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

Yang Cai  Haipeng Luo  Chen-Yu Wei  Weiqiang Zheng
Yale University  University of Southern California  University of Virginia  Yale University

Abstract

We study policy optimization algorithms for computing correlated equilibria in multiplayer 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).
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, 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 $\{1, 2, \ldots , n\}$ as $[n]$. For any set $A$, the probability simplex over $A$ is $\Delta_A := \{x \in \mathbb{R}^{|A|} : \sum_{a \in A} x[a] = 1, x[a] \geq 0, \forall a \in A\}$.

Multi-player General-Sum Markov Games

In this paper, we focus on finite-horizon $n$-player general-sum Markov games denoted as $\mathcal{M}(H, \mathcal{S}, \{A_i\}_{i \in [n]}, \mathcal{P}, \{r_i\}_{i \in [n]})$, 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 $|A_i| = A_i$ and we denote a joint action profile of all players as $a = (a_1, a_2, \ldots , a_n) \in \Pi_{i=1}^n A_i$; $\mathcal{P} = \{P_h\}_{h \in [H]}$ is the transition probabilities where $P_h(s' | s, a)$ specifies the probabilities of transition to state $s'$ in step $h + 1$ if players take the joint action $a$ at $s$ in step $h$; $r_i = \{r_{i,h}\}_{h \in [H]}$ are the reward function for player $i$ where $r_{i,h}(s, a) \in [0, 1]$ is the reward for player $i$ when players take the joint action $a$ 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 $a_{i,h} \in A_i$, then each player receives reward $r_{i,h}(s_h, a_{ih})$ and the game transits to the next state $s_{h+1} \sim P_h(\cdot | s_h, a_{ih})$. 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 \( \pi_i \) for player \( i \) is a collection of \( H \) maps \( \{ \pi_{i,h} : \Omega \times (\mathcal{S} \times \prod_{j=1}^{i-1} \mathcal{A}_j)^{h-1} \times \mathcal{S} \to \Delta_{\mathcal{A}_i} \}_{h \in [H]} \) where \( \pi_{i,h} \) maps a random sample \( \omega \) 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 \( \pi_i \), player \( i \) samples \( \omega \) at the beginning of the episode, then at each step \( h \), supposing the history is \( \tau_h := (s_1, a_1, \ldots, s_{j-1}, a_{j-1}) \), player \( i \) chooses action \( a_i,h \sim \pi_{i,h}(\cdot | \omega, \tau_h, s_h) \). We note that \( \omega \) 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 \( \pi_{i,h}(a | s_h) \) specifies the probability of taking action \( a \in \mathcal{A}_i \) at \((h, s_h)\).

A joint policy \( \pi \) is a set of policies denoted as \( \pi = (\pi_1 \circ \pi_2, \ldots \circ \pi_n) \) where the random sample \( \omega \) is shared among all players. We denote \( \pi_{i-1} := (\pi_1 \circ \ldots \circ \pi_{i-1} \circ \pi_{i+1} \circ \ldots \circ \pi_n) \) the joint policy that excludes player \( i \). When the random sample has a special form \( \omega = (\omega_1, \ldots, \omega_n) \) and for each \( i \in [n] \), \( \pi_i \) only uses the randomness in \( \omega_i \) 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 \( \pi_h(a | s) \) the probability of joint action \( a \) at state \( s \). The value function \( V^\pi_{i,h} : \mathcal{S} \to \mathbb{R} \) specifies the expected reward for player \( i \) if all players follow policy \( \pi \):

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

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

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

Strategy Modification and Correlated Equilibrium A strategy modification \( \phi_i \) for player \( i \) is a collection of maps \( \phi_i = \{ \phi_{i,h} : (\mathcal{S} \times \prod_{j=1}^{i-1} \mathcal{A}_j)^{h-1} \times \mathcal{S} \times \Delta_{\mathcal{A}_i} \to \Delta_{\mathcal{A}_i} \}_{h \in [H]} \) such that given history \( \tau_h \) 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.\(^2\) Any policy \( \pi_i \), the modified policy denoted as \( \phi_i \circ \pi_i \), changes the strategy \( \pi_{i,h}(\cdot | \omega, \tau_h, s_h) \) under random sample \( \omega \) and history \( \tau_h \) to another strategy \( \phi_{i,h}(\tau_h, s_h, \pi_{i,h}(\cdot | \omega, \tau_h, s_h)) \).

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 \( \pi \) is a correlated equilibrium (CE) if

\[
\max_{i \in [n]} \max_{\phi_i} \left( V_{i,1}^{(\phi_i \circ \pi_i) \circ \pi_i^{-1}}(s_1) - V_{i,1}^{\pi_i^{-1}}(s_1) \right) \leq 0. \]

A joint policy \( \pi \) is an \( \epsilon \)-approximate CE if \( \text{CEGap}(\pi) := \max_{i \in [n]} \max_{\phi_i} V_{i,1}^{(\phi_i \circ \pi_i) \circ \pi_i^{-1}}(s_1) - V_{i,1}^{\pi_i^{-1}}(s_1) \leq \epsilon \).

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 \( \pi \) is an \( \epsilon \)-approximate coarse correlated equilibrium if \( \max_{i \in [n]} \max_{\phi_i} V_{i,1}^{\pi_i \circ \pi_i^{-1}}(s_1) - V_{i,1}^{\pi_i^{-1}}(s_1) \leq \epsilon \).

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 \( [P_h V](s, a) := \mathbb{E}_{\pi_h(s, a)} V(s') \). For any Markov policy \( \pi_h(\cdot | s) \) and any \( Q \) function \( Q_{i,h}(\cdot | s) : \mathcal{S} \times \prod_{j=1}^{i-1} \mathcal{A}_j \to \mathbb{R} \), we denote \( [Q_{i,h} \pi_h](s) := (Q_{i,h}(s, \cdot), \pi_h(\cdot | s)) \). Similarly, for any joint policy \( \pi_{i-1,h}(\cdot | s) \) that excludes player \( i \), we denote \( [Q_{i-1,h} \pi_{i-1,h}](s, a) := (Q_{i,h}(s, a, \cdot), \pi_{i-1,h}(\cdot | s)) \).

### 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, \( \Phi \)-regret. For a set of strategy modifications \( \Phi = \{ \phi : \mathcal{X} \to \mathcal{X} \} \), the \( \Phi \)-regret of an algorithm \( \mathcal{R} \) over a time horizon \( T \) is defined as

\[
\text{reg}_T^\Phi := \max_{\phi \in \Phi} \sum_{t=1}^{T} \langle u_t, \phi(x_t) - x_t \rangle.
\]

An algorithm is no \( \Phi \)-regret if its \( \Phi \)-regret is sublinear in \( T \). The (external) regret denoted as \( \text{Reg}_T^\Phi \) is \( \Phi \)-regret when \( \Phi \) includes only constant transformations. The swap regret denoted as \( \text{SwapReg}_T^\Phi \) is \( \Phi \)-regret when \( \Phi \) includes all 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
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 \( \mathcal{R} \).

1: Initialize: For all \((i, h, s)\), initialize \( V_{i,h}^0(s) = H - h + 1 \), \( \mathcal{R}_i(h) \) as an instance of \( \mathcal{R} \) over decision set \( \Delta_{A_i} \), and \( \pi_{i,h}^0(s) \) as \( \mathcal{R}_i(h, s) \)'s initial output.

2: for \( t = 1, 2, \ldots, T \) do
3:     for all \((i, s, h)\) do
4:         Forward the utility vector \( u_{i,h,s}^{t-1} := w_{t-1} [r_h + \mathbb{E}_h V_{i,h+1}^{t-1}]\pi_{i,h}^{t-1}(s) \) to \( \mathcal{R}_i(h, s) \).
5:         Update \( \pi_{i,h}^t(s) \) according to \( \mathcal{R}_i(h, s) \).
6:     end for
7:     for all \((i, s, a)\) \in \([n] \times S \times A\), from \( h = H \) to 1:
8:         \[
         V_{i,h}^t(s) \leftarrow (1 - \alpha_t)V_{i,h}^{t-1}(s) + \alpha_t \left[ r_h + \mathbb{E}_h V_{i,h+1}^t \right] \pi_{i,h}^t(s).
         \]
9:     end for
10: Output \( \hat{\pi}_T = \pi_T^1 \) as defined in Algorithm 2.

Algorithm 2 Executing Policy \( \hat{\pi}_T \)

Require: Product policies \( \pi_{i,h}^t(\cdot | s') = \Pi_{h+1}^{n-1} \pi_{i,h}^t(\cdot | s') \) for all \((h', s', t') \in [H] \times S \times [T] \).

1: Sample \( j \in [t] \) with probability \( \Pr[j = i] = \alpha_j \) (see Equation (2) for definition).
2: Play policy \( \pi_j^1 \) at the \( h \)-th step of the game.
3: Play policy \( \hat{\pi}_{h+1}^1 \) for step \( h + 1 \).

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):

\[
\alpha_t = \frac{H + 1}{H + t}.
\]

The choice of \( \alpha_t = O(1/t) \) 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:

\[
\alpha_t' = \alpha_t, \quad \alpha_t'' = \alpha_t \Pi_{j=1}^{t-1} (1 - \alpha_j), \forall t \leq 1.
\]

and

\[
w_0 = w_1, \quad w_t = \frac{\alpha_t}{\alpha_t'}, \forall t \geq 1.
\]

After \( T \geq 1 \) episodes, Algorithm 1 outputs a joint policy \( \hat{\pi}_T \) as defined in Algorithm 2. The output policy is not a Markov policy and is defined recursively. Specifically, at each step \( h \), the policy \( \hat{\pi}_h^1 \) randomly selects a product policy from \( \{\pi_j^1\} \in [t] \) with probability \( \{\alpha_j\} \in [t] \) and plays policy \( \hat{\pi}_{h+1}^1 \) 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 \( Q(s, a) \) function has a size of \( S \cdot \prod_{i=1}^{n-1} A_i \), 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 \( (r_h + \mathbb{E}_h V_{i}^t) \pi_{-i,h}^t \) 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 \( \pi_{-i,h}(\cdot | s) \) 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 \( \pi_{-i,h}(\cdot | s) \) at \((h, s)\).

While we directly assume access to such oracles, both of them can be approximately implemented within \( \varepsilon > 0 \) error using \( \text{poly}(n, A_{\max}, S, H, 1/\varepsilon) \) 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 \( \hat{\pi}_T \) 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 $\{\alpha_i\}$ and $\{w_i\}$, an online learning algorithm $\mathcal{R}$

1: **Initialize:** For all $(i, h, s)$, initialize $Q_i^0(s, a) = H - h + 1$, $\mathcal{R}_{i, h}$ as an instance of $\mathcal{R}$ over decision set $\mathcal{A}_i$, and $\pi_i^0(\cdot | s)$ as $\mathcal{R}_{i, h}$’s initial output
2: for $t = 1, 2, \ldots, T$ do
3: for all $(i, s, h)$ do
4: Forward the utility vector $u_{i, h, s}^{t-1} \leftarrow w_{i, h}^{-1}(H - h + 1)(s, \cdot)$ to $\mathcal{R}_{i, h}$.
5: Update $\pi_i^{t, h} (\cdot | s)$ according to $\mathcal{R}_{i, h}$.
6: end for
7: for all $(i, s, a) \in [n] \times S \times A$, from $h = H$ to 1:
8: $Q_{i, h}(s, a) \leftarrow (1 - \alpha_t)Q_{i, h}^{t-1}(s, a) + \alpha_t(r_h + P_h(Q_{i, h}^{t+1}\pi_{i, h}^{t+1}))(s, a)$.
9: end for

Output $\hat{\pi}^T = \pi_i^T$ as defined in Algorithm 2.

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

$$\delta_h^i := \max_{\phi_i} \max_{s \in S} \left( \max_{\phi_i \{ \pi_{i, h}^T \} \circ \pi_{i, h}^T} (s) - V_{\phi_i}^{T}(s) \right).$$

In Lemma 6, we establish bounds on $\delta_h^i$ using weighted regret such that

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

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

### 3.2 Policy Update

We now turn to the design of the policy update, with the goal of minimizing the weighted swap regret $\text{reg}_h^i$. Each player $i$ maintains a Markov policy $\pi_{i, h}^T (\cdot | s)$ for every pair step $h$ and step $s$. During each episode $t \in [T]$, player $i$ uses $\pi_{i, h}^T$ using an online learning algorithm $\mathcal{R}$. 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 breakthroughs in normal-form games (Anagnostides et al., 2022b), we select $\mathcal{R}$ 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 $\mathcal{R}_{\text{swap}}$ from several external regret minimizers $\mathcal{R}_a$ for each action $a \in \mathcal{A}_i$; (2) employ weighted OFTRL with log barrier regularization for each external regret minimizer $\mathcal{R}_a$. 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 equilibria in multi-player general-sum Markov games.

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

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

$$\text{CEGap}(\hat{x}^T) \leq 8192H^{3.5}nA_{\text{max}}^3 \cdot (\log T)^2 / T.$$

We remark again that the previous best rate for finding CE is $\hat{O}(T^{-2})$ 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 $\hat{O}(T^{-\frac{3}{2}})$ to the weaker notion of CCE. Theorem 2 improves both results significantly and for the first time, shows a near-optimal $\hat{O}(T^{-1})$ convergence rates to CE/CCE in multi-player general-sum Markov games using a single loop and uncoupled policy optimization algorithm.$^3$

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 $\mathcal{R}_{\text{swap}}$ from a set of external regret minimizers $\mathcal{R}_{a} \forall a \in A$, one for each action $a \in A$. The resulting algorithm $\mathcal{R}_{\text{swap}}$ ensures $\text{SwapReg}^T = \sum_{a \in A} \text{Reg}_{a}^T$ (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: $x^0 := \arg\min_{x \in \mathcal{X}} \mathcal{R}(x)$ and for $t \geq 1$, the algorithm updates $x^t$ using step size $\eta_t > 0$.

$$x^t = \arg\max_{x \in \mathcal{X}} \left\{ \eta_t \left( x, m^t + \sum_{\tau=1}^{t-1} u^\tau \right) - \mathcal{R}(x) \right\}$$

(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., 2022) and let $\mathcal{R}$ be a self-concordant barrier. We first extend the RVU-bound established in (Anagnostides et al., 2022) for OFTRL with a constant step size to the case of variable step sizes.$^4$ Before stating the result, we first introduce some notations. We assume $\mathcal{X}$ has a nonempty interior $\text{int}(\mathcal{X})$. We say $\mathcal{R}$ is non-degenerate if its Hessian $\nabla^2 \mathcal{R}(x)$ is positive definite for all $x \in \text{int}(\mathcal{X})$. For any vector $u \in \mathbb{R}^d$, the primal local norm with respect to $x \in \text{int}(\mathcal{X})$ is defined as $\|u\|_x := \sqrt{u^T \nabla^2 \mathcal{R}(x) u}$ and the dual norm is defined as $\|u\|_{x^*} := \sqrt{u^T \nabla^2 \mathcal{R}(x)^{-1} u}$ 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 $\text{int}(\mathcal{X})$ and let $\eta_t > 0$ be such that $\eta_t \|u^t - m^t\|_{x^*} \leq \frac{1}{2}$ and $\|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{x^*} \leq \frac{1}{2}$ for all $t \in [T]$. Then, the regret of OFTRL with respect to any $x^* \in \text{int}(\mathcal{X})$ and under any sequence of utilities $u^1, \ldots, u^T$ can be bounded as

$$\text{Reg}^T(x^*) \leq \frac{\mathcal{R}(x^*)}{\eta_T} + 2 \sum_{t=1}^{T} \eta_t \|u^t - m^t\|_{x^*}^2$$

$$- \sum_{t=1}^{T} \left( \frac{1}{4 \eta_t} \|x^t - g^t\|_{x^*}^2 + \frac{1}{4 \eta_{t-1}} \|x^t - g^{t-1}\|_{g^{t-1}}^2 \right).$$

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 $\eta_t = \frac{\eta}{\log \tau}$ ($w_t$ 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.

$$x^t = \arg\max_{x \in \mathcal{X}} \left\{ \frac{\eta}{w_t} \left( x, w_t \hat{u}^{t-1} + \sum_{\tau=1}^{t-1} w_t \hat{u}^\tau \right) - \mathcal{R}(x) \right\}$$

(OFTRL-LogBar)

$^3$The term RVU is from Syrgkanis et al. (2015), which stands for “Regret bounded by Variation in Utilities”.

$^4$After our paper is accepted and posted on Arxiv, a concurrent work by Mao et al. (2024) was posted on Arxiv which proves a similar result to Theorem 2.
Note that we can not directly apply Theorem 3 to OFTRL-LogBar since the simplex $\Delta^d$ has empty interior. This can be addressed by a transformation on the relative interior $\text{relint}(\Delta^d)$ which preserves the regret (See Appendix C). The following lemma further verifies that OFTRL-LogBar with any $\eta \leq \frac{1}{10}$ satisfies the two stability conditions required by Theorem 3.

**Lemma 1.** Let $\eta_t = \frac{\eta}{m_t}$, $u^t = w_t \hat{u}^t$, 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

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

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

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

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

$$\left\| \eta m^t + (\eta - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau \right\|_{*, x, t} \leq \eta \|\hat{m}^t\|_{\infty} + \eta \left( \frac{1}{w_{t-1}} - \frac{1}{w_t} \right) \sum_{\tau=1}^{t-1} \|u^\tau \|_{\infty} \leq \eta + \eta \left( \frac{1}{w_{t-1}} - \frac{1}{w_t} \right) \sum_{\tau=1}^{t-1} \|u^\tau \|_{\infty} \leq \eta + \eta \cdot \frac{H + 1}{H} \leq 3\eta.$$

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

**Corollary 1.** Let $\eta \leq \frac{1}{10}$. Then, the regret of OFTRL-LogBar under any sequence of utilities $u^1, \ldots, u^T$ can be bounded as

$$\text{Reg}_t^T(x^*) \leq \frac{R(x^*)}{\eta T} + 2\sum_{t=1}^{T} \eta_t \|u^t - m^t\|_{*, x}^2 - \sum_{t=1}^{T} \frac{1}{16\eta T} \|x^t - x^{t-1}\|_{x, t}^2.$$

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

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

$$\frac{2|A|^2 \log T}{\eta T} + 2\sum_{t=1}^{T} \eta_t \|u^t - m^t\|_{*, x}^2 - \sum_{t=1}^{T} \frac{1}{16\eta T} \|x^t - x^{t-1}\|_{x, t}^2,$$

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

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

$$\|x^t - x^{t-1}\|_{t}^2 \leq \frac{64|A|}{\eta T} \sum_{a \in A} \|x^t_a - x^{t-1}_a\|_{x, t}^2.$$

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

$$\frac{2|A|^2 \log T}{\eta T} + 2\sum_{t=1}^{T} \eta_t \|u^t - m^t\|_{*, x}^2 - \sum_{t=1}^{T} \frac{1}{1024|A|\eta T} \frac{1}{1024|A|\eta T} \|x^t - x^{t-1}\|_{x, t}^2.$$

**4.2 Bounding Per-State Regret**

In this subsection, we prove upper bounds for $\text{reg}_{i,h}^t(s)$, the per-state regret for player $i \in [n]$ and any $(h, s, t) \in [H] \times [S] \times [T]$. 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., $\pi^t_i(h, \cdot | s)$ will be abbreviated as $\pi^t_i$ and $Q^t_i(h, s, \cdot)$ will be abbreviated as $Q^t_i$. We also overload $T$ be any iteration $T \geq 1$.

Recall the policy update step in Algorithm 3 where we feed $u^t_i = w_t \hat{u}^t_i (Q^t_i)^\top \pi^t_i$ to BM-OFTRL-LogBar (Algorithm 4). Thus we can relate $\text{reg}_{i,h}^t(s)$ to the regret incurred by OFTRL-LogBar for any $T \geq 1$ as follows:

$$\text{reg}_{i,h}^T(s) = \max_{\phi_i} \sum_{t=1}^{T} \phi_i \langle \phi_i \circ \pi^t_i - \pi^t_i, Q^t_i u^t_i \rangle = H \alpha^T_i \cdot \max_{\phi_i} \sum_{t=1}^{T} \phi_i \langle \phi_i \circ \pi^t_i - \pi^t_i, u^t_i \rangle,$$

where the second equality holds since $\alpha^T_i = \alpha^T_i w_t$ (defined in Equation (2) and Equation (3)). Now we apply
Theorem 4} and obtain that for any $T \geq 1$,

$$\text{SwapReg}_i^T = \max_{\phi_i} \sum_{t=1}^T \langle \phi_i \circ \pi_t^i - \pi_t^{i-1}, u_t^i \rangle$$

$$\leq \frac{2A_i^2 \log T}{\eta T} + \sum_{t=1}^T 2\eta \|u_t^i - u_{t-1}^i\|_2^2 - \sum_{t=1}^T \frac{\|\pi_t^i - \pi_t^{i-1}\|^2}{1024A_i \eta T}$$

$$\leq \frac{2w_TA_i^2 \log T}{\eta} + \frac{\sum_{t=1}^T \eta w_t}{H^2} \|Q_t^i \pi_t^i - Q_t^{i-1} \pi_t^{i-1}\|_2^2$$

$$- \sum_{t=1}^T \frac{w_t}{1024A_i \eta T} \|\pi_t^i - \pi_t^{i-1}\|_2^2,$$

where we use $\eta_t = \frac{\eta}{w_t}$, $u_t^i = w_t \frac{Q_t^i}{\eta T} \pi_t^i$ and $w_{t-1} = \frac{w_t}{\eta T}$ in the last equality. Using the fact that $\|Q_t^i \pi_t^i - Q_t^{i-1} \pi_t^{i-1}\|_2^2 \leq 2\|Q_t^i - Q_t^{i-1}\|_2^2 + 2\|Q_t^i \pi_t^i - Q_t^{i-1} \pi_t^{i-1}\|_2^2$ as well as $\|Q_t^i - Q_t^{i-1}\|_2^2 \leq \alpha_t^2 H^2$ and $\|Q_t^i \pi_t^i - Q_t^{i-1} \pi_t^{i-1}\|_2^2 \leq H^2 \|\pi_t^i - \pi_t^{i-1}\|_2^2$, we can further upper bound SwapReg$_i^T$ as

$$\frac{2w_TA_i^2 \log T}{\eta} + \frac{\eta \sum_{t=1}^T w_t(\alpha_T)^2}{H^2} + 4\eta \sum_{t=1}^T w_t \|\pi_t^i - \pi_t^{i-1}\|_2^2$$

$$- \sum_{t=1}^T \frac{w_t}{1024A_i \eta T} \|\pi_t^i - \pi_t^{i-1}\|_2^2.$$  \hspace{1cm} (6)

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

For term II, we have

$$\Pi = \sum_{t=1}^T w_t \|\pi_t^i - \pi_t^{i-1}\|_1$$

$$\leq (n-1) \sum_{t=1}^T \sum_{j \neq i}^T \|\pi_j^i - \pi_j^{i-1}\|_1,$$

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 $\sum_{t=1}^n \text{SwapReg}_{i}^T$ can be upper bounded by

$$\sum_{t=1}^n \text{SwapReg}_{i}^T \leq \frac{2w_T n A_{max}^2 \log T}{\eta} + 4\eta \sum_{t=1}^T w_t(\alpha_T)^2$$

$$+ \frac{\sum_{j=1}^n \sum_{t=1}^T (4\eta w_t n^2 - \frac{w_t}{1024A_j \eta T}) \|\pi_j^i - \pi_j^{i-1}\|^2}{1}$$

$$\leq \frac{2w_T n A_{max}^2 \log T}{\eta} + 4\eta \sum_{t=1}^T w_t(\alpha_T)^2$$

$$- 4\eta n^2 \sum_{j=1}^n \sum_{t=1}^T w_t \|\pi_j^i - \pi_j^{i-1}\|^2,$$

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

$$\Pi \leq n \sum_{j=1}^n \sum_{t=1}^T w_t \|\pi_j^i - \pi_j^{i-1}\|^2_1$$

$$\leq 8192 H n^2 A_{max}^4 \log T + \sum_{t=1}^T W_t(\alpha_T)^2.$$

Now we can plug the above bounds on terms I and II into the individual regret bound in Equation (6) and multiply $H \alpha_T$ (Equation (5)) to bound $\text{reg}_{i, h}(s)$. Since $\eta = \frac{1}{128n \sqrt{H} A_{max}}$ and $\alpha_T^2 w_T = \alpha_T \leq \frac{4H}{\alpha_T T}$, we finally get

$$\text{reg}_{i, h}(s) = H \alpha_T \cdot \text{SwapReg}_{i}^T$$

$$\leq \frac{2H A_{max}^2 \alpha_T \log T}{\eta} + 4\eta H \alpha_T^2 \cdot (I + II)$$

$$\leq 2048 n H^{3/2} A_{max}^3 \log T.$$  \hspace{1cm} (7)

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

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

$$\text{CEGap}(\bar{\pi}_T^T) \leq 2H \cdot \frac{1}{T} \sum_{t=1}^T \max_{h \in [H]} \text{reg}_h^t$$

$$\leq 4096 H^{3.5} n A_{max}^3 \cdot \frac{1}{T} \sum_{t=1}^T \log t$$

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

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 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 $A_{\text{max}}$, 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 interactions.

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


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


Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won’t result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]
Supplementary Materials for Near-Optimal Policy Optimization for Correlated Equilibrium in General-Sum Markov Games

Contents

1 Introduction 1
  1.1 Related Work ...................................................... 2

2 Preliminaries 2
  2.1 Online Learning and Regret ....................................... 3

3 Algorithm and Main Results 3
  3.1 Value Update ...................................................... 4
  3.2 Policy Update ...................................................... 5
  3.3 Main Results ...................................................... 5

4 Proof of the Main Result 6
  4.1 Optimistic Follow the Regularized Leader with Log Barrier Regularization ................................................. 6
  4.2 Bounding Per-State Regret ....................................... 7

5 Conclusion and Future Directions 9

A Properties of $\alpha^i_t$ and $w_i$ .................................. 13

B Missing Proofs in Section 3.1 ........................................ 13
  B.1 Equivalence between $V$ update and $Q$ update ............... 13
  B.2 Proof of Theorem 1 .................................................. 14

C Background on Self-Concordant Barriers 15
  C.1 Self-Concordant Functions ....................................... 15
  C.2 Self-Concordant Barriers and the Log Barrier ................. 16

D Missing Proofs in Section 4.1 ........................................ 17
  D.1 RVU for (OFTRL) with Decreasing Step Sizes .................... 17
  D.2 Proof of Theorem 3 .................................................. 18
  D.3 Proof of Corollary 1: RVU for (OFTRL-LogBar) .................. 19
  D.4 Proof of Lemma 2 .................................................... 19
  D.5 Proof of Theorem 4: RVU for Swap Regret ....................... 19
A Properties of $\alpha_i^t$ and $w_i$

We present several useful properties of the sequence $\{\alpha_i^t\}_{t \geq 1, 1 \leq i \leq t}$ and $\{w_i\}_{t \geq 1}$ 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 $\{\alpha_i^t\}_{t \geq 1, 1 \leq i \leq t}$ and $\{w_i\}_{t \geq 1}$ are

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

and

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

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

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

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

1. $\sum_{t=1}^{T} \alpha_i^t(\alpha_i)^2 \leq \frac{4H}{T}$.
2. $w_t = \frac{\alpha_i^t}{\alpha_t}$ 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 $\{\Delta_h^t\}_{h \geq 1}$ defined by

$$
\begin{align*}
\Delta_h^t &= \sum_{i=1}^{t} \alpha_i^t \Delta_h^i + \beta_t, \\
\Delta_h^{t+1} &= 0, \forall t,
\end{align*}
$$

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 $(i, s, h, a)$ and $t \in [T]$,

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

(8)

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

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

Moreover, it is easy to see $Q_{i,H}^t(s, a) = r_H(s, a)$ 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

$$
Q_{i,h}^t(s, a) = (1 - \alpha_t) Q_{i,h}^{t-1}(s, a) + \alpha_t \left[ r_h + \mathbb{P}_h [Q_{i,h+1}^{t-1} \pi_h^{i+1}](s, a) \right] = (1 - \alpha_t) \left[ r_h + \mathbb{P}_h V_{i,h+1}^{t-1} \right](s, a) + \alpha_t \left[ r_h + \mathbb{P}_h \left( r_{h+1} + \mathbb{P}_h V_{i,h+2}^t \pi_h^{i+1} \right) \right](s, a)
$$

(by induction hypothesis)

$$
= \left[ r_h + \mathbb{P}_h \left( (1 - \alpha_t) V_{i,h+1}^{t-1} + \alpha_t (r_{h+1} + \mathbb{P}_h V_{i,h+2}^t \pi_h^{i+1}) \right) \right](s, a)
$$

(by update rule of $V_{i,h+1}^t(s)$ in Algorithm 1)

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 $V_{i,h}^{t}(s)$ maintained in Algorithm 1 equals to $V_{i,h}^{t+1}(s)$ where the policy $\pi_{h}^{t}$ is defined in Algorithm 2. In Lemma 6, we prove a recursive inequality that bounds the CEGap of $\pi_{h}^{t}$ by weighted regret (as defined in (4)).

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

1. $V_{i,h}^{t}(s) = \sum_{j=1}^{t} \alpha_{h}^{j}[(r_{h} + P_{h}V_{i,h+1}^{j})\pi_{h}^{j}(s)]$.

2. $V_{i,h}^{t}(s) = V_{i,h}^{t+1}(s)$ while $\pi_{h}^{t}$ are defined in Algorithm 2.

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

$$V_{i,h}^{t}(s) = (1 - \alpha_{t})V_{i,h}^{t-1}(s) + \alpha_{t}[(r_{h} + P_{h}V_{i,h+1}^{t})\pi_{h}^{t}(s)].$$

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., 2021, Lemma 2.1).

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

$$\max_{s \in \mathcal{S}} \left( \max_{\phi_{i}} V_{i,h}^{\phi_{i} \circ \pi_{h}^{t}, \pi_{i,h}^{t}}(s) - V_{i,h}^{t}(s) \right) \leq \sum_{j=1}^{t} \alpha_{h}^{j} \sum_{s' \in \mathcal{S}} \left( \max_{\phi_{i}} V_{i,h+1}^{\phi_{i} \circ \pi_{h}^{j}, \pi_{i,h}^{j}}(s') - V_{i,h+1}^{j}(s') \right) + \text{reg}_{h}^{t}.$$

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

\[
\max_{\phi_{i}} V_{i,h}^{\phi_{i} \circ \pi_{h}^{t}, \pi_{i,h}^{t}}(s) - V_{i,h}^{t}(s) \\
\leq \max_{\phi_{i}} \left( (\phi_{i} \circ \pi_{i,h}^{t})(\cdot \mid s), \sum_{j=1}^{t} \alpha_{h}^{j} \left[ (r_{h} + \max_{\phi_{i}} P_{h}V_{i,h+1}^{j})\pi_{h}^{j}(s', \cdot) \right] \right) \\
- \sum_{j=1}^{t} \alpha_{h}^{j} \left( \pi_{i,h}^{j}(\cdot \mid s), (r_{h} + P_{h}V_{i,h+1}^{j})\pi_{h}^{j}(s', \cdot) \right) \\
\leq \sum_{j=1}^{t} \alpha_{h}^{j} \max_{s' \in \mathcal{S}} \left( \max_{\phi_{i}} V_{i,h+1}^{\phi_{i} \circ \pi_{h}^{j}, \pi_{i,h}^{j}}(s') - V_{i,h+1}^{j}(s') \right) \\
+ \max_{\phi_{i}} \sum_{j=1}^{t} \left( (\phi_{i} \circ \pi_{i,h}^{j})(\cdot \mid s) - \pi_{i,h}^{j}(\cdot \mid s), (r_{h} + P_{h}V_{i,h+1}^{j})\pi_{h}^{j}(s', \cdot) \right) \\
= \sum_{j=1}^{t} \alpha_{h}^{j} \max_{s' \in \mathcal{S}} \left( \max_{\phi_{i}} V_{i,h+1}^{\phi_{i} \circ \pi_{h}^{j}, \pi_{i,h}^{j}}(s') - V_{i,h+1}^{j}(s') \right) \\
+ \max_{\phi_{i}} \sum_{j=1}^{t} \left( (\phi_{i} \circ \pi_{i,h}^{j})(\cdot \mid s) - \pi_{i,h}^{j}(\cdot \mid s), (r_{h} + P_{h}V_{i,h+1}^{j})\pi_{h}^{j}(s', \cdot) \right) \\
\leq \sum_{j=1}^{t} \alpha_{h}^{j} \max_{s' \in \mathcal{S}} \left( \max_{\phi_{i}} V_{i,h+1}^{\phi_{i} \circ \pi_{h}^{j}, \pi_{i,h}^{j}}(s') - V_{i,h+1}^{j}(s') \right) + \text{reg}_{h}^{t},
\]

where in the first inequality we use the definition of the policy $\pi_{h}^{j}$ which plays $\pi_{i,h}^{j}$ with probability $\alpha_{h}^{j}$ for $j \in [t]$ and then plays $\pi_{h+1}^{j}$ afterwards; in the second inequality, we replace $\max_{\phi_{i}} V_{i,h+1}^{\phi_{i} \circ \pi_{h}^{j}, \pi_{i,h}^{j}}(s')$ with $V_{i,h+1}^{j}(s')$ and pay the difference; in the equality, we use $V_{i,h+1}^{j}(s') = V_{i,h+1}^{j}$ by Lemma 5; in the last inequality we use the definition of weighted regret (Equation (4)). \[\square\]
Proof of Theorem 1 Define $\delta^T_h := \max_{i \in [n]} \max_{s \in S} (\max_{\phi_i} V_{t,h}^{\phi_i} s) - V_{t,h}^{\pi_t} s)$. Then by Lemma 6 we have

$$\delta^T_h \leq \sum_{j=1}^t \alpha_j^T \delta^T_{h+1} + \reg^T_h.$$ 

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

$$\Delta^T_h \leq \sum_{j=1}^t \alpha_j^T \Delta^T_{h+1} + \reg^T_h.$$ 

Note that $\Delta^T_h \geq \delta^T_h$ for all $(h, t)$ and $\Delta^T_{h+1} \leq \Delta^T_{h}$ (by Lemma 4). It implies that

$$\Delta^T_h \leq \frac{1}{t} \sum_{j=1}^t \Delta^T_{h+1} \leq \frac{1}{t} \sum_{j=1}^t \alpha_j^T \Delta^T_{h+1} + \frac{1}{t} \sum_{j=1}^t \reg^T_h$$

$$\leq \frac{1}{t} \sum_{k=1}^t (\sum_{j=k}^t \alpha_j^T) \Delta^T_{h+1} + \frac{1}{t} \sum_{j=1}^t \reg^T_h$$

$$\leq (1 + \frac{1}{H}) \cdot \frac{1}{t} \sum_{j=1}^t \Delta^T_{h+1} + \frac{1}{t} \sum_{j=1}^t \reg^T_h$$

$$\leq (1 + \frac{1}{H})^2 \cdot \frac{1}{t} \sum_{j=1}^t \Delta^T_{h+2} + (1 + \frac{1}{H}) \cdot \frac{1}{t} \sum_{j=1}^t \reg^T_h + \frac{1}{t} \sum_{j=1}^t \reg^T_h$$

$$\leq \ldots$$

$$\leq \left( \sum_{h'=h}^H (1 + \frac{1}{H})^{h'-h} \right) \cdot \frac{1}{t} \sum_{j=1}^t \max_{h' \in [H]} \reg^T_{h'}$$

$$\leq (e - 1)H \cdot \frac{1}{t} \sum_{j=1}^t \max_{h' \in [H]} \reg^T_{h'}$$

$$\leq 2H \cdot \frac{1}{t} \sum_{j=1}^t \max_{h' \in [H]} \reg^T_{h'}.$$ 

Thus $\text{CEGap}(\pi^T) = \text{CEGap}(\pi^T) = \delta^T_1 \leq \Delta^T_1 \leq 2H \cdot \frac{1}{t} \sum_{j=1}^T \max_{h \in [H]} \reg^T_h$. 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 $f : Q \rightarrow \mathbb{R}$ in $C^3$ 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 $i \rightarrow \infty$ it holds that $f(x_i) \rightarrow \infty$.

2. The functions $f$ satisfies the inequality

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

for all $x \in Q$ and $u \in \mathbb{R}^d$. Here $D^k f(x)[u_1, \ldots, u_k]$ denotes the $k$-th-order differential of $f$ at point $x$ along the directions $u_1, \ldots, u_k$.
As an example, the log barrier for the non-negative ray, i.e., \( f : (0, \infty) \ni x \rightarrow -\log x \), 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 \( \|u\|_x := \sqrt{u^\top \nabla^2 f(x) u} \) and the dual norm as \( \|u\|_{x,*} := \sqrt{u^\top (\nabla^2 f(x))^{-1} u} \). We present some useful properties of self-concordant functions below.

**Lemma 7** (\cite{Nesterov2003}). 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** (\cite{Anagnostides2022}). 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** (\cite{Nesterov2003}). Let \( f \) be a self-concordant function such that \( \|\nabla f(x)\|_{*,x} \leq 1 \) 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** (\cite{Nemirovski2004}). Let \( \mathcal{X} \) be a convex and compact set with nonempty interior and \( f : \text{int}(\mathcal{X}) \rightarrow \mathbb{R} \) be a self-concordant function with \( x^* = \arg\min_x f(x) \). 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 \( f : \text{int}(\mathcal{X}) \rightarrow \mathbb{R} \) is called a \( \theta \)-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 \( u \in \mathbb{R}^d \),
\[
|Df(x)[u]| \leq \theta^{d/2} (D^2 f(x)[u, u])^{1/2}.
\]

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

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

**Definition 5** (Log Barrier Regularizer for the Simplex). For \( x \in \Delta^o \), 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 regularizer \( \mathcal{R} \) 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 \( \Delta^o \). 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 \(x, x' \in \text{int}(\Delta^n)\),
\[
\|x - x'\|^2_{\mathcal{R}, x} = \sum_{r=1}^{d} \left( \frac{x'[r] - x'[r]}{x'[r]} \right)^2.
\]

For any \(\tilde{u} \in \mathbb{R}^{d-1}\) and \(x \in \text{int}(\Delta^n)\),
\[
\|\tilde{u}\|_{*, \mathcal{R}, x} \leq \|u\|_{*, \mathcal{R}, x}.
\]

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 \(\eta_t = \frac{t}{w_t}\) where \(w_t = \frac{c_t'}{c_1'}\) for all \(t \geq 1\) (see Equation (1), (2) and (3) for definitions). Moreover, we remark that we usually write the utility and prediction vectors in the form of \(u^t = w_t \tilde{u}^t\), and \(m^t = w_t \tilde{m}^t\) such that \(\|\tilde{u}^t\|_\infty, \|\tilde{m}^t\|_\infty \leq 1\). We define \(\eta_0 = \eta_1, w_0 = w_1\), and \(x^0 := \text{argmin}_{x \in \mathcal{X}} \mathcal{R}(x)\).

\[
x^t = \text{argmax}_{x \in \mathcal{X}} \left\{ \Phi^t(x) := \eta_t \left( x, m^t + \sum_{\tau=1}^{t-1} u^\tau \right) - \mathcal{R}(x) \right\} \quad \text{(OFTRL)}
\]

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 \(u^t\).

\[
g^t = \text{argmax}_{g \in \mathcal{X}} \left\{ \Psi^t(g) := \eta_t \left( g, \sum_{\tau=1}^{t} u^\tau \right) - \mathcal{R}(g) \right\} \quad \text{(BTL)}
\]

We remark that both (OFTRL) and (BTL) are well-defined if \(\eta_t \|u^t - m^t\|_{*, x^t} \leq \frac{1}{2}\) 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 \(t \in [T]\), 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 \(\text{int}(\mathcal{X})\) and let \(\eta > 0\). Then, the regret of OFTRL with respect to any \(x^* \in \text{int}(\mathcal{X})\) and under any sequence of utilities \(u^1, \ldots, u^T\) can be bounded as

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

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

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

**Proof.** Fix any \(t \in [T]\). We first note that \(x^t - g^t = x^t - \text{argmin}\{-\Psi^t\}\) by the definition of \(\Psi^t\) in (BTL). We also have \(\Psi(x) = \Phi(x) + \eta_t(x, u^t - m^t)\). 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 \(\|\nabla \Psi^t(x^t)\|_{*, x^t} = \eta_t \|u^t - m^t\|_{*, x^t} \leq \frac{1}{2}\). Now we can use Lemma 9 to get
\[
\|x^t - g^t\|_{x^t} = \|x^t - \text{argmin}\{-\Psi^t\}\|_{x^t} \leq 2\|\nabla \Psi^t(x^t)\|_{*, x^t} = 2\eta_t \|u^t - m^t\|_{*, x^t}.
\]
This finishes the proof of the first inequality.

The proof of the second inequality follows the same idea. We first note that \( x^t - g^{t-1} = \arg\min \{-\Phi^t(x)\} - g^{t-1} \) by the definition of \( \Phi^t \) in (OFTRL). We also have \( \Phi^t(x) = \Psi^{t-1}(x) + \langle x, \eta_t m^t \rangle + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau \). 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 \( \|\nabla \Phi^t(g^{t-1})\|_{s,g^{t-1}} = \|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{s,g^{t-1}} \leq \frac{1}{T} \). Now we can use Lemma 9 to get

\[
\|x^t - g^{t-1}\|_{g^{t-1}} = \|g^{t-1} - \arg\min \{-\Phi^t\}\|_{g^{t-1}} \leq 2\|\nabla \Phi^t(g^{t-1})\|_{s,g^{t-1}} = 2\|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{s,g^{t-1}}.
\]

This finishes the proof of the second inequality.

\[\square\]

### D.2 Proof of Theorem 3

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

\[
\text{Reg}^T(x^*) \leq \frac{R(x^*)}{\eta_T} + \sum_{t=1}^T 2\|x^t - m^t\|_{s,x^t}^2 - \sum_{t=1}^T \frac{1}{\eta_t} \omega\left(\|x^t - g^t\|_{x^t}\right) - \frac{1}{\eta_{t-1}} \omega\left(\|x^t - g^{t-1}\|_{g^{t-1}}\right).
\]

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

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

\[\square\]

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 \( \text{int}(\mathcal{X}) \) such that \( \nabla^2 \mathcal{R}(\tilde{x}) \leq 2\nabla^2 \mathcal{R}(x) \) for any \( x, \tilde{x} \in \text{int}(\mathcal{X}) \) with \( \|x - \tilde{x}\|_{\tilde{x}} \leq \frac{1}{4} \). Moreover, let \( \eta_t > 0 \) be such that \( \eta_t \|u^t - m^t\|_{s,x^t} \leq \frac{1}{8} \) and \( \|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{s,g^{t-1}} \leq \frac{1}{2} \) for all \( t \in [T] \). Then, the regret of OFTRL with respect to any \( x^* \in \text{int}(\mathcal{X}) \) and under any sequence of utilities \( u^1, \ldots, u^T \) can be bounded as

\[
\text{Reg}^T(x^*) \leq \frac{R(x^*)}{\eta_T} + 2\sum_{t=1}^T \eta_t \|u^t - m^t\|_{s,x^t}^2 - \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\|x^t - x^{t-1}\|_{x^{t-1}}^2.
\]

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

\[
\|x^t - x^{t-1}\|_{x^{t-1}}^2 \leq 2\|x^t - g^{t-1}\|_{x^{t-1}}^2 + 2\|x^{t-1} - g^{t-1}\|_{g^{t-1}}^2 \leq 4\|x^t - g^{t-1}\|_{g^{t-1}}^2 + 4\|x^{t-1} - g^{t-1}\|_{x^{t-1}}^2.
\]

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

\[
\sum_{t=1}^T \left(\frac{1}{4\eta_t}\|x^t - g^t\|_{x^t}^2 + \frac{1}{4\eta_{t-1}}\|x^t - g^{t-1}\|_{g^{t-1}}^2\right) \geq \sum_{t=1}^T \left(\frac{1}{4\eta_{t-1}}\|x^t - g^t\|_{x^t}^2 + \frac{1}{4\eta_{t-1}}\|x^t - g^{t-1}\|_{g^{t-1}}^2\right)
\]

\[
\geq \sum_{t=1}^T \frac{1}{16\eta_{t-1}}\|x^t - x^{t-1}\|_{x^{t-1}}^2.
\]

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 int($\mathcal{X}$) such that $\nabla^2\mathcal{R}(\bar{x}) \preceq 2\nabla^2\mathcal{R}(x)$ for any $x, \bar{x} \in \text{int}(\mathcal{X})$ with $\|x - \bar{x}\|_2 \leq \frac{1}{2}$. Moreover, let $\eta_t > 0$ be such that $\eta_t\|u^t - m^t\|_{s,x^t} \leq \frac{1}{8}$ and $\|\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\|_{s,x^t} \leq \frac{1}{2}$ for all $t \in [T]$. Then
\[
\|x^t - x^{t-1}\|_{x^{t-1}} \leq 14\eta.
\]

Proof. Similar to the proof of Corollary 2, we have
\[
\|x^t - x^{t-1}\|_{x^{t-1}} \leq \|x^t - g^{t-1}\|_{x^{t-1}} + \|g^{t-1} - x^{t-1}\|_{x^{t-1}}
\leq 2\|x^t - g^{t-1}\|_{g^{t-1}} + \|g^{t-1} - x^{t-1}\|_{x^{t-1}}
\leq 4\left(\eta_t m^t + (\eta_t - \eta_{t-1}) \sum_{\tau=1}^{t-1} u^\tau\right) + 2\eta_t\|u^{t-1} - m^{t-1}\|_{s,x^{t-1}}
\leq 14\eta. \quad \text{(by Lemma 1)}
\]

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 $\sum_{a \in A} \mu^t_a \leq \frac{1}{2}$ where $\mu^t_a$ is defined as
\[
\mu^t_a := \max_{a' \in A} \left| 1 - \frac{x^t_a [a']}{x^{t-1}_a [a']} \right|.
\]
Recall the local norm induced by the log barrier regularization (Claim 1): $\|x - x'\|_x^2 = \sum_{a=1}^d \left(\frac{x[a] - x'[a]}{x[a]}\right)^2$. Thus
\[
\mu^t_a = \max_{a' \in A} \left| 1 - \frac{x^t_a [a']}{x^{t-1}_a [a']} \right| \leq \sqrt{\sum_{a \in A} \left(\frac{x^{t-1}_a [a'] - x^t_a [a']}{x^{t-1}_a [a']}\right)^2} = \|x^t_a - x^{t-1}_a\|_{x^{t-1}_a}.
\]
Now combining Corollary 3 and $\eta < \frac{1}{2\|\mathcal{X}\|}$, we get
\[
\sum_{a \in A} \mu^t_a \leq |A| \max_{a \in A} \|x^t_a - x^{t-1}_a\|_{x^{t-1}_a} \leq 14|A|\eta \leq \frac{1}{2}.
\]

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

Proof. Applying Corollary 2 and to each regret minimizer $\mathcal{R}_a$ gives the following regret guarantee for all $\eta \leq \frac{1}{28m}$:
\[
\text{Reg}_a^T(x^*_a) \leq \frac{R(x^*_a)}{\eta T} + 2\sum_{t=1}^T \eta_t \|u^t x^t[a] - m^t x^t[a]\|_{s,x^t} + 2\sum_{t=1}^T \frac{1}{16\eta_{t-1}} \|x^t_a - x^{t-1}_a\|_{x^{t-1}_a}^2
\leq \frac{R(x^*_a)}{\eta T} + 2\sum_{t=1}^T \eta_t (x^t[a])^2 \|u^t - m^t\|_\infty^2 - 2\sum_{t=1}^T \frac{1}{16\eta_{t-1}} \|x^t_a - x^{t-1}_a\|_{x^{t-1}_a}^2.
for any $x^*_a \in \text{relin}(\Delta_A)$. Following the same argument in (Anagnostides et al., 2022b, Lemma 4.2), we can also bound the diameter term $R(x^*_a)$ and get

$$
\text{Reg}_a^T \leq \frac{|A| \log T}{\eta T} + 2 \sum_{t=1}^{T} x^t[a] \|u^t\|_{\infty} + 2 \sum_{t=1}^{T} \eta_t(x^t[a])^2 \|u_t - m_t\|_{\infty}^2 - \sum_{t=1}^{T} \frac{1}{16 \eta_{t-1}} \|x^t[a] - x^{t-1}[a]\|_{x^t[a]}^2
$$

$$
\leq \frac{|A| \log T}{\eta T} + 2w_t + \sum_{t=1}^{T} \eta_t(x^t[a])^2 \|u_t - m_t\|_{\infty}^2 - \sum_{t=1}^{T} \frac{1}{16 \eta_{t-1}} \|x^t[a] - x^{t-1}[a]\|_{x^t[a]}^2
$$

$$
\leq \frac{2|A| \log T}{\eta T} + 2 \sum_{t=1}^{T} \eta_t(x^t[a])^2 \|u_t - m_t\|_{\infty}^2 - \sum_{t=1}^{T} \frac{1}{16 \eta_{t-1}} \|x^t[a] - x^{t-1}[a]\|_{x^t[a]}^2,
$$

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

$$
\text{SwapReg}^T \leq \sum_{a \in A} \text{reg}_a^T \leq \frac{2|A|^2 \log T}{\eta T} + 2 \sum_{t=1}^{T} \eta_t \|u_t - m_t\|_{\infty}^2 - \sum_{t=1}^{T} \frac{1}{1024|A| \eta_{t-1}} \|x^t - x^{t-1}\|_1^2.
$$

\[\square\]