$\widetilde{O}(T^{-1})$ Convergence to (Coarse) Correlated Equilibria in Full-Information General-Sum Markov Games

Weichao Mao   WEICHAO2@ILLINOIS.EDU
Haoran Qiu   HAORANQ4@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Chen Wang   CHEN.WANG1@IBM.COM
Hubertus Franke   FRANKEH@US.IBM.COM
IBM Research

Zbigniew Kalbarczyk   KALBARCZ@ILLINOIS.EDU
Tamer Başar   BASAR1@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Abstract

No-regret learning has a long history of being closely connected to game theory. Recent works have devised uncoupled no-regret learning dynamics that, when adopted by all the players in normal-form games, converge to various equilibrium solutions at a near-optimal rate of $\widetilde{O}(T^{-1})$, a significant improvement over the $O(1/\sqrt{T})$ rate of classic no-regret learners. However, analogous convergence results are scarce in Markov games, a more generic setting that lays the foundation for multi-agent reinforcement learning. In this work, we close this gap by showing that the optimistic-follow-the-regularized-leader (OFTRL) algorithm, together with appropriate value update procedures, can find $\widetilde{O}(T^{-1})$-approximate (coarse) correlated equilibria in full-information general-sum Markov games within T iterations. Numerical results are also included to corroborate our theoretical findings.[1]

Keywords: Learning in games, reinforcement learning, correlated equilibrium, no-regret learning

1. Introduction

Online learning has an intimate connection to game theory for finding solutions under various equilibrium concepts (Robinson, 1951). In no-regret learning, the learner aims to maximize its cumulative utility in response to the (possibly adversarial) outcome sequence generated by the environment. For normal-form games (NFGs), a folklore result states that if all the players adopt certain no-regret learning algorithms that have $O(\sqrt{T})$ regret guarantees against an adversarial environment, then they can find an $O(1/\sqrt{T})$-approximate Nash equilibrium in two-player zero-sum games or an $O(1/\sqrt{T})$-approximate (coarse) correlated equilibrium in general-sum games after T iterations (Hart and Mas-Colell, 2000; Cesa-Bianchi and Lugosi, 2006). A broad family of no-regret learning algorithms fit into this category, including the well-known multiplicative weight updates (Littlestone and Warmuth, 1994; Freund and Schapire, 1997), follow-the-regularized/perturbed-leader (Kalai and Vempala, 2005), and mirror descent (Nemirovskij and Yudin, 1983).

---

1. This preprint carries essentially the same results and title as one that was submitted to a conference on December 8, 2023, except the appearance here on page 3 of a post-submission note which provides a comparison with a recently (January 26, 2024) posted arXiv paper by Cai et al. (2024).

Table 1: No-regret learning convergence rates in NFGs and Markov games.

Learning objective	Normal-form games	Markov games
Nash equilibrium (two-player zero-sum)	$\widetilde{O}(T^{-1})$ (Daskalakis et al., 2011)	$O(T^{-1})$ (Yang and Ma, 2023)
Correlated equilibrium	$\widetilde{O}(T^{-1})$ (Anagnostides et al., 2022b)	$\widetilde{O}(T^{-1/4})$ (Erez et al., 2023) $\widetilde{O}(T^{-1})$ (Theorem 1)
Coarse correlated equilibrium	$\widetilde{O}(T^{-1})$ (Daskalakis et al., 2021)	$\widetilde{O}(T^{-3/4})$ (Zhang et al., 2022) $\widetilde{O}(T^{-1})$ (Theorem 2)

While the $O(\sqrt{T})$ regret is unimprovable against an adversarial environment, it need not be the case for learning equilibria in games because each player in a repeated game is not facing adversarial payoffs, but instead is interacting with other players who also exhibit certain learning behavior. Indeed, the seminal work (Daskalakis et al., 2011) developed an algorithm based on the Nesterov's excessive gap technique and established its $\widetilde{O}(T^{-1})$ convergence[2] to Nash equilibria (NE) in two-player zero-sum NFGs when the algorithm is adopted by both players. Recent works (Rakhlin and Sridharan, 2013; Syrgkanis et al., 2015; Foster et al., 2016; Chen and Peng, 2020; Daskalakis et al., 2021; Anagnostides et al., 2022a,b) significantly strengthened this line of results by devising other no-regret learning dynamics that find different equilibrium solutions at a faster rate than $O(1/\sqrt{T})$. Notably, Syrgkanis et al. (2015) showed that if all the players in a general-sum NFG employ an *optimistic* version of follow-the-regularizer-leader (henceforth OFTRL), the players' strategies converge to the set of coarse correlated equilibria (CCE) at a fast rate of $O(T^{-3/4})$; such a rate was later improved to $\widetilde{O}(T^{-1})$ by Daskalakis et al. (2021). More recently, the $\widetilde{O}(T^{-1})$ rate was established for swap regrets and correlated equilibria (CE) in NFGs (Anagnostides et al., 2022a,b).

Despite the encouraging fast convergence results in NFGs, very few results are known for the more challenging regime of Markov games (also known as stochastic games (Shapley, 1953)). The only exceptions include Zhang et al. (2022) and Yang and Ma (2023), who established the $\widetilde{O}(T^{-1})$ convergence of OFTRL (together with smooth value updates) to NE in two-player zero-sum full-information Markov games, matching the best rates in NFGs. As for general-sum Markov games, the best known results for CCE and CE are $\widetilde{O}(T^{-3/4})$ (Zhang et al., 2022) and $\widetilde{O}(T^{-1/4})$ (Erez et al., 2023), respectively, which largely lag behind their $\widetilde{O}(T^{-1})$ counterparts in NFGs. In fact, establishing $\widetilde{O}(T^{-1})$ convergence to CCE or CE in general-sum Markov games has been raised as an important open question by Yang and Ma (2023).

Contributions. In this work, we close this gap by developing no-regret learning algorithms with accompanying value update procedures and establishing their fast $\widetilde{O}(T^{-1})$ convergence to CCE or CE in general-sum Markov games. For CE (Section 3), we consider the OFTRL algorithm with a log-barrier regularizer, and integrate it with the celebrated external-to-swap-regret reduction (Blum and Mansour, 2007) and smooth value updates. Our $\widetilde{O}(T^{-1})$ convergence analysis builds on a Regret bounded by Variation in Utilities (RVU) property (Syrgkanis et al., 2015) for the weighted swap regret at each state. We make a seemingly trivial observation that swap regrets are always non-negative and use it to easily bound the second-order path lengths of the learning dynamics. For CCE (Section 4), we consider standard OFTRL with negative entropy regularization but combine it with a stage-based value update scheme. We show that this algorithm induces a no-*average*-regret problem

---

2. Throughout the paper, we use $\widetilde{O}(\cdot)$ to suppress the poly-logarithmic dependence on T.

within each stage, which allows us to apply existing analysis for the *individual* regret of the players (Daskalakis et al., 2021). Table 1 compares our results with the best-known convergence rates of no-regret learning in NFGs and Markov games. We further provide numerical results (Section 5) to validate the $\widetilde{O}(T^{-1})$ convergence behavior of our algorithms.

Post conference submission note. An independent paper by Cai et al. (2024) investigates the same problem of $\widetilde{O}(T^{-1})$ convergence to CE in general-sum Markov games. They also consider OFTRL with the log-barrier regularizer and smooth value updates and establish a similar convergence rate as in Theorem 1 of this paper. Interestingly, their algorithm performs V-value updates, which allows for a more preferred decentralized implementation, and they prove its equivalence to Q-value updates as we had done in this paper. Compared with their convergence rate for CE, however, our Theorem 1 happens to shave off an additional $\log T$ factor by using a more refined analysis of a weighted average of a sequence (Lemma 4 from Yang and Ma (2023)). Our work has further established the $\widetilde{O}(T^{-1})$ convergence to CCE (Section 4), which has not been considered by Cai et al. (2024).

Further related work. Learning in Markov games has been widely studied in multi-agent reinforcement learning (MARL) (Bai and Jin, 2020; Wei et al., 2021; Cen et al., 2021, 2022; Zhao et al., 2022; Cai et al., 2023; Wang et al., 2023; Mao et al., 2023). Closest to ours are the works by Liu et al. (2021); Jin et al. (2022); Song et al. (2022); Mao et al. (2022); Mao and Başar (2023) who have shown $\widetilde{O}(1/\sqrt{T})$ convergences to CCE or CE in multi-player general-sum Markov games under bandit feedback.

2. Preliminaries

No-regret learning. Let $\mathcal{A}$ be a finite set of actions. At each iteration $t \in \mathbb{N}_+$, a learning agent selects a strategy $\boldsymbol{x}^t \in \Delta(\mathcal{A})$ as a probability distribution over the action space. The environment returns a utility vector $\boldsymbol{u}^t \in \mathbb{R}^{|\mathcal{A}|}$, and the agent obtains a utility of $\langle \boldsymbol{x}^t, \boldsymbol{u}^t \rangle$. The classic notion of regret, or more generally Φ-regret (Greenwald and Jafari, 2003), is used to measure the performance of the learning agent in terms of the suboptimality in hindsight. Formally, given a sequence of strategies $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^T)$ over T iterations, the incurred Φ-regret is defined as

$$\mathrm{Reg}_\Phi^T := \max_{\phi^* \in \Phi} \sum_{t=1}^T \langle \phi^*(\boldsymbol{x}^t) - \boldsymbol{x}^t, \boldsymbol{u}^t \rangle. \tag{1}$$

In particular, Reg_Φ^T is called *external regret* (or simply *regret*) if Φ is the set of all constant functions $\{\phi : \phi(\boldsymbol{x}) = \phi(\boldsymbol{x}'), \forall \boldsymbol{x}, \boldsymbol{x}' \in \Delta(\mathcal{A})\}$, and *swap regret* if Φ is the set of all linear transformations $\{\phi : \phi(\boldsymbol{x}) = Q^\top \boldsymbol{x}, \text{ where } Q \text{ is a row stochastic matrix}\}$. One can see that swap regret is a more powerful notion of hindsight rationality by allowing a broader class of possible deviations.

Markov game. An N-player episodic Markov game is defined by a tuple $\mathbb{G} = (\mathcal{N}, H, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, \{r_i\}_{i=1}^N, P)$, where (1) $\mathcal{N} = \{1, 2, \ldots, N\}$ is the set of agents; (2) $H \in \mathbb{N}_+$ is the number of time steps in each episode; (3) $\mathcal{S}$ is the finite state space; (4) $\mathcal{A}_i$ is the finite action space for agent $i \in \mathcal{N}$; (5) $r_i : [H] \times \mathcal{S} \times \mathcal{A}_{\mathrm{all}} \to [0, 1]$ is the reward function for agent i, where $\mathcal{A}_{\mathrm{all}} = \times_{i=1}^N \mathcal{A}_i$ is the joint action space; and (6) $P : [H] \times \mathcal{S} \times \mathcal{A}_{\mathrm{all}} \to \Delta(\mathcal{S})$ is the state transition function. The agents interact in an unknown environment for T episodes. At each step $h \in [H]$, the agents observe the state $s_h \in \mathcal{S}$ and take actions $a_{h,i} \in \mathcal{A}_i$ simultaneously. Agent i then receives its reward $r_{h,i}(s_h, \boldsymbol{a}_h)$, where $\boldsymbol{a}_h = (a_{h,1}, \ldots, a_{h,N}) \in \mathcal{A}_{\mathrm{all}}$, and the environment transitions to the next state $s_{h+1} \sim P_h(\cdot | s_h, \boldsymbol{a}_h)$. We make a standard assumption (Jin et al., 2022; Song et al., 2022) that each episode starts from a fixed initial state s_1. Let $S = |\mathcal{S}|$, $A_i = |\mathcal{A}_i|$, and $A_{\max} = \max_{i \in \mathcal{N}} A_i$.

Policy and value function. A (Markov) policy $\pi_i \in \Pi_i : [H] \times \mathcal{S} \to \Delta(\mathcal{A}_i)$ for agent $i \in \mathcal{N}$ is a mapping from the time index and state space to a distribution over its own action space. Each agent seeks to find a policy that maximizes its own cumulative reward. A joint, product policy $\pi = (\pi_1, \ldots, \pi_N) \in \Pi$ induces a probability measure over the sequence of states and joint actions. We use the subscript $-i$ to denote the set of agents excluding agent i, i.e., $\mathcal{N} \backslash \{i\}$. We can rewrite $\pi = (\pi_i, \pi_{-i})$ using this convention. For a joint policy π, and for any $h \in [H]$, $s \in \mathcal{S}$, and $\boldsymbol{a} \in \mathcal{A}_{\text{all}}$, we define the value function and Q-function for agent i as

$$V_{h,i}^\pi(s) = \mathbb{E}\Big[\sum_{h'=h}^H r_{h',i}(s_{h'}, \boldsymbol{a}_{h'})|s_h = s\Big], Q_{h,i}^\pi(s, \boldsymbol{a}) = \mathbb{E}\Big[\sum_{h'=h}^H r_{h',i}(s_{h'}, \boldsymbol{a}_{h'})|s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\Big]. \quad (2)$$

For notational convenience, for any $V : \mathcal{S} \to \mathbb{R}$, we define $[P_h V](s, \boldsymbol{a}) := \mathbb{E}_{s' \sim P_h(\cdot|s, \boldsymbol{a})} [V(s')]$. For an arbitrary Q-function $Q_{h,i} : \mathcal{S} \times \mathcal{A}_{\text{all}} \to \mathbb{R}$, we write $[Q_{h,i} \pi_h](s) := \langle Q_{h,i}(s, \cdot), \pi_h(s, \cdot) \rangle$ and $[Q_{h,i} \pi_{h,-i}](s, a_i) := \langle Q_{h,i}(s, a_i, \cdot), \pi_{h,-i}(s, \cdot) \rangle$ for short.

Full-information feedback. Following Zhang et al. (2022); Yang and Ma (2023); Erez et al. (2023), we consider the full-information feedback setting where each agent can observe the expected rewards it would have received had it played any candidate action. In our formulation, this can be interpreted as an oracle from which each agent i can query $[Q_{h,i} \pi_{h,-i}](s, a_i)$ for each candidate action $a_i \in \mathcal{A}_i$ at any state $s \in \mathcal{S}$.

Correlated policy and (coarse) correlated equilibrium. We define $\pi = \{\pi_h : \mathbb{R} \times (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \Delta(\mathcal{A})\}_{h \in [H]}$ as a (non-Markov) *correlated policy*. Specifically, the agents first sample the value of $z \in \mathbb{R}$ from an underlying distribution using a common source of randomness (e.g., a common random seed), and then they can use z to coordinate their choices of actions. Such a virtual coordinator is crucial for learning CCE/CE as shown in the literature (Mao and Başar, 2023; Song et al., 2022; Jin et al., 2022), and is standard in decentralized learning (Bernstein et al., 2009; Arabneydi and Mahajan, 2015; Mao et al., 2020). For each $h \in [H]$, π_h maps from $z \in \mathbb{R}$ and a state-action history $(s_1, \boldsymbol{a}_1, \ldots, s_{h-1}, \boldsymbol{a}_{h-1}, s_h)$ to a distribution over the joint action space $\mathcal{A}_{\text{all}}$. Unlike product policies, such action distributions in general cannot be factorized into independent probability distributions over the individual action spaces. For a correlated policy π, we let π_i and π_{-i} be the proper marginal distributions of π whose outputs are restricted to $\Delta(\mathcal{A}_i)$ and $\Delta(\mathcal{A}_{-i})$, respectively. The value functions for non-Markov correlated policies at step $h = 1$ are defined similarly as those for product policies (2).

For any correlated policy $\pi = (\pi_i, \pi_{-i})$, the best response value of agent i is denoted by $V_{1,i}^{\dagger, \pi_{-i}}(s_1) := \sup_{\pi_i^\dagger} V_{1,i}^{\pi_i^\dagger, \pi_{-i}}(s_1)$, where the supremum is taken over all (non-Markov) policies of agent i independent of the randomness of π_{-i}. A policy $\pi_i^\dagger$ is agent i's best response to π_{-i} if it achieves the supremum. Given the PPAD-hardness of Nash equilibria (Daskalakis et al., 2009), people often study relaxed equilibrium concepts in general-sum games, such as CCE and CE.

Definition 1. *(ε-CCE) For any $\varepsilon > 0$, a correlated policy $\pi = (\pi_i, \pi_{-i})$ is an ε-approximate coarse correlated equilibrium if $V_{1,i}^{\pi_i, \pi_{-i}}(s_1) \geq V_{1,i}^{\dagger, \pi_{-i}}(s_1) - \varepsilon, \forall i \in \mathcal{N}$.*

To properly define a CE, we need to first specify the concept of a strategy modification. Formally, for agent i, a strategy modification $\phi_i = \{\phi_{h,i}^s : h \in [H], s \in \mathcal{S}\}$ is a set of mappings from agent i's action space to itself, i.e., $\phi_{h,i}^s : \mathcal{A}_i \to \mathcal{A}_i$. Given a strategy modification ϕ_i, whenever a policy π selects the joint action $\boldsymbol{a} = (a_1, \ldots, a_N)$ at step h and state s, the modified policy $\phi_i \diamond \pi$ will select $(a_1, \ldots, a_{i-1}, \phi_{h,i}^s(a_i), a_{i+1}, \ldots, a_N)$ instead. Let Φ_i denote the set of all possible

---

Algorithm 1 Optimistic follow-the-regularized-leader for correlated equilibria (agent i)

Initialize: $Q_{h,i}^0(s, \boldsymbol{a}) \leftarrow 0, \pi_{h,i}^0(s, a_i) \leftarrow 1/A_i, \forall s \in \mathcal{S}, h \in [H], a_i, a_i' \in \mathcal{A}_i, \boldsymbol{a} \in \mathcal{A}_{\text{all}};$

for *iteration* $t \leftarrow 1$ *to* T **do**

    Policy update:

    for *action* $a_i \in \mathcal{A}_i$ **do**

        $\ell_{h,i}^{t,a_i}(s, a_i') \leftarrow \sum_{j=1}^{t-1} w_j \pi_{h,i}^j(s, a_i)[Q_{h,i}^j \pi_{h,-i}^j](s, a_i') + w_t \pi_{h,i}^{t-1}(s, a_i)[Q_{h,i}^{t-1} \pi_{h,-i}^{t-1}](s, a_i');$

        $q_{h,i}^{t,a_i}(s, \cdot) \leftarrow \operatorname{argmax}_{\boldsymbol{x} \in \Delta(\mathcal{A}_i)} \left(\langle \boldsymbol{x}, \eta \ell_{h,i}^{t,a_i}(s, \cdot)/w_t \rangle - \mathcal{R}(\boldsymbol{x}) \right), \forall s \in \mathcal{S}, h \in [H];$

    Find $\pi_{h,i}^t$ such that $\pi_{h,i}^t(s, \cdot) = \sum_{a_i \in \mathcal{A}_i} \pi_{h,i}^t(s, a_i) q_{h,i}^{t,a_i}(s, \cdot), \forall s \in \mathcal{S}, h \in [H], a_i \in \mathcal{A}_i;$

    Value update:

    for $h \leftarrow H$ *to* 1 **do**

        $Q_{h,i}^t(s, \boldsymbol{a}) \leftarrow (1 - \alpha_t) Q_{h,i}^{t-1}(s, \boldsymbol{a}) + \alpha_t \left(r_{h,i} + P_h[Q_{h+1,i}^t \pi_{h+1}^t] \right)(s, \boldsymbol{a}), \forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}_{\text{all}};$

Output policy: $\bar{\pi} = \bar{\pi}_1^T$, where $\bar{\pi}_h^t$ is defined in Algorithm 2.

---

strategy modifications for agent i. A CE states that no agent has the incentive to deviate from a correlated policy π by using any strategy modification.

Definition 2. *(ε-CE) For any $\varepsilon > 0$, a correlated policy π is an ε-approximate correlated equilibrium if $V_{1,i}^\pi(s_1) \geq \max_{\phi_i \in \Phi_i} V_{1,i}^{\phi_i \diamond \pi}(s_1) - \varepsilon, \forall i \in \mathcal{N}$.*

3. Convergence to Correlated Equilibria

In this section, we present our optimistic follow-the-regularized-leader (OFTRL) algorithm for learning correlated equilibria in general-sum Markov games in Section 3.1, and then establish its $\widetilde{O}(T^{-1})$ convergence in Section 3.2.

3.1. Algorithm

Algorithm 1 describes the OFTRL procedure run by agent $i \in \mathcal{N}$. Since the algorithms run by all the agents are exactly symmetric, in the following, we only illustrate our algorithm using a single agent i as an example. Algorithm 1 consists of three major components: The policy update step that computes the strategy for each matrix game, the value update step that updates the (Q-)value functions, and the policy output step that generates a CE policy.

Policy update. At each fixed $(s, h) \in \mathcal{S} \times [H]$, the agents are essentially faced with a sequence of matrix games, where the payoff matrix for agent i in the t-th matrix game is given by the estimated Q-function $Q_{h,i}^t(s, \cdot)$ at the corresponding iteration t. For learning CE in matrix games, a folklore result suggests that each agent should employ a no-swap-regret learning algorithm. Specifically, suppose that each agent employs a no-swap-regret algorithm such that the cumulative swap regret up to time $T \in \mathbb{N}_+$ is upper bounded by SwapReg^T; then, the empirical distribution of the joint actions played by the players is an $(\text{SwapReg}^T/T)$-approximate CE (Hart and Mas-Colell, 2000).

For a fixed matrix game at $(s, h) \times \mathcal{S} \times [H]$, we follow the generic reduction introduced in (Blum and Mansour, 2007) to obtain a no-swap-regret learning algorithm $\mathscr{A}_{\text{swap}}$ from a no-(external-)regret base algorithm $\mathscr{A}$. Specifically, Blum and Mansour (2007) maintain a separate no-regret algorithm $\mathscr{A}_a$ for each candidate action $a \in \mathcal{A}_i$ of the agent. $\mathscr{A}_{\text{swap}}$ computes a strategy by combining the strategies of the A_i base algorithms. At time step $t \in [T]$, each base algorithm $\mathscr{A}_a$ outputs a distribution $q^{t,a}(\cdot) \in \Delta(\mathcal{A}_i)$, where $q^{t,a}(a')$ is the probability that it selects $a' \in \mathcal{A}_i$. Then, a (row)

Algorithm 2 Policy $\bar{\pi}_h^t$

Input: Policy trajectory $\{\pi_h^t\}_{h \in [H], t \in [T]}$ of Algorithm 1;
for *step $h' \leftarrow h$ to H* **do**
    Sample $\tau \in [t]$ with probability $\mathbb{P}(\tau = j) = \alpha_t^j$;
    Play policy $\pi_{h'}^\tau$ at step h';
    Set $t \leftarrow \tau$.

---

stochastic matrix $q^t \in \mathbb{R}^{A_i \times A_i}$ is constructed, where the a-th row of q^t is equal to the $q^{t,a}$ vector. $\mathscr{A}_{\text{swap}}$ obtains the action selection strategy by computing a stationary distribution[3] $\pi^t \in \Delta(\mathcal{A}_i)$ of q^t such that $(q^t)^\top \pi^t = \pi^t$. Upon receiving the payoff vector $\boldsymbol{u}^t \in \mathbb{R}^{A_i}$ (in the case of Algorithm 1, $\boldsymbol{u}^t = [Q_{h,i}^j \pi_{h,-i}^j](s, \cdot)$ for agent i) from the environment, $\mathscr{A}_{\text{swap}}$ returns to each $\mathscr{A}_a$ base algorithm a $\pi^t(a)$ fraction of the received utility, so that $\mathscr{A}_a$ is updated with a utility vector of $\pi^t(a)\boldsymbol{u}^t \in \mathbb{R}^{A_i}$. It is shown that $\mathscr{A}_{\text{swap}}$ guarantees no-swap-regret as long as each base algorithm $\mathscr{A}_a$ has sublinear (external) regret in T.

In Algorithm 1, we use weighted OFTRL as the no-regret base algorithm $\mathscr{A}$. OFTRL (Syrgkanis et al., 2015) extends the standard FTRL paradigm by maintaining a prediction sequence $\boldsymbol{m}^t$ of the utilities. Given a utility sequence $(\boldsymbol{u}^1, \ldots, \boldsymbol{u}^T)$, OFTRL computes the strategies by

$$\boldsymbol{x}^t := \operatorname*{argmax}_{\boldsymbol{x} \in \Delta(\mathcal{A}_i)} \left\{ \eta \Big\langle \boldsymbol{x}, \boldsymbol{m}^t + \sum_{j=1}^{t-1} \boldsymbol{u}^j \Big\rangle - \mathcal{R}(\boldsymbol{x}) \right\}, \tag{3}$$

where $\eta > 0$ is the learning rate, and $\mathcal{R}$ is the regularizer. In Algorithm 1, we instantiate (3) with $\boldsymbol{m}^t = \boldsymbol{u}^{t-1}$ and the log-barrier regularizer $\mathcal{R}(\boldsymbol{x}) = -\sum_{a_i \in \mathcal{A}_i} \log(\boldsymbol{x}[a_i])$. Such a log-barrier regularizer satisfies the self-concordant condition in Anagnostides et al. (2022b), which is used to establish the Regret bounded by Variation in Utilities (RVU) property (Syrgkanis et al., 2015) of the swap regret. Due to the time-varying learning rates in the value update step (to be discussed momentarily), we additionally use a weighted variant of OFTRL that considers a weighted sum over the utility sequence. The choice of the weights $\{w_j\}_{j \in [t]}$ will also be defined shortly. Combining the OFTRL base algorithm, the utility weights and the external-to-swap-regret reduction, we arrive at the policy update rule as presented in Algorithm 1. With the Blum and Mansour (2007) reduction, we name our no-swap-regret algorithm BM-OFTRL.

Value update. For any $(h, s, \boldsymbol{a})$, we update the Q-value estimates at each iteration in a Bellman manner using a weighted average of previous estimates. We perform incremental updates using the classic step size $\alpha_t = (H + 1)/(H + t)$ proposed by Jin et al. (2018). With this step size, the value update rule in Algorithm 1 effectively becomes:

$$Q_{h,i}^t(s, \boldsymbol{a}) = \sum_{j=1}^t \alpha_t^j \left(r_{h,i} + P_h[Q_{h+1,i}^j \pi_{h+1}^j] \right)(s, \boldsymbol{a}), \forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}_{\text{all}}, \tag{4}$$

where $\alpha_t^j := \alpha_j \prod_{j'=j+1}^t (1 - \alpha_{j'})$ and $\alpha_t^t := \alpha_t$. One can verify that $\sum_{j=1}^t \alpha_t^j = 1$. Given the time-varying weights α_t^j, to ensure that our policy update step is no-swap-regret in the matrix games defined by the Q-value estimates, we define the weights of our weighted OFTRL procedure in Algorithm 1 to be $w_j := \alpha_t^j/\alpha_t^1$ for any fixed $t \in [T]$.

---

3. It is known that such a distribution π^t exists and is computationally efficient.

Policy output. Our output policy $\bar{\pi}$ is a state-wise weighted average of the history policies, where the weights are again related to the step sizes α_t^j. The construction of $\bar{\pi}$ is formally defined in Algorithm 2, which is closely related to the "certified policies" from Bai et al. (2020). Specifically, Algorithm 2 takes the policy trajectory $\{\pi_h^t\}_{h \in [H], t \in [T]}$ of Algorithm 1 as input. For each step $h \in [H]$, Algorithm 2 randomly samples a joint policy from the policy trajectory using the sampling probabilities α_t^j and let all the agents play this joint policy at the given step. Similar to Song et al. (2022); Mao et al. (2022); Zhang et al. (2022), the constructed policy $\bar{\pi}$ is a correlated policy because the agents implicitly use a common source of randomness to select the same history iteration. We will show that the output policy constitutes an approximate CE. We also remark that existing results for learning Nash equilibria in two-player zero-sum Markov games (Zhang et al., 2022; Yang and Ma, 2023) do not require such a shared randomness and generally output Markov policies.

3.2. Analysis

In the following, we present the analysis of Algorithm 1. We use the following notion of CE-Gap to measure the distance of a correlated policy to a CE:

$$\text{CE-Gap}(\pi) := \max_{i \in \mathcal{N}} \max_{\phi_i \in \Phi_i} \left(V_{1,i}^{\phi_i \diamond \pi}(s_1) - V_{1,i}^{\pi}(s_1) \right),$$

where recall that Φ_i is the set of strategy modifications for agent i. The following theorem states that Algorithm 1 finds an $\widetilde{O}(T^{-1})$-approximate CE in T iterations.

Theorem 1. *If Algorithm 1 is run on an N-player episodic Markov game for T iterations with a learning rate $\eta = \frac{1}{256NH\sqrt{HA_{\max}}}$, the output policy $\bar{\pi}$ satisfies:*

$$\text{CE-Gap}(\bar{\pi}) \leq \frac{6144NH^{\frac{7}{2}}A_{\max}^{\frac{5}{2}}\log T}{T}.$$

Theorem 1 improves the existing $\widetilde{O}(T^{-1/4})$ rate (Erez et al., 2023) of no-regret learning to CE in full-information Markov games. The parameter dependences in Theorem 1 also match the best known rate for normal-form games (Anagnostides et al., 2022b), except that Theorem 1 introduces an additional $O(H^{\frac{7}{2}})$ dependence on the Markov game episode length. We remark that we make no effort to improve the constant factors in the bounds, which can certainly be tightened.

The proof structure of Theorem 1 is conceptually similar to those for learning Nash equilibria in two-player zero-sum Markov games (Zhang et al., 2022; Yang and Ma, 2023) . We first introduce a few notations to facilitate the proof. For any $(s, h) \in \mathcal{S} \times [H]$, we define the per-state weighted swap regret up to iteration $t \in [T]$ in the corresponding matrix game as

$$\text{SwapReg}_{h,i}^t(s) := \max_{\phi_{h,i}^s : \mathcal{A}_i \to \mathcal{A}_i} \sum_{j=1}^t \alpha_t^j \left\langle \phi_{h,i}^s \diamond \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^j(s, \cdot), [Q_{h,i}^j \pi_{h,-i}^j](s, \cdot) \right\rangle,$$

$$\text{SwapReg}_h^t := \max_{i \in \mathcal{N}} \max_{s \in \mathcal{S}} \text{SwapReg}_{h,i}^t(s).$$

For any $(h, t) \in [H] \times [T]$, we further define the best response CE value gap as

$$\delta_h^t := \max_{i \in \mathcal{N}} \max_{\phi_i} \max_{s \in \mathcal{S}} \left(V_{h,i}^{\phi_i \diamond \bar{\pi}_h^t}(s) - V_{h,i}^{\bar{\pi}_h^t}(s) \right),$$

where $\bar{\pi}_h^t$ is defined in Algorithm 2 and we slightly abuse the notation ϕ_i to denote a strategy modification that is only effective starting from step h. By the definition of δ_h^t and $\bar{\pi}$, one can easily

see that $\text{CE-Gap}(\bar{\pi}) = \text{CE-Gap}(\bar{\pi}_1^T) \le \delta_1^T$. To control δ_1^T, we first use the following lemma to establish the recursive relationship of the best response CE value gaps between two consecutive steps h and $h+1$:

Lemma 1. *(Recursion of best response CE value gaps) For any fixed* $(h,t) \in [H] \times [T]$, *we have*

$$\delta_h^t \le \sum_{j=1}^t \alpha_t^j \delta_{h+1}^j + \text{SwapReg}_h^t. \tag{5}$$

Therefore, upper bounding $\text{CE-Gap}(\bar{\pi})$ breaks down to controlling the per-state weighted swap regrets for every $(s,h) \in \mathcal{S} \times [H]$. We can further establish the upper bound of $\text{SwapReg}_{h,i}^t(s)$ in the next lemma. The proof of this lemma relies on an RVU bound for the swap regret of BM-OFTRL under time-varying learning rates in normal-form games.

Lemma 2. *(Per-state weighted swap regret bounds) For any* $t \in [T], h \in [H], s \in \mathcal{S}$ *and* $i \in \mathcal{N}$, *Algorithm 1 ensures that*

$$\text{SwapReg}_{h,i}^t(s) \le \frac{4A_i^2 H \log t}{\eta t} + \frac{32\eta H^3 N^2}{t} + 8\eta N H^2 \sum_{j=2}^t \sum_{k \ne i} \alpha_t^j \left\| \pi_{h,k}^j(s,\cdot) - \pi_{h,k}^{j-1}(s,\cdot) \right\|_1^2. \tag{6}$$

If $\eta \le \frac{1}{256NH\sqrt{HA_{\max}}}$, *we further have*

$$\sum_{i=1}^N \text{SwapReg}_{h,i}^t(s) \le \frac{4NA_{\max}^2 H \log t}{\eta t} + \frac{32\eta N H^2(N^2 + H)}{t}$$

$$- \frac{1}{2048\eta H} \sum_{i=1}^N \sum_{j=2}^t \frac{\alpha_t^j}{A_i} \left\| \pi_{h,i}^j(s,\cdot) - \pi_{h,i}^{j-1}(s,\cdot) \right\|_1^2. \tag{7}$$

We note that there is a discrepancy between (5) and (7). Specifically, (5) requires an upper bound for the *maximum* of the swap regrets over the agents while (7) controls the *sum* of them. This poses some additional challenges for learning NE (in zero-sum Markov games) or CCE in existing works (Zhang et al., 2022; Yang and Ma, 2023), because some players may experience negative regret (Hsieh et al., 2021) and the sum of regrets in general does not upper bound the maximum individual regret of the players. For CE, however, we can take advantage of a seemingly trivial property that the swap regret is always non-negative. This is in sharp contrast to the (external) regret and one can easily verify this property by letting all the strategy modifications $\phi_{h,i}^s$ in $\text{SwapReg}_{h,i}^t(s)$ be identity mappings. In this case, the discrepancy will not impede us as we can easily upper bound the maximum (5) by the sum (7), which already yields an $\widetilde{O}(t^{-1})$ convergence rate. Our proof of Theorem 1 instead follows a different route that upper bounds the second-order path lengths of the learning dynamics, which leads to an improved rate in terms of the dependence on N.

4. Convergence to Coarse Correlated Equilibria

4.1. Algorithm

Algorithm 3 describes the stage-based OFTRL procedure run by agent $i \in \mathcal{N}$ for learning CCE. Similar to Section 3, Algorithm 3 also consists of three components: policy update, value update, and policy output. The policy update step is standard OFTRL with a negative entropy regularizer, also known as the optimistic Hedge (see e.g., Chen and Peng (2020)). Our policy output step, formally described in Algorithm 4 (in Appendix C), is conceptually similar to Algorithm 2 for CE.

Algorithm 3 Stage-based OFTRL for coarse correlated equilibria (agent i)

Initialize: $Q_{h,i}^1(s, \boldsymbol{a}) \leftarrow 0, \pi_{h,i}^0(s, a_i) \leftarrow 1/A_i, \forall s \in \mathcal{S}, h \in [H], a_i, a_i' \in \mathcal{A}_i, \boldsymbol{a} \in \mathcal{A}_{\text{all}}$;

Set stage index $\tau \leftarrow 1, t_\tau^{\text{start}} \leftarrow 1$, and $L_\tau \leftarrow H$

for *iteration* $t \leftarrow 1$ *to* T **do**

    Policy update: For all $s \in \mathcal{S}, h \in [H]$, and $a_i \in \mathcal{A}_i$,

$$\ell_{h,i}^t(s, a_i) \leftarrow \sum_{t'=t_\tau^{\text{start}}}^{t-1} [Q_{h,i}^\tau \pi_{h,-i}^{t'}](s, a_i) + [Q_{h,i}^\tau \pi_{h,-i}^{t-1}](s, a_i);$$

$$\pi_{h,i}^t(s, \cdot) \leftarrow \underset{\boldsymbol{x} \in \Delta(\mathcal{A}_i)}{\text{argmax}} \left(\langle \boldsymbol{x}, \eta_\tau \ell_{h,i}^t(s, \cdot)/H \rangle - \mathcal{R}(\boldsymbol{x}) \right);$$

    if $t - t_\tau^{\text{start}} + 1 \geq L_\tau$ **then**

        $t_\tau^{\text{end}} \leftarrow t, t_{\tau+1}^{\text{start}} \leftarrow t + 1, L_{\tau+1} \leftarrow \lfloor (1 + 1/H) L_\tau \rfloor$;

        Value update: For each $h \in [H], s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}_{\text{all}}, i \in \mathcal{N}$:

$$Q_{h,i}^{\tau+1}(s, \boldsymbol{a}) \leftarrow \frac{1}{L_\tau} \sum_{t'=t_\tau^{\text{start}}}^{t_\tau^{\text{end}}} \left(r_{h,i} + P_h[Q_{h+1,i}^\tau \pi_{h+1}^{t'}] \right)(s, \boldsymbol{a});$$

        $\tau \leftarrow \tau + 1; \pi_{h,i}^t(s, a_i) \leftarrow 1/A_i, \forall s \in \mathcal{S}, h \in [H], a_i \in \mathcal{A}_i$

Output policy: Sample $t \sim \text{Unif}([T])$. Output $\bar{\pi} := \bar{\pi}_1^t$ where $\bar{\pi}_h^t$ is defined in Algorithm 4.

The value update step here is substantially different from that of Section 3. Rather than performing incremental updates as in Algorithm 1, we instead employ *stage-based* value updates by dividing the total T iterations into multiple stages and only updating the value estimates at the end of a stage. We use $\tau \in \mathbb{N}_+$ to index the stages and use L_τ to denote the length (i.e., number of iterations) of the τ-th stage. We set the lengths of the stages to grow exponentially at a rate of $(1 + 1/H)$ so that $L_{\tau+1} = \lfloor (1 + 1/H) L_\tau \rfloor$. The exponential growth ensures that the total T iterations can be covered by a small number of stages, while the $(1 + 1/H)$ growth rate guarantees that the value estimation error does not blow up during the H steps of recursion. Such a mechanism was initially proposed in single-agent RL (Zhang et al., 2020) and has later been advocated for creating a piece-wise stationary environment in MARL (Mao et al., 2022). The benefit of using stage-based value updates here is that we only need to bound the per-state *average* regret in the corresponding matrix games (in contrast to the weighted regret as in Section 3), which allows us to easily apply existing regret analysis results for normal-form games.

4.2. Analysis

We use the notion of CCE-Gap to measure the distance of a correlated policy to a CCE: $\text{CCE-Gap}(\pi)$ $:= \max_{i \in \mathcal{N}} (V_{1,i}^{\dagger, \pi_{-i}}(s_1) - V_{1,i}^\pi(s_1))$, where the best response value $V_{1,i}^{\dagger, \pi_{-i}}(s)$ is defined in Section 2. The following theorem shows that Algorithm 3 finds an $\widetilde{O}(T^{-1})$-approximate CCE in T iterations.

Theorem 2. *If Algorithm 3 is run on an N-player episodic Markov game for T iterations with a learning rate $\eta_\tau = \Theta(\frac{1}{N \log^4 L_\tau})$ in each stage τ, then the output policy $\bar{\pi}$ satisfies:*

$$\text{CCE-Gap}(\bar{\pi}) = O\left(\frac{NH^3 \log A_{\max} \cdot \log^5 T}{T} \right).$$
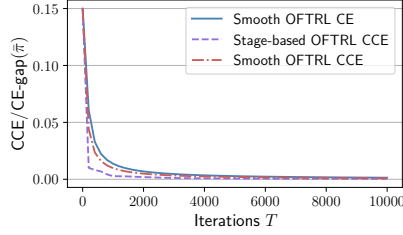
Figure 1: Convergence of CCE/CE-Gap($\bar{\pi}$)
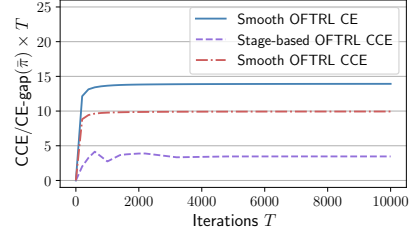
Figure 2: Convergence of CCE/CE-Gap($\bar{\pi}$) $\times T$

Theorem 2 improves the best-known rate of $\widetilde{O}(T^{-3/4})$ (Zhang et al., 2022) for OFTRL in general-sum Markov games. Compared to its counterpart $O(N \log A_{\max} \cdot \log^4 T / T)$ in normal-form games (Daskalakis et al., 2021), Theorem 2 incurs an extra $O(\log T)$ factor due to the stage-based value estimates. The proof of Theorem 2 starts by showing a recursive relationship of the best response CCE value gaps between two consecutive steps h and $h+1$. As a consequence of stage-based value updates, CCE-gap($\bar{\pi}$) breaks down to the sum of the per-state average regret over the stages, which allows us to apply each player's individual (average) regret bound in NFGs (Daskalakis et al., 2021) for each stage. The proof is then completed by upper bounding the total number of stages. We defer the complete proof of Theorem 2 to Appendix C due to space limitations.

5. Numerical Results

In this section, we numerically evaluate Algorithm 1 (denoted by "Smooth OFTRL CE") and Algorithm 3 ("Stage-based OFTRL CCE") to validate our $\widetilde{O}(T^{-1})$ theoretical convergence guarantees. Our simulations additionally consider an OFTRL algorithm with incremental value updates similar to that of Algorithm 1 for learning CCE ("Smooth OFTRL CCE"). We did not prove the convergence of such an algorithm but would be interested to see its numerical performance given its intuitive form. We conduct numerical studies on a simple general-sum Markov game with 2 players, 2 states, and 2 candidate actions for each player. Detailed definitions of the transition and reward functions of the game can be found in Appendix D. Figure 1 illustrates the convergence of the three algorithms to their corresponding equilibrium solutions as the number of iterations increases. To clearly demonstrate their convergence rates, we further plot the behavior of CCE/CE-Gap($\bar{\pi}$) $\times T$ as T increases. We can observe from Figure 2 that for all three algorithms, CCE/CE-Gap($\bar{\pi}$) $\times T$ essentially become a constant for any reasonably large value of T. This indicates that our algorithms indeed converge at a rate of $\widetilde{O}(T^{-1})$ numerically. We also observe that OFTRL with stage-based value updates numerically converges faster than its incrementally-updated counterpart despite using the same learning rate, which advocates the use of stage-based value updates in Markov games.

6. Concluding Remarks

In this paper, we have studied the fast convergence of no-regret learning in full-information general-sum Markov games and answer the open question of Yang and Ma (2023) in the affirmative. We have shown that within T iterations, BM-OFTRL with smooth value updates finds an $\widetilde{O}(T^{-1})$-approximate CE, and OFTRL with stage-based value updates finds an $\widetilde{O}(T^{-1})$-approximate CCE, both of which match the best-known rates in normal-form games. For future research, it would be interesting to investigate whether OFTRL with smooth value updates attains the same $\widetilde{O}(T^{-1})$

convergence to CCE as has been observed in our simulations. Another direction is to improve our rates in terms of the dependence on H and $\log T$, or to prove any lower bounds for them.

References

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *ACM SIGACT Symposium on Theory of Computing*, pages 736–749, 2022a.

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. *Advances in Neural Information Processing Systems*, 35:3292–3304, 2022b.

Jalal Arabneydi and Aditya Mahajan. Reinforcement learning in decentralized stochastic control systems with partial history sharing. In *American Control Conference*, pages 5449–5456, 2015.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33, 2020.

Daniel S Bernstein, Christopher Amato, Eric A Hansen, and Shlomo Zilberstein. Policy iteration for decentralized control of Markov decision processes. *Journal of Artificial Intelligence Research*, 34:89–132, 2009.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

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 *Conference on Neural Information Processing Systems*, 2023.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Near-optimal policy optimization for correlated equilibrium in general-sum Markov games. *arXiv preprint arXiv:2401.15240*, 2024.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

Shicong Cen, Yuejie Chi, Simon Shaolei Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum Markov games. In *International Conference on Learning Representations*, 2022.

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

Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems*, 33:18990–18999, 2020.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a Nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 235–254, 2011.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34:27604–27616, 2021.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum Markov games. In *International Conference on Machine Learning*, pages 9343–9373, 2023.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Amy Greenwald and Amir Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Conference on Learning Theory*, pages 2–12, 2003.

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

Wassily Hoeffding and J Wolfowitz. Distinguishability of sets of distributions. *The Annals of Mathematical Statistics*, 29(3):700–718, 1958.

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *Conference on Learning Theory*, pages 2388–2422, 2021.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? *Advances in Neural Information Processing Systems*, 31, 2018.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–A simple, efficient, decentralized algorithm for multiagent RL. In *ICLR Workshop on Gamification and Multiagent Solutions*, 2022.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

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

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

Weichao Mao, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Information state embedding in partially observable cooperative multi-agent reinforcement learning. In *IEEE Conference on Decision and Control*, pages 6124–6131, 2020.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Başar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049, 2022.

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravishankar Iyer, and Tamer Başar. Multi-agent meta-reinforcement learning: Sharper convergence rates with task similarity. *Advances in Neural Information Processing Systems*, 36, 2023.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem Complexity and Method Efficiency in Optimization. 1983.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, pages 296–301, 1951.

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

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*, 2022.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation. *arXiv preprint arXiv:2302.06606*, 2023.

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 *Conference on Learning Theory*, pages 4259–4299, 2021.

Yuepeng Yang and Cong Ma. $O(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum Markov games. In *International Conference on Learning Representations*, 2023.

Runyu Zhang, Qinghua Liu, Huan Wang, Caiming Xiong, Na Li, and Yu Bai. Policy optimization for Markov games: Unified framework and faster convergence. In *Advances in Neural Information Processing Systems*, 2022.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33: 15198–15207, 2020.

Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 2736–2761, 2022.

Appendix A. Technical Lemmas

Lemma 3. *(Extension of Theorem 4.3 in [Anagnostides et al. (2022b)](#) to time-varying learning rates) In a no-regret learning problem as defined in Section [2](#), suppose that BM-OFTRL ([3](#)) is run with log-barrier regularization and a time-varying learning rate $\eta_t \leq \frac{1}{128\sqrt{|\mathcal{A}|}}, \forall t \in [T]$. Then, for any $T \geq 2$, the swap regret is bounded by*

$$\mathrm{SwapReg}^T \leq \frac{2|\mathcal{A}|^2 \log T}{\eta_T} + 4\sum_{t=1}^{T} \eta_t \left\| \boldsymbol{u}^t - \boldsymbol{u}^{t-1} \right\|_\infty^2 - \frac{1}{2048|\mathcal{A}|} \sum_{t=1}^{T-1} \frac{1}{\eta_t} \left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^t \right\|_1^2.$$

Proof sketch. The proof follows a similar procedure as that of Theorem 4.3 in [Anagnostides et al. (2022b)](#), except that we need to re-derive their Theorems B.1 and 3.1 under a time-varying learning rate. We skip the proof here as such an extension is straightforward. □

Lemma 4. *(Theorem 3.1 from [Daskalakis et al. (2021)](#)) In a normal-form game with N players and A_i actions for player $i \in [N]$, suppose that all the players run OFTRL for T steps with negative entropy regularization and a learning rate $\eta = \Theta(\frac{1}{N\log^4 T})$. Then, there exists a constant $C > 1$ such that the regret of player i satisfies*

$$\mathrm{Reg}_i^T \leq CN \log A_i \cdot \log^4 T.$$

Appendix B. Proofs for Section [3](#)

Lemma 1. (Recursion of best response CE value gaps) For any fixed $(h, t) \in [H] \times [T]$, we have

$$\delta_h^t \leq \sum_{j=1}^{t} \alpha_t^j \delta_{h+1}^j + \mathrm{SwapReg}_h^t.$$

Proof. For any fixed $i \in \mathcal{N}$ and $s \in \mathcal{S}$, we know from the definition of $\bar{\pi}_h^t$ from Algorithm [2](#) that

$$V_{h,i}^{\bar{\pi}_h^t}(s) = \sum_{j=1}^{t} \alpha_t^j \left\langle \pi_{h,i}^j(s,\cdot), \left[\left(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}] \right) \pi_{h,-i}^j \right](s,\cdot) \right\rangle. \tag{8}$$

For a fixed $\bar{\pi}_h^t$, we use $\phi_i^\star$ to denote the best response strategy modification that maximizes the value function starting from step h. In this case, we know from the definition of the value function that

$$V_{h,i}^{\phi_i^\star \diamond \bar{\pi}_h^t}(s) = \max_{\phi_{h,i}^s : \mathcal{A}_i \to \mathcal{A}_i} \sum_{j=1}^{t} \alpha_t^j \left\langle \phi_{h,i}^s \diamond \pi_{h,i}^j(s,\cdot), \left[\left(r_{h,i} + [P_h V_{h+1,i}^{\phi_i^\star \diamond \bar{\pi}_{h+1}^j}] \right) \pi_{h,-i}^j \right](s,\cdot) \right\rangle$$

$$= \max_{\phi_{h,i}^s} \sum_{j=1}^{t} \alpha_t^j \left\langle \phi_{h,i}^s \diamond \pi_{h,i}^j(s,\cdot), \left[\left(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}] + [P_h V_{h+1,i}^{\phi_i^\star \diamond \bar{\pi}_{h+1}^j}] - [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}] \right) \pi_{h,-i}^j \right](s,\cdot) \right\rangle$$

$$\leq \max_{\phi_{h,i}^s} \sum_{j=1}^{t} \alpha_t^j \left\langle \phi_{h,i}^s \diamond \pi_{h,i}^j(s,\cdot), \left[(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}]) \pi_{h,-i}^j \right](s,\cdot) \right\rangle + \sum_{j=1}^{t} \alpha_t^j \max_{s' \in \mathcal{S}} \left(V_{h+1,i}^{\phi_i^\star \diamond \bar{\pi}_{h+1}^j} - V_{h+1,i}^{\bar{\pi}_{h+1}^j} \right)(s').$$

Subtracting (8) from the above equation leads to:

$$V_{h,i}^{\phi_i^\star \diamond \bar{\pi}_h^t}(s) - V_{h,i}^{\bar{\pi}_h^t}(s) \leq \sum_{j=1}^t \alpha_t^j \max_{s' \in \mathcal{S}} \left(V_{h+1,i}^{\phi_i^\star \diamond \bar{\pi}_{h+1}^j}(s') - V_{h+1,i}^{\bar{\pi}_{h+1}^j}(s') \right)$$

$$+ \max_{\phi_{h,i}^s} \sum_{j=1}^t \alpha_t^j \left\langle \phi_{h,i}^s \diamond \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^j(s, \cdot), \left[\left(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}] \right) \pi_{h,-i}^j \right](s, \cdot) \right\rangle. \quad (9)$$

In the following, we will show that (9) is equal to $\mathrm{SwapReg}_{h,i}^t(s)$. It suffices to show that $Q_{h,i}^t(s, \boldsymbol{a}) = \left(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^t}] \right)(s, \boldsymbol{a}), \forall t \in [T], \boldsymbol{a} \in \mathcal{A}_{\mathrm{all}}$. We prove this claim by backward induction over $h \in [H]$. Notice that the claim trivially holds for $h = H$ as $Q_{H,i}^t(s, \boldsymbol{a}) = r_{H,i}(s, \boldsymbol{a}), \forall t \in [T], \boldsymbol{a} \in \mathcal{A}_{\mathrm{all}}$. Suppose that the claim holds for h; then, for step $h - 1$, we have that

$$Q_{h-1,i}^t(s, \boldsymbol{a}) = \sum_{j=1}^t \alpha_t^j \left(r_{h-1,i} + P_{h-1}[Q_{h,i}^j \pi_h^j] \right)(s, \boldsymbol{a})$$

$$= r_{h-1,i}(s, \boldsymbol{a}) + P_{h-1} \left[\sum_{j=1}^t \alpha_t^j Q_{h,i}^j \pi_h^j \right](s, \boldsymbol{a})$$

$$= r_{h-1,i}(s, \boldsymbol{a}) + P_{h-1} \left[\sum_{j=1}^t \alpha_t^j \left(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}] \right) \pi_h^j \right](s, \boldsymbol{a})$$

$$= r_{h-1,i}(s, \boldsymbol{a}) + \left[P_{h-1} V_{h,i}^{\bar{\pi}_h^t} \right](s, \boldsymbol{a}),$$

where the first step is by (4), the second step changes the order of summation, the third step uses the induction hypothesis, and the last step is due to (8). This completes the proof of $Q_{h,i}^t(s, \boldsymbol{a}) = \left(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^t}] \right)(s, \boldsymbol{a})$. Substituting it back to (9), we obtain that

$$V_{h,i}^{\phi_i^\star \diamond \bar{\pi}_h^t}(s) - V_{h,i}^{\bar{\pi}_h^t}(s) \leq \sum_{j=1}^t \alpha_t^j \max_{s' \in \mathcal{S}} \left(V_{h+1,i}^{\phi_i^\star \diamond \bar{\pi}_{h+1}^j}(s') - V_{h+1,i}^{\bar{\pi}_{h+1}^j}(s') \right) + \mathrm{SwapReg}_{h,i}^t(s).$$

Since the above inequality holds for any $i \in \mathcal{N}$ and $s \in \mathcal{S}$, and since $V_{h+1,i}^{\phi_i^\star \diamond \bar{\pi}_{h+1}^j}(s') \leq \max_{\phi_i} V_{h+1,i}^{\phi_i \diamond \bar{\pi}_{h+1}^j}(s')$ at step $h + 1$, we can conclude that

$$\delta_h^t \leq \sum_{j=1}^t \alpha_t^j \delta_{h+1}^j + \mathrm{SwapReg}_h^t,$$

This completes the proof of the recursive relationship of best response CE value gaps. $\qquad\square$

Lemma 2. (Per-state weighted swap regret bounds) For any $t \in [T], h \in [H], s \in \mathcal{S}$ and $i \in \mathcal{N}$, Algorithm 1 ensures that

$$\mathrm{SwapReg}_{h,i}^t(s) \leq \frac{4A_i^2 H \log t}{\eta t} + \frac{32\eta H^3 N^2}{t} + 8\eta N H^2 \sum_{j=2}^t \sum_{k \in \mathcal{N}, k \neq i} \alpha_t^j \left\| \pi_{h,k}^j(s, \cdot) - \pi_{h,k}^{j-1}(s, \cdot) \right\|_1^2$$

$$- \frac{1}{2048\eta A_i} \sum_{j=2}^t \alpha_t^{j-1} \left\| \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^{j-1}(s, \cdot) \right\|_1^2.$$

16

Consequently, if $\eta \leq \frac{1}{256NH\sqrt{HA_{\max}}}$, we further have

$$\sum_{i=1}^{N} \text{SwapReg}_{h,i}^{t}(s) \leq \frac{4NA_{\max}^2 H \log t}{\eta t} + \frac{32\eta NH^2(N^2+H)}{t}$$
$$- \frac{1}{2048\eta H} \sum_{i=1}^{N} \sum_{j=2}^{t} \frac{\alpha_t^j}{A_i} \left\| \pi_{h,i}^{j}(s,\cdot) - \pi_{h,i}^{j-1}(s,\cdot) \right\|_1^2.$$

Proof. At each fixed $(s,h) \in \mathcal{S} \times [H]$, the agents essentially face a no-swap-regret learning problem in a matrix game, where the payoff matrix of agent i is $Q_{h,i}^{t}(s,\cdot)$ at iteration t. We can apply the weighted swap regret bound (Lemma 3 in Appendix A) of OFTRL under the Blum-Mansour reduction in normal-form games to obtain:

$$\text{SwapReg}_{h,i}^{t}(s) = \max_{\phi_{h,i}^s : \mathcal{A}_i \to \mathcal{A}_i} \sum_{j=1}^{t} \alpha_t^j \left\langle \phi_{h,i}^s \diamond \pi_{h,i}^{j}(s,\cdot) - \pi_{h,i}^{j}(s,\cdot), [Q_{h,i}^{j}\pi_{h,-i}^{j}](s,\cdot) \right\rangle$$

$$= \alpha_t^1 \max_{\phi_{h,i}^s : \mathcal{A}_i \to \mathcal{A}_i} \sum_{j=1}^{t} \left\langle \phi_{h,i}^s \diamond \pi_{h,i}^{j}(s,\cdot) - \pi_{h,i}^{j}(s,\cdot), w_j [Q_{h,i}^{j}\pi_{h,-i}^{j}](s,\cdot) \right\rangle \quad (10)$$

$$\leq \frac{2A_i^2 \alpha_t \log t}{\eta} + 4\sum_{j=1}^{t} \frac{\eta\alpha_t^1}{w_j} \left\| w_j[Q_{h,i}^{j}\pi_{h,-i}^{j}](s,\cdot) - w_j[Q_{h,i}^{j-1}\pi_{h,-i}^{j-1}](s,\cdot) \right\|_\infty^2$$

$$- \frac{\alpha_t^1}{2048\eta A_i} \sum_{j=2}^{t} w_{j-1} \left\| \pi_{h,i}^{j}(s,\cdot) - \pi_{h,i}^{j-1}(s,\cdot) \right\|_1^2, \quad (11)$$

where (10) is due to the choice of the weights $w_j = \alpha_t^j / \alpha_t^1$. (11) uses Lemma 3 from Appendix A, by instantiating $\boldsymbol{u}^j(\cdot)$ in Lemma 3 as $w_j[Q_{h,i}^{j}\pi_{h,-i}^{j}](s,\cdot)$, the prediction $\boldsymbol{m}^t = w_j[Q_{h,i}^{j-1}\pi_{h,-i}^{j-1}](s,\cdot)$, and the learning rate $\eta_j = \eta/w_j$. To further upper bound the above equation, notice that

$$\sum_{j=1}^{t} \frac{\eta\alpha_t^1}{w_j} \left\| w_j[Q_{h,i}^{j}\pi_{h,-i}^{j}](s,\cdot) - w_j[Q_{h,i}^{j-1}\pi_{h,-i}^{j-1}](s,\cdot) \right\|_\infty^2$$

$$= \sum_{j=1}^{t} \eta\alpha_t^1 w_j \left\| \left([Q_{h,i}^{j}\pi_{h,-i}^{j}] - [Q_{h,i}^{j-1}\pi_{h,-i}^{j}] + [Q_{h,i}^{j-1}\pi_{h,-i}^{j}] - [Q_{h,i}^{j-1}\pi_{h,-i}^{j-1}]\right)(s,\cdot) \right\|_\infty^2$$

$$\leq 2\sum_{j=1}^{t} \eta\alpha_t^1 w_j \left(\left\| Q_{h,i}^{j}(s,\cdot) - Q_{h,i}^{j-1}(s,\cdot) \right\|_\infty^2 + H^2 \left\| \pi_{h,-i}^{j}(s,\cdot) - \pi_{h,-i}^{j-1}(s,\cdot) \right\|_1^2 \right)$$

$$\leq 2\sum_{j=1}^{t} \eta\alpha_t^1 w_j (\alpha_j)^2 H^2 + 2\sum_{j=1}^{t} \eta\alpha_t^1 w_j H^2 \left\| \pi_{h,-i}^{j}(s,\cdot) - \pi_{h,-i}^{j-1}(s,\cdot) \right\|_1^2, \quad (12)$$

where the second step uses the observation that $(a+b)^2 \leq 2a^2 + 2b^2$, the Hölder's inequality, and the fact that $\|Q_{h,i}^{j-1}\|_\infty \leq H$. The third step is due to our value update rule in Algorithm 1, which

17

yields

$$
\begin{aligned}
\left\| Q_{h,i}^j(s,\cdot) - Q_{h,i}^{j-1}(s,\cdot) \right\|_\infty &= \left\| -\alpha_j Q_{h,i}^{j-1}(s,\cdot) + \alpha_j \left(r_{h,i} + P_h[Q_{h+1,i}^j \pi_{h+1}^j] \right)(s,\cdot) \right\|_\infty \\
&\leq \alpha_j \max \left\{ \left\| Q_{h,i}^{j-1}(s,\cdot) \right\|_\infty, \left\| \left(r_{h,i} + P_h[Q_{h+1,i}^j \pi_{h+1}^j] \right)(s,\cdot) \right\|_\infty \right\} \\
&\leq \alpha_j H.
\end{aligned}
$$

To continue from (12), we apply the properties that $w_j = \alpha_t^j / \alpha_t^1$ and $\sum_{j=1}^t \alpha_t^j (\alpha_j)^2 \leq \sum_{j=1}^t (\alpha_j)^2 / t \leq (H+2)/t \leq 3H/t$ (see Lemma 6 in Yang and Ma (2023) for a proof) to obtain:

$$
\begin{aligned}
(12) =& 2 \sum_{j=1}^t \eta \alpha_t^1 w_j (\alpha_j)^2 H^2 + 2 \sum_{j=1}^t \eta \alpha_t^1 w_j H^2 \left\| \pi_{h,-i}^j(s,\cdot) - \pi_{h,-i}^{j-1}(s,\cdot) \right\|_1^2 \\
\leq& \frac{6\eta H^3}{t} + 2 \sum_{j=1}^t \eta \alpha_t^1 w_j H^2 \left\| \pi_{h,-i}^j(s,\cdot) - \pi_{h,-i}^{j-1}(s,\cdot) \right\|_1^2 \\
\leq& \frac{6\eta H^3}{t} + 2\eta(N-1)H^2 \sum_{j=1}^t \alpha_t^j \sum_{k \in \mathcal{N}, k \neq i} \left\| \pi_{h,k}^j(s,\cdot) - \pi_{h,k}^{j-1}(s,\cdot) \right\|_1^2. \quad (13)
\end{aligned}
$$

In the last step, we used that the total variation between two product distributions is bounded by the sum of the total variations of each marginal distribution (see e.g. (Hoeffding and Wolfowitz, 1958)):

$$
\begin{aligned}
\left\| \pi_{h,-i}^j(s,\cdot) - \pi_{h,-i}^{j-1}(s,\cdot) \right\|_1^2 &= \left(\sum_{\boldsymbol{a}_{-i} \in \mathcal{A}_{-i}} \left| \pi_{h,-i}^j(s, \boldsymbol{a}_{-i}) - \pi_{h,-i}^{j-1}(s, \boldsymbol{a}_{-i}) \right| \right)^2 \\
&= \left(\sum_{\boldsymbol{a}_{-i} \in \mathcal{A}_{-i}} \left| \prod_{k \neq i} \pi_{h,k}^j(s, a_k) - \prod_{k \neq i} \pi_{h,k}^{j-1}(s, a_k) \right| \right)^2 \\
&\leq \left(\sum_{k \neq i} \left\| \pi_{h,k}^j(s, a_k) - \pi_{h,k}^{j-1}(s, a_k) \right\|_1 \right)^2 \\
&\leq (N-1) \sum_{k \neq i} \left\| \pi_{h,k}^j(s, a_k) - \pi_{h,k}^{j-1}(s, a_k) \right\|_1^2,
\end{aligned}
$$

and the last step is by the Cauchy–Schwarz inequality. Substituting (13) back to (11) leads to

$$
\begin{aligned}
\text{SwapReg}_{h,i}^t(s) \leq & \frac{2A_i^2 \alpha_t \log t}{\eta} + 8\eta(N-1)H^2 \sum_{j=1}^t \alpha_t^j \sum_{k \in \mathcal{N}, k \neq i} \left\| \pi_{h,k}^j(s,\cdot) - \pi_{h,k}^{j-1}(s,\cdot) \right\|_1^2 \\
& + \frac{24\eta H^3}{t} - \frac{\alpha_t^1}{2048\eta A_i} \sum_{j=2}^t w_{j-1} \left\| \pi_{h,i}^j(s,\cdot) - \pi_{h,i}^{j-1}(s,\cdot) \right\|_1^2 \\
\leq & \frac{4A_i^2 H \log t}{\eta t} + 8\eta(N-1)H^2 \sum_{j=2}^t \alpha_t^j \sum_{k \in \mathcal{N}, k \neq i} \left\| \pi_{h,k}^j(s,\cdot) - \pi_{h,k}^{j-1}(s,\cdot) \right\|_1^2 \\
& + \frac{32\eta H^2(H+N^2)}{t} - \frac{1}{2048\eta A_i} \sum_{j=2}^t \alpha_t^{j-1} \left\| \pi_{h,i}^j(s,\cdot) - \pi_{h,i}^{j-1}(s,\cdot) \right\|_1^2, \quad (14)
\end{aligned}
$$

where the second inequality uses $\alpha_t = (H+1)/(H+t) \leq 2H/t$. This step also takes out the term for $j = 1$ and upper bounds it by

$$8\eta(N-1)H^2\alpha_t^1 \sum_{k \in \mathcal{N}, k \neq i} \left\| \pi_{h,k}^1(s, \cdot) - \pi_{h,k}^0(s, \cdot) \right\|_1^2 \leq \frac{32\eta(N-1)^2H^2}{t},$$

using the fact that $\alpha_t^1 \leq 1/t$ (Lemma 6 in Yang and Ma (2023)). This proves the first claim in the lemma. To further establish the second statement, we sum over (14) to obtain

$$\sum_{i=1}^N \mathrm{SwapReg}_{h,i}^t(s) \leq \frac{4NA_i^2 H \log t}{\eta t} + 8\eta(N-1)^2 H^2 \sum_{i=1}^N \sum_{j=2}^t \alpha_t^j \left\| \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^{j-1}(s, \cdot) \right\|_1^2$$

$$+ \frac{32\eta NH^2(H+N^2)}{t} - \frac{1}{2048\eta} \sum_{i=1}^N \sum_{j=2}^t \frac{\alpha_t^{j-1}}{A_i} \left\| \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^{j-1}(s, \cdot) \right\|_1^2$$

$$\leq \frac{4NA_i^2 H \log t}{\eta t} + \frac{32\eta NH^2(H+N^2)}{t}$$

$$+ \sum_{i=1}^N \sum_{j=2}^t \left(8\eta(N-1)^2 H^2 - \frac{1}{2048\eta HA_i} \right) \alpha_t^j \left\| \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^{j-1}(s, \cdot) \right\|_1^2$$

$$\leq \frac{4NA_i^2 H \log t}{\eta t} + \frac{32\eta NH^2(H+N^2)}{t}$$

$$- \frac{1}{2048\eta H} \sum_{i=1}^N \sum_{j=2}^t \frac{\alpha_t^j}{A_i} \left\| \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^{j-1}(s, \cdot) \right\|_1^2,$$

where the second step uses the fact that $\alpha_t^{j-1}/\alpha_t^j = (j-1)/(H+j-1) \geq 1/H$, and the last step is due to the condition that $\eta \leq \frac{1}{256NH\sqrt{HA_{\max}}}$. $\qquad\square$

Theorem 1. If Algorithm 1 is run on an N-player episodic Markov game for T iterations with a learning rate $\eta = \frac{1}{256NH\sqrt{HA_{\max}}}$, the output policy $\bar{\pi}$ satisfies:

$$\mathrm{CE\text{-}Gap}(\bar{\pi}) \leq \frac{6144NH^{\frac{7}{2}}A_{\max}^{\frac{5}{2}} \log T}{T}.$$

Proof. Using (7) from Lemma 2, we upper bound the second-order path lengths by

$$8\eta NH^2 \sum_{i=1}^N \sum_{j=2}^t \alpha_t^j \left\| \pi_{h,i}^j(s, \cdot) - \pi_{h,i}^{j-1}(s, \cdot) \right\|_1^2$$

$$\leq 8\eta NH^2 \cdot 2048\eta HA_{\max} \left(\frac{4NA_{\max}^2 H \log t}{\eta t} + \frac{32\eta NH^2(N^2+H)}{t} \right),$$

where we used the crucial fact that the swap regret is non-negative. Substituting the above equation back to (6) yields

$$\mathrm{SwapReg}_{h,i}^t(s) \leq \frac{4A_i^2 H \log t}{\eta t} + \frac{32\eta H^3 N^2}{t} + \frac{2^{16}\eta N^2 H^4 A_{\max}^3 \log t}{t} + \frac{2^{19}\eta^3 N^4 H^6}{t}$$

$$\leq \frac{2048NH^{\frac{5}{2}} A_{\max}^{\frac{5}{2}} \log t}{t}, \tag{15}$$

19

where the second step uses $\eta = \frac{1}{256NH\sqrt{HA_{\max}}}$. Since (15) holds for any $i \in \mathcal{N}$ and $s \in \mathcal{S}$, we can apply it back to the recursion of best response CE value gaps from Lemma 1 to obtain

$$\delta_h^t \le \sum_{j=1}^{t} \alpha_t^j \delta_{h+1}^j + \frac{2048NH^{\frac{5}{2}}A_{\max}^{\frac{5}{2}}\log t}{t}.$$

Starting from $\delta_{H+1}^t = 0$, we can show via backward induction that for any $(h,t) \in [H] \times [T]$,

$$\delta_h^t \le \frac{6144NA_{\max}^{\frac{5}{2}}(H-h+1)H^{\frac{5}{2}}\log t}{t},$$

where we applied Lemma 4 from Yang and Ma (2023) that $\sum_{j=1}^{t} \alpha_t^j/j \le (1+\frac{1}{H})\frac{1}{t}$. We conclude the proof of the theorem by referring to the property that CE-Gap$(\bar{\pi}) \le \delta_1^T$. $\qquad\square$

Appendix C. Supplementary Material for Section 4

C.1. Policy Output Algorithm

Algorithm 4 Policy $\bar{\pi}_h^t$ for stage-based OFTRL

Input: Policy trajectory $\{\pi_h^t\}_{h\in[H],t\in[T]}$ of Algorithm 3;
for *step* $h' \leftarrow h$ *to* H **do**
    Uniformly sample j from $\{t_{\tau(t)-1}^{\text{start}}, t_{\tau(t)-1}^{\text{start}}+1, \ldots, t_{\tau(t)-1}^{\text{end}}\}$;
    Play policy $\pi_{h'}^j$ for step h';
    Set $t \leftarrow j$.

C.2. Proofs

Lemma 5. *(Recursion of best response CCE value gaps) For any fixed* $(h,t) \in [H] \times [T]$*, let* $\tau = \tau(t)$ *denote the stage of* t*. Then, we have*

$$\zeta_h^t \le \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{start}}^{t_{\tau-1}^{end}} \zeta_{h+1}^j + \text{Reg}_h^{\tau-1}.$$

Proof. For any fixed $i \in \mathcal{N}$ and $s \in \mathcal{S}$, we know from the definition of $\bar{\pi}_h^t$ from Algorithm 4 that

$$V_{h,i}^{\bar{\pi}_h^t}(s) = \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{start}}^{t_{\tau-1}^{end}} \left\langle \pi_{h,i}^j(s,\cdot), \left[\left(r_{h,i} + [P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}] \right) \pi_{h,-i}^j \right](s,\cdot) \right\rangle. \tag{16}$$

From the definition of the best response value function,

$$
\begin{aligned}
V_{h,i}^{\dagger,\bar{\pi}_{h,-i}^t}(s) &= \max_{\pi_i^\dagger(s,\cdot)\in\Delta(\mathcal{A}_i)} \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \left\langle \pi_i^\dagger(s,\cdot), \left[\left(r_{h,i}+[P_h V_{h+1,i}^{\dagger,\bar{\pi}_{h+1,-i}^j}]\right)\pi_{h,-i}^j\right](s,\cdot)\right\rangle \\
&= \max_{\pi_i^\dagger} \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \left\langle \pi_i^\dagger(s,\cdot), \left[\left(r_{h,i}+[P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}]-[P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}]+[P_h V_{h+1,i}^{\dagger,\bar{\pi}_{h+1,-i}^j}]\right)\pi_{h,-i}^j\right](s,\cdot)\right\rangle \\
&\leq \max_{\pi_i^\dagger} \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \left\langle \pi_i^\dagger(s,\cdot), \left[\left(r_{h,i}+[P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}]\right)\pi_{h,-i}^j\right](s,\cdot)\right\rangle \\
&\quad + \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \max_{s'\in\mathcal{S}} \left(V_{h+1,i}^{\dagger,\bar{\pi}_{h+1,-i}^j}-V_{h+1,i}^{\bar{\pi}_{h+1}^j}\right)(s').
\end{aligned}
$$

Subtracting (16) from the above equation leads to:

$$
\begin{aligned}
V_{h,i}^{\dagger,\bar{\pi}_{h,-i}^t}(s)-V_{h,i}^{\bar{\pi}_h^t}(s) &\leq \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \max_{s'\in\mathcal{S}} \left(V_{h+1,i}^{\dagger,\bar{\pi}_{h+1,-i}^j}-V_{h+1,i}^{\bar{\pi}_{h+1}^j}\right)(s') \\
&\quad + \max_{\pi_i^\dagger} \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \left\langle \pi_i^\dagger(s,\cdot)-\pi_{h,i}^j(s,\cdot), \left[\left(r_{h,i}+[P_h V_{h+1,i}^{\bar{\pi}_{h+1}^j}]\right)\pi_{h,-i}^j\right](s,\cdot)\right\rangle. \quad (17)
\end{aligned}
$$

Using a similar inductive argument as in the proof of Lemma 1, we can show that the term in (17) is equal to $\text{Reg}_{h,i}^{\tau-1}(s)$, which leads to

$$
V_{h,i}^{\dagger,\bar{\pi}_{h,-i}^t}(s) - V_{h,i}^{\bar{\pi}_h^t}(s) \leq \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \max_{s'\in\mathcal{S}} \left(V_{h+1,i}^{\dagger,\bar{\pi}_{h+1,-i}^j}-V_{h+1,i}^{\bar{\pi}_{h+1}^j}\right)(s') + \text{Reg}_{h,i}^{\tau-1}(s).
$$

Since the above inequality holds for any $i\in\mathcal{N}$ and $s\in\mathcal{S}$, we can conclude that

$$
\zeta_h^t \leq \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \zeta_{h+1}^j + \text{Reg}_h^{\tau-1}.
$$

This completes the proof of the recursive relationship of best response CCE value gaps. $\qquad\square$

Theorem 2. If Algorithm 3 is run on an N-player episodic Markov game for T iterations with a learning rate $\eta_\tau = \Theta(\frac{1}{N\log^4 L_\tau})$ in each stage τ, the output policy $\bar{\pi}$ satisfies:

$$
\text{CCE-Gap}(\bar{\pi}) = O\left(\frac{NH^3\log A_{\max}\cdot\log^5 T}{T}\right).
$$

Proof. We introduce a few more notations before presenting the proof. Let $\tau(t)$ denote the index of the stage that iteration t belongs to. We denote by $\bar{\tau}$ the total number of stages, i.e., $\bar{\tau} := \tau(T)$. For any (τ, h, s), we define the per-state (average) regret for player $i \in \mathcal{N}$ in the τ-th stage of the corresponding matrix game as

$$\text{Reg}_{h,i}^{\tau}(s) := \max_{\pi_i^{\dagger}(s,\cdot) \in \Delta(\mathcal{A}_i)} \frac{1}{L_{\tau}} \sum_{j=t_{\tau}^{\text{start}}}^{t_{\tau}^{\text{end}}} \left\langle \pi_i^{\dagger}(s,\cdot) - \pi_{h,i}^j(s,\cdot), [Q_{h,i}^{\tau} \pi_{h,-i}^j](s,\cdot) \right\rangle,$$

$$\text{Reg}_h^{\tau} := \max_{i \in \mathcal{N}} \max_{s \in \mathcal{S}} \text{Reg}_{h,i}^{\tau}(s),$$

where $Q_{h,i}^{\tau}$ is player i's Q-function estimate at stage τ. For any $(h, t) \in [H] \times [T]$ and for the policy $\bar{\pi}_h^t$ as defined in Algorithm 4, we define the best response CCE value gap as

$$\zeta_h^t := \max_{i \in \mathcal{N}} \max_{s \in \mathcal{S}} \left(V_{h,i}^{\dagger, \bar{\pi}_{h,-i}^t}(s) - V_{h,i}^{\bar{\pi}_h^t}(s) \right).$$

By the definition of $\bar{\pi}$ and ζ_h^t, we have

$$\text{CCE-gap}(\bar{\pi}) = \max_{i \in \mathcal{N}} \left(V_{1,i}^{\dagger, \bar{\pi}^{-i}}(s_1) - V_{1,i}^{\bar{\pi}}(s_1) \right)$$

$$\leq \frac{1}{T} \sum_{t=1}^{T} \max_{i \in \mathcal{N}} \max_{s \in \mathcal{S}} \left(V_{1,i}^{\dagger, \bar{\pi}_1^t, -i}(s) - V_{1,i}^{\bar{\pi}_1^t}(s) \right) \leq \frac{1}{T} \sum_{t=1}^{T} \zeta_1^t. \tag{18}$$

We use Lemma 5 to establish the following recursive relationship of the best response CCE value gaps between two consecutive steps h and $h+1$:

$$\zeta_h^t \leq \frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \zeta_{h+1}^j + \text{Reg}_h^{\tau-1}. \tag{19}$$

Hence, upper bounding CCE-gap($\bar{\pi}$) breaks down to controlling the per-state regret in the corresponding matrix game for each $(\tau, s, h) \in [\bar{\tau}] \times \mathcal{S} \times [H]$. In our stage-based OFTRL, since the reward matrix $Q_{h,i}^{\tau}$ in each stage is fixed and $\text{Reg}_{h,i}^{\tau}(s)$ is the standard (average) regret, we can readily apply the individual regret bound of each player when running OFTRL in normal-form games (Daskalakis et al., 2021). Specifically, Theorem 3.1 from Daskalakis et al. (2021) (restated as Lemma 4 in our Appendix A) shows that with a learning rate $\eta_{\tau} = \Theta(\frac{1}{N \log^4 L_{\tau}})$, there exists a constant $C > 1$ such that for any $(i, \tau, s, h) \in \mathcal{N} \times [\bar{\tau}] \times \mathcal{S} \times [H]$,

$$\text{Reg}_{h,i}^{\tau}(s) \leq \frac{C N H \log A_i \cdot \log^4 L_{\tau}}{L_{\tau}}. \tag{20}$$

Notice that we multiplied the regret bound by H because Daskalakis et al. (2021) assumes the rewards to be from $[0, 1]$ but our rewards lie in $[0, H]$. According to the definition in Algorithm 4, the behavior of the policy $\bar{\pi}_h^t$ is unchanged for all t within the same stage τ as it always uniformly samples a time index from the previous stage and plays the corresponding history policy. Consequently, the value estimation error ζ_h^t does not change within a stage $\tau(t)$; that is, ζ_h^t takes the same value for all $t \in [t_{\tau}^{\text{start}}, t_{\tau}^{\text{end}}]$. We occasionally slightly abuse the notation and use ζ_h^{τ} to denote the estimation

22

Table 2: Reward matrices for Player 1.

s_0	b_0	b_1
a_0	0.8	0.2
a_1	0.0	1.0

s_1	b_0	b_1
a_0	1.0	0.2
a_1	0.5	0.8

Table 3: Reward matrices for Player 2.

s_0	b_0	b_1
a_0	0.2	1.0
a_1	0.5	0.0

s_1	b_0	b_1
a_0	0.5	1.0
a_1	1.0	0.2

error for a stage τ. This immediately implies that $\frac{1}{L_{\tau-1}} \sum_{j=t_{\tau-1}^{\text{start}}}^{t_{\tau-1}^{\text{end}}} \zeta_{h+1}^j = \zeta_{h+1}^{\tau-1}$. Substituting (20) and the above equation back to the recursion (19), we obtain that

$$
\begin{aligned}
\zeta_h^t &\leq \zeta_{h+1}^{\tau-1} + \frac{CNH \log A_{\max} \cdot \log^4 L_{\tau-1}}{L_{\tau-1}} \\
&\leq \sum_{h'=h}^{H} \frac{CNH \log A_{\max} \cdot \log^4 T}{L_{\tau-h'+h-1}} \qquad (21) \\
&\leq \frac{3CNH^2 \log A_{\max} \cdot \log^4 T}{L_\tau}, \qquad (22)
\end{aligned}
$$

where the second step is by applying the inequality recursively over h, and the last step holds because our choice of the stage lengths $L_{\tau+1} = \lfloor (1+1/H)L_\tau \rfloor$ implies that

$$
\frac{1}{L_{\tau-h'+h-1}} \leq \frac{1}{L_\tau} \left(1 + \frac{1}{H}\right)^{h'-h+1} \leq \frac{1}{L_\tau} \left(1 + \frac{1}{H}\right)^H \leq \frac{3}{L_\tau}.
$$

We then substitute (22) back to (18) and change the counting method to obtain

$$
\begin{aligned}
\text{CCE-gap}(\bar\pi) \leq \frac{1}{T} \sum_{t=1}^{T} \zeta_1^t &\leq \frac{1}{T} \sum_{\tau=1}^{\bar\tau} \sum_{j=t_\tau^{\text{start}}}^{t_\tau^{\text{end}}} \frac{3CNH^2 \log A_{\max} \cdot \log^4 T}{L_\tau} \\
&\leq \frac{3CN\bar\tau H^2 \log A_{\max} \cdot \log^4 T}{T}.
\end{aligned}
$$

It remains to bound the total number of stages $\bar\tau$. Since the lengths of the stages increase exponentially as $L_{\tau+1} = \lfloor (1+1/H)L_\tau \rfloor$ and the $\bar\tau$ stages sum up to T iterations, by taking the sum of a geometric series, it suffices to find a value of $\bar\tau$ such that $(1+1/H)^{\bar\tau} \geq T/H$. Using the Taylor series expansion, one can show that $(1 + \frac{1}{H})^H \geq e - \frac{e}{2H}$, and hence any $\bar\tau \geq \frac{H \log T}{\log(e/2)}$ satisfies the condition. This completes the proof of the theorem. □

Appendix D. Simulation Details

Our numerical studies are conducted on a simple general-sum Markov game with 2 players, 2 states $\mathcal{S} = \{s_0, s_1\}$ and $H = 2$ steps per episode. Each player has 2 candidate actions $\mathcal{A} = \{a_0, a_1\}$ and $\mathcal{B} = \{b_0, b_1\}$, respectively. The reward matrices for Player 1 and Player 2 at the two states are given in Tables 2 and 3, respectively. The state transition function is defined as follows: In both states s_0 and s_1, if the two players take matching actions (namely (a_0, b_0) or (a_1, b_1)), the system stays at the current state with probability 0.8, and transitions to the other state with probability 0.2. On the other hand, if the two players take opposite actions (namely (a_0, b_1) or (a_1, b_0)), the environment will stay

at the current state with probability 0.2, and will transition to the other state with probability 0.8. We choose a constant learning rate $\eta = 0.2$ for all the three algorithms. We have also experimented with other choices of the transition and reward functions and have observed similar behavior as shown in Figures 1 and 2.