holds with probability at least $1-\delta$ that, for any $a^\star \in \mathcal{A}$,*

$$\sum_{j=1}^{n}\ell^j(a^j) - \sum_{j=1}^{n}\ell^j(a^\star) \le 2\sqrt{2An\log A} + \left(\sqrt{\frac{2An}{\log A}} + 1\right)\log\left(\frac{2}{\delta}\right).$$

Proof of Lemma C.1. To prove the lemma, we use the decomposition established in Eq. (8) to bound term-P and term-Q respectively.

For term-P. First, we use a decomposition established in part (a) of Lemma C.7, which is inspired by value difference lemma in Shani et al. (2020), to get

$$\begin{aligned}
\text{term-P} &= V^{\dagger,\hat{\pi}_{-i}}_{i,1}(s_1) - \hat{V}_{i,1}(s_1) \\
&= \mathbb{E}_{\mathbb{P},\mu_i^\dagger\times\hat{\pi}_{-i}}\left[\frac{1}{T}\sum_{t,h}\big(r_{i,h} + \hat{\mathbb{P}}^t_h \hat{V}_{i,h+1}\big)\big(s_h, a_{i,h}\times \boldsymbol{a}^t_{-i,h}\big) - \frac{1}{T}\sum_{t,h}\big(r_{i,h} + \hat{\mathbb{P}}^t_h \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}^t_h)\right] \\
&\quad + \mathbb{E}_{\mathbb{P},\mu_i^\dagger\times\hat{\pi}_{-i}}\left[\sum_{h}(r_{i,h} + \mathbb{P}_h \hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h) - \frac{1}{T}\sum_{t,h}\big(r_{i,h} + \hat{\mathbb{P}}^t_h \hat{V}_{i,h+1}\big)\big(s_h, a_{i,h}\times \boldsymbol{a}^t_{-i,h}\big)\right].
\end{aligned}$$

For the first part of RHS, we notice that it could be seen as the expectation of a summation of H average regret. For any $(i,h,s_h) \in [m] \times [H] \times \mathcal{S}$, if we instantiate ADV_BANDIT as Exp3-IX, then by Theorem C.2, setting $\eta = 2\gamma = \sqrt{\frac{2\log A}{AT}}$, it holds with probability at least $1 - \frac{p}{3mHS}$ that,

$$\frac{1}{T}\sum_t \big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)\big(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t\big) - \frac{1}{T}\sum_t \big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h^t)$$
$$\le 2\sqrt{2H^2 A_i \log A_i / T} + H\left(\sqrt{\frac{2A_i}{T\log A_i}} + \frac{1}{T}\right)\log\left(\frac{6mHS}{p}\right).$$

For the second part of the RHS, fix $(i,h,s_h,a_{i,h})$ and write

$$X_t := \big(r_{i,h} + \mathbb{P}_h \hat{V}_{i,h+1}\big)(s_h, a_{i,h} \times \pi_{-i,h}^t) - \big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t).$$

With the filtration introduced above, $\{X_t\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{F}_{h,s_h,t}\}_{t=0}^T$. By definition, we have $\hat{V}_{i,h+1} \in [0, H-h+1]$, so $|X_t| \le H$. For any $(i,h,s_h,a_{i,h}) \in [m] \times [H] \times \mathcal{S} \times \mathcal{A}_i$, by the Azuma–Hoeffding inequality (Theorem B.1) and the uniform-mixture definition of $\hat{\pi}_h$, it holds with probability at least $1 - \frac{p}{3mHSA}$ that

$$\big(r_{i,h} + \mathbb{P}_h \hat{V}_{i,h+1}\big)(s_h, a_{i,h} \times \hat{\pi}_{-i,h}) - \frac{1}{T}\sum_t \big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)\big(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t\big)$$
$$\le \sqrt{2H^2 \ln\left(\frac{3mHSA}{p}\right)/T}.$$

Combining above two and taking union bound over all $(i,h,s_h,a_{i,h}) \in [m] \times [H] \times \mathcal{S} \times \mathcal{A}_i$, we know it holds with probability at least $1 - 2/3p$ that, for $\iota = 2\log(6mHSA/p)$,

$$\begin{aligned}\text{term-P} &\le \mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}}\left[\sum_h 5\sqrt{H^2 A\iota/T}\right] + \mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}}\left[\sum_h \sqrt{H^2\iota/T}\right] \\ &= \widetilde{O}\left(\sqrt{H^4 A/T}\right).\end{aligned}$$

For the term-Q. By part (c) of Lemma C.7,

$$\begin{aligned}\text{term-Q} &= \hat{V}_{i,1}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) \\ &= \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\frac{1}{T}\sum_{t,h}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h^t) - \sum_h \big(r_{i,h} + \mathbb{P}_h \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h)\right].\end{aligned}$$

For each fixed (i,h,s_h), since $\hat{V}_{i,h+1}$ is $\mathcal{F}_h$-measurable, the differences

$$\big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h^t) - \big(r_{i,h} + \mathbb{P}_h \hat{V}_{i,h+1}\big)(s_h, \pi_h^t)$$

form a martingale difference sequence with respect to $\{\mathcal{F}_{h,s_h,t}\}_{t=0}^T$ and are bounded in absolute value by H. Thus the Azuma–Hoeffding inequality (Theorem B.1) gives, with probability at least $1 - p/(3mHS)$,

$$\frac{1}{T}\sum_t \big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h^t) - \big(r_{i,h} + \mathbb{P}_h \hat{V}_{i,h+1}\big)(s_h, \hat{\pi}_h)$$
$$\le \sqrt{2H^2 \ln(3mHS/p)/T},$$

using the uniform-mixture definition of $\hat{\pi}_h$. Therefore, taking union bound over all $(i,h,s_h) \in [m] \times [H] \times \mathcal{S}$, it holds with probability at least $1 - p/3$ that, for $\iota = 2\log(6mHSA/p)$,

$$\text{term-Q} \le O\left(\sum_h \sqrt{H^2\iota/T}\right) = \widetilde{O}\left(\sqrt{H^4/T}\right).$$

Finally, combining the two bounds, we have

$$\begin{aligned}\mathrm{CCEGap}(\hat{\pi}) &= V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) \\ &= \underbrace{V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \hat{V}_{i,1}(s_1)}_{\text{term-P}} + \underbrace{\hat{V}_{i,1}(s_1) - V_{i,1}^{\hat{\pi}}(s_1)}_{\text{term-Q}} \\ &\le \widetilde{O}\Big(\sqrt{H^4 A/T}\Big)\,.\end{aligned}$$

which completes the proof. □

Proof of Theorem 4.1. By Lemma C.1, we have

$$\mathrm{CCEGap}(\hat{\pi}) \le \widetilde{O}\left(\sqrt{\frac{H^4 A}{T}}\right).$$

Choosing $T = \widetilde{\Theta}(H^4 A/\varepsilon^2)$ sampling rounds per state-step pair gives $\mathrm{CCEGap}(\hat{\pi}) \le \varepsilon$. The total number of simulator queries is then $HST = \widetilde{\Theta}(H^5 SA/\varepsilon^2)$, establishing the claimed sample-complexity upper bound. □

C.2 Proof of Theorem 4.2

In this section, we prove the Theorem 4.2. Before we start, let's define some intermediate value used in our analysis:

$$\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_h) := \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)\big(s_h, \mu_i^\dagger \times \boldsymbol{a}_{-i,h}^t\big) \tag{9}$$

$$\overline{V}_{i,h}^{\hat{\pi}}(s_h) := \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \overline{V}_{i,h+1}^{\hat{\pi}}\big)\big(s_h, \boldsymbol{a}_h^t\big) \tag{10}$$

Both intermediate value functions have terminal value zero at step $H+1$ and lie in $[0, H-h+1]$ by backward induction, since the rewards lie in $[0,1]$. By their backward recursions and the best-response selection above, both $\overline{V}_{i,h+1}^{\hat{\pi}}$ and $\overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}$ are $\mathcal{F}_h$-measurable. The concentration inequalities below may therefore be applied conditionally on $\mathcal{F}_h$ using the local filtration $\{\mathcal{F}_{h,s,t}\}_{t=0}^{T}$; integrating these conditional bounds gives the stated unconditional failure probabilities. We also define the empirical transition kernel $\overline{\mathbb{P}}_h$ such that for any $V_{h+1} : \mathcal{S}_{h+1} \mapsto \mathbb{R}$

$$\overline{\mathbb{P}}_h V_{h+1}(s_h) = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbb{P}}_h^t V_{h+1}(s_h, \boldsymbol{a}_h^t) \tag{11}$$

We note that the two intermediate terms $\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}$, $\overline{V}_{i,h}^{\hat{\pi}}$ are introduced for analytical purpose only, and are not necessary for players to know.

Here we establish a high probability good event in the following lemma.

Lemma C.3 (Good Event for Section C.2)**.** *Assume $T \ge 16H^2 A$ and set $\lambda := 1/\sqrt{HT}$. Define the following set of events*

$$\begin{aligned}\mathcal{E}_1(i,h,s_h) := \Big\{ &\frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)\big(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t\big) - \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h^t) \\ &\le \sqrt{A W_{i,h}(s_h)/T}\,\iota_1 + 4H(A/T)^{3/4}\iota_1, \forall a_{i,h} \in \mathcal{A}_i\Big\}\,, \\ \mathcal{E}_2(i,h,s_h) := \Big\{ &\left|\frac{1}{T}\sum_{t}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h, \boldsymbol{a}_h^t) - \big(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h, \hat{\pi}_h)\right| \\ &\le \frac{1}{T}\sum_{t}\lambda \mathbb{V}_h^{\pi^t}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h) + \frac{\iota}{\lambda T}\Big\}\,,\end{aligned}$$

$$\mathcal{E}_3(i,h,s_h) := \Big\{ \Big| \big(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)(s_h, \mu \times \hat{\pi}_{-i,h}) - \frac{1}{T}\sum_t \big(r_{i,h} + \hat{\mathbb{P}}_h^t \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)(s_h, \mu \times \boldsymbol{a}_{-i,h}^t) \Big| \leq \frac{1}{T}\sum_{t=1}^T \lambda \, \mathbb{V}_h^{\mu \times \pi_{-i}^t}\big(r_{i,h} + \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)(s_h) + \frac{\iota}{\lambda T}, \forall \mu \in \Delta(\mathcal{A}_i) \Big\}.$$

Furthermore, defining

$$\mathcal{E}_1 := \cap_{(i,h,s_h)\in[m]\times[H]\times\mathcal{S}} \mathcal{E}_1(i,h,s_h)\,,$$
$$\mathcal{E}_2 := \cap_{(i,h,s_h)\in[m]\times[H]\times\mathcal{S}} \mathcal{E}_2(i,h,s_h)\,,$$
$$\mathcal{E}_3 := \cap_{(i,h,s_h)\in[m]\times[H]\times\mathcal{S}} \mathcal{E}_3(i,h,s_h)\,,$$

and $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we see that for any $p \in (0,1)$, $\mathcal{E}$ holds with probability at least $1-p$, for $\iota = 8\ln(6mHSA/p)/3$ and ι_1 is the logarithmic term defined in Corollary 3.2.

We note that, throughout this section, we fix the single deterministic parameter $\lambda := 1/\sqrt{HT}$. The proof of above lemma is deferred to Appendix C.3. On the good event $\mathcal{E}$ in Lemma C.3, now we have the following results.

Lemma C.4 (Optimism)**.** *On the good event $\mathcal{E}$, for any $i \in [m]$, $h \in [H]$, and $s_h \in \mathcal{S}$, it holds that*

$$\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_h) \leq \hat{V}_{i,h}(s_h)\,.$$

Lemma C.5 (Total Variance)**.** *On the good event $\mathcal{E}$, if $T \geq 16H^2A$, it holds that*

$$\mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\sum_h \mathbb{V}_h^{\hat{\pi}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h)\right] \leq \widetilde{O}(H^2)\,,$$

and

$$\mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}}\left[\sum_h \mathbb{V}_h^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)(s_h)\right] \leq \widetilde{O}(H^2)\,,$$

and

$$\mathbb{E}_{\overline{\mathbb{P}}}\left[\sum_h W_{i,h}(s_h)\right] \leq \widetilde{O}(H^2)\,.$$

Again we defer the proof of above two lemmas to Appendix C.3. With those results in hand, we could have the following lemma, which is the key for proof of Theorem 4.2.

Lemma C.6. *With $b_{i,h}(s) = \sqrt{AW_{i,h}(s)/T}\iota_1 + 4H(A/T)^{3/4}\iota_1$ and* ADV_BANDIT *instantiated by Algorithm 5, for any $p \in (0,1)$, $s_1 \in \mathcal{S}_1$, it holds with probability at least $1-p$ that, if $T \geq 16H^2A$, the output policy $\hat{\pi}$ of Algorithm 2 satisfies*

$$\max_{i\in[m]} V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) \leq \widetilde{O}\Big(\sqrt{H^3A/T}\Big)\,.$$

Proof. Let $\mathcal{E}$ be the good event defined in Lemma C.3, which holds with probability at least $1-p$. All inequalities in the remainder of this proof hold conditioned on $\mathcal{E}$. First, using the CCEGap decomposition established in Eq. (8), we bound term-P and term-Q respectively.

For term-P. Recall the definition of $\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_h)$ in Eq. (9), by which we further decompose term-P into two terms as following

$$\begin{aligned}\text{term-P} &= V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \hat{V}_{i,1}(s_1) \\ &= \underbrace{V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \overline{V}_{i,1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_1)}_{P_1} + \underbrace{\overline{V}_{i,1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_1) - \hat{V}_{i,1}(s_1)}_{P_2}.\end{aligned}$$

Then for P_1, we have

$$
\begin{aligned}
P_1 &= V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \overline{V}_{i,1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_1) \\
&= \mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\sum_h \big(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)(s_h, \boldsymbol{a}_h) - \frac{1}{T} \sum_{t,h} \big(r_{i,h} + \hat{\mathbb{P}}_h^t \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)\big(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t\big) \right] \\
&\le \mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\frac{1}{T} \sum_{t,h} \left(\lambda \mathbb{V}^{\mu_i^\dagger \times \pi_{-i}^t} \big(r_{i,h} + \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)(s_h) + \frac{\iota}{\lambda T} \right) \right] \\
&\le \mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\sum_h \left(\lambda \mathbb{V}^{\mu_i^\dagger \times \hat{\pi}_{-i}} \big(r_{i,h} + \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)(s_h) + \frac{\iota}{\lambda T} \right) \right] \\
&\le \widetilde{O}\left(\lambda H^2 + \frac{H}{\lambda T} \right),
\end{aligned}
$$

where the second equality is by part (b) of Lemma C.7, the first inequality holds under subevent $\mathcal{E}_3$ of $\mathcal{E}$ in Lemma C.3, the second inequality is by law of total variance, i.e. $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X]) \ge \mathbb{E}[\mathrm{Var}(Y|X)]$, and the last inequality is by Lemma C.5.

For P_2, by Lemma C.4, we have $P_2 = \overline{V}_{i,1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_1) - \hat{V}_{i,1}(s_1) \le 0$

Substituting $\lambda = 1/\sqrt{HT}$, we obtain

$$
\begin{aligned}
\text{term-P} &= P_1 + P_2 \\
&\le \widetilde{O}\Big(\sqrt{H^3/T}\Big).
\end{aligned}
$$

For the term-Q. Recall the definition of $\overline{V}_{i,1}^{\hat{\pi}}(s_1)$ in Eq. (10), by which we further decompose term-Q into two terms as following

$$
\begin{aligned}
\text{term-Q} &= \hat{V}_{i,1}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) \\
&= \underbrace{\hat{V}_{i,1}(s_1) - \overline{V}_{i,1}^{\hat{\pi}}(s_1)}_{Q_1} + \underbrace{\overline{V}_{i,1}^{\hat{\pi}}(s_1) - V_{i,1}^{\hat{\pi}}(s_1)}_{Q_2}.
\end{aligned}
$$

For Q_1, Eq. (7) and the empirical Bellman recursion for $\overline{V}^{\hat{\pi}}$ imply

$$
\hat{V}_{i,h}(s) - \overline{V}_{i,h}^{\hat{\pi}}(s) \le b_{i,h}(s) + \overline{\mathbb{P}}_h(\hat{V}_{i,h+1} - \overline{V}_{i,h+1}^{\hat{\pi}})(s).
$$

Unrolling this inequality and using the zero terminal values gives

$$
\begin{aligned}
Q_1 &= \hat{V}_{i,1}(s_1) - \overline{V}_{i,1}^{\hat{\pi}}(s_1) \\
&\le \sum_h \mathbb{E}_{\overline{\mathbb{P}}}[b_{i,h}(s_h)] \\
&= \sum_h \mathbb{E}_{\overline{\mathbb{P}}} \left[\sqrt{AW_{i,h}(s_h)/T}\iota_1 + 4H(A/T)^{3/4}\iota_1 \right] \\
&\le \sum_h \mathbb{E}_{\overline{\mathbb{P}}} \left[\widetilde{O}\left(\lambda\sqrt{A}W_{i,h}(s_h) + \frac{\sqrt{A}}{\lambda T} + H(A/T)^{3/4} \right) \right] \\
&\le \widetilde{O}\left(\lambda H^2\sqrt{A} + \frac{H\sqrt{A}}{\lambda T} + H^2(A/T)^{3/4} \right) \\
&\le \widetilde{O}\left(\lambda H^2\sqrt{A} + \frac{H\sqrt{A}}{\lambda T} + \sqrt{H^3A/T} \right),
\end{aligned}
$$

where the second step is by definition of $\overline{\mathbb{P}}$ in Eq. (11), the third step is by definition of $b_{i,h}(s_h)$ in Eq. (4) , the fourth step follows from AM-GM with the fixed parameter λ, the fifth step is by the third inequality in Lemma C.5, the last step uses $T \geq 16H^2A$ in the statement of this lemma.

For Q_2, we have

$$\begin{aligned}
&\overline{V}^{\hat{\pi}}_{i,1}(s_1) - V^{\hat{\pi}}_{i,1}(s_1) \\
&= \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\frac{1}{T}\sum_{t,h}\big(r_{i,h} + \hat{\mathbb{P}}^t_h \overline{V}^{\hat{\pi}}_{i,h+1}\big)(s_h, \boldsymbol{a}^t_h) - \sum_{h}\big(r_{i,h} + \mathbb{P}_h \overline{V}^{\hat{\pi}}_{i,h+1}\big)(s_h, \boldsymbol{a}_h)\right] \\
&\leq \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\frac{1}{T}\sum_{t,h}\left(\lambda \mathbb{V}^{\pi^t}_h\big(r_{i,h} + \overline{V}^{\hat{\pi}}_{i,h+1}\big)(s_h) + \frac{\iota}{\lambda T}\right)\right] \\
&\leq \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\sum_{h}\left(\lambda \mathbb{V}^{\hat{\pi}}_h\big(r_{i,h} + \overline{V}^{\hat{\pi}}_{i,h+1}\big)(s_h) + \frac{\iota}{\lambda T}\right)\right] \\
&\leq \widetilde{O}\left(\lambda H^2 + \frac{H}{\lambda T}\right),
\end{aligned}$$

where the first step is by part (c) of Lemma C.7 with $b_{i,h} = 0$, the first inequality holds under the subevent $\mathcal{E}_2$ of $\mathcal{E}$, the second inequality is by the law of total variance and the uniform-mixture definition of $\hat{\pi}_h$ for every $h \in [H]$, the last step is by Lemma C.5.

Combining the bounds of Q_1 and Q_2 and substituting $\lambda = 1/\sqrt{HT}$, we have

$$\text{term-Q} = \hat{V}_{i,1}(s_1) - V^{\hat{\pi}}_{i,1}(s_1) \leq \widetilde{O}\left(\sqrt{\frac{H^3A}{T}}\right).$$

Finally, combining all above, we have

$$\begin{aligned}
\text{CCEGap}(\hat{\pi}) &= V^{\dagger,\hat{\pi}_{-i}}_{i,1}(s_1) - V^{\hat{\pi}}_{i,1}(s_1) \\
&= \underbrace{V^{\dagger,\hat{\pi}_{-i}}_{i,1}(s_1) - \hat{V}_{i,1}(s_1)}_{\text{term-P}} + \underbrace{\hat{V}_{i,1}(s_1) - V^{\hat{\pi}}_{i,1}(s_1)}_{\text{term-Q}} \\
&\leq \widetilde{O}\Big(\sqrt{H^3/T}\Big) + \widetilde{O}\Big(\sqrt{H^3A/T}\Big) \\
&\leq \widetilde{O}\Big(\sqrt{H^3A/T}\Big).
\end{aligned}$$

which completes the proof. □

Finally, by Lemma C.6, we could easily prove Theorem 4.2 as following.

Proof of Theorem 4.2. By Lemma C.6, we have

$$\text{CCEGap}(\hat{\pi}) \leq \widetilde{O}\left(\sqrt{\frac{H^3A}{T}}\right).$$

Since $\varepsilon \leq \sqrt{H}$, a sufficiently large constant in $T = \widetilde{\Theta}(H^3A/\varepsilon^2)$ also ensures $T \geq 16H^2A$. Therefore, this choice of T sampling rounds per state-step pair gives an ε-approximate CCE output policy using $HST = \widetilde{\Theta}(H^4SA/\varepsilon^2)$ simulator queries, establishing the claimed sample-complexity upper bound.

□

C.3 Further Details for the Simulator Setting

Lemma C.7 (Dedicated Performance Difference Lemma)**.** *For the output policy $\hat{\pi}$ and estimator $\hat{V}_{i,h}$ of Algorithm 2, intermediate term $\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}$, it holds that, for any nonnegative choice of bonus $b_{i,h}$, the following relations hold:*

(a)

$$
\begin{aligned}
&V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \hat{V}_{i,1}(s_1) \\
\leq &\mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\frac{1}{T} \sum_{t,h} (r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1})(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t) - \frac{1}{T} \sum_{t,h} (r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h^t) \right] \\
&+ \mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\sum_h (r_{i,h} + \mathbb{P} \hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h) - \frac{1}{T} \sum_{t,h} (r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1})(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t) \right].
\end{aligned}
$$

(b)

$$
\begin{aligned}
&V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \overline{V}_{i,1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_1) \\
= &\mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\sum_h (r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}})(s_h, \boldsymbol{a}_h) - \frac{1}{T} \sum_{t,h} (r_{i,h} + \hat{\mathbb{P}}_h^t \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}})(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^t) \right].
\end{aligned}
$$

(c)

$$
\begin{aligned}
&\hat{V}_{i,1}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) \\
\leq &\mathbb{E}_{\mathbb{P},\hat{\pi}} \left[\frac{1}{T} \sum_{t,h} (r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h^t) - \sum_h (r_{i,h} + \mathbb{P}_h \hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h) + \sum_h b_{i,h}(s_h) \right].
\end{aligned}
$$

*Both inequalities in **(a)** and **(c)** become equalities when $b_{i,h} \equiv 0$ for all i, h. Here the expectations $\mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}}$, $\mathbb{E}_{\mathbb{P},\hat{\pi}}$ are taken over the state-action trajectories $(s_h, \boldsymbol{a}_h)$ induced by their respective policies, and action $a_{i,h}$ explicitly denotes the i-th player's action sampled within those trajectories.*

Proof. We first prove **(a)**. Following a similar telescoping structure as the extended Performance Difference Lemma(Theorem B.3), we apply the definition of $\hat{V}_{i,1}(s_1)$ in Eq. (7)

$$
\begin{aligned}
&V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \hat{V}_{i,1}(s_1) \\
\leq &\mathbb{E}_{\mu_i^\dagger \times \hat{\pi}_{-i}} \left[(r_{i,1} + \mathbb{P}_1 V_{i,2}^{\dagger,\hat{\pi}_{-i}})(s_1, \boldsymbol{a}_1) \right] - \frac{1}{T} \sum_{t=1}^T (r_{i,1} + \hat{\mathbb{P}}_1^t \hat{V}_{i,2})(s_1, \boldsymbol{a}_1^t) \\
= &\mathbb{E}_{\mu_i^\dagger \times \hat{\pi}_{-i}} \left[(r_{i,1} + \mathbb{P}_1 V_{i,2}^{\dagger,\hat{\pi}_{-i}})(s_1, \boldsymbol{a}_1) - (r_{i,1} + \mathbb{P}_1 \hat{V}_{i,2})(s_1, \boldsymbol{a}_1) \right] \\
&+ \mathbb{E}_{\mu_i^\dagger \times \hat{\pi}_{-i}} \left[(r_{i,1} + \mathbb{P}_1 \hat{V}_{i,2})(s_1, \boldsymbol{a}_1) - \frac{1}{T} \sum_{t=1}^T (r_{i,1} + \hat{\mathbb{P}}_1^t \hat{V}_{i,2})(s_1, a_{i,1} \times \boldsymbol{a}_{-i,1}^t) \right] \\
&+ \mathbb{E}_{\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\frac{1}{T} \sum_{t=1}^T (r_{i,1} + \hat{\mathbb{P}}_1^t \hat{V}_{i,2})(s_1, a_{i,1} \times \boldsymbol{a}_{-i,1}^t) \right] - \frac{1}{T} \sum_{t=1}^T (r_{i,1} + \hat{\mathbb{P}}_1^t \hat{V}_{i,2})(s_1, \boldsymbol{a}_1^t) \\
= &\mathbb{E}_{\mathbb{P},\mu_i^\dagger \times \hat{\pi}_{-i}} \left[V_{i,2}^{\dagger,\hat{\pi}_{-i}}(s_2) - \hat{V}_{i,2}(s_2) \right] \\
&+ \mathbb{E}_{\mu_i^\dagger \times \hat{\pi}_{-i}} \left[(r_{i,1} + \mathbb{P}_1 \hat{V}_{i,2})(s_1, \boldsymbol{a}_1) - \frac{1}{T} \sum_{t=1}^T (r_{i,1} + \hat{\mathbb{P}}_1^t \hat{V}_{i,2})(s_1, a_{i,1} \times \boldsymbol{a}_{-i,1}^t) \right] \\
&+ \mathbb{E}_{\mu_i^\dagger \times \hat{\pi}_{-i}} \left[\frac{1}{T} \sum_{t=1}^T (r_{i,1} + \hat{\mathbb{P}}_1^t \hat{V}_{i,2})(s_1, a_{i,1} \times \boldsymbol{a}_{-i,1}^t) \right] - \frac{1}{T} \sum_{t=1}^T (r_{i,1} + \hat{\mathbb{P}}_1^t \hat{V}_{i,2})(s_1, \boldsymbol{a}_1^t),
\end{aligned}
$$

where the three terms on the right-hand side are, respectively, the expected difference $V_{i,2}^{\dagger,\hat{\pi}_{-i}}(s_2) - \hat{V}_{i,2}(s_2)$, the sampling error, and the bandit regret.

Next, we unroll the relationship recursively from $h = 1$ to H. Utilizing the terminal condition $V_{i,H+1}^{\dagger,\hat{\pi}_{-i}}(s) = \hat{V}_{i,H+1}(s) = 0$, for all $i \in [m]$, $s \in \mathcal{S}$, the step-wise errors accumulate to yield the final decomposition

$$
\begin{aligned}
&V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \hat{V}_{i,1}(s_1) \\
\leq &\mathbb{E}_{\mathbb{P},\mu_i^\dagger\times\hat{\pi}_{-i}}\left[\frac{1}{T}\sum_{t,h}(r_{i,h} + \hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h, a_{i,h}\times \boldsymbol{a}_{-i,h}^t) - \frac{1}{T}\sum_{t,h}(r_{i,h} + \hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h^t)\right] \\
&+ \mathbb{E}_{\mathbb{P},\mu_i^\dagger\times\hat{\pi}_{-i}}\left[\sum_{h}(r_{i,h} + \mathbb{P}_h\hat{V}_{i,h+1})(s_h, \boldsymbol{a}_h) - \frac{1}{T}\sum_{t,h}(r_{i,h} + \hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h, a_{i,h}\times \boldsymbol{a}_{-i,h}^t)\right].
\end{aligned}
$$

Then we prove **(b)**. Much similar to proof of **(a)**, we expand the difference term at $h = 1$ by adding and subtracting an intermediate term Then unroll the relationship recursively from $h = 1$ to H and by the fact that, $V_{i,H+1}^{\dagger,\hat{\pi}_{-i}}(s) = \overline{V}_{i,H+1}^{\mu_i^\dagger\times\hat{\pi}_{-i}}(s) = 0$, for all $i \in [m]$, $s \in \mathcal{S}$, we get

$$
\begin{aligned}
&V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - \overline{V}_{i,1}^{\mu_i^\dagger\times\hat{\pi}_{-i}}(s_1) \\
=&(r_{i,1} + \mathbb{P}_1 V_{i,2}^{\dagger,\hat{\pi}_{-i}})(s_1, \mu_{i,1}^\dagger \times \hat{\pi}_{-i,1}) \\
&- \frac{1}{T}\sum_{t=1}^{T}(r_{i,1} + \hat{\mathbb{P}}_1^t\overline{V}_{i,2}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_1, \mu_{i,1}^\dagger \times \boldsymbol{a}_{-i,1}^t) \\
=&\mathbb{E}_{\mu_i^\dagger\times\hat{\pi}_{-i}}\left[(r_{i,1} + \mathbb{P}_1 V_{i,2}^{\dagger,\hat{\pi}_{-i}})(s_1, \boldsymbol{a}_1) - (r_{i,1} + \mathbb{P}_1\overline{V}_{i,2}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_1, \boldsymbol{a}_1)\right] \\
&+ \mathbb{E}_{\mu_i^\dagger\times\hat{\pi}_{-i}}\left[(r_{i,1} + \mathbb{P}_1\overline{V}_{i,2}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_1, \boldsymbol{a}_1) - \frac{1}{T}\sum_{t=1}^{T}(r_{i,1} + \hat{\mathbb{P}}_1^t\overline{V}_{i,2}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_1, a_{i,1}\times \boldsymbol{a}_{-i,1}^t)\right] \\
=&\mathbb{E}_{\mathbb{P},\mu_i^\dagger\times\hat{\pi}_{-i}}\left[V_{i,2}^{\dagger,\hat{\pi}_{-i}}(s_2) - \overline{V}_{i,2}^{\mu_i^\dagger\times\hat{\pi}_{-i}}(s_2)\right] \\
&+ \mathbb{E}_{\mu_i^\dagger\times\hat{\pi}_{-i}}\left[(r_{i,1} + \mathbb{P}_1\overline{V}_{i,2}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_1, \boldsymbol{a}_1) - \frac{1}{T}\sum_{t=1}^{T}(r_{i,1} + \hat{\mathbb{P}}_1^t\overline{V}_{i,2}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_1, a_{i,1}\times \boldsymbol{a}_{-i,1}^t)\right] \\
=&\mathbb{E}_{\mathbb{P},\mu_i^\dagger\times\hat{\pi}_{-i}}\left[\sum_{h}(r_{i,h} + \mathbb{P}_h\overline{V}_{i,h+1}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_h, \boldsymbol{a}_h) - \frac{1}{T}\sum_{t,h}(r_{i,h} + \hat{\mathbb{P}}_h^t\overline{V}_{i,h+1}^{\mu_i^\dagger\times\hat{\pi}_{-i}})(s_h, a_{i,h}\times \boldsymbol{a}_{-i,h}^t)\right].
\end{aligned}
$$

For **(c)**, we use the definition of $\hat{V}_{i,h}(s_h)$ in Eq. (7) and unroll it:

$$
\begin{aligned}
&\hat{V}_{i,1}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) \\
\leq &\frac{1}{T}\sum_{t=1}^{T}(r_{i,1} + \hat{\mathbb{P}}_1^t\hat{V}_{i,2})(s_1, \boldsymbol{a}_1^t) + b_{i,1}(s_1) - \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[(r_{i,1} + \mathbb{P}_1 V_{i,2}^{\hat{\pi}})(s_1, \boldsymbol{a}_1)\right] \\
= &\frac{1}{T}\sum_{t=1}^{T}(r_{i,1} + \hat{\mathbb{P}}_1^t\hat{V}_{i,2})(s_1, \boldsymbol{a}_1^t) - \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[(r_{i,1} + \mathbb{P}_1\hat{V}_{i,2})(s_1, \boldsymbol{a}_1)\right] + b_{i,1}(s_1) \\
&+ \mathbb{E}_{\hat{\pi}}\left[(r_{i,1} + \mathbb{P}_1\hat{V}_{i,2})(s_1, \boldsymbol{a}_1)\right] - \mathbb{E}_{\hat{\pi}}\left[(r_{i,1} + \mathbb{P}_1 V_{i,2}^{\hat{\pi}})(s_1, \boldsymbol{a}_1)\right] \\
= &\frac{1}{T}\sum_{t=1}^{T}(r_{i,1} + \hat{\mathbb{P}}_1^t\hat{V}_{i,2})(s_1, \boldsymbol{a}_1^t) - \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[(r_{i,1} + \mathbb{P}_1\hat{V}_{i,2})(s_1, \boldsymbol{a}_1)\right] + b_{i,1}(s_1) \\
&+ \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\hat{V}_{i,2}(s_2) - V_{i,2}^{\hat{\pi}}(s_2)\right]
\end{aligned}
$$

$$\leq \mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\frac{1}{T}\sum_{t,h}(r_{i,h}+\hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h,\boldsymbol{a}_h^t)-\sum_h(r_{i,h}+\mathbb{P}_h\hat{V}_{i,h+1})(s_h,\boldsymbol{a}_h)+\sum_h b_{i,h}(s_h)\right].$$

□

Proof of Lemma C.3. **For subevent $\mathcal{E}_1$.** With ADV_BANDIT instantiated as Algorithm 5, for any $(i,h,s_h)\in[m]\times[H]\times\mathcal{S}$, Corollary 3.2 and $A_i\leq A$ give, with probability at least $1-p/(3mHS)$,

$$\mathrm{Regret}_\star^T=\sum_{t=1}^T\ell^t(a_{i,h}^t)-\min_{a_{i,h}\in\mathcal{A}_i}\sum_{t=1}^T\ell^t(a_{i,h})\leq\sqrt{AQ_T}\iota_1+4A^{3/4}T^{1/4}\iota_1,$$

for $\ell^t(a_{i,h})=(H-(r_{i,h}+\hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h,a_{i,h}\times\boldsymbol{a}_{-i,h}^t))/H$. By definition, we know $0\leq r_{i,h}+\hat{\mathbb{P}}_h^t\hat{V}_{i,h+1}\leq H-h+1\leq H$, and therefore $0\leq\ell^t(a_{i,h})\leq 1$ for all $a_{i,h}\in\mathcal{A}_i$. Recall the definition of $W_{i,h}$ in Eq. (5), we have

$$Q_T=\sum_{t=1}^T\left(\ell^t(a_{i,h}^t)-\frac{1}{T}\sum_{\tau=1}^T\ell^\tau(a_{i,h}^\tau)\right)^2=\frac{T}{H^2}W_{i,h}(s_h).$$

Here and throughout this proof, ι_1 is the factor in Corollary 3.2 evaluated with confidence parameter $\delta=p/(3mHS)$, $n=T$, and action count A; since $A_i\leq A$, this is a common upper bound for all players. Multiplying the regret bound by H/T, for any $a_{i,h}\in\mathcal{A}_i$ we obtain

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^T(r_{i,h}+\hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h,a_{i,h}\times\boldsymbol{a}_{-i,h}^t)-\frac{1}{T}\sum_{t=1}^T(r_{i,h}+\hat{\mathbb{P}}_h^t\hat{V}_{i,h+1})(s_h,\boldsymbol{a}_h^t)\\&\leq\iota_1\sqrt{AW_{i,h}(s_h)/T}+4H(A/T)^{3/4}\iota_1,\end{aligned}$$

Taking union bound over all $(i,h,s_h)\in[m]\times[H]\times\mathcal{S}$, we know the subevent $\mathcal{E}_1$ holds with probability at least $1-p/3$.

For subevent $\mathcal{E}_2$. Fix (i,h,s_h). By the same filtration argument, the differences

$$\big(r_{i,h}+\hat{\mathbb{P}}_h^t\overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h,\boldsymbol{a}_h^t)-\big(r_{i,h}+\mathbb{P}_h\overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h,\pi_h^t)$$

form a martingale difference sequence with respect to $\{\mathcal{F}_{h,s_h,t}\}_{t=0}^T$, bounded in absolute value by H, with conditional variances

$$\mathbb{V}_h^{\pi_h^t}\big(r_{i,h}+\overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h).$$

Apply Theorem B.2 to both signs of these differences, with the fixed parameter λ and failure allowance $p/(6mHS)$ per sign. Since $T\geq 16H^2A$, we have $\lambda\leq 1/(4H)<1/H$. Using the uniform-mixture definition of $\hat{\pi}_h$ and $4\ln(6mHS/p)/3\leq\iota$, a union bound over the two signs gives the bound defining $\mathcal{E}_2(i,h,s_h)$ with probability at least $1-p/(3mHS)$. Taking a union bound over (i,h,s_h) yields $\mathbb{P}(\mathcal{E}_2)\geq 1-p/3$.

For subevent $\mathcal{E}_3$. Fix $(i,h,s_h,a_{i,h})$. By the sampling construction above, the differences

$$\big(r_{i,h}+\mathbb{P}_h\overline{V}_{i,h+1}^{\mu_i^\dagger\times\hat{\pi}_{-i}}\big)(s_h,a_{i,h}\times\pi_{-i,h}^t)-\big(r_{i,h}+\hat{\mathbb{P}}_h^t\overline{V}_{i,h+1}^{\mu_i^\dagger\times\hat{\pi}_{-i}}\big)(s_h,a_{i,h}\times\boldsymbol{a}_{-i,h}^t)$$

form a martingale difference sequence for the same local filtration, bounded in absolute value by H, with conditional variances

$$\mathbb{V}_h^{a_{i,h}\times\pi_{-i,h}^t}\big(r_{i,h}+\overline{V}_{i,h+1}^{\mu_i^\dagger\times\hat{\pi}_{-i}}\big)(s_h).$$

Applying Theorem B.2 to both signs with the same fixed parameter λ and failure allowance $p/(6mHSA)$ per sign gives the bound defining $\mathcal{E}_3(i,h,s_h)$ for each pure action, using the uniform-mixture definition of $\hat{\pi}_{-i,h}$ and $4\ln(6mHSA/p)/3\leq\iota$. A union bound over all pure actions and all (i,h,s_h) makes these bounds simultaneous with probability at least $1-p/3$, since $A_i\leq A$. On this event, convex averaging over $a_{i,h}\sim\mu$ and the law of total variance prove $\mathcal{E}_3$ simultaneously for every $\mu\in\Delta(\mathcal{A}_i)$, including data-dependent choices such as $\mu_{i,h}^\dagger(\cdot\mid s_h)$. Combining the three subevents proves the lemma.

□

Proof of Lemma C.4. First, for any $(i,s) \in [m] \times \mathcal{S}$, by definition we have $\overline{V}_{i,H+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s) = \hat{V}_{i,H+1}(s) = 0$. We proceed by backward induction. Assume the claimed inequality holds at step $h+1$. If $\hat{V}_{i,h}(s_h) = H-h+1$, it holds trivially at step h, since $\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_h) \le H-h+1$. Otherwise $\hat{V}_{i,h}(s_h) < H-h+1$, so clipping is inactive and the update equals the empirical mean plus $b_{i,h}(s_h)$. Then, conditioning on the subevent $\mathcal{E}_1$ of $\mathcal{E}$, we know that

$$
\begin{aligned}
&\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}(s_h) - \hat{V}_{i,h}(s_h) \\
&= \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}\big)\big(s_h, \mu_i^\dagger \times \boldsymbol{a}_{-i,h}^t\big) - \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h^t) - b_{i,h}(s_h) \\
&\le \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)\big(s_h, \mu_i^\dagger \times \boldsymbol{a}_{-i,h}^t\big) - \frac{1}{T}\sum_{t=1}^{T}\big(r_{i,h} + \hat{\mathbb{P}}_h^t \hat{V}_{i,h+1}\big)(s_h, \boldsymbol{a}_h^t) - b_{i,h}(s_h) \\
&\le \sqrt{AW_{i,h}(s_h)/T\iota_1} + 4H(A/T)^{3/4}\iota_1 - b_{i,h}(s_h) \le 0\,,
\end{aligned}
$$

where the first step is by definition of $\overline{V}_{i,h}^{\mu_i^\dagger \times \hat{\pi}_{-i}}$ in Eq. (9), the second step is by induction hypothesis, the third step holds under subevent $\mathcal{E}_1$, and the last step is by definition of $b_{i,h}$ in Eq. (4). □

Proof of Lemma C.5. **For the first inequality.** By definition of variance and expanding the square, for any $(i,h,s_h) \in [m] \times [H] \times \mathcal{S}$,

$$
\begin{aligned}
&\mathbb{V}_h^{\hat{\pi}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h) \\
&= \mathbb{P}_h^{\hat{\pi}}\big(\overline{V}_{i,h+1}^{\hat{\pi}}\big)^2(s_h) - \Big(\big(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h, \hat{\pi}_h)\Big)^2 \\
&\quad + \big(r_{i,h}^2 + 2r_{i,h}\mathbb{P}_h \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h, \hat{\pi}_h) \\
&\le \mathbb{P}_h^{\hat{\pi}}\big(\overline{V}_{i,h+1}^{\hat{\pi}}\big)^2(s_h) - \big(\overline{V}_{i,h}^{\hat{\pi}}(s_h)\big)^2 \\
&\quad + \big(\overline{V}_{i,h}^{\hat{\pi}}(s_h)\big)^2 - \Big(\big(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h, \hat{\pi}_h)\Big)^2 + 2H,
\end{aligned}
$$

where the last step subtracts and adds the intermediate squared value and uses $0 \le r_{i,h} \le 1$ and $0 \le \overline{V}_{i,h+1}^{\hat{\pi}} \le H-h$. For the squared-value difference on the right-hand side, both values lie in $[0,H]$, so Eq. (10) gives

$$
\begin{aligned}
&\Big|\big(\overline{V}_{i,h}^{\hat{\pi}}(s_h)\big)^2 - \Big(\big(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h, \hat{\pi}_h)\Big)^2\Big| \\
&\le 2H\Big|\overline{V}_{i,h}^{\hat{\pi}}(s_h) - \big(r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h, \hat{\pi}_h)\Big| \\
&\le 2H\left(\lambda \mathbb{V}_h^{\hat{\pi}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h) + \frac{\iota}{\lambda T}\right) \\
&\le \frac{1}{2}\mathbb{V}_h^{\hat{\pi}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h) + \frac{2H\iota}{\lambda T},
\end{aligned}
$$

where the second inequality uses $\mathcal{E}_2$ and the law of total variance, and the last inequality uses $2H\lambda \le 1/2$.

This further leads to

$$
\begin{aligned}
&\mathbb{V}_h^{\hat{\pi}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h) \\
&\le \mathbb{P}_h^{\hat{\pi}}\big(\overline{V}_{i,h+1}^{\hat{\pi}}\big)^2(s_h) - \big(\overline{V}_{i,h}^{\hat{\pi}}(s_h)\big)^2 \\
&\quad + \frac{1}{2}\mathbb{V}_h^{\hat{\pi}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h) + \frac{2H\iota}{\lambda T} + 2H,
\end{aligned}
$$

which can be rearranged to yield

$$
\begin{aligned}
&\mathbb{V}_h^{\hat{\pi}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h) \\
&\le 2\left(\mathbb{P}_h^{\hat{\pi}}\big(\overline{V}_{i,h+1}^{\hat{\pi}}\big)^2(s_h) - \big(\overline{V}_{i,h}^{\hat{\pi}}(s_h)\big)^2\right) + \frac{4H\iota}{\lambda T} + 4H.
\end{aligned}
$$

Substituting this and combining terms gives

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\sum_h \mathbb{V}_h^{\hat{\pi}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\hat{\pi}}\big)(s_h)\right] \\
&\le 2\mathbb{E}_{\mathbb{P},\hat{\pi}}\left[\sum_h \Big(\mathbb{P}_h^{\hat{\pi}}\big(\overline{V}_{i,h+1}^{\hat{\pi}}\big)^2(s_h) - \big(\overline{V}_{i,h}^{\hat{\pi}}(s_h)\big)^2\Big)\right] \\
&\quad + \sum_h \left(\frac{4H\iota}{\lambda T} + 4H\right) \\
&= 2\mathbb{E}_{\mathbb{P},\hat{\pi}}\Big[\big(\overline{V}_{i,H+1}^{\hat{\pi}}(s_{H+1})\big)^2 - \big(\overline{V}_{i,1}^{\hat{\pi}}(s_1)\big)^2\Big] + \widetilde{O}\big(H^2/(\lambda T) + H^2\big) \\
&\le \widetilde{O}(H^2),
\end{aligned}$$

where the last step uses the zero terminal value and, since $T \ge 16H^2A$, the bound

$$\frac{1}{\lambda T} = \sqrt{H/T} \le 1.$$

For the second inequality. Following analogous telescoping variance argument as used for the first inequality, replacing the evaluation policy $\hat{\pi}$ with the best-response product policy $\mu_i^\dagger \times \hat{\pi}_{-i}$ and substituting the corresponding intermediate value function $\overline{V}_{i,h+1}^{\mu_i^\dagger \times \hat{\pi}_{-i}}$ (as defined in Eq. (9)), the same bound follows with $\lambda = 1/\sqrt{HT}$. Details are omitted here for brevity.

For the third inequality. In this paragraph, we denote by $\overline{v}_i(s)$ and $\overline{w}_i(s)$ their values in Algorithm 2 immediately before the accumulators are reset at (h, s). By the update rule of $\hat{V}_{i,h}$ in Eq. (7),

$$0 \le \hat{V}_{i,h}(s) - \frac{\overline{v}_i(s)}{T} \le b_{i,h}(s), \qquad \hat{V}_{i,h}(s) + \frac{\overline{v}_i(s)}{T} \le 2H.$$

Consequently, we have $\hat{V}_{i,h}(s)^2 - (\overline{v}_i(s)/T)^2 \le 2Hb_{i,h}(s)$. Also, $0 \le r_{i,h} \le 1$ and $0 \le \hat{V}_{i,h+1} \le H - h$ imply $r_{i,h}^2 + 2r_{i,h}\hat{V}_{i,h+1} \le 2H$. Expanding the $\overline{w}_i(s)$ term in $W_{i,h}$, we then obtain

$$\begin{aligned}
W_{i,h}(s) &= \frac{\overline{w}_i(s)}{T} - \left(\frac{\overline{v}_i(s)}{T}\right)^2 \\
&\le \overline{\mathbb{P}}_h(\hat{V}_{i,h+1}^2)(s) - \left(\frac{\overline{v}_i(s)}{T}\right)^2 + 2H \\
&\le \overline{\mathbb{P}}_h(\hat{V}_{i,h+1}^2)(s) - \hat{V}_{i,h}(s)^2 + 2H(1 + b_{i,h}(s)).
\end{aligned}$$

By the definition of the bonus and AM-GM,

$$\begin{aligned}
2Hb_{i,h}(s) &= 2H\iota_1\sqrt{AW_{i,h}(s)/T} + 8H^2(A/T)^{3/4}\iota_1 \\
&\le \frac{W_{i,h}(s)}{2} + \frac{2H^2A\iota_1^2}{T} + 8H^2(A/T)^{3/4}\iota_1.
\end{aligned}$$

Combining the two bounds and rearranging yields

$$\begin{aligned}
W_{i,h}(s) \le 2\Big[\overline{\mathbb{P}}_h(\hat{V}_{i,h+1}^2)(s) - \hat{V}_{i,h}(s)^2\Big] \\
+ 4H + \frac{4H^2A\iota_1^2}{T} + 16H^2(A/T)^{3/4}\iota_1.
\end{aligned}$$

Taking expectations along the empirical transition kernels $\overline{\mathbb{P}}_h$ and summing over h telescopes the squared-value terms. Since $\hat{V}_{i,H+1} = 0$,

$$\begin{aligned}
\mathbb{E}_{\overline{\mathbb{P}}}\left[\sum_{h=1}^H W_{i,h}(s_h)\right] &\le 4H^2 + \frac{4H^3A\iota_1^2}{T} + 16H^3(A/T)^{3/4}\iota_1 \\
&\le \widetilde{O}(H^2),
\end{aligned}$$

where the last inequality uses $T \ge 16H^2A$. Thus we prove the claimed bound. □

D Missing Proofs for Online Setting

Notation. Before we start, we first introduce some notation here. We denote by $(s_h^t, \boldsymbol{a}_h^t)$ the actual state-action pair observed and chosen at step h of episode t. For any step $h \in [H]$, state $s_h \in \mathcal{S}$, let $\pi_{i,h}^t(\cdot|s_h) \in \Delta(\mathcal{A}_i)$ be the mixed strategy picked by Algorithm 3 for player i at episode t and denote by $\pi_{i,h}^t = \{\pi_{i,h}^t(\cdot|s_h) : s_h \in \mathcal{S}\}$ the actual policy at step h of episode t. We denote by $\overline{V}_{i,h}^t, b_{i,h}^t$ the $\overline{V}_{i,h}, b_h$ functions at the start of episode t. We denote by $\check{N}_h^t(s)$ the number of times the state s has been visited within the stage right *before* the current stage of (h, s) at the beginning of episode t and $t_{j,h}^t(s)$ the index of the episode that state s was visited the j-th time within that previous stage. For simplicity of notation, we use $\check{n}, t_j$ to denote $\check{N}_h^t(s)$, $t_{j,h}^t(s)$, omitting the dependence on s, h and t. We also denote by $\hat{\mathbb{P}}_h^t$ the empirical probability transition kernel such that

$$\hat{\mathbb{P}}_h^t(s_{h+1}|s_h, \boldsymbol{a}_h) = \begin{cases} 1 & \text{if } s_{h+1} = s_{h+1,s_h^t,\boldsymbol{a}}^t \\ 0 & \text{else} \end{cases}$$

where it worth to note that we assume the existence of sampling event $s_{h+1,s_h^t,\boldsymbol{a}}^t \sim \mathbb{P}_h(\cdot|, s_h^t, \boldsymbol{a})$ for any $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$ at step h of episode t, even though only the result for state action pair $(s_h^t, \boldsymbol{a}_h^t)$ is observed by the players.

We denote by $\mu_i^\dagger$ a best response policy of the i-th player against $\hat{\pi}_{-i}$ at the initial state s_1 such that

$$\mu_i^\dagger = \arg\max_{\pi_i'} V_{i,1}^{\pi_i' \times \hat{\pi}_{-i}}(s_1)\,,$$

where the tie could be broken arbitrarily.

Also, for any $Q : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, $V : \mathcal{S} \mapsto \mathbb{R}$ and $\pi : \mathcal{S} \mapsto \Delta(\boldsymbol{\mathcal{A}})$, we define for simplicity of notation that

$$\begin{aligned} Q(s, \pi) :=& \mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s, a)] \\ \mathbb{V}_h^\pi V(s) :=& \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim\mathbb{P}_h(\cdot|s,a)}\left[V(s')^2\right] - \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim\mathbb{P}_h(\cdot|s,a)}[V(s')]^2\,. \end{aligned}$$

Specifically, with slight abuse of notation, for the variance operator $\mathbb{V}^\pi$, we might also use it as normal variance in later analysis such that $\mathbb{V}_h^\pi(r + V)(s) = \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim\mathbb{P}_h(\cdot|s,a)}\left[\left(r(s, a) + V(s')\right)^2\right] - \mathbb{E}_{a\sim\pi(\cdot|s),\, s'\sim\mathbb{P}_h(\cdot|s,a)}[r(s, a) + V(s')]^2$.

Using this notation, the update rule of $\overline{V}_{i,h}$ and $W_{i,h}$ in Algorithm 3 at episode t can be written as following, for any $i \in [m], h \in [H]$,$s_h \in \mathcal{S}$

$$\overline{V}_{i,h}^t(s_h) = \min\left\{\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\,\overline{V}_{i,h+1}^{t_j}\right)\left(s_h, \boldsymbol{a}_h^{t_j}\right) + b_{i,h}^t(s_h),\, H - h + 1\right\} \tag{12}$$

$$\begin{aligned} W_{i,h}^t(s_h) =& \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h}(s_h, \boldsymbol{a}_h^{t_j}) + \overline{V}_{i,h+1}^{t_j}(s_{h+1}^{t_j})\right)^2 \\ &- \left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h}(s_h, \boldsymbol{a}_h^{t_j}) + \overline{V}_{i,h+1}^{t_j}(s_{h+1}^{t_j})\right)\right)^2\,. \end{aligned} \tag{13}$$

For any (i, h, s, t) with $\check{n} > 0$, the value of the certified policy $\hat{\pi}_h^t$ and the best-response value for player i satisfy the following:

$$V_{i,h}^{\hat{\pi}_h^t}(s) = \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \mathbb{P}_h V_{i,h+1}^{\hat{\pi}_{h+1}^{t_j}}\right)(s, \pi_h^{t_j}) \tag{14}$$

$$V_{i,h}^{\dagger,\hat{\pi}_{-i,h}^t}(s) \le \max_{a_i\in\mathcal{A}_i}\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \mathbb{P}_h V_{i,h+1}^{\dagger,\hat{\pi}_{-i,h+1}^{t_j}}\right)(s, a_i \times \pi_{-i,h}^{t_j}). \tag{15}$$

The second relation holds because the deviating player's current action cannot depend on the fresh index j, while allowing the future response to be optimized separately for each t_j can only increase its value. Both value functions are zero at $h = H + 1$. When $\check{n} = 0$, $\hat{\pi}_h^t$ prescribes independent uniform actions for the remaining steps.

D.1 Proof of Theorem 5.1

For analytical purposes only, we define the following intermediate terms:

$$\overline{V}_{i,h}^{\dagger,t}(s_h) := \min\Bigg\{ H - h + 1, \max_{a_i \in \mathcal{A}_i} \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \left(r_{i,h} + \hat{\mathbb{P}}_h^{t_j} \overline{V}_{i,h+1}^{\dagger,t_j} \right) \left(s_h, a_i \times \boldsymbol{a}_{-i,h}^{t_j} \right) + \frac{3\iota}{\lambda \check{n}} \Bigg\}, \tag{16}$$

$$\underline{V}_{i,h}^{t}(s_h) := \max\left\{ \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \left(r_{i,h} + \hat{\mathbb{P}}_h^{t_j} \underline{V}_{i,h+1}^{t_j} \right) \left(s_h, \boldsymbol{a}_h^{t_j} \right) - b_{i,h}^t(s_h),\, 0 \right\}. \tag{17}$$

For any $(i, h, s) \in [m] \times [H] \times \mathcal{S}$, set $\overline{V}_{i,h}^{\dagger,t}(s) = H - h + 1$ and $\underline{V}_{i,h}^{t}(s) = 0$ in the first stage of (h, s). Both intermediate values are zero at $h = H + 1$.

To prove Theorem 5.1, we decompose the CCE gap at the initial state s_1, using the intermediate terms defined in Eq. (16) and Eq. (17) in following way

$$\begin{aligned}
& V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) \\
&= \frac{1}{T} \sum_{t=1}^{T} V_{i,1}^{\mu_i^\dagger \times \hat{\pi}_{-i,1}^t}(s_1) - \frac{1}{T} \sum_{t=1}^{T} V_{i,1}^{\hat{\pi}_1^t}(s_1) \\
&\le \frac{1}{T} \sum_{t=1}^{T} \underbrace{\left(V_{i,1}^{\dagger,\hat{\pi}_{-i,1}^t}(s_1) - \overline{V}_{i,1}^{t}(s_1) \right)}_{\phi_{i,1}^t(s_1)} + \frac{1}{T} \sum_{t=1}^{T} \underbrace{\left(\overline{V}_{i,1}^{t}(s_1) - \underline{V}_{i,1}^{t}(s_1) \right)}_{\delta_{i,1}^t(s_1)} \\
&\quad + \frac{1}{T} \sum_{t=1}^{T} \underbrace{\left(\underline{V}_{i,1}^{t}(s_1) - V_{i,1}^{\hat{\pi}_1^t}(s_1) \right)}_{\psi_{i,1}^t(s_1)},
\end{aligned} \tag{18}$$

where the first step is by the definition that $\hat{\pi} = \mathrm{Unif}(\hat{\pi}_1^1, \ldots, \hat{\pi}_1^T)$, the inequality is by bounding each component by its own best-response value.

For brevity, we denote each part in the above decomposition separately and define $\phi_{i,h}^t : \mathcal{S} \to \mathbb{R}$, $\delta_{i,h}^t : \mathcal{S} \to \mathbb{R}$, $\psi_{i,h}^t : \mathcal{S} \to \mathbb{R}$ by

$$\begin{aligned}
\phi_{i,h}^t(s) &:= V_{i,h}^{\dagger,\hat{\pi}_{-i,h}^t}(s) - \overline{V}_{i,h}^{t}(s), \\
\delta_{i,h}^t(s) &:= \overline{V}_{i,h}^{t}(s) - \underline{V}_{i,h}^{t}(s), \\
\psi_{i,h}^t(s) &:= \underline{V}_{i,h}^{t}(s) - V_{i,h}^{\hat{\pi}_h^t}(s).
\end{aligned}$$

Next, with the following lemmas, we bound the three parts of the CCE gap respectively.

Lemma D.1 (Quasi-Optimism). *On the good event $\mathcal{E}$ in Lemma D.4, for any $(i, h, s, t) \in [m] \times [H] \times \mathcal{S} \times [T]$, if $\lambda \le 1/16H$,*

$$\phi_{i,h}^t(s_h) \le \widetilde{O}(\lambda H^2).$$

Lemma D.2 (Quasi-Pessimism). *On the good event $\mathcal{E}$ in Lemma D.4, for any $(i, h, s, t) \in [m] \times [H] \times \mathcal{S} \times [T]$, if $\lambda \le 1/16H$,*

$$\psi_{i,h}^t(s_h) \le \widetilde{O}(\lambda H^2).$$

Lemma D.3 (Cost of Bonus). *On the good event $\mathcal{E}$ in Lemma D.4, for any $(i,h) \in [m] \times [H]$,*

$$\sum_{t=1}^{T} \delta_{i,h}^{t}(s_h^t) \le \widetilde{O}\left(\lambda H^2 T + \frac{H^2 SA}{\lambda} + H^{11/4}(SA)^{3/4}T^{1/4}\right).$$

Proofs of above lemmas are deferred to Appendix D.2.

Note that many previous works in both single-agent and multi-agent environment begin their analysis for optimistic (or pessimistic in multi-agent environment) estimators by showing that $V^{\dagger,\hat{\pi}^t_{-i,h}}(s) - \overline{V}(s) \le 0$ (or $\underline{V}(s) - V^{\hat{\pi}^t_h}(s) \le 0$ for optimism). Quasi-optimism principle proposed by Lee and Oh (2025) suggests that, by maintaining quasi-optimistic estimated values that are still optimistic but within a certain range, we could actually remove the need of maintaining an empirical variance and simplify the analysis while still achieving the same regret bound for single-agent setting. Though we still need to maintain an empirical variance by the regret bound of the adversarial bandit subroutine, we utilize the quasi-optimism principle in our bonus design and analysis of Lemma D.1, Lemma D.2, Lemma D.3.

Proof of Theorem 5.1. Let $\mathcal{E}$ be the good event in Lemma D.4, which holds with probability at least $1-p$. For the remainder of this proof, we condition on $\mathcal{E}$. For any player $i \in [m]$, by decomposition in Eq. (18) and Lemma D.1, Lemma D.2, Lemma D.3, we have

$$\begin{aligned} V_{i,1}^{\dagger,\hat{\pi}_{-i}}(s_1) - V_{i,1}^{\hat{\pi}}(s_1) &\le \frac{1}{T}\sum_{t=1}^{T}\phi_{i,1}^{t}(s_1) + \frac{1}{T}\sum_{t=1}^{T}\delta_{i,1}^{t}(s_1) + \frac{1}{T}\sum_{t=1}^{T}\psi_{i,1}^{t}(s_1) \\ &\le \widetilde{O}\left(\lambda H^2 + \frac{H^2 SA}{\lambda T} + H^{11/4}(SA/T)^{3/4}\right). \end{aligned}$$

Taking the maximum over $i \in [m]$ and using $\lambda = \min\{1/(16H), \sqrt{SA/T}\}$, we obtain

$$\begin{aligned} \mathrm{CCEGap}(\hat{\pi}) &\le \widetilde{O}\left(H^2\sqrt{\frac{SA}{T}} + \frac{H^3 SA}{T} + H^{11/4}\left(\frac{SA}{T}\right)^{3/4}\right) \\ &\le \widetilde{O}\left(\sqrt{\frac{H^4 SA}{T}}\right). \end{aligned}$$

The last step holds for $T \ge H^3 SA$, since

$$\frac{H^3 SA}{T} \le H^2\sqrt{\frac{SA}{T}}, \qquad H^{11/4}\left(\frac{SA}{T}\right)^{3/4} \le H^2\sqrt{\frac{SA}{T}}.$$

Since $\varepsilon \le \sqrt{H}$, the choice $T = \widetilde{\Theta}(H^4 SA/\varepsilon^2)$ ensures $T \ge H^3 SA$. With the logarithmic factors in the preceding bound included in this budget, we obtain $\mathrm{CCEGap}(\hat{\pi}) \le \varepsilon$ on $\mathcal{E}$, completing the proof. □

D.2 Proofs of Main Lemmas

In this section, we prove Lemma D.1, Lemma D.2 and Lemma D.3 respectively.

First, we establish a high probability good event $\mathcal{E}$ in the following lemma, which we will condition on in the proof of Lemma D.1, Lemma D.2 and Lemma D.3.

Lemma D.4 (Good event). *For $\check{n} > 0$ and fixed algorithm parameter $0 < \lambda \le 1/(16H)$, define the following set of events:*

$$\mathcal{E}_1(i,h,s_h,t) := \left\{ \begin{aligned} &\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{t_j}\right)\left(s_h, \mu \times \boldsymbol{a}_{-i,h}^{t_j}\right) \\ &- \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{t_j}\right)\left(s_h, \boldsymbol{a}_h^{t_j}\right) \end{aligned} \right.$$

$$\leq \sqrt{AW_{i,h}^t(s_h)/\check{n}}\iota_1 + 4H(A/\check{n})^{3/4}\iota_1, \forall \mu \in \Delta(\mathcal{A}_i)\Big\},$$

$$\mathcal{E}_2(i,h,s_h,t) := \Bigg\{ \left| \frac{1}{\check{n}}\sum_{j=1}^{\check{n}} \big(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\underline{V}_{i,h+1}^{t_j}\big)\big(s_h, \boldsymbol{a}_h^{t_j}\big) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}} \big(r_{i,h} + \mathbb{P}_h\underline{V}_{i,h+1}^{t_j}\big)\big(s_h, \pi_h^{t_j}\big) \right|$$

$$\leq \frac{\lambda}{\check{n}}\sum_{j=1}^{\check{n}} \mathbb{V}_h^{\pi_h^{t_j}}\big(r_{i,h} + \underline{V}_{i,h+1}^{t_j}\big)(s_h) + \frac{\iota}{\lambda\check{n}}\Big\},$$

$$\mathcal{E}_3(i,h,s_h,t) := \Bigg\{ \left| \frac{1}{\check{n}}\sum_{j=1}^{\check{n}} \big(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)\big(s_h, \mu \times \boldsymbol{a}_{-i,h}^{t_j}\big) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}} \big(r_{i,h} + \mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big)\big(s_h, \mu \times \pi_{-i}^{t_j}\big) \right|$$

$$\leq \frac{\lambda}{\check{n}}\sum_{j=1}^{\check{n}} \mathbb{V}_h^{\mu\times\pi_{-i}^{t_j}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h) + \frac{\iota}{\lambda\check{n}}, \forall \mu \in \Delta(\mathcal{A}_i)\Big\}.$$

For $\check{n} = 0$, let these three events be the whole sample space. Furthermore, defining

$$\mathcal{E}_1 := \cap_{(i,h,s_h,t)\in[m]\times[H]\times\mathcal{S}\times[T]}\, \mathcal{E}_1(i,h,s_h,t),$$
$$\mathcal{E}_2 := \cap_{(i,h,s_h,t)\in[m]\times[H]\times\mathcal{S}\times[T]}\, \mathcal{E}_2(i,h,s_h,t),$$
$$\mathcal{E}_3 := \cap_{(i,h,s_h,t)\in[m]\times[H]\times\mathcal{S}\times[T]}\, \mathcal{E}_3(i,h,s_h,t).$$

and $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we see that $\mathcal{E}$ holds with probability at least $1-p$, for $\iota = \frac{4}{3}\log\left(\frac{6mHSAT}{p}\right)$, ι_1 as defined in the Corollary 3.2.

Detailed proof of Lemma D.4 is deferred to Appendix D.3. Conditioning on the good event $\mathcal{E}$ in Lemma D.4, we have the following result.

Remark 1. For the notation of variance $\mathbb{V}$ used in the lemma below, fix the certified policy $\hat{\pi}_h^t$ and a family of value functions $\{V^{t_j}\}_{j=1}^{\check{n}}$, with $\check{n} > 0$. Draw $j \sim \mathrm{Unif}([\check{n}])$, then $\boldsymbol{a} \sim \pi_h^{t_j}(\cdot \mid s_h)$ and $s' \sim \mathbb{P}_h(\cdot \mid s_h, \boldsymbol{a})$. We use the shorthand

$$\mathbb{V}_h^{\hat{\pi}_h^t}(r_{i,h} + V^{t_j})(s_h) := \mathrm{Var}_{j,\boldsymbol{a},s'}\big(r_{i,h}(s_h,\boldsymbol{a}) + V^{t_j}(s')\big).$$

The same index j selects both the policy and the value function. For $\mu \times \hat{\pi}_{-i,h}^t$, replace the action law by $\mu \times \pi_{-i,h}^{t_j}$; the same convention applies to V^{t_j} alone.

Lemma D.5 (Variance Bound)**.** *On the good event $\mathcal{E}$ in Lemma D.4, with the fixed parameter $0 < \lambda \leq 1/(16H)$, for any $(i,h,s_h,t) \in [m]\times[H]\times\mathcal{S}\times[T]$ with $\check{n} > 0$ and any $a_i \in \mathcal{A}_i$, we have*

$$\begin{aligned}
&\frac{1}{\check{n}}\sum_{j=1}^{\check{n}} \mathbb{V}_h^{a_i\times\pi_{-i,h}^{t_j}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h) \\
&\leq \mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\big(r_{i,h} + \overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h) \\
&\leq \frac{4}{\check{n}}\sum_{j=1}^{\check{n}} \mathbb{P}_h\left(\overline{V}_{i,h+1}^{\dagger,t_j}\right)^2(s_h, a_i\times\pi_{-i,h}^{t_j}) - 4\left(\overline{V}_{i,h}^{\dagger,t}(s_h)\right)^2 \\
&\quad + 8H\left[\overline{V}_{i,h}^{\dagger,t}(s_h) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h, a_i\times\boldsymbol{a}_{-i,h}^{t_j})\right] \\
&\quad + \frac{8H\iota}{\lambda\check{n}} + 8H,
\end{aligned}$$

and

$$\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{V}_h^{\pi^{t_j}}\big(r_{i,h}+\underline{V}_{i,h+1}^{t_j}\big)(s_h)\le\mathbb{V}_h^{\hat{\pi}_h^t}\big(r_{i,h}+\underline{V}_{i,h+1}^{t_j}\big)(s_h)$$
$$\le\frac{4}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\big(\underline{V}_{i,h+1}^{t_j}\big)^2(s_h,\pi^{t_j})-4\big(\underline{V}_{i,h}^{t}(s_h)\big)^2$$
$$+\frac{8H\iota}{\lambda\check{n}}+8Hb_{i,h}^t(s_h)+O(H)\,,$$

We also have

$$W_{i,h}^t(s_h)\le\frac{4}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}_{i,h+1}^{t_j}(s_{h+1}^{t_j})^2-4\overline{V}_{i,h}^t(s_h)^2$$
$$+\frac{16H^2A\iota_1^2}{\check{n}}+\frac{32H\iota}{\lambda\check{n}}+32H^2(A/\check{n})^{3/4}\iota_1+8H\,.$$

Again, the proof of Lemma D.5 are deferred to Appendix D.3. Equipped with above results, we now can prove Lemma D.1, Lemma D.2 and Lemma D.3

Proof of Lemma D.1. We decompose

$$\phi_{i,h}^t(s_h)=\underbrace{V_{i,h}^{\dagger,\hat{\pi}_{-i,h}^t}(s_h)-\overline{V}_{i,h}^{\dagger,t}(s_h)}_{\phi_{i,h}^{1,t}(s_h)}+\underbrace{\overline{V}_{i,h}^{\dagger,t}(s_h)-\overline{V}_{i,h}^{t}(s_h)}_{\phi_{i,h}^{2,t}(s_h)}.$$

We bound these two terms separately.

For the first term. We prove by backward induction on h, simultaneously for all t and s_h, that

$$\phi_{i,h}^{1,t}(s_h)+4\lambda\left(\overline{V}_{i,h}^{\dagger,t}(s_h)\right)^2\le 8\lambda H(H-h+1). \tag{19}$$

At $h=H+1$, both terms on the left are zero. If $\overline{V}_{i,h}^{\dagger,t}(s_h)=H-h+1$, then $\phi_{i,h}^{1,t}(s_h)\le 0$, and the above inequality holds trivially. Otherwise $\check{n}>0$ and upper clipping is inactive. Let $a_i^\star$ be a maximizer of the right-hand side of (15), so

$$\begin{aligned}
&\phi_{i,h}^{1,t}(s_h)\\
&=V_{i,h}^{\dagger,\hat{\pi}_{-i,h}^t}(s_h)-\overline{V}_{i,h}^{\dagger,t}(s_h)\\
&\le\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\mathbb{P}_hV_{i,h+1}^{\dagger,\hat{\pi}_{-i,h+1}^{t_j}}\big)(s_h,a_i^\star\times\pi_{-i,h}^{t_j})-\overline{V}_{i,h}^{\dagger,t}(s_h)\\
&=\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\phi_{i,h+1}^{1,t_j}(s_h,a_i^\star\times\pi_{-i,h}^{t_j})\\
&\quad+\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h,a_i^\star\times\pi_{-i,h}^{t_j})-\overline{V}_{i,h}^{\dagger,t}(s_h)\\
&\le\lambda\mathbb{V}_h^{a_i^\star\times\hat{\pi}_{-i,h}^t}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)+\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\phi_{i,h+1}^{1,t_j}(s_h,a_i^\star\times\pi_{-i,h}^{t_j})+\frac{\iota}{\lambda\check{n}}\\
&\quad-\left[\overline{V}_{i,h}^{\dagger,t}(s_h)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h,a_i^\star\times\boldsymbol{a}_{-i,h}^{t_j})\right].
\end{aligned}$$

where the first inequality is by (15), and the last equality is by the definition of $\phi_{i,h+1}^{1,t_j}$. The last inequality holds under $\mathcal{E}_3$, together with the law of total variance.

Applying the first inequality in Lemma D.5 with $a_i = a_i^\star$ gives

$$\begin{aligned}
&\phi_{i,h}^{1,t}(s_h) + 4\lambda\left(\overline{V}_{i,h}^{\dagger,t}(s_h)\right)^2 \\
&\le \frac{1}{\check{n}}\sum_{j=1}^{\check{n}} \mathbb{P}_h\left(\phi_{i,h+1}^{1,t_j} + 4\lambda\left(\overline{V}_{i,h+1}^{\dagger,t_j}\right)^2\right)(s_h, a_i^\star \times \pi_{-i,h}^{t_j}) \\
&\quad + (8\lambda H - 1)\left[\overline{V}_{i,h}^{\dagger,t}(s_h) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h, a_i^\star \times \boldsymbol{a}_{-i,h}^{t_j})\right] \\
&\quad + (8\lambda H + 1)\frac{\iota}{\lambda\check{n}} + 8\lambda H \\
&\le \frac{1}{\check{n}}\sum_{j=1}^{\check{n}} \mathbb{P}_h\left(\phi_{i,h+1}^{1,t_j} + 4\lambda\left(\overline{V}_{i,h+1}^{\dagger,t_j}\right)^2\right)(s_h, a_i^\star \times \pi_{-i,h}^{t_j}) + 8\lambda H \\
&\le 8\lambda H(H - h + 1).
\end{aligned}$$

The second step uses $\lambda \le 1/(16H)$ and the definition of $\overline{V}_{i,h}^{\dagger,t}(s_h)$ in (16). Since upper clipping is inactive, the bracketed difference is at least $3\iota/(\lambda\check{n})$, and

$$8\lambda H - 1 \le 0, \qquad 3(8\lambda H - 1) + (8\lambda H + 1) = 32\lambda H - 2 \le 0.$$

The last step follows from the induction hypothesis, proving (19).

For the second term. We prove $\phi_{i,h}^{2,t}(s_h) \le 0$ by induction on t, simultaneously for all states and steps. In the first stage the two estimates coincide. If $\overline{V}_{i,h}^{t}(s_h) = H - h + 1$, the claim follows from the range of the intermediate. Otherwise, every preceding-stage index satisfies $t_j < t$, and the induction hypothesis, followed by $\mathcal{E}_1$, gives

$$\begin{aligned}
\overline{V}_{i,h}^{\dagger,t}(s_h) - \overline{V}_{i,h}^{t}(s_h) \le\ & \max_{a_i \in \mathcal{A}_i} \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{t_j}\big)(s_h, a_i \times \boldsymbol{a}_{-i,h}^{t_j}) \\
&- \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{t_j}\big)(s_h, \boldsymbol{a}_h^{t_j}) + \frac{3\iota}{\lambda\check{n}} - b_{i,h}^t(s_h) \\
&\le \sqrt{AW_{i,h}^t(s_h)/\check{n}}\,\iota_1 + 4H(A/\check{n})^{3/4}\iota_1 + \frac{3\iota}{\lambda\check{n}} - b_{i,h}^t(s_h) \le 0.
\end{aligned}$$

The last step uses the bonus definition. Combining the two terms and dropping the nonnegative squared term proves $\phi_{i,h}^t(s_h) \le 8\lambda H^2$. □

Proof of Lemma D.2. We prove by backward induction on h, simultaneously for all t and s_h, that

$$\psi_{i,h}^t(s_h) + 4\lambda\left(\underline{V}_{i,h}^t(s_h)\right)^2 \le O(\lambda H(H - h + 1)).$$

At $h = H + 1$, both terms on the left are zero. If $\underline{V}_{i,h}^t(s_h) = 0$, the left-hand side is $-V_{i,h}^{\hat{\pi}_h^t}(s_h) \le 0$; this also covers the first stage. Otherwise $\check{n} > 0$ and lower clipping is inactive, so

$$\begin{aligned}
&\psi_{i,h}^t(s_h) \\
&= \underline{V}_{i,h}^t(s_h) - V_{i,h}^{\hat{\pi}_h^t}(s_h) \\
&\le \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\underline{V}_{i,h+1}^{t_j}\right)\left(s_h, \boldsymbol{a}_h^{t_j}\right) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \mathbb{P}_h V_{i,h+1}^{\hat{\pi}_{h+1}^{t_j}}\right)\left(s_h, \pi_h^{t_j}\right) - b_{i,h}^t(s_h) \\
&= \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\underline{V}_{i,h+1}^{t_j}\right)\left(s_h, \boldsymbol{a}_h^{t_j}\right) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \mathbb{P}_h\underline{V}_{i,h+1}^{t_j}\right)\left(s_h, \pi_h^{t_j}\right) \\
&\quad + \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \mathbb{P}_h\underline{V}_{i,h+1}^{t_j}\right)\left(s_h, \pi_h^{t_j}\right) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(r_{i,h} + \mathbb{P}_h V_{i,h+1}^{\hat{\pi}_{h+1}^{t_j}}\right)\left(s_h, \pi_h^{t_j}\right) - b_{i,h}^t(s_h)
\end{aligned}$$

$$\leq \frac{\lambda}{\check{n}} \sum_{j=1}^{\check{n}} \mathbb{V}_h^{\pi^{t_j}} \left(r_{i,h} + \underline{V}_{i,h+1}^{t_j} \right)(s_h) + \frac{\iota}{\lambda \check{n}} + \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \mathbb{P}_h \psi_{i,h+1}^{t_j}(s_h, \pi_h^{t_j}) - b_{i,h}^t(s_h).$$

where the second step is by the definition of $\underline{V}_{i,h}^t(s_h)$ in Eq. (17), the last inequality holds under the subevent $\mathcal{E}_2$ of the good event $\mathcal{E}$ in Lemma D.4.

Applying the second inequality in Lemma D.5 gives

$$\begin{aligned}
\psi_{i,h}^t(s_h) + 4\lambda \left(\underline{V}_{i,h}^t(s_h)\right)^2 &\leq \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \mathbb{P}_h \left(\psi_{i,h+1}^{t_j} + 4\lambda \left(\underline{V}_{i,h+1}^{t_j} \right)^2 \right)(s_h, \pi_h^{t_j}) \\
&\quad + \frac{8H\iota}{\check{n}} + \frac{\iota}{\lambda \check{n}} + (8\lambda H - 1) b_{i,h}^t(s_h) + O(\lambda H) \\
&\leq \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \mathbb{P}_h \left(\psi_{i,h+1}^{t_j} + 4\lambda \left(\underline{V}_{i,h+1}^{t_j} \right)^2 \right)(s_h, \pi_h^{t_j}) + O(\lambda H) \\
&\leq O(\lambda H (H - h + 1)).
\end{aligned}$$

The second step uses $\lambda \leq 1/(16H)$ and $b_{i,h}^t(s_h) \geq 4\iota/(\lambda \check{n})$ from Line 15 of Algorithm 3, since

$$\frac{8H\iota}{\check{n}} + \frac{\iota}{\lambda \check{n}} + (8\lambda H - 1) b_{i,h}^t(s_h) \leq (40\lambda H - 3) \frac{\iota}{\lambda \check{n}} \leq -\frac{\iota}{2\lambda \check{n}} \leq 0.$$

The last step follows from the induction hypothesis, with an implicit constant uniform in h. Dropping the nonnegative squared-value term proves the claim. □

Proof of Lemma D.3. Recall the definition of $\overline{V}_{i,h}^t$ and $\underline{V}_{i,h}^t$ in Eq. (12) (17), we have $\delta_{i,h}^t(s_h) = H - h + 1 \leq H$ in the first stage. For later stages, upper clipping can only decrease $\overline{V}$, and lower clipping can only increase $\underline{V}$, so the following bound holds, no matter either clip is active or not:

$$\begin{aligned}
\delta_{i,h}^t(s_h) &= \overline{V}_{i,h}^t(s_h) - \underline{V}_{i,h}^t(s_h) \\
&\leq 2b_{i,h}^t(s_h) + \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \delta_{i,h+1}^{t_j}(s_{h+1}^{t_j}) \\
&= \frac{8\iota}{\lambda \check{n}} + 2\sqrt{A W_{i,h}^t(s_h)/\check{n}\,\iota_1} + 8H(A/\check{n})^{3/4}\iota_1 + \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \delta_{i,h+1}^{t_j}(s_{h+1}^{t_j}) \\
&\leq \lambda W_{i,h}^t(s_h) + \frac{8\iota + A\iota_1^2}{\lambda \check{n}} + 8H(A/\check{n})^{3/4}\iota_1 + \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \delta_{i,h+1}^{t_j}(s_{h+1}^{t_j}).
\end{aligned}$$

where the step is by the definition of bonus $b_{i,h}^t(s_h)$ in Line 15 of Algorithm 3. To separate the analysis for the first variance term $W_{i,h}^t(s_h)$ and the rest, we define intermediate terms $\delta_{i,h}^{1,t}(s_h)$ and $\delta_{i,h}^{2,t}(s_h)$ as follows. At $h = H + 1$, both terms are zero. In the first stage of (h, s_h), set $\delta_{i,h}^{1,t}(s_h) = 0$ and $\delta_{i,h}^{2,t}(s_h) = H - h + 1$. For later stages, where $\check{n} > 0$, define

$$\begin{aligned}
\delta_{i,h}^{1,t}(s_h) &:= \lambda W_{i,h}^t(s_h) + \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \delta_{i,h+1}^{1,t_j}(s_{h+1}^{t_j}), \\
\delta_{i,h}^{2,t}(s_h) &:= \frac{8\iota}{\lambda \check{n}} + \frac{A\iota_1^2}{\lambda \check{n}} + 8H \left(\frac{A}{\check{n}} \right)^{3/4} \iota_1 + \frac{1}{\check{n}} \sum_{j=1}^{\check{n}} \delta_{i,h+1}^{2,t_j}(s_{h+1}^{t_j}).
\end{aligned}$$

Both terms are nonnegative, and backward induction gives $\delta_{i,h}^t(s_h) \leq \delta_{i,h}^{1,t}(s_h) + \delta_{i,h}^{2,t}(s_h)$ for every episode and state. Thus it suffices to bound the two terms separately.

For the first term. At $h = H + 1$, all terms are zero. In the first stage, $\lambda \leq 1/(16H)$ gives

$$\delta_{i,h}^{1,t}(s_h) + 4\lambda \left(\overline{V}_{i,h}^t(s_h) \right)^2 = 4\lambda (H - h + 1)^2 \leq \tfrac{1}{4} \delta_{i,h}^{2,t}(s_h).$$

For later stages, the third inequality in Lemma D.5 gives

$$\begin{aligned}
\delta_{i,h}^{1,t}(s_h^t) + 4\lambda\left(\overline{V}_{i,h}^{t}(s_h^t)\right)^2 &\le \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(\delta_{i,h+1}^{1,t_j} + 4\lambda\left(\overline{V}_{i,h+1}^{t_j}\right)^2\right)(s_{h+1}^{t_j}) \\
&\quad + \frac{16\lambda H^2 A\iota_1^2}{\check{n}} + \frac{32H\iota}{\check{n}} + 32\lambda H^2(A/\check{n})^{3/4}\iota_1 + 8\lambda H \\
&\le \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\left(\delta_{i,h+1}^{1,t_j} + 4\lambda\left(\overline{V}_{i,h+1}^{t_j}\right)^2\right)(s_{h+1}^{t_j}) \\
&\quad + \frac{A\iota_1^2}{16\lambda\check{n}} + \frac{2\iota}{\lambda\check{n}} + 2H(A/\check{n})^{3/4}\iota_1 + 8\lambda H \\
&\le O(\lambda H(H-h+1)) + \frac{1}{4}\delta_{i,h}^{2,t}(s_h^t)\,.
\end{aligned}$$

The second step uses $\lambda \le 1/(16H)$; its three nonconstant terms are bounded by one quarter of the corresponding nonrecursive terms defining $\delta_{i,h}^{2,t}$. In the last step, the induction hypothesis contributes one quarter of the averaged future δ^{2,t_j}, and the remaining $8\lambda H$ is absorbed into $O(\lambda H(H-h+1))$.

For the second term. First, reweighting the historical samples gives

$$\begin{aligned}
&\sum_{t:\,\check{n}_h^t>0}\frac{1}{\check{n}_h^t}\sum_{j=1}^{\check{n}_h^t}\delta_{i,h+1}^{2,t_j}(s_{h+1}^{t_j}) \\
&= \sum_{\tau=1}^{T}\delta_{i,h+1}^{2,\tau}(s_{h+1}^{\tau})\sum_{t:\,\check{n}_h^t>0}\frac{1}{\check{n}_h^t}\sum_{j=1}^{\check{n}_h^t}\mathbb{I}\{\tau=t_j\} \\
&\le \left(1+\frac{1}{H}\right)\sum_{\tau=1}^{T}\delta_{i,h+1}^{2,\tau}(s_{h+1}^{\tau}).
\end{aligned}$$

Indeed, a sample is used only in the stage immediately following the stage in which it was collected. The ratio of these stage lengths is at most $1+1/H$, and an unfinished final stage only reduces its total weight. Separating first-stage visits, we obtain

$$\begin{aligned}
\sum_{t=1}^{T}\delta_{i,h}^{2,t}(s_h^t) &= \sum_{t:\,\check{n}_h^t=0}(H-h+1) + \sum_{t:\,\check{n}_h^t>0}\left(\frac{1}{\check{n}_h^t}\sum_{j=1}^{\check{n}_h^t}\delta_{i,h+1}^{2,t_j}(s_{h+1}^{t_j})\right. \\
&\qquad\left. + \frac{8\iota + A\iota_1^2}{\lambda\check{n}_h^t} + 8H(A/\check{n}_h^t)^{3/4}\iota_1\right) \\
&\le \left(1+\frac{1}{H}\right)\sum_{t=1}^{T}\delta_{i,h+1}^{2,t}(s_{h+1}^{t}) + \sum_{t:\,\check{n}_h^t=0}(H-h+1) \\
&\quad + \sum_{t:\,\check{n}_h^t>0}\left(\frac{8\iota + A\iota_1^2}{\lambda\check{n}_h^t} + 8H(A/\check{n}_h^t)^{3/4}\iota_1\right) \\
&\le \sum_{h'=h}^{H}\left(1+\frac{1}{H}\right)^{h'-h}\left[\sum_{t:\,\check{n}_{h'}^t=0}(H-h'+1)\right. \\
&\qquad\left. + \sum_{t:\,\check{n}_{h'}^t>0}\left(\frac{8\iota + A\iota_1^2}{\lambda\check{n}_{h'}^t} + 8H(A/\check{n}_{h'}^t)^{3/4}\iota_1\right)\right] \\
&\le \widetilde{O}\left(H^3S + \frac{H^2SA}{\lambda} + H^{11/4}(SA)^{3/4}T^{1/4}\right) \\
&\le \widetilde{O}\left(\frac{H^2SA}{\lambda} + H^{11/4}(SA)^{3/4}T^{1/4}\right).
\end{aligned}$$

The second inequality unrolls the recurrence down to $H+1$, including the contribution at h; its weights are bounded by $(1+1/H)^H \le 3$. Each state-step pair has at most H first-stage visits, so their total contribution is $O(H^3S)$. The third inequality is by Lemma D.6 with $w_h(s)=1$ and the Hölder's inequality. Finally, we know $H^3S \le H^2SA/(16\lambda)$ since $A \ge 1$ and $\lambda \le 1/(16H)$, which justifies the last inequality above.

Combining the two terms, for every $h \in [H]$ we have

$$\begin{aligned}\sum_{t=1}^{T} \delta_{i,h}^{t}(s_h^t) &\le \sum_{t=1}^{T} \delta_{i,h}^{1,t}(s_h^t) + \sum_{t=1}^{T} \delta_{i,h}^{2,t}(s_h^t) \\ &\le O(\lambda H^2 T) + \frac{5}{4}\sum_{t=1}^{T} \delta_{i,h}^{2,t}(s_h^t) \\ &\le \widetilde{O}\left(\lambda H^2 T + \frac{H^2 SA}{\lambda} + H^{11/4}(SA)^{3/4}T^{1/4}\right),\end{aligned}$$

which completes the proof. □

D.3 Further Details for the Online Setting

Proof of Lemma D.4. For analysis, let $\mathcal{F}_{t,h}$ be the σ-algebra generated by the current state s_h^t, all algorithmic randomness revealed before the actions at step h of episode t, and all transition results from earlier training steps, including the counterfactual samples. In chronological order, these σ-algebras form an enlarged filtration. Conditional on $\mathcal{F}_{t,h}$, the players' actions have product law π_h^t, and the next-state samples $\{s_{h+1,s_h^t,\boldsymbol{a}}^t\}_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}$ are generated independently of the current action randomization, with each $s_{h+1,s_h^t,\boldsymbol{a}}^t$ having law $\mathbb{P}_h(\cdot \mid s_h^t, \boldsymbol{a})$. The players observe only $s_{h+1,s_h^t,\boldsymbol{a}_h^t}^t$; all these samples are included in the enlarged filtration after that step. By induction on episode indices, $\overline{V}_{i,h+1}^t$, $\overline{V}_{i,h+1}^{\dagger,t}$, and $\underline{V}_{i,h+1}^t$ are $\mathcal{F}_{t,h}$-measurable. Thus, centering their sampled returns by the displayed conditional expectations gives martingale differences. Below, fix a stage by its order of occurrence and use the filtration at successive visits to (h, s_h), stopped at the end of training, writing $t_1, \dots, t_{\check{n}}$ for its visit episodes on completion. Simultaneous bounds over completed stages then cover every episode t using them.

For event $\mathcal{E}_1$. For each fixed (i, h, s_h) and stage, Corollary 3.2 applied to the restarted ADV_BANDIT subroutine gives the following bound when that stage completes, except on an event of probability at most $p/(3mHST)$:

$$\mathrm{Regret}_\star^{\check{n}} = \sum_{j=1}^{\check{n}} \ell^j(a_{i,h}^{t_j}) - \sum_{j=1}^{\check{n}} \ell^j(a_{i,h}^\star) \le \sqrt{AQ_{\check{n}}}\iota_1 + 4A^{3/4}\check{n}^{1/4}\iota_1\,,$$

where $\ell^j(a_{i,h}) = (H - (r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{t_j})(s_h, a_{i,h} \times \boldsymbol{a}_{-i,h}^{t_j}))/H$, $\forall a_{i,h} \in \mathcal{A}_i$. The clipping ensures $\ell^j(a_{i,h}) \in [0,1]$, and the conditional sampling above makes ℓ^j independent of $a_{i,h}^{t_j}$ given $\mathcal{F}_{t_j,h}$. For the realized actions, $Q_{\check{n}} = \check{n}W_{i,h}^t(s_h)/H^2$. Here and throughout this proof, ι_1 is the factor in Corollary 3.2 evaluated with confidence parameter $\delta = p/(3mHST)$, action count A, and global horizon T; since $A_i \le A$ and each stage has at most T rounds, it upper-bounds the logarithmic factor for every bandit instance. Then we have, for any $\mu \in \Delta(\mathcal{A}_i)$

$$\begin{aligned}&\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{t_j})(s_h, \mu \times \boldsymbol{a}_{-i,h}^{t_j}) - \frac{1}{\check{n}}\sum_{j=1}^{\check{n}}(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{t_j})(s_h, \boldsymbol{a}_h^{t_j}) \\ \le\ &\sqrt{AW_{i,h}^t(s_h)/\check{n}}\iota_1 + 4H(A/\check{n})^{3/4}\iota_1\,,\end{aligned}$$

for $W_{i,h}^t(s_h)$ defined in Eq. (13). Taking a union bound over $(i, h, s_h) \in [m] \times [H] \times \mathcal{S}$ and at most T completed stages per state-step pair gives $\Pr(\mathcal{E}_1) \ge 1 - p/3$.

For event $\mathcal{E}_2$. For each fixed (i, h, s_h) and stage, the following is a martingale difference sequence with respect to the filtration at successive visits above:

$$\left\{(r_{i,h} + \hat{\mathbb{P}}_h^{t_j}\underline{V}_{i,h+1}^{t_j})(s_h, \boldsymbol{a}_h^{t_j}) - (r_{i,h} + \mathbb{P}_h\underline{V}_{i,h+1}^{t_j})(s_h, \pi_h^{t_j})\right\}_{j=1}^{\check{n}},$$

Each increment is bounded in absolute value by H, and $\lambda H \le 1/16$. Applying Theorem B.2 to both signs with allowance $p/(6mHST)$ per tail gives the following bound when that stage completes, except on an event of probability at most $p/(3mHST)$; the stated ι dominates the logarithmic factor:

$$\left|\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\underline{V}_{i,h+1}^{t_j}\big)\big(s_h,\boldsymbol{a}_h^{t_j}\big)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\mathbb{P}_h\underline{V}_{i,h+1}^{t_j}\big)\big(s_h,\pi_h^{t_j}\big)\right|$$
$$\le\frac{\lambda}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{V}_h^{\pi_h^{t_j}}\big(r_{i,h}+\underline{V}_{i,h+1}^{t_j}\big)(s_h)+\frac{\iota}{\lambda\check{n}}\,.$$

Taking a union bound over $(i,h,s_h)\in[m]\times[H]\times\mathcal{S}$ and at most T completed stages per state-step pair gives $\Pr(\mathcal{E}_2)\ge 1-p/3$.

For event $\mathcal{E}_3$. For each fixed (i,h,s_h), stage, and action $a_{i,h}\in\mathcal{A}_i$, $\overline{V}_{i,h+1}^{\dagger,t_j}$ is $\mathcal{F}_{t_j,h}$-measurable. Thus, the following is a martingale difference sequence with respect to the same filtration:

$$\Big\{\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)\big(s_h,a_{i,h}\times\boldsymbol{a}_{-i,h}^{t_j}\big)-\big(r_{i,h}+\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big)\big(s_h,a_{i,h}\times\pi_{-i}^{t_j}\big)\Big\}_{j=1}^{\check{n}}\,,$$

Applying the same two-sided bound with allowance $p/(6mHSAT)$ per tail gives, first for $\mu=a_{i,h}$, the following bound when that stage completes, except on an event of probability at most $p/(3mHSAT)$:

$$\left|\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)\big(s_h,\mu\times\boldsymbol{a}_{-i,h}^{t_j}\big)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big)\big(s_h,\mu\times\pi_{-i}^{t_j}\big)\right|$$
$$\le\frac{\lambda}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{V}_h^{\mu\times\pi_{-i}^{t_j}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)+\frac{\iota}{\lambda\check{n}}\,.$$

Taking a union bound over $(i,h,s_h)\in[m]\times[H]\times\mathcal{S}$, at most T completed stages per state-step pair, and at most A actions gives the bounds simultaneously for all these choices. Convexity of the absolute value and the law of total variance then give the displayed bound for every $\mu\in\Delta(\mathcal{A}_i)$, and hence $\Pr(\mathcal{E}_3)\ge 1-p/3$.

Finally, taking union bound over subevents $\mathcal{E}_1,\mathcal{E}_2,\mathcal{E}_3$ completes the proof. □

Proof of Lemma D.5. **For the first inequality.** Fix any $a_i\in\mathcal{A}_i$, and take $\mu=a_i$ in $\mathcal{E}_3$. First, by law of total variance, i.e. $\mathrm{Var}(Y)=\mathbb{E}[\mathrm{Var}(Y|X)]+\mathrm{Var}(\mathbb{E}[Y|X])\ge\mathbb{E}[\mathrm{Var}(Y|X)]$, we have

$$\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{V}_h^{a_i\times\pi_{-i,h}^{t_j}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)\le\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)\,.$$

which proves the first step. Next, by the fact that $\mathbb{E}\big[(X+Y-\mathbb{E}[X+Y])^2\big]\le 2\mathbb{E}\big[(X-\mathbb{E}[X])^2\big]+2\mathbb{E}\big[(Y-\mathbb{E}[Y])^2\big]$, we have

$$\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)\le 2\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\overline{V}_{i,h+1}^{\dagger,t_j}(s_h)+2\,.$$

For $\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\overline{V}_{i,h+1}^{\dagger,t_j}$, we have

$$\begin{aligned}
&\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\overline{V}_{i,h+1}^{\dagger,t_j}(s_h)\\
&=\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\Big(\overline{V}_{i,h+1}^{\dagger,t_j}\Big)^2\big(s_h,a_i\times\pi_{-i,h}^{t_j}\big)-\left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big(s_h,a_i\times\pi_{-i,h}^{t_j}\big)\right)^2\\
&=\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\Big(\overline{V}_{i,h+1}^{\dagger,t_j}\Big)^2\big(s_h,a_i\times\pi_{-i,h}^{t_j}\big)-\Big(\overline{V}_{i,h}^{\dagger,t}(s_h)\Big)^2
\end{aligned}$$

$$+\left(\overline{V}_{i,h}^{\dagger,t}(s_h)\right)^2-\left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big(s_h,a_i\times\pi_{-i,h}^{t_j}\big)\right)^2.$$

The bracketed difference below is nonnegative by (16), since every empirical average in that definition lies in $[0,H-h+1]$. By the good event $\mathcal{E}$, We know the last difference in RHS should satisfy

$$\begin{aligned}
&\left|\left(\overline{V}_{i,h}^{\dagger,t}(s_h)\right)^2-\left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}(s_h,a_i\times\pi_{-i,h}^{t_j})\right)^2\right|\\
\le{}&(2H-1)\left|\overline{V}_{i,h}^{\dagger,t}(s_h)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}(s_h,a_i\times\pi_{-i,h}^{t_j})\right|\\
\le{}&(2H-1)\left|\begin{aligned}&\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h,a_i\times\boldsymbol{a}_{-i,h}^{t_j})\\&-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h,a_i\times\pi_{-i,h}^{t_j})\end{aligned}\right|\\
&+(2H-1)\left[\overline{V}_{i,h}^{\dagger,t}(s_h)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h,a_i\times\boldsymbol{a}_{-i,h}^{t_j})\right]\\
&+2H-1\\
\le{}&\frac{1}{4}\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)\\
&+(2H-1)\left[\overline{V}_{i,h}^{\dagger,t}(s_h)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h,a_i\times\boldsymbol{a}_{-i,h}^{t_j})\right]\\
&+2H-1+\frac{2H\iota}{\lambda\check{n}}.
\end{aligned}$$

where the first inequality factors the difference of squares and uses the ranges $0\le\overline{V}_{i,h}^{\dagger,t}(s_h)\le H-h+1$ and $0\le\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}(s_h,a_i\times\pi_{-i,h}^{t_j})\le H-h$. The second step uses the nonnegative bracketed difference and $0\le r_{i,h}\le 1$. The last inequality holds under $\mathcal{E}_3$, together with the law of total variance and $(2H-1)\lambda\le 1/8\le 1/4$.

Finally, combining all above,

$$\begin{aligned}
&\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)\\
\le{}&2\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\overline{V}_{i,h+1}^{\dagger,t_j}(s_h)+2\\
={}&\frac{2}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\Big(\overline{V}_{i,h+1}^{\dagger,t_j}\Big)^2\big(s_h,a_i\times\pi_{-i,h}^{t_j}\big)-2\left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\overline{V}_{i,h+1}^{\dagger,t_j}\big(s_h,a_i\times\pi_{-i,h}^{t_j}\big)\right)^2+2\\
\le{}&\frac{2}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\Big(\overline{V}_{i,h+1}^{\dagger,t_j}\Big)^2\big(s_h,a_i\times\pi_{-i,h}^{t_j}\big)-2\Big(\overline{V}_{i,h}^{\dagger,t}(s_h)\Big)^2\\
&+\frac{1}{2}\mathbb{V}_h^{a_i\times\hat{\pi}_{-i,h}^{t}}\big(r_{i,h}+\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h)+4H+\frac{4H\iota}{\lambda\check{n}}\\
&+4H\left[\overline{V}_{i,h}^{\dagger,t}(s_h)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}_h^{t_j}\overline{V}_{i,h+1}^{\dagger,t_j}\big)(s_h,a_i\times\boldsymbol{a}_{-i,h}^{t_j})\right].
\end{aligned}$$

Here the bracketed difference is nonnegative and $2(2H-1)\le 4H$.

Rearranging terms, we have

$$\mathbb{V}_h^{a_i\times\hat{\pi}^t_{-i,h}}\big(r_{i,h}+\overline{V}^{\dagger,t_j}_{i,h+1}\big)(s_h)\le\frac{4}{\check{n}}\sum_{j=1}^{\check{n}}\mathbb{P}_h\Big(\overline{V}^{\dagger,t_j}_{i,h+1}\Big)^2\big(s_h,a_i\times\pi^{t_j}_{-i,h}\big)-4\Big(\overline{V}^{\dagger,t}_{i,h}(s_h)\Big)^2$$
$$+8H\left[\overline{V}^{\dagger,t}_{i,h}(s_h)-\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\big(r_{i,h}+\hat{\mathbb{P}}^{t_j}_h\overline{V}^{\dagger,t_j}_{i,h+1}\big)(s_h,a_i\times\boldsymbol{a}^{t_j}_{-i,h})\right]$$
$$+\frac{8H\iota}{\lambda\check{n}}+8H,$$

which completes the proof for the first inequality in the claim.

For the second inequality. It can be proven in a very similar way, we omit it here for simplicity.

For the third inequality. For $\check{n}>0$, Eq. (13), the empirical variance inequality, and $r_{i,h}\in[0,1]$ give

$$\begin{aligned}W^t_{i,h}(s_h)&\le\frac{2}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})^2-2\left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})\right)^2+2\\&=\frac{2}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})^2-2\overline{V}^t_{i,h}(s_h)^2\\&\quad+2\left[\overline{V}^t_{i,h}(s_h)^2-\left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})\right)^2\right]+2\\&\le\frac{2}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})^2-2\overline{V}^t_{i,h}(s_h)^2+4Hb^t_{i,h}(s_h)+4H\,.\end{aligned}$$

For the last step, upper clipping in Eq. (12) implies $\overline{V}^t_{i,h}(s_h)\le\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})+1+b^t_{i,h}(s_h)$. Since $0\le\overline{V}^t_{i,h}\le H-h+1$ and $0\le\overline{V}^{t_j}_{i,h+1}\le H-h$, factoring the squared-value difference gives

$$\overline{V}^t_{i,h}(s_h)^2-\left(\frac{1}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})\right)^2\le(2H-1)\big(1+b^t_{i,h}(s_h)\big).$$

Substituting the bonus in Line 15 of Algorithm 3 and retaining the squared-value difference gives

$$\begin{aligned}W^t_{i,h}(s_h)&\le\frac{2}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})^2-2\overline{V}^t_{i,h}(s_h)^2\\&\quad+\frac{16H\iota}{\lambda\check{n}}+4H\sqrt{AW^t_{i,h}(s_h)/\check{n}\iota_1}+16H^2(A/\check{n})^{3/4}\iota_1+4H\\&\le\frac{2}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})^2-2\overline{V}^t_{i,h}(s_h)^2+\frac{1}{2}W^t_{i,h}(s_h)\\&\quad+\frac{8H^2A\iota_1^2}{\check{n}}+\frac{16H\iota}{\lambda\check{n}}+16H^2(A/\check{n})^{3/4}\iota_1+4H\,,\end{aligned}$$

where the second inequality is AM–GM applied only to the square-root term. Rearranging the above inequality yields

$$\begin{aligned}W^t_{i,h}(s_h)&\le\frac{4}{\check{n}}\sum_{j=1}^{\check{n}}\overline{V}^{t_j}_{i,h+1}(s^{t_j}_{h+1})^2-4\overline{V}^t_{i,h}(s_h)^2\\&\quad+\frac{16H^2A\iota_1^2}{\check{n}}+\frac{32H\iota}{\lambda\check{n}}+32H^2(A/\check{n})^{3/4}\iota_1+8H\,,\end{aligned}$$

which completes the proof.

□

Lemma D.6 (Variant of Lemma 11 in Zhang et al. (2020)). *Let $\check{n}_h^t := \check{N}_h^t(s_h^t)$, and n_h^t denote the number of visits to (h, s_h^t) before its current stage. For any non-negative weights $\{w_h(s)\}_{s\in\mathcal{S},\, h\in[H]}$ and $\alpha \in (0,1)$, it holds that*

$$\sum_{t,h:\, n_h^t>0} \frac{w_h(s_h^t)}{(n_h^t)^\alpha} \le \frac{2^\alpha}{1-\alpha} \sum_{s,h} w_h(s) N_h^{(T+1)}(s)^{1-\alpha} \tag{20}$$

and

$$\sum_{t,h:\, \check{n}_h^t>0} \frac{w_h(s_h^t)}{(\check{n}_h^t)^\alpha} \le \frac{2^{2\alpha} H^\alpha}{1-\alpha} \sum_{s,h} w_h(s) N_h^{T+1}(s)^{1-\alpha}.$$

In case $\alpha = 1$, it holds that

$$\sum_{t,h:\, n_h^t>0} \frac{w_h(s_h^t)}{n_h^t} \le 2 \sum_{s,h} w_h(s) \log\bigl(N_h^{(T+1)}(s) \vee 1\bigr) \tag{21}$$

and

$$\sum_{t,h:\, \check{n}_h^t>0} \frac{w_h(s_h^t)}{\check{n}_h^t} \le 4H \sum_{s,h} w_h(s) \log\bigl(N_h^{(T+1)}(s) \vee 1\bigr).$$

Here $x \vee 1 := \max\{x, 1\}$.

Proof. Fix (h, s) and write $N = N_h^{(T+1)}(s)$. Its stage lengths satisfy $e_1 = H$ and $e_{\ell+1} = \lfloor (1 + 1/H) e_\ell \rfloor$. In stage $\ell \ge 2$, let $x_\ell = \sum_{k<\ell} e_k$; then $n_h^t = x_\ell$ and $\check{n}_h^t = e_{\ell-1}$ whenever $s_h^t = s$ in that stage. Both counts are zero in the first stage. If $N \le H$, the corresponding sums are empty, so assume $N > H$.

For each visited stage $\ell \ge 2$, let d_ℓ be its number of visits up to episode T, allowing the final stage to be incomplete. Since $d_\ell \le e_\ell \le 2e_{\ell-1} \le 2x_\ell$, convexity of $u \mapsto u^{-\alpha}$ gives, for $0 < \alpha \le 1$,

$$\frac{d_\ell}{x_\ell^\alpha} \le \frac{2^\alpha d_\ell}{(x_\ell + d_\ell/2)^\alpha} \le 2^\alpha \int_{x_\ell}^{x_\ell + d_\ell} u^{-\alpha}\, du.$$

These intervals partition $[H, N]$. Thus

$$\sum_{t:\, s_h^t = s,\, n_h^t>0} \frac{1}{(n_h^t)^\alpha} \le 2^\alpha \int_H^N u^{-\alpha}\, du \le \begin{cases} \dfrac{2^\alpha}{1-\alpha} N^{1-\alpha}, & 0 < \alpha < 1, \\ 2\log(N \vee 1), & \alpha = 1. \end{cases}$$

For the preceding-stage counts, the integer update with $e_k \ge H$ gives $2H(e_{k+1} - e_k) \ge e_{k+1}$. Summing this inequality yields

$$x_\ell \le H + 2H(e_{\ell-1} - H) \le 2He_{\ell-1}, \qquad \check{n}_h^t \ge \frac{n_h^t}{2H}.$$

Hence replacing n_h^t by $\check{n}_h^t$ costs at most a factor $(2H)^\alpha$, including for $\alpha = 1$. Multiplying by $w_h(s)$ and summing over (h, s) proves all four inequalities. □

E Adversarial Bandit with Second-Order Regret

In this section, we provide the full version of Algorithm 1, which is as presented in Algorithm 5, and proofs of the Theorem 3.1 and Corollary 3.2.

Notation. Before we start the proof, we define several key notations used throughout this section. Let $\mathbf{1} \in \mathbb{R}^A$ denote the all-ones vector. For any two vectors $x, y \in \mathbb{R}^A$, $\langle x, y \rangle$ denotes their standard inner product. Given a strictly convex and continuously differentiable regularizer function ψ defined on a convex set, the Bregman divergence between two points w, w' is defined as $D_\psi(w, w') := \psi(w) - \psi(w') - \langle \nabla\psi(w'), w - w' \rangle$.

In Algorithm 5, we specifically employ a time-varying, weighted log-barrier regularizer $\psi^j(w) = -\sum_{a\in\mathcal{A}} \frac{1}{\eta^j(a)} \ln w(a)$. To facilitate the variance-style analysis of this regularizer, for any vector $h \in \mathbb{R}^A$, we define the local norm with respect to the Hessian of ψ^j at a point w as $\|h\|_{(j,w)} := \|h\|_{\nabla^2\psi^j(w)} = \sqrt{h^\top \nabla^2\psi^j(w)h}$, and its corresponding dual norm as $\|h\|_{\star(j,w)} := \|h\|_{\nabla^{-2}\psi^j(w)} = \sqrt{h^\top \nabla^{-2}\psi^j(w)h}$.

Algorithm 5 Adversarial Bandits with Empirical Variance Bound

Input: Initial learning rate η, confidence level δ
Define truncated simplex $\Omega = \{w \in \Delta(\mathcal{A}) : w(a) \geq 1/(An),\ \forall a \in \mathcal{A}\}$, increase factor $\kappa = e^{1/\ln(eAn)}$.
Initialize: $w^1(a) = 1/A$, $\rho^1(a) = 2A$, and $\eta^1(a) = \eta$ for all $a \in \mathcal{A}$.
for $j = 1, \ldots, n$ **do**
  $\psi^j(w) \leftarrow -\sum_{a\in\mathcal{A}} \frac{1}{\eta^j(a)} \ln w(a)$.
  Choose the predictable scalar $m^j \in [0, 1]$ from past observations.
  Sample $a^j \sim w^j$ and receive bandit feedback $\ell^j(a^j)$.
  Construct loss estimator $\hat{\ell}^j(a)$ and bonus $\hat{b}^j(a)$ as in Eq. (2), Eq. (3) respectively.
  Compute

$$w^{j+1} \leftarrow \arg\min_{w\in\Omega} \langle w, \hat{\ell}^j + \hat{b}^j\rangle + D_{\psi^j}(w, w^j) \tag{22}$$

  for $a \in \mathcal{A}$ **do**
    if $\frac{1}{w^{j+1}(a)} > \rho^j(a)$ **then**
      Set $\rho^{j+1}(a) \leftarrow \frac{2}{w^{j+1}(a)}$ and $\eta^{j+1}(a) \leftarrow \eta^j(a)\kappa$.
    else
      Set $\rho^{j+1}(a) \leftarrow \rho^j(a)$ and $\eta^{j+1}(a) \leftarrow \eta^j(a)$.
    end if
  end for
end for

Throughout this appendix, we use the closed domain $\Omega = \{w \in \Delta(\mathcal{A}) : w(a) \geq 1/(An)\}$. The update computes w^{j+1} using ψ^j before the threshold test and learning-rate updates. The initialization and threshold updates ensure

$$1/w^j(a) \leq \rho^j(a) \leq 2An.$$

Each threshold increase more than doubles its value from the initial $2A$. Since $\rho^j(a) \leq 2An$, at most $\log_2 n$ increases are possible for each action. Consequently, $\eta^j(a) \leq \eta \exp(\log_2(n)/\ln(eAn)) < 5\eta$. These statements also hold at $n = 1$, when Ω contains only the uniform distribution.

E.1 Proof of Theorem 3.1

For any $a^\star \in \mathcal{A}$, let $u = (1 - \frac{1}{n})a^\star + \frac{1}{An}\mathbf{1} \in \Omega$, where we also regard $a^\star$ as a one-hot vector for simplicity of notation. To prove Theorem 3.1, we first notice that the regret can be decomposed as the following way

$$\begin{aligned}
\mathrm{Regret}^n_\star &= \sum_{j=1}^n \langle a^j - a^\star, \ell^j\rangle \\
&= \sum_{j=1}^n \left(\ell^j(a^j) + b^j(a^j) - \langle u, \hat{\ell}^j + \hat{b}^j\rangle\right) \\
&\quad + \sum_{j=1}^n \langle u, \hat{\ell}^j - \ell^j\rangle + \sum_{j=1}^n \langle u, \hat{b}^j - b^j\rangle + \sum_{j=1}^n \langle u - a^j, b^j\rangle + \sum_{j=1}^n \langle u - a^\star, \ell^j\rangle \\
&\leq \sum_{j=1}^n \left(\ell^j(a^j) + b^j(a^j) - \langle u, \hat{\ell}^j + \hat{b}^j\rangle\right)
\end{aligned}$$

$$+\sum_{j=1}^{n}\langle u,\hat{\ell}^j-\ell^j\rangle+\sum_{j=1}^{n}\langle u,\hat{b}^j-b^j\rangle+\sum_{j=1}^{n}\langle u-a^j,b^j\rangle+1\,. \tag{23}$$

The rest of this section is about bounding the first four terms. We begin with the following lemma to bound the first term.

Lemma E.1. *Assume that $m^j \in [0,1]$ is a predictable scalar, as in Theorem 3.1, and $\eta \le \min\{1/405, 1/\iota_0\}$. Algorithm 5 ensures, for all $u \in \Omega$,*

$$\sum_{j=1}^{n}\Big(\ell^j(a^j)+b^j(a^j)-\langle u,\hat{\ell}^j+\hat{b}^j\rangle\Big)$$
$$\le \frac{2A\ln(eAn)}{\eta}+30\eta\sum_{j=1}^{n}(\ell^j(a^j)-m^j)^2+30\eta^3 n\iota_0^2-\frac{\langle\rho^n,u\rangle}{10\eta\ln(eAn)}.$$

Proof of Lemma E.1. Since $\ell^j(a^j), m^j \in [0,1]$, we have $|\ell^j(a^j)-m^j|\le 1$ and $0\le b^j(a^j)\le \eta\iota_0$. Thus AM–GM gives

$$\big(\ell^j(a^j)-m^j+b^j(a^j)\big)^2 \le 2\big(\ell^j(a^j)-m^j\big)^2+2\eta^2\iota_0^2.$$

Applying Lemmas E.5, E.6, and E.8 gives

$$\sum_{j=1}^{n}\langle w^j-u,\hat{\ell}^j+\hat{b}^j\rangle$$
$$\le \frac{2A\ln(eAn)}{\eta}-\frac{\langle\rho^n,u\rangle}{10\eta\ln(eAn)}+\sum_{j=1}^{n}3\eta^j(a^j)\big(\ell^j(a^j)-m^j+b^j(a^j)\big)^2$$
$$\le \frac{2A\ln(eAn)}{\eta}-\frac{\langle\rho^n,u\rangle}{10\eta\ln(eAn)}+30\eta\sum_{j=1}^{n}\big(\ell^j(a^j)-m^j\big)^2+30\eta^3 n\iota_0^2,$$

where the last step is by $\eta^j(a)\le 5\eta$ and the definition of the correction bonus b^j. Since m^j is a scalar, we have $\langle w^j,\hat{\ell}^j+\hat{b}^j\rangle=\ell^j(a^j)+b^j(a^j)$, which proves the claim. □

To bound the second and third term, which could be seen as martingales, we introduce the generalized Freedman's theorem by Lee et al. (2020a).

Theorem E.2 (Generalized Freedman's inequality, Theorem 2.2 of Lee et al. (2020a))**.** *Let $\{X_j\}_{j=1}^n$ be an m.d.s. with respect to a filtration $\{\mathcal{F}_j\}_{j=1}^n$ such that $\mathbb{E}[X_j \mid \mathcal{F}_j]=0$. Suppose $B_j\in[1,b]$ for a fixed constant b, where B_j is $\mathcal{F}_j$-measurable and $X_j\le B_j$ holds almost surely. Then, with probability at least $1-\delta'$, we have*

$$\sum_{j=1}^{n}X_j \le C\left(\sqrt{8V\ln\left(\frac{C}{\delta'}\right)}+2B^*\ln\left(\frac{C}{\delta'}\right)\right),$$

where $V=\max\Big\{1,\sum_{j=1}^{n}\mathbb{E}[X_j^2\mid\mathcal{F}_j]\Big\}$, $B^*=\max_{j\in[n]}B_j$ and $C=\lceil\log_2 b\rceil\,\lceil\log_2(b^2n)\rceil$.

For the fourth term in the regret decomposition, we notice that it's the "regret" of bonus term. Here we could finally explain the function of the calibration bonus $b^j(a)$. Basically, it act as a structural counterbalance during the regret decomposition. Specifically, when we apply Freedman's inequality (Theorem E.2) to bound the second term $\sum_{j=1}^{n}\langle u,\hat{\ell}^j-\ell^j\rangle$, the concentration naturally yields a variance term scaling with $\sum_{j=1}^{n}\langle u,(\ell^j-m^j\mathbf{1})^2\rangle$, which depends on the fixed optimal arm u. However, by deliberately designing the true bonus $b^j(a)$ to contain a negative quadratic penalty $-(\ell^j(a)-m^j)^2$, the fourth term in our decomposition evaluates to exactly $\sum_{j=1}^{n}\langle u-a^j,b^j\rangle = \eta\sum_{j=1}^{n}\big((\ell^j(a^j)-m^j)^2-\langle u,(\ell^j-m^j\mathbf{1})^2\rangle\big)\,\iota_0$. The negative component gracefully and exactly cancels out the fixed-arm variance generated by the second term, while the remaining positive component naturally replaces it with the learner's empirical realized-loss variance, Q_n.

Finally, we prove Theorem 3.1 with Lemma E.1 and Theorem E.2.

Proof of Theorem 3.1. Recall the decomposition in Eq. (23),

$$\begin{aligned}\mathrm{Regret}_\star^n \le \sum_{j=1}^n \Big(\ell^j(a^j) + b^j(a^j) - \langle u, \hat{\ell}^j + \hat{b}^j\rangle\Big) \\ + \sum_{j=1}^n \langle u, \hat{\ell}^j - \ell^j\rangle + \sum_{j=1}^n \langle u, \hat{b}^j - b^j\rangle + \sum_{j=1}^n \langle u - a^j, b^j\rangle + 1\,.\end{aligned}$$

We then bound the first four terms in RHS separately.

For the first term. We have the bound in Lemma E.1.

For the second term. Fix an arm $a^\star$ and its smoothed comparator $u = (1-1/n)a^\star + \mathbf{1}/(An)$. Set

$$V_n = \sum_{j=1}^n \big(\ell^j(a^j) - m^j\big)^2, \quad U_n = \sum_{j=1}^n \langle u, (\ell^j - m^j\mathbf{1})^2\rangle.$$

First take a fixed base learning rate $0 < \eta \le 1/\iota_0$, which also ensures $\eta \le 1/405$. Let $\delta' = \delta/(2A)$. We apply Theorem E.2 to the loss and correction estimators, using the common range bound $b = 16An$ and hence $C = \lceil \log_2(16An)\rceil\lceil\log_2(256A^2n^3)\rceil$. Conditional expectations below are taken before drawing a^j, conditional on the past and the current loss vector. m^j and w^j are measurable at this point, and a^j has distribution w^j.

For the loss-estimation martingale, let

$$X_j = \langle u, \hat{\ell}^j - \ell^j\rangle = \frac{u(a^j)(\ell^j(a^j) - m^j)}{w^j(a^j)} - \langle u, \ell^j - m^j\mathbf{1}\rangle,$$

we have $\mathbb{E}_j X_j = 0$, $|X_j| \le 4\langle \rho^j, u\rangle$, and $\mathbb{E}_j X_j^2 \le \langle \rho^j, u\rangle\langle u, (\ell^j - m^j\mathbf{1})^2\rangle$, since $|\ell^j(a) - m^j| \le 1$ and $1/w^j(a) \le \rho^j(a)$. As $\langle \rho^j, u\rangle$ is nondecreasing, with probability at least $1-\delta'$, we have

$$\begin{aligned}&\sum_{j=1}^n X_j \\ &\le C\sqrt{8(\langle \rho^n, u\rangle U_n + 1)\ln(2CA/\delta)} + 8C\langle \rho^n, u\rangle \ln(2CA/\delta) \\ &\le \eta\iota_0 U_n + \frac{\langle \rho^n, u\rangle}{30\eta\ln(eAn)} + C\sqrt{8\ln(2CA/\delta)} + 8C\langle \rho^n, u\rangle\ln(2CA/\delta),\end{aligned}$$

where $\iota_0 = 60C^2\ln(2CA/\delta)\ln(eAn)$ and AM-GM gives the last inequality.

For the third term. Note that $b^j(a) = \eta\iota_0\big(1 - (\ell^j(a) - m^j)^2\big) \in [0, \eta\iota_0]$. The correction-estimation errors $\langle u, \hat{b}^j - b^j\rangle$ have conditional mean zero and satisfy

$$\begin{aligned}|\langle u, \hat{b}^j - b^j\rangle| &\le 6\eta\iota_0\langle \rho^j, u\rangle, \\ \mathbb{E}_j\big[\langle u, \hat{b}^j - b^j\rangle^2\big] &\le 9\eta^2\iota_0^2\langle \rho^j, u\rangle.\end{aligned}$$

Applying the same concentration inequality to $\langle u, \hat{b}^j - b^j\rangle/(\eta\iota_0)$ gives, with probability at least $1-\delta'$,

$$\begin{aligned}&\sum_{j=1}^n \langle u, \hat{b}^j - b^j\rangle \\ &\le \eta\iota_0 C\left(\sqrt{72\langle \rho^n, u\rangle n\ln(2CA/\delta)} + 12\langle \rho^n, u\rangle\ln(2CA/\delta)\right) \\ &\le \frac{\langle \rho^n, u\rangle}{30\eta\ln(eAn)} + 540\eta^3\iota_0^2C^2\ln(2CA/\delta)\ln(eAn)n \\ &\quad + 12\eta\iota_0 C\langle \rho^n, u\rangle\ln(2CA/\delta).\end{aligned}$$

The range choices in the two concentration applications are valid because $2A \le \langle \rho^j, u\rangle \le 2An$, so both $4\langle \rho^j, u\rangle$ and $6\langle \rho^j, u\rangle$ lie in $[1, 16An]$.

For the fourth term. The correction-regret term is exactly

$$\sum_{j=1}^{n}\langle u - a^j, b^j\rangle = \eta\iota_0(V_n - U_n).$$

Combining all above. Combining the decomposition in Eq. (23) with Lemma E.1, the U_n terms cancel. The remaining range terms satisfy

$$\begin{aligned}&8C\langle\rho^n, u\rangle \ln(2CA/\delta) + 12\eta\iota_0 C\langle\rho^n, u\rangle \ln(2CA/\delta)\\ &\le 20C\langle\rho^n, u\rangle \ln(2CA/\delta) \le \frac{\langle\rho^n, u\rangle}{30\eta \ln(eAn)},\end{aligned}$$

since $\eta\iota_0 \le 1$ and $C \ge 10$. Thus the three terms $\langle\rho^n, u\rangle/(30\eta\ln(eAn))$ are canceled by the negative term in Lemma E.1. Since $C\sqrt{8\ln(2CA/\delta)} + 1 \le 2A\ln(eAn)/\eta$, we obtain

$$\begin{aligned}\sum_{j=1}^{n}\langle a^j - a^\star, \ell^j\rangle &\le \frac{4A\ln(eAn)}{\eta} + (\iota_0 + 30)\eta V_n + \iota_0^2(30 + 540C^2\ln(2CA/\delta)\ln(eAn))\eta^3 n\\ &\le 10\iota_0^3\left(\frac{A}{\eta} + \eta V_n + \eta^3 n\right).\end{aligned}$$

A union bound over the two martingales and the A choices of $a^\star$ gives probability at least $1-\delta$.

For the oracle tuning in the theorem, choose

$$\eta = \min\left\{\frac{1}{\iota_0}, \sqrt{\frac{A}{V_n + \sqrt{An}}}\right\}.$$

Then $A/\eta \le A\iota_0 + \sqrt{AV_n} + A^{3/4}n^{1/4}$, $\eta V_n \le \sqrt{AV_n}$, and $\eta^3 n \le A^{3/4}n^{1/4}$. Consequently,

$$\mathrm{Regret}_\star^n \le 20\iota_0^4\left(A + \sqrt{AV_n} + A^{3/4}n^{1/4}\right).$$

Finally, losses in $[0,1]$ give the deterministic bound $\mathrm{Regret}_\star^n \le n$. Combining these bounds gives

$$\begin{aligned}\mathrm{Regret}_\star^n &\le 20\iota_0^4\left(\min\{n, A\} + \sqrt{AV_n} + A^{3/4}n^{1/4}\right)\\ &\le 40\iota_0^4\left(\sqrt{AV_n} + A^{3/4}n^{1/4}\right),\end{aligned}$$

where $\min\{n, A\} \le A^{3/4}n^{1/4}$ for all $A, n \ge 1$. This is the claimed bound with $\iota_1 = 40\iota_0^4$. □

Remark 2. The tuning of η with oracle knowledge of V_n can be avoided using Algorithm 6 in Appendix E.4. Thus Algorithm 6 may replace the oracle-tuned Algorithm 5 in the Markov-game algorithms with the same sample-complexity orders up to logarithmic factors.

E.2 Proof of Corollary 3.2

Finally, we could prove the main result Corollary 3.2 using Theorem 3.1.

Proof of Corollary 3.2. By Lemma E.3 with $z_j = \ell^j(a^j)$, $x_j = m^j$, $L = 1$, we have

$$\sum_{j=1}^{n}(\ell^j(a^j) - m^j)^2 - \sum_{j=1}^{n}(\ell^j(a^j) - m^*)^2 \le 4(\ln n + 1).$$

Therefore, by Theorem 3.1 and choosing $m^* = \overline{\ell}_{a^j}$, we get

$$\begin{aligned}\mathrm{Regret}_\star^n &\le \sqrt{A(Q_n + 4\ln n + 4)}\iota_1 + A^{3/4}n^{1/4}\iota_1\\ &\le \sqrt{AQ_n}\iota_1 + \sqrt{A(4\ln n + 4)}\iota_1 + A^{3/4}n^{1/4}\iota_1\\ &\le \sqrt{AQ_n}\iota_1 + 4A^{3/4}n^{1/4}\iota_1,\end{aligned}$$

where the last step is by the fact that $\ln n = 2\ln n^{1/2} \le 2n^{1/2} - 2$. □

E.3 Further Details for Adversarial Bandits

In this subsection, we provide supporting lemmas.

Lemma E.3 (FTL for Online Quadratic Optimization, Corollary 2.2 of Shalev-Shwartz (2012))**.** *For $l(x; z_j) = \|x - z_j\|^2$ where $\max_j \|z_j\| \le L$, if we choose*

$$x_j = \arg\min_x \sum_{\tau<j} l(x; z_\tau),$$

then

$$\begin{aligned}\mathrm{Reg}_n(x^*) &= \sum_{j=1}^n l(x_j; z_j) - \sum_{j=1}^n l(x^*; z_j) \\ &\le 4L^2\big(\ln n + 1\big).\end{aligned}$$

Lemma E.4 (Standard OMD analysis)**.** *Let Ω be a convex compact set in $\mathbb{R}^A$, ψ be a convex function on Ω, w' be any point in Ω, and $x \in \mathbb{R}^A$. If*

$$w^* = \arg\min_{w\in\Omega} \left\{\langle w, x\rangle + D_\psi(w, w')\right\},$$

then for any $u \in \Omega$,

$$\langle w^* - u, x\rangle \le D_\psi(u, w') - D_\psi(u, w^*) - D_\psi(w^*, w').$$

Lemma E.5 (One-step OMD regret)**.** *Let $m^j \in [0, 1]$. The update in Eq. (22) ensures, for every $u \in \Omega$,*

$$\begin{aligned}&\langle w^j - u, \hat{\ell}^j + \hat{b}^j\rangle \\ &\quad\le D_{\psi^j}(u, w^j) - D_{\psi^j}(u, w^{j+1}) - D_{\psi^j}(w^{j+1}, w^j) + \langle w^j - w^{j+1}, \hat{\ell}^j + \hat{b}^j - m^j\mathbf{1}\rangle.\end{aligned}$$

Proof. Apply Lemma E.4 with $w^* = w^{j+1}$, $w' = w^j$, and $x = \hat{\ell}^j + \hat{b}^j$. Since m^j is a scalar and $w^j, w^{j+1} \in \Delta(\mathcal{A})$, we have $\langle w^j - w^{j+1}, m^j\mathbf{1}\rangle = m^j \sum_a (w^j(a) - w^{j+1}(a)) = 0$. Therefore,

$$\langle w^j - u, \hat{\ell}^j + \hat{b}^j\rangle = \langle w^{j+1} - u, \hat{\ell}^j + \hat{b}^j\rangle + \langle w^j - w^{j+1}, \hat{\ell}^j + \hat{b}^j - m^j\mathbf{1}\rangle.$$

Combining this identity with Lemma E.4 proves the claim. □

Lemma E.6 (Following analysis of Lemma 2.1 from Lee et al. (2020a))**.** *Algorithm 5 ensures, for any $u \in \Omega$,*

$$\sum_{j=1}^n \Big(D_{\psi^j}(u, w^j) - D_{\psi^j}(u, w^{j+1})\Big) \le \frac{2A\ln(eAn)}{\eta} - \frac{\langle \rho^n, u\rangle}{10\eta\ln(eAn)}.$$

Proof. For $h(y) = y - 1 - \ln y$,

$$\begin{aligned}&\sum_{j=1}^n \Big(D_{\psi^j}(u, w^j) - D_{\psi^j}(u, w^{j+1})\Big) \\ &\le D_{\psi^1}(u, w^1) + \sum_{j=1}^{n-1}(D_{\psi^{j+1}}(u, w^{j+1}) - D_{\psi^j}(u, w^{j+1})) \qquad (D_{\psi^n}(u, w^{n+1}) \ge 0) \\ &= \frac{1}{\eta}\sum_{a\in\mathcal{A}} h\left(\frac{u(a)}{w^1(a)}\right) + \sum_{a\in\mathcal{A}}\sum_{j=1}^{n-1}\left(\frac{1}{\eta^{j+1}(a)} - \frac{1}{\eta^j(a)}\right) h\left(\frac{u(a)}{w^{j+1}(a)}\right).\end{aligned}$$

For the first term, since $u(a) \ge \frac{1}{An}$ and $w^1 = \frac{\mathbf{1}}{A}$

$$\frac{1}{\eta}\sum_{a\in\mathcal{A}} h\left(\frac{u(a)}{w^1(a)}\right) = \frac{1}{\eta}\sum_{a\in\mathcal{A}} -\ln(Au(a)) \le \frac{A\ln n}{\eta}.$$

For the second term, note that by the non-decreasing property of learning rate, it is always non-positive and we will only bound the last non-zero part of it. For each arm a, denote by $n(a)$ the number of its learning-rate increases during rounds $1, \dots, n-1$. If $n(a) \geq 1$, let $j_a \leq n-1$ be the last such round, so that $\eta^n(a) = \eta^{j_a+1}(a) = \kappa \eta^{j_a}(a)$. Then

$$\begin{aligned}
&\left(\frac{1}{\eta^{j_a+1}(a)} - \frac{1}{\eta^{j_a}(a)}\right) h\left(\frac{u(a)}{w^{j_a+1}(a)}\right) \\
&= \frac{1-\kappa}{\kappa^{n(a)}\eta} h\left(\frac{u(a)}{w^{j_a+1}(a)}\right) \leq -\frac{1}{5\eta \ln(eAn)} h\left(\frac{u(a)}{w^{j_a+1}(a)}\right) = -\frac{h(u(a)\rho^n(a)/2)}{5\eta \ln(eAn)},
\end{aligned}$$

where we use the fact that $1-\kappa \leq -\frac{1}{\ln(eAn)}$ and $\kappa^{n(a)} \leq 5$. Further, for $h(u(a)\rho^n(a)/2)$,

$$-h\left(\frac{u(a)\rho^n(a)}{2}\right) = \ln\left(\frac{u(a)\rho^n(a)}{2}\right) - \frac{u(a)\rho^n(a)}{2} + 1 \leq \ln(eAn) - \frac{u(a)\rho^n(a)}{2},$$

where the inequality follows from $\frac{u(a)\rho^n(a)}{2} \leq \frac{1}{w^{j_a+1}(a)} \leq An$. For arms with $n(a) = 0$, the sum contributes zero and $\rho^n(a) = \rho^1(a) = 2A$. Including the same negative term for these arms costs at most

$$\sum_{a:n(a)=0} \frac{u(a)\rho^n(a)}{10\eta \ln(eAn)} = \frac{A}{5\eta \ln(eAn)} \sum_{a:n(a)=0} u(a) \leq \frac{A}{5\eta \ln(eAn)}.$$

Finally, combining the above inequalities gives

$$\begin{aligned}
\sum_{j=1}^{n}\Big(D_{\psi^j}(u, w^j) - D_{\psi^j}(u, w^{j+1})\Big) &\leq \frac{A \ln n}{\eta} + \sum_{a \in \mathcal{A}} \left(\frac{1}{5\eta} - \frac{u(a)\rho^n(a)}{10\eta \ln(eAn)}\right) + \frac{A}{5\eta \ln(eAn)} \\
&\leq \frac{2A \ln(eAn)}{\eta} - \frac{\langle \rho^n, u \rangle}{10\eta \ln(eAn)}.
\end{aligned}$$

The last inequality uses $\ln n \leq \ln(eAn)$ and $\ln(eAn) \geq 1$.

□

Lemma E.7 (Specialization of Lemma 14 of Wei and Luo (2018))**.** *Let $m^j \in [0,1]$. Let w^{j+1} follow Eq. (22), where ψ^j is the log barrier and $\eta^j(a) \leq 1/81$ for every a. If $\|\hat{\ell}^j + \hat{b}^j - m^j \mathbf{1}\|_{\star(j,w^j)} \leq 1/3$, then*

$$\|w^{j+1} - w^j\|_{(j,w^j)} \leq 3\|\hat{\ell}^j + \hat{b}^j - m^j \mathbf{1}\|_{\star(j,w^j)}.$$

Proof. Since $\langle w, m^j \mathbf{1} \rangle = m^j$ is constant over Ω, Eq. (22) is equivalently

$$w^{j+1} = \arg\min_{w \in \Omega} \left\{ \langle w, \hat{\ell}^j + \hat{b}^j - m^j \mathbf{1} \rangle + D_{\psi^j}(w, w^j) \right\}.$$

The claim follows from Lemma 14 of Wei and Luo (2018), specialized to scalar optimistic prediction and the update vector $\hat{\ell}^j + \hat{b}^j - m^j \mathbf{1}$. □

Lemma E.8 (Stability)**.** *Assume that $m^j \in [0,1]$ is a predictable scalar. If $\eta \leq \min\{1/405, 1/\iota_0\}$, then Algorithm 5 guarantees*

$$\langle w^j - w^{j+1}, \hat{\ell}^j + \hat{b}^j - m^j \mathbf{1} \rangle \leq \sum_a 3\eta^j(a)\big(w^j(a)(\hat{\ell}^j(a) + \hat{b}^j(a) - m^j)\big)^2.$$

Proof. Since $\ell^j(a^j), m^j \in [0,1]$ and $\eta\iota_0 \leq 1$, we have $|\ell^j(a^j) - m^j| \leq 1$ and $0 \leq b^j(a^j) \leq \eta\iota_0 \leq 1$. Hence $|\ell^j(a^j) - m^j + b^j(a^j)| \leq 2$, and the squared dual norm satisfies

$$\begin{aligned}
\|\hat{\ell}^j + \hat{b}^j - m^j \mathbf{1}\|^2_{\star(j,w^j)} &= \sum_a \eta^j(a)\big(w^j(a)(\hat{\ell}^j(a) + \hat{b}^j(a) - m^j)\big)^2 \\
&= \eta^j(a^j)\big(\ell^j(a^j) - m^j + b^j(a^j)\big)^2 \\
&\leq 4\eta^j(a^j) \leq 20\eta \leq 1/9.
\end{aligned}$$

Also $\eta^j(a) \le 5\eta \le 1/81$. Thus Lemma E.7 applies. By Hölder's inequality,

$$\begin{aligned}\langle w^j - w^{j+1}, \hat{\ell}^j + \hat{b}^j - m^j \mathbf{1}\rangle &\le \|w^j - w^{j+1}\|_{(j,w^j)} \|\hat{\ell}^j + \hat{b}^j - m^j \mathbf{1}\|_{\star(j,w^j)} \\ &\le 3\|\hat{\ell}^j + \hat{b}^j - m^j \mathbf{1}\|^2_{\star(j,w^j)},\end{aligned}$$

which is the claimed bound. □

E.4 Adaptive Version of Algorithm 5

In this section, we utilize the doubling trick proposed in Lee et al. (2020b) to remove the need of tuning the learning rate η in terms of Q_n and provide an adaptive version of Algorithm 5 in Algorithm 6. For this adaptive run, evaluate the confidence factors ι_0 and ι_1 from Section 3 at allowance δ/n, while δ denotes the failure probability for the whole run. In particular,

$$\iota_0 = 60C^2 \ln(2CAn/\delta) \ln(eAn),$$

with the same global-horizon constant C as before. These factors, the domain, and κ remain fixed across epochs.

Algorithm 6 Adaptive version of Algorithm 5

Input: Initial learning rate η, confidence level δ
Initialize: $j = 1$.
Define $\kappa = e^{1/\ln(eAn)}$ and truncated simplex $\Omega = \{w \in \Delta(\mathcal{A}) : w(a) \ge 1/(An),\ \forall a \in \mathcal{A}\}$.
for $\lambda = 1, \dots$ **do**
  $n_\lambda = j - 1$, $w^j(a) = 1/A$, $\eta^j(a) = \eta$, $\rho^j(a) = 2A$, $\forall a \in \mathcal{A}$.
  while $j \le n$ **do**
    Execute Line 5 to Line 16
    if $\frac{A+1}{\eta} \le \eta \sum_{\tau=n_\lambda+1}^{j} (\ell^\tau(a^\tau) - m^\tau)^2 + \eta^3(j - n_\lambda)$ **then**
      $\eta \leftarrow \frac{\eta}{2}$, $j \leftarrow j + 1$
      Break
    end if
    $j \leftarrow j + 1$
  end while
  if $j > n$ **then**
    Break
  end if
end for

Theorem E.9. *Assume that $m^j \in [0,1]$ is a predictable scalar, as in Theorem 3.1. With $\eta = \min\{1/405, 1/\iota_0\}$, for any $\delta \in (0,1)$, Algorithm 6 guarantees with probability at least $1 - \delta$, for all $a^\star \in \mathcal{A}$, that*

$$\sum_{j=1}^{n} \big(\ell^j(a^j) - \ell^j(a^\star)\big) \le \widetilde{O}\left(\sqrt{A \sum_{j=1}^{n} (\ell^j(a^j) - m^j)^2} + A^{3/4} n^{1/4}\right).$$

Proof. Define the time steps between two resets an epoch and denote by λ the indices of these epochs, $\lambda^\star$ the index of the last epoch and η_λ the value of η during epoch λ so that $\eta_\lambda = 2^{1-\lambda}\eta_1$. Set $n_1 = 0$ and $n_{\lambda^\star+1} = n$, so the last epoch includes any final incomplete segment.

For each $\lambda \in [n]$ and fixed arm $a^\star$, take zero increments if that epoch does not occur, and use the comparator $u = (1 - 1/n)a^\star + \mathbf{1}/(An)$ from the fixed-rate proof. Apply Theorem E.2 to

$$\begin{aligned}&\mathbb{I}\{n_\lambda < j \le n_{\lambda+1}\}\langle u, \hat{\ell}^j - \ell^j\rangle, \\ &\frac{\mathbb{I}\{n_\lambda < j \le n_{\lambda+1}\}}{\eta_\lambda \iota_0}\langle u, \hat{b}^j - b^j\rangle.\end{aligned}$$

Whether round j belongs to epoch λ is known before its action is sampled, so these sequences remain martingale differences. Extend the epoch's threshold vector by its initial value before the epoch and freeze it after the epoch; the predictable range bounds stay in $[1, 16An]$ and are controlled by the

epoch's terminal thresholds. Hence the fixed-rate concentration and cancellation argument applies at its random endpoint. There are at most n nonempty epochs; allocating $\delta/(2An)$ to each of the two martingales, each arm, and each epoch gives, by a union bound, one event of probability at least $1-\delta$ on which the realized-regret bounds below hold simultaneously. The domain and logarithmic factors use the global horizon n, while local sums use the actual epoch length. The comparator smoothing cost in an epoch is at most $(n_{\lambda+1}-n_\lambda)/n$.

On this event, consider the realized regret in epoch λ.

$$\begin{aligned}
&\sum_{j=n_\lambda+1}^{n_{\lambda+1}} \left(\ell^j(a^j)-\ell^j(a^\star)\right) \\
&\le \widetilde{O}\left(\frac{A\ln(eAn)}{\eta_\lambda}+\eta_\lambda\sum_{j=n_\lambda+1}^{n_{\lambda+1}}(\ell^j(a^j)-m^j)^2+\eta_\lambda^3(n_{\lambda+1}-n_\lambda)\right) \\
&\le \widetilde{O}\Bigg(\frac{A\ln(eAn)}{\eta_\lambda}+\eta_\lambda\sum_{j=n_\lambda+1}^{n_{\lambda+1}-1}(\ell^j(a^j)-m^j)^2 \\
&\qquad\qquad +\eta_\lambda^3(n_{\lambda+1}-n_\lambda-1)+\eta_\lambda+\eta_\lambda^3\Bigg) \\
&\le \widetilde{O}\left(\frac{A\ln(eAn)}{\eta_\lambda}+\eta_\lambda+\eta_\lambda^3\right),
\end{aligned}$$

where the first inequality follows from the fixed-rate calculation in the proof of Theorem 3.1 on the common event above, and the third inequality is because the reset condition does not hold for $j=n_{\lambda+1}-1$. Summing over the epochs gives

$$\sum_{j=1}^{n}\left(\ell^j(a^j)-\ell^j(a^\star)\right)\le\widetilde{O}\left(\sum_{\lambda=1}^{\lambda^\star}\frac{A\ln(eAn)}{\eta_\lambda}+\eta_\lambda+\eta_\lambda^3\right)\le\widetilde{O}\left(\frac{A\ln(eAn)}{\eta_{\lambda^\star}}\right), \tag{24}$$

where the last step is by $\sum_{\lambda=1}^{\lambda^\star}\frac{1}{\eta_\lambda}=\sum_{\lambda=1}^{\lambda^\star}\frac{2^{\lambda-1}}{\eta_1}=\frac{2^{\lambda^\star}-1}{\eta_1}\le\frac{2}{\eta_{\lambda^\star}}$ and $\sum_{\lambda=1}^{\lambda^\star}\eta_\lambda\le 2\eta_1\le\widetilde{O}(1)$.

If $\lambda^\star=1$, Eq. (24) gives a $\widetilde{O}(A)$ bound, since $1/\eta_1$ is polylogarithmic. Combining this with the deterministic regret bound n and $\min\{n,A\}\le A^{3/4}n^{1/4}$ proves the claim in this case. Otherwise, for $\lambda^\star\ge 2$, using the reset condition, we have

$$\begin{aligned}
\frac{A}{\eta_{\lambda^\star-1}} &\le \eta_{\lambda^\star-1}\sum_{j=n_{\lambda^\star-1}+1}^{n_{\lambda^\star}}(\ell^j(a^j)-m^j)^2+\eta_{\lambda^\star-1}^3(n_{\lambda^\star}-n_{\lambda^\star-1}) \\
&\le \eta_{\lambda^\star-1}\sum_{j=1}^{n}(\ell^j(a^j)-m^j)^2+\eta_{\lambda^\star-1}^3 n \\
&\le 2\max\left\{\eta_{\lambda^\star-1}\sum_{j=1}^{n}(\ell^j(a^j)-m^j)^2,\eta_{\lambda^\star-1}^3 n\right\},
\end{aligned}$$

from which we could deduce that

$$\begin{aligned}
\frac{A}{\eta_{\lambda^\star-1}} &\le 2\max\left\{\sqrt{A\sum_{j=1}^{n}(\ell^j(a^j)-m^j)^2},A^{3/4}n^{1/4}\right\} \\
&\le 4\left(\sqrt{A\sum_{j=1}^{n}(\ell^j(a^j)-m^j)^2}+A^{3/4}n^{1/4}\right).
\end{aligned}$$

Further, by the fact that $\eta_{\lambda^\star}=\eta_{\lambda^\star-1}/2$ and plugging the above inequality back to Eq (24), we have

$$\sum_{j=1}^{n}\left(\ell^j(a^j)-\ell^j(a^\star)\right)\le\widetilde{O}\left(\sqrt{A\sum_{j=1}^{n}(\ell^j(a^j)-m^j)^2}+A^{3/4}n^{1/4}\right),$$

The common event is simultaneous over all arms, so maximizing the left side over $a^\star$ gives the same bound for $\mathrm{Regret}_\star^n$, completing the proof. $\square$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explicitly state our main theoretical contributions and sample complexity bounds in the Abstract and Section 1, which are rigorously supported by the theorems presented in Section 3 (Theorem 3.1 and Corollary 3.2), Section 4 (Theorems 4.1 and Theorem 4.2), and Section 5 (Theorem 5.1).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitations of our current theoretical guarantees in Section 6 (Conclusion), specifically noting the remaining $\widetilde{O}(H)$ sample complexity gap in our online setting results and our current restriction to coarse correlated equilibria (CCE) rather than correlated equilibria (CE).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We establish our formal problem setting and baseline assumptions (e.g., the generative model) in Section 2 (Preliminaries). Complete proofs for all theorems and corollaries are provided in the supplementary material, specifically in Appendix C (simulator setting), Appendix D (online setting), and Appendix E (adversarial bandit algorithm).

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper is purely theoretical and does not contain any empirical experiments or computational simulations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: This paper is purely theoretical and does not contain any empirical experiments or computational simulations.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper is purely theoretical and does not contain any empirical experiments or computational simulations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This paper is purely theoretical and does not contain any empirical experiments or computational simulations.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper is purely theoretical and does not contain any empirical experiments or computational simulations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper presents foundational theoretical research in multi-agent reinforcement learning and fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This paper presents foundational theoretical research on the sample complexity of Markov games and is not tied to any particular application or deployment that would yield direct societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper focuses exclusively on theoretical sample complexity bounds and does not release any models or datasets, meaning it poses no risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: This paper is purely theoretical and does not utilize any existing assets such as datasets, codebases, or pre-trained models.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper focuses exclusively on theoretical algorithmic design and sample complexity bounds, and does not introduce or release any new empirical assets such as datasets, code, or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This is a purely theoretical machine learning paper that does not involve crowdsourcing or any research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This theoretical research does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as a component of the core methodology, algorithmic design, or theoretical analysis in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.