INDEPENDENT REINFORCEMENT LEARNING IN DISCOUNTED MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we study *radically uncoupled* learning in discounted general-sum Markov games. Assuming "ETH for PPAD", we show that, for every fixed discount factor, there is no polynomial-time algorithm for computing inverse-polynomially accurate coarse correlated equilibria in discounted general-sum Markov games when players learn independently in decentralized settings. Complementing this hardness result, we present what appears to be the first *radically uncoupled* algorithm that achieves sub-exponential convergence to coarse correlated equilibria in discounted general-sum Markov games, under both full-feedback and partial-feedback settings. Our approach builds on a *layered* variant of optimistic mirror descent, equipped with an increasing step-size schedule designed to address the multi-agent setting.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) studies how multiple decision-making agents learn and act in a shared environment, where each agent seeks to optimize its own objective while adapting to the behavior of others (Buşoniu et al., 2008; Zhang et al., 2021). This framework has become a central modeling paradigm for large-scale decision-making systems involving both cooperation and competition. Its recent applications range from strategic game playing (Silver et al., 2016; Brown & Sandholm, 2018; Perolat et al., 2022; Kramár et al., 2022; Bakhtin et al., 2022) to large language models (Park et al., 2025; Wan et al., 2025), robotics (Levine et al., 2018), quantitative finance (Zhang et al., 2024), intelligent transportation (Shalev-Shwartz et al., 2016), cybersecurity (Malialis & Kudenko, 2015), and economic policy design (Zheng et al., 2022). These applications motivate algorithms that are not only efficient, but also robust to decentralization, and limited information.

A standard framework for the theoretical study of MARL is Markov games (Shapley, 1953; Littman, 1994). In a Markov game, interaction unfolds over a sequence of states. At each state, players choose actions simultaneously; each player receives a reward depending on the current state and the joint actions; and the next state is drawn from a transition kernel that also depends on the joint actions, which extends Markov decision processes (MDPs) to multi-agent environments while also extending normal-form games to dynamic environments. Thus, computing an equilibrium in a Markov game amounts to finding a fixed-point in a dynamic environment.

In this work, we focus on a fully decentralized information model for MARL, namely *independent* or *radically uncoupled* learning (Foster et al., 2023; Foster & Young, 2006). We study this model in the discounted setting, which is widely used in the control and reinforcement learning literatures (Daskalakis et al., 2023; Tessler et al., 2019; Huang et al., 2026; Pérolat et al., 2017; Sayin et al., 2021). Here, *radically uncoupled* refers to that players are not allowed to observe one another's policies, losses, or random bits. In particular, there is no available shared randomness, communication, or correlation device during learning. Consequently, each player only uses her own feedback and treats the other players as part of the environment. Independent learning dynamics are attractive because they scale without the communication overhead required by centralized coordination. For instance, in large-scale systems, aggregating global information or coordinating agents through a central learner can be costly or infeasible under bandwidth, latency, or privacy constraints Lian et al. (2017). Moreover, these informational restrictions rule out many of the mechanisms used in existing results for equilibrium learning in Markov games (Bai et al., 2020; Wei et al., 2021; Jin et al., 2023), motivating the learning under decentralized information structures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Beyond asymptotic convergence, one typically seeks algorithms that approximates equilibrium with controlled computational cost and sample complexity, for example with (quasi-)polynomial dependence on the problem parameters. To achieve these guarantees, centralized methods often rely on shared value estimates, policy information, or coordinated equilibrium-computation subroutines. In contrast, under radically uncoupled information, each player must both compute and statistically estimate her update using her own feedback. Therefore, the main challenge in the latter setting is not merely to design a convergent algorithm, but to obtain convergence with computational and statistical efficiency under radically uncoupled information. Therefore, we investigate the following question:

Are there (quasi-)efficient, radically uncoupled learning algorithms that converge to approximate equilibria in discounted general-sum Markov games?

In normal-form games, the connection between independent learning and equilibrium learning is well understood, as independent learning reduces to no-regret learning. If all players have vanishing external regret, then their average play converges to a coarse correlated equilibrium (CCE), and if all players have vanishing swap regret, their average play converges to a correlated equilibrium (CE) (Cesa-Bianchi & Lugosi, 2006; Blum & Mansour, 2007). Moreover, it has been shown that optimistic and regularized learning dynamics yield faster self-play rates (Daskalakis et al., 2011; Syrgkanis et al., 2015; Daskalakis et al., 2021; Anagnostides et al., 2022; Soleymani et al., 2025).

In contrast, the relationship between independent learning and equilibrium learning in Markov games remains much less understood. For two-player zero-sum Markov games there is a substantial body of work on decentralized learning, including settings with limited shared information and *radically uncoupled* algorithms with provable guarantees (Bai et al., 2020; Wei et al., 2021; Cai et al., 2023; Chen et al., 2024). Naturally, general-sum Markov games are more challenging, as there is no common value function or saddle-point structure. The challenge stems not only from the general-sum structure, but also from the dynamic nature of the environment: alternative policies may induce different distributions over future states, and these distributions depend jointly on the transition dynamics and the policies employed by the other players. Since the opponents may also change their policies over time, these distributions are themselves time-varying and unknown to players. This is a central obstruction to transferring the normal-form no-regret argument directly to Markov games.

The existing positive results for general-sum Markov games typically rely on coordination or post-processing mechanisms that are not present in the independent learning problems (Arslan & Yüksel, 2017; Jin et al., 2023; Song et al., 2022; Mao & Başar, 2023; Daskalakis et al., 2023). A frequently studied class of algorithms include V-learning and related algorithms (Jin et al., 2023; Song et al., 2022; Mao & Başar, 2023), which learn CCE or CE in decentralized MARL. The policies produced by these methods are represented as carefully specified distributions over the learning history and require the players to sample a common index during execution. Notably, the learning phase of these algorithms does not provide any online regret guarantees. Thus, while these methods are decentralized, they do not address the radically uncoupled learning objective considered in this paper. Being another related algorithm, the so-called SPoCMAR (Daskalakis et al., 2023) is a decentralized algorithm with polynomial sample and computational complexity for learning a nonstationary Markov-CCE, but the provided regret guarantees rely on the existence of common random bits, where Markov-CCE is an equilibrium with respect to the Markov policy deviations. Consequently, these works do not provide online regret guarantees for the independently played product policies during the learning process.

Another closely related precursor is Erez et al. (2025). Similar to our work, they study radically uncoupled learning in finite-horizon general-sum Markov games in the self-play setting. Their work, compared to ours, contains further restrictions. First, their regret bounds are obtained with respect to Markov policies, whereas we allow non-Markov policies. Second, their sublinear regret guarantee is obtained when agents may affect one another's losses but not one another's transition dynamics. In contrast, we study finite- and infinite-horizon discounted Markov games with action-dependent transition kernels and develop radically uncoupled learning algorithms with online regret guarantees.

On the other hand, there are results showing limitations in finding efficient algorithms for the MDPs and Markov games. In the single-agent case, no-regret learning in adversarially changing MDPs can be computationally intractable (Abbasi-Yadkori et al., 2013). Naturally, this intractability also arises in Markov games with adversarial opponents, where both computational and statistical barriers are known (Liu et al., 2022; Tian et al., 2021). Motivated by the fact that computing Nash equilibria is PPAD-complete in normal-form games (Daskalakis et al., 2009), it has been shown that similar

hardness constraints can persist in Markov games. In particular, learning stationary Markov-CCEs in discounted Markov games is PPAD-hard (Daskalakis et al., 2023). Moreover, in finite-horizon Markov games, any independent learning algorithm that is no regret is also PPAD-hard (Foster et al., 2023). Thus, we seek nonstationary policies that approximate CCE in discounted Markov games while avoiding the intractability barriers in the self-play setting, which complements the literature.

Our contributions in this paper can be summarized as follows.

1. In Section 2, we study finite-horizon general-sum Markov games under full-feedback. We introduce an algorithm that achieves an $\mathcal{O}\big(T^{-3/(3H+1)}\big)$ -approximate CCE after T episodes. In Subsection 2.2, under a reachability assumption, we also present a similar (high-probability CCE) result under partial feedback, up to extra polynomial terms.
2. In Section 3, we transfer the finite-horizon guarantees established in Section 2 to discounted Markov games (Theorem 3). In the discounted case our algorithms yield quasi-polynomial guarantees for learning inverse-polynomially accurate discounted-CCE for the full- and partial feedback cases for both finite-horizon and infinite-horizon Markov games (Corollaries 1 and 2).
3. In Section 4, we complement the quasi-polynomial upper bounds established in Section 3 for learning discounted-CCE with a worst-case complexity-theoretic lower bound under the assumption that "ETH for PPAD" holds (Babichenko et al., 2016). This result relies on realizing the problem as a *sparse* approximate CCE for discounted general-sum Markov games.

1.1 Notations

For any $q \in \mathbb{Z}_{\geq 1}$, we write $[q] := \{1, \ldots, q\}$. By $\Delta(X)$, we denote probability measures over the set X. Given $x \in \Delta(X)$, probability of an action a is $x(a)$. All logarithms are natural base.

1.2 Undiscounted Finite-horizon Markov games

Let m be a positive integer. An m-player finite-horizon Markov game (FHMG) is specified by the tuple $G = \big(H, \{S_h\}_{h=1}^{H+1}, \{A_i\}_{i=1}^{m}, \{P_h\}_{h=1}^{H}, \{\ell_h^i\}_{i\in[m],h\in[H]}, s_1\big)$. We will denote by $\mathcal{M} = [m]$ the set of players. The components of the the Markov game are defined as follows:

- Horizon length is denoted by $H \in \mathbb{N}_+$.
- We denote the states at layer h by S_h. The initial state s_1 is deterministic. The layer $H+1$ is terminal and has no actions or costs. We write $\mathcal{S} := \bigsqcup_{h=1}^{H+1} S_h$, and $S := |\mathcal{S}|$.
- Each player i has a finite action set A_i, with joint action space $A := \prod_{i=1}^{m} A_i$. We write $\boldsymbol{a} = (a_1, \ldots, a_m) \in A$, $\boldsymbol{a}_{-i} \in A_{-i} := \prod_{j\neq i} A_j$ for the action profile excluding player i, and by $A_{\max} := \max_{i\in[m]} |A_i|$, we denote the maximum cardinality among players.
- For any $h \in [H]$, by $P_h : S_h \times A \to \Delta(S_{h+1})$ we denote the (state) transition probability kernel.
- For each player $i \in [m]$ and layer $h \in [H]$, the cost function is $\ell_h^i : S_h \times A \to [0, 1]$.

General (non-Markov) policies. For $h \in [H]$, let $\mathcal{H}_h := S_1 \times A \times \cdots \times S_{h-1} \times A \times S_h$ denote the set of histories before the layer-h action. An element of $\mathcal{H}_h$ is written as $\tau_h = (s_1, \boldsymbol{a}_1, \ldots, s_{h-1}, \boldsymbol{a}_{h-1}, s_h)$. We let $s(\tau_h) \in S_h$ be the concluding state of the history $\tau_h \in \mathcal{H}_h$. A non-Markov policy of player i is denoted as $\pi^i = \{\pi_h^i\}_{h=1}^{H}$, $\pi_h^i(\cdot \mid \tau_h) \in \Delta(A_i)$ for every $\tau_h \in \mathcal{H}_h$. Let Π_i^{gen} denote the set of all such finite-horizon policies of player i, and by $\boldsymbol{\pi} \in \Pi^{\mathrm{gen}} := \prod_{i=1}^{m} \Pi_i^{\mathrm{gen}}$ define the set of non-Markov product policy profiles. In particular, for player $i \in \mathcal{M}$ a general (non-Markov) policy $\pi_i \in \Pi_i^{\mathrm{gen}}$ is specified by a collection $\pi^i = (\pi_1^i, \cdots, \pi_H^i)$, where $\pi_h^i : \tau_h \to \Delta(A_i)$. Under such a profile, conditional on the history, τ_h, the players' actions are sampled independently as $a_h^i \sim \pi_h^i(\cdot \mid \tau_h)$, $i \in [m], h \in [H]$.

Markov policies. A Markov policy of player i is a sequence $\pi^i = \{\pi_h^i\}_{h=1}^{H} \in \prod_{h=1}^{H} \Delta(A_i)^{S_h}$. Let Π_i^{markov} denote the set of all such finite-horizon policies of player i, and define $\Pi^{\mathrm{markov}} := \prod_{i=1}^{m} \Pi_i^{\mathrm{markov}}$ to denote the space of product Markov policies, where each agent i independently follows a policy in Π_i^{markov} for a given state. In particular, a Markov product policy $\boldsymbol{\pi} \in \Pi^{\mathrm{markov}}$ is given by $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_H) \in \prod_{j\in[m]} \Delta(A_j)^{S_h}$. At state $s \in S_h$, it induces the policy distribution $\boldsymbol{\pi}_h(\boldsymbol{a} \mid s) := \prod_{j=1}^{m} \pi_h^j(a_j \mid s)$, $\boldsymbol{a} = (a_1, \ldots, a_m) \in A$. For player i, write $\boldsymbol{\pi}^{-i} :=$

$(\pi^1,\ldots,\pi^{i-1},\pi^{i+1},\ldots,\pi^m)$ for the policy profile of all players except i. If ρ^i is another policy of player i, either Markov or non-Markov, we write $\rho^i \odot \boldsymbol{\pi}^{-i}$ for the profile obtained by replacing player i's policy with ρ^i and leaving the opponents' policies equal to $\boldsymbol{\pi}^{-i}$. In particular, if ρ^i is Markov, then for $s \in S_h$, $(\rho^i \odot \boldsymbol{\pi}^{-i})_h(\boldsymbol{a} \mid s) := \rho^i_h(a_i \mid s) \prod_{j \neq i} \pi^j_h(a_j \mid s)$.

Repeated interaction. The agents interact with environment over T episodes. At the beginning of episode $t \in [T]$, each player i chooses some $\pi^i_t \in \Pi^{\mathrm{markov}}_i$, which results in the product Markov profile $\boldsymbol{\pi}_t = (\pi^1_t, \ldots, \pi^m_t)$. Each episode starts from $s_{t,1} = s_1$. For each $h \in [H]$, player i samples $a^i_{t,h} \sim \pi^i_{t,h}(\cdot \mid s_{t,h})$, the joint action is $\boldsymbol{a}_{t,h} := (a^1_{t,h}, \ldots, a^m_{t,h})$, player i incurs cost $\ell^i_h(s_{t,h}, \boldsymbol{a}_{t,h})$, and the next state is sampled as $s_{t,h+1} \sim P_h(\cdot \mid s_{t,h}, \boldsymbol{a}_{t,h})$.

Definition 1 (Full-feedback)**.** *At the end of episode t, player i observes the exact $Q^i_t(s, a_i) = Q^{i,\boldsymbol{\pi}_t}_h(s, a_i),\ \forall h \in [H],\ s \in S_h,\ a_i \in A_i$.*

Definition 2 (Partial feedback)**.** *At the end of episode t, player i observes only its own trajectory information $\big(s_{t,1}, a^i_{t,1}, c^i_{t,1}, s_{t,2}, a^i_{t,2}, c^i_{t,2}, \ldots, s_{t,H}, a^i_{t,H}, c^i_{t,H}, s_{t,H+1}\big)$, where $c^i_{t,h} := \ell^i_h(s_{t,h}, \boldsymbol{a}_{t,h})$.*

For a policy profile $\boldsymbol{\pi} \in \Pi^{\mathrm{markov}}$, let $\mathbb{P}^{\boldsymbol{\pi}}$ denote the distribution over trajectories induced by the policy profile $\boldsymbol{\pi}$ and the transition kernels $\{P_h\}_{h=1}^H$, and define $V^{i,\boldsymbol{\pi}}_{H+1} \equiv 0$. All expectations, $\mathbb{E}^{\boldsymbol{\pi}}$, are taken with respect to this probability distribution. For $h \in [H]$, $s \in S_h$, and $i \in [m]$, we define the value function $V^{i,\boldsymbol{\pi}}_h(s) := \mathbb{E}^{\boldsymbol{\pi}}\big[\sum_{r=h}^H \ell^i_r(s_r, \boldsymbol{a}_r) \,|\, s_h = s\big]$. Similarly, the state-action value is given as $Q^{i,\boldsymbol{\pi}}_h(s, a_i) := \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}^{-i}_h(\cdot|s)}\big[\ell^i_h(s, \boldsymbol{a}) + \sum_{s' \in S_{h+1}} P_h(s' \mid s, \boldsymbol{a})\, V^{i,\boldsymbol{\pi}}_{h+1}(s')\big]$, where the opponents follow the $\boldsymbol{\pi}^{-i}$. Then, $V^{i,\boldsymbol{\pi}}_h(s) = \mathbb{E}_{a_i \sim \pi^i_h(\cdot|s)}\big[Q^{i,\boldsymbol{\pi}}_h(s, a_i)\big]$. For s_1, we have $V^{i,\boldsymbol{\pi}}(s_1) := V^{i,\boldsymbol{\pi}}_1(s_1)$.

The induced MDP of a player. For an episode t, a player i, with the fixed opponents' policies $\boldsymbol{\pi}^{-i}_t$, faces an induced finite-horizon MDP $M^i_t := \big(H, \{S_h\}_{h=1}^{H+1}, A_i, \{P^i_{t,h}\}_{h=1}^H, \{\ell^i_{t,h}\}_{h=1}^H, s_1\big)$, where, for $s \in S_h$, $a_i \in A_i$, and $s' \in S_{h+1}$, $\ell^i_{t,h}(s, a_i) := \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}^{-i}_{t,h}(\cdot|s)}\big[\ell^i_h\big(s, (a_i, \boldsymbol{a}_{-i})\big)\big]$, and $P^i_{t,h}(s' \mid s, a_i) := \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{\pi}^{-i}_{t,h}(\cdot|s)}\big[P_h\big(s' \mid s, (a_i, \boldsymbol{a}_{-i})\big)\big]$. For any $\mu^i \in \Pi^{\mathrm{markov}}_i$ in this induced MDP, define $V^{i,\mu^i}_{t,H+1} \equiv 0$ and, then, we have $Q^{i,\mu^i}_{t,h}(s, a_i) := \ell^i_{t,h}(s, a_i) + \sum_{s' \in S_{h+1}} P^i_{t,h}(s' \mid s, a_i)\, V^{i,\mu^i}_{t,h+1}(s')$, $V^{i,\mu^i}_{t,h}(s) := \mathbb{E}_{a_i \sim \mu^i_h(\cdot|s)}\big[Q^{i,\mu^i}_{t,h}(s, a_i)\big]$. These induced MDP quantities agree with the corresponding Markov-game quantities, i.e. $V^{i,\mu^i}_{t,h}(s) = V^{i,\mu^i \odot \boldsymbol{\pi}^{-i}_t}_h(s)$, $Q^{i,\mu^i}_{t,h}(s, a_i) = Q^{i,\mu^i \odot \boldsymbol{\pi}^{-i}_t}_h(s, a_i)$. Under the played policy π^i_t, we use the shorthand $V^{i,\pi}_{t,h}(s) := V^{i,\pi^i_t \odot \boldsymbol{\pi}^{-i}_t}_h(s)$, $Q^{i,\pi}_{t,h}(s, a_i) := Q^{i,\pi^i_t \odot \boldsymbol{\pi}^{-i}_t}_h(s, a_i)$, $s \in S_h$. Moreover, we drop policy-dependent indices from the value-functions' notation whenever they are clear from the context.

Distributional policies. We call distributions over general policies, i.e., elements of $\Delta(\Pi^{\mathrm{gen}})$, *distributional policies*. Playing a distributional policy $P \in \Delta(\Pi^{\mathrm{gen}})$ consists of first sampling a randomized policy $\boldsymbol{\pi} \sim P$, and then executing the $\boldsymbol{\pi}$. For any policy $\boldsymbol{\pi} \in \Pi^{\mathrm{gen}}$, let $\delta_{\boldsymbol{\pi}} \in \Delta(\Pi^{\mathrm{gen}})$ denote the distribution placing unit mass on $\boldsymbol{\pi}$. Accordingly, the empirical distribution over the iterates $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T \in \Pi^{\mathrm{gen}}$ is $\widehat{\Pi}_T := \frac{1}{T}\sum_{t=1}^T \delta_{\boldsymbol{\pi}_t} \in \Delta(\Pi^{\mathrm{gen}})$. For $\mathfrak{P} \in \Delta(\Pi^{\mathrm{gen}})$, we define $V^{i,\mathfrak{P}}(s_1) := \mathbb{E}_{\boldsymbol{\sigma} \sim \mathfrak{P}}\big[V^{i,\boldsymbol{\sigma}}(s_1)\big]$. Then, for a deviation $\mu^i \in \Pi^{\mathrm{gen}}_i$, we let $V^{i,\mu^i \odot \mathfrak{P}^{-i}}(s_1) := \mathbb{E}_{\boldsymbol{\sigma} \sim \mathfrak{P}}\big[V^{i,\mu^i \odot \boldsymbol{\sigma}^{-i}}(s_1)\big]$.

Definition 3 (External regret)**.** *Given product Markov policy profiles $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T \in \Pi^{\mathrm{markov}}$, the external regret of player i is* $\mathrm{Reg}^{\mathrm{gen},i}_T := \sup_{\mu^i \in \Pi^{\mathrm{gen}}_i} \sum_{t=1}^T \big[V^{i,\boldsymbol{\pi}_t}(s_1) - V^{i,\mu^i \odot \boldsymbol{\pi}^{-i}_t}(s_1)\big]$.

Definition 4 (Coarse Correlated Equilibrium)**.** *The empirical distribution $\widehat{\Pi}_T = \frac{1}{T}\sum_{t=1}^T \delta_{\boldsymbol{\pi}_t}$ is an ε-CCE if, for every player $i \in [m]$, $\sup_{\mu^i \in \Pi^{\mathrm{gen}}_i} \big[V^{i,\widehat{\Pi}_T}(s_1) - V^{i,\mu^i \odot \widehat{\Pi}^{-i}_T}(s_1)\big] \leq \varepsilon$.*

1.3 Discounted Markov games

We use the same notation for finite and infinite-horizon discounted games. Let $\bar{H} \in \mathbb{N}_+ \cup \{\infty\}$ denote the horizon. When $\bar{H} < \infty$, an m-player finite-horizon discounted Markov game (FHdMG) is specified by $G_{\gamma,\bar{H}} = \big(\bar{H}, \{S_h\}_{h=1}^{\bar{H}+1}, \{A_i\}_{i=1}^m, \{P_h\}_{h=1}^{\bar{H}}, \{\ell^i_h\}_{i \in [m], h \in [\bar{H}]}, \gamma, s_1\big)$. When $\bar{H} = \infty$, we recover the infinite-horizon discounted Markov game (IHdMG) $G_\gamma = \big(S, \{A_i\}_{i=1}^m, P, \{\ell^i\}_{i=1}^m, \gamma, s_1\big)$.

In both cases, let $\mathcal{M} = [m]$ be the set of players. The corresponding games are similar to the undiscounted case, except, at each layer h, instead of ℓ_h^i, each player incurs the cost $\gamma^{h-1}\ell_h^i$.

General (non-Markov) and Markov policies. For each $h \le \bar{H}$, let $\mathcal{H}_h^{\bar{H}} := S_1 \times A \times \cdots \times S_{h-1} \times A \times S_h$ be the set of histories before the time-h action. When $\bar{H} = \infty$, we set $S_h = S$ for every $h \ge 2$. An element of $\mathcal{H}_h^{\bar{H}}$ is written as $\tau_h = (s_1, \boldsymbol{a}_1, \ldots, s_{h-1}, \boldsymbol{a}_{h-1}, s_h)$. A non-Markov policy of player i over horizon $\bar{H}$, set of which is denoted as $\Pi_i^{\mathrm{gen},\bar{H}}$, is a sequence of maps $\mu^i = \{\mu_h^i\}_{h \le \bar{H}},\ \mu_h^i(\cdot \mid \tau_h) \in \Delta(A_i)$ for every $\tau_h \in \mathcal{H}_h^{\bar{H}}$. We define *profile* obtained from them as $\Pi^{\mathrm{gen},\bar{H}} := \prod_{i=1}^m \Pi_i^{\mathrm{gen},\bar{H}}$. A Markov policy of player i over horizon $\bar{H}$ is a sequence $\pi^i = \{\pi_h^i\}_{h \le \bar{H}},\ \pi_h^i(\cdot \mid s) \in \Delta(A_i),\ s \in S_h$, set of which is defined by $\Pi_i^{\mathrm{markov},\bar{H}}$. By $\Pi^{\mathrm{markov},\bar{H}} := \prod_{i=1}^m \Pi_i^{\mathrm{markov},\bar{H}}$, we denote Markov profiles.

Discounted values. For policy $\boldsymbol{\sigma} \in \Pi^{\mathrm{gen},\bar{H}}$ and a finite $L \le \bar{H}$, define the L-truncated discounted cost of player i by $J_{i,L}^\gamma(\boldsymbol{\sigma}) := \mathbb{E}^{\boldsymbol{\sigma}}\big[\sum_{h=1}^L \gamma^{h-1}\ell_h^i(s_h, \boldsymbol{a}_h) \,|\, s_1 = s_{\mathrm{init}}\big]$. When $\bar{H} = \infty$, we also write $J_i^\gamma(\boldsymbol{\sigma}) := J_{i,\infty}^\gamma(\boldsymbol{\sigma})$. Since $0 \le \ell_h^i \le 1$, we have $\sup_{\boldsymbol{\sigma}} J_{i,\bar{H}}^\gamma(\boldsymbol{\sigma}) \le \sum_{h=1}^{\bar{H}} \gamma^{h-1} \le (1-\gamma)^{-1}$.

Given $\mathfrak{P} \in \Delta(\Pi^{\mathrm{gen},\bar{H}})$, define $J_{i,\bar{H}}^\gamma(\mathfrak{P}) := \mathbb{E}_{\boldsymbol{\sigma}\sim\mathfrak{P}}\big[J_{i,\bar{H}}^\gamma(\boldsymbol{\sigma})\big]$. For a unilateral deviation $\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}$, define $J_{i,\bar{H}}^\gamma(\mu^i \odot \mathfrak{P}^{-i}) := \mathbb{E}_{\boldsymbol{\sigma}\sim\mathfrak{P}}\big[J_{i,\bar{H}}^\gamma(\mu^i \odot \boldsymbol{\sigma}^{-i})\big]$.

Definition 5. *For a sequence of $\bar{H}$-horizon product policies $\bar{\boldsymbol{\pi}}_1, \ldots, \bar{\boldsymbol{\pi}}_T$, the discounted regret of player i is defined as* $\mathrm{Reg}_{\gamma,\bar{H},T}^{\mathrm{gen},i} := \sup_{\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}} \sum_{t=1}^T \big[J_{i,\bar{H}}^\gamma(\bar{\boldsymbol{\pi}}_t) - J_{i,\bar{H}}^\gamma(\mu^i \odot \bar{\boldsymbol{\pi}}_t^{-i})\big]$.

Definition 6. *The empirical distribution $\widehat{\Pi}_T$ is an ε-discounted CCE if, for every player $i \in [m]$,* $\sup_{\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}} \big[J_{i,\bar{H}}^\gamma(\widehat{\Pi}_T) - J_{i,\bar{H}}^\gamma(\mu^i \odot \widehat{\Pi}_T^{-i})\big] \le \varepsilon$. *If $\bar{H} = \infty$, it is the discounted infinite-horizon CCE.*

Truncated Markov Games. Given $G_{\gamma,\bar{H}} = \big(\bar{H},\ \{S_h\}_{h=1}^{\bar{H}+1},\ \{A_i\}_{i=1}^m,\ \{P_h\}_{h=1}^{\bar{H}},\ \{\ell_h^i\}_{i\in[m],h\in[\bar{H}]},\ \gamma,\ s_1\big)$, for $L \le \bar{H}$, we define the corresponding L-step truncation as $G_{\gamma,\bar{H}}^{[L]} = \big(L,\ \{S_h\}_{h=1}^{L+1},\ \{A_i\}_{i=1}^m,\ \{P_h\}_{h=1}^L,\ \{c_h^i\}_{i\in[m],h\in[L]},\ \gamma, s_1\big)$, where $c_h^i = \gamma^{h-1}\ell_h^i$. For $\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}$, define $\mu^{i,[L]} := (\mu_h^i)_{h=1}^L \in \Pi_i^{\mathrm{gen},L}$. We define policy $\boldsymbol{\pi}^{L,\mathrm{ref}} \in \Pi_i^{\mathrm{gen},\bar{H}-L}$ for the layers after L, in which every player plays uniformly at every state. For a profile $\boldsymbol{\pi}^{[L]} \in \Pi_i^{\mathrm{gen},L}$, let $\mathrm{Ext}_L(\boldsymbol{\pi}^{[L]}) \in \Pi_i^{\mathrm{gen},\bar{H}}$ denote the $\bar{H}$-horizon profile that plays $\boldsymbol{\pi}^{[L]}$ for first L layers and then plays $\boldsymbol{\pi}^{L,\mathrm{ref}}$ afterward. Finally, the discounted tail after layer L or truncation error is $\tau_{L,\bar{H}}(\gamma) := \sum_{h=L+1}^{\bar{H}} \gamma^{h-1} \le \frac{\gamma^L}{1-\gamma}$.

2 Main Algorithm and Results for Episodic Markov Games

Let G be a FHMG with no discount and horizon length $\bar{H} < \infty$. In this section, we establish regret guarantees for G in the self-play setting that underlie our later IHdMG and FHdMG results. The performance metric is regret with respect to history-dependent deviations, which yields approximate CCE.

A unilateral deviation of policies in a Markov game may change the distribution of future states, and therefore the future decision points at which the deviation is evaluated. Hence, the comparison places more weight on histories that the deviating policy visits frequently and less weight on histories that it reaches rarely, which, for instance, does not happen in normal-form games.

From the learner's perspective, these history-occupancy weights are unknown and time-varying, as they depend on the transition kernel, the deviation, and the opponents' evolving policies. Thus, the analysis must control their *drift* alongside the online learning dynamics. We first address this in the full-feedback model and then extend the argument to partial feedback under a reachability condition.

2.1 Episodic Markov games under full-feedback

In this subsection, we present a provably convergent algorithm for independent self-play in game G under full feedback (see Algorithm 1). Here, independence means that each player observes only the state-action Markov profile from the previous episode. In the full-feedback setting, one has access to exact $Q_{t,h}^i(s,\cdot)$, which bypasses the need for estimation.

To ensure stability of policy updates, we introduce the smoothed entropy function as the regularizer choice. Given a smoothing parameter $\lambda_i > 0$, the smoothed negative-entropy regularizer over $\Delta(A_i)$ is defined as $\Psi_i(x) := \sum_{a \in A_i} \big(x(a) + \lambda_i\big) \log\big(x(a) + \lambda_i\big)$. The associated Bregman divergence over $\Delta(A_i) \times \Delta(A_i)$ is $D_i(x, y) = \sum_{a \in A_i} \big(x(a) + \lambda_i\big) \log \frac{x(a)+\lambda_i}{y(a)+\lambda_i}$. Here, the parameter λ_i is used to smooth the regularized on the boundary of $\Delta(A_i)$, similar to the fixed-share smoothing idea used in tracking and adaptive-regret algorithms Herbster & Warmuth (1998); Cesa-Bianchi et al. (2012); Si Salem et al. (2024).

Algorithm 1 Layered Optimistic Online Mirror Descent

1. **Input:** Horizon H, episode count T, base step-size $\eta_0 \in (0, 1]$.
2. Define $\alpha_h := \frac{3(H-h)+1}{3H+1},\ \eta_h := \eta_0 T^{-\alpha_h},\ \bar{x}^i(a_i) := \frac{1}{|A_i|}, \lambda_i = 1/|A_i|,\ a_i \in A_i, h \in [H]$.
3. Set $\widetilde{x}^{i,s}_{0,h} := \bar{x}^i \in \operatorname{argmin}_{x \in \Delta(A_i)} \Psi_i(x),\ Q^i_0(s, \cdot) \equiv 0 \in \mathbb{R}^{A_i}, \forall\, i \in [m], h \in [H]$, and $s \in S_h$.

   For $t = 1, \ldots, T$

   For each player $i \in [m]$, layer $h \in [H]$, and state $s \in S_h$

   i. $x^{i,s}_{t,h} = \operatorname{argmin}_{x \in \Delta(A_i)} \Big\{ \eta_h \left\langle Q^i_{t-1,h}(s, \cdot), x \right\rangle + D_i\Big(x, \widetilde{x}^{i,s}_{t-1,h}\Big) \Big\}$.

   ii. Set the policy at (h, s) by $\pi^i_{t,h}(\cdot \mid s) = x^{i,s}_{t,h}(\cdot)$.

   a. Execute $\boldsymbol{\pi}_t = (\pi^1_t, \ldots, \pi^m_t) \in \Pi^{\text{markov}}$, and get $Q^i_{t,h}(s, a_i),\ \forall i \in [m], h \in [H],\ s \in S_h,\ a_i \in A_i$.

   For each player $i \in [m]$, layer $h \in [H]$, and state $s \in S_h$

   i. $\widetilde{x}^{i,s}_{t,h} = \operatorname{argmin}_{x \in \Delta(A_i)} \Big\{ \eta_h \left\langle Q^i_{t,h}(s, \cdot), x \right\rangle + D_i\Big(x, \widetilde{x}^{i,s}_{t-1,h}\Big) \Big\}$.

The second design choice is the increasing step-size schedule to have iterations that are robust across the layers. In particular, by using smaller learning rates on earlier layers force the acquired policies during training to change slowly and remain robust on later epochs. In contrast, we use larger learning rates on later layers, allowing the downstream parts of the game to adapt more *quickly*. In a sense, this creates a nested *time-scale separation* across the horizon as deep layers react faster, while earlier layers remain comparatively stable. Lastly, as a consequence of optimism, this schedule turns the problem into a controlled collection of slowly varying weighted online learning problems. Similar ideas previously have appeared for learning-rate schemes used in adaptive online learning and adversarial bandit/MDP settings Bubeck et al. (2021); Agarwal et al. (2017); Wei & Luo (2018).

Theorem 1. *Define* $C^{\text{SE}}_* := \max_{i \in [m]} \big\{ \log(|A_i| + 1)[\frac{2H}{\eta_0} + 12mH^3] + \eta_0 H^3 + 144\eta_0^3 m^2 H^7 \big\}$. *Given* $\varepsilon \in (0, 1]$, *choose* $T_\varepsilon := \left\lceil \big(C^{\text{SE}}_* \varepsilon^{-1}\big)^{(3H+1)/3} \right\rceil$. *Then, after running Algorithm 1 for every player* $i \in [m]$, *for* T_ε *episodes, results in* $\text{Reg}^{\text{gen},i}_{T_\varepsilon} \le \varepsilon T_\varepsilon$, *i.e.,* $\widehat{\Pi}_{T_\varepsilon}$ *is an* ε*-CCE with respect to all history-dependent deviations.*

Proof. We refer to Appendix B for the proof. □

2.2 Episodic Markov games under partial feedback

In this subsection, we extend the results of Subsection 2.1 to the partial-feedback setting. We tackle the difficulties that come from restrictions on the access to Q-functions by estimating the Q-functions as follows. First, players will maintain a small action-exploration floor throughout the play by taking mirror descent steps over the simplex $\Delta^\zeta_i := \big\{x \in \Delta(A_i) : x(a_i) \ge \frac{\zeta}{|A_i|} \text{ for every } a_i \in A_i\big\}$ instead of $\Delta(A_i)$, where $\zeta \in (0, 1]$. Second, unlike the full-information case, the players update policies in *blocks*, see Erez et al. (2023). During each block, the same policy is repeated for several independent episodes, whose realized samples are then used to estimate the Q-functions through exploration. The policy is then updated using these approximations.

To ensure that players can explore the state space, we impose the following reachability condition.

Definition 7. *A FHMG is said to be* (κ, ζ)*-reachable if, for any* $\boldsymbol{\pi} \in \Pi^{\text{markov}}$ *satisfying* $\min_{(h,s,a_i) \in [H] \times S_h \times A_i} \pi^i_h(a_i \mid s) \ge \frac{\zeta}{|A_i|}, i \in [m]$, *we have* $\min_{h \in [H], s \in S_h} \Pr_{\boldsymbol{\pi}}(s_h = s) \ge \kappa$.

At block k, all players use the product Markov profile $\boldsymbol{\pi}_k$ for B independent episodes. For episode $r \in [B]$, let $(s_h^{k,r}, \boldsymbol{a}_h^{k,r})_{h=1}^H$ denote the realized state-action trajectory. For $h \in [H]$, $s \in S_h$, and $a_i \in A_i$, define $X_{k,r,h}^i(s,a_i) := \mathbf{1}\{s_h^{k,r} = s,\ a_h^{i,k,r} = a_i\}, Y_h^{i,k,r} := \sum_{u=h}^H \ell_u^i(s_u^{k,r}, \boldsymbol{a}_u^{k,r})$, and $N_{k,i}(s,a_i) := \sum_{r=1}^B X_{k,r,h}^i(s,a_i)$. We define the Q-estimate of player i at block k as

$$\widehat{Q}_{k,h}^i(s,a_i) := \begin{cases} \dfrac{\sum_{r=1}^B X_{k,r,h}^i(s,a_i) Y_h^{i,k,r}}{N_{k,i}(s,a_i)}, & N_{k,i}(s,a_i) > 0, \\ 0, & N_{k,i}(s,a_i) = 0. \end{cases} \tag{1}$$

Algorithm 2 in the partial-feedback setting, compared to Algorithm 1, differs in only two aspects: the use of a blocked estimation procedure and the usage of the estimated values.

Algorithm 2 Blocked Layered OOMD

1. **Input:** Horizon H, number of blocks K, block length B, $\zeta \in (0,1]$, $\eta_0 \in (0,1]$.
2. Define $\alpha_h := \frac{3(H-h)+1}{3H+1},\ \eta_h := \eta_0 K^{-\alpha_h},\ \bar{x}^i(a_i) := \frac{1}{|A_i|}, \lambda_i = 1/|A_i|,\ a_i \in A_i, h \in [H]$.
3. Set $\widetilde{x}_{0,h}^{i,s} = \bar{x}^i \in \operatorname{argmin}_{x \in \Delta_i^\zeta} \Psi_i(x),\ \widehat{Q}_0^i(s,\cdot) \equiv 0 \in \mathbb{R}^{A_i}, \forall\, i \in [m], h \in [H]$, and $s \in S_h$.

For $k = 1, \ldots, K$

For each player $i \in [m]$, layer $h \in [H]$, and state $s \in S_h$

i. $x_{k,h}^{i,s} = \operatorname{argmin}_{x \in \Delta_i^\zeta} \left\{ \eta_h \langle \widehat{Q}_{k-1,h}^i(s,\cdot), x \rangle + D_i(x, \widetilde{x}_{k-1,h}^{i,s}) \right\}$.

ii. Set $\pi_{k,h}^i(\cdot \mid s) = x_{k,h}^{i,s}(\cdot)$.

a. Execute B independent episodes using the product profile $\boldsymbol{\pi}_k = (\pi_k^1, \ldots, \pi_k^m) \in \Pi^{\text{markov}}$.

b. Each player i obtains the estimate $\widehat{Q}_{k,h}^i(s,a_i)$ for all $h \in [H]$, $s \in S_h$, and $a_i \in A_i$.

For each player $i \in [m]$, layer $h \in [H]$, and state $s \in S_h$

i. $\widetilde{x}_{k,h}^{i,s} = \operatorname{argmin}_{x \in \Delta_i^\zeta} \left\{ \eta_h \langle \widehat{Q}_{k,h}^i(s,\cdot), x \rangle + D_i(x, \widetilde{x}_{k-1,h}^{i,s}) \right\}$.

For the regret guarantee of Algorithm 2, we first introduce the following:

$$\widetilde{C}_*^{\text{SE}} := \max_{i \in [m]} \left\{ \frac{2H \log(|A_i|+1)}{\eta_0} + 12mH^3 \log(|A_i|+1) + \eta_0 H^3 + 288\eta_0^3 m^2 H^7 \right\},$$

Since we estimate Q-functions, the constant above differs from C_*^{SE} only in the last term.

Theorem 2. *Suppose the underlying Markov game is (κ, ζ_ϵ)-reachable, where $\zeta_\varepsilon := \varepsilon 8^{-1} H^{-2}$. Fix $\varepsilon, \delta \in (0,1$. Choose $K_\varepsilon := \left\lceil \left(\frac{4\widetilde{C}_*^{\text{SE}}}{\varepsilon} \right)^{(3H+1)/3} \right\rceil, B_\varepsilon := \left\lceil \frac{8A_{\max}}{\kappa \zeta_\varepsilon} \left(n_\varepsilon + u_\varepsilon \right) \right\rceil$, where $M_\varepsilon := K_\varepsilon m A_{\max} S, \xi_\varepsilon := \min\left\{ \frac{\varepsilon}{8H}, \sqrt{\frac{\varepsilon}{32\eta_0 H}} \right\}, u_\varepsilon := \log \frac{4M_\varepsilon}{\delta}, n_\varepsilon := \left\lceil \frac{H^2}{2\xi_\varepsilon^2} u_\varepsilon \right\rceil$. Then, running Algorithm 2 for every player $i \in [m]$, for $N_\varepsilon = \widetilde{O}\left(\frac{A_{\max} H^2}{\kappa \varepsilon} \left(\frac{H^2}{\xi_\varepsilon^2} + 1 \right) \left(\frac{4\widetilde{C}_*^{\text{SE}}}{\varepsilon} \right)^{(3H+1)/3} \right)$ episodes results in $\text{Reg}_{N_\varepsilon}^{\text{gen},i} \le \varepsilon N_\varepsilon$, w.p. at least $1-\delta$. Consequently, the empirical distribution $\widehat{\Pi}_{N_\varepsilon}$ is an ε-CCE.*

Proof. We refer to Appendix C for the proof. □

Thus, our regret guarantees for the Algorithm 2 inherits the same exponential dependence on the horizon length as in the full-feedback finite-horizon case, up to an additional polynomial overhead arising from exploration, and value-function estimation.

3 Approximating Equilibrium in Discounted Markov Games

In this section, we extend the finite-horizon non-discounted Markov game regret guarantees from Section 2 to discounted case. This extension is independent of the feedback model, and it only requires a finite-horizon regret guarantee on the truncated discounted game to obtain a *sub-exponential* bound.

3.1 A TRANSFER PRINCIPLE

For a given IHdMG, a policy in Π^{gen} cannot be represented by a finite-sized object, and computing stationary equilibria of discounted Markov games is PPAD-hard (Daskalakis et al., 2023). We therefore use time-truncated policies and *sparse* distributions over finite-horizon product-policy prefixes. We extend each prefix to a policy of size H.

Under discount, contribution of events that happen at later time are discounted exponentially under the objective functions of players. By using time-truncation, we transfer the regret guarantees obtained in Section 2 to discounted Markov games.

Theorem 3. *Let $\bar{H} \in \mathbb{N}_+ \cup \{\infty\}$, $\gamma \in (0,1)$, and fix $L \le \bar{H}$. Suppose that a learning procedure is run on the game $G^{[L]}_{\gamma,\bar{H}}$ for N episodes, and outputs $\{\boldsymbol{\pi}_1^{[L]}, \dots, \boldsymbol{\pi}_N^{[L]}\} \in \prod_{t\in[N]} \Pi^{\mathrm{markov},L}$. If for every player $i \in [m]$, we have $\sup_{\nu^i \in \Pi_i^{\mathrm{gen},L}} \frac{1}{N}\sum_{n=1}^N \left[J^\gamma_{i,L}(\boldsymbol{\pi}_n^{[L]}) - J^\gamma_{i,L}(\nu^i \odot \boldsymbol{\pi}_n^{[L],-i})\right] \le \varepsilon_{\mathrm{alg}}$, for each $n \in [N]$, with $\bar{\boldsymbol{\pi}}_n := \mathrm{Ext}_L(\boldsymbol{\pi}_n^{[L]})$ it holds $\sup_{i\in[m],\mu^i\in\Pi_i^{\mathrm{gen},\bar{H}}} \frac{1}{N}\sum_{n=1}^N \left[J^\gamma_{i,\bar{H}}(\bar{\boldsymbol{\pi}}_n) - J^\gamma_{i,\bar{H}}(\mu^i \odot \bar{\boldsymbol{\pi}}_n^{-i})\right] \le \varepsilon_{\mathrm{alg}} + \tau_{L,\bar{H}}(\gamma)$. Equivalently, $\frac{1}{N}\sum_{n=1}^N \delta_{\bar{\boldsymbol{\pi}}_n}$ is an $(\varepsilon_{\mathrm{alg}} + \tau_{L,\bar{H}}(\gamma))$-discounted CCE.*

Proof. We refer to Appendix D for the proof. □

For the rest of this section, by identifying a suitable time truncation for the target accuracy, we extend our results to learning frameworks by applying Algorithms 1 and 2 to time-truncated games.

3.2 FULL-FEEDBACK AND PARTIAL-FEEDBACK CONSEQUENCES

Let $C^{\mathrm{SE}}_*(L)$ denote the full-feedback constant in Theorem 1 with the horizon L, and set $A_{\log} := \max_{i\in[m]} \log(|A_i|+1)$. Furthermore, for $(\gamma, \bar{H}, \varepsilon) \in (0,1)\times(\mathbb{N}_+\cup\{\infty\})\times(0,1]$, define $H^\gamma_\varepsilon := \left\lceil \frac{\log\left(2/((1-\gamma)\varepsilon)\right)}{\log(1/\gamma)} \right\rceil$, $L_\varepsilon := \bar{H} \wedge H^\gamma_\varepsilon, T_\varepsilon := \left\lceil \left(\frac{2C^{\mathrm{SE}}_*(L_\varepsilon)}{\varepsilon}\right)^{(3L_\varepsilon+1)/3} \right\rceil$. Then, we have the following guarantee,

Corollary 1. *Run Algorithm 1 for T_ε iterations on $G^{[L_\varepsilon]}_{\gamma,\bar{H}}$, producing iterates $\{\boldsymbol{\pi}_t^{[L_\varepsilon]}\}_{t=1}^{T_\varepsilon}$, and define $\bar{\boldsymbol{\pi}}_t := \mathrm{Ext}_{L_\varepsilon}\left(\boldsymbol{\pi}_t^{[L_\varepsilon]}\right) \in \Pi^{\mathrm{markov},\bar{H}}$. Then $\frac{1}{T_\varepsilon}\sum_{t=1}^{T_\varepsilon} \delta_{\bar{\boldsymbol{\pi}}_t}$ is an ε-discounted CCE of $G_{\gamma,\bar{H}}$. Moreover, $T_\varepsilon \le \left[\left(A_{\log}\eta_0^{-1} + mA_{\log} + m^2 + 1\right)\min\{\bar{H}, 1+\log(1/\varepsilon)\}^7\varepsilon^{-1}\right]^{O\left(\min\{\bar{H},1+\log(1/\varepsilon)\}\right)}$.*

Proof. We refer to Appendix D.1 for the proof. □

In the partial-feedback case by $\widetilde{C}^{\mathrm{SE}}_*(L)$ we denote the $\widetilde{C}^{\mathrm{SE}}_*$ from Theorem 2 with horizon L. Moreover, we write $\mathcal{S}_L := \bigsqcup_{h=1}^{L+1} S_h$, and $S_L := |\mathcal{S}_L|$. For a target finite-horizon accuracy $\frac{\epsilon}{2} = \bar{\varepsilon}$, horizon L, and confidence level $\delta \in (0,1)$, choose $K_{\bar{\varepsilon},L} := \left\lceil \left(4\widetilde{C}^{\mathrm{SE}}_*(L)\bar{\varepsilon}^{-1}\right)^{(3L+1)/3} \right\rceil$, $\zeta_{\bar{\varepsilon},L} := \frac{\bar{\varepsilon}}{8L^2}$, $B_{\bar{\varepsilon},L,\delta} := \left\lceil \frac{8A_{\max}}{\kappa\zeta_{\bar{\varepsilon},L}}\left(n_{\bar{\varepsilon},L,\delta} + u_{\bar{\varepsilon},L,\delta}\right)\right\rceil$, where, $M_{\bar{\varepsilon},L} := K_{\bar{\varepsilon},L} m A_{\max} S_L$, $u_{\bar{\varepsilon},L,\delta} := \log\frac{4M_{\bar{\varepsilon},L}}{\delta}$, $n_{\bar{\varepsilon},L,\delta} := \left\lceil \frac{L^2}{2\xi^2_{\bar{\varepsilon},L}} u_{\bar{\varepsilon},L,\delta} \right\rceil$, $\xi_{\bar{\varepsilon},L} := \min\left\{\frac{\bar{\varepsilon}}{8L}, \sqrt{\frac{\bar{\varepsilon}}{32\eta_0 L}}\right\}$. Then, we have the following guarantee.

Corollary 2. *Assume that $G^{[L_\varepsilon]}_{\gamma,\bar{H}}$ is $(\kappa, \zeta_{\bar{\varepsilon},L_\varepsilon})$-reachable. Run Algorithm 2 on $G^{[L_\varepsilon]}_{\gamma,\bar{H}}$ with parameters $K_{\bar{\varepsilon},L_\varepsilon}$, $B_{\bar{\varepsilon},L_\varepsilon,\delta}$, and $\zeta_{\bar{\varepsilon},L_\varepsilon}$ for $N_{\varepsilon,\bar{H},\delta}$ episodes. For each $k \in [K_{\bar{\varepsilon},L_\varepsilon}]$, let $\bar{\pi}_k := \mathrm{Ext}_{L_\varepsilon}\left(\pi_k^{[L_\varepsilon]}\right) \in \Pi^{\mathrm{markov},\bar{H}}$. Then, with probability at least $1-\delta$, for every player $i \in [m]$, $\frac{1}{N_{\varepsilon,\bar{H},\delta}}\sum_{t=1}^{N_{\varepsilon,\bar{H},\delta}} \delta_{\bar{\pi}_t}$ is an ε-approximate CCE. Consequently, for constants defined as $\mathcal{C}_{\mathrm{band}} := \frac{2A_{\log}}{\eta_0} + 12mA_{\log} + \eta_0 + 288m^2\eta_0^3$, $\mathcal{L}_{\bar{H},\varepsilon} := \min\{\bar{H}, 1+\log(1/\varepsilon)\}$, and $\gamma < 1$, we have $N_{\varepsilon,\bar{H},\delta} \le \left[\frac{A_{\max}\mathcal{L}^2_{\bar{H},\varepsilon}}{\kappa\varepsilon}\left(1 + \frac{\mathcal{L}^4_{\bar{H},\varepsilon}}{\varepsilon^2} + \frac{\eta_0\mathcal{L}^3_{\bar{H},\varepsilon}}{\varepsilon}\right)\left(\mathcal{L}_{\bar{H},\varepsilon}\log\frac{\mathcal{C}_{\mathrm{band}}\mathcal{L}^7_{\bar{H},\varepsilon}}{\varepsilon} + \log\frac{mA_{\max}S}{\delta}\right)\right]\left[\frac{\mathcal{C}_{\mathrm{band}}\mathcal{L}^7_{\bar{H},\varepsilon}}{\varepsilon}\right]^{O(\mathcal{L}_{\bar{H},\varepsilon})}$.*

Proof. We refer to Appendix D.2 for the proof. □

4 Computational Lower Bound for Discounted Markov Games

In this section, we provide computational lower bounds for the task of finding a CCE in discounted Markov games to complement the quasi-polynomial upper bounds we have found in the earlier sections. To obtain these results, we extend the *sparse Markov CCE* studied in Foster et al. (2023) for finite-horizon undiscounted Markov games.

Definition 8. *Let $\gamma \in (0,1]$ be a discount factor (with 1 being excluded in the infinite-horizon case). Given m-player Markov game $G_{\gamma,\bar{H}}$ and parameters $T \in \mathbb{N}$ and $\varepsilon > 0$, which may depend on the size of $G_{\gamma,\bar{H}}$, the (T,ε)-DiscSparseMarkovCCE$_\gamma^{\mathrm{gen}}$ is the problem of finding a nonempty list $\mathcal{L} = (\boldsymbol{\sigma}^{(1)},\dots,\boldsymbol{\sigma}^{(r)}), 1 \le r \le T$, $\boldsymbol{\sigma}^{(t)} \in \Pi^{\mathrm{markov},\bar{H}}$, such that the policy $\widehat{\boldsymbol{\sigma}}_r := \frac{1}{r}\sum_{t=1}^{r} \delta_{\boldsymbol{\sigma}^{(t)}} \in \Delta\big(\Pi^{\mathrm{gen},\bar{H}}\big)$ is an ε-approximate discounted CCE of $G_{\gamma,\bar{H}}$, i.e., $\sup_{\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}} \sum_{t=1}^{r} \Big[J_{i,\bar{H}}^{\gamma}\big(\boldsymbol{\sigma}^{(t)}\big) - J_{i,\bar{H}}^{\gamma}\big(\mu^i \odot \boldsymbol{\sigma}^{(t),-i}\big)\Big] \le \varepsilon r$ for every player $i \in [m]$. We call such a sequence a T-sparse discounted CCE.*

Definition 8 formulates the equilibrium-computation objective studied in this section as a search problem over finite sequences of product Markov policy profiles.

Proposition 1 (Informal)**.** *Let $G_{\gamma,\bar{H}}$ be an m-player discounted Markov game, and fix $\varepsilon \in (0,1]$. Algorithms 1 and 2 solve the corresponding sparse CCE problem.*

Note that (T,ε)-DiscSparseMarkovCCE$_\gamma^{\mathrm{gen}}$ is a relaxation of the radically uncoupled learning task addressed by Algorithms 1 and 2 by extending policies. The search problem only asks for a finite sequence of product Markov policy profiles whose uniform empirical distribution is an ε-approximate discounted CCE. It does not require independency of the players; in particular, a centralized algorithm with full-information is also permitted as stated in Foster et al. (2023). Therefore, any lower bound provided for this search problem, is also valid for our radically uncoupled learning problem. Our main hardness result is the following, which depends on the Exponential Time Hypothesis for PPAD:

Theorem 4 (Informal)**.** *Given discount factor $\gamma \in (0,1) \cap \mathbb{Q}$ and support exponent C, if* ETH *for* PPAD *holds, then there exists a family of 2-player IHdMGs such that no polynomial-time algorithm solves $(|\widetilde{F}|^C, |\widetilde{F}|^{-K})$-DiscSparseMarkovCCE$_\gamma^{\mathrm{gen}}$ for all games $\widetilde{F}$ in this family, for any sufficiently large $K = K(\gamma, C)$.*

This result demonstrates that with our current understanding of computational complexity, it is *unlikely* that there exists a polynomial time algorithm for discounted Markov games. In particular, a quasi-polynomial time algorithm (which we have constructed) is *about* the best one can hope for. A similar result also holds for FHdMGs.

Corollary 3 (Informal)**.** *For any discount factor $\gamma \in (0,1)$ and support exponent $C_{\mathrm{fh}} > 0$, one cannot solve the approximate finite-horizon sparse CCE problem in polynomial time.*

The results above rely on a Markov game extension of a special $n \times n$ bimatrix game introduced in Rubinstein (2016). The construction of this bimatrix game as well as the properties of the corresponding Markov game is technical, so we omit details and refer to Appendix E for the complete proofs and formal statements of Proposition 1, Theorem 4 and Corollary 3. The assumption ETH for PPAD is needed to enforce a complexity lower-bound, which relies on (Rubinstein, 2016, Theorem 1.2).

5 Conclusion

In this work, we have shown that, under the so-called "ETH for PPAD", there does not exist any polynomial time algorithm for computing CCE in discounted general-sum Markov games. We have complemented this hardness result by providing a quasi-polynomial-time algorithm in the online self-play setting under both full- and partial feedback.

Our work suggests several future directions. First, a gap remains between our upper bounds and the computational lower bounds established in this paper; determining the precise complexity of computing sparse CCEs in discounted Markov games remains an open problem. Second, it would be interesting to determine whether polynomial-time radically uncoupled online-learning algorithms exist for general-sum Markov games when the policy classes are restricted to Markov policies.

AI USE STATEMENT

In this paper, we use large language models (LLMs). For example, by correcting grammatical errors, and producing codes needed for the numerical experiments.

REFERENCES

Yasin Abbasi-Yadkori, Peter L. Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS)*, 2013.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pp. 263–274, 2008.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, 2017.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Gürdal Arslan and Serdar Yüksel. Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4):1545–1558, 2017.

Yakov Babichenko, Christos Papadimitriou, and Aviad Rubinstein. Can almost everybody be almost happy? In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, 2016.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Tamer Başar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory*. SIAM, 1998.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.

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

Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *Journal of the ACM*, 68(4):1–35, 2021.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(2):156–172, 2008.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum Markov games with bandit feedback. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Near-optimal policy optimization for correlated equilibrium in general-sum Markov games. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In *Proceedings of the 26th Conference on Neural Information Processing Systems (NeurIPS)*, 2012.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. Last-iterate convergence of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2409.01447*, 2024.

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

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 235–254, 2011.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of Markov equilibrium in stochastic games. In *Proceedings of the 36th Conference on Learning Theory (COLT)*, pp. 4180–4234, 2023.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum Markov games. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum Markov games. *arXiv preprint arXiv:2207.14211*, 2025.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Dean P. Foster and H. Peyton Young. Regret testing: Learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1:341–367, 2006.

Dylan J. Foster, Noah Golowich, and Sham M. Kakade. Hardness of independent learning and sparse equilibrium computation in Markov games. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory – 2nd European Conference, EuroCOLT 1995, Proceedings*, volume 904 of *Lecture Notes in Computer Science*, pp. 23–37. Springer, 1995.

Amy Greenwald and Keith Hall. Correlated-Q learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 2003.

James Hannan. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games*, volume 3, pp. 97–139. Princeton University Press, 1957.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

Junling Hu and Michael P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069, 2003.

Shansong Huang, Virginia R. Young, and Bin Zou. Optimal investment, consumption, and insurance strategies under claim habit. Working paper, 2026.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*, 49(4): 2295–2322, 2023.

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R. McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, 13:7214, 2022.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4–5):421–436, 2018.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 157–163, 1994.

Michael L. Littman. Friend-or-foe Q-learning in general-sum games. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 2001.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning Markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Kleanthis Malialis and Daniel Kudenko. Distributed response to network intrusions using multiagent reinforcement learning. *Engineering Applications of Artificial Intelligence*, 41:270–284, 2015.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games and Applications*, 13(1):165–186, 2023.

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, and Tamer Başar. $\widetilde{O}(T^{-1})$ convergence to (coarse) correlated equilibria in full-information general-sum Markov games. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference (L4DC)*, pp. 361–374, 2024.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations (ICLR)*, 2019.

A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Christos Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and System Sciences*, 48(3):498–532, 1994.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E. Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 30215–30248, 2025.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning Nash equilibrium for general-sum Markov games from batch data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS)*, 2013.

Aviad Rubinstein. Settling the complexity of computing approximate two-player Nash equilibria. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 258–265, 2016.

Muhammed O. Sayin, Kaiqing Zhang, David S. Leslie, Tamer Başar, and Asuman Ozdaglar. Decentralized Q-learning in zero-sum Markov games. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. In *Learning, Inference and Control of Multi-Agent Systems Workshop at NeurIPS*, 2016.

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

Tareq Si Salem, Gözde Özcan, Iasonas Nikolaou, Evimaria Terzi, and Stratis Ioannidis. Online submodular maximization via online convex optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15038–15046, 2024.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

Ashkan Soleymani, Georgios Piliouras, and Gabriele Farina. Faster rates for no-regret learning in general games via cautious optimism. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 518–529, 2025.

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

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown Markov games. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Ziyu Wan, Yunxiang Li, Xiaoyu Wen, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, and Ying Wen. ReMA: Learning to meta-think for LLMs with multi-agent reinforcement learning. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, 2018.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, 2021.

Asrın Efe Yorulmaz and Tamer Başar. Near optimal convergence to coarse correlated equilibrium in general-sum Markov games. *arXiv preprint arXiv:2511.02157*, 2025.

Hengxi Zhang, Zhendong Shi, Yuanquan Hu, Wenbo Ding, Ercan E. Kuruoglu, and Xiao-Ping Zhang. Optimizing trading strategies in quantitative markets using multi-agent reinforcement learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 136–140, 2024.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Kyriakos G. Vamvoudakis, Bahare Kiumarsi, Sarangapani Jagannathan, and Hamidreza Modares (eds.), *Handbook of Reinforcement Learning and Control*, volume 325 of *Studies in Systems, Decision and Control*, pp. 321–384. Springer, Cham, 2021.

Kaiqing Zhang, Sham M. Kakade, Tamer Başar, and Lin F. Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24:1–53, 2023.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18), 2022.

A Related Work

The connection between no-regret learning and equilibrium computation is a cornerstone of learning in games. In repeated normal-form games, if every player uses an external no-regret algorithm with regret $\mathcal{O}(T^{\alpha})$, then the empirical distribution of play converges to an $\mathcal{O}(T^{\alpha-1})$-approximate CCE; in two-player zero-sum games this further implies convergence to the saddle-point/Nash value, while swap or internal regret yields CE in general-sum games (Hannan, 1957; Hart & Mas-Colell, 2000; Cesa-Bianchi & Lugosi, 2006; Blum & Mansour, 2007). Classical online learning algorithms such as multiplicative weights, online mirror descent, and follow-the-regularized-leader provide the underlying no-regret guarantees in the fully adversarial setting, yielding the standard $\widetilde{\mathcal{O}}(T^{-1/2})$ convergence rate to equilibrium (Freund & Schapire, 1995; Nemirovski & Yudin, 1983; Abernethy et al., 2008).

It is known that agents in the adversarial setting can be *pessimistic* when their learning algorithms interact with one another in self-play Farina et al. (2022). Starting with the optimistic dynamics of Daskalakis et al. (2011) for two-player zero-sum games and continuing through the RVU framework of Syrgkanis et al. (2015), optimism and stability have emerged as central tools for leveraging this predictability. In normal-form games, optimistic variants of FTRL and OMD, together with their regularized extensions, attain convergence rates to coarse correlated and correlated equilibria that improve upon the baseline $\widetilde{\mathcal{O}}(T^{-1/2})$ rate, with recent results approaching $O(1/T)$ rates in general-sum normal-form games (Rakhlin & Sridharan, 2013; Daskalakis et al., 2021; Anagnostides et al., 2022; Soleymani et al., 2025). Our work is motivated by this optimistic self-play perspective. However, extending these ideas to Markov games introduces a challenge absent from normal-form games: deviations in the joint policy profile affect not only instantaneous losses but also future state occupancies.

Early learning approaches in Markov games were mostly value-based. Specifically, the convergence of Nash-Q learning (Hu & Wellman, 2003) and Friend-or-Foe Q-learning (Littman, 2001), is established only under restrictions on the games encountered during learning. Related approaches based on CE computation further developed this equilibrium-computation perspective while maintaining similar structural requirements (Greenwald & Hall, 2003). A separate line of research on decentralized Q-learning with convergence guarantees for weakly acyclic stochastic teams and games also emerged, which includes stochastic team problems, but again relies on strong structural assumptions on the games (Arslan & Yüksel, 2017).

Compared to the general-sum case, the two-player zero-sum structure in Markov games has been studied more thoroughly. There exist tractable algorithms with optimism and value-iteration variants, achieve finite sample-complexity guarantees for learning approximate Nash equilibria (Bai et al., 2020; Zhang et al., 2023). Policy-optimization methods have also been analyzed in the decentralized learning setting, including independent policy-gradient and optimistic gradient-descent/ascent approaches (Daskalakis et al., 2020; Wei et al., 2021). Furthermore, Cai et al. (2023); Chen et al. (2024) have developed uncoupled algorithms for two-player zero-sum Markov games with bandit feedback and finite-time last-iterate guarantees, relaxing the coordination and prior-knowledge constraints used in Wei et al. (2021). From the value-learning angle, Sayin et al. (2021) develops a radically uncoupled Q-learning algorithm for discounted zero-sum Markov games with asymptotic convergence guarantees.

Even for multi-player general-sum normal-form games, Nash equilibrium computation is known to be intractable Daskalakis et al. (2009); and thus in Markov games provably efficient algorithms in the existing literature targets either CE or CCE. For instance, in sample-based MARL literature, V-learning (Jin et al., 2023; Song et al., 2022; Mao & Başar, 2023), SPoCMAR (Daskalakis et al., 2023) and related algorithms are utilized to compute CCE/CE policies in a decentralized setting and avoid the exponential dependence on the number of agents that appears in centralized settings Bai et al. (2020); Liu et al. (2021). Near-optimal $\widetilde{\mathcal{O}}(1/T)$ convergence rates for both CCE and CE obtained in the full-information setting by using either value-based or stage-based policy-optimization schemes (Cai et al., 2024; Mao et al., 2024; Yorulmaz & Başar, 2025). These results require post-processing of the policies after the learning phase, which relies on the so-called *shared randomness*. In contrast to our work, their learning phases do not provide any online regret guarantees, making their operating setting simpler than ours since they allow coordination mechanisms among the players.

B Convergence Analysis of Algorithm 1

B.1 Properties of smoothed entropy

Lemma 1. *Fix a player $i \in [m]$ and a smoothing parameter $\lambda_i > 0$. Define the smoothed negative-entropy regularizer*

$$\Psi_i(x) := \sum_{a \in A_i} \big(x(a) + \lambda_i\big) \log\big(x(a) + \lambda_i\big), \qquad x \in \Delta(A_i).$$

Then,

$$D_i(x, y) = \sum_{a \in A_i} \big(x(a) + \lambda_i\big) \log \frac{x(a) + \lambda_i}{y(a) + \lambda_i}. \tag{2}$$

Moreover, defining $\rho_i := 1 + |A_i|\lambda_i$, $\bar{\rho} := \max_{i \in [m]} \rho_i$, and $B_i := \rho_i \log \frac{1+\lambda_i}{\lambda_i}$, the following properties hold:

1. ***Boundedness.*** *For all $x, y \in \Delta(A_i)$, $D_i(x, y) \le B_i$.*
2. ***Strong convexity.*** *The function Ψ_i is $1/\rho_i$-strongly convex with respect to the $\|\cdot\|_1$ on $\Delta(A_i)$. Equivalently, for all $x, y \in \Delta(A_i)$, $D_i(x, y) \ge \frac{1}{2\rho_i}\|x - y\|_1^2$.*

Proof. For each $a \in A_i$, we have

$$\nabla_a \Psi_i(x) = \log\big(x(a) + \lambda_i\big) + 1.$$

Therefore,

$$\begin{aligned} D_i(x, y) = &\sum_{a \in A_i} \big(x(a) + \lambda_i\big) \log\big(x(a) + \lambda_i\big) - \sum_{a \in A_i} \big(y(a) + \lambda_i\big) \log\big(y(a) + \lambda_i\big) \\ &- \sum_{a \in A_i} \big(\log(y(a) + \lambda_i) + 1\big)\big(x(a) - y(a)\big). \end{aligned}$$

Since $x, y \in \Delta(A_i)$, we have $\sum_{a \in A_i} \big(x(a) - y(a)\big) = 0$. Therefore, we obtain (2). Next, we prove that D_i is a bounded function. For every $a \in A_i$, $x(a) + \lambda_i \le 1 + \lambda_i$, $y(a) + \lambda_i \ge \lambda_i$. Thus,

$$0 \le \frac{x(a) + \lambda_i}{y(a) + \lambda_i} \le \frac{1 + \lambda_i}{\lambda_i} \implies D_i(x, y) \le \sum_{a \in A_i} \big(x(a) + \lambda_i\big) \log \frac{1 + \lambda_i}{\lambda_i}.$$

Since

$$\sum_{a \in A_i} \big(x(a) + \lambda_i\big) = 1 + |A_i|\lambda_i = \rho_i, \tag{3}$$

D_i is bounded above by $\rho_i \log\left((1 + \lambda_i)/\lambda_i\right) = B_i$. Next, note that the Hessian of Ψ_i is diagonal matrix and given by $\nabla^2 \Psi_i(x) = \operatorname{diag}\left(1/(x(a) + \lambda_i)\right)_{a \in A_i}$. Thus, for any $z \in \mathbb{R}^{|A_i|}$, by Cauchy–Schwarz,

$$\|z\|_1^2 = \left(\sum_{a \in A_i} |z(a)|\right)^2 \le \left(\sum_{a \in A_i} (x(a) + \lambda_i)\right)\left(\sum_{a \in A_i} \frac{z(a)^2}{x(a) + \lambda_i}\right) = \rho_i \, z^\top \nabla^2 \Psi_i(x) z,$$

where the last inequality follows from (3). Therefore, Ψ_i is $1/\rho_i$-strongly convex with respect to the ℓ_1-norm, which implies $D_i(x, y) \ge \frac{1}{2\rho_i}\|x - y\|_1^2$. This completes the proof. □

B.2 Properties of OOMD with smoothed entropy

Lemma 2. *Fix player i, a base point $y \in \Delta(A_i)$, and $g, g' \in \mathbb{R}^{A_i}$. Define*

$$P_y(g) := \operatorname*{argmin}_{x \in \Delta(A_i)} \left\{\eta \langle g, x\rangle + D_i(x, y)\right\}.$$

Then

$$\|P_y(g) - P_y(g')\|_1 \le \eta \rho_i \|g - g'\|_\infty.$$

In particular,

$$\|P_y(g) - y\|_1 \le \eta \rho_i \|g\|_\infty.$$

Proof. The first claim follows from Proposition B.4 in Mertikopoulos et al. (2019). The second claim follows by taking $g' = 0$ and observing that $P_y(0) = y$, where P_y is a well-defined function. □

Lemma 3. *Let $\mathcal{X} \subseteq \mathbb{R}^{A_i}$ be a nonempty closed convex set, and let $\Psi : \mathcal{X} \to \mathbb{R}$ be a differentiable strictly convex function. Further, let $D(x, y) := \Psi(x) - \Psi(y) - \langle \nabla\Psi(y), x - y\rangle$, $x, y \in \mathcal{X}$, denote the associated Bregman divergence. Now, fix $g \in \mathbb{R}^d$ and $y \in \mathcal{X}$, and define $x^+ := \operatorname{argmin}_{x\in\mathcal{X}} \{\eta \langle g, x\rangle + D(x, y)\}$. Then, for every $u \in \mathcal{X}$,*

$$\eta \left\langle x^+ - u, g\right\rangle \le D(u, y) - D(u, x^+) - D(x^+, y).$$

Proof. The claim follows from Lemma 3.2. in Chen & Teboulle (1993).

□

Lemma 4. *Let $x_t, \widetilde{x}_t \in \Delta(A_i)$ be generated by $x_t = \operatorname{argmin}_{x\in\Delta(A_i)} \{\eta \langle M_t, x\rangle + D(x, \widetilde{x}_{t-1})\}$, $\widetilde{x}_t = \operatorname{argmin}_{x\in\Delta(A)} \{\eta \langle \ell_t, x\rangle + D(x, \widetilde{x}_{t-1})\}$, where $M_t = \ell_{t-1}$ and $\ell_0 = 0$. Suppose that the regularizer chosen for the iterations is the smoothed entropy regularizer with parameters $\lambda_i, \rho_i = 1 + |A_i|\lambda_i$, and Bregman divergence at most B. Then, for any $u \in \Delta(A_i)$ and any weights $q_t \in [0, 1]$,*

$$\sum_{t=1}^{T} q_t \langle x_t - u, \ell_t\rangle \le \frac{B(q_1 + \mathrm{TV}(q))}{\eta} + \frac{\eta\rho_i}{2} \sum_{t=1}^{T} q_t \|\ell_t - \ell_{t-1}\|_\infty^2,$$

where we define $\mathrm{TV}(q) := \sum_{t=1}^{T-1} |q_{t+1} - q_t|$.

Proof. Throughout the proof, we follow steps similar those in Rakhlin & Sridharan (2013); Syrgkanis et al. (2015); Erez et al. (2023).

Since $\widetilde{x}_t = \operatorname{argmin}_{x\in\Delta(A_i)} \{\eta \langle \ell_t, x\rangle + D(x, \widetilde{x}_{t-1})\}$, from Lemma 3, for every $u \in \Delta(A_i)$,

$$\langle \widetilde{x}_t - u, \ell_t\rangle \le \frac{D(u, \widetilde{x}_{t-1}) - D(u, \widetilde{x}_t)}{\eta} - \frac{D(\widetilde{x}_t, \widetilde{x}_{t-1})}{\eta}. \tag{4}$$

Next, decompose $\langle x_t - u, \ell_t\rangle = \langle \widetilde{x}_t - u, \ell_t\rangle + \langle x_t - \widetilde{x}_t, \ell_t\rangle$. Substituting (4) leads to

$$\langle x_t - u, \ell_t\rangle \le \frac{D(u, \widetilde{x}_{t-1}) - D(u, \widetilde{x}_t)}{\eta} + \langle x_t - \widetilde{x}_t, \ell_t\rangle - \frac{D(\widetilde{x}_t, \widetilde{x}_{t-1})}{\eta}. \tag{5}$$

Now split

$$\langle x_t - \widetilde{x}_t, \ell_t\rangle = \langle x_t - \widetilde{x}_t, \ell_t - M_t\rangle + \langle x_t - \widetilde{x}_t, M_t\rangle. \tag{6}$$

Since $x_t = \operatorname{argmin}_{x\in\Delta(A_i)} \{\eta \langle M_t, x\rangle + D(x, \widetilde{x}_{t-1})\}$, another application of Lemma 3, with comparator $u = \widetilde{x}_t$, yields

$$\langle x_t - \widetilde{x}_t, M_t\rangle \le \frac{D(\widetilde{x}_t, \widetilde{x}_{t-1})}{\eta} - \frac{D(\widetilde{x}_t, x_t)}{\eta} - \frac{D(x_t, \widetilde{x}_{t-1})}{\eta}. \tag{7}$$

Substituting (6) and (7) into (5) cancels out the term $D(\widetilde{x}_t, \widetilde{x}_{t-1})$, yielding

$$\langle x_t - u, \ell_t\rangle \le \frac{D(u, \widetilde{x}_{t-1}) - D(u, \widetilde{x}_t)}{\eta} + \langle x_t - \widetilde{x}_t, \ell_t - M_t\rangle - \frac{D(\widetilde{x}_t, x_t)}{\eta} - \frac{D(x_t, \widetilde{x}_{t-1})}{\eta}.$$

Finally, since Bregman divergences are nonnegative, we may drop the negative term, $-\frac{D(x_t, \widetilde{x}_{t-1})}{\eta}$, to obtain

$$\langle x_t - u, \ell_t\rangle \le \frac{D(u, \widetilde{x}_{t-1}) - D(u, \widetilde{x}_t)}{\eta} + \langle x_t - \widetilde{x}_t, \ell_t - M_t\rangle - \frac{D(\widetilde{x}_t, x_t)}{\eta}.$$

Due to Lemma 1, Ψ_i is $1/\rho_i$-strongly convex with respect to ℓ_1, $D(\widetilde{x}_t, x_t) \ge \frac{1}{2\rho_i} \|x_t - \widetilde{x}_t\|_1^2$. Therefore, by Young's inequality,

$$\langle x_t - \widetilde{x}_t, \ell_t - M_t\rangle - \frac{D(\widetilde{x}_t, x_t)}{\eta} \le \|x_t - \widetilde{x}_t\|_1 \|\ell_t - M_t\|_\infty - \frac{1}{2\eta\rho_i} \|x_t - \widetilde{x}_t\|_1^2 \le \frac{\eta\rho_i}{2} \|\ell_t - M_t\|_\infty^2.$$

Since $M_t = \ell_{t-1}$,

$$\langle x_t - u, \ell_t \rangle \le \frac{D(u, \widetilde{x}_{t-1}) - D(u, \widetilde{x}_t)}{\eta} + \frac{\eta \rho_i}{2} \|\ell_t - \ell_{t-1}\|_\infty^2 .$$

Multiplying by q_t and summing up yields

$$\sum_{t=1}^{T} q_t \langle x_t - u, \ell_t \rangle \le \frac{1}{\eta} \sum_{t=1}^{T} q_t (D_{t-1} - D_t) + \frac{\eta \rho_i}{2} \sum_{t=1}^{T} q_t \|\ell_t - \ell_{t-1}\|_\infty^2 ,$$

where $D_t := D(u, \widetilde{x}_t)$. Since $0 \le D_t \le B$, using summation by parts we get,

$$\sum_{t=1}^{T} q_t (D_{t-1} - D_t) = q_1 D_0 - q_T D_T + \sum_{t=1}^{T-1} (q_{t+1} - q_t) D_t \le B q_1 + B \,\mathrm{TV}(q).$$

This completes the proof. □

B.3 First order path of the played policies

Lemma 5. *Let $\pi^i_{t+1,h}(\cdot \mid s) \in \Pi_i^{\mathrm{markov}}$ be the policies stated in Algorithm 1. For every player i, layer h, state $s \in S_h$, and episode t,*

$$\left\|\pi^i_{t+1,h}(\cdot \mid s) - \pi^i_{t,h}(\cdot \mid s)\right\|_1 \le 3\rho_i H \eta_h.$$

Proof. Write $x_t = x^{i,s}_{t,h}$ and $\widetilde{x}_t = \widetilde{x}^{i,s}_{t,h}$. By Lemma 2, applied with base $\widetilde{x}_t$ and loss $Q^i_t(s, \cdot)$,

$$\|x_{t+1} - \widetilde{x}_t\|_1 \le \eta_h \rho_i \left\|Q^i_{t,h}(s, \cdot)\right\|_\infty \le \eta_h \rho_i H.$$

Again applying Lemma 2 to compare the reference point $\widetilde{x}_t$ and the played point x_t, both generated from the common base point $\widetilde{x}_{t-1}$ using the losses $Q^i_t(s, \cdot)$ and $Q^i_{t-1}(s, \cdot)$, respectively, we obtain:

$$\|\widetilde{x}_t - x_t\|_1 \le \eta_h \rho_i \left\|Q^i_{t,h}(s, \cdot) - Q^i_{t-1,h}(s, \cdot)\right\|_\infty \le 2\eta_h \rho_i H,$$

because both Q-functions lie in $[0, H]^{A_i}$. The triangle inequality gives $\|x_{t+1} - x_t\|_1 \le 3\rho_i H \eta_h$. Since $x_t = \pi^i_{t,h}(\cdot \mid s)$, the claim follows. □

For later use, we define $\psi^i_{t,h} := \max_{s \in S_h} \left\|\pi^i_{t+1,h}(\cdot \mid s) - \pi^i_{t,h}(\cdot \mid s)\right\|_1$. Therefore, Lemma 5 gives $\psi^i_{t,h} \le 3\bar{\rho} H \eta_h$.

B.4 History occupancy variation

Lemma 6. *Let $d^{i,\mu}_{t,h} \in \Delta(\mathcal{H}_h)$ denote the distribution over histories at the beginning of layer h induced by player i using $\mu^i \in \Pi_i^{\mathrm{markov}}$ and the opponents using $\boldsymbol{\pi}^{-i}_t$. Fix a player i, a layer h, and a policy $\mu_i \in \Pi_i^{\mathrm{gen}}$. Then, we have,*

$$\mathrm{TV}^{i,\mu}_h := \sum_{t=1}^{T-1} \left\|d^{i,\mu}_{t+1,h} - d^{i,\mu}_{t,h}\right\|_1 \le 3m\bar{\rho} H T \sum_{\ell=1}^{h-1} \eta_\ell.$$

Proof. Given $\tau_\ell \in \mathcal{H}_\ell$, $a_i \in A_i$, $\boldsymbol{a}_{-i} \in A_{-i}$, and $s' \in S_{\ell+1}$, define the next history, $\tau_{\ell+1} = (\tau_\ell, a_i, \boldsymbol{a}_{-i}, s') \in \mathcal{H}_{\ell+1}$. For episode t, define the history transition kernel $K^{i,\mu}_{t,\ell} : \mathcal{H}_\ell \to \Delta(\mathcal{H}_{\ell+1})$ by

$$K^{i,\mu}_{t,\ell}(\tau_{\ell+1} \mid \tau_\ell) := \sum_{\substack{a_i \in A_i,\ a_{-i} \in A_{-i},\ s' \in S_{\ell+1}: \\ \tau_{\ell+1} = (\tau_\ell, a_i, \boldsymbol{a}_{-i}, s')}} \mu^i(a_i \mid \tau_\ell)\, \boldsymbol{\pi}^{-i}_{t,\ell}(\boldsymbol{a}_{-i} \mid s(\tau_\ell))\, P_\ell(s' \mid s(\tau_\ell), a_i, \boldsymbol{a}_{-i}),$$

First, we demonstrate that $K^{i,\mu}_{t,\ell}$ is a stochastic kernel. For every fixed $\tau_\ell \in \mathcal{H}_\ell$,

$$\begin{aligned}\sum_{\tau_{\ell+1}\in\mathcal{H}_{\ell+1}} K^{i,\mu}_{t,\ell}(\tau_{\ell+1} \mid \tau_\ell) &= \sum_{a_i\in A_i}\sum_{\boldsymbol{a}_{-i}\in A_{-i}}\sum_{s'\in S_{\ell+1}} \mu^i(a_i \mid \tau_\ell)\boldsymbol{\pi}^{-i}_{t,\ell}(\boldsymbol{a}_{-i} \mid s(\tau_\ell))P_\ell(s' \mid s(\tau_\ell), \boldsymbol{a}_i, a_{-i}) \\ &= \sum_{a_i\in A_i} \mu^i(a_i \mid \tau_\ell) \sum_{\boldsymbol{a}_{-i}\in A_{-i}} \boldsymbol{\pi}^{-i}_{t,\ell}(\boldsymbol{a}_{-i} \mid s(\tau_\ell)) \sum_{s'\in S_{\ell+1}} P_\ell(s' \mid s(\tau_\ell), \boldsymbol{a}_i, a_{-i}) \\ &= 1.\end{aligned}$$

Denoting by $\delta_s \in \Delta(\mathcal{S})$ the probability distribution placing unit mass on s, we have $d^{i,\mu}_{t,1} = \delta_{s_1}$, For $h \geq 2$,

$$d^{i,\mu}_{t,h} = \delta_{s_1} K^{i,\mu}_{t,1} K^{i,\mu}_{t,2} \cdots K^{i,\mu}_{t,h-1}.$$

Here we view distributions as row vectors and kernels as row-stochastic matrices. We next compare $d^{i,\mu}_{t+1,h}$ with $d^{i,\mu}_{t,h}$. For notational simplicity, write $K_\ell := K^{i,\mu}_{t,\ell},\ K'_\ell := K^{i,\mu}_{t+1,\ell}$. Then

$$\begin{aligned}d^{i,\mu}_{t+1,h} - d^{i,\mu}_{t,h} &= \delta_{s_1}\left(K'_1K'_2\cdots K'_{h-1} - K_1K_2\cdots K_{h-1}\right) \\ &= \delta_{s_1}\left(\sum_{\ell=1}^{h-1} K_1\cdots K_{\ell-1}(K'_\ell - K_\ell)K'_{\ell+1}\cdots K'_{h-1}\right).\end{aligned}$$

Stochastic kernels are ℓ_1-contractions on signed measures, i.e., it holds that $\|\nu K\|_1 \leq \|\nu\|_1$ for every signed measure ν and every stochastic kernel K. Hence $\left\|\delta_{s_1}K_1\cdots K_{\ell-1}(K'_\ell - K_\ell)K'_{\ell+1}\cdots K'_{h-1}\right\|_1 \leq \|\delta_{s_1}K_1\cdots K_{\ell-1}(K'_\ell - K_\ell)\|_1 \leq \|K'_\ell - K_\ell\|_{\infty,1}$, where $\|K'_\ell - K_\ell\|_{\infty,1} := \max_{\tau_\ell\in\mathcal{H}_\ell}\sum_{\tau_{\ell+1}\in\mathcal{H}_{\ell+1}} |K'_\ell(\tau_{\ell+1} \mid \tau_\ell) - K_\ell(\tau_{\ell+1} \mid \tau_\ell)|$. Therefore,

$$\left\|d^{i,\mu}_{t+1,h} - d^{i,\mu}_{t,h}\right\|_1 \leq \sum_{\ell=1}^{h-1}\left\|\delta_{s_1}K_1\cdots K_{\ell-1}(K'_\ell - K_\ell)K'_{\ell+1}\cdots K'_{h-1}\right\|_1 \leq \sum_{\ell=1}^{h-1}\left\|K^{i,\mu}_{t+1,\ell} - K^{i,\mu}_{t,\ell}\right\|_{\infty,1}. \tag{8}$$

It remains to control $\left\|K^{i,\mu}_{t+1,\ell} - K^{i,\mu}_{t,\ell}\right\|_{\infty,1}$. Fix $\tau_\ell \in \mathcal{H}_\ell$ and write $s = s(\tau_\ell)$. Since μ^i and P_ℓ are the same in episodes t and $t+1$, only the opponents' product distribution changes. Therefore,

$$\begin{aligned}\left\|K^{i,\mu}_{t+1,\ell}(\cdot \mid \tau_\ell) - K^{i,\mu}_{t,\ell}(\cdot \mid \tau_\ell)\right\|_1 &= \sum_{\tau_{\ell+1}\in\mathcal{H}_{\ell+1}}\left|K^{i,\mu}_{t+1,\ell}(\tau_{\ell+1} \mid \tau_\ell) - K^{i,\mu}_{t,\ell}(\tau_{\ell+1} \mid \tau_\ell)\right| \\ &= \sum_{a_i}\sum_{\boldsymbol{a}_{-i}}\sum_{s'} \mu^i(a_i \mid \tau_\ell)P_\ell(s' \mid s, a_i, \boldsymbol{a}_{-i})\left|\boldsymbol{\pi}^{-i}_{t+1,\ell}(\boldsymbol{a}_{-i} \mid s) - \boldsymbol{\pi}^{-i}_{t,\ell}(\boldsymbol{a}_{-i} \mid s)\right| \\ &= \sum_{\boldsymbol{a}_{-i}}\left|\boldsymbol{\pi}^{-i}_{t+1,\ell}(\boldsymbol{a}_{-i} \mid s) - \boldsymbol{\pi}^{-i}_{t,\ell}(\boldsymbol{a}_{-i} \mid s)\right| \sum_{a_i}\mu^i(a_i \mid \tau_\ell)\sum_{s'}P_\ell(s' \mid s, a_i, \boldsymbol{a}_{-i}) \\ &= \left\|\boldsymbol{\pi}^{-i}_{t+1,\ell}(\cdot \mid s) - \boldsymbol{\pi}^{-i}_{t,\ell}(\cdot \mid s)\right\|_1 \leq \max_{s\in S_\ell}\left\|\boldsymbol{\pi}^{-i}_{t+1,\ell}(\cdot \mid s) - \boldsymbol{\pi}^{-i}_{t,\ell}(\cdot \mid s)\right\|_1.\end{aligned}$$

For product distributions, we have

$$\left\|\boldsymbol{\pi}^{-i}_{t+1,\ell}(\cdot \mid s) - \boldsymbol{\pi}^{-i}_{t,\ell}(\cdot \mid s)\right\|_1 \leq \sum_{j\neq i}\left\|\pi^j_{t+1,\ell}(\cdot \mid s) - \pi^j_{t,\ell}(\cdot \mid s)\right\|_1.$$

Since, $\psi^j_{t,\ell} := \max_{s\in S_\ell}\left\|\pi^j_{t+1,\ell}(\cdot \mid s) - \pi^j_{t,\ell}(\cdot \mid s)\right\|_1$, it holds that $\left\|K^{i,\mu}_{t+1,\ell} - K^{i,\mu}_{t,\ell}\right\|_{\infty,1} \leq \sum_{j\neq i}\psi^j_{t,\ell}$. For all $h > 1$, combining this with (8) gives

$$\left\|d^{i,\mu}_{t+1,h} - d^{i,\mu}_{t,h}\right\|_1 \leq \sum_{\ell=1}^{h-1}\sum_{j\neq i}\psi^j_{t,\ell}.$$

In the special case $h = 1$, the distribution is always $d^{i,\mu}_{t,1} = \delta_{s_1}$, and hence $\left\|d^{i,\mu}_{t+1,1} - d^{i,\mu}_{t,1}\right\|_1 = 0$. Finally, summing over $t = 1, \ldots, T-1$ gives

$$\mathrm{TV}^{i,\mu}_h = \sum_{t=1}^{T-1}\left\|d^{i,\mu}_{t+1,h} - d^{i,\mu}_{t,h}\right\|_1 \leq \sum_{t=1}^{T-1}\sum_{\ell=1}^{h-1}\sum_{j\neq i}\psi^j_{t,\ell} \leq \sum_{\ell=1}^{h-1}\sum_{j\neq i} 3\bar{\rho}HT\eta_\ell \leq 3m\bar{\rho}HT\sum_{\ell=1}^{h-1}\eta_\ell.$$

This completes the proof. □

B.5 VARIATION BOUNDS ON THE Q-VALUES

Lemma 7. *Fix a player i and two Markov joint policies $\boldsymbol{\pi} = (\pi^1, \dots, \pi^m)$, $\boldsymbol{\pi}' = (\pi^{1\prime}, \dots, \pi^{m\prime})$. Define $\Delta := \sum_{j=1}^m \left\|\pi^j - \pi^{j\prime}\right\|_{\infty,1}$, where $\left\|\pi^j - \pi^{j\prime}\right\|_{\infty,1} := \max_{h,s} \left\|\pi_h^j(\cdot \mid s) - \pi_h^{j\prime}(\cdot \mid s)\right\|_1$. Then, for every layer h and state s, $\left\|Q_h^{i,\boldsymbol{\pi}}(s,\cdot) - Q_h^{i,\boldsymbol{\pi}'}(s,\cdot)\right\|_\infty \le 2H^2\Delta$.*

Proof. Fix player i. Define $\Delta_h := \max_s \left\|Q_h^{i,\boldsymbol{\pi}}(s,\cdot) - Q_h^{i,\boldsymbol{\pi}'}(s,\cdot)\right\|_\infty$. It can be seen that, at layer $H+1$, both value functions are zero, and hence $\Delta_{H+1} = 0$. Now, fix $h \le H$, $s \in S_h$, and $a_i \in A_i$. Then,

$$Q_h^{i,\boldsymbol{\pi}}(s,a_i) - Q_h^{i,\boldsymbol{\pi}'}(s,a_i) = \mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h^{-i}}\left[\ell_h^i + \mathbb{E}_{s'} V_{h+1}^{i,\boldsymbol{\pi}}(s')\right] - \mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h'^{-i}}\left[\ell_h^i + \mathbb{E}_{s'} V_{h+1}^{i,\boldsymbol{\pi}'}(s')\right].$$

Add and subtract $\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h^{-i}}\left[\ell_h^i + \mathbb{E}_{s'} V_{h+1}^{i,\boldsymbol{\pi}'}(s')\right]$. Then,

$$\left|Q_h^{i,\boldsymbol{\pi}}(s,a_i) - Q_h^{i,\boldsymbol{\pi}'}(s,a_i)\right| \le \underbrace{\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h^{-i}}\left|\mathbb{E}_{s'}\left[V_{h+1}^{i,\boldsymbol{\pi}}(s') - V_{h+1}^{i,\boldsymbol{\pi}'}(s')\right]\right|}_{\text{future-value difference}} + \underbrace{\left|\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h^{-i}} f(\boldsymbol{a}_{-i}) - \mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h'^{-i}} f(\boldsymbol{a}_{-i})\right|}_{\text{distribution shift}}, \tag{9}$$

where $f(\boldsymbol{a}_{-i}) := \ell_h^i(s, a_i, \boldsymbol{a}_{-i}) + \mathbb{E}_{s'} V_{h+1}^{i,\boldsymbol{\pi}'}(s')$. For the future-value term, note that

$$\begin{aligned}
\left|V_{h+1}^{i,\boldsymbol{\pi}}(s') - V_{h+1}^{i,\boldsymbol{\pi}'}(s')\right| &= \left|\left\langle \pi_{h+1}^i(\cdot \mid s'), Q_{h+1}^{i,\boldsymbol{\pi}}(s',\cdot)\right\rangle - \left\langle \pi_{h+1}^{i\prime}(\cdot \mid s'), Q_{h+1}^{i,\boldsymbol{\pi}'}(s',\cdot)\right\rangle\right| \\
&\le \left|\left\langle \pi_{h+1}^i(\cdot \mid s'), Q_{h+1}^{i,\boldsymbol{\pi}}(s',\cdot) - Q_{h+1}^{i,\boldsymbol{\pi}'}(s',\cdot)\right\rangle\right| + \left|\left\langle \pi_{h+1}^i(\cdot \mid s') - \pi_{h+1}^{i\prime}(\cdot \mid s'), Q_{h+1}^{i,\boldsymbol{\pi}'}(s',\cdot)\right\rangle\right| \\
&\le \Delta_{h+1} + \left\|\pi_{h+1}^i(\cdot \mid s') - \pi_{h+1}^{i\prime}(\cdot \mid s')\right\|_1 \left\|Q_{h+1}^{i,\boldsymbol{\pi}'}(s',\cdot)\right\|_\infty.
\end{aligned}$$

Since per-layer losses lie in $[0,1]$, we have $\left\|Q_{h+1}^{i,\boldsymbol{\pi}'}(s',\cdot)\right\|_\infty \le H - h \le H$. Thus,

$$\left|V_{h+1}^{i,\boldsymbol{\pi}}(s') - V_{h+1}^{i,\boldsymbol{\pi}'}(s')\right| \le \Delta_{h+1} + H\left\|\pi_{h+1}^i(\cdot \mid s') - \pi_{h+1}^{i\prime}(\cdot \mid s')\right\|_1 \le \Delta_{h+1} + H\Delta.$$

Therefore, $\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h^{-i}}\left|\mathbb{E}_{s'}[V_{h+1}^{i,\boldsymbol{\pi}}(s') - V_{h+1}^{i,\boldsymbol{\pi}'}(s')]\right| \le \Delta_{h+1} + H\Delta$. For the distribution-shift term in (9), since $\|f\|_\infty \le H$, we have

$$\left|\mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h^{-i}} f(\boldsymbol{a}_{-i}) - \mathbb{E}_{\boldsymbol{a}_{-i}\sim\boldsymbol{\pi}_h'^{-i}} f(\boldsymbol{a}_{-i})\right| \le H\left\|\boldsymbol{\pi}_h^{-i}(\cdot \mid s) - \boldsymbol{\pi}_h'^{-i}(\cdot \mid s)\right\|_1.$$

Using the product-distributions properties we get,

$$\left\|\boldsymbol{\pi}_h^{-i}(\cdot \mid s) - \boldsymbol{\pi}_h'^{-i}(\cdot \mid s)\right\|_1 \le \sum_{j\ne i}\left\|\pi_h^j(\cdot \mid s) - \pi_h^{j\prime}(\cdot \mid s)\right\|_1 \le \Delta.$$

Hence $\Delta_h \le \Delta_{h+1} + 2H\Delta$. Since $\Delta_{H+1} = 0$, we have $\Delta_h \le 2H(H-h+1)\Delta$. In particular, it holds that $\Delta_h \le 2H^2\Delta$. □

Lemma 8. *Let $Q_{t,h}^i(s,\cdot)$ be the state-action values from Algorithm 1. For every player i, state s, layer h, and episode $t \ge 2$, $\left\|Q_{t,h}^i(s,\cdot) - Q_{t-1,h}^i(s,\cdot)\right\|_\infty^2 \le 36m^2\bar{\rho}^2H^6\eta_H^2$, and, with $Q_{0,h}^i(s,\cdot) = 0$, $\left\|Q_{1,h}^i(s,\cdot) - Q_{0,h}^i(s,\cdot)\right\|_\infty^2 \le H^2$.*

Proof. The upper bound from Lemma 7 gives

$$\left\|Q_{t,h}^i(s,\cdot) - Q_{t-1,h}^i(s,\cdot)\right\|_\infty \le 2H^2\sum_{j=1}^m\left\|\pi_t^j - \pi_{t-1}^j\right\|_{\infty,1} = 2H^2\sum_{j=1}^m \max_{h,s\in S_h}\left\|\pi_{t,h}^j(\cdot \mid s) - \pi_{t-1,h}^j(\cdot \mid s)\right\|_1.$$

By Lemma 5 and the monotonicity $\eta_1 \le \cdots \le \eta_H$, it holds that $\left\|\pi_t^j - \pi_{t-1}^j\right\|_{\infty,1} \le 3\bar{\rho}H\eta_H$, which directly implies the desired result. □

B.6 WEIGHTED VALUE-DIFFERENCE DECOMPOSITION

Lemma 9. *Fix an episode t, a policy $\boldsymbol{\pi}_t \in \Pi^{\text{markov}}$, a player i, and a history-dependent policy $\mu^i \in \Pi_i^{\text{gen}}$. Let $d_{t,h}^{i,\mu}$ be the distribution of the history τ_h under $\mu^i \odot \boldsymbol{\pi}_t^{-i}$. Then,*

$$V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu^i}(s_1) = \sum_{h=1}^{H} \sum_{\tau_h \in \mathcal{H}_h} d_{t,h}^{i,\mu}(\tau_h) \left\langle Q_{t,h}^i(s(\tau_h), \cdot), \pi_{t,h}^i(\cdot \mid s(\tau_h)) - \mu_h^i(\cdot \mid \tau_h) \right\rangle.$$

Proof. Let $W_h(\tau_h)$ denote the expected *continuation* loss from history $\tau_h \in \mathcal{H}_h$ when player i follows the comparator $\mu^i \in \Pi_i^{\text{gen}}$ from layer h onward and the opponents follow $\boldsymbol{\pi}_t^{-i}$. Thus $W_{H+1}(\tau_{H+1}) = 0$, and for $\tau_h \in \mathcal{H}_h$ with current state $s = s(\tau_h)$,

$$W_h(\tau_h) = \sum_{\substack{a_i \in A_i,\ a_{-i} \in A_{-i} \\ s' \in S_{h+1}}} \mu_h^i(a_i \mid \tau_h)\, \boldsymbol{\pi}_{t,h}^{-i}(a_{-i} \mid s)\, P_h(s' \mid s, (a_i, a_{-i}))\, \left[\ell_h^i(s, (a_i, a_{-i})) + W_{h+1}(\tau_{h+1})\right]. \tag{10}$$

where $\tau_{h+1} = (\tau_h, a_i, \boldsymbol{a}_{-i}, s')$ denotes the next history after player i chooses a_i, the opponents' policies are according to $\boldsymbol{\pi}_t^{-i}$, and the next state is s'. In contrast, the for the Q-function we have:

$$Q_{t,h}^i(s, a_i) = \sum_{\substack{a_{-i} \in A_{-i} \\ s' \in S_{h+1}}} \boldsymbol{\pi}_{t,h}^{-i}(a_{-i} \mid s)\, P_h(s' \mid s, (a_i, a_{-i}))\, \left[\ell_h^i(s, (a_i, a_{-i})) + V_{t,h+1}^{i,\pi}(s')\right]. \tag{11}$$

Let $D_h(\tau_h) := V_{t,h}^{i,\pi}(s(\tau_h)) - W_h(\tau_h)$. Using $V_{t,h}^{i,\pi}(s) = \langle Q_{t,h}^i(s, \cdot), \pi_{t,h}^i(\cdot \mid s)\rangle$, subtracting (10) and add and subtract $\langle Q_{t,h}^i(s, \cdot), \mu_h^i(\cdot \mid \tau_h)\rangle$. This gives

$$\begin{aligned} D_h(\tau_h) = {} & \left\langle Q_{t,h}^i(s, \cdot), \pi_{t,h}^i(\cdot \mid s) - \mu_h^i(\cdot \mid \tau_h) \right\rangle \\ & + \sum_{\substack{a_i \in A_i,\ a_{-i} \in A_{-i} \\ s' \in S_{h+1}}} \mu_h^i(a_i \mid \tau_h)\, \boldsymbol{\pi}_{t,h}^{-i}(a_{-i} \mid s)\, P_h(s' \mid s, (a_i, a_{-i})) D_{h+1}(\tau_{h+1}). \end{aligned} \tag{12}$$

The second term is the expectation of D_{h+1} under the full-history. Multiplying (12) by $d_{t,h}^{i,\mu}(\tau_h)$ and summing over histories yields

$$\begin{aligned} \mathbb{E}_{\tau_h \sim d_{t,h}^{i,\mu}} D_h(\tau_h) = {} & \sum_{\tau_h \in \mathcal{H}_h} d_{t,h}^{i,\mu}(\tau_h) \left\langle Q_{t,h}^i(s(\tau_h), \cdot), \pi_{t,h}^i(\cdot \mid s(\tau_h)) - \mu_h^i(\cdot \mid \tau_h) \right\rangle \\ & + \mathbb{E}_{\tau_{h+1} \sim d_{t,h+1}^{i,\mu}} D_{h+1}(\tau_{h+1}). \end{aligned}$$

Summing over $h = 1, \ldots, H$ telescopes. At the initial history $d_{t,1}^{i,\mu}(\tau_h)$, $D_1(s_1) = V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu_i}(s_1)$. At the terminal layer, $D_{H+1} = 0$. This completes the proof. □

B.7 PROOF OF THEOREM 1

Fix a player i and an arbitrary policy $\mu^i \in \Pi_i^{\text{gen}}$. For each layer h, define

$$R_{i,h}(\mu^i) := \sum_{\tau_h \in \mathcal{H}_h} \sum_{t=1}^{T} d_{t,h}^{i,\mu}(\tau_h) \left\langle Q_{t,h}^i(s(\tau_h), \cdot), \pi_{t,h}^i(\cdot \mid s(\tau_h)) - \mu_h^i(\cdot \mid \tau_h) \right\rangle.$$

Now, apply Lemma 4 separately for every history $\tau_h \in \mathcal{H}_h$, with $x_t = \pi_{t,h}^i(\cdot \mid s(\tau_h))$, $u = \mu_h^i(\cdot \mid \tau_h)$, $\ell_t = Q_{t,h}^i(s(\tau_h), \cdot)$, $q_t = d_{t,h}^{i,\mu}(\tau_h)$, and learning rate η_h. Summing over $\tau \in \mathcal{H}_h$ yields

$$\begin{aligned} R_{i,h}(\mu^i) \le {} & \frac{B_i}{\eta_h} \sum_{\tau_h \in \mathcal{H}_h} \left(d_{1,h}^{i,\mu}(\tau_h) + \sum_{t=1}^{T-1} |d_{t+1,h}^{i,\mu}(\tau_h) - d_{t,h}^{i,\mu}(\tau_h)| \right) \\ & + \frac{\eta_h \rho_i}{2} \sum_{\tau_h \in \mathcal{H}_h} \sum_{t=1}^{T} d_{t,h}^{i,\mu}(\tau_h) \left\| Q_{t,h}^i(s(\tau_h), \cdot) - Q_{t-1,h}^i(s(\tau_h), \cdot) \right\|_\infty^2. \end{aligned}$$

Since $\sum_{\tau_h \in \mathcal{H}_h} d_{1,h}^{i,\mu}(\tau_h) = 1$ and is the same at every episode, Lemma 8 gives

$$R_{i,h}(\mu_i) \le \frac{B_i(1 + \mathrm{TV}_h^{i,\mu})}{\eta_h} + \frac{1}{2}\eta_h \rho_i H^2 + 18\eta_h \rho_i m^2 \bar{\rho}^2 H^6 \eta_H^2 T.$$

We now sum over h. First,

$$\sum_{h=1}^{H} \frac{B_i}{\eta_h} = \frac{B_i}{\eta_0} \sum_{h=1}^{H} T^{\alpha_h} \le \frac{HB_i}{\eta_0} T^{\beta},$$

where $\beta = \frac{3H-2}{3H+1}$. Second, for $h \ge 2$, Lemma 6 gives

$$\frac{B_i \, \mathrm{TV}_h^{i,\mu}}{\eta_h} \le 3mB_i \bar{\rho} HT \frac{\sum_{\ell=1}^{h-1} \eta_\ell}{\eta_h}.$$

Since $\eta_1 \le \cdots \le \eta_H$, $\sum_{\ell=1}^{h-1} \eta_\ell \le H\eta_{h-1}$. Moreover, $\frac{\eta_{h-1}}{\eta_h} = T^{-(\alpha_{h-1} - \alpha_h)} = T^{-3/(3H+1)}$. Therefore,

$$\frac{B_i \, \mathrm{TV}_h^{i,\mu}}{\eta_h} \le 3mB_i \bar{\rho} H^2 T^{1-3/(3H+1)} = 3mB_i \bar{\rho} H^2 T^{\beta},$$

and hence

$$\sum_{h=1}^{H} \frac{B_i \, \mathrm{TV}_h^{i,\mu}}{\eta_h} \le 3mB_i \bar{\rho} H^3 T^{\beta}.$$

Third,

$$\sum_{h=1}^{H} \frac{1}{2}\eta_h \rho_i H^2 \le \frac{1}{2}\rho_i H^2 \cdot H\eta_H \le \frac{1}{2}\eta_0 \rho_i H^3 T^{\beta},$$

where we used $T^{-\alpha_H} \le T^{\beta}$ for $T \ge 1$. Fourth,

$$\begin{aligned}
\sum_{h=1}^{H} 18\eta_h \rho_i m^2 \bar{\rho}^2 H^6 \eta_H^2 T &= 18\rho_i m^2 \bar{\rho}^2 H^6 T \eta_H^2 \sum_{h=1}^{H} \eta_h \le 18\rho_i m^2 \bar{\rho}^2 H^6 T \eta_H^2 \cdot H\eta_H \\
&\le 18\eta_0^3 \rho_i m^2 \bar{\rho}^2 H^7 T^{1-3\alpha_H} \\
&\le 18\eta_0^3 \rho_i m^2 \bar{\rho}^2 H^7 T^{\beta},
\end{aligned}$$

where the last line follows from $\alpha_H = 1/(3H+1)$, $1 - 3\alpha_H = \frac{3H-2}{3H+1} = \beta$. Choosing $\lambda_i = \frac{1}{|A_i|}$, gives $\rho_i = 1 + |A_i|\lambda_i = 2$, $\bar{\rho} = 2$, and $B_i = 2\log(|A_i| + 1)$. Combining all four pieces results in

$$\begin{aligned}
\sum_{h=1}^{H} R_{i,h}(\mu^i) &\le \left(\frac{HB_i}{\eta_0} + 3mB_i \bar{\rho} H^3 + \frac{1}{2}\eta_0 \rho_i H^3 + 18\eta_0^3 \rho_i m^2 \bar{\rho}^2 H^7 \right) T^{\beta}. \\
&\le \left(\frac{2H\log(|A_i| + 1)}{\eta_0} + 12mH^3 \log(|A_i| + 1) + \eta_0 H^3 + 144\eta_0^3 m^2 H^7 \right) T^{\beta}.
\end{aligned}$$

Define

$$C_i^{\mathrm{SE}} := \frac{2H\log(|A_i| + 1)}{\eta_0} + 12mH^3 \log(|A_i| + 1) + \eta_0 H^3 + 144\eta_0^3 m^2 H^7.$$

By Lemma 9, we have

$$\sum_{t=1}^{T} \left(V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu_i}(s_1) \right) = \sum_{h=1}^{H} R_{i,h}(\mu^i).$$

The left-hand side is precisely the regret of player i with respect to the comparator policy μ_i. Since the bound on $\sum_{t=1}^{T} \left(V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu_i}(s_1) \right)$ holds uniformly for all $\mu_i \in \Pi_i^{\mathrm{gen}}$, taking the supremum over $\mu_i \in \Pi_i^{\mathrm{gen}}$ yields

$$\mathrm{Reg}_T^{\mathrm{gen},i} = \sup_{\mu_i \in \Pi_i^{\mathrm{gen}}} \sum_{t=1}^{T} \left(V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu_i}(s_1) \right) \le C_i^{\mathrm{SE}} T^{\beta}, \qquad \beta = \frac{3H-2}{3H+1}.$$

Now we let $C_*^{\mathrm{SE}} := \max_{i\in[m]} C_i^{\mathrm{SE}}$, $T_\varepsilon := \left\lceil \left(\frac{C_*^{\mathrm{SE}}}{\varepsilon}\right)^{(3H+1)/3} \right\rceil$. In particular, since $\beta - 1 = -3/(3H+1)$, this choice ensures that, for every player $i \in [m]$,

$$\frac{\mathrm{Reg}_{T_\varepsilon}^{\mathrm{gen},i}}{T_\varepsilon} \le C_i^{\mathrm{SE}} T_\varepsilon^{-3/(3H+1)} \le C_*^{\mathrm{SE}} T_\varepsilon^{-3/(3H+1)} \le \varepsilon.$$

Equivalently, $\mathrm{Reg}_{T_\varepsilon}^{\mathrm{gen},i} \le \varepsilon T_\varepsilon$. The ε-CCE guarantee follows by applying the standard no-regret-to-CCE equivalence to the empirical distribution $\widehat{\Pi}_{T_\varepsilon}$ (Cesa-Bianchi & Lugosi, 2006).

C Convergence Analysis of Algorithm 2

C.1 Comparator smoothing

Algorithm 2 plays policies in the restricted simplex Δ_i^ζ. We therefore first compare an arbitrary history-dependent policy with its ζ-smoothed version. For $\mu^i \in \Pi_i^{\mathrm{gen}}$, define

$$\mu_h^{i,\zeta}(\cdot \mid \tau_h) := (1-\zeta)\mu_h^i(\cdot \mid \tau_h) + \zeta\ \mathrm{Unif}(A_i), \qquad h \in [H],\ \tau_h \in \mathcal{H}_h.$$

Then $\mu_h^{i,\zeta}(\cdot \mid \tau_h) \in \Delta_i^\zeta$ for every admissible history, and we have the following guarantee for the *smoothing*.

Lemma 10. *Fix player i, a product Markov opponent profile $\boldsymbol{\pi}^{-i}$, and a history-dependent policy $\mu^i \in \Pi_i^{\mathrm{gen}}$. Consequently, for every product Markov profile $\boldsymbol{\pi}$, it holds that*

$$V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu^i}(s_1) \le V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu^{i,\zeta}}(s_1) + 2\zeta H^2.$$

Proof. For $r = 0, 1, \dots, H$, define a policy $\nu^{i,r} \in \Pi_i^{\mathrm{gen}}$ by

$$\nu_h^{i,r}(\cdot \mid \tau_h) := \begin{cases} \mu_h^{i,\zeta}(\cdot \mid \tau_h), & h \le r, \\ \mu_h^i(\cdot \mid \tau_h), & h > r. \end{cases}$$

Then $\nu^{i,0} = \mu^i$ and $\nu^{i,H} = \mu^{i,\zeta}$. Hence

$$V_{t,1}^{i,\mu^{i,\zeta}}(s_1) - V_{t,1}^{i,\mu^i}(s_1) = \sum_{r=1}^{H} \left[V_{t,1}^{i,\nu^{i,r}}(s_1) - V_{t,1}^{i,\nu^{i,r-1}}(s_1) \right]. \tag{13}$$

The policies $\nu^{i,r}$ and $\nu^{i,r-1}$ agree on all layers $\ell < r$, and their only possible difference is at layer r. For $\tau_r \in \mathcal{H}_r$, let $Q_{t,r}^i(\tau_r, \cdot)$ denote the Q-function whose a_i-th coordinate is the expected payoff obtained by taking a_i at τ_r, then following μ^i, while the opponents follow $\boldsymbol{\pi}^{-i}$. Since the remaining total cost is at most H, $\|Q_{t,r}^i(\tau_r, \cdot)\|_\infty \le H$. Moreover, by the definition of the smoothed policy, $\mu_r^{i,\zeta}(\cdot \mid \tau_r) - \mu_r^i(\cdot \mid \tau_r) = \zeta\big(\mathrm{Unif}(A_i) - \mu_r^i(\cdot \mid \tau_r)\big)$. Hence,

$$\left\|\mu_r^{i,\zeta}(\cdot \mid \tau_r) - \mu_r^i(\cdot \mid \tau_r)\right\|_1 = \zeta \left\|\mathrm{Unif}(A_i) - \mu_r^i(\cdot \mid \tau_r)\right\|_1 \le 2\zeta.$$

Thus, we get

$$\begin{aligned} \left| V_{t,1}^{i,\nu^{i,r}}(s_1) - V_{t,1}^{i,\nu^{i,r-1}}(s_1) \right| &\le \sum_{\tau_r \in \mathcal{H}_r} d_{t,r}^{i,\nu^{i,r-1}}(\tau_r) \|Q_{t,r}(\tau_r, \cdot)\|_\infty \left\|\mu_r^{i,\zeta}(\cdot \mid \tau_r) - \mu_r^i(\cdot \mid \tau_r)\right\|_1 \\ &\le \sum_{\tau_r \in \mathcal{H}_r} d_{t,r}^{i,\nu^{i,r-1}}(\tau_r) \cdot H \cdot 2\zeta = 2\zeta H. \end{aligned}$$

Summing over $r = 1, \dots, H$ by (13), we obtain

$$\left| V_{t,1}^{i,\mu^{i,\zeta}}(s_1) - V_{t,1}^{i,\mu^i}(s_1) \right| \le 2\zeta H^2.$$

The final claim follows by adding and subtracting $V_{t,1}^{i,\mu^{i,\zeta}}(s_1)$ to $V_{t,1}^{i,\pi}(s_1) - V_{t,1}^{i,\mu^i}(s_1)$. □

C.2 UNIFORM ESTIMATION OF Q-VALUES

At block k, conditional on the past, the block policy π_k is fixed and the B trajectories in that block are independent. The reachability condition, together with the action-probability floor for player i, gives a uniform lower bound on the probability of observing each state–player-i-action pair. We state the resulting estimation guarantee for the player i's Q-function estimates as follows.

Lemma 11. *Fix $K, B \in \mathbb{N}$, $\delta \in (0,1)$, and $\xi > 0$. $M_K := KmA_{\max}S$, $u_K := \log \frac{4M_K}{\delta}$, $n_Q := \left\lceil \frac{H^2}{2\xi^2} u_K \right\rceil$, $p_{\min} := \frac{\kappa\zeta}{A_{\max}}$. If $B \geq \left\lceil \frac{8}{p_{\min}}(n_Q + u_K) \right\rceil$, then, with probability at least $1-\delta$,*

$$\max_{\substack{k\in[K], i\in[m], h\in[H], \\ s\in S_h, a\in A_i}} \left| \widehat{Q}^i_{k,h}(s,a) - Q^i_{k,h}(s,a) \right| \leq \xi.$$

Proof. First, fix a tuple (k, i, h, s, a_i), with $s \in S_h$. Then, let $\mathcal{F}_{k-1}$ be the sigma-field generated by all randomness before block k. Let $\mathbb{P}$ denote the probability measure induced by Algorithm 2 and the game itself. Conditional on $\mathcal{F}_{k-1}$ under $\mathbb{P}$, the policy π_k is fixed and the B trajectories in block k are independent.

For $r \in [B]$, write $X_r := X^i_{k,r,h}(s, a_i) = \mathbf{1}\{s^{k,r}_h = s, a^{i,k,r}_h = a_i\}$. By (1) and the definition of the count, $N_{k,i}(s, a_i) = \sum_{r=1}^B X_r$. Conditional on $\mathcal{F}_{k-1}$, the episodes are independent and identically distributed under the fixed profile $\boldsymbol{\pi}_k$. Hence $X_1, \ldots, X_B$ are conditionally i.i.d. Bernoulli random variables with success probability $p_{k,i,h,s,a_i} := \mathbb{P}(s^{k,r}_h = s,\ a^{i,k,r}_h = a_i \,|\, \mathcal{F}_{k-1})$.

By the reachability assumption and the action-floor condition,

$$p_{k,i,h,s,a_i} = \mathbb{P}(s^{k,r}_h = s \mid \mathcal{F}_{k-1})\pi^i_{k,h}(a_i \mid s) \geq \kappa \frac{\zeta}{|A_i|} \geq \frac{\kappa\zeta}{A_{\max}} =: p_{\min}$$

$\mathbb{P}$-almost surely. Hence, under the conditional law $\mathbb{P}(\cdot \mid \mathcal{F}_{k-1})$,

$$N_{k,i}(s, a_i) \sim \mathrm{Bin}(B, p_{k,i,h,s,a_i}), \qquad p_{k,i,h,s,a_i} \geq p_{\min} \quad \mathbb{P}\text{-a.s.}$$

Since the lower-tail probability of a binomial random variable is non-increasing in its success probability, it follows that, for every threshold n,

$$\mathbb{P}(N_{k,i}(s, a_i) < n \mid \mathcal{F}_{k-1}) \leq \Pr(\mathrm{Bin}(B, p_{\min}) < n)$$

$\mathbb{P}$-almost surely. Since $Bp_{\min} \geq 8(n_Q + u_K)$, we have $n_Q \leq Bp_{\min}/2$. Therefore, Chernoff's inequality gives

$$\begin{aligned}\mathbb{P}(N_{k,i}(s, a_i) < n_Q \mid \mathcal{F}_{k-1}) &\leq \Pr\left(\mathrm{Bin}(B, p_{\min}) < \frac{1}{2}Bp_{\min}\right) \\ &\leq \exp\left(-\frac{Bp_{\min}}{8}\right) \leq e^{-u_K},\end{aligned}$$

$\mathbb{P}$-almost surely. Now, let $A := \{N_{k,i}(s, a_i) \geq n_Q\}$. On A, the estimator is

$$\widehat{Q}^i_{k,h}(s, a_i) = \frac{1}{N_{k,i}(s, a_i)} \sum_{r=1}^B X_r Y^{i,k,r}_h.$$

Fix a realization $x = (x_1, \ldots, x_B) \in \{0,1\}^B$ having positive conditional probability given $\mathcal{F}_{k-1}$, and satisfying $n(x) := \sum_{r=1}^B x_r \geq n_Q$. Conditional on $\mathcal{F}_{k-1}$, the pairs $(X_r, Y^{i,k,r}_h)$, $r \in [B]$, are independent, since each pair is a function of the trajectory in episode r. Conditioning further on $X_r = x_r$ for all $r \in [B]$ therefore preserves independence across episodes. For every r with $x_r = 1$, this independence gives

$$\begin{aligned}\mathbb{E}\left[Y^{i,k,r}_h \,\middle|\, \mathcal{F}_{k-1},\ X_j = x_j,\ j \in [B]\right] &= \mathbb{E}\left[Y^{i,k,r}_h \,\middle|\, \mathcal{F}_{k-1},\ X_r = 1\right] \\ &= \mathbb{E}\left[\sum_{u=h}^H \ell^i_u(s^{k,r}_u, \boldsymbol{a}^{k,r}_u) \,\middle|\, \mathcal{F}_{k-1},\ s^{k,r}_h = s,\ a^{i,k,r}_h = a_i\right] \\ &= Q^i_{k,h}(s, a_i).\end{aligned}$$

Moreover, $0 \le Y_h^{i,k,r} \le H - h + 1 \le H$. Therefore, Hoeffding's inequality gives

$$\mathbb{P}\left(\left|\frac{1}{\sum_{r=1}^{B} x_r}\sum_{r=1}^{B} x_r Y_h^{i,k,r} - Q_{k,h}^i(s,a_i)\right| > \xi \,\middle|\, \mathcal{F}_{k-1},\ X_r = x_r,\ r \in [B]\right) \le 2\exp\left(-\frac{2n_Q\xi^2}{H^2}\right).$$

Averaging over all realizations $x_1, \dots, x_B$ with $\sum_{r=1}^{B} x_r \ge n_Q$, we obtain

$$\mathbb{P}\left(\left|\widehat{Q}_{k,h}^i(s,a_i) - Q_{k,h}^i(s,a_i)\right| > \xi,\ A \,\middle|\, \mathcal{F}_{k-1}\right) \le 2\exp\left(-\frac{2n_Q\xi^2}{H^2}\right) \le 2e^{-u_K}.$$

Combining the bound on $A^c = \{N_{k,i}(s,a_i) < n_Q\}$ with the bound on the estimation error on A, we obtain, for the fixed tuple (k,i,h,s,a_i),

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{Q}_{k,h}^i(s,a_i) - Q_{k,h}^i(s,a_i)\right| > \xi \,\middle|\, \mathcal{F}_{k-1}\right) &\le \mathbb{P}(A^c \mid \mathcal{F}_{k-1}) + \mathbb{P}\left(\left|\widehat{Q}_{k,h}^i(s,a_i) - Q_{k,h}^i(s,a_i)\right| > \xi,\ A \,\middle|\, \mathcal{F}_{k-1}\right) \\ &\le e^{-u_K} + 2e^{-u_K} = 3e^{-u_K},\end{aligned}$$

$\mathbb{P}$-almost surely. Taking expectations on both sides gives the unconditional bound

$$\mathbb{P}\left(\left|\widehat{Q}_{k,h}^i(s,a_i) - Q_{k,h}^i(s,a_i)\right| > \xi\right) \le 3e^{-u_K}.$$

Now take a union bound over all tuples (k,i,h,s,a_i) with $k \in [K],\ i \in [m],\ h \in [H],\ s \in S_h, a_i \in A_i$. The number of such tuples is at most $M_K = KmA_{\max}S$. Therefore,

$$\mathbb{P}\left(\max_{\substack{k\in[K],i\in[m],h\in[H],\\ s\in S_h,\ a_i\in A_i}} \left|\widehat{Q}_{k,h}^i(s,a_i) - Q_{k,h}^i(s,a_i)\right| > \xi\right) \le 3M_K e^{-u_K} = 3M_K\frac{\delta}{4M_K} \le \delta.$$

Equivalently, with probability at least $1-\delta$,

$$\max_{\substack{k\in[K],i\in[m],h\in[H],\\ s\in S_h,\ a_i\in A_i}} \left|\widehat{Q}_{k,h}^i(s,a_i) - Q_{k,h}^i(s,a_i)\right| \le \xi.$$

This proves the result.

□

C.3 Regret with uniformly accurate estimates

We next prove a deterministic regret bound for Algorithm 2 on the event that all estimated Q-vectors are uniformly accurate. Define $\mathcal{E}_\xi := \left\{\max_{k,i,h,s,a_i}\left|\widehat{Q}_{k,h}^i(s,a_i) - Q_{k,h}^i(s,a_i)\right| \le \xi\right\}$. On this event, the regret analysis of the estimated-Q functions differs from the full-feedback case only through the estimation-error terms that appear.

Proposition 2. *Suppose Algorithm 2 is run for K blocks. On the event $\mathcal{E}_\xi$, for every player $i \in [m]$,*

$$\sup_{\mu^i \in \Pi_i^{\mathrm{gen}}} \frac{1}{K}\sum_{k=1}^{K}\left[V_{k,1}^{i,\pi}(s_1) - V_{k,1}^{i,\mu^i}(s_1)\right] \le \widetilde{C}_i^{\mathrm{SE}} K^{-3/(3H+1)} + 2H\xi + 2\zeta H^2 + 8\eta_0 H\xi^2 K^{-\alpha_H},$$

where $\widetilde{C}_i^{\mathrm{SE}} := \frac{2H\log(|A_i|+1)}{\eta_0} + 12mH^3\log(|A_i|+1) + \eta_0 H^3 + 288\eta_0^3 m^2 H^7$.

Proof. Fix player i and a comparator $\mu^i \in \Pi_i^{\mathrm{gen}}$. Let $\mu^{i,\zeta}$ be its ζ-smoothed version defined as in Appendix C.1. By Lemma 10, it is enough to control the regret against $\mu^{i,\zeta}$ and then add $2\zeta H^2$ to the average regret.

For block k, let $d_{k,h}^{i,\mu^\zeta} \in \Delta(\mathcal{H}_h)$ be the history distribution at layer h generated when player i uses $\mu^{i,\zeta}$ and the opponents use $\boldsymbol{\pi}_k^{-i}$ Applying Lemma 9 to the block profile $\boldsymbol{\pi}_k$ gives

$$V_{k,1}^{i,\pi^i}(s_1) - V_{k,1}^{i,\mu^{i,\zeta}}(s_1) = \sum_{h=1}^{H}\sum_{\tau_h\in\mathcal{H}_h} d_{k,h}^{i,\mu^\zeta}(\tau_h)\left\langle Q_{k,h}^i(s(\tau_h),\cdot), \pi_{k,h}^i(\cdot \mid s(\tau_h)) - \mu_h^{i,\zeta}(\cdot \mid \tau_h)\right\rangle,$$

On the event $\mathcal{E}_\xi$, for every history τ_h, it holds almost surely that

$$\left|\left\langle Q^i_{k,h}(s(\tau_h),\cdot) - \widehat{Q}^i_{k,h}(s(\tau_h),\cdot), \pi^i_{k,h}(\cdot \mid s(\tau_h)) - \mu^{i,\zeta}_h(\cdot \mid \tau_h)\right\rangle\right| \le \xi \left\|\pi^i_{k,h}(\cdot \mid s(\tau_h)) - \mu^{i,\zeta}_h(\cdot \mid \tau_h)\right\|_1 \le 2\xi.$$

Since $d^{i,\mu^\zeta}_{k,h} \in \Delta(\mathcal{H}_h)$ has total mass one at each layer, replacing $Q^i_{k,h}$ by $\widehat{Q}^i_{k,h}$ costs at most $2HK\xi$ in cumulative regret.

It remains to bound the corresponding weighted regret expression with $\widehat{Q}^i_{k,h}$ in place of $Q^i_{k,h}$. Lemma 4 is stated for the full simplex, but its proof uses only first-order optimality over a convex decision set, the Bregman diameter bound $D_i \le B_i$, and $1/\rho_i$-strong convexity of the smoothed entropy regularizer. Hence the same bound applies verbatim on the restricted simplex Δ^ζ_i. Applying Lemma 4 separately to every history $\tau \in \mathcal{H}_h$, with

$$x_k = \pi^i_{k,h}(\cdot \mid s(\tau_h)), \quad u = \mu^{i,\zeta}_h(\cdot \mid \tau_h), \quad \ell_k = \widehat{Q}^i_{k,h}(s(\tau_h),\cdot), \quad q_k = d^{i,\mu^\zeta}_{k,h}(\tau_h), \quad M_k = \widehat{Q}^i_{k-1,h}(s(\tau_h),\cdot),$$

and summing over histories gives

$$R^{\mathrm{est}}_{i,h}(\mu^{i,\zeta}) \le \frac{B_i(1+\mathrm{TV}^{i,\mu^\zeta}_h)}{\eta_h} + \frac{\eta_h\rho_i}{2}\sum_{\tau\in\mathcal{H}_h}\sum_{k=1}^{K} d^{i,\mu^\zeta}_{k,h}(\tau)\left\|\widehat{Q}^i_{k,h}(s(\tau),\cdot) - \widehat{Q}^i_{k-1,h}(s(\tau),\cdot)\right\|^2_\infty, \tag{14}$$

where $\mathrm{TV}^{i,\mu^\zeta}_h := \sum_{k=1}^{K-1}\left\|d^{i,\mu^\zeta}_{k+1,h} - d^{i,\mu^\zeta}_{k,h}\right\|_1$.

We next bound first order path of the policies generated by estimated Q-functions. The proof of Lemma 5, which is based on Lemma 2, only uses the fact that the feedback vectors lie in $[0,H]^{A_j}$. Since $\widehat{Q}^j_{k,h} \in [0,H]^{A_j}$, repeating the same calculation, for every player j, layer h, state $s \in S_h$, and block k, we obtain

$$\left\|\pi^j_{k+1,h}(\cdot \mid s) - \pi^j_{k,h}(\cdot \mid s)\right\|_1 \le 3\rho_j H\eta_h \le 3\bar{\rho}H\eta_h. \tag{15}$$

Applying the Lemma 6 to the block sequence $\pi_1,\dots,\pi_K$, and using (15), yields

$$\mathrm{TV}^{i,\mu^\zeta}_h \le 3m\bar{\rho}HK\sum_{\ell<h}\eta_\ell.$$

We now control the variation of the estimated Q-function. For the exact values, Lemma 7 and (15) imply, for $k \ge 2$,

$$\begin{aligned}\left\|Q^i_{k,h}(s,\cdot) - Q^i_{k-1,h}(s,\cdot)\right\|_\infty &\le 2H^2\sum_{j=1}^{m}\|\pi^j_k - \pi^j_{k-1}\|_{\infty,1}\\ &\le 2H^2\cdot m\cdot 3\bar{\rho}H\eta_H = 6m\bar{\rho}H^3\eta_H.\end{aligned}$$

On the good event $\mathcal{E}_\xi$, for $k \ge 2$,

$$\left\|\widehat{Q}^i_{k,h}(s,\cdot) - \widehat{Q}^i_{k-1,h}(s,\cdot)\right\|_\infty \le \left\|Q^i_{k,h}(s,\cdot) - Q^i_{k-1,h}(s,\cdot)\right\|_\infty + 2\xi \le 6m\bar{\rho}H^3\eta_H + 2\xi$$

Then, we have

$$\left\|\widehat{Q}^i_{k,h}(s,\cdot) - \widehat{Q}^i_{k-1,h}(s,\cdot)\right\|^2_\infty \le 72m^2\bar{\rho}^2H^6\eta_H^2 + 8\xi^2,$$

where the last line follows from the inequality $(a+b)^2 \le 2a^2+2b^2$. On the first block, $\widehat{Q}^i_{0,h} \equiv 0$ and $\widehat{Q}^i_{1,h} \in [0,H]^{A_i}$, so $\left\|\widehat{Q}^i_{1,h}(s,\cdot) - \widehat{Q}^i_{0,h}(s,\cdot)\right\|^2_\infty \le H^2$. Substituting these estimates into the weighted OOMD bound (14) gives, on every layer h,

$$R^{\mathrm{est}}_{i,h}(\mu^{i,\zeta}) \le \frac{B_i}{\eta_h} + \frac{B_i\,\mathrm{TV}^{i,\mu^\zeta}_h}{\eta_h} + \frac{1}{2}\eta_h\rho_iH^2 + 36\eta_h\rho_im^2\bar{\rho}^2H^6\eta_H^2K + 4\eta_h\rho_i\xi^2K. \tag{16}$$

We now sum over $h \in [H]$, using the same bounds on the learning-rate related terms as in the proof of Theorem 1 in Appendix B. Namely, with $\beta = (3H-2)/(3H+1)$,

$$\sum_{h=1}^{H} \frac{B_i}{\eta_h} \le \frac{HB_i}{\eta_0} K^{\beta}, \quad \sum_{h=1}^{H} \frac{B_i \,\mathrm{TV}_h^{i,\mu^\zeta}}{\eta_h} \le 3mB_i\bar{\rho}H^3K^{\beta}, \quad \sum_{h=1}^{H} \frac{1}{2}\eta_h\rho_i H^2 \le \frac{1}{2}\eta_0\rho_i H^3 K^{\beta}, \tag{17}$$

and

$$\sum_{h=1}^{H} 36\eta_h\rho_i m^2\bar{\rho}^2 H^6 \eta_H^2 K \le 36\eta_0^3\rho_i m^2\bar{\rho}^2 H^7 K^{\beta}. \tag{18}$$

Then the final term in (16) satisfies

$$\sum_{h=1}^{H} 4\eta_h\rho_i\xi^2 K \le 4\eta_0\rho_i H\xi^2 K^{1-\alpha_H}. \tag{19}$$

Choosing $\lambda_i = \frac{1}{|A_i|}$, gives $\rho_i = 1+|A_i|\lambda_i = 2,\ \bar{\rho} = 2$, and $B_i = 2\log(|A_i|+1)$. Then, substituting these bounds into the (17), (18), (19) yields,

$$\sum_{h=1}^{H} R_{i,h}^{\mathrm{est}}(\mu^{i,\zeta}) \le \widetilde{C}_i^{\mathrm{SE}} K^{\beta} + 8\eta_0 H\xi^2 K^{1-\alpha_H}. \tag{20}$$

Finally, combining (20) with the error from comparing the played policies against $\mu^{i,\zeta}$, namely $2\zeta H^2 K$, and the error from Q-function estimation, namely $2HK\xi$, we obtain,

$$\sum_{k=1}^{K} \left[V_{k,1}^{i,\pi}(s_1) - V_{k,1}^{i,\mu^i}(s_1)\right] \le \widetilde{C}_i^{\mathrm{SE}} K^{\beta} + 2HK\xi + 2\zeta H^2 K + 8\eta_0 H\xi^2 K^{1-\alpha_H}.$$

Dividing by K and taking the supremum over $\mu^i \in \Pi_i^{\mathrm{gen}}$ proves the proposition. □

C.4 Proof of Theorem 2

Proof. Apply Lemma 11 with $K = K_\varepsilon,\ B = B_\varepsilon,\ \zeta = \zeta_\varepsilon,\ \xi = \xi_\varepsilon$. By the definition of B_ε, the event $\mathcal{E}_{\xi_\varepsilon}$ holds with probability at least $1-\delta$. On this event, Proposition 2 gives, for every player $i \in [m]$,

$$\sup_{\mu^i \in \Pi_i^{\mathrm{gen}}} \frac{1}{K_\varepsilon} \sum_{k=1}^{K_\varepsilon} \left[V_{k,1}^{i,\pi}(s_1) - V_{k,1}^{i,\mu^i}(s_1)\right] \le \widetilde{C}_i^{\mathrm{SE}} K_\varepsilon^{-3/(3H+1)} + 2H\xi_\varepsilon + 2\zeta_\varepsilon H^2 + 8\eta_0 H\xi_\varepsilon^2 K_\varepsilon^{-\alpha_H}.$$

- By the definition of K_ε, $\widetilde{C}_i^{\mathrm{SE}} K_\varepsilon^{-3/(3H+1)} \le \widetilde{C}_*^{\mathrm{SE}} K_\varepsilon^{-3/(3H+1)} \le \frac{\varepsilon}{4}$.
- By the definition of ξ_ε, $2H\xi_\varepsilon \le \frac{\varepsilon}{4}$.
- By the definition of ζ_ε, $2\zeta_\varepsilon H^2 = \frac{\varepsilon}{4}$.
- Finally, since $K_\varepsilon^{-\alpha_H} \le 1$, $\rho_i = 2$, and $\xi_\varepsilon^2 \le \varepsilon/(32\eta_0 H)$, $8\eta_0 H\xi_\varepsilon^2 K_\varepsilon^{-\alpha_H} \le 8\eta_0 H\xi_\varepsilon^2 \le \frac{\varepsilon}{4}$.

Thus, with probability at least $1-\delta$, for every player $i \in [m]$,

$$\sup_{\mu^i \in \Pi_i^{\mathrm{gen}}} \frac{1}{K_\varepsilon} \sum_{k=1}^{K_\varepsilon} \left[V_{k,1}^{i,\pi}(s_1) - V_{k,1}^{i,\mu^i}(s_1)\right] \le \varepsilon. \tag{21}$$

It remains to translate this block-level guarantee into a guarantee over the actual trajectory episodes. By construction, during block k the algorithm plays the same product Markov policy π_k for all B_ε episodes in that block. Therefore

$$\frac{1}{N_\varepsilon} \sum_{n=1}^{N_\varepsilon} \left[V_n^{i,\pi}(s_1) - V_n^{i,\mu^i}(s_1)\right] = \frac{1}{K_\varepsilon B_\varepsilon} \sum_{k=1}^{K_\varepsilon} \sum_{r=1}^{B_\varepsilon} \left[V_{k,1}^{i,\pi}(s_1) - V_{k,1}^{i,\mu^i}(s_1)\right] = \frac{1}{K_\varepsilon} \sum_{k=1}^{K_\varepsilon} \left[V_{k,1}^{i,\pi}(s_1) - V_{k,1}^{i,\mu^i}(s_1)\right].$$

Hence the bound given in (21) holds when the average is taken over the $N_\varepsilon = K_\varepsilon B_\varepsilon$ actual trajectory episodes. Moreover, the empirical distribution policies over $\pi_1, \dots, \pi_{N_\varepsilon}$, is identical to the empirical distribution over the block policies $\pi_1, \dots, \pi_{K_\varepsilon}$, since each policy π_k appears exactly B_ε times. Thus, the empirical distribution over block policies is an ε-approximate CCE.

Finally, the total number of episodes is $N_\varepsilon = K_\varepsilon B_\varepsilon$. By the definitions of K_ε, B_ε, ζ_ε, n_ε, u_ε, and using the fact $\xi_\varepsilon \le \varepsilon/(8H)$,

$$N_\varepsilon = \widetilde{O}\left(\frac{A_{\max} H^6}{\kappa \varepsilon^3}\left(\frac{4\widetilde{C}_*^{\mathrm{SE}}}{\varepsilon}\right)^{(3H+1)/3}\right),$$

where $\widetilde{O}(\cdot)$ hides logarithmic factors in $(m, S_{\mathrm{nt}}, A_{\max}, K_\varepsilon, 1/\delta)$, and

$$\widetilde{C}_*^{\mathrm{SE}} := \max_{i\in[m]}\left\{\frac{2H\log(|A_i|+1)}{\eta_0} + 12mH^3\log(|A_i|+1) + \eta_0 H^3 + 288\eta_0^3 m^2 H^7\right\}.$$

This completes the proof.

□

D Proofs for the Approximation of Discounted Markov Games (Section 3)

This appendix proves the truncation statements used in Section 3. Throughout this appendix, we fix the discount factor $\gamma \in (0,1)$. The argument is written once for the unified discounted model with original horizon $\bar{H} \in \mathbb{N}_+ \cup \{\infty\}$. The case $\bar{H} < \infty$ corresponds to the finite-horizon discounted games, while $\bar{H} = \infty$ corresponds to the discounted infinite-horizon games.

Lemma 12. *For every $\bar{H}$-horizon policy profile $\boldsymbol{\sigma} \in \Pi^{\mathrm{gen},\bar{H}}$, every $L \le \bar{H}$, and every player $i \in [m]$,*

$$0 \le J^\gamma_{i,\bar{H}}(\boldsymbol{\sigma}) - J^\gamma_{i,L}(\boldsymbol{\sigma}^{[L]}) \le \tau_{L,\bar{H}}(\gamma) \le \frac{\gamma^L}{1-\gamma}.$$

Proof. The policy profile $\boldsymbol{\sigma}$ and its truncation $\boldsymbol{\sigma}^{[L]}$ have the same first L decision rules. Therefore they induce the same distribution over trajectories up to the beginning of layer $L+1$. Hence,

$$J^\gamma_{i,\bar{H}}(\boldsymbol{\sigma}) - J^\gamma_{i,L}(\boldsymbol{\sigma}^{[L]}) = \mathbb{E}^{\boldsymbol{\sigma}}\left[\sum_{h=L+1}^{\bar{H}} \gamma^{h-1}\ell^i_h(s_h, \boldsymbol{a}_h) \,\middle|\, s_1\right].$$

The right-hand side is nonnegative. Since $0 \le \ell^i_h \le 1$, we have

$$\sum_{h=L+1}^{\bar{H}} \gamma^{h-1}\ell^i_h(s_h,\boldsymbol{a}_h) \le \sum_{h=L+1}^{\infty}\gamma^{h-1} = \frac{\gamma^L}{1-\gamma},$$

which concludes the proof. □

Lemma 13. *Let $\boldsymbol{\pi}^{[L]} \in \Pi^{\mathrm{markov},L}$, be an L-step product Markov profile, and define $\bar{\boldsymbol{\pi}} := \mathrm{Ext}_L(\boldsymbol{\pi}^{[L]})$. Then,*

$$J^\gamma_{i,\bar{H}}(\bar{\boldsymbol{\pi}}) - J^\gamma_{i,\bar{H}}(\mu^i \odot \bar{\boldsymbol{\pi}}^{-i}) \le J^\gamma_{i,L}(\boldsymbol{\pi}^{[L]}) - J^\gamma_{i,L}(\mu^{i,[L]} \odot \boldsymbol{\pi}^{[L],-i}) + \tau_{L,\bar{H}}(\gamma).$$

Proof. Applying Lemma 12 with $\boldsymbol{\pi} = \bar{\boldsymbol{\pi}}$, we obtain

$$J^\gamma_{i,\bar{H}}(\bar{\boldsymbol{\pi}}) \le J^\gamma_{i,L}(\boldsymbol{\pi}^{[L]}) + \tau_{L,\bar{H}}(\gamma). \tag{22}$$

On the other hand, the first L decision rules of $\mu^i \odot \bar{\boldsymbol{\pi}}^{-i}$ are exactly those of $\mu^{i,[L]} \odot \boldsymbol{\pi}^{[L],-i}$. Since all costs are nonnegative, ignoring the layers after L can only decrease the term $J^\gamma_{i,\bar{H}}(\mu^i \odot \bar{\boldsymbol{\pi}}^{-i})$. Therefore,

$$J^\gamma_{i,\bar{H}}(\mu^i \odot \bar{\boldsymbol{\pi}}^{-i}) \ge J^\gamma_{i,L}(\mu^{i,[L]} \odot \boldsymbol{\pi}^{[L],-i}). \tag{23}$$

Then, subtracting (23) from (22) proves the claim. □

Proof of Theorem 3. Fix a player $i \in [m]$ and a policy $\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}$. For each output profile $\boldsymbol{\pi}_n^{[L]} \in \Pi^{\mathrm{markov},L}$, define $\bar{\boldsymbol{\pi}}_n := \mathrm{Ext}_L(\pi_n^{[L]})$. By Lemma 13, for every n,

$$J_{i,\bar{H}}^{\gamma}(\bar{\boldsymbol{\pi}}_n) - J_{i,\bar{H}}^{\gamma}(\mu^i \odot \bar{\boldsymbol{\pi}}_n^{-i}) \le J_{i,L}^{\gamma}(\boldsymbol{\pi}_n^{[L]}) - J_{i,L}^{\gamma}(\mu^{i,[L]} \odot \boldsymbol{\pi}_n^{[L],-i}) + \tau_{L,\bar{H}}(\gamma).$$

Averaging over $n = 1, \dots, N$ gives

$$\frac{1}{N}\sum_{n=1}^{N}\left[J_{i,\bar{H}}^{\gamma}(\bar{\boldsymbol{\pi}}_n) - J_{i,\bar{H}}^{\gamma}(\mu^i \odot \bar{\boldsymbol{\pi}}_n^{-i})\right] \le \underbrace{\frac{1}{N}\sum_{n=1}^{N}\left[J_{i,L}^{\gamma}(\boldsymbol{\pi}_n^{[L]}) - J_{i,L}^{\gamma}(\mu^{i,[L]} \odot \boldsymbol{\pi}_n^{[L],-i})\right]}_{*} + \tau_{L,\bar{H}}(\gamma). \tag{24}$$

Since $\mu^{i,[L]} \in \Pi_i^{\mathrm{gen},L}$, the assumed finite-horizon regret guarantee implies that the term $(*)$ in (24) is at most $\varepsilon_{\mathrm{alg}}$, simultaneously for all players $i \in [m]$. The resulting bound is uniform over μ^i, and thus taking the supremum over $\Pi_i^{\mathrm{gen},\bar{H}}$ proves the statement. □

D.1 Proof for the full-feedback case (Corollary 1)

We first note that Algorithm 1 can be applied to game $G_{\gamma,\bar{H}}^{[L_\varepsilon]}$, as $G_{\gamma,\bar{H}}^{[L_\varepsilon]}$ is simply an undiscounted finite-horizon Markov game with costs $c_h^i(s,a) = \gamma^{h-1}\ell_h^i(s,a)$, and these costs remain in $[0,1]$. Hence Theorem 1 applies directly with the horizon parameter H replaced by L_ε.

Proof of Corollary 1. We apply Theorem 3 when Algorithm 1 runs on the truncated discounted game $G_{\gamma,\bar{H}}^{[L_\varepsilon]}$. By Theorem 1, applied with horizon L_ε, the finite-horizon average regret on the truncated game $G_{\gamma,\bar{H}}^{[L_\varepsilon]}$ is bounded by $\varepsilon_{\mathrm{alg}} = C_*^{\mathrm{SE}}(L_\varepsilon)T_\varepsilon^{-3/(3L_\varepsilon+1)}$. The definition of T_ε gives

$$C_*^{\mathrm{SE}}(L_\varepsilon)T_\varepsilon^{-3/(3L_\varepsilon+1)} \le \frac{\varepsilon}{2}.$$

It remains to bound the truncation error. If $L_\varepsilon = \bar{H} < \infty$, then the truncated game $G_{\gamma,\bar{H}}^{[L_\varepsilon]}$, is the same as the original game $G_{\gamma,\bar{H}}$ and $\tau_{L_\varepsilon,\bar{H}}(\gamma) = 0$. Otherwise, by the definition of $L_\varepsilon = \min\{\bar{H}, H_\varepsilon^\gamma\}$, we have $L_\varepsilon = H_\varepsilon^\gamma$. In that case, by Lemma 12

$$\tau_{L_\varepsilon,\bar{H}}(\gamma) \le \frac{\gamma^{L_\varepsilon}}{1-\gamma} \le \frac{\varepsilon}{2},$$

where the last inequality follows from the choice $H_\varepsilon^\gamma = \left\lceil \frac{\log\left(\frac{2}{(1-\gamma)\varepsilon}\right)}{\log(1/\gamma)} \right\rceil$. Therefore, for every player $i \in [m]$, Theorem 3 gives

$$\sup_{\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}} \sum_{t=1}^{T_\varepsilon}\left[J_{i,\bar{H}}^{\gamma}(\bar{\boldsymbol{\pi}}_t) - J_{i,\bar{H}}^{\gamma}(\mu^i \odot \bar{\boldsymbol{\pi}}_t^{-i})\right] \le \varepsilon T_\varepsilon.$$

The discounted CCE conclusion in Corollary 1 follows from the standard no-regret-to-CCE equivalence (Cesa-Bianchi & Lugosi, 2006) applied to the empirical distribution of joint policies over $\bar{\boldsymbol{\pi}}_1, \dots, \bar{\boldsymbol{\pi}}_{T_\varepsilon}$.

We now justify the upper bound for the total number of episodes. Since $\log(1/\gamma) \ge 1-\gamma$, we have

$$L_\varepsilon \le L_{\gamma,\varepsilon}^{\bar{H}} := \min\left\{\bar{H},\, 1 + \frac{1}{1-\gamma}\log\left(\frac{2}{(1-\gamma)\varepsilon}\right)\right\},$$

with the convention that when $\bar{H} = \infty$, $L_{\gamma,\varepsilon}^{\bar{H}} = 1 + \frac{1}{1-\gamma}\log\left(\frac{2}{(1-\gamma)\varepsilon}\right)$. Under the choice $\lambda_i = 1/|A_i|$, from Theorem 1 we have,

$$C_*^{\mathrm{SE}}(L) \le \left(\frac{2A_{\log}}{\eta_0} + 12mA_{\log} + \eta_0 + 144\eta_0^3 m^2\right) L^7, \qquad A_{\log} := \max_{i\in[m]} \log(|A_i|+1).$$

Substituting this estimate into the definition of T_ε, using $\lceil x \rceil \le 1 + x$, gives, for every fixed $\gamma < 1$,

$$T_\varepsilon \le \left[\frac{\left(\frac{A_{\log}}{\eta_0} + mA_{\log} + m^2 + 1 \right) \min\{\bar{H}, 1 + \log(1/\varepsilon)\}^7}{\varepsilon} \right]^{O(\min\{\bar{H}, 1+\log(1/\varepsilon)\})}.$$

This concludes the proof. □

D.2 Proof of the partial-feedback case (Corollary 2)

Now, similar to Subsection D.1, Theorem 2 can be applied to the game $G^{[L_\varepsilon]}_{\gamma,\bar{H}}$, as its scaled costs $c^i_h(s,a) = \gamma^{h-1}\ell^i_h(s,a)$ still lie in $[0,1]$. Hence Theorem 2 applies directly with the horizon parameter H replaced by L_ε and the $(\kappa, \zeta_{\bar{\varepsilon},L})$-reachability condition for $G^{[L_\varepsilon]}_{\gamma,\bar{H}}$.

Proof of Corollary 2. Set $L = L_\varepsilon,\ \bar{\varepsilon} = \frac{\varepsilon}{2}$. By assumption, the truncated discounted game $G^{[L]}_{\gamma,\bar{H}}$ is $(\kappa, \zeta_{\bar{\varepsilon},L})$-reachable. Therefore, applying Theorem 2 to $G^{[L]}_{\gamma,\bar{H}}$ with stated parameters, we obtain that, with probability at least $1-\delta$, for every player $i \in [m]$,

$$\sup_{\nu^i \in \Pi_i^{\mathrm{gen},L}} \frac{1}{N_{\varepsilon,\bar{H},\delta}} \sum_{k=1}^{K_{\bar{\varepsilon},L_\varepsilon}} \sum_{t=1}^{B_{\varepsilon,L_\varepsilon,\delta}} \left[J^\gamma_{i,L}(\boldsymbol{\pi}^{[L]}_k) - J^\gamma_{i,L}(\nu^i \odot \boldsymbol{\pi}^{[L],-i}_k) \right] \le \bar{\varepsilon} = \frac{\varepsilon}{2}.$$

By Theorem 3, the average regret of the extended block policies $\bar{\boldsymbol{\pi}}_k := \mathrm{Ext}_L(\boldsymbol{\pi}^{[L]}_k)$ is at most $\frac{\varepsilon}{2} + \tau_{L,\bar{H}}(\gamma)$. It remains to bound the truncation error. If $L_\varepsilon = \bar{H} < \infty$, then the truncation is the original game and $\tau_{L_\varepsilon,\bar{H}}(\gamma) = 0$. Hence

$$\tau_{L,\bar{H}}(\gamma) \le \frac{\gamma^L}{1-\gamma} \le \frac{\varepsilon}{2}$$

by the Lemma 12. Consequently, with probability at least $1-\delta$, for every player $i \in [m]$, Theorem 3 gives,

$$\sup_{\mu^i \in \Pi_i^{\mathrm{gen},\bar{H}}} \frac{1}{N_{\varepsilon,\bar{H},\delta}} \sum_{k=1}^{K_{\bar{\varepsilon},L_\varepsilon}} \sum_{t=1}^{B_{\varepsilon,L_\varepsilon,\delta}} \left[J^\gamma_{i,\bar{H}}(\bar{\pi}_k) - J^\gamma_{i,\bar{H}}(\mu^i \odot \bar{\pi}_k^{-i}) \right] \le \varepsilon.$$

The discounted CCE statement follows from the standard no-regret-to-CCE equivalence (Cesa-Bianchi & Lugosi, 2006) applied to the empirical distribution over the extended block policies.

It remains to justify the upper bound on $N^{\mathrm{band}}_{\varepsilon,\bar{H},\delta}$. We have, $L_\varepsilon = O(\mathcal{L}_{\bar{H},\varepsilon}),\ \mathcal{L}_{\bar{H},\varepsilon} := \min\{\bar{H}, 1 + \log(1/\varepsilon)\}$, with the convention $\mathcal{L}_{\infty,\varepsilon} = 1 + \log(1/\varepsilon)$. Then,

$$K_{\bar{\varepsilon},L_\varepsilon} = \left\lceil \left(\frac{4\widetilde{C}^{\mathrm{SE}}_*(L_\varepsilon)}{\bar{\varepsilon}} \right)^{(3L_\varepsilon+1)/3} \right\rceil \le \left[\frac{\mathcal{C}_{\mathrm{band}} \mathcal{L}^7_{\bar{H},\varepsilon}}{\varepsilon} \right]^{O(\mathcal{L}_{\bar{H},\varepsilon})}, \tag{25}$$

where, $\mathcal{C}_{\mathrm{band}} := \frac{2A_{\log}}{\eta_0} + 12mA_{\log} + \eta_0 + 288m^2\eta_0^3$. Next, by the definitions $\zeta_{\bar{\varepsilon},L} = \frac{\bar{\varepsilon}}{8L^2},\ \xi_{\bar{\varepsilon},L} = \min\left\{ \frac{\bar{\varepsilon}}{8L}, \sqrt{\frac{\bar{\varepsilon}}{32\eta_0 L}} \right\}$, we have, up to universal constants,

$$\frac{1}{\zeta_{\bar{\varepsilon},L_\varepsilon}} \lesssim \frac{\mathcal{L}^2_{\bar{H},\varepsilon}}{\varepsilon}, \qquad \frac{L^2_\varepsilon}{\xi^2_{\bar{\varepsilon},L_\varepsilon}} \lesssim \frac{\mathcal{L}^4_{\bar{H},\varepsilon}}{\varepsilon^2} + \frac{\eta_0 \mathcal{L}^3_{\bar{H},\varepsilon}}{\varepsilon}. \tag{26}$$

Using the definitions $M_{\bar{\varepsilon},L} = K_{\bar{\varepsilon},L} mA_{\max} S_L,\ u_{\bar{\varepsilon},L,\delta} = \log \frac{4M_{\bar{\varepsilon},L}}{\delta}$, we obtain

$$u_{\bar{\varepsilon},L_\varepsilon,\delta} \le O\left(\mathcal{L}_{\bar{H},\varepsilon} \log \frac{\mathcal{C}_{\mathrm{band}} \mathcal{L}^7_{\bar{H},\varepsilon}}{\varepsilon} + \log \frac{mA_{\max} S}{\delta} \right). \tag{27}$$

Finally, substituting (26),(27) into $B_{\bar{\varepsilon},L_\varepsilon,\delta} = \left\lceil \frac{8A_{\max}}{\kappa\zeta_{\bar{\varepsilon},L_\varepsilon}} \left(n_{\bar{\varepsilon},L_\varepsilon,\delta} + u_{\bar{\varepsilon},L_\varepsilon,\delta}\right) \right\rceil$, $n_{\bar{\varepsilon},L_\varepsilon,\delta} = \left\lceil \frac{L_\varepsilon^2}{2\xi^2_{\bar{\varepsilon},L_\varepsilon}} u_{\bar{\varepsilon},L_\varepsilon,\delta} \right\rceil$, up to universal constants we get;

$$B_{\bar{\varepsilon},L_\varepsilon,\delta} \lesssim \left\lceil \frac{A_{\max}}{\kappa} \frac{\mathcal{L}^2_{\bar{H},\varepsilon}}{\varepsilon} \left(\frac{\mathcal{L}^4_{\bar{H},\varepsilon}}{\varepsilon^2} + \frac{\eta_0 \mathcal{L}^3_{\bar{H},\varepsilon}}{\varepsilon} + 1 \right) u_{\bar{\varepsilon},L_\varepsilon,\delta} \right\rceil, \tag{28}$$

and multiplying (28) by (25), yields

$$N_{\varepsilon,\bar{H},\delta} \leq \left\lceil \frac{A_{\max}\mathcal{L}^2_{\bar{H},\varepsilon}}{\kappa\varepsilon} \left(1 + \frac{\mathcal{L}^4_{\bar{H},\varepsilon}}{\varepsilon^2} + \frac{\eta_0\mathcal{L}^3_{\bar{H},\varepsilon}}{\varepsilon}\right) \times \left(\mathcal{L}_{\bar{H},\varepsilon} \log \frac{\mathcal{C}_{\text{band}}\mathcal{L}^7_{\bar{H},\varepsilon}}{\varepsilon} + \log \frac{mA_{\max}S}{\delta}\right) \right] \left[\frac{\mathcal{C}_{\text{band}}\mathcal{L}^7_{\bar{H},\varepsilon}}{\varepsilon} \right]^{O(\mathcal{L}_{\bar{H},\varepsilon})}$$

This concludes the proof. □

E Proofs for the Computational Lower Bounds (Section 4)

This appendix contains the complete proofs of the computational results stated in Section 4. Appendix E.1 first verifies that Algorithms 1 and 2 solve the sparse discounted-CCE search problem introduced in Definition 8. Starting in Appendix E.2, we develop the lower-bound reduction. We first state the underlying complexity assumption, identify the hard family of *source* bimatrix games, and fix our encoding conventions. Appendix E.3 recalls the logarithmic-horizon Markov-game construction, while Appendix E.4 embeds this construction into an infinite-horizon game with a fixed discount factor. Finally, Appendix E.5 combines these ingredients to prove Theorem 4.

E.1 Proof of Algorithms 1 and 2 solving the sparse discounted-CCE problem

Proposition 1 (Formal)**.** Let $G_{\gamma,\bar{H}}$ be an m-player discounted Markov game, and fix $\varepsilon \in (0,1]$.

1. Let L_ε and T_ε be chosen as in Corollary 1. Run Algorithm 1 on the truncated Markov game $G^{[L_\varepsilon]}_{\gamma,\bar{H}}$, and define $\bar{\boldsymbol{\pi}}_t := \mathrm{Ext}_{L_\varepsilon}\left(\boldsymbol{\pi}_t^{[L_\varepsilon]}\right)$, $t \in [T_\varepsilon]$. Then $\left(\bar{\boldsymbol{\pi}}_1, \ldots, \bar{\boldsymbol{\pi}}_{T_\varepsilon}\right)$ is a solution to $(T_\varepsilon,\varepsilon)$-DiscSparseMarkovCCE$^{\text{gen}}_\gamma$.
2. Suppose that the reachability condition of Corollary 2 holds, and choose the parameters L_ε, K_ε, and B_ε as in Corollary 2. Run Algorithm 2 on $G^{[L_\varepsilon]}_{\gamma,\bar{H}}$, and define $\bar{\boldsymbol{\pi}}_k := \mathrm{Ext}_{L_\varepsilon}\left(\boldsymbol{\pi}_k^{[L_\varepsilon]}\right)$, $k \in [K_\varepsilon]$. Let $N_\varepsilon := K_\varepsilon B_\varepsilon$, and define the policy sequence by $\boldsymbol{\sigma}^{(k-1)B_\varepsilon + r} := \bar{\boldsymbol{\pi}}_k$, $k \in [K_\varepsilon]$, $r \in [B_\varepsilon]$. Then, with probability $1-\delta$, $\left(\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^{N_\varepsilon}\right)$ is a solution to $(N_\varepsilon,\varepsilon)$-DiscSparseMarkovCCE$^{\text{gen}}_\gamma$.

Proof of Proposition 1. We verify separately the policy-class requirement and the discounted CCE requirement appearing in Definition 8.

First consider Algorithm 1. At every episode t, player i, layer h, and state s, the algorithm sets $\pi^{i,[L_\varepsilon]}_{t,h}(\cdot \mid s) = x^{i,s}_{t,h} \in \Delta(\mathcal{A}_i)$ independently for each instance. Hence $\pi^{i,[L_\varepsilon]}_t = \left\{\pi^{i,[L_\varepsilon]}_{t,h}\right\}_{h=1}^{L_\varepsilon} \in \Pi^{\text{markov},L_\varepsilon}_i$, and therefore $\boldsymbol{\pi}^{[L_\varepsilon]}_t = \left(\pi^{1,[L_\varepsilon]}_t, \ldots, \pi^{m,[L_\varepsilon]}_t\right) \in \Pi^{\text{markov},L_\varepsilon}$. The continuation used by $\mathrm{Ext}_{L_\varepsilon}$ is itself a fixed product Markov policy. Consequently, $\bar{\boldsymbol{\pi}}_t = \mathrm{Ext}_{L_\varepsilon}\left(\boldsymbol{\pi}^{[L_\varepsilon]}_t\right) \in \Pi^{\text{markov},\bar{H}}$ for every $t \in [T_\varepsilon]$. Corollary 1 gives, for every player $i \in [m]$, $\sup_{\mu^i \in \Pi^{\text{gen},\bar{H}}_i} \sum_{t=1}^{T_\varepsilon} \left[J^\gamma_{i,\bar{H}}(\bar{\boldsymbol{\pi}}_t) - J^\gamma_{i,\bar{H}}\left(\mu^i \odot \bar{\boldsymbol{\pi}}_t^{-i}\right)\right] \leq \varepsilon T_\varepsilon$. Thus the sequence $\left(\bar{\boldsymbol{\pi}}_1, \ldots, \bar{\boldsymbol{\pi}}_{T_\varepsilon}\right)$ satisfies both requirements of Definition 8, and therefore solves $(T_\varepsilon,\varepsilon)$-DiscSparseMarkovCCE$^{\text{gen}}_\gamma$.

We next consider Algorithm 2. At every block k, player i, layer h, and state s, the algorithm sets $\pi^{i,[L_\varepsilon]}_{k,h}(\cdot \mid s) = x^{i,s}_{k,h} \in \Delta^\zeta_i \subseteq \Delta(\mathcal{A}_i)$ independently for each instance. Therefore, $\boldsymbol{\pi}^{[L_\varepsilon]}_k \in \Pi^{\text{markov},L_\varepsilon}$, $\bar{\boldsymbol{\pi}}_k = \mathrm{Ext}_{L_\varepsilon}\left(\boldsymbol{\pi}^{[L_\varepsilon]}_k\right) \in \Pi^{\text{markov},\bar{H}}$. Since the policy $\bar{\boldsymbol{\pi}}_k$ is played throughout block k, every entry of the sequence $\boldsymbol{\sigma}^{(k-1)B_\varepsilon + r} := \bar{\boldsymbol{\pi}}_k$ belongs to $\Pi^{\text{markov},\bar{H}}$. By Corollary 2, with probability at least $1-\delta$, for any player $i \in [m]$, $\sup_{\mu^i \in \Pi^{\text{gen},\bar{H}}_i} \sum_{n=1}^{N_\varepsilon} \left[J^\gamma_{i,\bar{H}}(\boldsymbol{\sigma}^n) - J^\gamma_{i,\bar{H}}\left(\mu^i \odot \boldsymbol{\sigma}^{n,-i}\right)\right] \leq \varepsilon N_\varepsilon$. Hence the episode-level sequence solves $(N_\varepsilon,\varepsilon)$-DiscSparseMarkovCCE$^{\text{gen}}_\gamma$ on this event. □

E.2 Game size and encoding conventions

We first state the complexity-theoretic assumption underlying the lower bound and the bimatrix-game hardness result that serves as the source of our reduction.

Introduced by Papadimitriou (1994), PPAD is a widely studied complexity class for equilibrium computation in game theory. For the sake of completeness, we provide a complete statement of the so-called Exponential Time Hypothesis for PPAD. We refer to Papadimitriou (1994) and Rubinstein (2016) for further details.

Definition 9 (EndOfALine; (Daskalakis et al., 2009))**.** *Given two binary circuits S, P, each with m-input bits and m output bits such that $P(0^m) = 0^m \neq S(0^m)$, find an input $x \in \{0,1\}^m$ such that $P(S(x)) \neq x$ or $S(P(x)) \neq x \neq 0^m$.*

Assumption 1 (ETH for PPAD; Babichenko et al. (2016))**.** *Solving the* PPAD*-complete problem* EndOfALine*, on instances of size $\tilde{n}$, requires time $2^{\Omega(\tilde{n})}$.*

Our hardness results rely on the following consequence of the "ETH for PPAD".

Theorem 5 (Rubinstein (2016))**.** *Assume* ETH *for* PPAD*. Then, there exists a universal constant $\varepsilon_\star > 0$ such that every algorithm that computes an $\varepsilon_\star$-approximate Nash equilibrium of every two-player $n \times n$ bimatrix game requires time $n^{\log^{1-o(1)} n}$.*

Let $\mathfrak{M}_{\mathrm{Rub}}$ denote the family of bimatrix games produced by the hard-instance construction underlying Theorem 5; see, in particular, Rubinstein (2016, Lemma 8.3). We refer to a game $\widehat{M} \in \mathfrak{M}_{\mathrm{Rub}}$ as a *source game* and write n for its number of pure actions per player. Theorem 5 is a worst-case lower bound over this family, whose action sizes are unbounded. Furthermore, we emphasize that, to remain consistent with the previous sections, we will use these games under the *loss* convention instead of the *utility-payoff* convention, since this conversion neither changes the equilibrium points nor the size of the games in view of Lemma 17 and the subsequent results. For brevity, for further details regarding the exact construction of the game, we refer to Rubinstein (2016).

For an encoded object z, we write $\mathrm{bits}(z)$ for the length of its binary encoding. Rational numbers are encoded in reduced form: a rational number $q \in \mathbb{Q}$ is written as $q = p/r$, where $p \in \mathbb{Z}$, $r \in \mathbb{N}$, and $\gcd(|p|, r) = 1$. We define $\mathrm{bits}(q) := 1 + \lceil \log_2(|p| + 1) \rceil + \lceil \log_2(r + 1) \rceil$.

For a finite-horizon Markov game $G_H = \left(H, \{S_h\}_{h=1}^{H+1}, \{A_i\}_{i=1}^m, \{P_h\}_{h=1}^H, \{\ell_h^i\}_{i\in[m],h\in[H]}, s_1\right)$, we define the term $\beta(G_H)$ as

$$\beta(G_H) := \max\left\{ \max_{h\in[H],s\in S_h,\boldsymbol{a}\in A,\ s'\in S_{h+1}} \mathrm{bits}\left(P_h(s' \mid s,\boldsymbol{a})\right), \max_{i\in[m],h\in[H],s\in S_h,\boldsymbol{a}\in A} \mathrm{bits}\left(\ell_h^i(s,\boldsymbol{a})\right) \right\}.$$

Thus, $\beta(G_H)$ is the largest bit length of one transition-probability entry or one loss entry. We define $|G_H| := \max\{H, S, A_{\max}, \beta(G_H)\}$. Consequently, G_H can be described using $|G_H|^{O(1)}$ bits for fixed m. For the special case of $H = 1$, we simply denote the size of a normal-form game by $|G|$ under this convention.

Similarly, for a discounted infinite-horizon Markov game $G_\gamma = \left(S, \{A_i\}_{i=1}^m, P, \{\ell^i\}_{i=1}^m, \gamma, s_1\right)$, we define the term $\beta(G_\gamma)$ as

$$\beta(G_\gamma) := \max\left\{ \mathrm{bits}(\gamma), \max_{s\in S,\boldsymbol{a}\in A,s'\in S} \mathrm{bits}\left(P(s' \mid s,\boldsymbol{a})\right), \max_{i\in[m],s\in S,\boldsymbol{a}\in A} \mathrm{bits}\left(\ell^i(s,\boldsymbol{a})\right) \right\}.$$

We then define $|G_\gamma| := \max\{S, A_{\max}, \beta(G_\gamma)\}$. If γ is fixed before the input is given, then $\mathrm{bits}(\gamma) = O(1)$, and including $\mathrm{bits}(\gamma)$ in $\beta(G_\gamma)$ changes $|G_\gamma|$ only by a fixed constant factor inside the maximum.

For a finite-horizon discounted Markov game $G_{\gamma,\bar{H}}$, we use the same convention as above with the horizon parameter included: $|G_{\gamma,\bar{H}}| := \max\left\{\bar{H}, S, A_{\max}, \beta(G_{\gamma,\bar{H}})\right\}$, and

$$\beta(G_{\gamma,\bar{H}}) := \max\left\{ \mathrm{bits}(\gamma), \max_{\substack{h\in[\bar{H}],s\in S_h,\\ \boldsymbol{a}\in A,\ s'\in S_{h+1}}} \mathrm{bits}\left(P_h(s' \mid s,\boldsymbol{a})\right), \max_{\substack{i\in[m],h\in[\bar{H}],\\ s\in S_h,\boldsymbol{a}\in A}} \mathrm{bits}\left(\ell_h^i(s,\boldsymbol{a})\right) \right\}.$$

Lemma 14 (Size of the source game)**.** *Let $\widehat{M} \in \mathfrak{M}_{\mathrm{Rub}}$ be a* source *two-player game that belongs to the family of bimatrix games described in Lemma 8.3 of Rubinstein (2016), and let n be the number of pure actions per player in $\widehat{M}$. Then, there exist absolute constants $q_{\mathrm{src}}, n_{\mathrm{src}} \geq 1$, such that $n \leq |\widehat{M}| \leq (n+2)^{q_{\mathrm{src}}}$ for all $n \geq n_{\mathrm{src}}$.*

Proof outline. The lower bound follows immediately from the definition of $|\widehat{M}|$. For the upper bound, observe that the payoff functions of the bimatrix game constructed in Lemma 8.3 of Rubinstein (2016) consist of two types of components: a scaled bipartite polymatrix game constructed in Section 7 of Rubinstein (2016), and the payoffs of the two auxiliary *Althöfer* games.

We first consider the polymatrix component. The bipartite polymatrix game constructed in Section 7 of Rubinstein (2016) contains number of players that is at most polynomial in n for each *side* of the bipartite polymatrix game. Moreover, as described in Sections 7.2–7.4 of Rubinstein (2016), each polymatrix payoff entry is computed from a number of quantities that is polynomially bounded by n. Moreover, all numerical quantities are represented at the fixed precision. Therefore, there exist an absolute constant r_1 and a sufficiently large absolute threshold n_{src} such that every payoff entry of the bipartite polymatrix game has binary encoding length at most $(n+2)^{r_1}$ whenever $n \geq n_{\mathrm{src}}$.

Then, the main payoff component is obtained by multiplying the bipartite polymatrix payoff by factors of the form λn_A or λn_B, where λ is an absolute constant and n_A, n_B are the numbers of players on the two sides of the bipartite polymatrix game. It follows that the binary encoding length of every payoff entry arising from the scaled polymatrix component is at most $(n+2)^{r_2}$ for some absolute constant r_2 and all $n \geq n_{\mathrm{src}}$, after sufficiently large n_{src}.

We next consider the two *Althöfer*-game components used in Lemma 8.3 of Rubinstein (2016). Their payoff entries belong to $\{0,1\}$, and therefore have constant binary encoding length. Since each final bimatrix payoff is the sum of one scaled polymatrix payoff and a constant number of Althöfer payoffs, there exists an absolute constant r_0 such that $\beta(\widehat{M}) \leq (n+2)^{r_0}$ for all $n \geq n_{\mathrm{src}}$, where $\beta(\widehat{M})$ denotes the maximum binary encoding length of any payoff entry of M. Consequently, $|\widehat{M}| = \max\{n, \beta(\widehat{M})\} \leq \max\{n, (n+2)^{r_0}\} \leq (n+2)^{q_{\mathrm{src}}}$ for a sufficiently large absolute constant q_{src}.

Finally, the payoff range of the *source* game need not initially be normalized to $[0,1]$. Since the payoffs in the construction are bounded by absolute constants, such a normalization can be carried out by an affine transformation with fixed coefficients. This changes the payoff bit complexity by at most an absolute additive amount and therefore does not affect the asserted polynomial relation for sufficiently large n. □

E.3 The logarithmic-horizon undiscounted Markov game

We use a finite-horizon Markov game construction from Foster et al. (2023). Let $M = (L^1, L^2)$ be a bimatrix $n \times n$ game. For M, we identify a corresponding finite-horizon Markov game $F_H(M)$ as follows, where $H \geq 2$ is a fixed even (numbered) horizon length.

In $F_H(M)$, both players share the same state and action sets. The admissible action set is $[n]$ at every non-terminal layer. For odd (numbered) layers h, the state set is a singleton $S_h = \{s_h^\star\}$, and for even (numbered) layers h, the state set is $S_h = [n] \times [n]$.

For each layer h, the state transition probability kernels $P_h \times [n] \times [n] \to \Delta(S_{h+1})$ are "deterministic" and are defined as follows. For an odd layer h, if the current state is $s_h^\star$ and the players choose the joint action $(a,b) \in [n] \times [n]$, then the next state records this joint action: $P_h\big((a,b) \mid s_h^\star, (a,b)\big) = 1$. For an even layer $h < H$, the transition probability is independent of both the current state and the joint action: for every $(x,y) \in [n] \times [n]$ and every $(a,b) \in [n] \times [n]$, $P_h\big(s_{h+1}^\star \mid (x,y), (a,b)\big) = 1$. All other transition probabilities are zero.

Furthermore, at an odd layer h, if the players choose the action profile $(a,b) \in [n] \times [n]$, then the player i incurs cost $\ell_h^i(s_h^\star, (a,b)) := \frac{1}{H} L^i(a,b)$, At an even layer $h < H$, all costs are zero and the next state is the next odd-layer singleton $s_{h+1}^\star$.

Let $\mathcal{O}_H := \{1, 3, 5, \ldots, H-1\}$ be the set of odd layers. For a given product Markov profile $\boldsymbol{\pi}_t$ and odd layer $h \in \mathcal{O}_H$, define the corresponding local mixed strategies of players as $x_{t,h} := \pi_t^1(\cdot \mid s_h^\star)$,

and $y_{t,h} := \pi_t^2(\cdot \mid s_h^\star)$. Finally, it can be clearly seen that the described construction takes polynomial time.

Lemma 15 (Lemma D.4 Foster et al. (2023)). *Let M be a bimatrix game. There exists an absolute constant $c_F > 0$, independent of any problem parameter, such that the following holds. Let $H \geq 2$ be even, $T \in \mathbb{N}_+$, and let $F_H(M)$ be the finite-horizon Markov game constructed as described in Appendix E.3. Suppose, $L_H = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T)$ is an $\varepsilon_\star/4$-approximate T-sparse CCE of $F_H(M)$, against non-Markov deviations. If $T < \exp(c_F \varepsilon_\star^2 H)$, then there exist $t \in [T]$ and $h \in \mathcal{O}_H$ such that $(x_{t,h}, y_{t,h})$ is an $\varepsilon_\star$-approximate Nash equilibrium of M, where $\mathcal{O}_H$ is the set of odd numbered layers.*

E.4 Embedding into a fixed-discount game

We extend the construction of $F_H(M)$ to the infinite-horizon setting. This enables us to establish a connection between the notion of CCE in discounted infinite-horizon Markov games and the notion of Nash equilibrium in the associated normal-form game M.

For a finite-horizon Markov game $G_H = \Big(H, \{S_h\}_{h=1}^{H+1}, \{A_i\}_{i=1}^m, \{P_h\}_{h=1}^H, \{\ell_h^i\}_{i\in[m],h\in[H]}, s_1\Big)$, we define a corresponding infinite-horizon discounted Markov game $\widetilde{G}_{H,\gamma} = (\widetilde{S}, \{A_i\}_{i=1}^m, \widetilde{P}, \{\widetilde{\ell}^i\}_{i=1}^m, \gamma, s_1)$, where $\gamma \in (0,1) \cap \mathbb{Q}$ as follows.

The state space is $\widetilde{S} := \{(h,s) : h \in [H],\ s \in S_h\} \cup \{\perp\}$, where $\perp$ is an absorbing state. The action sets of the players stay the same. In this configuration, we define the transition kernel of the players in $\widetilde{G}_{H,\gamma}$ as,

$$\widetilde{P}(x' \mid x, a) = \begin{cases} P_h(s' \mid s, a), & \text{if } x = (h,s),\ h < H,\ x' = (h+1, s'), \\ 1, & \text{if } x = (H,s),\ x' = \perp, \\ 1, & \text{if } x = \perp,\ x' = \perp, \\ 0, & \text{otherwise.} \end{cases}$$

Choose a scaling parameter $\alpha := \gamma^{H-1}$, and define the cost function of player i as

$$\widetilde{\ell}^i((h,s),a) := \alpha\gamma^{-(h-1)}\ell_h^i(s,a) = \gamma^{H-h}\ell_h^i(s,a), \qquad h \in [H],$$

while $\widetilde{\ell}^i(\perp, \cdot) :\equiv 0$. Moreover, as $\widetilde{G}_{H,\gamma}$ can be constructed from G_H in polynomial time, since it only adds one absorbing state and rescales the finitely many transition and cost entries of G_H. Then, we have the following result.

Lemma 16. *For every policy profile $\boldsymbol{\sigma} \in \Pi^{\text{markov}}$ in $\widetilde{G}_{H,\gamma}$ and every player i, $J_i^\gamma(\boldsymbol{\sigma}; \widetilde{G}_\gamma) = \alpha V^{i,\boldsymbol{\sigma}^{[H]}}(s_1; G_H)$. Moreover, for every sparse list $\widetilde{L} = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_T)$ where, for all $t \in [T]$ $\sigma_t \in \Pi^{\text{markov},\infty}$, it holds that, if $\widetilde{L}$ is an ε-approximate discounted CCE of $\widetilde{G}_{H,\gamma}$, then $\widetilde{L}^{[H]}$ is an (ε/α)-approximate CCE of G_H.*

Proof. For the first H layers, the infinite-horizon game $\widetilde{G}_{H,\gamma}$ exactly emulates G_H. By construction, after the layer H, the state of the players in the game $\widetilde{G}_{H,\gamma}$ permanently moves to $\perp$, where all future costs of the players are zero. Along every simulated state-action pair, i.e. $((h, s_h), a_h)$, for $h \leq H$, we obtain

$$\gamma^{h-1}\widetilde{\ell}^i((h,s_h),a_h) = \gamma^{h-1}\alpha\gamma^{-(h-1)}\ell_h^i(s_h,a_h) = \alpha\ell_h^i(s_h,a_h).$$

Taking the expectation of the expression above, we obtain

$$J_i^\gamma(\boldsymbol{\sigma}; \widetilde{G}_\gamma) = \alpha\mathbb{E}^{\boldsymbol{\sigma}^{[H]}}\left[\sum_{h=1}^H \ell_h^i(s_h,a_h)\right] = \alpha V^{i,\boldsymbol{\sigma}^{[H]}}(s_1; G_H).$$

Now fix a list $\widetilde{L}$. Then, for any infinite-horizon deviation $\mu^i \in \Pi_i^{\text{gen},\infty}$ we get the following equality

$$\frac{1}{T}\sum_{t=1}^T \Big[J_i^\gamma(\boldsymbol{\sigma}_t; \widetilde{G}_\gamma) - J_i^\gamma(\mu^i \odot \boldsymbol{\sigma}_t^{-i}; \widetilde{G}_\gamma)\Big] = \alpha\frac{1}{T}\sum_{t=1}^T \Big[V^{i,\boldsymbol{\sigma}_t^{[H]}}(s_1; G_H) - V^{i,\mu^{i,[H]}\odot\boldsymbol{\sigma}_t^{[H],-i}}(s_1; G_H)\Big]. \tag{29}$$

It can be seen that, every H-step deviation can be extended arbitrarily to infinite horizon after the absorbing layer H, and the discounted value of the players remains the same because all post-H costs are zero. Thus, taking the supremum over deviations on the LHS of (29) is equivalent to taking the supremum over H-step deviations on the right. This completes the proof. □

Now, we describe the expression sizes of the source game $\widehat{M}$ and its corresponding encoding $\widetilde{F}_{H,\gamma}(\widehat{M})$ in the next two lemmas, Lemmas 14 and 17.

Lemma 17 (Size of the discounted embedding)**.** *Fix a rational discount factor $\gamma \in (0,1) \cap \mathbb{Q}$. Let $q_{\mathrm{src}} \geq 1$ be the absolute constant from Lemma 14. Define $d_0 := \max\{4, q_{\mathrm{src}} + 2\}$. Then, for every finite constant $B_H \geq 2$, there exists an integer $n_{\mathrm{size}} = n_{\mathrm{size}}(\gamma, B_H)$ such that the following holds for every $n \geq n_{\mathrm{size}}$. Let $\widehat{M}$ be the* source *two-player game described in Lemma 14, with n pure actions per player, and let $H \geq 2$ be an even integer satisfying $H \leq B_H \log(n+2)$. Let $F_H(\widehat{M})$ be the finite-horizon Markov game constructed according to Subsection E.3, and let $\widetilde{F}_{H,\gamma}(\widehat{M})$ be the discounted game obtained from $F_H(\widehat{M})$ as described in Subsection E.4. Then $n \leq |\widetilde{F}_{H,\gamma}(\widehat{M})| \leq (n+2)^{d_0}$.*

Proof. To construct $\widetilde{F}_{H,\gamma}(\widehat{M})$ from $\widehat{M}$, we first denote it in *loss* convention through well known conversions from payoffs to losses. By Lemma 14, the source game satisfies $|\widehat{M}| \leq (n+2)^{q_{\mathrm{src}}}$. Then, every loss entry of $\widehat{M}$ has bit length at most

$$\max_{i \in \{1,2\}} \max_{a,b \in [n]} \mathrm{bits}(L^i(a,b)) \leq (n+2)^{q_{\mathrm{src}}}.$$

We now bound the size of $\widetilde{F}_{H,\gamma}(\widehat{M})$. First, consider the state and action sets. In $F_H(\widehat{M})$, every odd layer is a singleton and every even layer has state set $[n] \times [n]$. Hence, $\sum_{h=1}^{H} |S_h| = \frac{H}{2} + \frac{H}{2} n^2$. Recall that the state space of the players in the game $\widetilde{F}_{H,\gamma}(\widehat{M})$ is $\widetilde{S} := \{(h,s) : h \in [H],\ s \in S_h\} \cup \{\perp\}$. Therefore

$$|\widetilde{S}| = 1 + \frac{H}{2} + \frac{H}{2} n^2 \leq 1 + H(n+2)^2.$$

Since $H \leq B_H \log(n+2)$, we have $H \leq n+2$ whenever $n \geq n_{\mathrm{size}}$. Consequently,

$$|\widetilde{S}| \leq 1 + (n+2)^3 \leq (n+2)^4. \tag{30}$$

Moreover, both players in $\widetilde{F}_{H,\gamma}(\widehat{M})$ have the action set $[n]$, and thus $A_{\max} = n$. This already gives the lower bound

$$|\widetilde{F}_{\gamma}(\widehat{M})| = \max\{|\widetilde{S}|, A_{\max}, \beta(\widetilde{G}_{\gamma}(\widehat{M})\} \geq A_{\max} = n.$$

It remains to bound $\beta(\widetilde{F}_{H,\gamma}(\widehat{M}))$. The transition probabilities of the players in $\widetilde{F}_{H,\gamma}(\widehat{M})$ are all either 0 or 1, and hence the transition entries have constant bit length. We only need to bound the loss entries. For $h \in [H]$, the discounted embedding uses losses of the form $\widetilde{\ell}^i((h,s),a) = \gamma^{H-h} \ell_h^i(s,a)$, and $\widetilde{\ell}^i(\perp, a) = 0$. Thus every nonzero loss entry in $\widetilde{F}_{H,\gamma}(\widehat{M})$ has the form $\gamma^{H-h} \frac{1}{H} L^i(a,b)$. Since $\gamma \in (0,1) \cap \mathbb{Q}$ is fixed, there is a constant $c_\gamma \geq 1$, depending only on γ, such that $\mathrm{bits}(\gamma^k) \leq c_\gamma(k+1)$ for all $k \geq 0$. Also, $\mathrm{bits}\left(\frac{1}{H}\right) \leq O(\log(H+1))$. Furthermore, there is a universal constant $c_{\mathrm{rat}} \geq 1$ such that, for every finite collection of rationals $r_1, \ldots, r_k$, $\mathrm{bits}\left(\prod_{\ell=1}^{k} r_\ell\right) \leq c_{\mathrm{rat}} \left(1 + \sum_{\ell=1}^{k} \mathrm{bits}(r_\ell)\right)$. Then, there exists a universal constant $c_{\mathrm{rat}} \geq 1$ such that every loss entry of $\widetilde{F}_{H,\gamma}(\widehat{M})$ satisfies

$$\begin{aligned}
\mathrm{bits}\left(\gamma^{H-h} \frac{2}{H} L^i(a,b)\right) &\leq c_{\mathrm{rat}} \left(c_\gamma(H+1) + O(\log(H+1)) + \mathrm{bits}(L^i(a,b)) + 1\right) \\
&\leq c_{\mathrm{rat}} \left(c_\gamma(H+1) + O(\log(H+1)) + (n+2)^{q_{\mathrm{src}}} + 1\right).
\end{aligned}$$

Since $H \leq B_H \log(n+2)$, for sufficiently large n_{size} we have

$$\beta(\widetilde{F}_{H,\gamma}(\widehat{M})) \leq (n+2)^{q_{\mathrm{src}}+2} \tag{31}$$

for all $n \geq n_{\mathrm{size}}$. Combining (30) and (31) we obtain

$$|\widetilde{F}_{\gamma}(\widehat{M})| \leq \max\{(n+2)^4, n, (n+2)^{q_{\mathrm{src}}+2}\} \leq (n+2)^{d_0},$$

where $d_0 = \max\{4, q_{\mathrm{src}} + 2\}$. Together with $|\widetilde{F}_{H,\gamma}(\widehat{M})| \geq n$, we have $n \leq |\widetilde{F}_{H,\gamma}(\widehat{M})| \leq (n+2)^{d_0}$. □

E.5 Proof of Theorem 4 and Corollary 3

Theorem 4 (Formal)**.** Fix a constant discount factor $\gamma \in (0,1) \cap \mathbb{Q}$ and a support exponent $C > 0$. We define $A_H := \frac{2(Cd_0+1)}{c_F \varepsilon_\star^2}$, $B_H := A_H + 2$, $a_\gamma = B_H \log(\frac{1}{\gamma})$. By $\widehat{M}$ denote a bimatrix game belonging to the family of games in Lemma 14. Under Assumption 1, there exists an accuracy exponent $K = K(\gamma, C) \geq a_\gamma + 1$ such that, for any constructed discounted Markov game $\widetilde{F}_{H,\gamma}(\widehat{M})$, satisfying $n \geq n_{\mathrm{src}}, n_{\mathrm{size}}(\gamma, B_H)$, $n^{-K}(n+2)^{a_\gamma} \leq \frac{\varepsilon_\star}{4}$, and such that there exists an even integer H with $B_H \log(n+2) \geq H \geq A_H \log(n+2)$, there exists a discounted two-player general-sum Markov game such that no polynomial-time algorithm solves $\big(|\widetilde{F}_{H,\gamma}(\widehat{M})|^C, |\widetilde{F}_{H,\gamma}(\widehat{M})|^{-K}\big)$-DiscSparseMarkovCCE$_\gamma^{\mathrm{gen}}$.

Proof of Theorem 4. Suppose not. Take $K := a_\gamma + 1$ that violates the running hypothesis. Let $\widehat{M}$ denote the bimatrix game whose infinite-horizon discounted extension under γ admits a polynomial-time algorithm $\mathcal{A}$ that solves $\big(|\widetilde{F}_{H,\gamma}(\widehat{M})|^C, |\widetilde{F}_{H,\gamma}(\widehat{M})|^{-K}\big)$-DiscSparseMarkovCCE$_\gamma^{\mathrm{gen}}$ with the aforementioned properties. Running $\mathcal{A}$ on $\widetilde{F}_{H,\gamma}(\widehat{M})$, we obtain a list $\widetilde{L} = (\boldsymbol{\sigma}_1, \cdots, \boldsymbol{\sigma}_T)$ such that

$$T \leq \lceil |\widetilde{F}_{H,\gamma}(\widehat{M})|^C \rceil, \qquad \max_{i \in \{1,2\}} \mathrm{Gap}_i^\gamma(\widetilde{L}; \widetilde{F}_{H,\gamma}(\widehat{M})) \leq |\widetilde{F}_{H,\gamma}(\widehat{M})|^{-K},$$

whose restrictions to the first H-layers is denoted by $L_H := \widetilde{L}^{[H]} = (\boldsymbol{\sigma}_1^{[H]}, \ldots, \boldsymbol{\sigma}_T^{[H]})$ that satisfies the properties in Lemma 16. In particular, it holds that

$$\max_{i \in \{1,2\}} \mathrm{Gap}_i^H(L_H; F_H(\widehat{M})) = \gamma^{-H+1} \max_{i \in \{1,2\}} \mathrm{Gap}_i^\gamma(\widetilde{L}; \widetilde{F}_{H,\gamma}(\widehat{M})) \leq |\widetilde{F}_{H,\gamma}(\widehat{M})|^{-K} \gamma^{-H+1}.$$

Since $|\widetilde{F}_{H,\gamma}(\widehat{M})| \geq n$ and $H \leq B_H \log(n+2)$ by Lemma 17,

$$\begin{aligned} |\widetilde{F}_{H,\gamma}(\widehat{M})|^{-K} \gamma^{-(H-1)} &\leq n^{-K} \gamma^{-H} \\ &= n^{-K} \exp(H \log(1/\gamma)) \\ &\leq n^{-K} \exp(B_H \log(n+2) \log(1/\gamma)) \\ &= n^{-K} (n+2)^{a_\gamma} \leq \frac{\varepsilon_\star}{4}. \end{aligned}$$

Therefore L_H is an $\varepsilon_\star/4$-approximate T-sparse CCE of $F_H(\widehat{M})$, against non-Markov deviations. Since $T \leq \lceil |\widetilde{F}_{H,\gamma}(\widehat{M})|^C \rceil$, $|\widetilde{F}_{H,\gamma}(\widehat{M})| \leq (n+2)^{d_0}$, and that $H \geq A_h \log(n+2)$, a straightforward inequality implies that $T < \exp(c_F \varepsilon_\star^2 H)$, meaning that Lemma 15 is applicable. In particular, through $\mathcal{A}$, one obtains an $\varepsilon_\star$-approximate Nash equilibrium of $\widehat{M}$ in polynomial time.

Note that $TH \leq (n+2)^{Cd_0+1} O(\log n)$, which can be written in polynomial time in n. Furthermore, for any given mixed strategy in $\widehat{M}$, its Nash gap can be checked in polynomial time. Since $\mathcal{A}$ runs in polynomial time and $|\widetilde{F}_{H,\gamma}(\widehat{M})| \leq (n+2)^{d_0}$, all output probabilities have polynomial bit length in n. Thus enumerating the candidate pairs and outputting one with gap at most $\varepsilon_\star$ is feasible in polynomial time. Construction of $F_H(\widehat{M})$ and $\widetilde{F}_{H,\gamma}(\widehat{M})$ takes polynomial time by Lemmas 15 and 17. In particular, we have obtained a contradiction to Theorem 5. □

Corollary 3 (Formal)**.** Fix a constant discount factor $\gamma \in (0,1) \cap \mathbb{Q}$ and a support exponent $C_{\mathrm{fh}} > 0$. Under Assumption 1, there exists a finite-horizon discounted Markov game and an accuracy exponent $K_{\mathrm{fh}} = K_{\mathrm{fh}}(\gamma, C_{\mathrm{fh}}) > 0$ such that no polynomial-time algorithm can solve the following problem: given a game $G_{\gamma,\bar{H}}$ of input length at most N, output a list of at most $N^{C_{\mathrm{fh}}}$ product Markov profiles whose $\bar{H}$-horizon discounted CCE gap, against all history-dependent deviations, is at most $N^{-K_{\mathrm{fh}}}$.

Proof of Corollary 3. Due to equilibrium correspondence between $F_H(\widehat{M})$ and $\widetilde{F}_{H,\gamma}(\widehat{M})$, given in Lemma 16, Theorem 4 implies that no polynomial time algorithm is feasible for a class of Markov games $F_H(\widehat{M})$ under some restrictions on H and n. In such a game $F_H(\widehat{M})$, suppose that player i has an objective function of the form $J_{i,H} = \mathbb{E}\left[\sum_{h=1}^H \gamma^{H-1} \ell_h^i(x_h, a_h)\right]$. Note that $J_{i,H} = \mathbb{E}\left[\sum_{h=1}^H \gamma^{h-1} \gamma^{H-1} \frac{\ell_h^i(x_h,a_h)}{\gamma^{h-1}}\right]$, and thus rewriting the cost of player i instead as

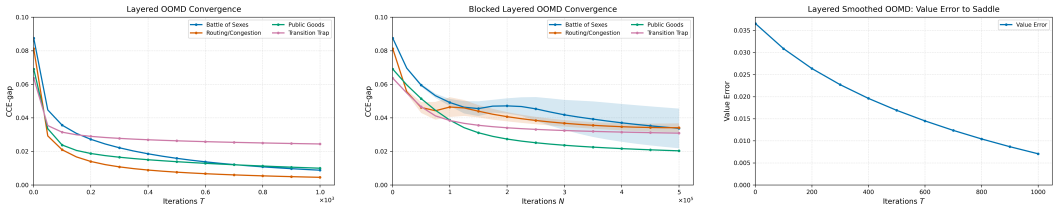

Figure 1: Convergence of the proposed learning dynamics. Panels (a) and (b) show the behavior of the learning dynamics for Algorithm 1 and Algorithm 2, respectively. Panel (c) reports the value error for the discounted linear-quadratic benchmark relative to the closed-form saddle-point value.

$\{\ell_h^i \gamma^{H-1}/\gamma^{h-1}\}_{h=1}^H$, we obtain that there is no polynomial time algorithm for a discounted Markov game. □

F Numerical Experiments

In this section, we evaluate the independent learning dynamics under both information structures considered in the paper. First, we run Algorithm 1 with full-feedback, as in Theorem 1. Second, we run Algorithm 2 using only independently sampled partial feedbacks, as in Theorem 2. Finally, we use a two-state discounted linear-quadratic zero-sum Markov game with a closed-form stationary solution as a benchmark for the discounted truncation approximation result stated in Corollary 1.

For the finite-horizon episodic Markov games, we consider a battle-of-the-sexes game, a congestion game, a three-player public-goods game, and a transition-trap game, which are given in Appendix F. Every state appearing in these games has occupancy probability at least 0.05 under every product Markov profile. Hence, all four games are $(0.05, \zeta)$-reachable for every $\zeta \in (0, 1]$. Given the episode-indexed sequence of product Markov profiles $(\rho_n)_{n=1}^N$, we report $\mathrm{Gap}_N := \max_{i\in[m]} \sup_{\mu^i \in \Pi_i^{\mathrm{gen}}} \frac{1}{N} \sum_{n=1}^N \left[V^{i,\rho_n}(s_1) - V^{i,\mu^i \odot \rho_n^{-i}}(s_1) \right]$. The full-feedback experiment uses $T = 1000$, $\eta_0 = 0.5$, and evaluates the gap every 50 episodes. The partial-feedback experiment uses $K = 1000,\ B = 500,\ \zeta = 0.05,\ \eta_0 = 0.1$, and evaluates the gap every 50 blocks. We perform 20 independent runs of the partial-feedback experiment. Figure 1 reports their mean gap, with the shaded region representing one empirical standard deviation. In all four games, the CCE gap decreases under both feedback models supporting the Theorems 1 and 2.

The final experiment studies a two-player, two-state, zero-sum discounted linear-quadratic Markov game with continuous scalar actions. The benchmark is chosen so that its infinite-horizon discounted saddle-point solution can be computed in closed form. In the simulation, we discretize the action interval while ensuring that the grid contains the closed-form saddle-point actions. Let $J_0^\star(s_0)$ denote the normalized discounted saddle-point value obtained from the closed-form solution. We report $\mathrm{Err}_{T'} := \left| \frac{1}{T'} \sum_{t=1}^{T'} J_0(\pi_t) - J_0^\star(s_0) \right|$, where $J_0(\pi_t)$ is the value of the finite approximation used in the simulation. Figure 1 shows that this error decreases rapidly. We refer to Appendix F.3 for details.

F.1 Implementation and evaluation protocol

Both finite-horizon experiments use the four games specified below without any changes. The two experiments differ only in the feedback available to the players.

F.2 Finite-Horizon Episodic Games

Below, we list the transition and cost tables used for the finite-horizon experiments.

F.2.1 Markov battle-of-the-sexes chain

This is a two-player, two-action, horizon-2 game with state layers $S_1 = \{0\},\ S_2 = \{1, 2\},\ S_3 = \{3\}$. At the initial state 0, the transition and costs are

Joint action a	Transition to $(1,2)$	$c_1(0,a)$	$c_2(0,a)$
$(0,0)$	$(0.9, 0.1)$	0.05	0.20
$(1,1)$	$(0.1, 0.9)$	0.20	0.05
$(0,1),(1,0)$	$(0.5, 0.5)$	0.55	0.55

At the second layer, the game terminates deterministically. Explicitly, for $s = 1$, costs are

$$c_1(1,a) = \begin{cases} 0.15, & a = (0,0), \\ 0.35, & a = (1,1), \\ 0.65, & a \in \{(0,1),(1,0)\}, \end{cases} \qquad c_2(1,a) = \begin{cases} 0.35, & a = (0,0), \\ 0.15, & a = (1,1), \\ 0.65, & a \in \{(0,1),(1,0)\}. \end{cases}$$

For $s = 2$,

$$c_1(2,a) = \begin{cases} 0.15, & a = (1,1), \\ 0.35, & a = (0,0), \\ 0.65, & a \in \{(0,1),(1,0)\}, \end{cases} \qquad c_2(2,a) = \begin{cases} 0.35, & a = (1,1), \\ 0.15, & a = (0,0), \\ 0.65, & a \in \{(0,1),(1,0)\}. \end{cases}$$

F.2.2 ROUTING/CONGESTION GAME

This is a two-player, two-action, horizon-3 game with layers $S_1 = \{0\},\ S_2 = \{1,2\},\ S_3 = \{3,4\},\ S_4 = \{5\}$. Actions are interpreted as route choices: 0 is an upper/fast route and 1 is a lower/safe route. Congestion occurs when both players choose the same action.

At $h = 1$, let $n_0(a) = \mathbf{1}\{a_1 = 0\} + \mathbf{1}\{a_2 = 0\}$ and $\kappa(a) = \mathbf{1}\{a_1 = a_2\}$. The transition is

$$P(1 \mid 0, a) = \begin{cases} 0.75, & n_0(a) \le 1, \\ 0.25, & n_0(a) = 2, \end{cases} \qquad P(2 \mid 0, a) = 1 - P(1 \mid 0, a), \tag{32}$$

with player-i cost $c_i(0,a) = 0.10 + 0.45\kappa(a) + 0.05a_i$. At $h = 2$, for $s \in \{1,2\}$,

$$P(3 \mid s, a) = \begin{cases} 0.85, & s = 1,\ \kappa(a) = 0, \\ 0.35, & s = 1,\ \kappa(a) = 1, \\ 0.35, & s = 2,\ \kappa(a) = 0, \\ 0.15, & s = 2,\ \kappa(a) = 1, \end{cases} \qquad P(4 \mid s, a) = 1 - P(3 \mid s, a), \tag{33}$$

with base cost 0.05 at state 1 and 0.20 at state 2: $c_i(s,a) = \mathrm{base}(s) + 0.50\kappa(a) + 0.05\mathbf{1}\{a_i = 0\}$. At $h = 3$, the game terminates. Costs are $c_i(s,a) = 0.10 + 0.55\kappa(a) + p_i(s,a)$, where $p_i(s,a) = 0.10$ if $(s = 3, a_i = 1)$ or $(s = 4, a_i = 0)$, and $p_i(s,a) = 0$ otherwise.

F.2.3 THREE-PLAYER PUBLIC-GOODS GAME

This game has $m = 3$, binary actions, horizon 2, and layers $S_1 = \{0\},\ S_2 = \{1,2\},\ S_3 = \{3\}$. Action 1 means contribute. Let $k(a) = a_1 + a_2 + a_3$ be the number of contributors. At the first layer,

$$P(1 \mid 0, a) = \begin{cases} 0.85, & k(a) \ge 2, \\ 0.20, & k(a) < 2, \end{cases} \qquad P(2 \mid 0, a) = 1 - P(1 \mid 0, a), \tag{34}$$

with $c_i(0,a) = 0.08 + 0.22a_i + 0.35\frac{\max\{0, 2-k(a)\}}{2}$. At the second layer, the game terminates. If $s = 1$, then

$$c_i(1,a) = 0.10 + 0.18a_i + 0.08\mathbf{1}\{k(a) < 1\}. \tag{35}$$

If $s = 2$, then

$$c_i(2,a) = 0.55 + 0.15a_i - 0.10\min\{k(a), 2\}. \tag{36}$$

F.2.4 TRANSITION-TRAP GAME

This is a two-player, two-action, horizon-3 game with layers $S_1 = \{0\},\ S_2 = \{1,2\},\ S_3 = \{3,4\},\ S_4 = \{5\}$. Action 0 is a shortcut and action 1 is safe. The shortcut gives lower immediate cost but increases the probability of entering a dangerous or trapped state.

At $h = 1$, let $n_0(a) = \mathbf{1}\{a_1 = 0\} + \mathbf{1}\{a_2 = 0\}$. The probability of moving to state 2 is

$$p_{\text{danger}}(n_0) = \begin{cases} 0.10, & n_0 = 0, \\ 0.50, & n_0 = 1, \\ 0.80, & n_0 = 2. \end{cases} \tag{37}$$

The transition is $P(2 \mid 0, a) = p_{\text{danger}}(n_0(a))$, and the immediate cost is 0.05 for choosing action 0 and 0.13 for choosing action 1.

At $h = 2$, from $s = 1$ the trap probability is $(0.05, 0.20, 0.50)$ for $n_0 = 0, 1, 2$, while from $s = 2$ it is $(0.10, 0.50, 0.80)$. The immediate cost is 0.06 for action 0 and 0.20 for action 1, with an additional 0.02 cost in state 2.

At $h = 3$, the game terminates. In state 3, every action has cost 0.20. In the trap state 4, choosing shortcut action 0 costs 0.80, while action 1 costs 0.60.

F.3 Discounted LQ Markov Game

The discounted case is a two-player, two-state, zero-sum Markov game. The state is $s \in \{0, 1\}$, player 0 chooses $u \in [-U, U]$, and player 1 chooses $v \in [-U, U]$. Player 0 minimizes the raw cost

$$g_s(u, v) = \frac{r}{2}u^2 - \frac{r}{2}v^2 + b_s uv + a_s u + d_s v + k_s, \tag{38}$$

while player 1 minimizes $-g_s(u, v)$. The transition is

$$P(s' = 1 \mid s, u, v) = p_s + \alpha_s u + \beta_s v. \tag{39}$$

For the numerical experiment, the raw zero-sum costs have been transformed by

$$c_0(s, u, v) = \frac{1}{2} + \frac{g_s(u, v)}{2M}, \qquad c_1(s, u, v) = \frac{1}{2} - \frac{g_s(u, v)}{2M}. \tag{40}$$

The parameters are

s	r	b_s	a_s	d_s	k_s	p_s	α_s	β_s	U	γ	M
0	1.50	0.35	1.10	0.90	−0.35	0.22	0.07	−0.055	1	0.90	2.40
1	1.50	−0.30	−1.05	−0.95	0.45	0.78	−0.05	0.060	1	0.90	2.40

Also, our infinite-horizon benchmark can be viewed as a finite-state stochastic dynamic-game analogue of the classical two-player zero-sum LQ dynamic games studied in Başar & Olsder (1998).

F.3.1 Closed-form saddle-point solution

Let V_s be the raw, unnormalized infinite-horizon discounted value of player 0 at state s, and define $\Delta := V_1 - V_0$. The Bellman-Isaacs saddle-point equation is $V_s = \min_u \max_v \{g_s(u, v) + \gamma \mathbb{E}[V_{s'} \mid s, u, v]\}$. Thus,

$$\mathbb{E}[V_{s'} \mid s, u, v] = (1 - p_s - \alpha_s u - \beta_s v)V_0 + (p_s + \alpha_s u + \beta_s v)V_1 = V_0 + (p_s + \alpha_s u + \beta_s v)\Delta. \tag{41}$$

Therefore the Bellman-Isaacs objective equals

$$\gamma V_0 + \gamma p_s \Delta + k_s + \frac{r}{2}u^2 - \frac{r}{2}v^2 + b_s uv + A_s(\Delta)u + D_s(\Delta)v, \tag{42}$$

where $A_s(\Delta) := a_s + \gamma \alpha_s \Delta$, $D_s(\Delta) := d_s + \gamma \beta_s \Delta$. Thus

$$V_s = \gamma V_0 + \gamma p_s \Delta + k_s + \phi_s(\Delta), \tag{43}$$

where $\phi_s(\Delta) := \min_u \max_v \left\{ \frac{r}{2}u^2 - \frac{r}{2}v^2 + b_s uv + A_s(\Delta)u + D_s(\Delta)v \right\}$. The first-order saddle-point equations are

$$ru + b_s v + A_s(\Delta) = 0, \qquad -rv + b_s u + D_s(\Delta) = 0. \tag{44}$$

Solving these gives

$$u_s^*(\Delta) = -\frac{rA_s(\Delta) + b_s D_s(\Delta)}{r^2 + b_s^2}, \quad v_s^*(\Delta) = \frac{rD_s(\Delta) - b_s A_s(\Delta)}{r^2 + b_s^2}. \tag{45}$$

Substituting (45) into the optimized quadratic yields

$$\phi_s(\Delta) = \frac{1}{2}\frac{r\big(D_s(\Delta)^2 - A_s(\Delta)^2\big) - 2b_s A_s(\Delta) D_s(\Delta)}{r^2 + b_s^2}. \tag{46}$$

Subtracting (43) for $s = 0$ from the same equation for $s = 1$ yields the scalar equation

$$\Delta = \gamma(p_1 - p_0)\Delta + (k_1 - k_0) + \phi_1(\Delta) - \phi_0(\Delta). \tag{47}$$

Since ϕ_s is quadratic in Δ, equation (47) is a scalar quadratic. We solve this equation numerically, and select the feasible root whose saddle-point actions lie in $[-U, U]$, and then recovers $V_0 = \frac{\gamma p_0 \Delta + k_0 + \phi_0(\Delta)}{1-\gamma}$, $V_1 = V_0 + \Delta$. For the transformed normalized costs, the corresponding values are

$$J_0^*(s) = \frac{1}{2} + \frac{(1-\gamma)V_s}{2M}, \qquad J_1^*(s) = \frac{1}{2} - \frac{(1-\gamma)V_s}{2M}. \tag{48}$$