Near-Optimal Policy Optimization for Correlated Equilibrium in General-Sum Markov Games

Yang Cai
Yale University

Haipeng Luo
University of Southern California

Chen-Yu Wei
University of Virginia

Weiqiang Zheng
Yale University

Abstract

We study policy optimization algorithms for computing correlated equilibria in multi-player general-sum Markov Games. Previous results achieve $\tilde{O}(T^{-1/2})$ convergence rate to a correlated equilibrium and an accelerated $\tilde{O}(T^{-3/4})$ convergence rate to the weaker notion of coarse correlated equilibrium. In this paper, we improve both results significantly by providing an uncoupled policy optimization algorithm that attains a near-optimal $\tilde{O}(T^{-1})$ convergence rate for computing a correlated equilibrium. Our algorithm is constructed by combining two main elements (i) smooth value updates and (ii) the *optimistic-follow-the-regularized-leader* algorithm with the log barrier regularizer.

1 Introduction

How does a multi-agent system evolve when each agent independently updates their policy based on their own utility? Can the system converge to an equilibrium, and if so, how quickly? These questions lie at the heart of game theory, economics, and learning theory, and have stimulated decades of research. For example, in normal-form games, it is well-known that when each agent employs a standard online learning algorithm with low external regret or low swap regret, the empirical distribution of their joint strategy profile converges to a coarse correlated equilibrium (CCE) or a correlated equilibrium (CE) respectively.

While $\sqrt{T}$ (external/swap) regret is minimax optimal after T interactions in the adversarial environment, it is possible to achieve strictly better regret in a normal-form game when each agent employs the same no-regret algorithm. For example, Syrgkanis et al. (2015) show that $T^{1/4}$ external regret can be achieved by the *optimistic* online mirror descent (OOMD) algorithm or the *optimistic* follow-the-regularized-leader (OFTRL) algorithm. Various improvements on this result have been proposed over the past few years, with the most recent one by Anagnostides et al. (2022a,b) achieving a near-optimal $\log(T)$ bound for both the external and swap regret. A direct corollary of this result is that when all agents employ the corresponding algorithm, the empirical distribution of their joint strategy profile converges to a CCE or CE, respectively, at a rate of $\tilde{O}(T^{-1})$.

However, achieving similar results for the more general setting of Markov games – the focus of this work, is much more challenging. Importantly, unlike the normal-form game setting, achieving $o(T)$ regret has been shown to be both statistically and computationally intractable for Markov games (Tian et al., 2021; Foster et al., 2023). This significant difference leads to considerably different algorithms for Markov games, the majority of which are aimed at finding an approximate equilibrium directly. We review this line of work in Section 1.1 and only point out here that, with an oracle access to the reward and transition function of the Markov game (see Remark 1), the state-of-the-art uncoupled learning dynamic converges to a CCE at a rate of $T^{-3/4}$ (Zhang et al., 2022) and to a CE at a rate of $T^{-1/2}$ (Jin et al., 2021; Song et al., 2021; Mao and Başar, 2023), both of which are substantially slower than the aforementioned $\tilde{O}(T^{-1})$ rate for normal-form games.

In this work, we close this gap by proposing an uncoupled policy optimization algorithm that converges to a CE (thus also to the weaker notion of CCE) at a near-optimal rate of $\log^2(T)/T = \tilde{O}(T^{-1})$, significantly improving existing results. Our algorithm builds upon the OFTRL framework with smooth value updates similar to Zhang et al. (2022), but importantly also incorporates the technique of using the log barrier as a regularizer from the recent work of Anagnostides et al. (2022b).

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume TBD. Copyright 2024 by the author(s).

1.1 Related Work

Learning in normal-form games The connection between no-regret learning algorithms and finding equilibria in games dates back to the seminal work of Freund and Schapire (1999), which shows that in a two-player zero-sum games, if both players have an external regret bound of R after T rounds, then their average strategy is an R/T-approximate Nash equilibrium (NE). For general-sum games, similar connections hold between the external regret and CCE, and also between the stronger notion of swap regret and CE (Stoltz and Lugosi, 2007; Blum and Mansour, 2007).

As mentioned, while in the worst case the best possible regret bounds are of order $\sqrt{T}$, the players could enjoy even lower regret when all of them employ the same algorithm, since this usually leads to an overall stable environment. Such results were pioneered by Daskalakis et al. (2011); Rakhlin and Sridharan (2013) for two-player zero-sum games, and extended by Syrgkanis et al. (2015) for general-sum games. Various improvements have been made over the past few years (Chen and Peng, 2020; Daskalakis et al., 2021; Anagnostides et al., 2022a,b).

Learning in Markov games Markov game (Shapley, 1953) is a general framework for modeling multi-agent sequential decision making problems. A line of earlier development has already focused on designing decentralized learning algorithms that offer better scalability (Littman, 1994; Littman et al., 2001; Bowling and Veloso, 2001), but the convergence guarantees are often only asymptotic. Inspired by recent advances in online optimization and the empirical success of multi-agent reinforcement learning through self-play (Silver et al., 2017; Vinyals et al., 2019; Bard et al., 2020), there is a surge of research trying to sharpen the theoretical guarantees for multi-agent learning in Markov games, especially in the decentralized setting (Bai et al., 2020; Wei et al., 2021; Sayin et al., 2021; Mao and Başar, 2023; Jin et al., 2021; Song et al., 2021; Kao et al., 2022; Leonardos et al., 2021; Ding et al., 2022; Erez et al., 2023; Cui et al., 2023; Wang et al., 2023). Below, we only highlight the most relevant ones.

As mentioned, Tian et al. (2021) show that no-regret learning is generally impossible when against arbitrary opponents. However, this does not preclude the possibility of enjoying low regret against Markov policies when all players employ the same algorithm. Indeed, Erez et al. (2023) design a policy optimization algorithm (also using techniques from Anagnostides et al. (2022b)) that achieves $\tilde{O}(T^{3/4})$ swap regret (a weaker notion of swap regret that concerns only Markov policy deviations) assuming the same oracle access to reward/transition as we do. This implies $\tilde{O}(T^{-1/4})$ convergence to a certain kind of CE that only allows Markov policy deviation, a notion weaker than ours.

Recent findings by Foster et al. (2023) indicate that the previously mentioned results by Erez et al. (2023) is unlikely to hold when general deviations are permitted. More explicitly, under standard computational complexity assumptions,[1] no polynomial-time algorithm can be no-regret in general-sum Markov games when executed independently by all players, even if the algorithm designer knows the game.

Due to such impossibility results, most algorithms directly aim at finding CCE/CE without considering the regret of the players. Specifically, the certified policy output by the V-learning algorithm (Jin et al., 2021; Song et al., 2021; Mao and Başar, 2023) has been proven to converge to a CCE/CE at a rate of $T^{-1/2}$, which is optimal when the players need to learn all the game parameters through interactions with the environment. In the simpler scenario where an oracle access to reward/transition is given, the best currently known rate is $T^{-3/4}$ for finding a CCE (Zhang et al., 2022), and $T^{-1/2}$ for finding a CE (still by V-learning). In this work, we improve both rates to $\log^2(T)/T$.

2 Preliminaries

For a positive integer n, we denote the set $\{1, 2, \dots, n\}$ as $[n]$. For any set $\mathcal{A}$, the probability simplex over $\mathcal{A}$ is $\Delta_{\mathcal{A}} := \{x \in \mathbb{R}^{|\mathcal{A}|} : \sum_{a\in\mathcal{A}} x[a] = 1, x[a] \geq 0, \forall a \in \mathcal{A}\}$.

Multi-player General-Sum Markov Games In this paper, we focus on finite-horizon n-player general-sum Markov games denoted as $\mathcal{M}(H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[n]}, \mathbb{P}, \{r_i\}_{i\in[n]})$, where H is the length of the horizon; $\mathcal{S}$ is the set of states with size $|\mathcal{S}| = S$; A_i is the the action set of player i with size $|\mathcal{A}_i| = A_i$ and we denote a joint action profile of all players as $\boldsymbol{a} = (a_1, a_2, \dots, a_n) \in \Pi_{i=1}^n \mathcal{A}_i$; $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ is the transition probabilities where $\mathbb{P}_h(s' \mid s, \boldsymbol{a})$ specifies the probabilities of transition to state s' in step $h+1$ if players take the joint action $\boldsymbol{a}$ at s in step h; $r_i = \{r_{i,h}\}_{h\in[H]}$ are the reward function for player i where $r_{i,h}(s, \boldsymbol{a}) \in [0, 1]$ is the reward for player i when players take the joint action $\boldsymbol{a}$ at s in step h. In each episode, we assume the game starts at s_1 without loss of generality. In each step $h \in [H]$, each player observes the current state s_h and chooses an action $a_{i,h} \in \mathcal{A}_i$, then each player receives reward $r_{i,h}(s_h, \boldsymbol{a}_h)$ and the game transits to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, \boldsymbol{a}_h)$. The episode ends after H steps.

[1] This is based on the assumption that PPAD-hard problems are not solvable in polynomial time.

Policies and Value Functions A (random) policy π_i for player i is a collection of H maps $\{\pi_{i,h} : \Omega \times (\mathcal{S} \times \Pi_{i=1}^n \mathcal{A}_i)^{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}_i}\}_{h\in[H]}$ where $\pi_{i,h}$ maps a random sample ω from a probability space, a history of length $h-1$, and the current state to a probability distribution (mixed strategy) over $\mathcal{A}_i$. To execute policy π_i, player i samples ω at the beginning of the episode, then at each step h, supposing the history is $\tau_h := (s_1, \boldsymbol{a}, \ldots, s_{j-1}, \boldsymbol{a}_{h-1})$, player i chooses action $a_{i,h} \sim \pi_{i,h}(\cdot \mid \omega, \tau_h, s_h)$. We note that ω is shared across all steps $h \in [H]$. A *Markov policy* for player i is collection of H history independent maps $\pi_i = \{\pi_{i,h} : \mathcal{S} \to \Delta_{\mathcal{A}_i}\}$, where $\pi_{i,h}(a \mid s_h)$ specifies the probability of taking action $a \in \mathcal{A}_i$ at (h, s_h).

A joint policy π is a set of policies denoted as $\pi = \pi_1 \odot \pi_2 \odot \ldots \odot \pi_n$ where the same random sample ω is shared among all players. We denote $\pi_{-i} := \pi_1 \odot \ldots \pi_{i-1} \odot \pi_{i+1} \odot \ldots \odot \pi_n$ the joint policy that excludes player i. When the random sample has a special form $\omega = (\omega_1, \ldots, \omega_n)$ and for each $i \in [n]$, π_i only uses the randomness in ω_i that is independent of $\{\omega_j\}_{j\neq i}$, then the joint policy is a *product policy* and we denote it as $\pi = \pi_1 \times \pi_2 \times \ldots \times \pi_n$. We denote $\pi_h(\boldsymbol{a}|s)$ the probability of joint action $\boldsymbol{a}$ at state s. The value function $V_{i,h}^\pi : \mathcal{S} \to \mathbb{R}$ specifies the expected reward for player i if all players follow policy π:

$$V_{i,h}^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, \boldsymbol{a}_{h'}) \mid s_h = s\right].$$

The goal of player i is to maximize their own value function $V_{1,h}^\pi(s_1)$. The Q function at step h is defined as

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

Strategy Modification and Correlated Equilibrium A strategy modification ϕ_i for player i is a collection of maps $\phi_i = \{\phi_{i,h} : (\mathcal{S} \times \Pi_{i=1}^n \mathcal{A}_i)^{h-1} \times \mathcal{S} \times \Delta_{\mathcal{A}_i} \to \Delta_{\mathcal{A}_i}\}$ such that given history τ_h and state s_h, each map $\phi_{i,h}(\tau_h, s_h, \cdot) : \Delta_{\mathcal{A}_i} \to \Delta_{\mathcal{A}_i}$ is a linear transformation.[2] For any policy π_i, the modified policy denoted as $\phi_i \diamond \pi_i$ changes the strategy $\pi_{i,h}(\omega, \tau_h, s_h)$ under random sample ω and history τ_h to another strategy $\phi_{i,h}(\tau_h, s_h, \pi_{i,h}(\omega, \tau_h, s_h))$.

A correlated equilibrium is a joint policy where no player can increase their value by any strategy modification. Formally, it is defined as

Definition 1 (Correlated Equilibrium)**.** *A joint policy π is a correlated equilibrium (CE) if* $\max_{i\in[n]} \max_{\phi_i} V_{i,1}^{(\phi_i \diamond \pi_i)\odot \pi_{-i}}(s_1) - V_{i,1}^\pi(s_1) \leq 0$. *A joint policy π is an ϵ-approximate CE if* $\mathrm{CEGap}(\pi) := \max_{i\in[n]} \max_{\phi_i} V_{i,1}^{(\phi_i \diamond \pi_i)\odot \pi_{-i}}(s_1) - V_{i,1}^\pi(s_1) \leq \epsilon$.

A coarse correlated equilibrium is a joint policy where no player can increase their value by playing any other independent policy. Formally, it is defined as

Definition 2 (Coarse Correlated Equilibrium)**.** *A joint policy π is an ϵ-approximate coarse correlated equilibrium if* $\max_{i\in[n]} \max_{\pi_i'} V_{i,1}^{\pi_i' \times \pi_{-i}}(s_1) - V_{i,1}^\pi(s_1) \leq \ \epsilon$.

We remark that by definition a CE is also a CCE. In the rest of the paper, we focus on CE only, but the same results apply to CCE clearly.

Additional Notations Define $A_{\max} = \max_{i\in[n]} A_i$. For any value function $V : \mathcal{S} \to \mathbb{R}$, we define $[\mathbb{P}_h V](s, \boldsymbol{a}) := \mathbb{E}_{s'\sim \mathbb{P}_h(s,\boldsymbol{a})} V(s')$. For any Markov policy $\pi_h(\cdot \mid s)$ and any Q function $Q_{i,h}(\cdot,\cdot) : \mathcal{S} \times \Pi_{j=1}^n \mathcal{A}_j \to \mathbb{R}$, we denote $[Q_{i,h}\pi_h](s) := \langle Q_{i,h}(s,\cdot), \pi_h(\cdot \mid s)\rangle$. Similarly, for any joint policy $\pi_{-i,h}(\cdot \mid s)$ that excludes player i, we denote $[Q_{i,h}\pi_{-i,h}](s, a_i) := \langle Q_{i,h}(s, a_i, \cdot), \pi_{-i,h}(\cdot \mid s)\rangle$.

2.1 Online Learning and Regret

In a (linear) online learning setting, at each day $t \in \mathbb{N}$, the learner chooses a strategy x^t from a compact and convex set $\mathcal{X} \subseteq \mathbb{R}^d$ while the adversary picks a reward vector $u^t \in \mathbb{R}^d$. Then the learner gets reward $\langle u^t, x^t\rangle$ and the reward vector u^t as feedback. The goal of an online learning algorithm is to minimize *regret*, or more generally, Φ*-regret*. For a set of strategy modifications $\Phi = \{\phi : \mathcal{X} \to \mathcal{X}\}$, the Φ-regret of an algorithm $\mathfrak{R}$ over a time horizon T is defined as

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

An algorithm is *no Φ-regret* if its Φ-regret is sublinear in T. The *(external) regret* denoted as Reg^T is Φ-regret when Φ includes only constant transformations. The *swap regret* denoted as $\mathrm{SwapReg}^T$ is Φ-regret when Φ includes all possible linear transformations. The swap regret is non-negative since we can choose the identity transformation such that $\phi(x) = x$ for all $x \in \mathcal{X}$.

3 Algorithm and Main Results

In this section, we present a policy optimization algorithm (Algorithm 1) for learning correlated equilibrium in multi-player general-sum Markov games. Algorithm 1 is a single-loop algorithm where on each

[2] On the other hand, the set of strategy modifications studied in (Erez et al., 2023) is $\{\phi_i : \mathcal{S} \times \mathcal{A}_i \to \mathcal{A}_i\}$, which is a strict subset of ours and thus induces a weaker notion of correlated equilibrium.

Algorithm 1 Policy optimization in Markov games with V value update

```
Require: step size $\eta > 0$, weights $\{\alpha_t\}$ and $\{w_t\}$, an online learning algorithm $\mathfrak{R}$.
1: Initialize: For all $(i, h, s)$, initialize $V_{i,h}^0(s) = H - h + 1$, $\mathfrak{R}_{i,h,s}$ as an instance of $\mathfrak{R}$ over decision set $\Delta_{\mathcal{A}_i}$, and $\pi_{i,h}^0(\cdot \mid s)$ as $\mathfrak{R}_{i,h,s}$'s initial output
2: for $t = 1, 2, \ldots, T$ do
3:     for all $(i, s, h)$ do
4:         Forward the utility vector $u_{i,h,s}^{t-1} := \frac{w_{t-1}}{H}[(r_h + \mathbb{P}_h V_{i,h+1}^{t-1})\pi_{-i,h}^{t-1}](s, \cdot)$ to $\mathfrak{R}_{i,h,s}$.
5:         Update $\pi_{i,h}^t(\cdot \mid s)$ according to $\mathfrak{R}_{i,h,s}$.
6:     end for
7:     for all $(i, s, \boldsymbol{a}) \in [n] \times \mathcal{S} \times \mathcal{A}$, from $h = H$ to 1:
           $V_{i,h}^t(s) \leftarrow (1 - \alpha_t)V_{i,h}^{t-1}(s) + \alpha_t\big[(r_h + \mathbb{P}_h V_{i,h+1}^t)\pi_h^t\big](s)$.
8: end for
Output $\widehat{\pi}^T = \widehat{\pi}_1^T$ as defined in Algorithm 2.
```

Algorithm 2 Executing Policy $\widehat{\pi}_h^t$

```
Require: Product policies $\pi_{h'}^{t'}(\cdot \mid s') = \Pi_{i=1}^n \pi_{i,h'}^{t'}(\cdot \mid s')$ for all $(h', s', t') \in [H] \times \mathcal{S} \times [T]$.
1: Sample $j \in [t]$ with probability $\Pr[j = i] = \alpha_j^i$ (see Equation (2) for definition).
2: Play policy $\pi_h^j$ at the $h$-th step of the game.
3: Play policy $\widehat{\pi}_{h+1}^j$ for step $h + 1$.
```

step-state pair $(h, s) \in [H] \times [\mathcal{S}]$, each player employs a no-regret algorithm over its own action set following the online learning protocol described in Section 2.1 with some reward vectors carefully constructed from a smooth update. We explain both the value update and the policy update below.

3.1 Value Update

Each player maintains V value function $V_{i,h}^t$ and conducts *smooth value update* with the following learning rates (Line 7 of Algorithm 1):

$$\alpha_t = \frac{H + 1}{H + t}. \tag{1}$$

The choice of $\alpha_t = O(\frac{1}{t})$ is proposed by Jin et al. (2018) and adopted in many subsequent works (Jin et al., 2021; Wei et al., 2021; Zhang et al., 2022; Yang and Ma, 2023). This choice ensures conservative updates of value functions and hence stabilizes the update of policies. We also define a group of auxiliary weights:

$$\alpha_t^t = \alpha_t, \quad \alpha_t^i = \alpha_i \Pi_{j=i+1}^t (1 - \alpha_j), \forall i \leq t - 1, \tag{2}$$

and

$$w_0 = w_1, \quad w_t = \frac{\alpha_t^t}{\alpha_t^1}, \quad \forall t \geq 1. \tag{3}$$

After $T \geq 1$ episodes, Algorithm 1 outputs a joint policy $\widehat{\pi}^T$ as defined in Algorithm 2. The output policy is not a Markov policy and is defined recursively. Specifically, at each step h, the policy $\widehat{\pi}_h^t$ randomly selects a product policy from $\{\pi_h^j\}_{j \in [t]}$ with probability $\{\alpha_t^j\}_{j \in [t]}$ and plays policy $\widehat{\pi}_{h+1}^j$ onward.

Remark 1. *Algorithm 1 is an adaptation of (Zhang et al., 2022, Algorithm 12), a policy optimization algorithm originally designed for learning the coarse correlated equilibrium. The original algorithm performs Q value update, whereas our adaptation focuses on V value update. An equivalent version of Algorithm 1 that employs Q value update is presented in Algorithm 3, with its equivalence proven in Proposition 1.*

The main distinction between the two algorithms lies in their function sizes. The $Q(s, \boldsymbol{a})$ function has a size of $S \cdot \Pi_{i=1}^n A_i$, which grows exponentially with the number of agents, leading to the so-called curse of multi-agents. In contrast, the $V(s)$ function is significantly more compact with a size of S, effectively bypassing the curse of multi-agents (Jin et al., 2021).

Furthermore, Algorithm 1 offers a notable advantage: it supports a decentralized *implementation. This means each player does not need explicit knowledge of other players' policies. The update steps in Algorithm 1 only require $(r_h + \mathbb{P}_h V_i)\pi_{-i,h}$ for any value function V_i and the policies of other players $\pi_{-i,h}$. This can be efficiently computed with access to:*

1. *A reward oracle that provides the expected reward vector for player i based on the policies of other players $\pi_{-i,h}(\cdot \mid s)$ at (h, s).*

2. *A transition oracle that offers the distribution of s_{h+1} based on player i's action a_i and the policies of other players $\pi_{-i,h}(\cdot \mid s)$ at (h, s).*

While we directly assume access to such oracles, both of them can be approximately implemented within $\varepsilon > 0$ error using $\mathrm{poly}(n, A_{\max}, S, H, 1/\varepsilon)$ *samples.*

Given the equivalence between Algorithm 1 and Algorithm 3, any guarantee for Algorithm 3 also holds for Algorithm 1. We will thus focus on Algorithm 3 in the rest of the paper.

Bounding Correlated Equilibrium Gap by Per-State Regret We first show a general result that the output policy $\widehat{\pi}^T$ of Algorithm 3 is an approximate correlated equilibrium as long as each player has

Algorithm 3 Policy optimization in Markov games with Q value update (Zhang et al., 2022)

```
Require: step size $\eta > 0$, weights $\{\alpha_t\}$ and $\{w_t\}$, an online learning algorithm $\mathfrak{R}$
Initialize: For all $(i, h, s)$, initialize $Q^0_{i,h}(s, \boldsymbol{a}) = H - h + 1$, $\mathfrak{R}_{i,h,s}$ as an instance of $\mathfrak{R}$ over decision set $\Delta_{\mathcal{A}_i}$, and $\pi^0_{i,h}(\cdot \mid s)$ as $\mathfrak{R}_{i,h,s}$'s initial output
for $t = 1, 2, \ldots, T$ do
    for all $(i, s, h)$ do
        Forward the utility vector $u^{t-1}_{i,h,s} \leftarrow \frac{w_{t-1}}{H} Q^{t-1}_{i,h} \pi^{t-1}_{-i,h}(s, \cdot)$ to $\mathfrak{R}_{i,h,s}$.
        Update $\pi^t_{i,h}(\cdot \mid s)$ according to $\mathfrak{R}_{i,h,s}$.
    end for
    for all $(i, s, \boldsymbol{a}) \in [n] \times \mathcal{S} \times \mathcal{A}$, from $h = H$ to 1:
        $Q^t_{i,h}(s, \boldsymbol{a}) \leftarrow (1 - \alpha_t) Q^{t-1}_{i,h}(s, \boldsymbol{a}) + \alpha_t (r_h + P_h[Q^t_{i,h+1} \pi^t_{h+1}])(s, \boldsymbol{a})$.
end for
Output $\widehat{\pi}^T = \widehat{\pi}^T_1$ as defined in Algorithm 2.
```

low *per-state weighted swap regret*. Formally, we define the per-state weighted swap regret (per-state regret for short) up to time $t \geq 1$ with respect to weights $\{\alpha^i_t\}_{i \in [t]}$ as $\text{reg}^t_{i,h}(s) :=$

$$\max_{\phi_i} \sum_{j=1}^{t} \alpha^j_t \left\langle Q^j_{i,h}(s, \cdot), ((\phi_i \diamond \pi^j_{i,h}) \odot \pi^j_{-i,h})(\cdot \mid s) - \pi^j_h(\cdot \mid s) \right\rangle.$$

We also define reg^t_h as the maximum weighted regret over all players and all states:

$$\text{reg}^t_h := \max_{s \in \mathcal{S}} \max_{i \in [n]} \text{reg}^t_{i,h}(s). \tag{4}$$

Theorem 1. *Suppose that the per-state regret has upper bounds $\text{reg}^t_h \leq \overline{\text{reg}}^t_h$ for all $(h, t) \in [H] \times [T]$ where $\overline{\text{reg}}^t_h$ is non-increasing in t: $\overline{\text{reg}}^t_h \geq \overline{\text{reg}}^{t+1}_h$. Then the output policy of Algorithm 3 satisfies*

$$\text{CEGap}(\widehat{\pi}^T) \leq 2H \cdot \frac{1}{T} \sum_{t=1}^{T} \max_{h \in [H]} \overline{\text{reg}}^t_h.$$

for all $T \geq 2$.

Theorem 1 states that $\text{CEGap}(\widehat{\pi}^T)$ can be bounded by the average weighted regret $O(\frac{1}{T} \sum_{t=1}^{T} \max_{h \in [H]} \overline{\text{reg}}^t_h)$. Thus, for any algorithm $\mathfrak{R}$ chosen in the policy update step, as long as the weighted average regret is sublinear, the output policy is an approximate correlated equilibrium. However, we emphasize that minimizing weighted swap regret $\overline{\text{reg}}^t_h$ with respect to $\{\alpha^i_t\}_{i \in [t]}$ requires careful design and analysis of the algorithm.

Proof Overview For $h \in [H]$, we define the reward difference between policy $\widehat{\pi}^t_h$ and a best strategy modification over any player $i \in [m]$ and state $s \in \mathcal{S}$ as:

$$\delta^t_h := \max_{i \in [n]} \max_{s \in \mathcal{S}} \left(\max_{\phi_i} V^{(\phi_i \diamond \widehat{\pi}^t_{i,h}) \odot \widehat{\pi}^t_{-i,h}}_{i,h}(s) - V^{\widehat{\pi}^t_h}_{i,h}(s) \right).$$

In Lemma 6, we establish bounds on δ^t_h using weighted regret such that

$$\delta^t_h \leq \sum_{j=1}^{t} \alpha^j_t \delta^j_{h+1} + \text{reg}^t_h.$$

Then Theorem 1 follows by applying the above inequality recursively to bound $\text{CEGap}(\widehat{\pi}^T) = \delta^T_1$.

3.2 Policy Update

We now turn to the design of the policy update, with the goal of minimizing the weighted swap regret $\overline{\text{reg}}^t_h$. Each player i maintains a Markov policy $\pi^t_{i,h}(\cdot \mid s)$ for every pair step h and step s. During each episode $t \in [T]$, player i updates $\pi^t_{i,h}$ using an online learning algorithm $\mathfrak{R}$. Previous works such as (Zhang et al., 2022; Yang and Ma, 2023) adopted the optimistic follow-the-regularized-Leader (OFTRL) algorithm (Syrgkanis et al., 2015) with entropy regularization, which is a no external regret algorithm. However, inspired by recent breakthrough in normal-form games (Anagnostides et al., 2022b), we select $\mathfrak{R}$ to be a specific no swap regret algorithm as outlined in Algorithm 4. Specifically, Algorithm 4 (1) uses the template introduced by Blum and Mansour (2007), which constructs a no swap regret algorithm $\mathfrak{R}_{swap}$ from several external regret minimizers $\mathfrak{R}_a$ for each action $a \in \mathcal{A}_i$; (2) employ *weighted* OFTRL with log barrier regularization for each external regret minimizer $\mathfrak{R}_a$. It is has been shown that with *constant* step size, OFTRL with log barrier regularization guarantees $O(\log T)$ individual swap regret in general-sum games (Anagnostides et al., 2022b). We extend their analysis to the more challenging Markov games with *decreasing* step size and provide bounds for *weighted swap regret*. A detailed discussion and analysis of Algorithm 4 are presented in Section 4.1.

3.3 Main Results

We present our main result on the convergence of Algorithm 3 to correlated equilibrium in multi-player general-sum Markov games.

Theorem 2. *For an n-player general-sum Markov game and any $T \geq 2$, when $\mathfrak{R}$ = Algorithm 4 with step size $\eta = \frac{1}{128 n \sqrt{H} A_{\max}}$, the output policy $\widehat{\pi}^T$ of ei-*

Algorithm 4 BM-OFTRL-Log-Bar

```
Require: Action set $\mathcal{A}$, step size $\eta$, weights $\{w_t\}$.
Initialization: Initialize $\mathfrak{R}_a$ as an instance of OFTRL-LogBar for each $a \in \mathcal{A}$.
for $t = 1, 2, \ldots, T$ do
    Get $x_a^t$ from $\mathfrak{R}_a$ for all $a \in \mathcal{A}$; Construct a (row) stochastic matrix $M^t \in \mathbb{S}^{|\mathcal{A}|\times|\mathcal{A}|}$ where the row that corresponds to $a \in \mathcal{A}$ is equal to $x_a^t$; Output strategy $x^t \in \Delta_{\mathcal{A}}$ so that $M^t x^t = x^t$.
    Get reward vector $u^t$; Forward $u_a^t := x^t[a] \cdot u^t$ to $\mathfrak{R}_a$ for each $a \in \mathcal{A}$.
end for
```

ther Algorithm 1 or Algorithm 3 satisfies

$$\mathrm{CEGap}(\widehat{\pi}^T) \leq 8192 H^{3.5} n A_{\max}^3 \cdot \frac{(\log T)^2}{T}.$$

We remark again that the previous best rate for finding CE is $\tilde{O}(T^{-\frac{1}{2}})$ achieved by the V-learning algorithm (Jin et al., 2021; Song et al., 2021; Mao and Başar, 2023), and Zhang et al. (2022) provides a faster convergence rate of $\tilde{O}(T^{-\frac{3}{4}})$ to the weaker notion of CCE. Theorem 2 improves both results significantly and for the first time, shows a near-optimal $\tilde{O}(T^{-1})$ convergence rates to CE/CCE in multi-player general-sum Markov games using a single loop and uncoupled policy optimization algorithm.

4 Proof of the Main Result

In this section, we provide a sketch of our analysis along with more explanation on the algorithm design. We first recall that Algorithm 4 applies the template by Blum and Mansour (2007) that constructs a swap regret minimizer $\mathfrak{R}_{swap}$ from a set of external regret minimizers $\{\mathfrak{R}_a\}_{a\in\mathcal{A}}$, one for each action $a \in \mathcal{A}$. The resulting algorithm $\mathfrak{R}_{swap}$ ensures $\mathrm{SwapReg}^T = \sum_{a\in\mathcal{A}} \mathrm{Reg}_a^T$ (Blum and Mansour, 2007).

4.1 Optimistic Follow the Regularized Leader with Log Barrier Regularization

An important component of Algorithm 4 is the optimistic follow the regularized leader algorithm with variable step size $\{\eta_t\}$. The optimistic follow the regularized leader (OFTRL) algorithm (Syrgkanis et al., 2015) over strategy set $\mathcal{X}$ and with a regularizer $\mathcal{R} : \mathcal{X} \to \mathbb{R}$ is defined as follows: $x^0 := \mathrm{argmin}_{x\in\mathcal{X}} \mathcal{R}(x)$ and for $t \geq 1$, the algorithm updates x^t using step size $\eta_t > 0$

$$x^t = \operatorname*{argmax}_{x\in\mathcal{X}} \left\{ \eta_t \left\langle x, m^t + \sum_{\tau=1}^{t-1} u^\tau \right\rangle - \mathcal{R}(x) \right\} \tag{OFTRL}$$

Previous works (Zhang et al., 2022; Yang and Ma, 2023) choose $\mathcal{R}$ to be a strongly convex function such as the entropy regularization. Here we follow (Anagnostides et al., 2022b) and let $\mathcal{R}$ be a *self-concordant barrier*. We first extend the RVU-bound established in (Anagnostides et al., 2022b) for OFTRL with a *constant* step size to the case of *variable* step sizes.[3] Before stating the result, we first introduce some notations. We assume $\mathcal{X}$ has a nonempty interior $\mathrm{int}(\mathcal{X})$. We say $\mathcal{R}$ is non-degenerate if its Hessian $\nabla^2\mathcal{R}(x)$ is positive definite for all $x \in \mathrm{int}(\mathcal{X})$. For any vector $u \in \mathbb{R}^d$, the primal *local norm* with respect to $x \in \mathrm{int}(\mathcal{X})$ is defined as $\|u\|_x := \sqrt{u^\top \nabla^2\mathcal{R}(x) u}$ and the dual norm is defined as $\|u\|_{*,x} := \sqrt{u^\top (\nabla^2\mathcal{R}(x))^{-1} u}$ when $\mathcal{R}$ is non-degenerate. We also use g^t to denote the sequence produced by *Be-the-Leader* (BTL) algorithm.

Theorem 3 (RVU for Self-Concordant Barrier with decreasing step size)**.** *Suppose that $\mathcal{R}$ is a non-degenerate self-concordant barrier function for $\mathrm{int}(\mathcal{X})$ and let $\eta_t > 0$ be such that $\eta_t\|u^t - m^t\|_{*,x^t} \leq \frac{1}{2}$ and $\|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{*,g^{t-1}} \leq \frac{1}{2}$ for all $t \in [T]$. Then, the regret of OFTRL with respect to any $x^* \in \mathrm{int}(\mathcal{X})$ and under any sequence of utilities $u^1, \ldots, u^T$ can be bounded as*

$$\mathrm{Reg}^T(x^*) \leq \frac{R(x^*)}{\eta_T} + 2\sum_{t=1}^T \eta_t \|u^t - m^t\|_{*,x^t}^2 - \sum_{t=1}^T \left(\frac{1}{4\eta_t} \|x^t - g^t\|_{x^t}^2 + \frac{1}{4\eta_{t-1}} \|x^t - g^{t-1}\|_{g^{t-1}}^2 \right).$$

Log Barrier Regularization Now we describe the implementation of OFTRL in Algorithm 4. We choose $\mathcal{R}$ to be the *log barrier* over the simplex $\mathcal{X} = \Delta^d$ defined as $\mathcal{R}(x) = -\sum_{r=1}^d \log x[r]$. For $t \geq 1$, the step size is $\eta_t = \frac{\eta}{w_t}$ (w_t is defined in (3)) for some $\eta > 0$. In order to minimize the *weighted* regret, we also equip the utilities vectors with weights $\{w_t\}$ so that $u^t = w_t \hat{u}^t$ and $m^t = w_t \hat{m}^t$ with $\|\hat{u}^t\|_\infty \leq 1$ and $\|\hat{m}^t\|_\infty \leq 1$. The prediction vector $\hat{m}^t$ is chosen to be $\hat{u}^{t-1}$. We denote the resulting algorithm OFTRL-LogBar.

$$x^t = \operatorname*{argmax}_{x\in\mathcal{X}} \left\{ \frac{\eta}{w_t} \left\langle x, w_t \hat{u}^{t-1} + \sum_{\tau=1}^{t-1} w_\tau \hat{u}^\tau \right\rangle - \mathcal{R}(x) \right\} \tag{OFTRL-LogBar}$$

Note that we can not directly apply Theorem 3 to OFTRL-LogBar since the simplex Δ^d has empty interior. This can be addressed by a transformation on the relative interior $\mathrm{relint}(\Delta^d)$ which preserves the regret (See Appendix C). The following lemma further

[3] The term RVU is from Syrgkanis et al. (2015), which stands for "Regret bounded by Variation in Utilities".

verifies that OFTRL-LogBar with any $\eta \leq \frac{1}{16}$ satisfies the two stability conditions required by Theorem 3.

Lemma 1. *Let $\eta_t = \frac{\eta}{w_t}$, $u^t = w_t \hat{u}^t$, and $m^t = w_t \hat{m}^t$ such that $\|\hat{u}^t\|_\infty, \|\hat{m}^t\|_\infty \leq 1$. Then the iterates of OFTRL-LogBar satisfy $\eta_t \|u^t - m^t\|_{*,x^t} \leq 2\eta$ and*

$$\left\| \eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau \right\|_{*,g^{t-1}} \leq 3\eta.$$

Proof. By definition, for any vector $u \in \mathbb{R}^d$ and $x \in \mathrm{int}(\Delta^d)$, it holds that $\|u\|_{*,x} \leq \|u\|_\infty$. Then we have

$$\eta_t \|u^t - m^t\|_{*,x^t} = \eta \|\hat{u}^t - \hat{m}^t\|_{*,x^t} \leq \eta \|\hat{u}^t - \hat{m}^t\|_\infty \leq 2\eta.$$

Using properties of $\{w_t\}$ (Lemma 4), we have

$$\begin{aligned}
&\left\| \eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau \right\|_{*,g^{t-1}} \\
&\leq \eta \|\hat{m}^t\|_\infty + \eta \left\| \left(\frac{1}{w_{t-1}} - \frac{1}{w_t} \right) \sum_{\tau=1}^{t-1} w^i \hat{u}^\tau \right\|_\infty \\
&\leq \eta + \eta \left(\frac{1}{w_{t-1}} - \frac{1}{w_t} \right) \sum_{\tau=1}^{t-1} w^i \\
&\leq \eta + \eta \cdot \frac{H+1}{H} \leq 3\eta. \qquad \square
\end{aligned}$$

Combining Theorem 3 and Lemma 1 with additional analysis, we have the following RVU bound for OFTRL-LogBar.

Corollary 1. *Let $\eta \leq \frac{1}{16}$. Then, the regret of OFTRL-LogBar under any sequence of utilities $u^1, \ldots, u^T$ can be bounded as*

$$\begin{aligned}
\mathrm{Reg}^T(x^*) \leq \frac{R(x^*)}{\eta_T} + 2 \sum_{t=1}^{T} \eta_t \|u^t - m^t\|_{*,x^t}^2 \\
- \sum_{t=1}^{T} \frac{1}{16\eta_{t-1}} \|x^t - x^{t-1}\|_{x^{t-1}}^2,
\end{aligned}$$

for any $x^* \in \mathrm{relint}(\Delta^d)$, where $\|x^t - x^{t-1}\|_{x^{t-1}}^2 := \sum_{r=1}^{d} (\frac{x^t[r] - x^{t-1}[r]}{x^{t-1}[r]})^2$.

Swap Regret Applying Corollary 1 to each $\mathfrak{R}_a$, then the swap regret $\mathrm{SwapReg}^T = \sum_{a \in \mathcal{A}} \mathrm{Reg}_a^T$ of Algorithm 4 is upper bounded by

$$\begin{aligned}
\frac{2|\mathcal{A}|^2 \log T}{\eta_T} + 2 \sum_{t=1}^{T} \eta_t \|u^t - m^t\|_\infty^2 \\
- \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \frac{\|x_a^t - x_a^{t-1}\|_{x_a^{t-1}}^2}{16\eta_{t-1}},
\end{aligned}$$

where we get a $\log T$ factor due to the diameter of the log barrier regularizer. Following techniques developed in (Anagnostides et al., 2022b), we show a key lemma that lower bounds $\sum_{a \in \mathcal{A}} \frac{1}{16\eta_{t-1}} \|x_a^t - x_a^{t-1}\|_{x_a^{t-1}}^2$ by movement in the output strategies $\|x^t - x^{t-1}\|_1^2$.

Lemma 2. *Suppose $\eta < \frac{1}{28|\mathcal{A}|}$. Then the iterates of Algorithm 4 satisfies for any $t \in [T]$,*

$$\|x^t - x^{t-1}\|_1^2 \leq 64|\mathcal{A}| \sum_{a \in \mathcal{A}} \|x_a^t - x_a^{t-1}\|_{x_a^{t-1}}^2.$$

Combining the above then gives a RVU bound for swap regret of Algorithm 4 with variable step size.

Theorem 4 (RVU for swap regret)**.** *Let $\eta < \frac{1}{28|\mathcal{A}|}$. Then for any $t \in [T]$, the swap regret of Algorithm 4 is at most*

$$\frac{2|\mathcal{A}|^2 \log T}{\eta_T} + \sum_{t=1}^{T} 4\eta_t \|u^t - m^t\|_\infty^2 - \sum_{t=1}^{T} \frac{\|x^t - x^{t-1}\|_1^2}{1024|\mathcal{A}|\eta_{t-1}}.$$

4.2 Bounding Per-State Regret

In this subsection, we prove upper bounds for $\mathrm{reg}_{i,h}^t(s)$, the per-state regret for player $i \in [n]$ and any $(h, s, t) \in [H] \times [\mathcal{S}] \times [T]$. For simplicity of notation, throughout this subsection, we fix (i, h, s) and omit the subscripts (h, s) within the policies and Q-functions, i.e., $\pi_{i,h}^t(\cdot \mid s)$ will be abbreviated as π_i^t and $Q_{i,h}^t(s, \cdot)$ will be abbreviated as Q_i^t. We also overload T be any iteration $T \geq 1$.

Recall the policy update step in Algorithm 3 where we feed $u_i^t = w_t \frac{1}{H} (Q_i^t)^\top \pi_{-i}^t$ to BM-OFTRL-LogBar (Algorithm 4). Thus we can relate $\mathrm{reg}_{i,h}^T(s)$ to the regret incurred by OFTRL-Log-Bar for any $T \geq 1$ as follows:

$$\begin{aligned}
\mathrm{reg}_{i,h}^T(s) &= \max_{\phi_i} \sum_{t=1}^{T} \alpha_T^t \langle \phi_i \diamond \pi_i^t - \pi_i^t, Q_i^t \pi_{-i}^t \rangle \\
&= H\alpha_T^1 \cdot \underbrace{\max_{\phi_i} \sum_{t=1}^{T} \langle \phi_i \diamond \pi_i^t - \pi_i^t, u_i^t \rangle}_{\mathrm{SwapReg}_i^T}, \qquad (5)
\end{aligned}$$

where the second equality holds since $\alpha_T^t = \alpha_T^1 w_t$ (defined in Equation (2) and Equation (3)). Now we apply

Theorem 4 and obtain that for any $T \ge 1$,

$$
\begin{aligned}
\mathrm{SwapReg}_i^T &= \max_{\phi_i} \sum_{t=1}^T \left\langle \phi_i \diamond \pi_i^t - \pi_i^t, u_i^t \right\rangle \\
&\le \frac{2A_i^2 \log T}{\eta_T} + \sum_{t=1}^T 2\eta_t \left\| u_i^t - u_i^{t-1} \right\|_\infty^2 - \sum_{t=1}^T \frac{\left\| \pi_i^t - \pi_i^{t-1} \right\|_1^2}{1024 A_i \eta_{t-1}} \\
&\le \frac{2w_T A_i^2 \log T}{\eta} + \sum_{t=1}^T \frac{2\eta w_t}{H^2} \left\| Q_i^t \pi_{-i}^t - Q_i^{t-1} \pi_{-i}^{t-1} \right\|_\infty^2 \\
&\qquad - \sum_{t=1}^T \frac{w_t}{1024 A_i \eta H} \left\| \pi_i^t - \pi_i^{t-1} \right\|_1^2,
\end{aligned}
$$

where we use $\eta_t = \frac{\eta}{w_t}$, $u_i^t = w_t \frac{1}{H} (Q_i^t)^\top \pi_{-i}^t$ and $w_{t-1} = \frac{w_t (t-1)}{H+t-1} \ge \frac{w_t}{H}$ in the last equality. Using the fact that $\left\| Q_i^t \pi_{-i}^t - Q_i^{t-1} \pi_{-i}^{t-1} \right\|_\infty^2 \le 2 \left\| Q_i^t - Q_i^{t-1} \right\|_\infty^2 + 2 \left\| Q_i^t (\pi_{-i}^t - \pi_{-i}^{t-1}) \right\|_\infty^2$ as well as $\left\| Q_i^t - Q_i^{t-1} \right\|_\infty^2 \le \alpha_t^2 H^2$ and $\left\| Q_i^t (\pi_{-i}^t - \pi_{-i}^{t-1}) \right\|_\infty^2 \le H^2 \left\| \pi_{-i}^t - \pi_{-i}^{t-1} \right\|_1^2$, we can further upper bound $\mathrm{SwapReg}_i^T$ as

$$
\begin{aligned}
&\frac{2w_T A_i^2 \log T}{\eta} + 4\eta \underbrace{\sum_{t=1}^T w_t (\alpha_t)^2}_{\mathbf{I}} + 4\eta \underbrace{\sum_{t=1}^T w_t \left\| \pi_{-i}^t - \pi_{-i}^{t-1} \right\|_1^2}_{\mathbf{II}} \\
&- \sum_{t=1}^T \frac{w_t}{1024 A_i \eta H} \left\| \pi_i^t - \pi_i^{t-1} \right\|_1^2. \qquad (6)
\end{aligned}
$$

Now we focus on upper bounding term **I** and **II**. From Lemma 4, we know $\sum_{t=1}^T \alpha_T^t (\alpha_t)^2 \le \frac{4H}{T}$. This further implies $\mathbf{I} \le \frac{4H}{\alpha_T^1 T}$ since $\alpha_T^t = \alpha_T^1 w_t$.

For term **II**, we have

$$
\begin{aligned}
\mathbf{II} &= \sum_{t=1}^T w_t \left\| \pi_{-i}^t - \pi_{-i}^{t-1} \right\|_1^2 \le \sum_{t=1}^T w_t \left(\sum_{j \ne i} \left\| \pi_j^t - \pi_j^{t-1} \right\|_1 \right)^2 \\
&\le (n-1) \sum_{t=1}^T w_t \sum_{j \ne i} \left\| \pi_j^t - \pi_j^{t-1} \right\|_1^2,
\end{aligned}
$$

where the first inequality holds since the total variational distance between two product distribution is bounded by the sum of total variational distance between each marginal distribution.

Then the total swap regret $\sum_{i=1}^n \mathrm{SwapReg}_i^T$ can be upper bounded by

$$
\begin{aligned}
\sum_{i=1}^n \mathrm{SwapReg}_i^T &\le \frac{2w_T n A_{\max}^2 \log T}{\eta} + 4\eta n \sum_{t=1}^T w_t (\alpha_t)^2 \\
&+ \sum_{j=1}^n \sum_{t=1}^T \left(4\eta w_t n^2 - \frac{w_t}{1024 A_j \eta H} \right) \left\| \pi_j^t - \pi_j^{t-1} \right\|_1^2 \\
&\le \frac{2w_T n A_{\max}^2 \log T}{\eta} + 4\eta n \sum_{t=1}^T w_t (\alpha_t)^2 \\
&- 4\eta n^2 \sum_{j=1}^n \sum_{t=1}^T w_t \left\| \pi_j^t - \pi_j^{t-1} \right\|_1^2,
\end{aligned}
$$

since $\eta = \frac{1}{128 n \sqrt{H} A_{\max}}$. Since the swap regret is non-negative, the above inequality implies

$$
\begin{aligned}
\mathbf{II} &\le n \sum_{j=1}^n \sum_{t=1}^T w_t \left\| \pi_j^t - \pi_j^{t-1} \right\|_1^2 \\
&\le 8192 H n^2 A_{\max}^4 w_T \log T + \underbrace{\sum_{t=1}^T w_t (\alpha_t)^2}_{=\mathbf{I}}.
\end{aligned}
$$

Now we can plug the above bounds on terms **I** and **II** into the individual regret bound in Equation (6) and multiply $H\alpha_T^1$ (Equation (5)) to bound $\mathrm{reg}_{i,h}^T(s)$. Since $\eta = \frac{1}{128 n \sqrt{H} A_{\max}}$ and $\alpha_T^1 w_T = \alpha_T \le \frac{2H}{T}$, we finally get

$$
\begin{aligned}
\mathrm{reg}_{i,h}^T(s) &= H\alpha_T^1 \cdot \mathrm{SwapReg}_i^T \\
&\le \frac{2H A_{\max}^2 \alpha_T \log T}{\eta} + 4\eta H \alpha_T^1 \cdot (\mathbf{I} + \mathbf{II}) \\
&\le 2048 n H^{5/2} A_{\max}^3 \frac{\log T}{T}. \qquad (7)
\end{aligned}
$$

Since the above holds for all $(i, s, h) \in [n] \times [\mathcal{S}] \times [H]$ and $T \ge 1$, we conclude that the maximum weighted regret over all players, all states, and all steps is $\max_{h \in [H]} \overline{\mathrm{reg}}_h^t \le O(\frac{\log t}{t})$.

Proof of Theorem 2 Combining Theorem 1 and the weighted regret upper bound in Equation (7), we conclude that

$$
\begin{aligned}
\mathrm{CEGap}(\widehat{\pi}^T) &\le 2H \cdot \frac{1}{T} \sum_{t=1}^T \max_{h \in [H]} \overline{\mathrm{reg}}_h^t \\
&\le 4096 H^{3.5} n A_{\max}^3 \cdot \frac{1}{T} \sum_{t=1}^T \frac{\log t}{t} \\
&\le 8192 H^{3.5} n A_{\max}^3 \cdot \frac{(\log T)^2}{T}.
\end{aligned}
$$

This completes the proof.

5 Conclusion and Future Directions

In this work, we propose a policy optimization algorithm with near-optimal $\tilde{O}(T^{-1})$ convergence rate to correlated equilibrium in general-sum Markov games. Our result improves the results and answers the open questions in previous works (Zhang et al., 2022; Yang and Ma, 2023).

A natural future direction is to further improve the convergence rates with respect to the number of iterations T. We remark that shaving the $\operatorname{poly}\log T$ factors is challenging even in normal-form games. Other directions include improving the dependence on the horizon H and size of action set $A_{\max}$, and generalizing our results in the setting with oracle access to reward/transition to the sample-based setting where all game parameters are unknown and have to be learned from iteractions.

Acknowledgements HL is supported by NSF award IIS-1943607 and a Google Research Scholar Award. YC and WZ are supported by the NSF Award CCF-1942583 (CAREER).

References

Anagnostides, I., Daskalakis, C., Farina, G., Fishelson, M., Golowich, N., and Sandholm, T. (2022a). Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*.

Anagnostides, I., Farina, G., Kroer, C., Lee, C.-W., Luo, H., and Sandholm, T. (2022b). Uncoupled learning dynamics with $o(\log t)$ swap regret in multiplayer games. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., et al. (2020). The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216.

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

Bowling, M. and Veloso, M. (2001). Rational and convergent learning in stochastic games. In *International joint conference on artificial intelligence*, volume 17, pages 1021–1026. Citeseer.

Chen, X. and Peng, B. (2020). Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:18990–18999.

Cui, Q., Zhang, K., and Du, S. (2023). Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2651–2652. PMLR.

Daskalakis, C., Deckelbaum, A., and Kim, A. (2011). Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM.

Daskalakis, C., Fishelson, M., and Golowich, N. (2021). Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems (NeurIPS)*.

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. (2022). Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR.

Erez, L., Lancewicki, T., Sherman, U., Koren, T., and Mansour, Y. (2023). Regret minimization and convergence to equilibria in general-sum markov games. In *International Conference on Machine Learning*, pages 9343–9373. PMLR.

Foster, D. J., Golowich, N., and Kakade, S. M. (2023). Hardness of independent learning and sparse equilibrium computation in Markov games. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10188–10221. PMLR.

Freund, Y. and Schapire, R. E. (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? *Advances in neural information processing systems*, 31.

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

Kao, H., Wei, C.-Y., and Subramanian, V. (2022). Decentralized cooperative reinforcement learning with hierarchical information structure. In *International Conference on Algorithmic Learning Theory*, pages 573–605. PMLR.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2021). Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*.

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

Littman, M. L. et al. (2001). Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328.

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

Nemirovski, A. (2004). Interior point polynomial time methods in convex programming. *Lecture notes*, 42(16):3215–3224.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Rakhlin, S. and Sridharan, K. (2013). Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26.

Sayin, M., Zhang, K., Leslie, D., Basar, T., and Ozdaglar, A. (2021). Decentralized q-learning in zero-sum markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334.

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

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

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

Stoltz, G. and Lugosi, G. (2007). Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59(1):187–208.

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. (2015). Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W. M., Dudzik, A., Huang, A., Georgiev, P., Powell, R., et al. (2019). Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2:20.

Wang, Y., Liu, Q., Bai, Y., and Jin, C. (2023). Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. *arXiv preprint arXiv:2302.06606*.

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

Yang, Y. and Ma, C. (2023). \$o(tˆ{-1})\$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. In *The Eleventh International Conference on Learning Representations*.

Zhang, R., Liu, Q., Wang, H., Xiong, C., Li, N., and Bai, Y. (2022). Policy optimization for markov games: Unified framework and faster convergence. *Advances in Neural Information Processing Systems*, 35:21886–21899.

Supplementary Materials for Near-Optimal Policy Optimization for Correlated Equilibrium in General-Sum Markov Games

Contents

A Properties of α_t^i and w_i

We present several useful properties of the sequence $\{\alpha_t^i\}_{t\geq 1, 1\leq i\leq t}$ and $\{w_t\}_{t\geq 1}$ in the following lemmas that are known in previous works (Jin et al., 2018; Zhang et al., 2022). We first recall that $\alpha_t = \frac{H+1}{H+t}$ for all $t \geq 1$. The definitions of $\{\alpha_t^i\}_{t\geq 1, 1\leq i\leq t}$ and $\{w_t\}_{t\geq 1}$ are

$$\alpha_t^t = \alpha_t, \quad \alpha_t^i = \alpha_i \Pi_{j=i+1}^t (1-\alpha_j), \forall i \leq t-1,$$

and

$$w_t = \frac{\alpha_t^t}{\alpha_t^1}.$$

Lemma 3 ((Jin et al., 2018))**.** *The sequence $\{\alpha_t^i\}_{t\geq 1, 1\leq i\leq t}$ satisfies the following:*

$$\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}, \forall i \geq 1.$$

Lemma 4 ((Zhang et al., 2022))**.** *The sequence $\{\alpha_t^i\}_{t\geq 1, 1\leq i\leq t}$ and $\{w_t\}_{t\geq 1}$ satisfies the following:*

1. $\sum_{t=1}^T \alpha_T^t (\alpha_t)^2 \leq \frac{4H}{T}$.
2. $w^t = \frac{\alpha_T^t}{\alpha_T^1}$ *for all* $T \geq t$.
3. $(\frac{1}{w_{t-1}} - \frac{1}{w_t}) \sum_{i=1}^{t-1} w^i = \frac{H+1}{H}$.
4. *Given a sequence* $\{\Delta_h^t\}_{h,t}$ *defined by*

$$\begin{cases} \Delta_h^t = \sum_{i=1}^t \alpha_t^i \Delta_{h+1}^i + \beta_t, \\ \Delta_{H+1}^t = 0, \forall t, \end{cases}$$

where $\{\beta_t\}$ *is non-increasing in* t, *then* $\Delta_h^{t+1} \leq \Delta_h^t$ *for all* $t \geq 1$ *and* $h \in [H+1]$.

B Missing Proofs in Section 3.1

B.1 Equivalence between V update and Q update

In this section, we prove the equivalence between Algorithm 1 and Algorithm 3.

Proposition 1 (Equivalence between V update and Q update)**.** *Algorithm 1 and Algorithm 3 are equivalent in the sense that they produce the same sequence of policies $\{\pi_h^t\}(s,\cdot)$.*

Proof. It suffices to prove that the Q value in Algorithm 3 and V value in Algorithm 1 satisfies the following: for any $(i, s, h, \boldsymbol{a})$ and $t \in [T]$,

$$Q_{i,h}^t(s,\boldsymbol{a}) = \left[r_h + \mathbb{P}_h V_{i,h+1}^t\right](s,\boldsymbol{a}). \tag{8}$$

Note that the above holds for $t = 0$ according to the initialization step in both algorithms. Since $\alpha_t = 1$ and $Q_{i,H+1}^1(s,\boldsymbol{a}) = V_{i,H+1}^t(s) = 0$, we can also verify by induction that Equation (8) holds for $t = 1$:

$$Q_{i,h}^1(s,\boldsymbol{a}) = \left[r_h + \mathbb{P}_h[Q_{i,h+1}^1 \pi_{h+1}^1]\right](s,\boldsymbol{a}) = \left[r_h + \mathbb{P}_h V_{i,h+1}^t\right](s,\boldsymbol{a}).$$

Moreover, it is easy to see $Q_{i,H}^t(s,\boldsymbol{a}) = r_H(s,\boldsymbol{a})$ for all $t \geq 1$. Now we conduct induction on both t and h. We assume Equation (8) holds for $(t-1, h)$ and $(t, h+1)$, then for (t, h), we have

$$\begin{aligned}
Q_{i,h}^t(s,\boldsymbol{a}) &= (1-\alpha_t) Q_{i,h}^{t-1}(s,\boldsymbol{a}) + \alpha_t (r_h + \mathbb{P}_h[Q_{i,h+1}^t \pi_{h+1}^t])(s,\boldsymbol{a}) \\
&= (1-\alpha_t)\Big[r_h + \mathbb{P}_h V_{i,h+1}^{t-1}\Big](s,\boldsymbol{a}) + \alpha_t (r_h + \mathbb{P}_h[(r_{h+1} + \mathbb{P}_{h+1} V_{i,h+2}^t)\pi_{h+1}^t])(s,\boldsymbol{a}) \\
&\qquad \text{(by induction hypothesis)} \\
&= \Big[r_h + \mathbb{P}_h\Big((1-\alpha_t)V_{i,h+1}^{t-1} + \alpha_t(r_{h+1} + \mathbb{P}_{h+1}V_{i,h+2}^t)\pi_{h+1}^t\Big)\Big](s,\boldsymbol{a}) \\
&= \left[r_h + \mathbb{P}_h V_{i,h+1}^t\right](s,\boldsymbol{a}). \qquad \text{(by update rule of } V_{i,h+1}^t(s) \text{ in Algorithm 1)}
\end{aligned}$$

This completes the proof. □

B.2 Proof of Theorem 1

We need the following two technical lemmas in the proof of Theorem 1. In Lemma 5, we show that the value function $V^t_{i,h}(s)$ maintained in Algorithm 1 equals to $V^{\widehat{\pi}^t_h}_{i,h}(s)$ where the policy $\widehat{\pi}^t_h$ is defined in Algorithm 2. In Lemma 6, we prove a recursive inequality that bounds the CEGap of $\widehat{\pi}^t_h$ by weighted regret (as defined in (4)).

Lemma 5. *For all player i and $(h,s) \in [H+1] \times \mathcal{S}$ and $V^t_{i,h}(s)$ being the V values maintained in Algorithm 1, it holds that*

1. $V^t_{i,h}(s) = \sum_{j=1}^t \alpha_t^j [(r_h + \mathbb{P}_h V^j_{i,h+1})\pi^j_h](s)$.
2. $V^t_{i,h}(s) = V^{\widehat{\pi}^t_h}_{i,h}(s)$ *while $\widehat{\pi}^t_h$ are defined in Algorithm 2.*

Proof. Recall the update rule of V value in Algorithm 1:

$$V^t_{i,h}(s) = (1-\alpha_t)V^{t-1}_{i,h}(s) + \alpha_t\left[(r_h + \mathbb{P}_h V^t_{i,h+1})\pi^t_h\right](s).$$

Then the first claim holds by applying the above recursively for $j \in [t]$.

Given the first claim and the equivalence between Algorithm 1 and Algorithm 3, the second claim follows from (Zhang et al., 2022, Lemma G.1). □

Lemma 6. *For the policy $\widehat{\pi}^t_h$ defined in ..., we have for all $(i,h,t) \in [n] \times [H] \times [T]$ that*

$$\max_{s\in\mathcal{S}}\left(\max_{\phi_i} V^{\phi_i \diamond \widehat{\pi}^t_{i,h}, \widehat{\pi}^t_{-i,h}}_{i,h}(s) - V^{\widehat{\pi}^t_h}_{i,h}(s)\right) \le \sum_{j=1}^t \alpha_t^j \max_{s'\in\mathcal{S}}\left(\max_{\phi'_i} V^{\phi_{i'} \diamond \widehat{\pi}^j_{i,h+1} \times \widehat{\pi}^j_{-i,h+1}}_{i,h+1}(s') - V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1}(s')\right) + \mathrm{reg}^t_h.$$

Proof. Fix $(i,h,t) \in [n] \times [H] \times [T]$. We have for all state $s \in \mathcal{S}$ that

$$\begin{aligned}
&\max_{\phi_i} V^{\phi_i \diamond \widehat{\pi}^t_{i,h}, \widehat{\pi}^t_{-i,h}}_{i,h}(s) - V^{\widehat{\pi}^t_h}_{i,h}(s) \\
&\le \max_{\phi_i}\left\langle (\phi_i \diamond \pi^j_{i,h})(\cdot \mid s), \sum_{j=1}^t \alpha_t^j \left[\left(r_h + \max_{\phi_{i'}} \mathbb{P}_h V^{\phi_{i'} \diamond \widehat{\pi}^j_{i,h+1} \times \widehat{\pi}^j_{-i,h+1}}_{i,h+1}\right)\pi^j_{-i,h}\right](s,\cdot)\right\rangle \\
&\qquad - \sum_{j=1}^t \alpha_t^j \left\langle \pi^j_{i,h}(\cdot \mid s), \left[\left(r_h + \mathbb{P}_h V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1}\right)\pi^j_{-i,h}\right](s,\cdot)\right\rangle \\
&\le \sum_{j=1}^t \alpha_t^j \max_{s'\in\mathcal{S}}\left(\max_{\phi'_i} V^{\phi_{i'} \diamond \widehat{\pi}^j_{i,h+1} \times \widehat{\pi}^j_{-i,h+1}}_{i,h+1}(s') - V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1}(s')\right) \\
&\qquad + \max_{\phi_i} \sum_{j=1}^t \left\langle (\phi_i \diamond \pi^j_{i,h})(\cdot \mid s) - \pi^j_{i,h}(\cdot \mid s), \left[\left(r_h + \mathbb{P}_h V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1}\right)\pi^j_{-i,h}\right](s,\cdot)\right\rangle \\
&= \sum_{j=1}^t \alpha_t^j \max_{s'\in\mathcal{S}}\left(\max_{\phi'_i} V^{\phi_{i'} \diamond \widehat{\pi}^j_{i,h+1} \times \widehat{\pi}^j_{-i,h+1}}_{i,h+1}(s') - V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1}(s')\right) \\
&\qquad + \underbrace{\max_{\phi_i} \sum_{j=1}^t \left\langle (\phi_i \diamond \pi^j_{i,h})(\cdot \mid s) - \pi^j_{i,h}(\cdot \mid s), \left[\left(r_h + \mathbb{P}_h V^j_{i,h+1}\right)\pi^j_{-i,h}\right](s,\cdot)\right\rangle}_{\mathrm{reg}^t_{i,h}(s)} \\
&\le \sum_{j=1}^t \alpha_t^j \max_{s'\in\mathcal{S}}\left(\max_{\phi'_i} V^{\phi_{i'} \diamond \widehat{\pi}^j_{i,h+1} \times \widehat{\pi}^j_{-i,h+1}}_{i,h+1}(s') - V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1}(s')\right) + \mathrm{reg}^t_h,
\end{aligned}$$

where in the first inequality we use the definition of the policy $\widehat{\pi}^t_h$ which plays π^j_h with probability α^j_t for $j \in [t]$ and then plays $\widehat{\pi}^j_{h+1}$ afterwards; in the second inequality, we replace $\max_{\phi'_i} V^{\phi_{i'} \diamond \widehat{\pi}^j_{i,h+1} \times \widehat{\pi}^j_{-i,h+1}}_{i,h+1}(s')$ with $V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1}(s')$ and pay the difference; in the equality, we use $V^{\widehat{\pi}^j_{i,h+1}}_{i,h+1} = V^j_{i,h+1}$ by Lemma 5; in the last inequality we use the definition of weighted regret (Equation (4)). □

Proof of Theorem 1 Define $\delta_h^t := \max_{i\in[n]} \max_{s\in\mathcal{S}} (\max_{\phi_i} V_{i,h}^{\phi_i \diamond \widehat{\pi}_{i,h}^t, \widehat{\pi}_{-i,h}^t}(s) - V_{i,h}^{\widehat{\pi}_h^t}(s))$. Then by Lemma 6 we have

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

Now let us define an auxiliary sequence $\{\Delta_h^t\}_{h,t}$ such that $\Delta_{H+1}^t = 0$ for all t and

$$\Delta_h^t \le \sum_{j=1}^t \alpha_t^j \Delta_{h+1}^j + \overline{\text{reg}}_h^t.$$

Note that $\Delta_h^t \ge \delta_h^t$ for all (h,t) and $\Delta_h^{t+1} \le \Delta_h^t$ (by Lemma 4). It implies that

$$\begin{aligned}
\Delta_h^t \le \frac{1}{t}\sum_{j=1}^t \Delta_h^j &\le \frac{1}{t}\sum_{j=1}^t\sum_{k=1}^j \alpha_j^k \Delta_{h+1}^k + \frac{1}{t}\sum_{j=1}^t \overline{\text{reg}}_h^j \\
&\le \frac{1}{t}\sum_{k=1}^t (\sum_{j=k}^t \alpha_j^k)\Delta_{h+1}^k + \frac{1}{t}\sum_{j=1}^t \overline{\text{reg}}_h^j \\
&\le (1+\frac{1}{H})\cdot\frac{1}{t}\sum_{j=1}^t \Delta_{h+1}^j + \frac{1}{t}\sum_{j=1}^t \overline{\text{reg}}_h^j && \text{(Lemma 3)} \\
&\le (1+\frac{1}{H})^2\cdot\frac{1}{t}\sum_{j=1}^t \Delta_{h+2}^j + (1+\frac{1}{H})\cdot\frac{1}{t}\sum_{j=1}^t \overline{\text{reg}}_{h+1}^j + \frac{1}{t}\sum_{j=1}^t \overline{\text{reg}}_h^j \\
&\le \dots \\
&\le \left(\sum_{h'=h}^H (1+\frac{1}{H})^{h'-h}\right)\cdot\frac{1}{t}\sum_{j=1}^t \max_{h'\in[H]} \overline{\text{reg}}_{h'}^j \\
&\le (e-1)H\cdot\frac{1}{t}\sum_{j=1}^t \max_{h'\in[H]} \overline{\text{reg}}_{h'}^j \\
&\le 2H\cdot\frac{1}{t}\sum_{j=1}^t \max_{h'\in[H]} \overline{\text{reg}}_{h'}^j.
\end{aligned}$$

Thus $\text{CEGap}(\pi^T) = \text{CEGap}(\pi_1^T) = \delta_1^T \le \Delta_1^T \le 2H\cdot\frac{1}{T}\sum_{t=1}^T \max_{h\in[H]} \overline{\text{reg}}_h^t$. This completes the proof.

C Background on Self-Concordant Barriers

In this section, we present the necessary background on self-concordant barriers and properties of the log barrier that we use in Algorithm 4. We refer the readers to (Nesterov, 2003; Nemirovski, 2004) for a more comprehensive overview of self-concordant barriers.

C.1 Self-Concordant Functions

Definition 3 (Self-Concordant Function)**.** *Let $Q \subseteq \mathbb{R}^d$ be a nonempty open and convex set. A convex function $f: Q \to \mathbb{R}$ in $\mathcal{C}^3$ is called* self-concordant *on Q if it satisfies the following two properties:*

1. *For every sequence $\{x_i \in Q\}_{i=1}^\infty$ converging to a boundary point of Q as $i \to \infty$ it holds that $f(x_i) \to \infty$.*
2. *The functions f satisfies the inequality*

$$|D^3 f(x)[u,u,u]| \le 2(D^2 f(x)[u,u])^{3/2},$$

for all $x \in Q$ and $u \in \mathbb{R}^d$. Here $D^k f(x)[u_1, \dots, u_k]$ denotes the k-th-order differential of f at point x along the directions $u_1, \dots, u_k$.

As an example, the log barrier for the non-negative ray, i.e., $f : (0, \infty) \ni x \to -\log x$, is self-concordant. In the following, we assume f is *non-degenerate*, in the sense that the Hessian $\nabla^2 f(x)$ is positive definite for all $x \in \text{dom} f$. In this context, for any vector $u \in \mathbb{R}^d$, we can define the primal *local norm* with respect to $x \in \text{int}(\mathcal{X})$ as $\|u\|_x := \sqrt{u^\top \nabla^2 \mathcal{R}(x) u}$ and the dual norm as $\|u\|_{*,x} := \sqrt{u^\top (\nabla^2 \mathcal{R}(x))^{-1} u}$. We present some useful properties of self-concordant functions below.

Lemma 7 ((Nesterov, 2003))**.** *Let f be a self-concordant function. Then, for any $x, x' \in \text{dom} f$,*

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle + \omega(\|x' - x\|_x),$$

where $\omega(s) := s - \log(1 + s)$.

Fact 1 ((Anagnostides et al., 2022b))**.** *Let $\omega(s) = s - \log(1 + s)$. Then,*

$$\omega(s) \geq \frac{s^2}{2(1 + s)}.$$

In particular, for $s \in [0, 1]$, it holds that $\omega(s) \geq \frac{s^2}{4}$.

Lemma 8 ((Nesterov, 2003))**.** *Let f be a self-concordant function such that $\|\nabla f(x)\|_{*,x} < 1$ for some $x \in \text{dom} f$. Then the optimization problem*

$$\min_{x \in \text{dom} f} f(x)$$

has a unique solution.

Lemma 9 ((Nemirovski, 2004))**.** *Let $\mathcal{X}$ be a convex and compact set with nonempty interior and $f : \text{int}(\mathcal{X}) \to \mathbb{R}$ be a self-concordant function with $x^* = \text{argmin}_x f(x)$. Then for any $x \in \text{dom} f$ such that $\|\nabla f(x)\|_{*,x} \leq \frac{1}{2}$, it holds that*

$$\|x - x^*\|_x \leq 2\|\nabla f(x)\|_{*,x}, \quad \|x - x^*\|_{x^*} \leq 2\|\nabla f(x)\|_{*,x}.$$

C.2 Self-Concordant Barriers and the Log Barrier

Definition 4 (Self-Concordant Barrier)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex and compact set with nonempty interior* $\text{int}(\mathcal{X})$*. A function $f : \text{int}(\mathcal{X}) \to \mathbb{R}$ is called a θ-self-concordant barrier for $\mathcal{X}$ if*

1. *f is a self-concordant function on* $\text{int}(\mathcal{X})$*;*
2. *for all $x \in \text{int}(\mathcal{X})$ and $u \in \mathbb{R}^d$,*

$$|Df(x)[u]| \leq \theta^{\frac{1}{2}} (D^2 f(x)[u, u])^{1/2}.$$

As an example, the log barrier for the non-negative ray, i.e., $f : (0, \infty) \ni x \to -\log x$, is a 1-self-concordant barrier.

Next, we introduce the log barrier regularizer on the simplex. To address the issue that the simplex Δ^d has empty interior, we will restrict the problem to the domain $\Delta^\circ := \{x \in \mathbb{R}^{d-1}_{\geq 0} : \sum_{r=1}^{d-1} x[r] \leq 1\}$. For notational convenience, we also denote $x[d] = 1 - \sum_{r=1}^{d-1} x[r]$. The log barrier regularizer for Δ° is defined as follows.

Definition 5 (Log Barrier Regularizer for the Simplex)**.** *For $x \in \Delta^\circ$, the log barrier regularizer is*

$$\mathcal{R}(x) := -\sum_{r=1}^{d-1} \log(x[r]) - \log(1 - \sum_{r=1}^{d-1} x[r]). \tag{9}$$

It can be shown that $\mathcal{R}$ is a d-self-concordant barrier. Since the regualarizer $\mathcal{R}$ takes a $(d-1)$-dimensional vector as input while the regret minimizer receives a d-dimensional utility vector $u \in \mathbb{R}^d$, we first explain how the regret minimizer operates on Δ°. Upon receiving a utility vector $u \in \mathbb{R}^d$, the algorithm first constructs $\tilde{u} \in \mathbb{R}^{d-1}$ so that $\tilde{u}[r] = u[r] - u[d]$, for all $r \in [d-1]$. It is clear that the regret incurred is preserved after the transformation. For the purpose of analysis, we also introduce an auxiliary regularizer $\tilde{\mathcal{R}}$:

$$\tilde{\mathcal{R}}(x) := -\sum_{r=1}^{d} \log x[r]. \tag{10}$$

The following claim characterizes and relates the local norm induced by $\mathcal{R}$ and $\tilde{\mathcal{R}}$

Claim 1 ((Anagnostides et al., 2022b)**.** *For any $x, x' \in \operatorname{int}(\Delta^\circ)$.*

$$\|x - x'\|^2_{\mathcal{R},x} = \sum_{r=1}^{d} \left(\frac{x[r] - x'[r]}{x[r]} \right)^2.$$

For any $\tilde{u} \in \mathbb{R}^{d-1}$ and $x \in \operatorname{int}(\Delta^\circ)$,

$$\|\tilde{u}\|_{*,\mathcal{R},x} \le \|u\|_{*,\tilde{\mathcal{R}},x}.$$

D Missing Proofs in Section 4.1

We recall the update rule of Optimistic Follow the Regularized Leader (OFTRL) algorithm. In this section, we focus (OFTRL) with decreasing step size $\eta_t = \frac{\eta}{w_t}$ where $w_t = \frac{\alpha_t^t}{\alpha_t^1}$ for all $t \ge 1$ (see Equation (1), (2) and (3) for definitions). Moreover, we remark that we usually write the utility and prediction vectors in the form of $u^t = w_t \hat{u}^t$, and $m^t = w_t \hat{m}^t$ such that $\|\hat{u}^t\|_\infty, \|\hat{m}^t\|_\infty \le 1$. We define $\eta_0 = \eta_1$, $w_0 = w_1$, and $x^0 := \operatorname{argmin}_{x \in \mathcal{X}} \mathcal{R}(x)$.

$$x^t = \operatorname*{argmax}_{x \in \mathcal{X}} \left\{ \Phi^t(x) := \eta_t \left\langle x, m^t + \sum_{\tau=1}^{t-1} u^\tau \right\rangle - \mathcal{R}(x) \right\} \tag{OFTRL}$$

We also define $\{g^t\}$ as the sequence produced by the conceptual algorithm Be-the-Leader (BTL), which updates g^t with the information of utility vector u^t.

$$g^t = \operatorname*{argmax}_{g \in \mathcal{X}} \left\{ \Psi^t(g) := \eta_t \left\langle g, \sum_{\tau=1}^{t} u^\tau \right\rangle - \mathcal{R}(g) \right\} \tag{BTL}$$

We remark that both (OFTRL) and (BTL) are well-defined if $\eta_t \|u^t - m^t\|_{*,x^t} \le \frac{1}{2}$ and $\|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{*,g^{t-1}} \le \frac{1}{2}$ for all $t \in [T]$, which can be verified using Lemma 8 and Lemma 10.

D.1 RVU for (OFTRL) with Decreasing Step Sizes

Theorem 5 (Adapted from Theorem B.1 in (Anagnostides et al., 2022b))**.** *Suppose that $\mathcal{R}$ is a non-degenerate self-concordant barrier function for $\operatorname{int}(\mathcal{X})$ and let $\eta > 0$. Then,the regret of OFTRL with respect to any $x^* \in \operatorname{int}(\mathcal{X})$ and under any sequence of utilities $u^1, \dots, u^T$ can be bounded as*

$$\frac{R(x^*)}{\eta_T} + \sum_{t=1}^{T} \|u^t - m^t\|_{*,x^t} \|x^t - g^t\|_{x^t} - \sum_{t=1}^{T} \frac{1}{\eta_t} \omega(\|x^t - g^t\|_{x^t}) - \frac{1}{\eta_{t-1}} \omega\Big(\|x^t - g^{t-1}\|_{g^{t-1}}\Big)$$

where the function $\omega(\cdot)$ is defined in Fact 1.

Lemma 10 (Stability)**.** *Let $\eta_t > 0$ be such that $\eta_t \|u^t - m^t\|_{*,x^t} \le \frac{1}{2}$ and $\|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{*,g^{t-1}} \le \frac{1}{2}$. Then we have*

$$\|x^t - g^t\|_{x^t} \le 2\eta_t \|u^t - m^t\|_{*,x^t},$$

$$\|x^t - g^{t-1}\|_{g^{t-1}} \le 2 \left\| \eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau \right\|_{*,g^{t-1}}.$$

Proof. Fix any $t \in [T]$. We first note that $x^t - g^t = x^t - \operatorname{argmin}\{-\Psi^t\}$ by the definition of Ψ^t in (BTL). We also have $\Psi^t(x) = \Phi^t(x) + \eta_t \langle x, u^t - m^t \rangle$. Thus we have

$$\nabla \Psi^t(x^t) = \nabla \Phi^t(x^t) + \eta_t (u^t - m^t) = \eta_t (u^t - m^t),$$

where the second inequality holds since $\nabla \Phi^t(x^t) = 0$ by the first order optimality condition of the optimization problem associated with (OFTRL). By assumption, it further implies that $\|\nabla \Psi^t(x^t)\|_{*,x^t} = \eta_t \|u^t - m^t\|_{*,x^t} \le \frac{1}{2}$. Now we can use Lemma 9 to get

$$\|x^t - g^t\|_{x^t} = \|x^t - \operatorname{argmin}\{-\Psi^t\}\|_{x^t} \le 2 \|\nabla \Psi^t(x^t)\|_{*,x^t} = 2\eta_t \|u^t - m^t\|_{*,x^t}.$$

This finishes the proof of the first inequality.

The proof of the second inequality follows the same idea. We first note that $x^t - g^{t-1} = \operatorname{argmin}\{-\Phi^t(x)\} - g^{t-1}$ by the definition of Φ^t in (OFTRL). We also have $\Phi^t(x) = \Psi^{t-1}(x) + \langle x, \eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau \rangle$. Thus we have

$$\nabla\Phi^t(g^{t-1}) = \nabla\Psi^{t-1}(g^{t-1}) + \eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau = \eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau,$$

where the second inequality holds since $\nabla\Psi^{t-1}(g^{t-1}) = 0$ by the first order optimality condition of the optimization problem associated with (BTL). By assumption, it further implies that $\|\nabla\Phi^t(g^{t-1})\|_{*,g^{t-1}} = \|\eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau\|_{*,g^{t-1}} \le \frac{1}{2}$. Now we can use Lemma 9 to get

$$\left\|x^t - g^{t-1}\right\|_{g^{t-1}} = \left\|g^{t-1} - \operatorname{argmin}\{-\Phi^t\}\right\|_{g^{t-1}} \le 2\left\|\nabla\Phi^t(g^{t-1})\right\|_{*,g^{t-1}} = 2\left\|\eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau\right\|_{*,g^{t-1}}.$$

This finishes the proof of the second inequality. $\square$

D.2 Proof of Theorem 3

Proof. Combining Theorem 5 with the fact that $\|x^t - g^t\|_{x^t} \le 2\eta_t \|u^t - m^t\|_{*,x^t}$ (by Lemma 10) gives

$$\operatorname{Reg}^T(x^*) \le \frac{R(x^*)}{\eta_T} + \sum_{t=1}^T 2\eta_t \left\|u^t - m^t\right\|_{*,x^t}^2 - \sum_{t=1}^T \frac{1}{\eta_t}\omega\left(\left\|x^t - g^t\right\|_{x^t}\right) - \frac{1}{\eta_{t-1}}\omega\Big(\left\|x^t - g^{t-1}\right\|_{g^{t-1}}\Big).$$

Moreover, by Lemma 10, we know $\|x^t - g^t\|_{x^t} \le 1$ and $\|x^t - g^{t-1}\|_{g^{t-1}} \le 1$. Then by Fact 1, we get

$$\operatorname{Reg}^T(x^*) \le \frac{R(x^*)}{\eta_T} + 2\sum_{t=1}^T \eta_t \left\|u^t - m^t\right\|_{*,x^t}^2 - \sum_{t=1}^T \left(\frac{1}{4\eta_t}\left\|x^t - g^t\right\|_{x^t}^2 + \frac{1}{4\eta_{t-1}}\left\|x^t - g^{t-1}\right\|_{g^{t-1}}^2\right).$$

$\square$

The following corollary is useful to apply the RVU property to (OFTRL) with the log barrier regularization.

Corollary 2. *Suppose that $\mathcal{R}$ is a non-degenerate self-concordant barrier function for $\operatorname{int}(\mathcal{X})$ such that $\nabla^2\mathcal{R}(\tilde{x}) \preceq 2\nabla^2\mathcal{R}(x)$ for any $x, \tilde{x} \in \operatorname{int}(\mathcal{X})$ with $\|x - \tilde{x}\|_{\tilde{x}} \le \frac{1}{4}$. Moreover, let $\eta_t > 0$ be such that $\eta_t\|u^t - m^t\|_{*,x^t} \le \frac{1}{8}$ and $\|\eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau\|_{*,g^{t-1}} \le \frac{1}{2}$ for all $t \in [T]$. Then, the regret of OFTRL with respect to any $x^* \in \operatorname{int}(\mathcal{X})$ and under any sequence of utilities $u^1, \dots, u^T$ can be bounded as*

$$\operatorname{Reg}^T(x^*) \le \frac{R(x^*)}{\eta_T} + 2\sum_{t=1}^T \eta_t \left\|u^t - m^t\right\|_{*,x^t}^2 - \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\left\|x^t - x^{t-1}\right\|_{x^{t-1}}^2.$$

Proof. By Lemma 10, we have $\|x^{t-1} - g^{t-1}\|_{x^{t-1}} \le 2\eta_{t-1}\|u^{t-1} - m^{t-1}\|_{*,x^{t-1}} \le \frac{1}{4}$ for all t. Thus by assumption, we have$\nabla^2\mathcal{R}(\widetilde{x^{t-1}}) \preceq 2\nabla^2\mathcal{R}(g^{t-1})$. It further implies $\|x^t - g^{t-1}\|_{x^{t-1}} \le 2\|x^t - g^{t-1}\|_{g^{t-1}}$. Thus we have

$$\begin{aligned}\left\|x^t - x^{t-1}\right\|_{x^{t-1}}^2 &\le 2\left\|x^t - g^{t-1}\right\|_{x^{t-1}}^2 + 2\left\|x^{t-1} - g^{t-1}\right\|_{x^{t-1}}^2 \\ &\le 4\left\|x^t - g^{t-1}\right\|_{g^{t-1}}^2 + 4\left\|x^{t-1} - g^{t-1}\right\|_{x^{t-1}}^2.\end{aligned}$$

Since the step size $\{\eta_t\}$ is non-increasing, we have

$$\begin{aligned}\sum_{t=1}^T \left(\frac{1}{4\eta_t}\left\|x^t - g^t\right\|_{x^t}^2 + \frac{1}{4\eta_{t-1}}\left\|x^t - g^{t-1}\right\|_{g^{t-1}}^2\right) &\ge \sum_{t=1}^T \left(\frac{1}{4\eta_{t-1}}\left\|x^t - g^t\right\|_{x^t}^2 + \frac{1}{4\eta_{t-1}}\left\|x^t - g^{t-1}\right\|_{g^{t-1}}^2\right) \\ &\ge \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\left\|x^t - x^{t-1}\right\|_{x^{t-1}}^2.\end{aligned}$$

Combining the above with Theorem 3 finishes the proof. $\square$

Corollary 3. *Suppose that $\mathcal{R}$ is a non-degenerate self-concordant barrier function for* $\mathrm{int}(\mathcal{X})$ *such that* $\nabla^2\mathcal{R}(\tilde{x}) \leq 2\nabla^2\mathcal{R}(x)$ *for any* $x, \tilde{x} \in \mathrm{int}(\mathcal{X})$ *with* $\|x - \tilde{x}\|_{\tilde{x}} \leq \frac{1}{4}$*. Moreover, let* $\eta_t > 0$ *be such that* $\eta_t\|u^t - m^t\|_{*,x^t} \leq \frac{1}{8}$ *and* $\|\eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau\|_{*,g^{t-1}} \leq \frac{1}{2}$ *for all* $t \in [T]$*. Then*

$$\left\|x^t - x^{t-1}\right\|_{x^{t-1}} \leq 14\eta.$$

Proof. Similar to the proof of Corollary 2, we have

$$\begin{aligned}
\left\|x^t - x^{t-1}\right\|_{x^{t-1}} &\leq \left\|x^t - g^{t-1}\right\|_{x^{t-1}} + \left\|g^{t-1} - x^{t-1}\right\|_{x^{t-1}} \\
&\leq 2\left\|x^t - g^{t-1}\right\|_{g^{t-1}} + \left\|g^{t-1} - x^{t-1}\right\|_{x^{t-1}} \\
&\leq 4\left\|\eta_t m^t + (\eta_t - \eta_{t-1})\sum_{\tau=1}^{t-1} u^\tau\right\|_{*,g^{t-1}} + 2\eta_{t-1}\left\|u^{t-1} - m^{t-1}\right\|_{*,x^{t-1}} \\
&\leq 14\eta. && \text{(by Lemma 1)}
\end{aligned}$$

□

D.3 Proof of Corollary 1: RVU for (OFTRL-LogBar)

Combining Corollary 2 and Claim 1 directly leads to the RVU bound in Corollary 1.

D.4 Proof of Lemma 2

Proof. The proof follows from the proof of Lemma 4.2 in Anagnostides et al. (2022b), which shows that it suffices to prove $\sum_{a\in\mathcal{A}} \mu_a^t \leq \frac{1}{2}$ where μ_a^t is defined as

$$\mu_a^t := \max_{a'\in\mathcal{A}} \left|1 - \frac{x_a^t[a']}{x_a^{t-1}[a']}\right|.$$

Recall the local norm induced by the log barrier regularization (Claim 1): $\|x - x'\|_x^2 = \sum_{r=1}^d \left(\frac{x[r]-x'[r]}{x[r]}\right)^2$. Thus

$$\mu_a^t = \max_{a'\in\mathcal{A}} \left|1 - \frac{x_a^t[a']}{x_a^{t-1}[a']}\right| \leq \sqrt{\sum_{a'\in\mathcal{A}} \left(\frac{x_a^{t-1}[a'] - x_a^t[a']}{x_a^{t-1}[a']}\right)^2} = \left\|x_a^t - x_a^{t-1}\right\|_{x_a^{t-1}}.$$

Now combining Corollary 3 and $\eta < \frac{1}{28|\mathcal{A}|}$, we get

$$\sum_{a\in\mathcal{A}} \mu_a^t \leq |\mathcal{A}| \max_{a\in\mathcal{A}} \left\|x_a^t - x_a^{t-1}\right\|_{x_a^{t-1}} \leq 14|\mathcal{A}|\eta \leq \frac{1}{2}.$$

□

D.5 Proof of Theorem 4: RVU for Swap Regret

Proof. Applying Corollary 2 and to each regret minimizer $\mathfrak{R}_a$ gives the following regret guarantee for all $\eta \leq \frac{1}{28m}$:

$$\begin{aligned}
\mathrm{Reg}_a^T(x_a^*) &\leq \frac{R(x_a^*)}{\eta_T} + 2\sum_{t=1}^T \eta_t\left\|u^t x^t[a] - m^t x^t[a]\right\|_{*,x_a^t}^2 - \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\left\|x_a^t - x_a^{t-1}\right\|_{x_a^{t-1}}^2 \\
&\leq \frac{R(x_a^*)}{\eta_T} + 2\sum_{t=1}^T \eta_t (x^t[a])^2\left\|u^t - m^t\right\|_\infty^2 - \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\left\|x_a^t - x_a^{t-1}\right\|_{x_a^{t-1}}^2
\end{aligned}$$

for any $x_a^* \in \mathrm{relint}(\Delta_{\mathcal{A}})$. Following the same argument in (Anagnostides et al., 2022b, Lemma 4.2), we can also bound the diameter term $R(x_a^*)$ and get

$$\begin{aligned}
\mathrm{Reg}_a^T &\le \frac{|\mathcal{A}|\log T}{\eta_T} + \frac{2}{T}\sum_{t=1}^T x^t[a]\left\|u^t\right\|_\infty + 2\sum_{t=1}^T \eta_t (x^t[a])^2\left\|u^t - m^t\right\|_\infty^2 - \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\left\|x_a^t - x_a^{t-1}\right\|_{x_a^{t-1}}^2 \\
&\le \frac{|\mathcal{A}|\log T}{\eta_T} + 2w_t + 2\sum_{t=1}^T \eta_t (x^t[a])^2\left\|u^t - m^t\right\|_\infty^2 - \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\left\|x_a^t - x_a^{t-1}\right\|_{x_a^{t-1}}^2 \\
&\le \frac{2|\mathcal{A}|\log T}{\eta_T} + 2\sum_{t=1}^T \eta_t (x^t[a])^2\left\|u^t - m^t\right\|_\infty^2 - \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\left\|x_a^t - x_a^{t-1}\right\|_{x_a^{t-1}}^2,
\end{aligned}$$

where we use $w_T = \frac{\eta}{\eta_T} \le \frac{1}{\eta_T}$ in the last inequality. Summing the above inequality for all $a \in \mathcal{A}$ and applying Lemma 2 gives

$$\mathrm{SwapReg}^T \le \sum_{a\in\mathcal{A}} \mathrm{reg}_a^T \le \frac{2|\mathcal{A}|^2\log T}{\eta_T} + 2\sum_{t=1}^T \eta_t\left\|u^t - m^t\right\|_\infty^2 - \sum_{t=1}^T \frac{1}{1024|\mathcal{A}|\eta_{t-1}}\left\|x^t - x^{t-1}\right\|_1^2.$$

□