Near Optimal Convergence to Coarse Correlated Equilibrium in General-Sum Markov Games

Asrın Efe Yorulmaz AY20@ILLINOIS.COM
University of Illinois Urbana-Champaign

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

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

No-regret learning dynamics play a central role in game theory, enabling decentralized convergence to equilibrium for concepts such as Coarse Correlated Equilibrium (CCE) or Correlated Equilibrium (CE). In this work, we improve the convergence rate to CCE in general-sum Markov games, reducing it from the previously best-known rate of $\mathcal{O}(\log^5 T/T)$ to a sharper $\mathcal{O}(\log T/T)$. This matches the best known convergence rate for CE in terms of T, number of iterations, while also improving the dependence on the action set size from polynomial to polylogarithmic—yielding exponential gains in high-dimensional settings. Our approach builds on recent advances in adaptive step-size techniques for no-regret algorithms in normal-form games, and extends them to the Markovian setting via a stage-wise scheme that adjusts learning rates based on real-time feedback. We frame policy updates as an instance of Optimistic Follow-the-Regularized-Leader (OFTRL), customized for value-iteration-based learning. The resulting self-play algorithm achieves, to our knowledge, the fastest known convergence rate to CCE in Markov games.

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

1. Introduction

Multi-agent systems are increasingly at the forefront of real-world applications, from autonomous driving Shalev-Shwartz et al. (2016), smart grids Chen et al. (2022) and from LLMs Wan et al. (2025); Park et al. (2025) to distributed robotics Levine et al. (2017) and financial markets Zhang et al. (2024). In these environments, agents must learn to make sequential decisions while accounting for the presence of other strategic agents whose actions affect the shared outcome. This interplay between individual learning and collective behavior gives rise to a central class of problems known as *multi-agent reinforcement learning* (MARL) Zhang et al. (2021). The rise of MARL has been motivated not only by its broad applicability but also by the theoretical challenge of designing decentralized algorithms that ensure meaningful long-term behavior in interactive settings.

To model such scenarios, *Markov games*—also known as stochastic games— offer a principled generalization of both Markov decision processes and normal-form games Shapley (1953); Littman (1994) . These games capture the temporal evolution of state, the strategic nature of agent interactions, and the dependence of rewards on joint actions. As such, they provide a natural framework for analyzing multi-agent learning dynamics in dynamic environments. However, understanding what equilibria and how fast these equilibria emerge under different decentralized learning methods in Markov games remains an open and fundamental question.

Table 1: Comparison of convergence rates in normal-form and Markov games

Aimed Equilibria	**Normal-form Games**	**Markov Games**
NE	$\mathcal{O}\left(\frac{\log \lvert A_{\max}\rvert \log T}{T}\right)$ Daskalakis et al. (2011)	$\mathcal{O}\left(\frac{\log \lvert A_{\max}\rvert}{T}\right)$ Yang and Ma (2023)
CE	$\mathcal{O}\left(\frac{\lvert A_{\max}\rvert^{2.5} \log T}{T}\right)$ Anagnostides et al. (2022)	$\mathcal{O}\left(\frac{\lvert A_{\max}\rvert^{2.5} \log T}{T}\right)$ Mao et al. (2024)
CCE	$\mathcal{O}\left(\frac{(\log \lvert A_{\max}\rvert)^2 \log T}{T}\right)$ Soleymani et al. (2025)	$\mathcal{O}\left(\frac{(\log \lvert A_{\max}\rvert)^2 \log T}{T}\right)$ Theorem 4

In normal-form games (NFGs), it is well established that when all players follow no-regret algorithms with $\mathcal{O}(\sqrt{T})$ regret against adversarial opponents, their joint play converges to an $\mathcal{O}(1/\sqrt{T})$-approximate equilibrium—specifically, a Nash equilibrium (NE) in the two-player zero-sum case and a (coarse) correlated equilibrium (CCE) CE, in general-sum settings (Hart and Mas-Colell, 2000; Cesa-Bianchi and Lugosi, 2006). It is well known that multiplicative weights update (MWU) Yoav Freund (1995), online mirror descent (OMD) Nemirovski and Yudin (1983) and follow-the-regularized leader (FTRL) algorithms Abernethy et al. (2008) all fall within this category.

Although the regret for adversarial case is non-improvable, recent advances have further sharpened convergence rates for self-play algorithms in NFGs. Initiated by the seminal work of Daskalakis et al. (2011), which established a convergence rate of $\tilde{\mathcal{O}}(1/T)$ to NE in the two-player zero-sum setting, subsequent studies (Rakhlin and Sridharan, 2013; Syrgkanis et al., 2015; Daskalakis et al., 2021; Anagnostides et al., 2022; Soleymani et al., 2025) provided more refined analyses and faster convergence guarantees compared to the baseline rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$ for general NFGs. In particular, Syrgkanis et al. (2015) demonstrated that when all players in a NFG employ the Optimistic FTRL (OFTRL) algorithm, their strategies converge to a CCE at a rate of $\mathcal{O}(T^{-3/4})$, enabled by the regret bounded by Variation in Utilities (RVU) framework. More recently, Soleymani et al. (2025) established the best-known convergence rate to CCE in NFGs to date, namely $\mathcal{O}\big(\log T/T\big)$, in the self-play setting. They showed that adaptive learning rule corresponds to a specific instantiation of the OFTRL algorithm with a tailored regularizer, which facilitates the improved convergence bound.

Recent advances in learning theory have brought attention to achieve faster convergence rates in Markov games, extending the foundational $\tilde{\mathcal{O}}(1/T)$ convergence results from NFGs to dynamic multi-agent environments. Notable progress has been made for various equilibrium concepts: NE in two-player zero-sum Markov games Yang and Ma (2023), CE in general-sum Markov games Cai et al. (2024); Mao et al. (2024), and CCE in general-sum Markov games Mao et al. (2024). However, despite this progress, there remain limitations in the existing CCE convergence analyses. In particular, the best known rate for CCE learning, $\mathcal{O}((\log T)^5/T)$, lagged behind the CE convergence rate of $\mathcal{O}(\log T/T)$, a gap that undermines the appeal of CCE. Previous CCE results, such as Mao et al. (2024), were based on stage-based frameworks that required long and inflexible stage lengths—typically of the order $T \gg Cn(\log T)^4$—as inherited from Daskalakis et al. (2021), where C is a very large constant (see Lemmas 4.2 and C.4 in Daskalakis et al. (2021)). Importantly, it has been shown that achieving independent no-regret algorithms in Markov games is both statistically and computationally hard Foster et al. (2023); Tian et al. (2021). Consequently, prior work has focused on designing algorithms that rely on shared randomness between agents as we do.

In this work, we resolve these issues by introducing a self-play algorithm for general-sum Markov games that achieves a CCE convergence rate of $\mathcal{O}(\log T/T)$, thereby closing the gap with the CE literature in terms of time-horizon dependence. Crucially, our method not only accelerates convergence in T, but also achieves exponential improvement in dependence on the action space size compared to best known CE convergence rate. This makes learning CCE computationally viable in high-dimensional settings where computing CE is often infeasible. As CCE encompasses a richer set of decentralized strategies and allows for more flexible agent behavior—particularly in environments where correlation is difficult to coordinate—our results enhance the appeal of learning coarse equilibria in Markov games. Our approach builds on a *dynamic step-size adaptation* scheme, which was proposed by Soleymani et al. (2025) and shown to be equivalent to a particular instantiation of the OFTRL algorithm with a regularizer satisfying key smoothness properties. By establishing a RVU inequality under time-varying step sizes, and coupling it with value iteration procedure tailored to the episodic Markov games, we derive aforementioned convergence bounds.

2. Preliminaries

2.1. Multi-player General-sum Markov Games

We consider an N-player episodic Markov game defined by the tuple $\mathcal{G} = \big([N], H, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, \{r_i\}_{i=1}^N, \{P_h\}_{h=1}^H\big)$, where $[N] := \{1, \dots, N\}$ denotes the set of players, $H \in \mathbb{N}^+$ is the episode length (horizon), $\mathcal{S}$ is a finite state space, $\mathcal{A}_i$ is the finite action set of player i, and $\mathcal{A} := \prod_{i=1}^N \mathcal{A}_i$ is the joint action space. The per-step reward function for player i is given by $r_i : [H] \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, and the transition dynamics at step $h \in [H]$ are specified by $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. The agents interact in an unknown environment for T episodes. At each step $h \in [H]$, the system is in state $s_h \in \mathcal{S}$. Each player selects an action $a_{i,h} \in \mathcal{A}_i$, resulting in the joint action $a_h = (a_{1,h}, \dots, a_{N,h}) \in \mathcal{A}$. Player i receives reward $r_{i,h}(s_h, a_h)$, and the next state $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ is sampled. We make the assumption of, Song et al. (2022); Jin et al. (2022); Mao and Başar (2023), the episode beginning at a fixed initial state $s_1 \in \mathcal{S}$. Finally, we let $S = |\mathcal{S}|$, $A_i = |\mathcal{A}_i|$, $A_{\max} = \max_{i \in [N]} A_i$.

2.2. Policies and Value Functions

A (Markov) policy for player $i \in [N]$ is a sequence of functions $\pi_i = \{\pi_{i,h}\}_{h=1}^H$, where each $\pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)$ assigns a distribution over actions based on the current state at step h. A joint policy $\pi = (\pi_1, \dots, \pi_N)$ specifies a probability measure over the trajectory of states and joint actions. We write $\pi = (\pi_i, \pi_{-i})$ to distinguish player i's policy from the other players. Given a policy π, we define the V-function and Q-function for player i at step $h \in [H]$ and state $s \in \mathcal{S}$ as

$$V_{i,h}^{\pi}(s) := \mathbb{E}^{\pi}\Big[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, a_{h'}) \Big| s_h = s\Big], Q_{i,h}^{\pi}(s,a) := \mathbb{E}^{\pi}\Big[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, a_{h'}) \Big| s_h = s,\ a_h = a\Big] \quad (1)$$

where $a \in \mathcal{A}$ is a joint action taken at state s. For any $V_{i,h}$, we define the one-step Bellman operator as $[P_h V](s,a) := \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')]$, and for any $Q_{i,h}$, we define the expected values under policies as $[Q_{i,h}\pi_h](s) := \langle Q_{i,h}(s, \cdot),\ \pi_h(\cdot|s)\rangle$ and $[Q_{i,h}\pi_{-i,h}](s, a_i) := \langle Q_{i,h}(s, a_i, \cdot),\ \pi_{-i,h}(\cdot|s)\rangle$.

2.3. Decentralized Information Feedback

In our setting, we assume that each player i has access to the necessary information to compute their value function $V_{i,h}(s)$ at each stage h and state s. The update for $V_{i,h}(s)$ depends only on the

expected value of $r_h + P_h V_{i,h+1}$ under the joint policy of the other players, $\pi_{-i,h}(\cdot \mid s)$. In particular, for any fixed (s, a_i), we assume access to a *reward oracle* that returns $\mathbb{E}_{a_{-i}\sim\pi_{-i,h}}[r_i(s, a_i, a_{-i})]$, and a *transition oracle* that returns the marginal distribution $\mathbb{E}_{a_{-i}\sim\pi_{-i,h}}[P_h(\cdot \mid s, a_i, a_{-i})]$.

2.4. Correlated Policies and Coarse Correlated Equilibrium

We now extend our policy class to allow for coordination through shared randomness. A *correlated policy* π comprises a sequence of decision rules $\pi_h : \Omega \times (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \Delta(\mathcal{A}), h = [H]$, where Ω is the space of random seeds. At the start of an episode, a seed $\omega \in \Omega$ is drawn. Then, at each time step h, given the current state s_h and the history $(s_1, a_1, \ldots, s_{h-1}, a_{h-1})$, the joint action is generated by $a_h \sim \pi_h\big(\cdot \,\big|\, \omega, (s_1, a_1, \ldots, s_{h-1}, a_{h-1}), s_h\big)$. Because the same ω is used throughout the episode, this mechanism can induce arbitrary correlation across players' choices, For a given correlated policy π, denote by π_{-i} the marginal strategy of all players except i. We then define player i's *best-response value*, against π_{-i} as $V_{i,1}^{\dagger,\pi_{-i}}(s_1) := \sup_{\pi_i'} V_{i,1}^{(\pi_i', \pi_{-i})}(s_1)$, where the supremum is over all (non-Markov) policies for player i. Furthermore, since computing a Nash equilibrium is known to be PPAD-hard in general, the computation of coarse correlated equilibria has become a central focus in the literature. We therefore provide its definition below:

Definition 1 (ε-Coarse Correlated Equilibrium) *A correlated policy π is an ε-approximate coarse correlated equilibrium if, for every player $i \in [N]$, $V_{i,1}^{\pi}(s_1) \geq V_{i,1}^{\dagger,\pi_{-i}}(s_1) - \varepsilon$.*

2.5. Regret and Learning Feedback

We first recall the notion of *external regret*. Let $\mathcal{A}_i$ be the finite action set of player i, and $\Delta(\mathcal{A}_i)$ its probability simplex. At each round t, the agent selects a mixed strategy $x_i^{(t)} \in \Delta(\mathcal{A}_i)$, receives a utility vector $\nu_i^{(t)} \in \mathbb{R}^{|\mathcal{A}_i|}$, and earns payoff $\langle x_i^{(t)}, \nu_i^{(t)} \rangle$. The *external regret* over T rounds is

$$\mathrm{Reg}_i^T := \max_{a_i \in \Delta(\mathcal{A}_i)} \sum_{t=1}^{T} \left[\langle a_i, \nu_i^{(t)} \rangle - \langle x_i^{(t)}, \nu_i^{(t)} \rangle \right], \tag{2}$$

which compares the agent's policy to the best fixed action in hindsight. An algorithm has no external regret if $\mathrm{Reg}_i^T = o(T)$. We now extend this notion to Markov games, where each round t, each player $i \in [N]$ and step $h \in [H]$ observes the expected utility vector, $\nu_{i,h}^{(t)}(s, \cdot) := \left[Q_{i,h}^{(t)} \pi_{-i,h}^{(t)}\right](s, \cdot)$, $\forall s \in \mathcal{S}$. For each (s, h) pair, the *weighted external regret* incurred by player i is defined as

$$\mathrm{reg}_{i,h}^t(s) := \max_{\pi_{i,h}^\dagger \in \Delta(\mathcal{A}_i)} \sum_{j=1}^{t} \alpha_t^j \left\langle \pi_{i,h}^\dagger - \pi_{i,h}^j, \left[Q_{i,h}^{(j)} \pi_{-i,h}^{(j)}\right](s, \cdot) \right\rangle \tag{3}$$

where $\{\alpha_t^j\}_{j=1}^t$ is a set of non-negative weights summing to one, and $\pi_{i,h}^\dagger$ is player i's best response to $\pi_{-i,h}^{(j)}$ at step h. Furthermore, we define the worst-case regret at step h as $\mathrm{reg}_h^t := \max_{i\in[N]} \max_{s\in\mathcal{S}} \mathrm{reg}_h^t(i, s)$. Then, we define the CCE-Gap as the "distance" of a policy to CCE as

$$\text{CCE-Gap}(\bar{\pi}) := \max_{i \in [N]} \left[V_{i,1}^{\dagger,\bar{\pi}_{-i}}(s_1) - V_{i,1}^{\bar{\pi}}(s_1) \right] \tag{4}$$

3. Algorithm and the Main Result

In this section, we first present Algorithm 1, which yields the CCE-gap bound stated in Theorem 4, through the established equivalence between Algorithm 1 and Algorithm 3 for multi-player general-

sum Markov games. Since the algorithms run by all the agents are symmetric, we only illustrate our algorithm using a single agent i. Algorithm 1 consists of three major components: The policy update step that computes the strategy for each matrix game, the value update step that updates the value functions, and the policy output step that generates a CCE policy.

Algorithm 1 Markov-Game DLRC-OMWU with V-Updates

Initialize $d=|A_i|$, $U_{i,h}^{(t=1)}(s,\cdot), u_{i,h}^{(t=0)}(s,\cdot) \leftarrow \mathbf{0}^d$, and $V_{i,h}^{(0)}(s) \leftarrow 0\ \forall h \in [H], \forall s \in \mathcal{S}, i \in [N]$
for $t = 1$ to T **do**
  Policy update:
  $\mathcal{R}_{i,h}^{(t)}(s,\cdot) \leftarrow \eta_t\big(U_{i,h}^{(t)}(s,\cdot) + \frac{w_t}{w_{t-1}} u_{i,h}^{(t-1)}(s,\cdot)\big)$
  $\lambda^{(t)} = \arg\max_{\lambda \in (0,1]} \big[(\alpha-1)\log\lambda + \log \sum_{a_i' \in A_i} e^{\lambda\, \mathcal{R}_{i,h}^{(t)}(s,a_i')}\big]$
  for all $a_i \in A_i$ update policies: $\pi_{i,h}^{(t)}(a_i|s) \leftarrow \frac{e^{\lambda^{(t)} \mathcal{R}_{i,h}^{(t)}(s,a_i)}}{\sum_{a_i' \in A_i} e^{\lambda^{(t)} \mathcal{R}_{i,h}^{(t)}(s,a_i')}}$
  Value update: for $h = H \to 1$:
$$V_{i,h}^{(t)}(s) \leftarrow (1-\alpha_t)\, V_{i,h}^{(t-1)}(s) + \alpha_t \big[(r_{i,h} + P_h V_{i,h+1}^{(t)})\, \pi_h^{(t)}\big](s)$$
  Set: $u_{i,h}^{(t)}(s,\cdot) := w_t \Big[\big[(r_{i,h} + P_h V_{i,h+1}^{(t)})\, \pi_{-i,h}^{(t)}\big](s,\cdot) - V_{i,h}^{(t)}(s)\mathbf{1}_d \Big]$
  Update utility vector: $U_{i,h}^{(t+1)}(s,\cdot) = U_{i,h}^{(t)}(s,\cdot) + u_{i,h}^{(t)}(s,\cdot)$
end for
return policy $\bar{\pi} = \bar{\pi}_1^T$, where $\bar{\pi}_h^t$ specified in Algorithm 2

Algorithm 2 Roll-out procedure $\bar{\pi}_h^t$ for evaluation

Require: policy stream $\{\pi_h^t\}_{h\in[H],t\in[T]}$ from Algorithm 1
for $h' = h, \ldots, H$ **do**
  Sample $j \in [T]$ with probability $\mathbb{P}(j=i) = \alpha_T^i$
  Execute policy $\pi_{h'}^j$ at step h'
  Play policy $\bar{\pi}_{h+1}^j$ onward
end for

3.1. Policy Update.

At every state-step pair (s,h), the agents engage in a sequence of matrix games, where, each agents' payoff matrix is determined by the estimates of the V-functions. It is well known that when all players employ no-regret algorithms in NFGs, their time-averaged joint strategy forms a $\frac{\mathrm{Reg}_T}{T}$-approximate CCE Cesa-Bianchi and Lugosi (2006). Thus, for each state-stage pair, we treat the local interaction as a matrix game. Then, we introduce our algorithm, the Markov-Game Dynamic Learning-Rate Control Optimistic Multiplicative Weights Update (MG-DLRC-OMWU). Our proposed algorithm is adapted from the work of Soleymani et al. (2025), which poses equilibrium learning as a learning rate control problem. The underlying idea is penalization of excessively neg-

ative regret, which limits the exceptional actions. This approach is conceptually linked to replicator dynamics in evolutionary game theory Weibull (1997), where updates are based on "harmony".

The core of our algorithm is a variant of OMWU that incorporates an adaptive regularizer. For each agent i, and (s,h), the algorithm maintains two components; a cumulative dual vector $U_{i,h}^{(t)}(s,\cdot)$, and a regret correction vector $u_{i,h}^{(t-1)}(s,\cdot)$. Upon observing $\nu_{i,h}^{(t-1)}(s,\cdot) = \big[(r_h + P_h V_{h+1,i}^{(t-1)})\,\pi_{-i,h}^{(t-1)}\big](s,\cdot)$, we have $u_{i,h}^{(t-1)}(s,\cdot) := w_{t-1}\big(\nu_{i,h}^{(t-1)}(s,\cdot) - \langle \nu_{i,h}^{(t-1)}, \pi_{i,h}^{(t-1)}\rangle(s)\mathbf{1}_d\big)$. Then, we form an optimistic estimate $\mathcal{R}_{i,h}^{(t)}(s,\cdot) := \eta_t\big(U_{i,h}^{(t)} + \frac{w_t}{w_{t-1}} u_{i,h}^{(t-1)}\big)(s,\cdot)$, where $\eta_t := \frac{\eta}{w_t}$, and weights are the value update rates in Algorithm 1. Then, the policy $\pi^{(t)}$ is computed as follows,

$$\pi_{i,h}^{(t)}(a_i|s) := \frac{\exp\left(\lambda^{(t)}\mathcal{R}_{i,h}^{(t)}(s,a_i)\right)}{\sum_{a_i'\in A_i}\exp\left(\lambda^{(t)}\mathcal{R}_{i,h}^{(t)}(s,a_i')\right)} \tag{5}$$

A key point in Algorithm 1 is the dynamic learning-rate control scheme for selecting $\lambda^{(t)}$. This scheme adapts the learning rate based on the magnitude of the optimistic regret. If the rewards are already large, indicating a volatile learning phase between the players, a conservative fixed learning rate η is used. Otherwise, the learning rate is optimized to balance the learning progress against the stability of the updates. More clearly, when the parameter $\tilde{\alpha}$ is chosen to be on the order of $\Theta(\log^2 d + \log d)$, it can be shown that $\lambda^{(t)}$ update at Line 5 in Algorithm 1 is equivalent to:

$$\lambda^{(t)} = \begin{cases} 1, & \text{if } \max_{a_i'\in A_i}\mathcal{R}_{i,h}^{(t)}(s,a_i') \geq -\beta\log|\mathcal{A}_i|, \\ \arg\max\limits_{\lambda\in(0,1]}\left\{(\tilde{\alpha}-1)\log\lambda + \log\sum\limits_{a_i'\in A_i} e^{\lambda\,\mathcal{R}_{i,h}^{(t)}(s,a_i')}\right\}, & \text{otherwise} \end{cases} \tag{6}$$

where β, η_t, are hyperparameters chosen accordingly. For the analysis of the Algorithm 1 we use a weighted time-dependent learning rate schedule within the equivalent OFTRL algorithm, which extends the stationary learning rate analysis in Soleymani et al. (2025). This formulation lets players adapt to the non-stationary dynamics while preserving the regret-minimization principles. The analysis of the equivalent algorithm, Algorithm 3, lets us to provide the RVU inequality in Theorem 6, as the RVU property is key to achieving sublinear regret and ensuring equilibrium convergence.

3.2. Value update.

Each player maintains V value function $V_{i,h}^{(t)}(s)$ for every for every (h,s) pair and conducts smooth value update with the following learning rates: $\alpha_t := \frac{H+1}{H+t}$, proposed by Jin et al. (2018), which guarantees stability across long horizons, and adopted by the wide range of works in the literature Zhang et al. (2022); Yang and Ma (2023); Cai et al. (2024); Mao et al. (2024); Jin et al. (2022) Under this step size, the V-update at round t corresponds to the weighted average:

$$V_{i,h}^{(t)}(s) = \sum_{j=1}^{t}\alpha_t^j \cdot \Big[(r_{i,h} + P_h V_{i,h+1}^{(j)})\pi_h^{(j)}\Big](s) \tag{7}$$

where the time-dependent coefficients α_t^j are defined as $\alpha_t^j := \alpha_j\prod_{k=j+1}^{t}(1-\alpha_k)$, for $j<t$, and $\alpha_t^t := \alpha_t$. This update ensures that $\sum_{j=1}^{t}\alpha_t^j = 1$, so that the estimate remains a proper average. In this work, we also adopt the same weight sequence $\{\alpha_t^j\}$, which is used to construct the utility weights w_j in the MG-DLRC-OMWU procedure via the relation $w_j := \alpha_t^j/\alpha_t^1$ for $j \leq t \in [T]$.

3.3. Policy output.

The final joint policy $\bar{\pi}_h^t$ is constructed by aggregating the history of policies across time using the same weights α_j^t that govern the value updates. Formally, at the initial step $h \in [H]$, we sample an iteration index $j \in [t]$ with probability proportional to α_t^j, and execute the joint policy $\pi_h^{(j)}$ at that step. Subsequently, the process continues by executing $\bar{\pi}_{h+1}^j$ at the next step of the same episode, and proceeds similarly at each following step. This procedure is summarized in Algorithm 2 and to our knowledge it was first proposed in Bai et al. (2020). Since all players sample from the same index j at each step, the resulting policy $\bar{\pi}$ is correlated across players similar to Zhang et al. (2022).

3.4. Analysis

Now, we present Algorithm 3 for ease of analysis, which is shown to be equivalent to Algorithm 1 in Lemmas 2 and 3. Algorithm 3 employs FTRL updates with a specific regularizer, and uses Q-updates equivalent to the V-updates. This enables the use of stability arguments to analyze Algorithm 1 within a FTRL framework. We state the equivalence in policy updates in Lemma 2.

Algorithm 3 MG-DLRC-OFTRL

Initialize: $d=|A_i|$, $U_{i,h}^{(t=1)}(s,\cdot), u_{i,h}^{(t=0)}(s,\cdot) \leftarrow \mathbf{0}^d$, and $Q_{i,h}^{(0)}(s,\cdot) \leftarrow \mathbf{0}^d \; \forall h \in [H], \forall s \in \mathcal{S}, i \in [N]$
Define regularizer: $\Psi^{(t)}(y) = -\tilde{\alpha} \log\big(\sum_{j=1}^d y[j]\big) + \frac{1}{\sum_{j=1}^d y[j]} \sum_{j=1}^d y[j] \log y[j]$
for $t = 1$ **to** T **do**
  Policy update:
  Optimistic signal: $\mathcal{R}_{i,h}^{(t)}(s,\cdot) \leftarrow \eta_t \big(U_{i,h}^{(t)} + \frac{w_t}{w_{t-1}} u_{i,h}^{(t-1)}\big)(s,\cdot)$
  OFTRL update: $y_{i,h}^{(t)}(s,\cdot) = \arg\max_{y \in (0,1]\Delta^d} \big[\langle \mathcal{R}_{i,h}^{(t)}, y\rangle - \Psi^{(t)}(y)\big]$
  Recover policy for all $a_i \in A_i$: $\pi_{i,h}^{(t)}(a_i|s) \;=\; \frac{y_{i,h}^{(t)}(s,a_i)}{\sum_{a_i' \in A_i} y_{i,h}^{(t)}(s,a_i')}$
  Value update: for $h = H \to 1$:
  $Q_{i,h}^{(t)}(s,a) = (1-\alpha_t)\, Q_{i,h}^{(t-1)}(s,a) + \alpha_t \big[r_{i,h} + P_h[Q_{i,h+1}^t \pi_{h+1}^t]\big](s,a)$
  Set: $u_{i,h}^{(t)}(s,\cdot) = w_t \big[\, [Q_{i,h}^{(t)} \pi_{h,-i}^{(t)}](s,\cdot) - [Q_{i,h}^{(t)} \pi_h^{(t)}](s) \mathbf{1}_d\big]$
  Update utility vector: $U_{i,h}^{(t+1)}(s,\cdot) = U_{i,h}^{(t)}(s,\cdot) + u_{i,h}^{(t)}(s,\cdot)$
end for
return policy $\bar{\pi} = \bar{\pi}_1^T$, where $\bar{\pi}_h^t$ specified in Algorithm 2

Lemma 2 (Equivalence of DLRC–OMWU Formulations with Time-Varying Step-Size) *For each agent i, and (s,h), define $\mathcal{R}^{(t)} := \frac{\eta}{w_t}\Big(U^{(t)} + \frac{w_t}{w_{t-1}} u^{(t-1)}\Big)$, $u^{(t)} := w_t\Big(\nu^{(t)} - \langle \nu^{(t)}, \pi_i^{(t)}\rangle \mathbf{1}_d\Big)$, $U^{(t)} := \sum_{\tau=1}^{t-1} u^{(\tau)}$. First, the Lines 5 and 6 of Algorithm 1 are equivalent to (8). Then, with the variable change $y^{(t)} = \lambda^{(t)} x^{(t)}$, $\lambda^{(t)} = \sum_k y^{(t)}[k]$ two optimization problems given below are equivalent:*

$$1. \quad \big(\lambda^{(t)}, x^{(t)}\big) = \arg \max_{\lambda \in (0,1],\, x \in \Delta^d} \Big\{ \lambda \,\langle \mathcal{R}^{(t)}, x\rangle + (\tilde{\alpha} - 1) \log \lambda - \sum_{k=1}^d x[k] \log x[k] \Big\} \tag{8}$$

$$2. \quad y^{(t)} = \arg \max_{y \in (0,1]\Delta^d} \Big\{ \langle \mathcal{R}^{(t)}, y\rangle + \tilde{\alpha} \log\big(\textstyle\sum_k y[k]\big) - \frac{1}{\sum_k y[k]} \sum_{k=1}^d y[k] \log y[k] \Big\} \tag{9}$$

By Lemma 12 and Corollary 13 in Appendix B, the update in (9) is equivalent to a one-dimensional maximization over $\lambda \in (0, 1]$, given in (6), together with the closed-form update in (5). As the objective in (6) is strongly concave by Lemma 14, the update in (9) is efficiently computable.

On the other hand, for the value updates we motivate our V-value in Algorithm 1 as maintaining a state-action value function $Q_{i,h}(s, a)$ requiring storing and updating values for each joint action $a = (a_1, \ldots, a_n)$, leading to a space complexity of $|\mathcal{S}| \cdot \prod_{j=1}^{n} |\mathcal{A}_j|$. This exponential dependence on the number of agents often renders Q-based methods impractical in scaled multi-agent settings. Thus, we maintain a compact state-value function $V_{i,h}(s)$ that depends only on the state s for each player i, requiring only $\mathcal{O}(|\mathcal{S}|)$ space, following the ideas from Zhang et al. (2022); Cai et al. (2024). It is also important to note that this V-based approach also enables a decentralized implementation: as the update for $V_{i,h}(s)$ requires only the expected value of $(r_h + P_h V_{i,h+1})$ under the joint policy of the other players, $\pi_{-i,h}(\cdot \mid s)$. Crucially, this expectation can be obtained without explicitly reconstructing $\pi_{-i,h}$, by interacting with the environment or using standard black-box feedback mechanisms. In particular, we assume access to a *reward oracle* that returns $\mathbb{E}_{a_{-i} \sim \pi_{-i,h}}[r_i(s, a_i, a_{-i})]$, and a *transition oracle* that returns the $\mathbb{E}_{a_{-i} \sim \pi_{-i,h}}[P_h(\cdot \mid s, a_i, a_{-i})]$ for any fixed (s, a_i), or that both can be approximated through sampling in a model-free setting. As a result, each player can update its value function $V_{i,h}$ using only its own local trajectory data, without needing access to the policies of the other agents. This allows the overall learning process to be implemented in a fully decentralized manner. Following this, we state the equivalence between the Q and V-based updates in Lemma 3:

Lemma 3 *The value updates in Algorithm 3 (the Q-update) and Algorithm 1 (the V-update) are equivalent. Specifically, for all $t \in [T]$, $h \in [H]$, $s \in \mathcal{S}$, and $a_i \in \mathcal{A}_i$, the iterates satisfy*

$$Q_{i,h}^t(s, a_i) = r_h(s, a_i) + \left(P_h V_{i,h+1}^t\right)(s, a_i) \quad \text{and} \quad V_{i,h}^t(s) = \left\langle \pi_{i,h}^t(\cdot|s), Q_{i,h}^t(s, \cdot) \right\rangle,$$

3.5. Learning–Rate Selection and Equivalent Optimistic FTRL View

By the proven equivalence of Algorithms 1 and 3, we now state the CCE convergence of Algorithm 3 in multi-player general-sum Markov games, which therefore also holds for Algorithm 1.

Theorem 4 (Regret Bounds for MG-DLRC-OMWU) *Suppose that n players engage in self-play in a general-sum Markov game with an action set of size bounded by $|\mathcal{A}_{\max}|$, over T rounds. If each player follows Algorithm 3, equivalently Algorithm 1, with parameters $\beta \geq 70$, $\tilde{\alpha} = \beta \log^2 |\mathcal{A}_{\max}| + 2 \log |\mathcal{A}_{\max}| + 2$, and $\eta = 1/24H\sqrt{H}N$, then the output policy $\bar{\pi}$ satisfies:*

$$\textit{CCE-Gap}(\bar{\pi}) \leq \frac{864 H^{\frac{7}{2}} N\big(\tilde{\alpha} \log T + 2 \log |\mathcal{A}_{\max}| + 2\big)}{T} \tag{10}$$

Theorem 4 improves the best-known convergence bound to coarse correlated equilibrium (CCE) from $\mathcal{O}((\log T)^5/T)$, as established in Mao et al. (2024), to $\mathcal{O}(\log T/T)$. This matches the optimal convergence rate for CCE in NFGs, in terms of dependence on both the number of actions and the time horizon. Notably, the analysis in Mao et al. (2024) builds upon the work of Daskalakis et al. (2021) on the multiplicative weights update (MWU) algorithm, which introduces a large constant factor on the order of $C \sim 10^8$ and guarantees convergence only when $T \geq C|\mathcal{A}_{\max}|H^4$.

In contrast, Theorem 4 closes the gap between the best-known convergence rates for coarse correlated and correlated equilibria. While the existence of algorithms, which converge to CE, is already known to achieve a rate of $\mathcal{O}(\log T/T)$, our result shows that such rates are also attainable

for the more general CCE set. This aligns with empirical observations reported in Mao et al. (2024), suggesting that faster convergence to CCE is indeed achievable in practice. Also, our results improve mentioned CE convergence bounds in terms of $|\mathcal{A}_{\max}|$. The proof of Theorem 4 follows a similar structure to prior works such as Mao et al. (2024). We begin with defining the quantity, $\delta_h^t := \max_{s\in\mathcal{S},i\in[n]}\big(V_{i,h}^{\dagger,\bar{\pi}^t_{-i,h}}(s) - V_{i,h}^{\bar{\pi}_h^t}(s)\big)$. Then, we have the following recursive bound:

Lemma 5 *For the policy $\bar{\pi}_h^t$, in Algorithm 2, for all $(i,h,t)\in[n]\times[H]\times[T]$ we have*

$$\delta_h^t \le \sum_{j=1}^{t}\alpha_t^j\delta_{h+1}^t + \max_{s\in\mathcal{S},i\in[n]}\mathrm{reg}_{i,h}^t(s), \tag{11}$$

Thus, upper bounding CCE-Gap$(\bar{\pi})$ reduces to controlling the per-state weighted regrets for each player and every $(s,h)\in\mathcal{S}\times[H]$. We next provide a regret bound for each (i,h,s), derived using an RVU-type inequality for the MG-DLRC-OMWU algorithm under time-varying learning rates.

Theorem 6 (RVU bound for MG-DLRC-OMWU with time-varying η_t) *Let $\beta\ge 70$, and $\kappa^{(t)} = \frac{w_t}{w_{t-1}}$. For each (s,h), consider the* inner *OFTRL process in Algorithm 3, with iterates $y^{(t)}, u^{(t)}$ as in Lemma 2. Then, the cumulative regret $\tilde{\mathrm{Reg}}(T) := \max_{y^*\in[0,1]\Delta^d}\sum_{t=1}^{T}\langle y - y^{(t)}, u^{(t)}\rangle$ incurred up to horizon T obeys*

$$\tilde{\mathrm{Reg}}(T) \le 2\|u^{(t)}\|_\infty + \frac{\tilde{\alpha}\log T + 2\log|\mathcal{A}_{\max}|}{\eta_{T+1}} + \sum_{t=1}^{T}\eta_t\|u^{(t)} - \kappa^{(t)}u^{(t-1)}\|_\infty^2 - \frac{1}{20}\sum_{t=1}^{T-1}\frac{\|x^{(t+1)}-x^{(t)}\|_1^2}{\eta_t} \tag{12}$$

Now, using the given Theorem 6 we derive the following regret bound for each (i,h,s).

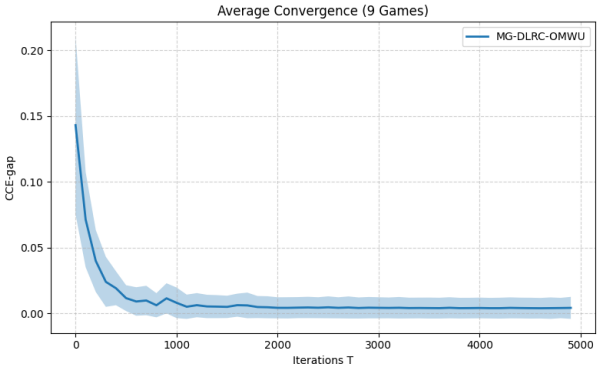
Lemma 7 (Per-state weighted regret bounds) *Fix an episode $h\in[H]$, state $s\in\mathcal{S}$, agent $i\in\mathcal{N}$, and horizon $T\ge 2$. Run Algorithm 3 with a base learning rate $\eta>0$ and weights $w_j = \frac{\alpha_t^j}{\alpha_t^1}$, Then,*

$$\begin{aligned}\mathrm{reg}_{i,h}^t(s) \le{}& \frac{2H\big(\tilde{\alpha}\log t + 2\log|A_{\max}| + 6H\eta\big)}{\eta\, t} + 12\eta\, H^2(N-1)\sum_{j=2}^{t-1}\sum_{k\neq i}\alpha_t^j\big\|\pi_{h,k}^j - \pi_{h,k}^{j-1}\big\|_1^2 \\ &+ \frac{12\eta H^2(3H+4N^2)}{t} - \frac{1}{24\,\eta H}\sum_{j=2}^{t-1}\alpha_t^j\,\big\|\pi_{i,h}^j - \pi_{i,h}^{j-1}\big\|_1^2.\end{aligned} \tag{13}$$

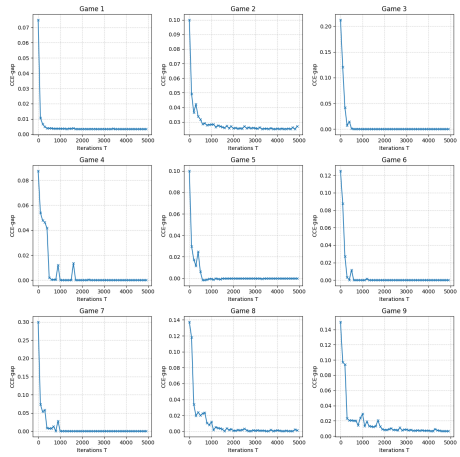
Moreover, summing over all agents and setting $\eta = 1/(24H\sqrt{HN})$ yields:

$$\sum_{i=1}^{N}\mathrm{reg}_{i,h}^t(s) \le \frac{2HN\big(\tilde{\alpha}\log t + 2\log|\mathcal{A}_{\max}| + 6\eta^2 HN(3H+4N^2) + 6H\eta\big)}{\eta t} - \frac{\sum_{i=1}^{N}\sum_{j=2}^{t-1}\alpha_t^j\Big\|[\pi_{i,h}^j - \pi_{i,h}^{j-1}]\Big\|_1^2}{48\eta H} \tag{14}$$

We note a structural distinction between the recursive bound in (11) and the regret guarantee in (14). Specifically, the former requires bounding the *maximum* external regret across agents, while the latter controls only the *sum* of regrets. This is typically problematic since external regret can be negative, and the sum does not necessarily upper bound the maximum. However, in our setting this issue is circumvented by the fact that the RVU bound in Theorem 6 applies to policy iterates $y^{(t)}$ of the OFTRL process, which are guaranteed to yield non-negative regrets due to the structure of the MG-DLRC-OFTRL update. This non-negativity, formally established in Proposition 20 in Appendix B, stands in sharp contrast to general external regret and allows us to upper bound the



(*a*) Trajectory of the average CCE-gap, over 9 games



(*b*) CCE-gap across 9 games

maximum regret by summing the external regret bound over all players. Thus, using the RVU bound we can bound the second order path length and derive the final bound as stated in Theorem 4, which has been stated in Appendix C. This completes the high-level overview of our analysis.

4. Numerical Results

In this section, we numerically evaluate our proposed algorithm MG-DLRC-OMWU on a set of general-sum Markov games. Each environment consists of 2 players, 2 states, and 2 actions per player, with a horizon length of $H = 2$. The rewards are generated within the interval of $[0, 1]$ for each trial, while the transitions are fixed: the system stays in the current state with probability 0.8 and transitions to the other state with probability 0.2. Figure 4-(a) reports the trajectory of the average CCE-gap over 9 independent simulations. In that figure, we show the mean convergence trajectory across 9 games, and the shaded region represents one standard deviation around the mean. In each chosen game, we observe that the CCE-gap converges with a rate of $\mathcal{O}(\log T/T)$. Finally, we show the individual trajectories across the 9 game instances in Figure 4-(b).

5. Conclusion

In this work, we have introduced a policy optimization algorithm that achieves a convergence rate of $\mathcal{O}(\log T/T)$ to the CCE in general-sum Markov games. This result improves upon the best-known convergence rate of $\mathcal{O}((\log T)^5/T)$ for CCE learning, while matching the fastest convergence rate for the CE in general-sum Markov games, both established by Mao et al. (2024). While achieving constant regret remains an open problem in general-sum Markov games, recent advances in the zero-sum setting, Yang and Ma (2023), provide a promising foundation. In particular, ideas inspired by social computation theory have shown that promoting coordination among agents can lead to improved convergence guarantees. Such ideas may be useful to achieve constant regret in general-sum settings as well. Moreover, future directions include improving the convergence rates by enhancing the algorithm's dependence on the horizon H and the action space $|A_{\max}|$. Another extension would be moving from the oracle setting to a sample-based setting where the game parameters must be learned through interaction.

Acknowledgments

Research of the authors was supported in part by the Army Research Office (ARO) Grant Number W911NF-24-1-0085.

References

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

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS '22)*, 2022.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS '20)*, 2020.

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

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006. ISBN 0521841089.

Xin Chen, Guannan Qu, Yujie Tang, Steven Low, and Na Li. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. *IEEE Transactions on Smart Grid*, 13(4):2935–2958, 2022.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. *Near-Optimal No-Regret Algorithms for Zero-Sum Games*, pages 235–254. Society for Industrial and Applied Mathematics, 2011.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS '21)*, 2021.

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

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

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

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Proceedings of the 32th Conference on Neural Information Processing Systems (NeurIPS '18)*, 2018.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent RL. In *ICLR Workshop on Gamification and Multiagent Solutions (GAMMAS '22)*, 2022.

Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2017.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. *Proceedings of 11th International Machine Learning Conference*, pages 157–163, 1994.

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

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, and Tamer Başar. $\widetilde{O}(T^{-1})$ convergence to (coarse) correlated Equilibria in full-information general-sum Markov games. *Proc. Machine Learning Research, 2024 (6th Annual Conf Learning for Dynamics and Control (L4DC), Oxford, England, July 15-17)*, pages 181–296, 2024.

A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983. ISBN 9780471103455.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. MAPoRL: Multi-agent post-co-training for collaborative large language Models with reinforcement learning. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Proceedings of the 26th Conference on Neural Information Processing Systems (NeurIPS '13)*, 2013.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *In Learning, Inference and Control of Multi-Agent Systems Workshop, (NeurIPS'16)*, 2016.

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

Ashkan Soleymani, Georgios Piliouras, and Gabriele Farina. Faster rates for no-regret learning in general games via cautious optimism. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, page 518–529. ACM, June 2025.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations (ICLR '22)*, 2022.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. *Proceedings of the 28th Conference on Neural Information Processing Systems (NeurIPS '15)*, 2015.

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

Ziyu Wan, Yunxiang Li, Xiaoyu Wen, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, and Ying Wen. ReMA: Learning to meta-think for LLMs with multi-agent reinforcement learning. *arXiv preprint arXiv:2503.09501*, 2025.

Jörgen W Weibull. *Evolutionary Game Theory*. MIT Press, USA, 1997. ISBN 9780262731218.

Yuepeng Yang and Cong Ma. $O(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum Markov games. *International Conference on Learning Representations (ICLR '23)*, 2023.

Robert E Schapire Yoav Freund. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory - 2nd European Conference, EuroCOLT 1995, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 23–37. Springer Verlag, 1995.

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

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *In K. G. Vamvoudakis et al. (eds.) Handbook of Reinforcement Learning and Control, Studies in System, Decision and Control 325*, pages 321–384, 2021.

Runyu Zhang, Qinghua Liu, Huan Wang, Caiming Xiong, Na Li, and Yu Bai. Policy optimization for Markov games: Unified framework and faster convergence. *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS '22)*, 2022.

Appendix A. Technical Tools and Their Proofs

Lemma 3 (Equivalence of V and Q-Updates)

Algorithms 3 (Q-based) and 1 (V-based) generate the identical policy sequence $\{\pi_h^t\}(s,\cdot)$. *Equivalently, for all agents* i, *states* s, *steps* h, *actions* a, *and rounds* t, *we have*

$$Q_{i,h}^t(s,a) \;=\; r_{i,h}(s,a) \;+\; \big[P_h\, V_{i,h+1}^t\big](s,a). \tag{15}$$

Proof The base case $t=0$ holds by initialization, and for $t=1$, as we have $\alpha_1=1$, equality holds by definition. Furthermore, it can also be seen that we have $Q_{i,H}^t(s,a)=r_{i,H}(s,a)$. Then, suppose that (15) holds for all rounds up to $t-1$ and all levels $\geq h+1$. Then, writing the Q-value update from Algorithm 3,

$$Q_{i,h}^t(s,a) \;=\; (1-\alpha_t)\, Q_{i,h}^{t-1}(s,a) \;+\; \alpha_t\Big(r_{i,h}(s,a) + P_h\,[Q_{i,h+1}^t\,\pi_{h+1}^t](s,a)\Big).$$

By induction on $(t-1,h)$ and on $(t,h+1)$, each occurrence of $Q_{i,h+1}^t$ is replaced by $Q_{i,h+1}^t(s,a) = r_{i,h+1}(s,a) + [P_{h+1}\, V_{i,h+2}^t](s,a)$, and thus we have;

$$\begin{aligned}
Q_{i,h}^t(s,a) &= (1-\alpha_t)[r_{i,h} + P_h V_{i,h+1}^{t-1}](s,a) + \alpha_t(r_{i,h} + P_h[(r_{i,h+1} + P_{h+1}V_{i,h+2}^t)\pi_{h+1}^t]) \\
&= \Big[r_{i,h} + P_h\Big((1-\alpha_t)V_{i,h+1}^{t-1} + \alpha_t\big[\big(r_{i,h+1} + P_{h+1}(V_{i,h+2}^t)\big)\pi_{h+1}^t\big]\Big)\Big] \\
&= r_{i,h}(s,a) \;+\; \big[P_h\, V_{i,h+1}^t\big](s,a),
\end{aligned}$$

where, the final step is due to the update rule on the V-values in Algorithm 1. This closes the inductive step. Hence the proof is complete. ■

Now, we establish following the two lemmas, for the recursive regret bounds.

Lemma 8 (Equivalence of value functions) *For Algorithm 2, we have for all players* $i\in[m]$ *and all* $(h,s,t)\in[H+1]\times S\times[T]$, *that,* $V_{i,h}^t(s) = V_{i,h}^{\bar\pi_h^t}(s)$.

Proof We prove this by backward induction on $h\in[H+1]$. The claim trivially holds for the base case $h=H+1$, since all values are zero. Now, suppose that the claim holds for step $h+1$ and all $(s,t)\in S\times[T]$. For step h and any fixed (s,t), we have:

$$\begin{aligned}
V_{i,h}^t(s) &= \sum_{j=1}^t \alpha_t^j \left\langle Q_{i,h}^j, \pi_h^j \right\rangle(s) && \text{(i)} \\
&= \sum_{j=1}^t \alpha_t^j \left\langle r_h + P_h V_{i,h+1}^j, \pi_h^j \right\rangle(s) \\
&= \sum_{j=1}^t \alpha_t^j \left\langle r_h + P_h V_{i,h+1}^{\bar\pi_{h+1}^j}, \pi_h^j \right\rangle(s) && \text{(ii)} \\
&= V_{i,h}^{\bar\pi_h^t}(s), && \text{(iii)}
\end{aligned}$$

where the first and third steps follow from definition and the second step is due to the induction step. This proves the claim for step h and thus completes the proof by induction. ■

Lemma 5 *For the policy* $\bar\pi_h^t$ *defined in Algorithm 2, we have, for all* $(i,h,t)\in[n]\times[H]\times[T]$, *that the CCE gap is bounded recursively:*

$$\max_{s\in\mathcal{S},i\in[n]}\left[V_{i,h}^{\dagger,\bar\pi_{-i,h}^t}(s) - V_{i,h}^{\bar\pi_h^t}(s)\right] \leq \sum_{j=1}^t \alpha_t^j \max_{s'\in\mathcal{S},i\in[n]}\left[V_{i,h+1}^{\dagger,\bar\pi_{-i,h+1}^j}(s') - V_{i,h+1}^{\bar\pi_{h+1}^j}(s')\right] + \max_{s\in\mathcal{S},i\in[n]} \mathrm{reg}_{i,h}^t(s).$$

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

$$V_{i,h}^{\dagger,\bar{\pi}^t_{-i,h}}(s) - V_{i,h}^{\bar{\pi}^t_h}(s) = \max_{\pi^\dagger_{i,h}\Delta(\mathcal{A}_i)} \sum_{j=1}^t \alpha_t^j \mathbb{E}_{\pi^\dagger_{i,h}\times\pi^j_{-i,h}}\left[r_h + P_h V_{i,h+1}^{\dagger,\bar{\pi}^j_{-i,h+1}}\right](s) - \sum_{j=1}^t \alpha_t^j \mathbb{E}_{\pi^j_h}\left[r_h + P_h V_{i,h+1}^{\bar{\pi}^j_{h+1}}\right](s)$$

$$\leq \sum_{j=1}^t \alpha_t^j \max_{s'\in S}\left[V_{i,h+1}^{\dagger,\bar{\pi}^j_{-i,h+1}}(s') - V_{i,h+1}^{\bar{\pi}^j_{h+1}}(s')\right] + \max_{\pi^\dagger_{i,h}\Delta(\mathcal{A}_i)} \sum_{j=1}^t \alpha_t^j \left\langle \pi^\dagger_{i,h}(\cdot|s) - \pi^j_{i,h}(\cdot|s), \left[(r_h + P_h V_{i,h+1}^{\bar{\pi}^j_{h+1}})\pi^j_{-i,h}\right](s,\cdot)\right\rangle$$

$$= \sum_{j=1}^t \alpha_t^j \max_{s'\in S}\left[V_{i,h+1}^{\dagger,\bar{\pi}^j_{-i,h+1}}(s') - V_{i,h+1}^{\bar{\pi}^j_{h+1}}(s')\right] + \underbrace{\max_{\pi^\dagger_{i,h}\Delta(\mathcal{A}_i)} \sum_{j=1}^t \alpha_t^j \left\langle \pi^\dagger_{i,h} - \pi^j_{i,h}, \left[(r_h + P_h V^j_{i,h+1})\pi^j_{-i,h}\right]\right\rangle}_{\mathrm{reg}^t_{i,h}(s)}$$

$$\leq \sum_{j=1}^t \alpha_t^j \max_{s'\in S}\left[V_{i,h+1}^{\dagger,\bar{\pi}^j_{-i,h+1}}(s') - V_{i,h+1}^{\bar{\pi}^j_{h+1}}(s')\right] + \mathrm{reg}^t_{i,h}(s).$$

Taking $\max_{s\in\mathcal{S},i\in[n]}$ of both sides concludes the proof. ■

We next present some basic algebraic properties of the weights $\alpha_t = \frac{H+1}{H+t}$, $\{\alpha_t^i\}_{t\geq 1,\, 1\leq i\leq t}$ and $\{w_t\}_{t\geq 1}$, which will be used in later proofs. We define them as:

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

Lemma 9 *Let $H \geq 1$, and for each $t \geq 1$ let $\alpha_t > 0$ be the step–size above. Then, for every integer $T \geq 1$, the following properties hold:*

1. $\sum_{j=1}^T \alpha_T^j = 1.$
2. *The sequence $j \mapsto \alpha_T^j$ is non-decreasing in j.*
3. $\sum_{j=1}^T (\alpha_T^j)^2 \;\leq\; \sum_{j=1}^T (\alpha_j)^2 \;\leq\; H+2.$
4. *For any non-increasing sequence $\{b_j\}_{j=1}^T$, $\sum_{j=1}^T \alpha_T^j\, b_j \;\leq\; \frac{1}{T}\sum_{j=1}^T b_j$.*
5. $\alpha_t^1 \leq \frac{1}{t}$
6. $\sum_{j=1}^T \frac{\alpha_T^j}{j} \;\leq\; \frac{1+\frac{1}{H}}{T}.$
7. $\sum_{j=1}^T \alpha_T^j\,(\alpha_j)^2 \;\leq\; \frac{3H}{T}.$

Proof Proof of the first five properties can be found in Lemma 6 of Yang and Ma (2023). We prove the fifth property as follows. First recall the closed form

$$\alpha_t^j = \;(H+1)\,\frac{(t-1)!\,(H+j-1)!}{(j-1)!\,(H+t)!}, \qquad 1 \leq j \leq t.$$

Hence

$$\begin{aligned}\sum_{j=1}^{t}\frac{\alpha_t^j}{j} &= (H+1)\,\frac{(t-1)!}{(H+t)!}\sum_{j=1}^{t}\frac{(H+j-1)!}{(j-1)!\,j}\\ &= (H+1)\,\frac{(t-1)!}{(H+t)!}(H-1)!\sum_{j=1}^{t}\binom{H+j-1}{j}\\ &= (H+1)\,\frac{(t-1)!\,(H-1)!}{(H+t)!}\left[\binom{H+t}{t}-1\right]\\ &= \frac{H+1}{H\,t}\;-\;(H+1)\,\frac{(t-1)!\,(H-1)!}{(H+t)!}\;\le\;\frac{1+\frac{1}{H}}{t}.\end{aligned}$$

where we used "hockey-stick" identity in the third step. For the last property, we know that the sequence $\{(\alpha_j)^2\}$ is non-increasing. Using the fourth property, with $b_j = (\alpha_j)^2$, along with the third property, yields

$$\sum_{j=1}^{t}\alpha_t^j(\alpha_j)^2 \;\le\; \frac{1}{t}\sum_{j=1}^{t}(\alpha_j)^2 \;\le\; \frac{H+2}{t}.$$

For all $H \ge 1$, one has $H+2 \le 3H$, and thus $\sum_{j=1}^{t}\alpha_t^j(\alpha_j)^2 \;\le\; \frac{3H}{t}$. ■

Lemma 10 (Pinsker's Inequality) *For discrete distributions p, q on support size d, we have the* $\|p-q\|_1^2 \;\le\; 2\,\mathrm{KL}(p\|q)$.

Lemma 11 (Entropy difference) *For discrete random variables p, q on support size d, we have* $\big|H(p)-H(q)\big| \;\le\; (\log d)\,\sqrt{2\,\mathrm{KL}(p\|q)}$.

Appendix B. Proof of RVU bound with time-varying learning rates

In this appendix, we present time-varying analogues of the lemmas used to derive the RVU bound in Soleymani et al. (2025), to ensure completeness. For a more detailed exposition and the corresponding results under a constant step-size η, we refer the reader to Soleymani et al. (2025). Throughout the rounds $t \in [T]$, we allow the learning-rate cap to vary, writing $\eta_t \in (0,1]$. Also, as we use the mentioned RVU bound to upper bound regret for all (s,h) pairs, we do not explicitly denote them throughout the appendix. Furthermore, we let $|\mathcal{A}_i| = d$. The optimistic FTRL step in lifted coordinates therefore uses the regularizer

$$\psi(y) \;:=\; -\tilde{\alpha}\,\log\Big(\Lambda(y)\Big) \;+\; \frac{1}{\Lambda(y)}\sum_{k=1}^{d} y[k]\log y[k], \qquad \Lambda(y) := \sum_{k=1}^{d} y[k], \tag{16}$$

where $\tilde{\alpha} = \beta\log^2 d + 2\log d + 2$ and the update rule

$$y^{(t)} \;=\; \arg\max_{y\in(0,1]\Delta^d}\Big\{\langle r^{(t)},\,y\rangle - \psi(y)\Big\}. \tag{17}$$

Also, we define

$$\mathcal{R}^{(t)} := \frac{\eta}{w_t}\Big(U^{(t)} + \frac{w_t}{w_{t-1}}u^{(t-1)}\Big), \quad u^{(t)} := w_t\Big(\nu^{(t)} - \langle\nu^{(t)}, x^{(t)}\rangle\mathbf{1}_d\Big), \quad U^{(t)} := \sum_{\tau=1}^{t-1}u^{(\tau)}. \tag{18}$$

Furthermore, we have $\|\nu^{(t)}\|_\infty \leq H$, due to $\nu^{(t)} = Q^t_{i,h}\pi^t_{-i}$, and $x^{(t)} = \pi^{(t)}_{i,h}$ due to Algorithm 1. Finally, we define sets, $\Delta_d := \{ x \in \mathbb{R}^d_{\geq 0} : \langle \mathbf{1}, x\rangle = 1 \}$, $(0,1]\Delta_d := \{ y \in \mathbb{R}^d_{+} : \langle \mathbf{1}, y\rangle \leq 1 \}$, and $[0,1]\Delta_d := \{ y \in \mathbb{R}^d_{\geq 0} : \langle \mathbf{1}, y\rangle \leq 1 \}$.

Since the RVU bound is derived using aspects from both Algorithm 1 and 3, we first establish the equivalence of the policy update steps between Algorithms 1 and 3. To this end, we prove Lemma 2 in Lemma 12 and Corollary 13, which formally demonstrate the equivalence of their policy update procedures.

Lemma 12 (Equivalence of DLRC-OMWU Formulations with Time-Varying Step-Size) *The following two optimization problems are equivalent:*

1. ***DLRC–OFTRL in*** (λ, x)***-space***

$$\left(\lambda^{(t)}, x^{(t)}\right) \;=\; \arg\max_{\substack{\lambda\in(0,1]\\ x\in\Delta^d}} \left\{ \lambda\langle \mathcal{R}^{(t)}, x\rangle + (\tilde{\alpha}-1)\log\lambda - \textstyle\sum_{k=1}^d x[k]\log x[k] \right\}. \tag{19}$$

2. ***Lifted optimistic FTRL in*** y***-space***

$$y^{(t)} \;=\; \arg\max_{y\in(0,1]\Delta^d} \left\{ \langle \mathcal{R}^{(t)}, y\rangle + \tilde{\alpha}\log\big(\textstyle\sum_k y[k]\big) - \frac{1}{\sum_k y[k]} \sum_{k=1}^d y[k]\log y[k] \right\}, \tag{20}$$

with the variable change $y^{(t)} = \lambda^{(t)}x^{(t)}$ *and* $\lambda^{(t)} = \sum_k y^{(t)}[k]$.

Proof *Let* $y = \lambda x$. *Then,* $\sum_k y[k] = \lambda$ *and* $y \in (0,1]\Delta^d \iff \lambda \in (0,1], x \in \Delta^d$. *By direct algebra,*

$$\begin{aligned}
\langle \mathcal{R}^{(t)}, y\rangle + \tilde{\alpha}\log\big(\textstyle\sum_k y[k]\big) - \frac{1}{\sum_k y[k]}\sum_k y[k]\log y[k] &= \langle \mathcal{R}^{(t)}, \lambda x\rangle + \tilde{\alpha}\log\lambda - \frac{1}{\lambda}\sum_k (\lambda x[k])\log(\lambda x[k]) \\
&= \lambda\langle \mathcal{R}^{(t)}, x\rangle + \tilde{\alpha}\log\lambda - \sum_k x[k]\big(\log\lambda + \log x[k]\big) \\
&= \lambda\langle \mathcal{R}^{(t)}, x\rangle + (\tilde{\alpha}-1)\log\lambda - \sum_k x[k]\log x[k].
\end{aligned}$$

Thus, under the bijection $y = \lambda x$, *the two problems are equivalent.* ■

Corollary 13 (Softmax Structure and Learning Rate Maximization) *Let* $(\lambda^{(t)}, x^{(t)})$ *be the solution to the DLRC–OFTRL problem in Lemma 12. Then:*

(i) The policy $x^{(t)}$ *is given by a softmax:*

$$x^{(t)}[k] = \frac{\exp\left(\lambda^{(t)}\mathcal{R}^{(t)}[k]\right)}{\sum_{j=1}^d \exp\left(\lambda^{(t)}\mathcal{R}^{(t)}[j]\right)}. \tag{21}$$

(ii) The learning rate $\lambda^{(t)}$ *is the solution to the following univariate maximization:*

$$\lambda^{(t)} = \arg\max_{\lambda\in(0,1]} \left\{ \log\left(\sum_{k=1}^d \exp\left(\lambda\mathcal{R}^{(t)}[k]\right)\right) + (\tilde{\alpha}-1)\log\lambda \right\}. \tag{22}$$

Proof

(i) Fix t and write down the Lagrangian of (19) with multiplier μ for the simplex constraint $\sum_k x[k] = 1$ as:

$$\mathcal{L}(\lambda, x, \mu) = \lambda\langle \mathcal{R}^{(t)}, x\rangle + (\tilde{\alpha} - 1)\log\lambda - \sum_k x[k]\log x[k] + \mu\Big(1 - \textstyle\sum_k x[k]\Big).$$

KKT stationarity in x gives $\lambda^{(t)}\mathcal{R}^{(t)}[k] - \log x^{(t)}[k] - 1 - \mu = 0$, and thus $x^{(t)}[k] \propto \exp\big(\lambda^{(t)}\mathcal{R}^{(t)}[k]\big)$. Normalizing over k yields the soft-max form (21).

(ii) Given $\lambda^{(t)}$ from (19), plugging the soft-max expression (21) back into the objective of (19) and maximizing over λ recovers exactly the univariate problem (22), which is the dynamic learning-rate rule of MG-DLRC-OMWU. This completes the proof.

■

Now, before proceeding with the RVU analysis, we provide the following Lemma 14 to argue proposed policy iterations are efficiently computable.

Lemma 14 *For any fixed reward vector $R \in \mathbb{R}^d$, define*

$$g_{\mathcal{R}}(\lambda) := (\tilde{\alpha} - 1)\log\lambda + \log\Big(\sum_{k=1}^{d} e^{\lambda\mathcal{R}^{(t)}[k]}\Big), \qquad \lambda \in (0, 1].$$

Suppose $\tilde{\alpha} > 1 + \log^2 d$. Then the $g_{\mathcal{R}}$ is strongly concave on $(0, 1]$; in particular,

$$g''_{\mathcal{R}}(\lambda) \le -\frac{\tilde{\alpha} - 1 - \log^2 d}{\lambda^2} < 0, \qquad \forall\lambda \in (0, 1].$$

Proof The derivative formulas are

$$g'_{\mathcal{R}}(\lambda) = \frac{\sum_{k=1}^{d}\mathcal{R}^{(t)}[k]e^{\lambda\mathcal{R}^{(t)}[k]}}{\sum_{k=1}^{d} e^{\lambda\mathcal{R}^{(t)}[k]}} + \frac{\tilde{\alpha} - 1}{\lambda},$$

and

$$\frac{\partial^2 f(\lambda; \mathcal{R})}{\partial\lambda^2} = \frac{\sum_{k=1}^{d}\mathcal{R}^{(t)}[k]^2 e^{\lambda\mathcal{R}^{(t)}[k]}}{\sum_{k=1}^{d} e^{\lambda\mathcal{R}^{(t)}[k]}} - \left(\frac{\sum_{k=1}^{d}\mathcal{R}^{(t)}[k]e^{\lambda\mathcal{R}^{(t)}[k]}}{\sum_{k=1}^{d} e^{\lambda\mathcal{R}^{(t)}[k]}}\right)^2 - \frac{\alpha - 1}{\lambda^2}. \tag{23}$$

Define

$$x[k] := \frac{e^{\lambda\mathcal{R}^{(t)}[k]}}{\sum_{j=1}^{d} e^{\lambda\mathcal{R}^{(t)}[j]}}, \qquad \Gamma := \sum_{k=1}^{d} e^{\lambda\mathcal{R}^{(t)}[k]}.$$

Then $x \in \Delta^d$ and

$$\mathcal{R}^{(t)}[k] = \frac{1}{\lambda}\big(\log x[k] + \log\Gamma\big).$$

Substituting, we obtain

$$\frac{\partial^2 f(\lambda; \mathcal{R})}{\partial\lambda^2} = \frac{1}{\lambda^2}\left[\sum_{k=1}^{d} x[k](\log x[k] + \log\Gamma)^2 - \Big(\sum_{k=1}^{d} x[k](\log x[k] + \log\Gamma)\Big)^2\right] - \frac{\alpha - 1}{\lambda^2} \tag{24}$$

$$= \frac{1}{\lambda^2}\left[\sum_{k=1}^{d} x[k]\log^2 x[k] - \Big(\sum_{k=1}^{d} x[k]\log x[k]\Big)^2\right] - \frac{\alpha - 1}{\lambda^2}. \tag{25}$$

Then, this yields $\frac{\partial^2 f(\lambda;\mathcal{R})}{\partial\lambda^2} \le \frac{1}{\lambda^2}\big(\log^2 d - (\alpha - 1)\big)$. which is strictly negative under $\tilde{\alpha} > 1 + \log^2 d$. Hence g_R is strongly concave on $(0, 1]$. ■

Now, as the equivalence between policy update steps of Algorithms 1 and 3 has proven, we proceed with the analysis.

Theorem 15 (Sensitivity of learning rates on regrets) *There exists a universal constant $\beta \geq 70$ such that for $\eta = \frac{1}{24H\sqrt{HN}}$, $\tilde{\alpha} \geq 2+2\log d+\beta \log^2 d$, the following property holds. Let $\mathcal{R}, \mathcal{R}' \in \mathbb{R}^d$ be such that $\|\mathcal{R} - \mathcal{R}'\|_\infty \leq 2H\eta$, and let $\hat{\lambda}, \hat{\lambda}'$ be the corresponding learning rates defined as*

$$\hat{\lambda} = \arg\max_{t\in(0,1]} f(t;\mathcal{R}), \qquad \hat{\lambda}' = \arg\max_{t\in(0,1]} f(t;\mathcal{R}'),$$

where the function f is given by $f(\lambda;\mathcal{R}) := (\tilde{\alpha}-1)\log\lambda + \log\left(\sum_{k=1}^d e^{\lambda\mathcal{R}[k]}\right)$. Then, $\hat{\lambda}$ and $\hat{\lambda}'$ are multiplicatively stable; specifically,

$$\frac{7}{10} \leq \frac{\hat{\lambda}}{\hat{\lambda}'} \leq \frac{7}{5}.$$

Proof The result follows directly by extending Theorem 3.5 and the preceding lemmas from Soleymani et al. (2025), where the reward signals satisfy the uniform bound $\|\nu^{(t)}\|_\infty \leq 1$, to our case with the bounds $\|\nu^{(t)}\|_\infty \leq H$, and step size η/w_t. ■

Theorem 16 (Strong convexity of the time–varying regularizer) *Fix $d \geq 2$ and set $\alpha = 2 + 2\log d + \beta\log^2 d$ with $\beta \geq 70$. For every round t and every $y \in (0,1]\Delta^d$, the Hessian of (16) satisfies*

$$\nabla^2\psi(y) \;\succeq\; \frac{1}{2}\,\mathrm{diag}\Big(\tfrac{1}{y[1]\,\Lambda(y)}, \ldots, \tfrac{1}{y[d]\,\Lambda(y)}\Big), \tag{26}$$

Proof Write $x[k] = y[k]/\Lambda(y) \in \Delta^d$. The first-order partial derivative of (16) is

$$\frac{\partial\psi}{\partial y[i]} = -\frac{\tilde{\alpha}}{\Lambda(y)} - \frac{1}{\Lambda(y)^2}\sum_k y[k]\log y[k] + \frac{1+\log y[i]}{\Lambda(y)}.$$

Differentiating again gives, for every $i, j \in [d]$,

$$\frac{\partial^2\psi}{\partial y[i]\partial y[j]} = \frac{\tilde{\alpha} - 2 + 2\sum_k x[k]\log x[k]}{\Lambda(y)^2} \;-\; \frac{\log x[i] + \log x[j]}{\Lambda(y)^2} \;+\; \frac{\mathbf{1}_{i=j}}{y[i]\Lambda(y)}.$$

Substitute $\tilde{\alpha} = 2 + 2\log d + \alpha'$ with $\alpha' \geq 2\log^2 d$. For any vector $v \in \mathbb{R}^d$,

$$\Lambda(y)^2\, v^\top\nabla^2\psi(y)\, v \;\geq\; \alpha'\big(\textstyle\sum_k v[k]\big)^2 \;+\; \sum_k \frac{v[k]^2}{x[i]} \;-\; 2\Big(\sum_k v[k]\log x[i]\Big)\Big(\sum_j v[j]\Big).$$

The final mixed term is controlled by $-2(\sum_k v[k]\log x[i])(\sum_k v[k]) \geq -2\log^2 d\,(\sum_k v[k])^2 - \sum_k \frac{v[k]^2}{2x[i]}$, and this is exactly absorbed by the choice $\alpha' \geq 2\log^2 d$. Hence, $\Lambda(y)^2\, v^\top\nabla^2\psi(y)v \geq \sum_k \frac{v[k]^2}{2x[i]}$, which is equivalent to (26), which concludes the proof. ■

Now, for the rest of this appendix section, we introduce the following notation. For any $y, z \in (0,1]\Delta^d$, we define $x[k] = \frac{y[k]}{\Lambda(y)}$, $\theta[k] = \frac{z[k]}{\Lambda(z)}$, $\rho := \frac{\Lambda(z)}{\Lambda(y)}$.

Proposition 17 (Decomposition of the Time–Varying Bregman Divergence) *The Bregman divergence induced by ψ satisfies*

$$D_\psi(z\,\|\,y) = (\tilde{\alpha}-1)\,D_{\log}(\Lambda(z)\|\Lambda(y)) \;+\; \rho\,KL\big(\theta\,\|\,x\big) \;+\; (1-\rho)\,\big[\,H(\theta) - H(x)\big]\,, \tag{27}$$

where $D_{\log}(u\|v) = \log\frac{v}{u} + \frac{u}{v} - 1$ is the log–regularizer divergence, $KL(\theta\|x) = \sum_k \theta[k]\log\frac{\theta[k]}{x[k]}$ is the Kullback–Leibler divergence, and $H(x) = -\sum_k x[k]\log x[k]$ is the entropy.

Proof Write the gradient of ψ:

$$\frac{\partial \psi}{\partial y[i]} = -\frac{\tilde{\alpha}-1}{\Lambda(y)} - \frac{1}{\Lambda(y)^2}\sum_k y[k]\log y[k] + \frac{\log y[i]}{\Lambda(y)}.$$

Using $x[i] = y[i]/\Lambda(y)$ and $\Lambda(y) > 0$,

$$\frac{\partial \psi}{\partial y[i]} = -\frac{\tilde{\alpha}-1}{\Lambda(y)} - \frac{1}{\Lambda(y)}\sum_k x[k]\log x[k] + \frac{\log x[i]}{\Lambda(y)}.$$

By definition,

$$D_\psi(z\|y) = \big[\psi(z) - \psi(y)\big] - \sum_i \frac{\partial \psi}{\partial y[i]}\,(z[i]-y[i]).$$

Insert the explicit forms of ψ and its gradient, factor out $\Lambda(y)$, and rearrange terms; after straightforward algebra one obtains

$$D_\psi(z\|y) = (\tilde{\alpha}-1)(\rho-1) - (\tilde{\alpha}-1)\log\rho + (\rho-1)\sum_k x[k]\log x[k] + \sum_k \theta[k]\log\theta[k] - \rho\sum_k \theta[k]\log x[k].$$

Then, adding and subtracting $\rho\sum_k \theta[k]\log\theta[k]$, and grouping the terms accordingly we get:

$$D_\psi(z\|y) = (\tilde{\alpha}-1)D_{\log}(\Lambda(z)\|\Lambda(y)) + \frac{\Lambda(z)}{\Lambda(y)}\mathrm{KL}(\theta\|x) + (1-\rho)(H(\theta)-H(x))$$

where, $D_{\log}$ is the Bregman divergence induced by $-\log(x)$ function due to its strict convexity. Thus, the proof is concluded. ■

Proposition 18 (Strong convexity on the lifted simplex) *For all $y, z \in (0,1]\Delta^d$,*

$$D_\psi(y\,\|\,z) \;\geq\; \frac{1}{2}\,\|y-z\|_1^2. \tag{28}$$

Proof By Theorem 16, for any $\nu \in \mathbb{R}^d$,

$$\begin{aligned}\nu^\top \nabla^2\psi(y)\,\nu \;&\geq\; \frac{1}{2}\sum_{i=1}^d \frac{\nu_i^2}{y_i\,\Lambda(y)} \\ &\geq\; \frac{1}{2}\sum_{i=1}^d \frac{\nu_i^2}{y_i} \geq \frac{\Lambda(y)}{2}\sum_{i=1}^d \frac{\nu_i^2}{y_i}\end{aligned}$$

Since $0 < \Lambda(y) \leq 1$, and by Cauchy–Schwarz we get,

$$\frac{\Lambda(y)}{2}\sum_{i=1}^d \frac{\nu_i^2}{y_i} \;\geq\; \frac{1}{2}\Big(\sum_{i=1}^d |\nu_i|\Big)^2 = \frac{1}{2}\|\nu\|_1^2.$$

Hence, $\nu^\top\nabla^2\psi(y)\,\nu \;\geq\; \frac{1}{2}\|\nu\|_1^2$, and ψ is $(\frac{1}{2})$–strongly convex w.r.t. $\|\cdot\|_1$. The result follows from

$$D_\psi(y\|z) = \psi(y) - \psi(z) - \langle\nabla\psi(z), y-z\rangle \geq \frac{1}{2}\|y-z\|_1^2$$

■

Proposition 19 (Curvature on the action simplex under mass stability) *Let $y, z \in (0,1]\Delta^d$ with masses $\rho := \Lambda(z)/\Lambda(y) \in [1-\varepsilon, 1+\varepsilon]$, where $\varepsilon \in (0, \frac{2}{5})$. Then, for every $t \geq 1$,*

$$D_\psi(z\,\|\,y) \;\geq\; \frac{1-\varepsilon}{4}\,\|\theta - x\|_1^2.$$

Proof By Proposition 17,

$$D_\psi(z\|y) = (\tilde{\alpha}-1)\, D_{\log}(\Lambda(z)\|\Lambda(y)) \;+\; \rho\,\mathrm{KL}(\theta \,\|\, x) \;+\; (1-\rho)\big[H(\theta) - H(x)\big].$$

Then, we write:

$$\begin{aligned}
D_\psi(z\,\|\,y) \;&\geq\; \beta\log^2 d\left(\log\left(\frac{1}{\rho}\right)+\rho-1\right) + (1-\rho)\big(H(\theta)-H(x)\big) + \rho\,\mathrm{KL}(\theta\,\|\,x)\\
&\geq \frac{1}{4}\beta\log^2 d\left(1-\frac{1}{\rho}\right)^2 - \rho\Big|1-\frac{1}{\rho}\Big|\log d\,\sqrt{2\mathrm{KL}(\theta\|x)} + \frac{2\rho^2}{\beta}\mathrm{KL}(\theta\,\|\,x) + \left(\rho-\frac{2\rho^2}{\beta}\right)\mathrm{KL}(\theta\|x)\\
&\geq \left(\frac{1}{2}\sqrt{\beta}\log d\left|1-\frac{1}{\rho}\right| \;-\; \rho\sqrt{\frac{2\mathrm{KL}(\theta\,\|\,x)}{\beta}}\right)^2 + \frac{\rho}{2}\,\mathrm{KL}(\theta\,\|\,x)\\
&\geq \frac{1}{4}(1-\epsilon)\,\|\theta-x\|_1^2,
\end{aligned}$$

where the first step is due to the definition of D_{log}. The second step is due to fact $\log(\frac{1}{\rho})+\rho-1 \geq (1-\frac{1}{\rho})^2$ for $\rho \in [1-\epsilon, 1+\epsilon]$ and Lemma 11 which implies $(1-\rho)[H(\theta)-H(x)] \geq -|1-\rho|\log d\sqrt{2\mathrm{KL}(\theta\|\,x)}$. The third step is due to the fact that $(\rho - \frac{2\rho^2}{\beta}) > \frac{\rho}{2}$, when $\beta \geq 20$. The last step is due to Lemma 10. ■

To analyze MG-DLRC-OMWU equivalently, we start the analysis by taking a closer look at (20). To analyze Reg_T, defined in (2), we first study the nonnegative regret defined by $\tilde{\mathrm{Reg}}(T) := \max_{y^*\in[0,1]\Delta^d}\sum_{t=1}^T\langle u^{(t)}, y^* - y^{(t)}\rangle$.

Proposition 20 *For any time horizon $T \in \mathbb{N}$, we have $\tilde{\mathrm{Reg}}(T) = \max\{0, \mathrm{Reg}(T)\}$. As a result, $\tilde{\mathrm{Reg}}(T) \geq 0$ and $\tilde{\mathrm{Reg}}(T) \geq \mathrm{reg}^T_{i,h}(s)$.*

Proof By definition of the reward signal $u^{(t)} = w_t\Big(\nu^{(t)} - \langle\nu^{(t)}, x^{(t)}\rangle\mathbf{1}_d\Big)$ and the induced action $x^{(t)} = \frac{y^{(t)}}{\langle y^{(t)},\mathbf{1}\rangle}$, we have:

$$\begin{aligned}
\tilde{\mathrm{Reg}}(T) &= \max_{y^*\in[0,1]\Delta^d}\sum_{t=1}^T\langle u^{(t)}, y^* - y^{(t)}\rangle\\
&= \max_{y^*\in[0,1]\Delta^d}\sum_{t=1}^T w_t\left\langle \nu^{(t)} - \langle\nu^{(t)}, x^{(t)}\rangle\mathbf{1}_d, y^* - y^{(t)}\right\rangle\\
&= \max_{y^*\in[0,1]\Delta^d}\sum_{t=1}^T w_t\left(\langle\nu^{(t)}, y^*\rangle - \langle\nu^{(t)}, x^{(t)}\rangle\langle\mathbf{1}_d, y^*\rangle\right).
\end{aligned}$$

Since $y^* \in \Delta^d$ implies $\langle\mathbf{1}_d, y^*\rangle = 1$, the above simplifies to

$$\tilde{\mathrm{Reg}}(T) \geq \left(\max_{y^*\in\Delta^d}\sum_{t=1}^T w_t\big(\langle\nu^{(t)}, y^*\rangle - \langle\nu^{(t)}, x^{(t)}\rangle\big)\right) = \frac{\mathrm{reg}^T_{i,h}(s)}{\alpha_T^1}.$$

On the other hand, we clearly have $\tilde{\mathrm{Reg}}(T) \geq 0$ by choosing $y^* = 0$ as the comparator. ■

This proposition is important as it implies that any RVU bounds on $\tilde{\mathrm{Reg}}(t)$ directly translate into nonnegative RVU bounds on $\mathrm{reg}^t_{i,h}(s)$. Now, define for every $t \geq 1$,

$$F_t(y) \;:=\; -\eta_t\big\langle U^{(t)} + \kappa^{(t)}u^{(t-1)},\, y\big\rangle + \psi(y), \quad G_t(z) \;:=\; -\eta_t\big\langle U^{(t)}, z\big\rangle + \psi(z), \tag{29}$$

where $\kappa^{(t)} = \frac{w_t}{w_{t-1}}$. The lifted OFTRL iterate and its FTRL proxy are respectively given as $y^{(t)} = \arg\min_{y\in(0,1]\Delta^d} F_t(y)$, $z^{(t)} = \arg\min_{z\in(0,1]\Delta^d} G_t(z)$. Then, first we present the following lemma:

Lemma 21 *Given any convex function $F : \Omega \to \mathbb{R}$ defined on the compact set Ω, the minimizer $z^* = \arg\min_{z\in\Omega} F(z)$ satisfies $F(z^*) \leq F(z) - D_F(z \parallel z^*) \quad \forall z \in \Omega$, where D_F is the Bregman divergence induced by the function F.*

Proof By definition of the Bregman divergence, and first-order optimality conditions, we have

$$F(z^*) = F(z) - \langle \nabla F(z^*), z - z^* \rangle - D_F(z \parallel z^*). \leq F(z) - D_F(z \parallel z^*),$$

which proves the claim. ■

Following, Lemma 21, we state state the following lemma:

Lemma 22 (OFTRL one–step inequality with time–varying step-size) *For any $y \in (0,1]\Delta^d$ and any horizon $T \geq 1$, the following inequality holds.*

$$\sum_{t=1}^{T}\langle y - y^{(t)}, u^{(t)}\rangle \leq \frac{\psi(y)}{\eta_{T+1}} - \frac{\psi(y^{(1)})}{\eta_1} + \sum_{t=1}^{T}\langle z^{(t+1)} - y^{(t)},\, u^{(t)} - \kappa^{(t)}u^{(t-1)}\rangle - \sum_{t=1}^{T}\frac{1}{\eta_t}\Big[D_\psi(y^{(t)}\|z^{(t)}) + D_\psi(z^{(t+1)}\|y^{(t)})\Big] + \sum_{t=1}^{T}\left[\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right]\psi(z^{(t+1)}) \tag{30}$$

Proof By Lemma 21,

$$G_t(z^{(t)}) \leq G_t(y^{(t)}) - D_\psi(y^{(t)}\|z^{(t)}), \tag{31}$$

$$F_t(y^{(t)}) \leq F_t(z^{(t+1)}) - D_\psi(z^{(t+1)}\|y^{(t)}) \tag{32}$$

Moreover,

$$G_t(y^{(t)}) = F_t(y^{(t)}) + \eta_t\langle \kappa^{(t)}u^{(t-1)},\, y^{(t)}\rangle. \tag{33}$$

Furthermore, for any $w \in (0,1]\Delta^d$,

$$\frac{1}{\eta_t}F_t(w) = \frac{1}{\eta_{t+1}}G_{t+1}(w) + \langle u^{(t)} - \kappa^{(t)}u^{(t-1)},\, w\rangle + \Delta_t\,\psi(w), \qquad \Delta_t := \frac{1}{\eta_t} - \frac{1}{\eta_{t+1}} \leq 0, \tag{34}$$

which follows by expanding F_t, G_{t+1} and using $U^{(t+1)} = U^{(t)} + u^{(t)}$. Then, dividing (31)–(32) by η_t, substituting (33), and then applying (34) at $w = z^{(t+1)}$:

$$\begin{aligned}
\frac{1}{\eta_t}G_t(z^{(t)}) &\leq \frac{1}{\eta_t}F_t(y^{(t)}) + \langle \kappa^{(t)}u^{(t-1)}, y^{(t)}\rangle - \frac{1}{\eta_t}D_\psi(y^{(t)}\|z^{(t)}) \\
&\leq \frac{1}{\eta_t}F_t(z^{(t+1)}) - \frac{1}{\eta_t}D_\psi(y^{(t)}\|z^{(t)}) - \frac{1}{\eta_t}D_\psi(z^{(t+1)}\|y^{(t)}) + \langle \kappa^{(t)}u^{(t-1)}, y^{(t)}\rangle \\
&= \frac{1}{\eta_{t+1}}G_{t+1}(z^{(t+1)}) + \langle u^{(t)} - \kappa^{(t)}u^{(t-1)}, z^{(t+1)}\rangle + \Delta_t\,\psi(z^{(t+1)}) \\
&\quad - \frac{1}{\eta_t}\Big[D_\psi(y^{(t)}\|z^{(t)}) + D_\psi(z^{(t+1)}\|y^{(t)})\Big] + \langle \kappa^{(t)}u^{(t-1)}, y^{(t)}\rangle
\end{aligned}$$

Grouping the linear terms yields

$$\begin{aligned}
\frac{1}{\eta_t}G_t(z^{(t)}) \leq \frac{1}{\eta_{t+1}}G_{t+1}(z^{(t+1)}) &+ \langle y^{(t)}, u^{(t)}\rangle + \langle z^{(t+1)} - y^{(t)},\, u^{(t)} - \kappa^{(t)}u^{(t-1)}\rangle \\
&+ \Delta_t\,\psi(z^{(t+1)}) - \frac{1}{\eta_t}\Big[D_\psi(y^{(t)}\|z^{(t)}) + D_\psi(z^{(t+1)}\|y^{(t)})\Big]
\end{aligned} \tag{35}$$

Summing (35) over $t = 1, \dots, T$ telescopes the G-terms:

$$\frac{1}{\eta_1} G_1\big(z^{(1)}\big) \le \frac{1}{\eta_{T+1}} G_{T+1}\big(z^{(T+1)}\big) + \sum_{t=1}^{T} \big\langle y^{(t)}, u^{(t)} \big\rangle + \sum_{t=1}^{T} \big\langle z^{(t+1)} - y^{(t)},\, u^{(t)} - \kappa^{(t)} u^{(t-1)} \big\rangle + \sum_{t=1}^{T} \Delta_t\, \psi\big(z^{(t+1)}\big) - \sum_{t=1}^{T} \frac{1}{\eta_t} \Big[D_\psi(\cdot) + D_\psi(\cdot) \Big]$$

Since $U^{(1)} = 0$, $G_1 = \psi$ and $z^{(1)} = y^{(1)} \in \arg\min_\Omega \psi$, and thus $\frac{1}{\eta_1} G_1(z^{(1)}) = \frac{\psi(y^{(1)})}{\eta_1}$. By optimality of $z^{(T+1)}$,

$$\frac{1}{\eta_{T+1}} G_{T+1}\big(z^{(T+1)}\big) \le \frac{1}{\eta_{T+1}} G_{T+1}(y) = -\big\langle U^{(T+1)}, y \big\rangle + \frac{\psi(y)}{\eta_{T+1}}.$$

Insert these, and use $\sum_{t=1}^{T} \langle y, u^{(t)} \rangle = \langle U^{(T+1)}, y \rangle$ to obtain

$$\sum_{t=1}^{T} \big\langle y - y^{(t)}, u^{(t)} \big\rangle \le \frac{\psi(y)}{\eta_{T+1}} - \frac{\psi\big(y^{(1)}\big)}{\eta_1} + \sum_{t=1}^{T} \Delta_t\, \psi\big(z^{(t+1)}\big) + \sum_{t=1}^{T} \big\langle z^{(t+1)} - y^{(t)},\, u^{(t)} - \kappa^{(t)} u^{(t-1)} \big\rangle - \sum_{t=1}^{T} \frac{1}{\eta_t} \Big[D_\psi\big(y^{(t)} \big\| z^{(t)}\big) + D_\psi\big(z^{(t+1)} \big\| y^{(t)}\big) \Big]$$

which is precisely (30). ■

Lemma 23 *We have the following inequality for iterative Bregman divergences of regularizer ψ:*

$$\sum_{t=1}^{T} \frac{1}{\eta_t} \Big[D_\psi\big(y^{(t)} \,\|\, z^{(t)}\big) + D_\psi\big(z^{(t+1)} \,\|\, y^{(t)}\big) \Big] \;\ge\; \sum_{t=1}^{T} \frac{1}{2\eta_t} \Big(\|y^{(t)} - z^{(t)}\|_1^2 + \|z^{(t+1)} - y^{(t)}\|_1^2 \Big).$$

Proof By repeated application of Proposition 18, we have for each $t \in [T]$:

$$\frac{1}{\eta_t} \Big[D_\psi\big(y^{(t)} \,\|\, z^{(t)}\big) + D_\psi\big(z^{(t+1)} \,\|\, y^{(t)}\big) \Big] \;\ge\; \frac{1}{2\eta_t} \|y^{(t)} - z^{(t)}\|_1^2 + \frac{1}{2\eta_t} \|z^{(t+1)} - y^{(t)}\|_1^2.$$

Summing this inequality over all $t = 1$ to T concludes the proof. ■

Lemma 24 *If β is large enough ($\beta \ge 70$), then*

$$\sum_{t=1}^{T} \frac{1}{\eta_t} \Big[D_\psi\big(y^{(t)} \,\|\, z^{(t)}\big) + D_\psi\big(z^{(t+1)} \,\|\, y^{(t)}\big) \Big] \;\ge\; \sum_{t=1}^{T-1} \frac{1}{10\eta_t} \Big(\|x^{(t+1)} - \theta^{(t+1)}\|_1^2 + \|\theta^{(t+1)} - x^{(t)}\|_1^2 \Big).$$

Proof By Theorem 15, we know the stability ratio $\rho := \frac{\Lambda(z)}{\Lambda(y)} \in [1 - \epsilon, 1 + \epsilon]$ with $\epsilon = \frac{2}{5}$. Using Proposition 19, this implies

$$D_\psi(z \,\|\, y) \ge \frac{1}{4}(1 - \epsilon) \|\theta - x\|_1^2.$$

Applying this with $z := z^{(t+1)}$, $y := y^{(t)}$, $\theta := \theta^{(t+1)}$, and $x := x^{(t)}$, we obtain:

$$\frac{1}{\eta_t} D_\psi\big(z^{(t+1)} \,\|\, y^{(t)}\big) \ge \frac{1}{4\eta_t}(1 - \epsilon) \|\theta^{(t+1)} - x^{(t)}\|_1^2 = \frac{3}{20\eta_t} \|\theta^{(t+1)} - x^{(t)}\|_1^2 > \frac{1}{10\eta_t} \|\theta^{(t+1)} - x^{(t)}\|_1^2.$$

Similarly, we get

$$\frac{1}{\eta_t} D_\psi\big(y^{(t+1)} \,\|\, z^{(t+1)}\big) > \frac{1}{10\eta_t}\|x^{(t+1)} - \theta^{(t+1)}\|_1^2.$$

Combining both and summing over $t = 1$ to $T-1$ yields the result. ■

Theorem 6 (RVU bound for MG-DLRC-OMWU with time-varying η_t) *Let $\beta \geq 70$, and assume that the reward signals obey $\big\|\nu^{(t)}\big\|_\infty \leq H$ for every $t \in [T]$. Then, the cumulative regret incurred by the* inner *OFTRL process up to horizon T obeys*

$$\tilde{\mathrm{Reg}}(T) := \sum_{t=1}^{T}\Big\langle y - y^{(t)}, u^{(t)}\Big\rangle \leq 2\|u^{(t)}\|_\infty \;+\; \frac{\tilde{\alpha}\log T + 2\log d}{\eta_{T+1}} \;+\; \sum_{t=1}^{T}\eta_t\|u^{(t)} - \kappa^{(t)}u^{(t-1)}\|_\infty^2 - \frac{1}{20}\sum_{t=1}^{T-1}\frac{\|x^{(t+1)} - x^{(t)}\|_1^2}{\eta_t} \tag{36}$$

Proof For an arbitrary $y \in \Omega$, define the smoothed comparator $y' := \frac{T-1}{T}\, y + \frac{1}{T}\, y^{(1)} \in \Omega$ (where $y^{(1)} := \arg\min_{y\in\Omega} \psi_1(y)$). Then

$$\sum_{t=1}^{T}\big\langle y - y^{(t)}, u^{(t)}\big\rangle \;\leq\; 2\|u^{(t)}\|_\infty + \sum_{t=1}^{T}\big\langle y' - y^{(t)}, u^{(t)}\big\rangle. \tag{37}$$

Lemma 22 with $y = y'$ yields

$$\sum_{t=1}^{T}\big\langle y' - y^{(t)}, u^{(t)}\big\rangle \;\leq\; \underbrace{\frac{\psi(y')}{\eta_{T+1}} - \frac{\psi_1\big(y^{(1)}\big)}{\eta_1}}_{(I)} \;+\; \underbrace{\sum_{t=1}^{T}\big\langle z^{(t+1)} - y^{(t)},\, u^{(t)} - \kappa^{(t)}u^{(t-1)}\big\rangle}_{(II)} - \underbrace{\sum_{t=1}^{T}\frac{1}{\eta_t}\Big[D_\psi\big(y^{(t)} \,\|\, z^{(t)}\big) + D_\psi\big(z^{(t+1)} \,\|\, y^{(t)}\big)\Big]}_{(III)} + \underbrace{\sum_{t=1}^{T}\left[\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right]\psi\big(z^{(t+1)}\big)}_{(IV)}. \tag{38}$$

We aim to bound the difference (I). Let $\Lambda(y) := \sum_{k=1}^{d} y[k]$ and $H(y) := \sum_{k=1}^{d} y[k]\log y[k]$. Recall $\psi_t(y) = -\tilde{\alpha}\,\log\Lambda(y) + \frac{1}{\Lambda(y)}H(y)$ and set $S' := \Lambda(y')$, $S_1 := \Lambda\big(y^{(1)}\big) = 1$. Then,

$$(I) = \frac{\psi_T(y')}{\eta_{T+1}} - \frac{\psi_1(y^{(1)})}{\eta_1} = \underbrace{-\frac{\tilde{\alpha}}{\eta_{T+1}}\log S'}_{\text{(A)}} \;+\; \underbrace{\frac{1}{\eta_{T+1}S'}\,H(y')}_{\text{(B)}} \;+\; \underbrace{\frac{\tilde{\alpha}}{\eta_1}\log S_1}_{\text{(C)}} \;-\; \underbrace{\frac{1}{\eta_1 S_1}\,H\big(y^{(1)}\big)}_{\text{(D)}}. \tag{39}$$

By linearity of Λ, $S' = \Lambda(y') = \frac{T-1}{T}\,\Lambda(y) + \frac{1}{T}\,\Lambda(y^{(1)})$. Since $y \in \Omega$ implies $0 \leq \Lambda(y) \leq 1$ and $\Lambda(y^{(1)}) = 1$, we obtain

$$\frac{1}{T} \;\leq\; S' \;\leq\; 1 \quad\Longrightarrow\quad 0 \;\leq\; -\log S' \;\leq\; \log T.$$

Hence (A) $= -\frac{\tilde{\alpha}}{\eta_{T+1}}\log S' \;\leq\; \frac{\tilde{\alpha}\log T}{\eta_{T+1}}$. Then, write $p'_k := y'[k]/S'$ so that $\sum_k p'_k = 1$ and

$$\frac{H(y')}{S'} = \log S' + \sum_{k=1}^{d} p'_k \log p'_k \;\leq\; \log S' \;\leq\; 0.$$

Thus (B) $= \frac{H(y')}{\eta_{T+1}S'} \le 0$. For $y^{(1)}$, set $p_k^{(1)} := y^{(1)}[k]/S_1 = y^{(1)}[k]$. Since $\sum_k p_k^{(1)} \log p_k^{(1)} \ge -\log d$,

$$-H\big(y^{(1)}\big) = -\sum_k y^{(1)}[k] \log y^{(1)}[k] \le \log d,$$

and with $S_1 = 1$, (D) $= -\frac{1}{\eta_1 S_1} H\big(y^{(1)}\big) \le \frac{\log d}{\eta_1}$. Finally, (C) $= \frac{\tilde{\alpha}}{\eta_1} \log S_1 = 0$. Summing over gives

$$(I) \;\le\; \frac{\tilde{\alpha} \log T + \log d}{\eta_{T+1}}. \tag{40}$$

Now, we bound term (II)). For each t, apply Hölder and Young with the *local* step-size η_t: $\big|\langle a, b\rangle\big| \le \frac{\|a\|_1^2}{4\eta_t} + \eta_t \|b\|_\infty^2$. Set $a = z^{(t+1)} - y^{(t)}$ and $b = u^{(t)} - \kappa^{(t)} u^{(t-1)}$:

$$(II) \le \sum_{t=1}^{T} \Big[\frac{\|z^{(t+1)} - y^{(t)}\|_1^2}{4\eta_t} + \eta_t \|u^{(t)} - \kappa^{(t)} u^{(t-1)}\|_\infty^2 \Big]. \tag{41}$$

Next, we bound term (III)). Combining Lemma 23 and Lemma 24, we have:

$$\begin{aligned}(III) &\le -\sum_{t=1}^{T} \frac{\|y^{(t)} - z^{(t)}\|_1^2 + \|z^{(t+1)} - y^{(t)}\|_1^2}{4\eta_t} - \sum_{t=1}^{T-1} \frac{\|x^{(t+1)} - \theta^{(t+1)}\|_1^2 + \|\theta^{(t+1)} - x^{(t)}\|_1^2}{20\eta_t} \\ &\le -\sum_{t=1}^{T} \frac{\|y^{(t)} - z^{(t)}\|_1^2 + \|z^{(t+1)} - y^{(t)}\|_1^2}{4\eta_t} - \sum_{t=1}^{T-1} \frac{\|x^{(t+1)} - x^{(t)}\|_1^2}{20\eta_t},\end{aligned} \tag{42}$$

where the last step is due to the triangle inequality. Finally, we combine (II) and (III)):

$$(II) + (III) \le \sum_{t=1}^{T} \eta_t \|u^{(t)} - \kappa^{(t)} u^{(t-1)}\|_\infty^2 \;-\; \frac{1}{20} \sum_{t=1}^{T-1} \frac{\|x^{(t+1)} - x^{(t)}\|_1^2}{\eta_t} \tag{43}$$

Now, we bound term (IV). Write $s := \Lambda(y) = \sum_k y[k]$ and $p_k := y[k]/s$ whenever $s > 0$. Then,

$$\psi(y) \;=\; -(\tilde{\alpha} - 1)\, \log s \;+\; \sum_{k=1}^{d} p_k \log p_k.$$

Note that $\sum_k p_k \log p_k \in [-\log d,\, 0]$, with the minimum $-\log d$ attained at the uniform $p_k \equiv 1/d$; the first term $-(\tilde{\alpha} - 1) \log s$ is nonnegative for $s \in (0, 1]$ and strictly decreasing in s. Hence the global infimum of ψ over Ω is achieved at $s = 1$ and uniform p, i.e., at $y^\star = (1/d, \ldots, 1/d)$ with $\inf_{y \in \Omega} \psi(y) \;=\; \psi(y^\star) \;=\; -\log d$. Since $\Delta_t \le 0$ and $\psi\big(z^{(t+1)}\big) \ge \inf_{y \in \Omega} \psi(y)$, we have for each t: $\Delta_t\, \psi\big(z^{(t+1)}\big) \;\le\; \Delta_t \cdot \inf_{y \in \Omega} \psi(y)$. Summing over T and using $\sum_{t=1}^{T} \Delta_t = \frac{1}{\eta_1} - \frac{1}{\eta_{T+1}}$ yields

$$\sum_{t=1}^{T} \Delta_t\, \psi\big(z^{(t+1)}\big) \;\le\; \Big(\frac{1}{\eta_1} - \frac{1}{\eta_{T+1}}\Big)\, (-\log d) \;=\; \Big(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_1}\Big)\, \log d. \tag{44}$$

Then, inserting (37),(40),(44),(43), we get:

$$\tilde{\mathrm{Reg}}(T) \;\le\; 2\|u^{(t)}\|_\infty + \frac{\tilde{\alpha} \log T + 2 \log d}{\eta_{T+1}} + \sum_{t=1}^{T} \eta_t \|u^{(t)} - \kappa^{(t)} u^{(t-1)}\|_\infty^2 - \frac{1}{20} \sum_{t=1}^{T-1} \frac{\|x^{(t+1)} - x^{(t)}\|_1^2}{\eta_t}$$

■

Lemma 25 *Assume that $\|\nu^{(t)}\|_\infty \le H$ for all $t \in [T]$. Then,*

$$\big\|u^{(t)} - \kappa^{(t)} u^{(t-1)}\big\|_\infty^2 \;\le\; w_t^2 \Big(6\, \|\nu^{(t)} - \nu^{(t-1)}\|_\infty^2 \;+\; 4\, H^2\, \|x^{(t)} - x^{(t-1)}\|_1^2\Big) \tag{45}$$

Proof Since $u^{(t)} = w_t\Big(\nu^{(t)} \;-\; \langle \nu^{(t)},\, x^{(t)}\rangle \mathbf{1}_d\Big)$, and $\kappa^{(t)} = \frac{w_t}{w_{t-1}}$, we have

$$\begin{aligned}
\left\|u^{(t)} - \kappa^{(t)} u^{(t-1)}\right\|_\infty^2 &= \left\| w_t\big(\nu^{(t)} - \langle \nu^{(t)}, x^{(t)}\rangle \mathbf{1}_d\big) - w_t\big(\nu^{(t-1)} - \langle \nu^{(t-1)}, x^{(t-1)}\rangle \mathbf{1}_d\big)\right\|_\infty^2 \\
&\le w_t^2 \left(\left\|\nu^{(t)} - \nu^{(t-1)}\right\|_\infty \;+\; \left|\langle \nu^{(t)}, x^{(t)}\rangle - \langle \nu^{(t-1)}, x^{(t-1)}\rangle\right| \right)^2 \\
&\le w_t^2 \Big(2\left\|\nu^{(t)} - \nu^{(t-1)}\right\|_\infty^2 \;+\; 2\left|\langle \nu^{(t)}, x^{(t)}\rangle - \langle \nu^{(t-1)}, x^{(t-1)}\rangle\right|^2 \Big) \\
&\le w_t^2 \Big(2\left\|\nu^{(t)} - \nu^{(t-1)}\right\|_\infty^2 + 4\left|\langle \nu^{(t)}, x^{(t)} - x^{(t-1)}\rangle\right|^2 + 4\left|\langle \nu^{(t)} - \nu^{(t-1)}, x^{(t-1)}\rangle\right|^2 \Big) \\
&\le w_t^2 \Big(2\left\|\nu^{(t)} - \nu^{(t-1)}\right\|_\infty^2 + 4\left\|\nu^{(t)}\right\|_\infty^2 \left\|x^{(t)} - x^{(t-1)}\right\|_1^2 + 4\left\|\nu^{(t)} - \nu^{(t-1)}\right\|_\infty^2 \Big) \\
&\le w_t^2 \Big(6\left\|\nu^{(t)} - \nu^{(t-1)}\right\|_\infty^2 \;+\; 4\,H^2 \left\|x^{(t)} - x^{(t-1)}\right\|_1^2 \Big)
\end{aligned}$$

We used triangle inequality in the first step, Young's inequality in the second and third steps, and Hölder's inequality with $\|\nu^{(t)}\|_\infty \le H$ in the fourth step. In the final step, we leverage the $\|\nu\|_\infty < H$, and group the terms which concludes the proof. ■

Appendix C. Proof of Theorem 4

In this appendix, we first present the proof of Lemma 7, where we establish the non-negative regret bound for each (s,h) pair. We then leverage this per-state regret bound to control the second-order path length of the policies and recursively bound the CCE-gap in Theorem 4, which allows us to conclude the convergence result for the CCE-gap.

Lemma 7 (Per-state weighted regret bounds: revised RVU) *Fix an episodic step $h \in [H]$, state $s \in \mathcal{S}$, agent $i \in \mathcal{N}$, and horizon $T \ge 2$. Run Algorithm 3 with a* constant *base learning rate $\eta > 0$ and the usual weights $w_j = \alpha_t^j / \alpha_t^1$, where $\alpha_t = (H+1)/(H+t)$. Let $|A_{\max}| := |\mathcal{A}_i|$ and let be $\tilde{\alpha}$ the constant appearing in the Theorem 6. Then, for every $t \in [T]$,*

$$\begin{aligned}
\mathrm{reg}_{i,h}^t(s) \le & \frac{2H\big(\tilde{\alpha}\log t + 2\log|A_{\max}| + 6H\eta\big)}{\eta\, t} + 12\eta\, H^2(N-1)\sum_{j=2}^{t-1}\sum_{k\neq i}\alpha_t^j \left\|\pi_{h,k}^j - \pi_{h,k}^{j-1}\right\|_1^2 \\
& + \frac{12\eta H^2(3H+4N^2)}{t} - \frac{1}{24\,\eta H}\sum_{j=2}^{t-1}\alpha_t^{\,j}\left\|\pi_{i,h}^j - \pi_{i,h}^{j-1}\right\|_1^2. \qquad (46)
\end{aligned}$$

Moreover, summing (46) over all agents and choosing $\eta = \dfrac{1}{24H\sqrt{HN}}$ yields

$$\begin{aligned}
\sum_{i=1}^N \mathrm{reg}_{i,h}^t(s) \le & \frac{2HN\big(\tilde{\alpha}\log t + 2\log|A_{\max}| + 6H\eta\big)}{\eta\, t} + \frac{12\eta H^2 N(3H+4N^2)}{t} \qquad (47) \\
& - \frac{1}{48\eta H}\sum_{i=1}^N\sum_{j=2}^{t-1}\alpha_t^j\left\|\pi_{i,h}^j - \pi_{i,h}^{j-1}\right\|_1^2
\end{aligned}$$

Proof We fix (s,h) and see each episode index j as a full-information normal-form game with payoff vector $\nu_j = [Q_{i,h}^j \pi_{h,-i}^j](s,\cdot)$, and $u^{(t)} = w_t\Big(\nu^{(t)} \;-\; \langle \nu^{(t)},\, x^{(t)}\rangle \mathbf{1}_d\Big)$. By definition of the

weighted regret term we have:

$$\begin{aligned}
\mathrm{reg}^t_{i,h}(s) &:= \max_{\pi^\dagger_{i,h}\in\Delta(\mathcal{A}_i)} \sum_{j=1}^{t} \alpha_t^j \Big\langle \pi^\dagger_{i,h} - \pi^j_{i,h}, \Big[Q^{(j)}_{i,h}\pi^{(j)}_{-i,h}\Big](s,\cdot)\Big\rangle = \alpha_t^1 \max_{\pi^\dagger_{i,h}\in\Delta(\mathcal{A}_i)} \sum_{j=1}^{t} \Big\langle \pi^\dagger_{i,h} - \pi^j_{i,h}, w_j\Big[Q^{(j)}_{i,h}\pi^{(j)}_{-i,h}\Big](s,\cdot)\Big\rangle \\
&\le \alpha_t^1 \left[2\|u^{(t)}\|_\infty + \frac{\tilde{\alpha}\log t + 2\log d}{\eta_{t+1}} + \sum_{j=1}^{t} \eta_j \|u^{(j)} - \kappa^{(j)} u^{(j-1)}\|_\infty^2 - \frac{1}{20}\sum_{j=1}^{t-1} \frac{\|\pi^{j+1}_{i,h} - \pi^j_{i,h}\|_1^2}{\eta_j} \right] \\
&\le \alpha_t^1 \left[2\|u^{(t)}\|_\infty + \frac{\tilde{\alpha}\log t + 2\log d}{\eta_{t+1}} + \sum_{j=1}^{t} \eta_j w_j^2 \Big(6\,\|\nu^{(j)} - \nu^{(j-1)}\|_\infty^2 + 4\,H^2\,\|\pi^j_{i,h} - \pi^{j-1}_{i,h}\|_1^2\Big) \right. \\
&\quad \left. - \frac{1}{20}\sum_{j=1}^{t-1} \frac{\|\pi^{j+1}_{i,h} - \pi^j_{i,h}\|_1^2}{\eta_j} \right] \\
&\le \alpha_t^1 \left[5Hw_t + 4H^2\eta w_1 + \frac{\tilde{\alpha}\log t + 2\log d}{\eta_{t+1}} + \sum_{j=1}^{t-1} 6\eta w_j \Big\| \Big[Q^{(j)}_{i,h}\pi^{(j)}_{-i,h}\Big](s,\cdot) - \Big[Q^{(j-1)}_{i,h}\pi^{(j-1)}_{-i,h}\Big](s,\cdot)\Big\|_\infty^2 \right. \\
&\quad \left. - \sum_{j=1}^{t-1} \frac{w_{j+1}}{24H\eta}\|\pi^{j+1}_{i,h} - \pi^j_{i,h}\|_1^2 \right]
\end{aligned}$$

where the first step is due to $w_j = \frac{\alpha_t^j}{\alpha_t^1}$ and the second step is due to Theorem 6 and Proposition 20. Furthermore, the third step is due to Lemma 25, and the final step is due to step size $\eta = \frac{1}{24H\sqrt{HN}}$. Since $1/\eta_{t+1} = w_{t+1}/\eta$, $\alpha_t^t = \alpha_t = \frac{H+1}{H+t} \le \frac{2H}{t}$ and $\alpha_t^1 w_t = \alpha_t^t$,

$$\alpha_t^1 \left(\frac{\tilde{\alpha}\log t + 2\log|A_{\max}|}{\eta_{t+1}}\right) = \frac{\tilde{\alpha}\log t + 2\log|A_{\max}|}{\eta}\,\alpha_{t+1}^{t+1}\gamma \;\le\; \frac{2H(\tilde{\alpha}\log t + 2\log|A_{\max}|)}{\eta\, t},$$

where $\gamma = \big(\frac{H+1+t}{t}\big)$ with the order of H. Likewise, $\alpha_t^1(5Hw_t + 4H^2\eta w_1) \le 6H\,\alpha_t^1 w_t \le \frac{12H^2}{t}$; these become the first term of (46). For the utility terms, start with the exact decomposition. Then, taking sup-norms and using $\|A\,x\|_\infty \le \|A\|_\infty\|x\|_1$:

$$\begin{aligned}
\|\big(Q^j_{i,h} - Q^{j-1}_{i,h}\big)\pi^j_{h,-i} + Q^{j-1}_{i,h}\big(\pi^j_{h,-i} - \pi^{j-1}_{h,-i}\big)\|_\infty &\le \|(Q^j_{i,h} - Q^{j-1}_{i,h})\,\pi^j_{h,-i}\|_\infty \\
&\quad + \|Q^{j-1}_{i,h}\,(\pi^j_{h,-i} - \pi^{j-1}_{h,-i})\|_\infty \\
&\le \alpha_j\,H \;+\; H\,\|\pi^j_{h,-i} - \pi^{j-1}_{h,-i}\|_1,
\end{aligned}$$

In the first step, we applied the triangle inequality; and in the second step, we used Hölder's inequality for the $(\|\cdot\|_\infty, \|\cdot\|_1)$ norm pair; and invoked the Bellman update guarantee $\|Q^j_{i,h} - Q^{j-1}_{i,h}\|_\infty \le \alpha_j H$, along with the bounds $\|Q^{j-1}_{i,h}\|_\infty \le H$ and $\|\pi^j_{h,-i}\|_1 \le 1$. Next squaring both sides and applying $(a+b)^2 \le 2a^2 + 2b^2$, yields:

$$\|\nu^{(j)} - \nu^{(j-1)}\|_\infty^2 \;\le\; \big(\alpha_j H + H\|\pi^j_{h,-i} - \pi^{j-1}_{h,-i}\|_1\big)^2 \;\le\; 2(\alpha_j H)^2 \;+\; 2H^2\,\|\pi^j_{h,-i} - \pi^{j-1}_{h,-i}\|_1^2.$$

Hence

$$\begin{aligned}
6\eta\,\alpha_t^1 \sum_{j=1}^{t-1} w_j \|\nu^{(j)} - \nu^{(j-1)}\|_\infty^2 &\le 12\eta\,\alpha_t^1 \sum_{j=1}^{t-1} w_j \Big((\alpha_j H)^2 + H^2\|\pi^j_{h,-i} - \pi^{j-1}_{h,-i}\|_1^2\Big) \\
&= 12\eta H^2 \sum_{j=1}^{t-1} \alpha_t^j\,(\alpha_j)^2 \;+\; 12\eta H^2 \sum_{j=1}^{t-1} \alpha_t^j\,\|\pi^j_{h,-i} - \pi^{j-1}_{h,-i}\|_1^2.
\end{aligned}$$

By Lemma 9 we have $\sum_{j=1}^{t} \alpha_t^j \, (\alpha_j)^2 \leq \frac{3H}{t}$, and utilizing the total-variation bound Hoeffding and Wolfowitz (1958):

$$\begin{aligned}\Big\|\pi_{h,-i}^{j}(\cdot|s) - \pi_{h,-i}^{j-1}(\cdot|s)\Big\|_1^2 &= \Big(\sum_{a_{-i}\in A_{-i}} \Big|\prod_{k\neq i}\pi_{h,k}^{j}(a_k|s) - \prod_{k\neq i}\pi_{h,k}^{j-1}(a_k|s)\Big|\Big)^2 \\ &\leq \Big(\sum_{k\neq i}\big\|\pi_{h,k}^{j}(\cdot|s) - \pi_{h,k}^{j-1}(\cdot|s)\big\|_1\Big)^2 \\ &\leq (N-1)\sum_{k\neq i}\big\|\pi_{h,k}^{j}(\cdot|s) - \pi_{h,k}^{j-1}(\cdot|s)\big\|_1^2,\end{aligned}$$

we get,

$$\begin{aligned}6\eta\,\alpha_t^1\sum_{j=1}^{t-1} w_j\|\nu^{(j)} - \nu^{(j-1)}\|_\infty^2 \;&\leq\; \frac{36\,\eta\,H^3}{t} \;+\; 12\,\eta\,H^2\,(N-1)\sum_{j=1}^{t-1}\alpha_t^j\sum_{k\neq i}\big\|\pi_{h,k}^{j} - \pi_{h,k}^{j-1}\big\|_1^2, \\ &\leq \frac{12\eta H^2(3H+4N^2)}{t} + 12\eta H^2(N-1)\sum_{j=2}^{t-1}\alpha_t^j\sum_{k\neq i}\big\|\pi_{h,k}^{j} - \pi_{h,k}^{j-1}\big\|_1^2,\end{aligned}$$

where, we have used the following inequality along with the fact $\alpha_t^1 \leq \frac{1}{t}$ by Lemma 9,

$$12\,\eta\,H^2\,(N-1)\alpha_t^1\sum_{k\neq i}\big\|\pi_{h,k}^{1} - \pi_{h,k}^{0}\big\|_1^2 \;\leq\; \frac{48\eta(N-1)^2H^2}{t}$$

leading to the two middle terms in (46). For the final term, we have:

$$\begin{aligned}-\frac{\alpha_t^1}{24H}\sum_{j=1}^{t-1}\frac{\|\pi_{i,h}^{j+1} - \pi_{i,h}^{j}\|_1^2}{\eta_{j+1}} = -\frac{1}{24\eta H}\sum_{j=1}^{t-1}\alpha_t^1 w_{j+1}\|\pi_{i,h}^{j+1} - \pi_{i,h}^{j}\|_1^2 &= -\frac{1}{24\eta H}\sum_{j=1}^{t-1}\alpha_t^{j+1}\|\pi_{i,h}^{j+1} - \pi_{i,h}^{j}\|_1^2, \\ &= -\frac{1}{24\,\eta H}\sum_{j=2}^{t}\alpha_t^j\|\pi_{i,h}^{j} - \pi_{i,h}^{j-1}\|_1^2\end{aligned}$$

Then, summing up the upper bounds for all four terms we get:

$$\begin{aligned}\mathrm{reg}_{i,h}^t(s) \leq &\frac{2H\big(\tilde{\alpha}\log t + 2\log|A_{\max}| + 6H\eta\big)}{\eta\,t} + 12\eta\,H^2(N-1)\sum_{j=2}^{t-1}\sum_{k\neq i}\alpha_t^j\big\|\pi_{h,k}^{j} - \pi_{h,k}^{j-1}\big\|_1^2 \\ &+ \frac{12\eta H^2(3H+4N^2)}{t} - \frac{1}{24\,\eta H}\sum_{j=2}^{t-1}\alpha_t^j\,\big\|\pi_{i,h}^{j} - \pi_{i,h}^{j-1}\big\|_1^2. \qquad (48)\end{aligned}$$

Finally, summing over all players we get:

$$\begin{aligned}\sum_{i=1}^{N}\mathrm{reg}_{i,h}^t(s) \;\leq\; &\frac{2HN\big(\tilde{\alpha}\log t + 2\log|A_{\max}| + 6H\eta\big)}{\eta\,t} \;+\; \frac{12\eta H^2N(3H+4N^2)}{t} \\ &+\; 12\,\eta\,H^2(N-1)^2\sum_{i=1}^{N}\sum_{j=2}^{t-1}\alpha_t^j\big\|\pi_{i,h}^{j}(\cdot|s) - \pi_{i,h}^{j-1}(\cdot|s)\big\|_1^2 \\ &-\; \frac{1}{24\,\eta H}\sum_{i=1}^{N}\sum_{j=2}^{t-1}\alpha_t^j\big\|\pi_{i,h}^{j}(\cdot|s) - \pi_{i,h}^{j-1}(\cdot|s)\big\|_1^2. \qquad (49)\end{aligned}$$

Choosing the learning-rate such as $\eta = 1/\big(24\,H\sqrt{HN}\big)$ yields the following inequality

$$\sum_{i=1}^{N}\mathrm{reg}_{i,h}^{t}(s)\leq\frac{2HN\big(\tilde{\alpha}\log t+2\log|A_{\max}|+6H\eta\big)}{\eta\,t}+\frac{12\eta H^2N(3H+4N^2)}{t}-\frac{\sum_{i=1}^{N}\sum_{j=2}^{t-1}\alpha_t^j\big\|\pi_{i,h}^{j}-\pi_{i,h}^{j-1}\big\|_1^2}{48\eta H}$$

which leads to (47) and concludes the proof. ■

Theorem 4 (Regret Bounds for MG-DLRC-OMWU) *If the Algorithm 3, equivalently Algorithm 1, is run on an N-player episodic Markov game for T iterations with parameters $\beta \geq 70$, $\tilde{\alpha} = \beta\log^2|\mathcal{A}_{\max}| + 2\log|\mathcal{A}_{\max}| + 2$, and $\eta = 1/24H\sqrt{HN}$, the output policy $\bar{\pi}$ satisfies:*

$$\text{CCE-Gap}(\bar{\pi}) \leq \frac{864H^{\frac{7}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}| + 2\big)}{T}$$

Proof We know that the right hand side of the inequality given by (49) in Lemma 7 are guaranteed to be non-negative due to Theorem 6 and Proposition 20. Then, we can write the following inequality for $12\eta H^2N\sum_{i=1}^{N}\sum_{j=2}^{t-1}\alpha_t^j\big\|\pi_{i,h}^{j}-\pi_{i,h}^{j-1}\big\|_1^2$ using $\eta = \frac{1}{24H\sqrt{HN}}$ and inequality (49) :

$$\begin{aligned}12\eta H^2N\sum_{i=1}^{N}\sum_{j=2}^{t-1}\alpha_t^j\big\|\pi_{i,h}^{j}-\pi_{i,h}^{j-1}\big\|_1^2 &\leq 576\eta^2H^3N\left[\frac{2HN\big(\tilde{\alpha}\log t+2\log|A_{\max}|+6H\eta\big)}{\eta\,t}+\frac{12\eta H^2N(3H+4N^2)}{t}\right]\\ &\leq \frac{2H\big(\tilde{\alpha}\log t+2\log|A_{\max}|+6H\eta\big)}{\eta\,t}+\frac{12\eta H^2(3H+4N^2)}{t}\end{aligned}$$

Then plugging last inequality into (48) and getting rid of negative terms lead us to the following per-state regret per player:

$$\begin{aligned}\mathrm{reg}_{i,h}^{t}(s) &\leq \frac{4H\big(\tilde{\alpha}\log t+2\log|A_{\max}|+6H\eta\big)}{\eta\,t}+\frac{24\eta H^2(3H+4N^2)}{t} &(50)\\ &\leq \frac{4H\big(\tilde{\alpha}\log t+2\log|A_{\max}|+6H\eta\big)}{\eta\,t}+\frac{3H^2+4HN}{t} &(51)\\ &\leq \frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log t+2\log|A_{\max}|+2\big)}{t} &(52)\end{aligned}$$

where the last step is due to $\eta = \frac{1}{24H\sqrt{HN}}$. Then, from Lemma 5, for $1\leq h\leq H,\ 1\leq t\leq T$, we have

$$\delta_h^t \ \leq\ \mathrm{reg}_{i,h}^{t}(s)\ +\ \sum_{j=1}^{t}\alpha_t^j\,\delta_{h+1}^{j},\qquad\text{and}\qquad\delta_{H+1}^{t}=0. \tag{53}$$

Also due to Lemma 9 we have $\sum_{j=1}^{t}\alpha_t^j = 1$ and $\sum_{j=1}^{t}\alpha_t^j\frac{1}{j}\ \leq\ \Big(1+\frac{1}{H}\Big)\frac{1}{t}$. Define $\gamma\ :=\ 1+\frac{1}{H}$. Then, we claim that for every $1\leq h\leq H,\ 1\leq t\leq T$,

$$\delta_h^t\ \leq\ \sum_{h'=h}^{H}\gamma^{\,2(H-h'+0.5)}\,\frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log t+2\log|A_{\max}|+2\big)}{t} \tag{54}$$

holds. We proceed by backward induction on h. *Base case is* $h = H+1$. Because $\delta_{H+1}^{t}=0$ and $\gamma^{\,2(H-H-1+0.5)}=\gamma^{-1}$, (54) becomes

$$0\leq\gamma^{-1}\frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log t+2\log|A_{\max}|+2\big)}{t},$$

which is true. For the induction step, assume (54) holds for level $h+1$. Using (53) and the induction hypothesis, for any t we have

$$\delta_h^t \leq \mathrm{reg}_{i,h}^t(s) + \sum_{j=1}^{t} \alpha_t^j \sum_{h'=h+1}^{H} \gamma^{2(H-h'+0.5)} \frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log t + 2\log|A_{\max}|+2\big)}{j}$$

$$\leq \mathrm{reg}_{i,h}^t(s) + \sum_{h'=h+1}^{H} \gamma^{2(H-h'+0.5)} \Big[\sum_{j=1}^{t} \alpha_t^j \frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{j}\Big]$$

$$\leq \frac{96HH^{\frac{5}{2}}N\big(\tilde{\alpha}\log t+2\log|A_{\max}|+2\big)}{t} + \sum_{h'=h+1}^{H}\gamma^{2(H-h'+0.5)}\Big[\sum_{j=1}^{t}\alpha_t^j\frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log T+2\log|A_{\max}|+2\big)}{j}\Big]$$

$$\delta_h^t \leq \frac{96HH^{\frac{5}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{t}\Big[1 + \sum_{h'=h+1}^{H}\gamma^{2(H-h'+1)}\Big]$$

$$= \frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{t}\sum_{h'=h}^{H}\gamma^{2(H-h')}$$

$$\leq \frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{t}\sum_{h'=h}^{H}\gamma^{2(H-h'+0.5)}$$

which concludes the induction step. This then leads to;

$$\delta_1^T \leq \frac{96H^{\frac{5}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{T}\sum_{h'=1}^{H}\Big(1+\frac{1}{H}\Big)^{2(H-h'+0.5)}$$

$$\leq \frac{96H^{\frac{7}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{T}\Big(1+\frac{1}{H}\Big)^{2H}$$

$$\leq \frac{96e^2H^{\frac{7}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{T} \leq \frac{864H^{\frac{7}{2}}N\big(\tilde{\alpha}\log T + 2\log|A_{\max}|+2\big)}{T}$$

Then, by referring to the property that CCE-Gap$(\bar{\pi}) \leq \delta_1^T$, we have

$$\text{CCE-Gap}(\bar{\pi}) \leq \frac{864H^{\frac{7}{2}}N\big(\tilde{\alpha}\log T + 2\log|\mathcal{A}_{\max}|+2\big)}{T} \tag{55}$$

By Lemma 12, Corollary 13, and Lemma 3, the policy and value-iteration steps of Algorithms 3 and 1 coincide. As, both Algorithms 3 and 1 use Algorithm 2 for policy execution they are equivalent. Thus, inequality in (55) for Algorithm 3 also holds for Algorithm 1, which completes the proof. ■