Regret Minimization and Convergence to Equilibria in General-sum Markov Games

Liad Erez [*1] **Tal Lancewicki** [*1] **Uri Sherman** [*1] **Tomer Koren** [1 2] **Yishay Mansour** [1 2]

Abstract

An abundance of recent impossibility results establish that regret minimization in Markov games with adversarial opponents is both statistically and computationally intractable. Nevertheless, none of these results preclude the possibility of regret minimization under the assumption that all parties adopt the same learning procedure. In this work, we present the first (to our knowledge) algorithm for learning in general-sum Markov games that provides sublinear regret guarantees when executed by all agents. The bounds we obtain are for *swap regret*, and thus, along the way, imply convergence to a *correlated* equilibrium. Our algorithm is decentralized, computationally efficient, and does not require any communication between agents. Our key observation is that online learning via policy optimization in Markov games essentially reduces to a form of *weighted* regret minimization, with *unknown* weights determined by the path length of the agents' policy sequence. Consequently, controlling the path length leads to weighted regret objectives for which sufficiently adaptive algorithms provide sublinear regret guarantees.

1. Introduction

Multiagent reinforcement learning (MARL; see Busoniu et al., 2008; Zhang et al., 2021) studies statistical and computational properties of learning setups that consist of multiple agents interacting within a dynamic environment. One of the most well studied models for MARL is Markov Games (also known as *stochastic games*, introduced originally by Shapley, 1953), which can be seen as a generalization of a Markov Decision Process (MDP) to the multiagent setup. In this model, the transition dynamics are governed by the *joint* action profile of all agents, implying that the environment as perceived by any individual agent is non-stationary. While providing powerful modeling capabilities, this comes at the cost of marked challenges in algorithm design. Furthermore, in its full generality the model considers multiplayer *general-sum* games, where it is well-known that computing a Nash equilibrium is computationally intractable already in the simpler model of normal form games (Daskalakis et al., 2009; Chen et al., 2009).

Contemporary research works that study general-sum Markov games consider objectives that roughly fall into one of two categories; sample complexity of learning an approximate (coarse) correlated equilibrium, or regret against an arbitrary opponent. The sample complexity setup assumes all players learn using the same algorithm, while in the regret minimization setting, where the vast majority of results are negative (e.g., Bai et al., 2020; Tian et al., 2021; Liu et al., 2022), the opponents are assumed to be adversarial, and in particular do not use the same algorithm as the learner nor attempt to minimize their regret. Curiously, developing (or, asking if there exist) algorithms that minimize individual regret given that all players adopt the same algorithm has been largely overlooked. Considering the intrinsic nature of MARL problems, where agents learn interactively from experience, it is of fundamental interest not only to arrive at an equilibrium, but to control the loss incurred during the learning process. Moreover, this is precisely the objective considered by a long line of works into learning in normal form games (Syrgkanis et al., 2015; Chen & Peng, 2020; Daskalakis et al., 2021; Anagnostides et al., 2022b). Thus, we are motivated to ask;

Can we design algorithms for learning in general-sum Markov games that, when adopted by all agents, provide sublinear individual regret guarantees?

In this work, we answer the above question affirmatively, and present the first (to our knowledge) algorithm for general-sum Markov games which guarantees sublinear regret compared to the best fixed Markov policy in hindsight. We consider finite horizon Markov games in two settings;

[*]Equal contribution [1]Blavatnik School of Computer Science, Tel Aviv University, Israel [2]Google Research, Tel Aviv. Correspondence to: Liad Erez <liaderez@mail.tau.ac.il>, Tal Lancewicki <lancewicki@mail.tau.ac.il>, Uri Sherman <urisherman@mail.tau.ac.il>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

full-information, where access to exact state-action value functions is available, and the unknown model setup (where each agent only observes loss values of visited state-action pairs) with a minimum reachability assumption. In both cases, our algorithm is decentralized and does not require any form of communication between agents. In addition, our bounds apply to the general notion of *swap regret* (Blum & Mansour, 2007), and therefore imply that the empirical distribution (over time steps) of policy profiles generated by our algorithm converges to a correlated equilibrium as the game progresses.

To achieve our results, we make the following observations. In a Markov game, from the point of view of any individual agent, the environment reduces to a single agent MDP in any given episode. When considering multiple episodes, the environment perceived by any individual agent is non-stationary, with the path length of the sequence of policies generated by fellow agents determining the total variation of MDP dynamics. Our first key observation is that, when executing a policy optimization routine (Shani et al., 2020; Cai et al., 2020) in a non-stationary MDP (and thus in a Markov game), the per state objective becomes one of *weighted regret* with weights *unknown* to the learner. Importantly, the total variation of weights in these objectives is governed by the degree of non-stationarity (and in turn, by the path length of the other agents' policies). Therefore, a possible approach would be to provide all agents with an algorithm which has the following two properties; (1) the path length of generated policies is well bounded, and (2) the per state weighted regret is bounded in terms of the total variation of weights (and thus in terms of the policy path length). Indeed, we prove that a carefully designed instantiation of policy optimization with *optimistic-online-mirror-descent* (OOMD; Rakhlin & Sridharan, 2013) produces a bounded path length policy sequence, and simultaneously exhibits the required weighted regret bounds.

Our approach builds on recent progress on decentralized learning in normal form games (Daskalakis et al., 2021; Anagnostides et al., 2022b;a). The work of Anagnostides et al. (2022b) demonstrated that *optimistic-follow-the-regularized-leader* (OFTRL; Syrgkanis et al., 2015), combined with log-barrier regularization and the no-swap-regret meta algorithm of Blum & Mansour (2007), leads to well bounded path length in general-sum normal form games. However, their techniques do not readily extend to the Markov game setup; indeed, FTRL-based algorithms are not sufficiently adaptive and at least in standard form cannot be tuned to satisfy weighted regret bounds. In fact, weighted regret is a generalization of the previously studied objective of *adaptive regret* (Hazan & Seshadhri, 2009), and it can be shown FTRL-based algorithms do not even satisfy this weaker notion (see Hazan & Seshadhri, 2009 and a more elaborate discussion in Appendix G). Evidently, however, an OOMD-based algorithm can be made sufficiently adaptive *and* produce iterates of bounded path length. When all agents adopt our proposed algorithm, the path length of the generated policy sequence remains well bounded, leading to moderate non-stationarity and low total variation per state weighted regret problems, which allows properly tuned mirror descent steps—crucially, *without* knowledge of the weights—to obtain sublinear regret. Notably, while much of the previous works (e.g., Syrgkanis et al., 2015; Chen & Peng, 2020; Anagnostides et al., 2022b) employ optimistic online algorithms and path length dependent regret bounds to improve upon naive square-root regret, in Markov games, with our approach, these are actually crucial for obtaining any form of sublinear regret.

1.1. Summary of contributions

To summarize, we present a decentralized algorithm for (multiplayer, tabular, and episodic) Markov games, with the following guarantees when adopted by all agents.

- In the full-information setting, with access to exact state-action value functions, the individual swap regret of every agent is $\widetilde{O}(T^{3/4})$ (see Section 3).
- In the unknown model setup, subject to reachability assumptions (see Assumption 4.1), we obtain individual swap regret of $\widetilde{O}(T^{8/9})$ (see Section 4).
- In the special case of full-information independent transition function where agents only affect the loss functions of each other but not the transition dynamics, our algorithm guarantees $O(\log T)$ individual regret. The result is relatively straight forward given our analysis for general Markov games, and we defer the formal setting and proofs to Appendix I.
- As an immediate implication, we obtain that the joint empirical distribution of policy profiles produced by our algorithm converges to a correlated equilibrium, at a rate of $\widetilde{O}(T^{-1/4})$ in the full-information setting, $\widetilde{O}(T^{-1/9})$ in the unknown model setting, and $\widetilde{O}(1/T)$ in the independent transition function setting.

1.2. Related work

Learning in Markov games. The framework of Markov games was originally introduced by Shapley (1953). The majority of studies consider learning Nash equilibria in two-player zero-sum Markov games, and may be roughly categorized by assumptions made on the model. The full-information setting, where the transition function is known and/or some sort of minimum state reachability is assumed, has gained much of the earlier attention (Littman, 1994; Littman et al., 2001; Brafman & Tennenholtz, 2002; Hu & Wellman, 2003; Hansen et al., 2013; Wei et al., 2017), as well as more recent (Daskalakis et al., 2020; Wei et al., 2021; Cen et al., 2021; Zhao et al., 2022; Alacaoglu et al., 2022; Zhang et al., 2022). The unknown model setup, where the burden of exploration is entirely in the hands of the agent, has been a target of several recent papers focusing on sample complexity (Sidford et al., 2020; Bai et al., 2020; Xie et al., 2020; Zhang et al., 2020; Liu et al., 2021).

The work of Wei et al. (2021) considers zero-sum games with model assumptions similar to ours, and present an optimistic gradient descent-ascent policy optimization algorithm with a smoothly moving critic. They obtain last iterate convergence to a Nash equilibrium at a rate of $\widetilde{O}(T^{-1/2})$ for the full-information setting, which immediately implies individual regret of $\widetilde{O}(T^{1/2})$. In the unknown model setup with reachability assumptions, their algorithm obtains a last iterate guarantee that implies $\widetilde{O}(T^{7/8})$ regret. Also noteworthy, Tian et al. (2021) consider the zero-sum unknown model setting, and develop an algorithm that provides a $O(T^{2/3})$ regret guarantee when comparing to the minimax game value. Tian et al. (2021) also present a certain extension of their result to general-sum games, however their definition of regret in this case does not translate to the usual notion of regret even when all players adopt their algorithm.

Learning in *general-sum* Markov games has been comparatively less explored. The work of Liu et al. (2021) presented a *centralized* algorithm in the unknown model setup with optimal sample complexity guarantees in terms of the number of episodes, but exponential dependence on the number of agents. Following their work, several recent papers (Jin et al., 2021; Song et al., 2021; Mao & Başar, 2022) independently develop variants of V-learning, a decentralized algorithm for learning unknown general-sum Markov games. After T episodes, their algorithms output a (non-Markov) $O(T^{-1/2})$-coarse correlated equilibrium, without dependence on the number of agents. Jin et al. (2021) and Song et al. (2021) also present extensions for obtaining approximate (non-coarse) correlated equilibrium with similar guarantees. Later, Mao et al. (2022) further propose simplifications to the V-learning algorithmic and analysis framework. Notably though, the output of these algorithms is linear in the number of episodes (as it includes the history of all policies), and it is unclear what are the online guarantees of these methods. In a full-information setting similar to ours, the recent work of Zhang et al. (2022) presents an algorithm that outputs a $\tilde{O}(T^{-3/4})$-optimal policy after T episodes for general-sum games; notably, however, it is unclear whether their algorithm provides regret guarantees.

No-regret learning in games. Theoretically understanding no-regret dynamics in multiplayer games has been a topic of vast interest in recent years (e.g., Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015; Foster et al., 2016; Chen & Peng, 2020; Daskalakis et al., 2021; Anagnostides et al., 2022b; Piliouras et al., 2021). The main focus in most of these works is to analyze the performance of optimistic variants of online learning algorithms such as FTRL and OMD in multiplayer normal form games, and ultimately prove regret bounds which are vastly better than the naive $O(\sqrt{T})$ guarantee achievable in adversarial environments. The state-of-the-art result in this setting was established by Anagnostides et al. (2022b) who proposed an algorithm which guarantees $O(\log T)$ swap regret in general-sum games. Some of these results have been extended to more general classes of games such as extensive-form games (Farina et al., 2022b; Anagnostides et al., 2022a) and convex games (Farina et al., 2022a). In this work we adopt some of the techniques presented by Anagnostides et al. (2022b) in order to establish sublinear swap regret guarantees in general-sum Markov games.

Hardness results for Markov games. Learning in Markov games is considered a notoriously challenging problem, and several learning objectives have been shown in previous works to be either computationally or statistically hard. For instance, Bai et al. (2020) show that computing the best response policy in zero-sum Markov games against an adversarial opponent is at least as hard as learning parities with noise, a problem conjectured to be computationally hard. Tian et al. (2021) and Liu et al. (2022) show a regret lower bound of $\Omega(\min\{\sqrt{2^H T}, T\})$ for zero-sum episodic Markov games with an unknown transition function, where the opponent is restricted to Markov policies. We note that these hardness results do not directly impact our goal of no-regret learning in general-sum Markov games, as they consider the setting of facing an arbitrary opponent which is only constrained to play Markov policies. By contrast, our main result shows that each player's individual regret is sublinear in T as long as the other players' policies have a well bounded second-order path length, which is a property enforced by our choice of algorithm for all players. Additionally, Daskalakis et al. (2022) show that the problem of computing a coarse correlated equilibrium comprised of stationary Markov policies in a general-sum infinite horizon Markov game is computationally hard. We consider a setting of regret minimization in layered episodic Markov games, and though our policies of interest are stationary, they do not translate into stationary policies in a corresponding infinite horizon Markov game. Hence, this lower bound is not applicable in the setting we consider here. Finally, in a recent work, Foster et al. (2023) consider a more general notion of regret in general sum Markov games, where each player's benchmark policy may be non-Markovian. Specifically, they show that obtaining sublinear regret in this regime is both computationally and statistically hard. In the setting we consider in this paper, each player's performance is compared to the best Markov policy in hindsight, and therefore we manage to obtain sublinear regret despite this result.

Zero sum Markov games. The recent work of Zhang et al. (2022) that was mentioned earlier also presents an algorithm for the zero sum setting with an $\tilde{O}(T^{-5/6})$ convergence rate. This was later improved in Yang & Ma (2022) to $\tilde{O}(T^{-1})$ and then later by Cen et al. (2022) who also establish last iterate convergence (at the same rate) and improve dependence on the horizon. It is worth noting Pattathil et al. (2022) who study the zero sum setting with function approximation, and Giannou et al. (2022) who prove *local* convergence to Nash equilibria for independent policy gradients in a general sum Markov game setting.

Regret in the policy revealing setting. The recent works of Liu et al. (2022); Zhan et al. (2022) explore the *policy-revealing* setting, where agents share their policies after every episode. Liu et al. (2022) give both positive and negative results on regret guarantees in this setup for zero-sum games, with the regret upper bounds depending on the cardinality of either the baseline or opponent policy classes. Zhan et al. (2022) extend their work and present an algorithm for the policy-revealing setting with function approximation, which achieves no-regret in general-sum games in face of arbitrary opponents, as long as these reveal their policies at the end of each episode. Importantly, in both Liu et al. (2022) and Zhan et al. (2022) the computational complexity depends on the cardinality of the baseline policy class, and thus their algorithm is inefficient whenever the baseline policy class is the class of all Markov policies, as in our case.

We remark that a possible approach for regret minimization in Markov games is via a decentralized algorithm with a sample complexity guarantee (e.g. Jin et al. (2021)), together with an explore-exploit mechanism. However, such an approach would require a correlation device between agents during the exploit phase, an thus would be effectively centralized. For a further discussion, see Appendix A.

2. Preliminaries

Markov games. An m-player general-sum finite horizon Markov game is defined by the tuple $\left(H, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, P, \{\ell^i\}_{i=1}^m\right)$. H is the horizon; $\mathcal{S}$ is set of states of size S partitioned as $\mathcal{S} = \bigcup_{h=1}^{H+1} \mathcal{S}_h$, where $\mathcal{S}_1 = \{s_1\}$ and $\mathcal{S}_{H+1} = \{s_{H+1}\}$; $\mathcal{A}_i$ is the set of actions of agent i of size A_i, and the joint action space is denoted by $\mathcal{A} := \bigtimes_{i=1}^m \mathcal{A}_i$. Further, P is the transition kernel, where given the state at time h, $s \in \mathcal{S}_h$, and a joint action profile $\mathbf{a} \in \mathcal{A}$, $P(\cdot \mid s, \mathbf{a}) \in \Delta_{\mathcal{S}_{h+1}}$ is the probability distribution over the next state, where given some set C, $\Delta_C := \left\{p : C \to [0,1] \mid \sum_{x \in C} p(x) = 1\right\}$ denotes the probability simplex over C. Finally, $\ell^i : \mathcal{S} \times \mathcal{A} \to [0,1]$ denotes the cost function of agent i. A policy for player i is a function $\pi^i(\cdot \mid \cdot) : \mathcal{A}_i \times \mathcal{S} \to [0,1]$, such that $\pi^i(\cdot \mid s) \in \Delta_{\mathcal{A}_i}$ for all $s \in \mathcal{S}$. Given a policy profile $\boldsymbol{\pi} = (\pi^1, ..., \pi^m)$, player $i \in [m]$, state $s \in \mathcal{S}_h$ and action $a \in \mathcal{A}_i$, we define the value function and the Q-function of agent i by:

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

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

Interaction protocol. The agents interact with the Markov game over the course of T episodes. At the beginning of each episode $t \in [T]$ every agent chooses a policy π_t^i. Then, all agents start at the initial state s_1, and for each time step $h = 1, 2, \ldots, H$, each player draws an action $a_h^i \sim \pi_t^i(\cdot \mid s_h)$ and the agents transition together to the next state $s_{h+1} \sim P(\cdot \mid s_h, \mathbf{a}_h)$ where $\mathbf{a}_h = (a_h^1, ..., a_h^m)$. At the end of the episode, agent i incurs a loss given by $\sum_{h=1}^H \ell^i(s_h, \mathbf{a}_h)$, and observes feedback that differs between two distinct settings we consider. In the ***full-information*** setup, agent $i \in [m]$ observes the exact state-action value functions; $Q^{i,\pi_t}(s,a), \quad \forall s, a \in \mathcal{S} \times \mathcal{A}_i$. In the ***unknown model*** setup, agent i only observes losses for the state and action profiles visited in the episode; $\{\ell^i(s_h, \mathbf{a}_h)\}_{h=1}^H$, and we additionally assume the Markov game satisfies a minimum reachability condition (Assumption 4.1).

Learning objective. Given an agent $i \in [m]$ and policy profile $\boldsymbol{\pi} = (\pi^1, ..., \pi^m) = \pi^1 \odot \pi^2 \odot \ldots \odot \pi^m$, we will be interested in the policy profile excluding i which we denote by $\boldsymbol{\pi}^{-i} := \pi^1 \odot \ldots \pi^{i-1} \odot \pi^{i+1} \odot \ldots \odot \pi^m$. For policy profile $\boldsymbol{\pi}$ and a player i policy $\pi \in \mathcal{S} \to \mathcal{A}_i$, we let $\pi \odot \boldsymbol{\pi}^{-i} = \pi^1 \odot \ldots \pi^{i-1} \odot \pi \odot \pi^{i+1} \odot \ldots \odot \pi^m$ denote the joint policy formed by replacing π^i with π. Given an episode t and policy profile $\boldsymbol{\pi}_t = (\pi_t^1, ..., \pi_t^m)$, we will be interested in the single agent MDP induced by $\boldsymbol{\pi}_t^{-i}$. This induced MDP is specified by $M_t^i := (H, \mathcal{S}, \mathcal{A}_i, P_t^i, \ell_t^i)$, where $\ell_t^i(s,a) := \mathbb{E}_{\mathbf{a} \sim \boldsymbol{\pi}_t(\cdot|s)}\left[\ell^i(s, \mathbf{a}) \mid a^i = a\right]$ and $P_t^i(\cdot \mid s, a) = \mathbb{E}_{\mathbf{a} \sim \boldsymbol{\pi}_t(\cdot|s)}\left[P(\cdot \mid s, \mathbf{a}) \mid a^i = a\right]$ define agent i's induced loss vector and transition kernel respectively. Furthermore, we denote the value and action-value functions of a policy $\pi \in \mathcal{S} \to \Delta_{\mathcal{A}_i}$ in this MDP by

$$V_t^{i,\pi}(s) := V^{i, \pi \odot \boldsymbol{\pi}_t^{-i}}(s); Q_t^{i,\pi}(s,a) := Q^{i, \pi \odot \boldsymbol{\pi}_t^{-i}}(s,a),$$

where $s \in \mathcal{S}_h$ and $a \in \mathcal{A}_i$. Given our definitions above, a standard argument shows that $V_t^{i,\pi}(s) = \mathbb{E}\left[\sum_{h'=h}^H \ell_t^i(s_{h'}, a_{h'}) \mid P_t^i, \pi, s_h = s\right]$, and $Q_t^{i,\pi}(s,a) = \mathbb{E}\left[\sum_{h'=h}^H \ell_t^i(s_{h'}, a_{h'}) \mid P_t^i, \pi, s_h = s, a_h = a\right]$. We note that we sometimes use the shorthand $V_t^i(\cdot)$ for $V_t^{i,\pi_t^i}(\cdot)$ and

$Q_t^i(\cdot,\cdot)$ for $Q_t^{i,\pi_t^i}(\cdot,\cdot)$. Given a jointly generated policy sequence $\{\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_T\}$, our primary performance measure is the individual swap regret of each player i, defined as

$$\mathrm{Swap}\mathfrak{R}_T^i = \max_{\phi_\star^i\in\{\mathcal{S}\times\mathcal{A}_i\to\mathcal{A}_i\}}\left\{\sum_{t=1}^T\left(V_t^{i,\pi_t^i}(s_1)-V_t^{i,\phi_\star^i(\pi_t^i)}(s_1)\right)\right\}, \tag{1}$$

where we slightly overload notation and define a policy swap function $\phi\colon\mathcal{S}\times\mathcal{A}_i\to\mathcal{A}_i$ applied to a policy $\pi\in\mathcal{S}\to\Delta_{\mathcal{A}_i}$ as follows;

$$\phi(\pi)(a\mid s)\coloneqq\sum_{a'\colon\phi(s,a')=a}\pi(a'\mid s).$$

That is, the distribution $\phi(\pi)(\cdot\mid s)$ is formed by sampling $a\sim\pi(\cdot\mid s)$ and then replacing it with $\phi(s,a)\in\mathcal{A}_i$. Similarly, given an action swap function $\psi^i\colon\mathcal{A}_i\to\mathcal{A}_i$, we slightly overload notation when applying it to $x\in\Delta_{\mathcal{A}_i}$ by defining $\psi^i(x)(a)=\sum_{a'\colon\psi^i(a')=a}x(a')$. We remark this notion of regret is strictly stronger (in the sense that it is always greater or equal) than the external regret against the best fixed Markov policy in hindsight, defined by;

$$\mathfrak{R}_T^i\coloneqq\max_{\pi_\star^i\in\{\mathcal{S}\to\Delta_{\mathcal{A}_i}\}}\left\{\sum_{t=1}^T\left(V_t^{i,\pi_t^i}(s_1)-V_t^{i,\pi_\star^i}(s_1)\right)\right\}. \tag{2}$$

Finally, a joint policy distribution Π is an ε-approximate *Markov correlated equilibrium* if for any player i,

$$\mathbb{E}_{\boldsymbol{\pi}\sim\Pi}\left[\max_{\phi_i}\left(V^{i,\boldsymbol{\pi}}(s_1)-V^{i,(\phi_i(\pi^i)\odot\boldsymbol{\pi}^{-i})}(s_1)\right)\right]\le\varepsilon. \tag{3}$$

It is straightforward to show that if all players achieve swap regret of $O(\varepsilon T)$ over T episodes, then the distribution given by sampling $\boldsymbol{\pi}_t$ with $t\sim[T]$ uniformly constitutes an ε-approximate Markov correlated equilibrium (Blum & Mansour, 2007).

Additional notation and definitions. We denote the size of the largest action set as $A\coloneqq\max_i A_i$. In addition, we let $q_t^{i,\pi}$ denote the state-occupancy measure of policy π in M_t^i;

$$q_t^{i,\pi}(s)\coloneqq\Pr(s_h=s\mid P_t^i,\pi).$$

For any pair of policies $\pi,\tilde\pi\in\mathcal{S}\to\Delta_{\mathcal{A}_i}$ of player i, we define

$$\|\pi-\tilde\pi\|_{\infty,1}\coloneqq\max_{s\in\mathcal{S}}\|\pi(\cdot\mid s)-\tilde\pi(\cdot\mid s)\|_1,$$

and for any $P,\tilde P\in\mathcal{S}\times\mathcal{A}_i\to\Delta_{\mathcal{S}}$,

$$\|P-\tilde P\|_{\infty,1}\coloneqq\max_{s\in\mathcal{S},a\in\mathcal{A}_i}\|P(\cdot\mid s,a)-\tilde P(\cdot\mid s,a)\|_1.$$

In addition, for a sequence of policies $\pi_0,\pi_1,\ldots,\pi_T$, we refer to the quantities $\sum_{t=1}^T\|\pi_t-\pi_{t-1}\|_1$ and $\sum_{t=1}^T\|\pi_t-\pi_{t-1}\|_1^2$ as their *(first-order) path length* and *second-order path length* respectively.

Finally, the notations $\widetilde O(\cdot),\widetilde\Omega(\cdot),\widetilde\Theta(\cdot)$ and $\lesssim$ hide constant and poly-logarithmic factors.

Optimistic online mirror descent Let $\mathcal{X}\subset\Delta_d$ be a convex subset of the d-dimensional simplex, and $\ell_1,\ldots,\ell_T\in[0,1]^d$ be an online loss sequence. Optimistic online mirror descent (OOMD) over $\mathcal{X}$ with convex regularizer $R\colon\mathcal{X}\to\mathbb{R}$ and learning rate $\eta>0$ is defined as follows:

$$\begin{aligned}
&\tilde x_0\leftarrow\arg\min_{x\in\mathcal{X}}R(x);\\
t=1,\ldots,T;\quad &x_t\leftarrow\arg\min_{x\in\mathcal{X}}\left\{\langle\tilde\ell_t,x\rangle+\frac{1}{\eta}D_R(x,\tilde x_{t-1})\right\},\\
&\tilde x_t\leftarrow\arg\min_{x\in\mathcal{X}}\left\{\langle\ell_t,x\rangle+\frac{1}{\eta}D_R(x,\tilde x_{t-1})\right\}.
\end{aligned}$$

We instantiate OOMD with 1-step recency bias, meaning $\tilde\ell_1\coloneqq 0$ and $\tilde\ell_t\coloneqq\ell_{t-1}$ for $t\ge 2$. The primal and dual local norms induced by the regularizer R are denoted;

$$\forall v\in\mathbb{R}^d,x\in\mathcal{X},\qquad\|v\|_x=\sqrt{v^\top\nabla^2R(x)v};\qquad\|v\|_{*,x}=\sqrt{v^\top(\nabla^2R(x))^{-1}v}.$$

For the most part, we will employ the log-barrier regularization specified by

$$\forall x\in\mathcal{X},\quad R(x)=\sum_{a\in[d]}\log\frac{1}{x(a)}. \tag{4}$$

The Bregman divergence induced by the log-barrier is given by $D_R(x,y)=\sum_{a\in[d]}\log\frac{y(a)}{x(a)}+\frac{x(a)-y(a)}{y(a)}$, and the local norms by

$$\|v\|_x=\sqrt{\sum_{a\in[d]}\frac{v(a)^2}{x(a)^2}};\ \|v\|_{*,x}=\sqrt{\sum_{a\in[d]}v(a)^2x(a)^2},$$

$\forall v\in\mathbb{R}^d,x\in\mathcal{X}$. Finally, throughout, we refer to the γ-*truncated* simplex, defined by

$$\Delta_d^\gamma\coloneqq\{x\in\Delta_d\mid x(a)\ge\gamma,\quad\forall a\in[d]\}. \tag{5}$$

3. Algorithm and main result

In this section, we present our algorithm and outline the analysis establishing the regret bound. We propose a policy optimization method with a carefully designed regret minimization algorithm employed in each state. Specifically, inspired by the work of Anagnostides et al. (2022b), we equip the swap regret algorithm of Blum & Mansour (2007) with a variant of optimistic online mirror descent over the truncated action simplex. This choice has two important properties; First, it can be shown that online mirror descent (as well as its optimistic variant) with some tuning satisfies weighted regret bounds of the form that emerges from non-stationarity in MDP dynamics which directly depends on the path length of the joint policy sequence. Additionally, the second-order path length of the generated policy sequence is $O(\log T)$, a fact we establish by suitable modifications of the arguments presented in Anagnostides et al. (2022b).

Algorithm 1 Policy Optimization by Swap Regret Minimization

1: **input:** $H, \mathcal{S}, \mathcal{A}_i, T$ agent index i, parameter $\gamma > 0$, learning rate $\eta > 0$, regularizer $R(\cdot)$.
2: **initialization:** π_1^i is the uniform policy. For every $s \in \mathcal{S}$ and every $a \in \mathcal{A}_i$ initialize $\tilde{x}_0^{i,s,a} = \arg\min_{x \in \Delta_{\mathcal{A}_i}^{\gamma}} R(x)$.
3:
4: **for** $t = 1$ to T **do**
5: Play policy π_t^i
6: Observe an ε-approximation of Q_t^i denoted by $\hat{Q}_t^i$
7: Incur the expected loss of the policy π_t^i with respect to the losses ℓ_t^i: $V_t^i(s_0)$.
8:
9: # Optimistic OMD step
10: For every s, a perform an optimistic OMD update with the loss vector $g_t^{i,s,a} := \pi_t^i(a \mid s)\hat{Q}_t^i(s, \cdot)$:

$$\tilde{x}_t^{i,s,a} = \arg\min_{x \in \Delta_{\mathcal{A}_i}^{\gamma}} \left\{ \eta \left\langle x, g_t^{i,s,a} \right\rangle + D_R\left(x, \tilde{x}_{t-1}^{i,s,a}\right) \right\}$$
$$x_{t+1}^{i,s,a} = \arg\min_{x \in \Delta_{\mathcal{A}_i}^{\gamma}} \left\{ \eta \left\langle x, g_t^{i,s,a} \right\rangle + D_R\left(x, \tilde{x}_t^{i,s,a}\right) \right\}$$

11: # Policy update
12: **for** $s \in \mathcal{S}$ **do**
13: # Calculate $\pi_{t+1}^i(\cdot \mid s)$ - the stationary distribution corresponding to $\left\{x_{t+1}^{i,s,a}\right\}_{a \in \mathcal{A}_i}$
14: Let B be the matrix whose rows are $\left\{x_{t+1}^{i,s,a}\right\}_{a \in \mathcal{A}_i}$
15: Compute $\pi_{t+1}^i(\cdot \mid s) \in \Delta_{\mathcal{A}_i}^{\gamma}$ by solving

$$B\pi_{t+1}^i(\cdot \mid s) = \pi_{t+1}^i(\cdot \mid s)$$

16: **end for**
17: **end for**

We remark that the policy update step in Algorithm 1 can be performed in polynomial time, since it only requires solving a system of linear equations under linear inequality constraints. We refer to the components of the algorithm which perform the OOMD steps at a given state and action (see line 10 of Algorithm 1) as *base algorithms*. On the level above the base algorithms, the components which perform the policy update from each state are referred to as *state algorithms*. The guarantee of Algorithm 1 is provided in the statement of Theorem 3.1 below. The important consequence is that when the Markov game's transition kernel is known, i.e., the players have access to the accurate induced Q-functions, sub-linear regret of $\widetilde{O}(T^{3/4})$ is achieved. By employing relatively standard arguments this may be extended (under a suitable reachability assumption) to a sub-linear regret bound in the unknown dynamics setting, which we present in Section 4.

Theorem 3.1. *Assume $H \geq 2$, and that all players adopt Algorithm 1 with log-barrier regularization (Equation (4)), $\gamma \leq 1/2A$ (recall that we define $A = \max_i A_i$) and step size $\eta = \frac{1}{96H^2 m\sqrt{SA}}$, and that $\|\hat{Q}_t^i - Q_t^i\|_\infty \leq \varepsilon$ for all i, t. Then, the swap regret of every player i is bounded as*

$$\begin{aligned}\mathrm{Swap}\mathfrak{R}_T^i \lesssim\ & H^4 S A^3 m^2 \sqrt{T}/\gamma + mH\sqrt{S}A^{3/2}\varepsilon T/\gamma \\ & + \gamma A H^2 T + mH^2 S^{3/2} A^{7/2}.\end{aligned}$$

In particular, if $\varepsilon = 0$, for the choice of $\gamma = mH\sqrt{S}AT^{-1/4}$, and $T \geq 16m^4H^4S^2A^8$ we obtain;

$$\mathrm{Swap}\mathfrak{R}_T^i = \tilde{O}\left(mH^3\sqrt{S}A^2T^{3/4}\right).$$

Theorem 3.1 hinges on two separate avenues in the analysis, each of which corresponds to the two properties mentioned earlier. The first, outlined in Section 3.1 and Section 3.2, establishes the individual regret may be bounded by the path length of the jointly generated policy sequence. The second avenue, presented in Section 3.3 largely follows the techniques of Anagnostides et al. (2022b), and establishes the jointly generated policy sequence indeed has a well bounded path length. The proof of Theorem 3.1 proceeds by standard arguments building on the aforementioned parts of the analysis, and is deferred to Section 3.4. To conclude this section, we obtain as an immediate corollary that the players' joint policy sequence converges to an approximate Markov correlated equilibrium (Equation (3)) of the Markov game.

Corollary 3.2 (Convergence to a correlated equilibrium)**.** *If all players adopt Algorithm 1 with the parameters and conditions specified in Theorem 3.1, then for every player i it holds that*

$$\begin{aligned}&\mathbb{E}_{t\sim[T]}\left[\max_{\phi_i}\left(V_t^{i,\pi_t^i}(s_1) - V_t^{i,\phi_i(\pi_t^i)}(s_1)\right)\right] \\ &\quad = \tilde{O}\left(mH^3\sqrt{S}A^2T^{-1/4}\right),\end{aligned}$$

where $t \sim [T]$ denotes the uniform distribution over $[T] = \{1, 2, \ldots, T\}$.

The proof follows immediately from the observation that the left-hand-side is $\mathrm{Swap}\mathfrak{R}_T^i/T$, and applying the result of Theorem 3.1.

3.1. Policy optimization in non-stationary MDPs and weighted regret

The analysis of policy optimization algorithms is often built upon a fundamental regret decomposition known as the value difference lemma (see Lemma F.1). In the regime of regret minimization with a stationary transition function, the value difference lemma leads to S regret expressions, each weighted by a constant factor, and thus amenable to standard analysis (e.g., Shani et al., 2020; Cai et al., 2020). By contrast, in the Markov game setup, the single agent induced MDP is essentially non-stationary, thus applying the value difference lemma results in a sum of *weighted regret* expressions, with weights given by the (changing) occupancy measure of the benchmark policy, and in particular are not known to the learner. A natural complexity measure of weighted regret minimization is the total variation of weights — roughly speaking, the larger the total variation the harder the problem becomes (for example, note that with zero total variation the problem reduces to standard regret minimization).

Thus, we are motivated to consider a weighted regret objective with unknown weights, which may be seen as a generalization to the previously studied notion of *adaptive regret* (Hazan & Seshadhri, 2009). As it turns out, the widely used FTRL for example, might suffer linear regret even if the weights' total variation is as small as a constant (see the discussion in Hazan & Seshadhri (2009) and in Appendix G). On the other hand, we show that OMD-style algorithms are resilient to weighting with small variations, as long as the Bregman divergence between the benchmark policy and the iterates of the algorithm is bounded and the step size is chosen appropriately. Lemma 3.3 below establishes a weighted regret bound for optimistic OMD. The proof follows from arguments typical to OMD analyses, and is thus deferred to Appendix B.

Lemma 3.3. *Assume we run OOMD on a sequence of losses $\{\ell_t\}_{t=1}^T$ with a regularizer $R\colon \mathcal{X} \to \mathbb{R}$ that is 1-strongly convex w.r.t. $\|\cdot\|$. Then, for any $x^\star \in \mathcal{X}$ and any weight sequence $\{q_t\}_{t=1}^T$, it holds that*

$$\sum_{t=1}^T q_t \left\langle x_t - x^\star, \ell_t \right\rangle \le \frac{q_1 D_R(x^\star, \tilde{x}_0)}{\eta} + \frac{1}{\eta}\sum_{t=1}^T (q_{t+1} - q_t) D_R(x^\star, \tilde{x}_t) + \frac{\eta}{2}\sum_{t=1}^T q_t \|\ell_t - \tilde{\ell}_t\|_*^2.$$

3.2. Bounding regret by path length

Our algorithm runs an instance of the meta algorithm of Blum & Mansour (2007) in each state with OOMD as a base algorithm. It is straightforward to show that the meta algorithm inherits the desired property of weighting-resilience (see Theorem D.1). By relating the weights $q_t^{i,\star}(s)$ with the *first-order* path length we are able to show the following regret bound.

Theorem 3.4. *Suppose every player i adopts Algorithm 1 with log-barrier regularization (Equation (4)) and $\gamma \le 1/2A_i$, and that $\|\hat{Q}_t^i - Q_t^i\|_\infty \le \varepsilon$ for all t. Then, assuming $H \ge 2$, the swap-regret of player i is bounded as*

$$\mathrm{Swap}\mathfrak{R}_T^i \lesssim \frac{SA_i^2}{\eta} + \frac{A_i H^2}{\eta\gamma}\sum_{j=1}^m \sum_{t=1}^T \|\pi_{t+1}^j - \pi_t^j\|_{\infty,1} + \eta m S A_i H^4 \sum_{t=1}^T \sum_{j=1}^m \|\pi_{t+1}^j - \pi_t^j\|_{\infty,1}^2 + \mathcal{E},$$

where $\mathcal{E} := O(\eta\varepsilon^2 ST + \varepsilon HT + \gamma AH(H+\varepsilon)T)$.

Below we provide the main ideas of the proof of Theorem 3.4; for a full proof, see Appendix C.

Proof (sketch). We begin by invoking a standard value difference lemma (Lemma F.1) and using the fact that $\hat{Q}_t^i$ is an ε-approximation of Q_t^i to bound the swap regret as follows:

$$\sum_{t=1}^T V_t^{i,\pi_t^i}(s_1) - V_t^{i,\phi_\star(\pi_t^i)}(s_1) \le \sum_{s\in\mathcal{S}}\sum_{t=1}^T q_t^{i,\star}(s)\left\langle \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star^{i,s}(\pi_t^i(\cdot \mid s))\right\rangle + \varepsilon HT.$$

We now make use of Theorem D.1 which is based on an argument by Blum & Mansour (2007), and relates the per-state weighted swap-regret to the sum of per-state weighted external regrets of the base algorithms:

$$\sum_{s\in\mathcal{S}}\sum_{t=1}^T q_t^{i,\star}(s)\left\langle \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star^{i,s}(\pi_t^i(\cdot \mid s))\right\rangle \le \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^T q_t^{i,\star}(s)\left\langle g_t^{i,s,a}, x_t^{i,s,a} - x_{\star,\gamma}^{i,s,a}\right\rangle + 2\gamma A_i H(H+\varepsilon)T.$$

We can now invoke Lemma 3.3 which bounds the weighted regret of optimistic OMD, to bound the last term by;

$$\le 2\gamma A_i H(H+\varepsilon)T + \frac{SA_i^2 \ln\frac{1}{\gamma}}{\eta} + \frac{1}{\eta}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^T (q_{t+1}^{i,\star}(s) - q_t^{i,\star}(s)) D_R(x_{\star,\gamma}^{i,s,a}, \tilde{x}_t^{i,s,a}) + \frac{\eta}{2}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^T \|g_{t+1}^{i,s,a} - g_t^{i,s,a}\|_\infty^2,$$

where we have used the fact that the first Bregman term is bounded by $A_i \log \frac{1}{\gamma}$. We proceed by using the boundedness of the Bregman terms over the γ-truncated simplex, and applying Lemma F.3 which relates the weights' total variation to the first-order path length:

$$\frac{1}{\eta}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}(q_{t+1}^{i,\star}(s)-q_t^{i,\star}(s))D_R(x_{\star,\gamma}^{i,s,a},\tilde{x}_t^{i,s,a})$$
$$\leq \frac{3A_i}{\eta\gamma}\sum_{t=1}^{T}\|q_{t+1}^{i,\star}-q_t^{i,\star}\|_1$$
$$\leq \frac{3A_iH^2}{\eta\gamma}\sum_{t=1}^{T}\sum_{j\in[m]}\|\pi_{t+1}^j-\pi_t^j\|_{\infty,1}.$$

We conclude by using Lemma B.1 which relates the second-order loss variation term to the second-order path length:

$$\frac{\eta}{2}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}\|g_{t+1}^{i,s,a}-g_t^{i,s,a}\|_\infty^2$$
$$\leq 4\eta mSA_iH^4\sum_{t=1}^{T}\sum_{j=1}^{m}\|\pi_{t+1}^j-\pi_t^j\|_{\infty,1}^2+4\eta\varepsilon^2ST.$$

Plugging both bounds into the swap regret bound, we conclude the proof. □

3.3. Bounding the path length

In this section, we provide a brief outline of the path length analysis, central to our regret bound. At a high level, the arguments closely follow those of Anagnostides et al. (2022b), with suitable adjustments made to accommodate the elements in which our setting differs. Specifically, these include the use of mirror descent (rather than OFTRL) over the truncated simplex, the fact that a single iterate of each player is comprised of outputs from a collection of S state algorithms, and that each of these algorithms optimizes w.r.t. approximate Q-functions rather than the true ones. Below, we state the main theorem and subsequently describe the analysis at a high level, with most of the technical details deferred to Appendix D.

Theorem 3.5. *If each player uses Algorithm 1 with $\eta = \frac{1}{96H^2m\sqrt{SA}}$ then the following second-order path length bound holds on the jointly generated policy sequence;*

$$\sum_{t=1}^{T}\sum_{i=1}^{m}\left\|\pi_{t+1}^i-\pi_t^i\right\|_{\infty,1}^2 \lesssim SA^3m+\frac{\varepsilon^2T}{mH^4}.$$

Similar to Anagnostides et al. (2022b), Theorem 3.5 is derived by establishing that the swap regret (defined in formally in Appendix D) of each state algorithm satisfies an RVU property (**R**egret bounded by **V**ariation in **U**tilities; originally defined in Syrgkanis et al. (2015)) with suitable norms. Following that, we use the non-negativity of the swap regret in order to arrive at the desired conclusion by a simple algebraic manipulation. In order to establish the aforementioned RVU property of the state algorithms, the first step consists of showing the base algorithms satisfy an RVU property with the local norms induced by the log-barrier.

Lemma 3.6 (RVU property of OOMD with log-barrier and local norms)**.** *Let $[d]$ represent an action set with d actions, and consider an online loss sequence $g_1,\ldots,g_T\in[0,H]^d$. Assume we run OOMD over $\mathcal{X}=\Delta_d^\gamma$ with log-barrier regularization and learning rate $\eta\leq\frac{1}{64H}$. Then, for all $x^\star\in\mathcal{X}$ the following regret bound holds:*

$$\sum_{t=1}^{T}\left\langle x_t-x^\star,g_t\right\rangle\leq\frac{d\log\frac{1}{\gamma}}{\eta}+4\eta\sum_{t=1}^{T}\|g_t-g_{t-1}\|_{*,x_t}^2$$
$$-\frac{1}{576\eta}\sum_{t=1}^{T}\|x_t-x_{t-1}\|_{x_t}^2,$$

where we define $p_0=\arg\min_{x\in\mathcal{X}}R(x)$.

Using Lemma 3.6 and the swap regret guarantee of Blum & Mansour (2007), we arrive at an RVU-like property for each state algorithm's swap regret, one that involves the local norms of base algorithms; $\mathrm{Swap}\mathfrak{R}_T^{i,s}\leq$

$$\frac{|\mathcal{A}_i|^2\log\frac{1}{\gamma}}{\eta}+4\eta\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_i}\left\|g_t^{i,s,a}-g_{t-1}^{i,s,a}\right\|_{*,x_t^{i,s,a}}^2$$
$$-\frac{1}{576\eta}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_i}\left\|x_t^{i,s,a}-x_{t+1}^{i,s,a}\right\|_{x_t^{i,s,a}}^2.$$

The negative term in the right-hand side can be converted to the state algorithm's second-order path length in L_1-norm using arguments similar to those of Anagnostides et al. (2022b) (see Lemma D.5). Each base algorithms' path length of the loss vectors may also be related to its second-order policy path length, by invoking Lemma B.1. This, combined with some standard arguments, leads to a swap-regret RVU property of each state algorithm w.r.t. the $\|\cdot\|_{\infty,1}$ and $\|\cdot\|_1$ norms.

Theorem 3.7. *If all players play according to Algorithm 1 with $\eta\leq\frac{1}{128H}$, then for any $i\in[m]$, $s\in\mathcal{S}$, we have that the swap regret in s, $\mathrm{Swap}\mathfrak{R}_T^{i,s}$, is bounded by,*

$$\frac{|\mathcal{A}_i|^2\log\frac{1}{\gamma}}{\eta}+4H^4m\eta\sum_{t=1}^{T}\sum_{j=1}^{m}\left\|\pi_{t+1}^j-\pi_t^j\right\|_{\infty,1}^2$$
$$-\frac{1}{576\eta|\mathcal{A}_i|}\sum_{t=1}^{T}\left\|\pi_t^i(\cdot\mid s)-\pi_{t+1}^i(\cdot\mid s)\right\|_1^2+36\eta\varepsilon^2T.$$

At this point, Theorem 3.5 follows easily from Theorem 3.7, by summing the swap regret upper bounds over all states and players, and rearranging.

3.4. Proof of main theorem

Having established that the regret of Algorithm 1 may be bounded by the path length of the generated policy sequence (Theorem 3.4), and that the second order path length is well bounded (Theorem 3.5), the proof of our main theorem combines both results using relatively standard arguments. Notably, Theorem 3.4 bounds the regret by the sum of both the first and second-order path lengths while Theorem 3.5 provides only a second order bound. Thus, a $\sqrt{mT}$ factor is ultimately incurred in the final bound.

Proof of Theorem 3.1. By Theorem 3.5,

$$\sum_{t=1}^{T}\sum_{i=1}^{m}\left\|\pi^i_{t+1}-\pi^i_t\right\|_{\infty,1} \le \sqrt{mT\sum_{t,i}\left\|\pi^i_{t+1}-\pi^i_t\right\|^2_{\infty,1}}$$
$$\lesssim m\sqrt{S}A^{3/2}\sqrt{T}+\frac{\varepsilon T}{H^2}.$$

Plugging in Theorem 3.4 gives us;

$$\mathrm{Swap}\mathfrak{R}^i_T \lesssim \frac{SA_i^2}{\eta}+\frac{H^2\sqrt{S}A^{\frac{5}{2}}m}{\eta\gamma}\sqrt{T}+\frac{A_i\varepsilon T}{\eta\gamma} + \eta SA\varepsilon^2T+\eta H^4S^2A_i^4m^2+\mathcal{E}.$$

Finally, setting $\eta=\frac{1}{96H^2m\sqrt{SA}}$ completes the proof. □

4. Regret analysis with unknown dynamics

In this section we analyze the setting of unknown dynamics and bandit feedback - i.e., every player only observes the losses of the state-action pairs in the trajectory visited in each episode. Our analysis requires the following assumption (similar to Wei et al., 2021) of a minimum reachability property of the dynamics:

Assumption 4.1 (β-reachability)**.** There exists a constant $\beta>0$ such that for every set of Markov policies for the m players $\{\pi^i\}_{i=1}^m$ corresponding to a joint policy $\boldsymbol{\pi}$, it holds that $q^{\boldsymbol{\pi}}(s)\ge\beta$ for all $s\in\mathcal{S}$, where $q^{\boldsymbol{\pi}}(s)$ is the probability of reaching state s when each player i plays according to the policy π^i.

We now consider a modification Algorithm 1 which employs a standard blocking mechanism, where inside each block the players do not change their policies, and the mirror descent updates take place only at the end of each block. This allows the players to obtain good approximations of their respective Q functions, as provided by the next lemma.

Lemma 4.2. *With blocks of size $\widetilde{\Theta}(\frac{H^2}{\gamma\beta\varepsilon^2})$, we have that with probability at least $1-2\delta$, the following holds: For every player $i\in[m]$, block index $j=1,\dots,T/B$ and state-action pair (s,a) it holds that*

$$\left|\hat{Q}^i_{t_j}(s,a)-Q^i_{t_j}(s,a)\right|\le\varepsilon.$$

The proof of the lemma, along with the rest of the technical details of this section are deferred to Appendix E. Our guarantee for this setting now follows by applying Theorem 3.1 with the approximation guarantee assured by the above lemma.

Theorem 4.3. *If each player i uses the blocked version of Algorithm 1 with appropriate choice of η,γ,ϵ and blocks of size $\widetilde{\Theta}(\frac{H^2}{\gamma\beta\varepsilon^2})$, then for sufficiently large T, we have that with probability at least $1-\delta$ the swap regret of every player i is bounded as*

$$\mathrm{Swap}\mathfrak{R}^i_T\lesssim H^{22/9}S^{1/3}A^{14/9}m^{2/3}\beta^{-1/9}T^{8/9}.$$

We conclude this section with our convergence guarantee, which follows straightforwardly from Theorem 4.3, in a similar manner as did Corollary 3.2 from Theorem 3.1.

Corollary 4.4 (Convergence to a correlated equilibrium with an unknown model)**.** *Under the conditions and setting of parameters specified in Theorem 4.3, with probability at least $1-\delta$, for every player i it holds that*

$$\mathbb{E}_{t\sim[T]}\left[\max_{\phi_i}\left(V^{i,\pi^i_t}_t(s_1)-V^{i,\phi_i(\pi^i_t)}_t(s_1)\right)\right] = \tilde{O}\left(H^{22/9}S^{1/3}A^{14/9}m^{2/3}\beta^{-1/9}T^{-1/9}\right),$$

where $t\sim[T]$ denotes the uniform distribution over $[T]=\{1,2,\dots,T\}$.

5. Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), the Israel Science Foundation (ISF, grant number 993/17), Len Blavatnik and the Blavatnik Family foundation, the Adelis Research Fund for Artificial Intelligence, the Yandex Initiative for Machine Learning at Tel Aviv University, and from the Tel Aviv University Center for AI and Data Science (TAD).

References

Alacaoglu, A., Viano, L., He, N., and Cevher, V. A natural actor-critic framework for zero-sum markov games. In *International Conference on Machine Learning*, pp. 307–366. PMLR, 2022.

Anagnostides, I., Farina, G., Kroer, C., Celli, A., and Sandholm, T. Faster no-regret learning dynamics for extensive-form correlated and coarse correlated equilibria. *arXiv preprint arXiv:2202.05446*, 2022a.

Anagnostides, I., Farina, G., Kroer, C., Lee, C.-W., Luo, H., and Sandholm, T. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. *arXiv preprint arXiv:2204.11417*, 2022b.

Anantharam, V. and Tsoucas, P. A proof of the markov chain tree theorem. *Statistics & Probability Letters*, 8(2): 189–192, 1989.

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

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

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Busoniu, L., Babuska, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

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

Cen, S., Chi, Y., Du, S. S., and Xiao, L. Faster last-iterate convergence of policy optimization in zero-sum markov games. *arXiv preprint arXiv:2210.01050*, 2022.

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

Chen, X., Deng, X., and Teng, S.-H. Settling the complexity of computing two-player nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

Daskalakis, C., Fishelson, M., and Golowich, N. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34, 2021.

Daskalakis, C., Golowich, N., and Zhang, K. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

Farina, G., Anagnostides, I., Luo, H., Lee, C.-W., Kroer, C., and Sandholm, T. Near-optimal no-regret learning for general convex games. *arXiv preprint arXiv:2206.08742*, 2022a.

Farina, G., Lee, C.-W., Luo, H., and Kroer, C. Kernelized multiplicative weights for 0/1-polyhedral games: Bridging the gap between learning in extensive-form and normal-form games. In *International Conference on Machine Learning*, pp. 6337–6357. PMLR, 2022b.

Foster, D. J., Li, Z., Lykouris, T., Sridharan, K., and Tardos, E. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Foster, D. J., Golowich, N., and Kakade, S. M. Hardness of independent learning and sparse equilibrium computation in markov games. *arXiv preprint arXiv:2303.12287*, 2023.

Giannou, A., Lotidis, K., Mertikopoulos, P., and Vlatakis-Gkaragkounis, E.-V. On the convergence of policy gradient methods to nash equilibria in general stochastic games. In *NeurIPS 2022-36th International Conference on Neural Information Processing Systems*, pp. 1–43, 2022.

Hansen, T. D., Miltersen, P. B., and Zwick, U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

Hazan, E. and Seshadhri, C. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, pp. 393–400, 2009.

Hu, J. and Wellman, M. P. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

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

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33: 16557–16566, 2020.

Lee, C.-W., Luo, H., and Zhang, M. A closer look at small-loss bounds for bandits with graph feedback. In *Conference on Learning Theory*, pp. 2516–2564. PMLR, 2020.

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

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

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Liu, Q., Wang, Y., and Jin, C. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *International Conference on Machine Learning*, pp. 14036–14053. PMLR, 2022.

Mao, W. and Başar, T. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pp. 1–22, 2022.

Mao, W., Yang, L., Zhang, K., and Basar, T. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 15007–15049. PMLR, 2022.

Pattathil, S., Zhang, K., and Ozdaglar, A. Symmetric (optimistic) natural policy gradient for multi-agent learning with parameter convergence. *arXiv preprint arXiv:2210.12812*, 2022.

Piliouras, G., Sim, R., and Skoulakis, S. Optimal no-regret learning in general games: Bounded regret with unbounded step-sizes via clairvoyant MWU. *arXiv preprint arXiv:2111.14737*, 2021.

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

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

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

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002. PMLR, 2020.

Song, Z., Mei, S., and Bai, Y. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2021.

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast convergence of regularized learning in games. *arXiv preprint arXiv:1507.00407*, 2015.

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

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291. PMLR, 2018.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pp. 3674–3682. PMLR, 2020.

Yang, Y. and Ma, C. $o(t^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. *arXiv preprint arXiv:2209.12430*, 2022.

Zhan, W., Lee, J. D., and Yang, Z. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. *arXiv preprint arXiv:2206.01588*, 2022.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent RL in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

Zhang, R., Liu, Q., Wang, H., Xiong, C., Li, N., and Bai, Y. Policy optimization for markov games: Unified framework and faster convergence. *arXiv preprint arXiv:2206.02640*, 2022.

Zhao, Y., Tian, Y., Lee, J., and Du, S. Provably efficient policy optimization for two-player zero-sum markov games. In *International Conference on Artificial Intelligence and Statistics*, pp. 2736–2761. PMLR, 2022.

A. Reduction to low sample complexity

As mentioned previously, there are several works that propose decentralized algorithms with sample-complexity guarantees. Given this, a plausible approach to obtain an online result is to employ an explore-exploit type black-box reduction; learn a good equilibrium in an initial period, then have all players play the joint policy output for the rest of the game. Applying this procedure to V-learning (Jin et al., 2021) for example, and optimizing the length of the initial policy learning period yields an $\widetilde{O}(T^{2/3})$ regret bound for all players. *However*, this approach would necessitate a *centralized* coordination device (such as shared randomness) which our approach does not require. The equilibrium output by the aforementioned algorithms is of course correlated; hence, the exploit phase of the procedure in question is inherently centralized. This highlights an important feature of obtaining correlated equilibrium in a truly online manner; that it circumvents the need for explicit correlation which would be necessary otherwise.

B. Weighted regret analysis (proofs for Section 3.1)

Lemma B.1. *Assuming $H \geq 2$ and $\|\hat{Q}_t - Q_t\|_\infty \leq \varepsilon$, it holds that for all $i \in [m], s \in \mathcal{S}, a \in \mathcal{A}_i$;*

$$\|g_{t+1}^{i,s,a} - g_t^{i,s,a}\|_\infty \leq 2H^2 \sum_{j=1}^m \|\pi_{t+1}^j - \pi_t^j\|_{\infty,1} + 2\varepsilon\pi_{t+1}^i(a \mid s).$$

Proof of Lemma B.1. First, note that,

$$\begin{aligned}
\left\|g_{t+1}^{i,s,a} - g_t^{i,s,a}\right\|_\infty &= \left\|\pi_{t+1}^i(a \mid s)\hat{Q}_{t+1}^i(s,\cdot) - \pi_t^i(a \mid s)\hat{Q}_t^i(s,\cdot)\right\|_\infty \\
&\leq \pi_{t+1}^i(a \mid s)\left\|\hat{Q}_{t+1}^i(s,\cdot) - \hat{Q}_t^i(s,\cdot)\right\|_\infty + \hat{Q}_t^i(s,\cdot)|\pi_{t+1}^i(a \mid s) - \pi_t^i(a \mid s)| \\
&\leq \pi_{t+1}^i(a \mid s)\left\|Q_{t+1}^i(s,\cdot) - Q_t^i(s,\cdot)\right\| + H|\pi_{t+1}^i(a \mid s) - \pi_t^i(a \mid s)| \\
&\quad + \pi_{t+1}^i(a \mid s)\left\|\hat{Q}_{t+1}^i(s,\cdot) - Q_{t+1}^i(s,\cdot)\right\|_\infty + \pi_{t+1}^i(a \mid s)\left\|\hat{Q}_t^i(s,\cdot) - Q_t^i(s,\cdot)\right\|_\infty \\
&\leq \pi_{t+1}^i(a \mid s)\left\|Q_{t+1}^i(s,\cdot) - Q_t^i(s,\cdot)\right\|_\infty + H|\pi_{t+1}^i(a \mid s) - \pi_t^i(a \mid s)| + 2\varepsilon\pi_{t+1}^i(a \mid s)
\end{aligned}$$

To bound the difference in Q-values, observe that for any $a' \in \mathcal{A}_i$, by Lemma F.2 and Lemma F.4;

$$\begin{aligned}
Q_{t+1}^i(s,a') &- Q_t^i(s,a') \\
&= Q^{i,\pi_{t+1}^i}(s,a';M_{t+1}^i) - Q^{i,\pi_t^i}(s,a';M_t^i) \\
&\leq H^2\left\|\pi_{t+1}^i - \pi_t^i\right\|_{\infty,1} + (H^2+1)\left\|P_{M_{t+1}^i} - P_{M_t^i}\right\|_{\infty,1} + (H+1)\left\|\ell_{M_{t+1}^i} - \ell_{M_t^i}\right\|_\infty \\
&\leq H^2\left\|\pi_{t+1}^i - \pi_t^i\right\|_{\infty,1} + (H^2+1)\sum_{j\neq i}\|\pi_{t+1}^j - \pi_t^j\|_{\infty,1} + (H+1)\left\|\ell_{M_{t+1}^i} - \ell_{M_t^i}\right\|_\infty \\
&\leq H^2\left\|\pi_{t+1}^i - \pi_t^i\right\|_{\infty,1} + (H^2+H+2)\sum_{j\neq i}\|\pi_{t+1}^j - \pi_t^j\|_{\infty,1}.
\end{aligned}$$

Thus, by our assumption that $H \geq 2$, we have

$$\begin{aligned}
\|g_{t+1}^{i,s,a} - g_t^{i,s,a}\|_\infty &\leq (H^2+H+2)\sum_{j\neq i}\|\pi_{t+1}^j - \pi_t^j\|_{\infty,1} + (H^2+H)\|\pi_{t+1}^i - \pi_t^i\|_{\infty,1} + 2\varepsilon\pi_{t+1}^i(a \mid s) \\
&\leq 2H^2\sum_{j=1}^m\|\pi_{t+1}^j - \pi_t^j\|_{\infty,1} + 2\varepsilon\pi_{t+1}^i(a \mid s).
\end{aligned}$$

□

Proof of Lemma 3.3. Following similar arguments made in Rakhlin & Sridharan (2013); Syrgkanis et al. (2015), we have

$$\left\langle \ell_t, x_t - x^\star \right\rangle = \left\langle \ell_t, \tilde{x}_t - x^\star \right\rangle + \langle \tilde{\ell}_t, x_t - \tilde{x}_t \rangle + \langle \ell_t - \tilde{\ell}_t, x_t - \tilde{x}_t \rangle,$$

and, from optimality conditions of the first and second optimization steps, respectively;

$$\begin{aligned}
\langle \tilde{\ell}_t, x_t - \tilde{x}_t \rangle &\leq \frac{1}{\eta}\left(D_R(\tilde{x}_t, \tilde{x}_{t-1}) - D_R(\tilde{x}_t, x_t) - D_R(x_t, \tilde{x}_{t-1})\right) \\
\left\langle \ell_t, \tilde{x}_t - x^\star \right\rangle &\leq \frac{1}{\eta}\left(D_R(x^\star, \tilde{x}_{t-1}) - D_R(x^\star, \tilde{x}_t) - D_R(\tilde{x}_t, \tilde{x}_{t-1})\right).
\end{aligned}$$

This implies

$$\begin{aligned}
\left\langle x_t - x^\star, \ell_t \right\rangle &\le \langle \ell_t - \tilde{\ell}_t, x_t - \tilde{x}_t \rangle \\
&\quad + \frac{1}{\eta}\left(D_R(x^\star, \tilde{x}_{t-1}) - D_R(x^\star, \tilde{x}_t) - D_R(\tilde{x}_t, x_t) - D_R(x_t, \tilde{x}_{t-1}) \right) \\
&\le \frac{\eta}{2}\|\ell_t - \tilde{\ell}_t\|_*^2 + \frac{1}{2\eta}\|x_t - \tilde{x}_t\|^2 \\
&\quad + \frac{1}{\eta}\left(D_R(x^\star, \tilde{x}_{t-1}) - D_R(x^\star, \tilde{x}_t) - \frac{1}{2}\|\tilde{x}_t - x_t\|^2 - D_R(x_t, \tilde{x}_{t-1}) \right) \\
&= \frac{\eta}{2}\|\ell_t - \tilde{\ell}_t\|_*^2 \\
&\quad + \frac{1}{\eta}\left(D_R(x^\star, \tilde{x}_{t-1}) - D_R(x^\star, \tilde{x}_t) - D_R(x_t, \tilde{x}_{t-1}) \right)
\end{aligned}$$

Multiplying both sides by q_t and summing over t, we obtain

$$\sum_{t=1}^T q_t \left\langle x_t - x^\star, \ell_t \right\rangle \le \frac{q_1 D_R(x^\star, \tilde{x}_0)}{\eta} + \frac{1}{\eta}\sum_{t=1}^T (q_{t+1} - q_t) D_R(x^\star, \tilde{x}_t) + \frac{\eta}{2}\sum_{t=1}^T q_t \|\ell_t - \tilde{\ell}_t\|_*^2.$$

□

C. Bounding regret by path length (proofs for Section 3.2)

Proof of Theorem 3.4. Let $\phi^i_\star \colon \mathcal{S} \times \mathcal{A}_i \to \mathcal{A}_i$ be any policy swap function, specifying S action swap functions $\phi^{i,s}_\star := \phi^i_\star(\cdot, s) \colon \mathcal{A}_i \to \mathcal{A}_i$. By Lemma F.1 (value difference), we have;

$$\begin{aligned}
&\sum_{t=1}^T V_t^{i,\pi_t^i}(s_1) - V_t^{i,\phi_\star(\pi_t^i)}(s_1) \\
&= \sum_{s\in\mathcal{S}} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle Q_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star^{i,s}(\pi_t^i(\cdot \mid s)) \right\rangle \\
&= \sum_{s\in\mathcal{S}} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle Q_t^i(s,\cdot) - \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star^{i,s}(\pi_t^i(\cdot \mid s)) \right\rangle \\
&\qquad + \sum_{s\in\mathcal{S}} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star^{i,s}(\pi_t^i(\cdot \mid s)) \right\rangle \\
&\le \varepsilon H T + \sum_{s\in\mathcal{S}} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star^{i,s}(\pi_t^i(\cdot \mid s)) \right\rangle. \qquad (6)
\end{aligned}$$

Now, let $x_{\star,\gamma}^{i,s,a} \in \arg\min_{x\in\Delta^\gamma_{\mathcal{A}_i}} \|x - \phi_\star^{i,s}(e_a)\|_1$, and observe that our assumption $\gamma \le 1/2A_i$ implies (see Lemma H.1) $\|x_{\star,\gamma}^{i,s,a} - \phi_\star^{i,s}(e_a)\|_1 \le 2\gamma A_i$. Thus, by Theorem D.1, it follows that for all s;

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star^{i,s}(\pi_t^i(\cdot \mid s)) \right\rangle \\
&\quad \le \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_i} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle g_t^{i,s,a}, x_t^{i,s,a} - \phi_\star^{i,s}(e_a) \right\rangle \\
&\quad = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_i} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle g_t^{i,s,a}, x_t^{i,s,a} - x_{\star,\gamma}^{i,s,a} \right\rangle + \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_i} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle g_t^{i,s,a}, x_{\star,\gamma}^{i,s,a} - \phi_\star^{i,s}(e_a) \right\rangle \\
&\quad \le \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_i} \sum_{t=1}^T q_t^{i,\star}(s) \left\langle g_t^{i,s,a}, x_t^{i,s,a} - x_{\star,\gamma}^{i,s,a} \right\rangle + 2\gamma A_i H(H+\varepsilon)T,
\end{aligned}$$

where in the last transition we use Hölder's inequality and that $\sum_s q_t^{i,\star}(s) = H$ for all t. By Lemma 3.3, this can be further bounded by;

$$
\begin{aligned}
&\le 2\gamma A_i H(H+\varepsilon)T + \frac{SA_i^2 \ln \frac{1}{\gamma}}{\eta} \\
&+ \frac{1}{\eta}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}(q_{t+1}^{i,\star}(s) - q_t^{i,\star}(s))D_R(x_{\star,\gamma}^{i,s,a}, \tilde{x}_t^{i,s,a}) + \frac{\eta}{2}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}\|g_{t+1}^{i,s,a} - g_t^{i,s,a}\|_\infty^2,
\end{aligned}
\tag{7}
$$

where we have used the fact that $\tilde{x}_0^{i,s,a} \in \arg\min_x R(x)$ and $x_{\star,\gamma}^{i,s,a} \in \Delta_{\mathcal{A}_i}^{\gamma}$ implies $D_R(x_{\star,\gamma}^{i,s,a}, \tilde{x}_0^{i,s,a}) \le A_i \log\frac{1}{\gamma}$. Proceeding, to bound the first series term we use the boundedness of the Bregman terms ensured by the truncated simplex (see Lemma H.1), and then apply Lemma F.3 which relates the total variation of the weights to the first-order path length;

$$
\begin{aligned}
\frac{1}{\eta}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}(q_{t+1}^{i,\star}(s) - q_t^{i,\star}(s))D_R(x_{\star,\gamma}^{i,s,a}, \tilde{x}_t^{i,s,a}) &\le \frac{3A_i}{\eta\gamma}\sum_{t=1}^{T}\|q_{t+1}^{i,\star} - q_t^{i,\star}\|_1 \\
&\le \frac{3A_i H^2}{\eta\gamma}\sum_{t=1}^{T}\sum_{j\in[m]}\|\pi_{t+1}^j - \pi_t^j\|_{\infty,1}.
\end{aligned}
$$

Finally, we show in Lemma B.1 that the second-order loss variation term (second series term in Equation (7)) is bounded by the second-order path length up to an additive lower order term. Formally,

$$
\begin{aligned}
\frac{\eta}{2}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}\|g_{t+1}^{i,s,a} - g_t^{i,s,a}\|_\infty^2 &\le \frac{\eta}{2}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}\left(2H^2\sum_{j=1}^{m}\|\pi_{t+1}^j - \pi_t^j\|_{\infty,1} + 2\varepsilon\pi_{t+1}^i(a \mid s)\right)^2 \\
&\le 4\eta m S A_i H^4 \sum_{t=1}^{T}\sum_{j=1}^{m}\|\pi_{t+1}^j - \pi_t^j\|_{\infty,1}^2 + 4\eta\varepsilon^2 ST.
\end{aligned}
$$

Plugging the bounds from the last two displays into Equation (7), the result follows. □

D. Path length analysis (proofs for Section 3.3)

In this section, we provide the full technical details of the analysis outlined in Section 3.3. As mentioned, the arguments mostly follow those of Anagnostides et al. (2022b), who prove a similar result in the setting of general-sum games with access to the exact induced utility functions. The analysis hinges on establishing an RVU property for the swap regret, and then exploiting the fact that it is non-negative. The swap regret of a sequence of iterates $x_1, \ldots, x_T \in \Delta_d$ w.r.t. a sequence of linear losses $g_1, \ldots, g_T \in [0,1]^d$, is defined by;

$$
\mathrm{Swap}\mathfrak{R}_T = \max_{\phi\in\{[d]\to[d]\}}\left\{\sum_{t=1}^{T}\langle g_t, x_t - \phi(x_t)\rangle\right\}.
\tag{8}
$$

Recall we slightly overload notation when applying a swap function $\phi\colon [d] \to [d]$ to $x \in \Delta_d$ as follows;

$$
\phi(x)(a) = \sum_{a':\phi(a')=a} x(a').
$$

That is, the distribution $\phi(x) \in \Delta_d$ is formed by sampling $a \sim x$ and then replacing it with $\phi(a) \in [d]$. We note that, since losses are linear, the external regret is obtained by mapping all actions to the one optimal in hindsight, and taking ϕ to be the identity mapping we see the swap regret is never negative, hence $\mathrm{Swap}\mathfrak{R}_T \ge \max(\mathfrak{R}_T, 0)$. The next theorem is due to Blum & Mansour (2007) and provides for an essential building block in our analysis. We formulate the theorem in the context of our state algorithms, and incorporate the weighting of instantaneous regret, for which the original arguments go through with no modification. The proof below is provided for completeness.

Theorem D.1 (Blum & Mansour (2007)). *Let $i \in [m]$, $s \in \mathcal{S}$. For any weight sequence $\{q_t\}_{t=1}^T, q_t \in [0,1]$, and any action swap function $\phi_\star \colon \mathcal{A}_i \to \mathcal{A}_i$, we have that when player i runs Algorithm 1, the following holds;*

$$\sum_{t=1}^T q_t \left\langle \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - \phi_\star(\pi_t^i(\cdot \mid s)) \right\rangle = \sum_{a\in\mathcal{A}_i} \sum_{t=1}^T q_t \left\langle g_t^{i,s,a}, x_t^{i,s,a} - \phi_\star(e_a) \right\rangle,$$

where $e_a(a') = \mathbb{I}\{a' = a\}$.

Proof of Theorem D.1. The policy played by the state algorithm of player i at s satisfies $\pi_t^i(a' \mid s) = \sum_{a\in\mathcal{A}_i} \pi_t^i(a \mid s) x^{i,s,a}(a')$ (see line 10 of Algorithm 1), thus;

$$\begin{aligned}
\left\langle \pi_t^i(\cdot \mid s), \hat{Q}_t^i(s,\cdot) \right\rangle &= \sum_{a'\in\mathcal{A}_i} \pi_t^i(a' \mid s) \hat{Q}_t^i(s,a') \\
&= \sum_{a'\in\mathcal{A}_i} \sum_{a\in\mathcal{A}_i} \pi_t^i(a \mid s) x^{i,s,a}(a') \hat{Q}_t^i(s,a') \\
&= \sum_{a\in\mathcal{A}_i} \sum_{a'\in\mathcal{A}_i} x_t^{i,s,a}(a') \pi_t^i(a \mid s) \hat{Q}_t^i(s,a') \\
&= \sum_{a\in\mathcal{A}_i} \left\langle x_t^{i,s,a}, g_t^{i,s,a} \right\rangle,
\end{aligned}$$

where we use the definition of $g_t^{i,s,a}$ in Algorithm 1. In addition, we have

$$\begin{aligned}
\left\langle \phi_\star(\pi_t^i(\cdot \mid s)), \hat{Q}_t^i(s,\cdot) \right\rangle = \left\langle \pi_t^i(\cdot \mid s), \hat{Q}_t^i(s,\phi_\star(\cdot)) \right\rangle &= \sum_{a\in\mathcal{A}_i} \pi_t^i(a \mid s) \hat{Q}_t^i(s,\phi_\star(a)) \\
&= \sum_{a\in\mathcal{A}_i} \pi_t^i(a \mid s) \left\langle \phi_\star(e_a), \hat{Q}_t^i(s,\cdot) \right\rangle, \\
&= \sum_{a\in\mathcal{A}_i} \left\langle \phi_\star(e_a), g_t^{i,s,a} \right\rangle.
\end{aligned}$$

Combining the above two displays, the result follows. □

Before proving Theorem 3.7 we give the proof of Lemma 3.6, demonstrating the RVU property of OOMD with local norms. The analysis is similar to arguments made in previous works, such as Wei & Luo (2018, Theorem 7).

Proof of Lemma 3.6. Denote:

$$\tilde{x}_t = \arg\min_{x\in\mathcal{X}} \left\{ \eta \langle g_t, x \rangle + D_R(x, \tilde{x}_{t-1}) \right\},$$

where $\tilde{x}_0 = \arg\min_{x\in\mathcal{X}} R(x)$. We bound the instantaneous regret as follows:

$$\left\langle x_t - x^\star, g_t \right\rangle = \left\langle \tilde{x}_t - x^\star, g_t \right\rangle + \langle x_t - \tilde{x}_t, g_{t-1} \rangle + \langle x_t - \tilde{x}_t, g_t - g_{t-1} \rangle.$$

Using first-order optimality conditions and the three point identity we have:

$$\left\langle \tilde{x}_t - x^\star, g_t \right\rangle \le \frac{1}{\eta} \left(D_R(x^\star, \tilde{x}_{t-1}) - D_R(x^\star, \tilde{x}_t) - D_R(\tilde{x}_t, \tilde{x}_{t-1}) \right),$$

and similarly:

$$\begin{aligned}
\langle x_t - \tilde{x}_t, g_{t-1} \rangle &\le \frac{1}{\eta} \left(D_R(\tilde{x}_t, \tilde{x}_{t-1}) - D_R(\tilde{x}_t, x_t) - D_R(\tilde{x}_t, x_{t-1}) \right) \\
&= \frac{1}{\eta} D_R(\tilde{x}_t, \tilde{x}_{t-1}) - \frac{1}{2\eta} \|\tilde{x}_t - x_t\|_{y_t}^2 - \frac{1}{2\eta} \|\tilde{x}_t - x_{t-1}\|_{z_t}^2,
\end{aligned}$$

where $y_t \in [\tilde{x}_t, x_t]$ and $z_t \in [x_{t-1}, \tilde{x}_t]$. Note that by Lemma D.3 and the condition on η, it holds that $\frac{1}{y_t(a)} \geq \frac{1}{2x_t(a)}$ for all $a \in \mathcal{A}$ and $\frac{1}{z_t(a)} \geq \frac{1}{6x_t(a)}$. Combining this with the above we obtain

$$\langle x_t - \tilde{x}_t, g_{t-1} \rangle \leq \frac{1}{\eta} D_R\left(\tilde{x}_t, \tilde{x}_{t-1}\right) - \frac{1}{8\eta} \left\|\tilde{x}_t - x_t\right\|_{x_t}^2 - \frac{1}{72\eta} \left\|\tilde{x}_t - x_{t-1}\right\|_{x_t}^2 .$$

Also, by Hölder's inequality and Young's inequality:

$$\begin{aligned}\langle x_t - \tilde{x}_t, g_t - g_{t-1} \rangle &\leq \left\|g_t - g_{t-1}\right\|_{*,x_t} \cdot \left\|x_t - \tilde{x}_t\right\|_{x_t} \\ &\leq 4\eta \left\|g_t - g_{t-1}\right\|_{*,x_t}^2 + \frac{1}{16\eta} \left\|x_t - \tilde{x}_t\right\|_{x_t}^2 .\end{aligned}$$

Summing the above over t we obtain:

$$\sum_{t=1}^{T} \left\langle \tilde{x}_t - x^\star, g_t \right\rangle \leq \frac{D_R(x^\star, \tilde{x}_0)}{\eta} + 4\eta \sum_{t=1}^{T} \left\|g_t - g_{t-1}\right\|_{*,x_t}^2 - \frac{1}{72\eta} \sum_{t=1}^{T} \left\|\tilde{x}_t - x_{t-1}\right\|_{x_{t-1}}^2 - \frac{1}{72\eta} \sum_{t=1}^{T} \left\|x_t - \tilde{x}_t\right\|_{x_t}^2 .$$

First note that by definition of $\tilde{x}_0$ and first-order optimality conditions, it holds that

$$D_R(x^\star, \tilde{x}_0) \leq R(x^\star) - R(\tilde{x}_0) \leq d \log \frac{1}{\gamma}.$$

Now note:

$$\left\|x_t - x_{t-1}\right\|_{x_t}^2 \leq 2 \left\|\tilde{x}_t - x_{t-1}\right\|_{x_t}^2 + 2 \left\|x_t - \tilde{x}_t\right\|_{x_t}^2 .$$

Using Lemma D.3 again we have $\left\|\tilde{x}_t - x_{t-1}\right\|_{x_t}^2 \leq 4 \left\|\tilde{x}_t - x_{t-1}\right\|_{x_{t-1}}^2$ which gives

$$\left\|x_t - x_{t-1}\right\|_{x_t}^2 \leq 8 \left\|\tilde{x}_t - x_{t-1}\right\|_{x_{t-1}}^2 + 8 \left\|x_t - \tilde{x}_t\right\|_{x_t}^2 .$$

Combining all of the above we conclude the proof. □

Next, we provide the proof of the principal theorem establishing an RVU property of each state algorithm.

Proof of Theorem 3.7. Using Theorem D.1 and Lemma 3.6, for any swap function ϕ,

$$\begin{aligned}\sum_{t=1}^{T} \left\langle \pi_t^i(\cdot \mid s) - \phi(\pi_t^i(\cdot \mid s)), \hat{Q}_t^i(s, \cdot) \right\rangle &\leq \sum_{a \in \mathcal{A}_i} \sum_{t=1}^{T} \left\langle x_t^{i,s,a} - \phi(\boldsymbol{e}_a), g_t^{i,s,a} \right\rangle \\ &\leq \frac{|\mathcal{A}_i|^2 \log \frac{1}{\gamma}}{\eta} + 4\eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_i} \left\|g_t^{i,s,a} - g_{t-1}^{i,s,a}\right\|_{*,x_t^{i,s,a}}^2 - \frac{1}{576\eta} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_i} \left\|x_t^{i,s,a} - x_{t+1}^{i,s,a}\right\|_{x_t^{i,s,a}}^2 , \\ &\leq \frac{|\mathcal{A}_i|^2 \log \frac{1}{\gamma}}{\eta} + 4\eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_i} \sum_{a' \in \mathcal{A}_i} (x_t^{i,s,a}(a'))^2 \left\|g_t^{i,s,a} - g_{t-1}^{i,s,a}\right\|_\infty^2 - \frac{1}{576\eta} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_i} \left\|x_t^{i,s,a} - x_{t+1}^{i,s,a}\right\|_{x_t^{i,s,a}}^2 , \\ &\leq \frac{|\mathcal{A}_i|^2 \log \frac{1}{\gamma}}{\eta} + 4\eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_i} \left\|g_t^{i,s,a} - g_{t-1}^{i,s,a}\right\|_\infty^2 - \frac{1}{576\eta} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_i} \left\|x_t^{i,s,a} - x_{t+1}^{i,s,a}\right\|_{x_t^{i,s,a}}^2 , \end{aligned} \tag{9}$$

Now, from Corollary D.4,

$$\begin{aligned}&\left|1 - \frac{x_{t+1}^{i,s,a}(a')}{x_t^{i,s,a}(a')}\right| \leq 32\eta \left(\left\|g_{t-1}^{i,s,a}\right\|_\infty + \left\|g_{t-2}^{i,s,a}\right\|_\infty\right) \leq 32\eta H \left(\pi_{t-1}^i(a \mid s) + \pi_{t-2}^i(a \mid s)\right) \\ \Longrightarrow &\sum_{a \in \mathcal{A}_i} \max_{a'} \left|1 - \frac{x_{t+1}^{i,s,a}(a')}{x_t^{i,s,a}(a')}\right| \leq \frac{1}{2}.\end{aligned}$$

Therefore, using Lemma D.5,

$$\left\|\pi_t^i(\cdot \mid s) - \pi_{t+1}^i(\cdot \mid s)\right\|_1^2 \leq |\mathcal{A}_i| \sum_{a \in \mathcal{A}_i} \left\|x_t^{i,s,a} - x_{t+1}^{i,s,a}\right\|_{x_t^{i,s,a}}^2 . \tag{10}$$

Combining Equations (9) and (10) and Lemma B.1 completes the proof. □

We conclude with the proof of Theorem 3.5, which follows easily from Theorem 3.7.

Proof of Theorem 3.5. From Theorem 3.7 and the fact that swap regret is non-negative, we have

$$\begin{aligned} 0 &\le \sum_{i=1}^{m}\sum_{s\in\mathcal{S}} \mathrm{Swap}\mathfrak{R}_T^{i,s} \\ &\le \frac{mSA^2\log\frac{1}{\gamma}}{\eta} + 36\eta\varepsilon^2 mST + \sum_{t=1}^{T}\sum_{i=1}^{m}\left(4\eta H^4 m^2 S - \frac{1}{576\eta A}\right)\left\|\pi_{t+1}^i - \pi_t^i\right\|_{\infty,1}^2, \end{aligned}$$

where we also used the fact that $\sum_{s\in\mathcal{S}}\left\|\pi_t^i(\cdot\mid s) - \pi_{t+1}^i(\cdot\mid s)\right\|_1^2 \ge \left\|\pi_{t+1}^i - \pi_t^i\right\|_{\infty,1}^2$. Setting $\eta = \frac{1}{96H^2m\sqrt{SA}}$ and rearranging the terms completes the proof. □

D.1. Log barrier lemmas for Theorem 3.7

The following Lemmas D.2 and D.3 follows by the proof technique of Jin & Luo (2020, Lemma 12); Lee et al. (2020, Lemma 9).

Lemma D.2. *Let $F\colon \Delta_d \to \mathbb{R}$ defined as $F(x) = \eta\langle x,\ell\rangle + D_R(x,x')$ for some $x'\in\Delta_d$, where R is the log-barrier regularization. Suppose $\|\ell\|_\infty \le H$, and that $\eta \le \frac{1}{8H}$. Then, for any $x''\in\Delta_d$ such that $\|x''-x'\|_{x'} = 8\eta\,\|\ell\|_\infty$, we have*

$$F(x'') \ge F(x').$$

Proof. By second-order Taylor expansion of F around x', there exist ξ is on the line segment between x'' and x' such that,

$$\begin{aligned} F(x'') &= F(x') + \nabla F(x')^\top(x''-x') + \frac{1}{2}(x''-x')^\top\nabla^2F(\xi)(x''-x'), \\ &= F(x') + \eta\langle \ell, x''-x'\rangle + \frac{1}{2}(x''-x')^\top\nabla^2F(\xi)(x''-x') \\ &\ge F(x') - \eta\|\ell\|_{*,x'}\|x''-x'\|_{x'} + \frac{1}{2}(x''-x')^\top\nabla^2F(\xi)(x''-x') \\ &= F(x') - 8\eta^2\,\|\ell\|_\infty\,\|\ell\|_{*,x'} + \frac{1}{2}\|x''-x'\|_\xi^2 \\ &\ge F(x') - 8\eta^2\,\|\ell\|_\infty^2 + \frac{1}{2}\|x''-x'\|_\xi^2 \end{aligned} \tag{11}$$

The second equality is since $\nabla F(x') = \eta\ell$, the first inequality is Hölder inequality, the last equality is since $\|x''-x'\|_{x'} = 8\eta\,\|\ell\|_\infty$ and $\nabla^2 F = \nabla^2 R$, and the last inequality is since $\|\ell\|_{*,x'} \le \|\ell\|_\infty$. Now, for all a, since $\eta \le \frac{1}{8H}$:

$$\frac{|\xi(a)-x'(a)|}{x'(a)} \le \frac{|x''(a)-x'(a)|}{x'(a)} \le \|x''-x'\|_{x'} \le 8\eta H \le 1.$$

In particular, $\xi(a) \le 2x'(a)$, which implies that

$$\begin{aligned} \|x''-x'\|_\xi^2 &= \sum_a\left(\frac{x''(a)-x'(a)}{\xi(a)}\right)^2 \\ &\ge \frac{1}{4}\sum_a\left(\frac{x''(a)-x'(a)}{x'(a)}\right)^2 \\ &= \frac{1}{4}\|x''-x'\|_{x'}^2 = 16\eta^2\,\|\ell\|_\infty^2. \end{aligned}$$

Plugging back in Equation (11),

$$F(x'') \ge F(x') - 8\eta^2\,\|\ell\|_\infty^2 + 16\eta^2\,\|\ell\|_\infty^2 \ge F(x').$$

□

Lemma D.3. *Let $x' \in \Delta_d$, $F\colon \Delta_d \to \mathbb{R}$ be defined as in Lemma D.2 and let $x^+ = \arg\min_x F(x)$. If $\eta \le \frac{1}{8H}$ then for any $a \in [d]$ it holds that*

$$(1 - 8\eta \|\ell\|_\infty)\, x'(a) \le x^+(a) \le (1 + 8\eta \|\ell\|_\infty)\, x'(a).$$

Proof. We first show that $\|x^+ - x'\|_{x'} \le 8\eta \|\ell\|_\infty$. Assume otherwise: $\|x^+ - x'\|_{x'} > 8\eta \|\ell\|_\infty$. Then for some $\lambda \in (0,1)$ and $x'' = \lambda x^+ + (1-\lambda)x'$ we have,

$$\|x'' - x'\|_{x'} = 8\eta \|\ell\|_\infty .$$

From Lemma D.2, $F(x') \le F(x'')$. Since F is strongly convex and x^+ is the (unique) minimizer of F,

$$F(x') \le F(x'') \le \lambda F(x^+) + (1-\lambda)F(x') < \lambda F(x') + (1-\lambda)F(x') = F(x'),$$

which is a contradiction. Hence, $\|x^+ - x'\|_{x'} \le 8\eta \|\ell\|_\infty$ and for any a,

$$\left(\frac{x^+(a) - x'(a)}{x'(a)}\right)^2 \le \sum_{\tilde a} \left(\frac{x^+(\tilde a) - x'(\tilde a)}{x'(\tilde a)}\right)^2 \le (8\eta \|\ell\|_\infty)^2.$$

By rearranging the inequality above we obtain the statement of the lemma. $\square$

Corollary D.4. *Assume that $\{x_t\}_{t=1}^T$ are iterates of OOMD with log-barrier and $\eta \le \frac{1}{64H}$, then for any t,*

$$1 - 32\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty) \le \frac{x_{t+1}(a)}{x_t(a)} \le 1 + 32\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty).$$

Proof. By Lemma D.3 we have,

$$\begin{aligned}
1 - 8\eta \|\ell_{t-1}\|_\infty &\le \frac{x_t(a)}{\tilde x_{t-1}(a)} \le 1 + 8\eta \|\ell_{t-1}\|_\infty \\
1 - 8\eta \|\ell_{t-1}\|_\infty &\le \frac{\tilde x_{t-1}(a)}{\tilde x_{t-2}(a)} \le 1 + 8\eta \|\ell_{t-1}\|_\infty \\
(1 + 8\eta \|\ell_{t-2}\|_\infty)^{-1} &\le \frac{\tilde x_{t-2}(a)}{x_{t-1}(a)} \le (1 - 8\eta \|\ell_{t-2}\|_\infty)^{-1}.
\end{aligned}$$

Hence,

$$\frac{x_t(a)}{x_{t-1}(a)} = \frac{x_t(a)}{\tilde x_{t-1}(a)} \cdot \frac{\tilde x_{t-1}(a)}{\tilde x_{t-2}(a)} \cdot \frac{\tilde x_{t-2}(a)}{x_{t-1}(a)} \le \frac{(1 + 8\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty))^2}{1 - 8\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty)} \le 1 + 32\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty)$$

where the last is since $\frac{(1+x)^2}{1-x} \le 1 + 4x$ for $x \in (0, 1/5]$. In a similar way,

$$\frac{x_t(a)}{x_{t-1}(a)} = \frac{x_t(a)}{\tilde x_{t-1}(a)} \cdot \frac{\tilde x_{t-1}(a)}{\tilde x_{t-2}(a)} \cdot \frac{\tilde x_{t-2}(a)}{x_{t-1}(a)} \ge \frac{(1 - 8\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty))^2}{1 + 8\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty)} \ge 1 - 24\eta(\|\ell_{t-1}\|_\infty + \|\ell_{t-2}\|_\infty)$$

where the last is since $\frac{(1-x)^2}{1+x} \le 1 - 3x$ for all $x > 0$. $\square$

The following lemma is a slight generalization of (Anagnostides et al., 2022b, Lemma 4.2), which applies for any sequence of sufficiently stable base iterates (not necessarily OFTRL generated).

Lemma D.5 (Anagnostides et al. (2022b))**.** *Fix some vectors $x_{t,a} \in \Delta_d$ for $t \in [T]$ and $a \in [d]$. Let $M_t \in \mathbb{R}^{d\times d}$ a matrix who's rows are $x_{t,a}$ and let $x_t \in \Delta_d$ be vectors such that $M_t^\top x_t = x_t$. If $\sum_{a\in[d]} \max_{a'} \left|\frac{x_{t-1,a}(a') - x_{t,a}(a')}{x_{t-1,a}(a')}\right| \le 1/2$, then*

$$\|x_t - x_{t-1}\|_1^2 \le 64A \sum_{a\in[d]} \left\|x_{t,a} - x_{t-1,a}\right\|_{x_{t-1,a}}^2$$

where the local norms here are those induced by the log-barrier regularizer (Equation (4)).

Proof. Let $\mathbb{T}_a$ be the set of all directed trees over $[d]$ (i.e., each directed tree has no directed cycles, each node $a' \neq a$ has exactly 1 outgoing edge and a has no outgoing edges). By the Markov chain tree theorem (Anantharam & Tsoucas, 1989) $x_t(a) = \frac{w_t(a)}{W_t}$ where

$$w_t(a) = \sum_{\mathcal{T}\in\mathbb{T}_a} \prod_{(u,v)\in E(\mathcal{T})} x_{t,u}(v) \qquad \text{and,} \qquad W_t = \sum_a w_t(a).$$

Let $\mu_{t,a} := \max_{a'} \left|1 - \frac{x_{t,a}(a')}{x_{t-1,a}(a')}\right| \leq \mu$. In particular, $1 - \mu_{t,a} \leq \frac{x_{t,a}(a')}{x_{t-1,a}(a')} \leq 1 + \mu_{t,a}$ which implies

$$\begin{aligned}
w_t(a) &\leq \sum_{\mathcal{T}\in\mathbb{T}_a} \prod_{(u,v)\in E(\mathcal{T})} x_{t,u}(v) \\
&\leq \sum_{\mathcal{T}\in\mathbb{T}_a} \prod_{(u,v)\in E(\mathcal{T})} (1+\mu_{t,u}) x_{t-1,u}(v) \\
&\leq \prod_{a'\in[d]} (1+\mu_{t,a'}) \sum_{\mathcal{T}\in\mathbb{T}_a} \prod_{(u,v)\in E(\mathcal{T})} x_{t-1,u}(v) \\
&= \prod_{a'\in[d]} (1+\mu_{t,a'}) w_{t-1}(a) \\
&\leq \exp\left(\sum_{a'\in[d]} \mu_{t,a'}\right) w_{t-1}(a).
\end{aligned}$$

This also implies that $W_t \leq \exp\left(\sum_{a'\in[d]} \mu_{t,a'}\right) W_{t-1}$. In a similar way,

$$\begin{aligned}
w_t(a) &\geq \prod_{a'\in[d]} (1-\mu_{t,a'}) w_{t-1}(a) \\
&\geq \prod_{a'\in[d]} \exp\left(-2\sum_{a'\in[d]} \mu_{t,a'}\right) w_{t-1}(a)
\end{aligned}$$

where the last uses the fact that $1 - x \geq e^{-2x}$ for $x \in [0, 1/2]$ and that $\sum_{a'\in[d]} \mu_{t,a'} \leq 1/2$. Similarly, $W_t \geq \exp\left(-2\sum_{a'\in[d]} \mu_{t,a'}\right) W_{t-1}$. Combining the inequities above we get,

$$\begin{aligned}
x_t(a) - x_{t-1}(a) &= \frac{w_t(a)}{W_t} - x_{t-1}(a) \\
&\leq \frac{\exp\left(\sum_{a'\in[d]} \mu_{t,a'}\right) w_{t-1}(a)}{\exp\left(-2\sum_{a'\in[d]} \mu_{t,a'}\right) W_{t-1}} - x_{t-1}(a) \\
&\leq x_{t-1}(a)\left(\exp\left(3\sum_{a'\in[d]} \mu_{t,a'}\right) - 1\right) \\
&\leq 8x_{t-1}(a) \sum_{a'\in[d]} \mu_{t,a'},
\end{aligned}$$

where the last holds since $e^x - 1 \leq \frac{8}{3}x$ for $x \in [0, 2/3]$ and $\sum_{a'\in[d]} \mu_{t,a'} \leq 1/2$. In similar way,

$$x_{t-1}(a) - x_t(a) \leq x_{t-1}(a)\left(1 - \exp\left(-3\sum_{a'\in[d]} \mu_{t,a'}\right)\right) \leq 3x_{t-1}(a) \sum_{a'\in[d]} \mu_{t,a'}.$$

From the last to we have $|x_{t-1}(a) - x_t(a)| \leq 8 x_{t-1}(a) \sum_{a' \in [d]} \mu_{t,a'}$ and so,

$$\begin{aligned}
\|x_{t-1} - x_t\|_1^2 &\leq 64 \left(\sum_{a \in [d]} \mu_{t,a} \right)^2 \\
&\leq 64A \sum_{a \in [d]} (\mu_{t,a})^2 \\
&\leq 64A \sum_{a \in [d]} \sum_{a' \in [d]} \left(\frac{x_{t-1,a}(a') - x_{t,a}(a')}{x_{t-1,a}(a')} \right)^2 \\
&= 64A \sum_{a \in [d]} \left\| x_{t,a} - x_{t-1,a} \right\|_{x_{t-1,a}}^2 .
\end{aligned}$$

□

E. Unknown dynamics regret analysis (proof for Section 4)

In this section we provide the full technical details involved in the proof of Theorem 4.3. As mentioned, we modify Algorithm 1 to a blocking type algorithm as follows: After each policy update, the players use the same policy for B episodes and use the trajectories observed in those B episodes to estimate their Q-function. We later set B such that with high probability, every player samples the loss of every state-action pair (s, a) at least $\approx 1/\varepsilon^2$ times, which is enough to guarantee that with high probability $\hat{Q}_t(s, a)$ is an ε-approximation of $Q_t(s, a)$.

More formally, after a policy update at the end of episode t_j for $j = 1, 2, \ldots, T/B$, each player uses the policy $\pi^i_{t_j}$ for B consecutive episodes to obtain the trajectories

$$\left\{ \left(s_1^\tau, a_1^\tau, \ldots, s_{H-1}^\tau a_{H-1}^\tau, s_H^\tau \right) \right\}_{\tau = t_j+1, t_j+2, \ldots, t_j+B},$$

and then constructs the following estimator for the Q-function at each state-action pair (s, a) where s is in layer h:

$$\hat{Q}^i_{t_j}(s, a) = \frac{1}{B} \sum_{\tau = t_j+1}^{t_j+B} \mathbb{I}\left[s_h^\tau = s, a_h^\tau = a \right] \sum_{h'=h}^{H-1} \ell^i_\tau(s_{h'}^\tau, a_{h'}^\tau).$$

Note that for each state-action pair (s, a) this is an unbiased estimator for $Q^i_{t_j}(s, a)$. We refer to the time periods in between each pair of episodes t_j and t_{j+1} as "blocks". As a first step, we establish that w.h.p. these estimates are indeed good approximations of the true Q-functions.

E.1. The good event

Lemma E.1. *Denote by $n^i_j(s, a)$ the number of times player i reached the state-action pair (s, a) in episodes $t_j+1, \ldots, t_j+B$. Then with probability at least $1 - \delta$, for every player i, block index $j = 1, \ldots, T/B$ and state-action pair (s, a) it holds that*

$$n^i_j(s, a) \geq \frac{\gamma \beta B}{2} - \log \frac{mSAT}{\delta}.$$

Proof. First note that for all $s \in \mathcal{S}$ and $a \in \mathcal{A}_i$,

$$\pi^i_{t_j}(a \mid s) \geq \gamma.$$

This follows immediately from the fact that $\pi^i_{t_j}(\cdot \mid s)$ is a stationary distribution corresponding to the base iterates $\left\{ x^{i,s,a}_{t_j}(\cdot) \right\}$ in which every action is taken with probability of at least γ. Therefore, if we denote by $X_\tau(s)$ the indicator variable of reaching state s at episode τ, we have

$$\begin{aligned}\mathbb{E}_{t_j}\left[n^i_j(s,a)\right] &= \sum_{\tau=t_j+1}^{t_j+B} \Pr\left[X_\tau(s)=1\right]\pi^i_{t_j}(a \mid s)\\ &\geq \gamma \sum_{\tau=t_j+1}^{t_j+B} \Pr\left[X_\tau(s)=1\right]\\ &\geq \gamma\beta B,\end{aligned}$$

where the last inequality follows from Assumption 4.1. By invoking Lemma F.4 in Dann et al. (2017) we obtain that with probability at least $1-\frac{\delta}{mSAT}$ it holds that

$$n^i_j(s,a) \geq \frac{\gamma\beta B}{2} - \log\frac{mSAT}{\delta},$$

and we conclude the proof with a union bound over i, s, a and j. □

Proof of Lemma 4.2. We set $B = \frac{2H^2\ln\frac{mSAT}{\delta}}{\gamma\beta\varepsilon^2}$. Denote by G_1 the event that for all i,j,s,a it holds that $n^i_j(s,a)\geq N$, where $N=\frac{H^2\ln\frac{mSAT}{\delta}}{2\varepsilon^2}$. By Lemma E.1 and the choice of B, the event G_1 holds with probability at least $1-\delta$. Therefore for the rest of the proof we assume G_1 holds. Note that conditioned on G_1, the estimator of the Q-function $\hat{Q}^i_{t_j}(s,a)$ is an average of at least N i.i.d random variables, each bounded in $[0,H]$ and each with expected value $Q^i_{t_j}(s,a)$. Therefore, using Hoeffding's inequality and our setting of N, with probability at least $1-\frac{\delta}{mSAT}$ it holds that

$$\left|\hat{Q}^i_{t_j}(s,a) - Q^i_{t_j}(s,a)\right| \leq \varepsilon.$$

Taking a union bound over i,s,a,j we conclude the proof. □

E.2. Regret bound

To prove our regret bound, we begin with a standard lemma which relates the swap regret of the blocked version of Algorithm 1 where each player plays the same policy for B episodes after each policy update, to the swap regret guarantee of Algorithm 1, assuming at each episode the players observe ε-approximations of their Q-functions.

Lemma E.2. *If the non-blocked version of Algorithm 1 has a swap regret guarantee of* $\mathrm{Swap}\mathfrak{R}^i_K \leq R(m,H,S,A,K)$ *when run for K episodes, then the swap regret of the blocked version of Algorithm 1 can be bounded by*

$$B\cdot R\left(m,H,S,A,\frac{T}{B}\right).$$

Proof. Let $\{\pi^i_t\}$ denote the policies played by player i when executing the blocked with estimations version of the algorithm, and let $r_t := V^{i,\pi^i_t}_t(s_1) - V^{i,\phi(\pi^i_\star)}_t(s_1)$ denote the per round instantaneous regret w.r.t. swap function ϕ. Then

$$\mathrm{Swap}\mathfrak{R}^i_T = \sum_{t=1}^{T} r_t = \sum_{j=1}^{T/B}\sum_{\tau=t_j+1}^{t_j+B} r_\tau = \sum_{j=1}^{T/B}\sum_{\tau=t_j+1}^{t_j+B} r_{t_j} = B\sum_{j=1}^{T/B} r_{t_j} \leq B\cdot R\left(m,H,S,A,\frac{T}{B}\right),$$

where the last inequality is due to the fact that the sequence of policies $\{\pi^i_{t_j}\}_{j\in[B]}$ corresponds to running Algorithm 1 over T/B episodes. □

Given Lemma E.2 and the individual swap regret bound in Theorem 3.1, we are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. We make the following parameter choices; block size $B=\frac{2H^2\ln\frac{mSAT}{\delta}}{\gamma\beta\varepsilon^2}$, step size $\eta=\frac{1}{96H^2m\sqrt{SA}}$, $\varepsilon = 6H^2\sqrt{m}S^{\frac14}A^{\frac34}\left(\ln\frac{mSAT}{\delta}\ln\frac{1}{\gamma}\right)^{\frac14}(\beta\gamma T)^{-\frac14}$ and $\gamma = H^{\frac49}S^{\frac13}A^{\frac59}m^{\frac23}\beta^{-\frac19}\left(\ln\frac{mSAT}{\delta}\ln T\right)^{\frac19}T^{-\frac19}$. Further, we assume $T\geq 512H^4S^3A^{14}m^6\beta^{-1}\ln\frac{mSAT}{\delta}\ln T$, so that $\gamma\leq 1/2A$.

Now, using Theorem 3.1, Lemma 4.2, Lemma E.2 and the choice of B we obtain that with probability at least $1-2\delta$, player i's swap regret can be bounded as follows:

$$\begin{aligned}\mathrm{Swap}\mathfrak{R}_T^i \leq\ & \frac{2\cdot 10^4 H^5 S A^3 m^2 \sqrt{\ln\frac{1}{\gamma}\ln\frac{mSAT}{\delta}}}{\varepsilon\gamma^{\frac{3}{2}}\sqrt{\beta}}\sqrt{T} + \frac{600 m H\sqrt{S}A^{\frac{3}{2}}}{\gamma}\varepsilon T + 2\gamma A H^2 T \\ & + \frac{300 m H^4 S^{\frac{3}{2}} A^{\frac{7}{2}}\ln\frac{1}{\gamma}\ln\frac{mSAT}{\delta}}{\gamma\beta\varepsilon^2}.\end{aligned}$$

Our choice of ε balances the first two terms, leading to;

$$\begin{aligned}\mathrm{Swap}\mathfrak{R}_T^i \leq\ & 7000 H^3 S^{\frac{3}{4}} A^{\frac{9}{4}} m^{\frac{3}{2}}\left(\ln\frac{mSAT}{\delta}\ln T\right)^{\frac{1}{4}}\beta^{-\frac{1}{4}}\gamma^{-\frac{5}{4}}T^{\frac{3}{4}} + 2\gamma A H^2 T \\ & + 10 S A^2\sqrt{\ln\frac{mSAT}{\delta}\ln T}\,\beta^{-\frac{1}{2}}\gamma^{-\frac{1}{2}}\sqrt{T},\end{aligned}$$

where we used the fact that $\gamma \geq \frac{1}{T}$. We now use our choice of γ to obtain the desired regret bound. □

F. Elementary MDP Lemmas

In this section, we prove some basic lemmas relating variations in state visitation measures, losses and dynamics to the movement (changes in policies) of players. Recall we let ℓ_t^i, P_t^i, M_t^i denote respectively the loss, dynamics, and MDP tuple $M_t^i := (H, \mathcal{S}, \mathcal{A}_i, P_t^i, \ell_t^i)$ of the single agent induced MDP of player i at round t. Further, for any (single agent) transition function P, policy $\pi \in \mathcal{S} \to \Delta_{\mathcal{A}_i}$ and state $s \in \mathcal{S}_h$, we denote

$$\begin{aligned} q_P^\pi(s,a) &:= \Pr(s_h = s, a_h = a \mid P, \pi, s_1), \\ q_P^\pi(s) &:= \Pr(s_h = s \mid P, \pi, s_1). \end{aligned}$$

When P is clear from context we may omit the subscript and write q^π for q_P^π. Further, for any single agent MDP $M = (H, \mathcal{S}, \mathcal{A}_i, P, \ell)$, we write $V(\cdot; M), Q(\cdot,\cdot; M)$ to denote respectively the state and state-action value functions of M. We may omit M and write $V(\cdot), Q(\cdot,\cdot)$ when M is clear from context. For $s \in \mathcal{S}$, we let $h(s) := h$ s.t. $s \in \mathcal{S}_h$. With this notation in place, we have for a policy $\pi \in \mathcal{S} \to \Delta_{\mathcal{A}_i}$:

$$\begin{aligned} V^\pi(s; M) &:= \mathbb{E}\Big[\sum_{h=h(s)}^{H} \ell(s_h, a_h) \mid P, \pi, s_{h(s)} = s\Big], \\ Q^\pi(s, a; M) &:= \mathbb{E}\Big[\sum_{h=h(s)}^{H} \ell(s_h, a_h) \mid P, \pi, s_{h(s)} = s, a_{h(s)} = a\Big]. \end{aligned}$$

We begin with value difference lemmas which are typical in single agent MDP analyses. The proofs below are provided for completeness; see also Shani et al. (2020); Cai et al. (2020) for similar arguments.

Lemma F.1 (value-difference)**.** *The following holds.*

1. *For any MDP $M = (\mathcal{S}, \mathcal{A}_i, H, P, \ell)$, and pair of policies $\pi, \tilde{\pi} \in \mathcal{S} \to \Delta_{\mathcal{A}_i}$, we have*

$$\begin{aligned} V^\pi(s_1) - V^{\tilde{\pi}}(s_1) &= \mathbb{E}\left[\sum_{h=1}^{H} \langle Q^\pi(s_h, \cdot), \pi(\cdot \mid s_h) - \tilde{\pi}(\cdot \mid s_h)\rangle \mid \tilde{\pi}\right] \\ &= \sum_{s \in \mathcal{S}} q^{\tilde{\pi}}(s)\, \langle Q^\pi(s, \cdot), \pi(\cdot \mid s) - \tilde{\pi}(\cdot \mid s)\rangle \\ &\leq H^2 \left\|\pi - \tilde{\pi}\right\|_{\infty,1}. \end{aligned}$$

2. *For any two MDPs* $M = (H, S, \mathcal{A}_i, P, \ell), \tilde{M} = (H, S, \mathcal{A}_i, \tilde{P}, \tilde{\ell})$, $V^\pi(\cdot) \coloneqq V^\pi(\cdot; M), \tilde{V}^\pi(\cdot) \coloneqq V^\pi(\cdot; \tilde{M})$, *and policy* $\pi \in \mathcal{S} \to \Delta_{\mathcal{A}_i}$, *we have*

$$
\begin{aligned}
V^\pi(s_1) - \tilde{V}^\pi(s_1) &= \mathbb{E}_{\tilde{P},\pi}\left[\sum_{h=1}^{H} \ell(s_h, a_h) - \tilde{\ell}(s_h, a_h) + \sum_{s' \in \mathcal{S}_{h+1}} (P(s'|s_h, a_h) - \tilde{P}(s'|s_h, a_h))V^\pi(s')\right] \\
&\le H \left\|\ell - \tilde{\ell}\right\|_\infty + H^2 \left\|P - \tilde{P}\right\|_{\infty,1}
\end{aligned}
$$

Proof. For (1), observe that for $s \in \mathcal{S}_l$;

$$
\begin{aligned}
V^\pi(s) - V^{\tilde{\pi}}(s) &= \langle Q^\pi(s, \cdot), \pi(\cdot|s) - \tilde{\pi}(\cdot|s)\rangle + \left\langle Q^\pi(s, \cdot) - Q^{\tilde{\pi}}(s, \cdot), \tilde{\pi}(\cdot|s)\right\rangle \\
&= \langle Q^\pi(s, \cdot), \pi(\cdot|s) - \tilde{\pi}(\cdot|s)\rangle + \mathbb{E}_{a' \sim \tilde{\pi}(\cdot|s)}\left[\mathbb{E}_{s' \sim P(\cdot|s,a')}\left[V^\pi(s') - V^{\tilde{\pi}}(s')\right]\right] \\
&= \langle Q^\pi(s, \cdot), \pi(\cdot|s) - \tilde{\pi}(\cdot|s)\rangle + \mathbb{E}\left[V^\pi(s_{l+1}) - V^{\tilde{\pi}}(s_{l+1}) \mid \tilde{\pi}, s_l = s\right].
\end{aligned}
$$

Applying the relation recursively, we obtain for $l = 1$;

$$
\begin{aligned}
V^\pi(s_1) - V^{\tilde{\pi}}(s_1) &= \mathbb{E}\left[\sum_{h=1}^{H} \langle Q^\pi(s_h, \cdot), \pi(\cdot|s_h) - \tilde{\pi}(\cdot|s_h)\rangle \mid \tilde{\pi}, s_1\right] \\
&= \sum_{h=l}^{H} \sum_{s \in \mathcal{S}_h} \Pr(s_h = s \mid, s_1, \tilde{\pi}) \langle Q^\pi(s, \cdot), \pi(\cdot|s) - \tilde{\pi}(\cdot|s)\rangle \\
&= \sum_{s \in \mathcal{S}} \Pr(s_h = s \mid, s_1, \tilde{\pi}) \langle Q^\pi(s, \cdot), \pi(\cdot|s) - \tilde{\pi}(\cdot|s)\rangle \\
&= \sum_{s \in \mathcal{S}} q^{\tilde{\pi}}(s) \langle Q^\pi(s, \cdot), \pi(\cdot|s) - \tilde{\pi}(\cdot|s)\rangle \\
&\le H \sum_{s \in \mathcal{S}} q^{\tilde{\pi}}(s) \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1 \\
&\le H \sum_{s \in \mathcal{S}} q^{\tilde{\pi}}(s) \|\pi - \tilde{\pi}\|_{\infty,1} \le H^2 \|\pi - \tilde{\pi}\|_{\infty,1},
\end{aligned}
$$

which completes the proof of (1). For (2), let $s \in \mathcal{S}_l$ and observe;

$$
V^\pi(s) - \tilde{V}^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[\ell(s, a) - \tilde{\ell}(s, a) + \sum_{s' \in \mathcal{S}_{l+1}} P(s'|s, a)V^\pi(s') - \tilde{P}(s'|s, a)\tilde{V}^\pi(s')\right]
$$

Further, we have

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}_{l+1}} P(s'|s, a)V^\pi(s') - \tilde{P}(s'|s, a)\tilde{V}^\pi(s') \\
&\qquad = \sum_{s' \in \mathcal{S}_{l+1}} \left(P(s'|s, a) - \tilde{P}(s'|s, a)\right) V^\pi(s') + \tilde{P}(s'|s, a)\left(V^\pi(s') - \tilde{V}^\pi(s')\right),
\end{aligned}
$$

and combining this with the previous equation we get

$$
\begin{aligned}
V^\pi(s) - \tilde{V}^\pi(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)}\left[\ell(s, a) - \tilde{\ell}(s, a) + \sum_{s' \in \mathcal{S}_{l+1}} \left(P(s'|s, a) - \tilde{P}(s'|s, a)\right) V^\pi(s')\right] \\
&\quad + \mathbb{E}\left[V^\pi(s_{l+1}) - \tilde{V}^\pi(s_{l+1}) \mid \tilde{P}, \pi, s_l = s\right].
\end{aligned}
$$

Applying this recursivly with $l = 1$, the first part of (2) follows. For the second part;

$$\begin{aligned}
&\mathbb{E}_{\tilde{P},\pi}\left[\sum_{h=1}^{H} \ell(s_h, a_h) - \tilde{\ell}(s_h, a_h) + \sum_{s' \in \mathcal{S}_{h+1}} (P(s'|s_h, a_h) - \tilde{P}(s'|s_h, a_h))V^{\pi}(s')\right] \\
&\leq \mathbb{E}_{\tilde{P},\pi}\left[\sum_{h=1}^{H} \left\|\ell - \tilde{\ell}\right\|_{\infty} + H\left\|P(\cdot|s_h, a_h) - \tilde{P}(\cdot|s_h, a_h)\right\|_1\right] \\
&\leq \mathbb{E}_{\tilde{P},\pi}\left[H\left\|\ell - \tilde{\ell}\right\|_{\infty} + H\sum_{h=1}^{H}\left\|P - \tilde{P}\right\|_{\infty,1}\right] \\
&= H\left\|\ell - \tilde{\ell}\right\|_{\infty} + H^2\left\|P - \tilde{P}\right\|_{\infty,1}
\end{aligned}$$

□

Lemma F.2 (action-value-difference)**.** *Let $M = (S, \mathcal{A}_i, H, P, \ell)$, $\tilde{M} = (S, \mathcal{A}_i, H, \tilde{P}, \tilde{\ell})$ be two MDPs, and $\pi, \tilde{\pi} \in \mathcal{S} \to \Delta_{\mathcal{A}_i}$ be a pair of policies. Then for all $s \in \mathcal{S}, a \in \mathcal{A}_i$, we have;*

$$Q^{\pi}(s, a; M) - Q^{\tilde{\pi}}(s, a; \tilde{M}) \leq H^2\left\|\pi - \tilde{\pi}\right\|_{\infty,1} + (H^2 + 1)\left\|P - \tilde{P}\right\|_{\infty,1} + (H + 1)\left\|\ell - \tilde{\ell}\right\|_{\infty}.$$

Proof. By Lemma F.1, we have

$$\begin{aligned}
V^{\pi}(s) - \tilde{V}^{\tilde{\pi}}(s) &= V^{\pi}(s) - V^{\tilde{\pi}}(s) + V^{\tilde{\pi}}(s) - \tilde{V}^{\tilde{\pi}}(s) \\
&\leq H^2\left\|\pi - \tilde{\pi}\right\|_{\infty,1} + H^2\left\|P - \tilde{P}\right\|_{\infty,1} + H\left\|\ell - \tilde{\ell}\right\|_{\infty}.
\end{aligned}$$

Thus, let $s \in \mathcal{S}_h, a \in \mathcal{A}_i$, and observe;

$$\begin{aligned}
Q^{\pi}(s, a; M) - Q_h^{\tilde{\pi}}(s, a; \tilde{M}) &= \ell(s, a) - \tilde{\ell}(s, a) \\
&\quad + \mathbb{E}_{s' \sim P(\cdot|s,a)} V^{\pi}(s'; M) - \mathbb{E}_{s' \sim \tilde{P}(\cdot|s,a)} V^{\tilde{\pi}}(s'; \tilde{M}) \\
&= \ell(s, a) - \tilde{\ell}(s, a) \\
&\quad + \sum_{s' \in \mathcal{S}_{h+1}} P(s' \mid s, a)V^{\pi}(s'; M) - \tilde{P}(s' \mid s, a)V^{\tilde{\pi}}(s'; \tilde{M}) \\
&\leq |\ell(s, a) - \tilde{\ell}(s, a)| \\
&\quad + \sum_{s' \in \mathcal{S}_{h+1}} P(s' \mid s, a)|V^{\pi}(s'; M) - V^{\tilde{\pi}}(s'; \tilde{M})| \\
&\quad + \sum_{s' \in \mathcal{S}_{h+1}} V^{\tilde{\pi}}(s'; \tilde{M})|P(s' \mid s, a) - \tilde{P}(s' \mid s, a)| \\
&\leq |\ell(s, a) - \tilde{\ell}(s, a)| \\
&\quad + H^2\left\|\pi - \tilde{\pi}\right\|_{\infty,1} + H^2\left\|P - \tilde{P}\right\|_{\infty,1} + H\left\|\ell - \tilde{\ell}\right\|_{\infty} \\
&\quad + H\left\|P(\cdot \mid s, a) - \tilde{P}(\cdot \mid s, a)\right\|_1 \\
&\leq H^2\left\|\pi - \tilde{\pi}\right\|_{\infty,1} + (H^2 + 1)\left\|P - \tilde{P}\right\|_{\infty,1} + (H + 1)\left\|\ell - \tilde{\ell}\right\|_{\infty}
\end{aligned}$$

□

Lemma F.3. *For any policy $\mu \colon \mathcal{S} \to \mathcal{A}_i$, player $i \in [m]$, we have*

$$\left\|q^{\mu}_{P^i_{t+1}} - q^{\mu}_{P^i_t}\right\|_{\infty} \leq H^2 \sum_{j \neq i} \|\pi^j_{t+1} - \pi^j_t\|_{\infty,1}.$$

Proof. Follows by combining Lemma F.5 and Lemma F.4;

$$\left\|q^{\mu}_{P^i_{t+1}} - q^{\mu}_{P^i_t}\right\|_1 \leq H^2\left\|P^i_{t+1} - P^i_t\right\|_{\infty,1} \leq H^2 \sum_{j \neq i} \|\pi^j_{t+1} - \pi^j_t\|_{\infty,1}.$$

□

Lemma F.4. *It holds that for all $i \in [m]$, $s \in \mathcal{S}, a \in \mathcal{A}_i$;*

$$\|P^i_{t+1}(\cdot \mid s,a) - P^i_t(\cdot \mid s,a)\|_1 \le \sum_{j \neq i} \|\pi^j_{t+1}(\cdot \mid s) - \pi^j_t(\cdot \mid s)\|_1,$$
$$|\ell^i_{t+1}(s,a) - \ell^i_t(s,a)| \le \sum_{j \neq i} \|\pi^j_{t+1}(\cdot \mid s) - \pi^j_t(\cdot \mid s)\|_1.$$

Proof. For the losses, observe;

$$\begin{aligned}
\ell^i_{t+1}(s,a) - \ell^i_t(s,a) &= \mathbb{E}_{\mathbf{a}^{-i} \sim \boldsymbol{\pi}^{-i}_{t+1}} \ell^i(s,a,\mathbf{a}^{-i}) - \mathbb{E}_{\mathbf{a}^{-i} \sim \boldsymbol{\pi}^{-i}_{t}} \ell^i(s,a,\mathbf{a}^{-i}) \\
&= \sum_{\mathbf{a}^{-i} \in \mathcal{A}^{-i}} \left(\boldsymbol{\pi}^{-i}_{t+1}(\mathbf{a}^{-i} \mid s) - \boldsymbol{\pi}^{-i}_{t}(\mathbf{a}^{-i} \mid s)\right) \ell^i(s,a,\mathbf{a}^{-i}) \\
&\le \left\|\boldsymbol{\pi}^{-i}_{t+1}(\cdot \mid s) - \boldsymbol{\pi}^{-i}_{t}(\cdot \mid s)\right\|_1 .
\end{aligned}$$

For the induced transition function, note that for any $h \in [H]$, we have

$$\begin{aligned}
&\sum_{s' \in \mathcal{S}_{h+1}} P^i_{t+1}(s' \mid s,a) - P^i_t(s' \mid s,a) \\
&\qquad= \sum_{s' \in \mathcal{S}_{h+1}} \mathbb{E}_{\mathbf{a}^{-i} \sim \boldsymbol{\pi}^{-i}_{t+1}(\cdot|s)} \left[P(s' \mid s,a,\mathbf{a}^{-i})\right] - \mathbb{E}_{\mathbf{a}^{-i} \sim \boldsymbol{\pi}^{-i}_{t}(\cdot|s)} \left[P(s' \mid s,a,\mathbf{a}^{-i})\right] \\
&\qquad= \sum_{s' \in \mathcal{S}_{h+1}} \sum_{\mathbf{a}^{-i} \in \mathcal{A}^{-i}} (\boldsymbol{\pi}^{-i}_{t+1}(\mathbf{a}^{-i} \mid s) - \boldsymbol{\pi}^{-i}_{t}(\mathbf{a}^{-i} \mid s)) P(s' \mid s,a,\mathbf{a}^{-i}) \\
&\qquad= \sum_{\mathbf{a}^{-i} \in \mathcal{A}^{-i}} (\boldsymbol{\pi}^{-i}_{t+1}(\mathbf{a}^{-i} \mid s) - \boldsymbol{\pi}^{-i}_{t}(\mathbf{a}^{-i} \mid s)) \sum_{s' \in \mathcal{S}_{h+1}} P(s' \mid s,a,\mathbf{a}^{-i}) \\
&\qquad= \left\|\boldsymbol{\pi}^{-i}_{t+1}(\cdot \mid s) - \boldsymbol{\pi}^{-i}_{t}(\cdot \mid s)\right\|_1
\end{aligned}$$

By Lemma H.2, we have

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

and the result follows. □

Lemma F.5. *For any policy $\pi \in \mathcal{S} \to \mathcal{A}_i$ and single agent transition functions $P, \tilde{P}$, it holds that*

$$\|q^\pi_P - q^\pi_{\tilde{P}}\|_1 \le H^2 \|P - \tilde{P}\|_{\infty,1},$$
$$\|q^\pi_P - q^\pi_{\tilde{P}}\|_\infty \le H \|P - \tilde{P}\|_{\infty,1}.$$

Proof. Let $L \in [H]$, $z \in \mathcal{S}_L$, set loss function $\ell_z(s,a) = \mathbb{I}\{s = z\}$, and consider the two MDPs $M_z = (H, \mathcal{S}, \mathcal{A}_i, P, \ell_z)$ and $\tilde{M}_z = (H, \mathcal{S}, \mathcal{A}_i, \tilde{P}, \tilde{\ell}_z)$ with value functions $V_z, \tilde{V}_z$ respectively. Then, we have for any $s \in \mathcal{S}_h$, $V^\pi_z(s) = \Pr(s_L = z \mid s_h = s, P, \pi)$, and $\tilde{V}_z(s) = \Pr(s_L = z \mid s_h = s, \tilde{P}, \pi)$, which also implies $V^\pi_z(s_1) = q^\pi_P(z)$ and $\tilde{V}^\pi_z(s_1) = q^\pi_{\tilde{P}}(z)$. Thus, by Lemma F.1, we have;

$$\begin{aligned}
&q^\pi_P(z) - q^\pi_{\tilde{P}}(z) \\
&\quad= \sum_{h=1}^{L} \sum_{s_h, a_h} q^\pi_{\tilde{P}}(s_h, a_h) \sum_{s_{h+1}} (P^\pi(s_{h+1} \mid s_h, a_h) - \tilde{P}^\pi(s_{h+1} \mid s_h, a_h)) \Pr(s_L = z \mid s_{h+1}, P, \pi).
\end{aligned}$$

Taking absolute values and summing the above over $z \in \mathcal{S}_L$ we obtain

$$\begin{aligned}
\sum_{z \in \mathcal{S}_L} |q^\pi_P(z) - q^\pi_{\tilde{P}}(z)| &\le \sum_{h=1}^{L} \sum_{s_h, a_h} q^\pi_{\tilde{P}}(s_h, a_h) \sum_{s_{h+1}} |P^\pi(s_{h+1} \mid s_h, a_h) - \tilde{P}^\pi(s_{h+1} \mid s_h, a_h)| \\
&\le L \|P - \tilde{P}\|_{\infty,1}.
\end{aligned}$$

Hence,

$$\|q_P^\pi - q_{\tilde{P}}^\pi\|_1 = \sum_{L=1}^{H} \sum_{z \in \mathcal{S}_L} |q_P^\pi(z) - q_{\tilde{P}}^\pi(z)| \leq H^2 \|P - \tilde{P}\|_{\infty,1}.$$

□

G. FTRL lower bound in non-stationary MDP

In the following, we provide an example demonstrating that FTRL-based policy optimization does not adapt to non-stationary dynamics, at least not in the sense discussed here.

In a nutshell, since FTRL considers the entire sequence of past loss functions, it may not pick up on the change in the long term reward in a timely fashion. Indeed, since the policy optimization paradigm prescribes a per state objective that effectively ignores the visitation frequency to that state, FTRL allows past losses (induced by action-value functions from previous episodes) *that may be irrelevant* to bias the policy towards suboptimal actions for a prohibitively large number of episodes. Loosely speaking, this behavior is due to the fact that in contrast to OMD, FTRL is insensitive to the order of losses.

Notably, the failure of FTRL is strongly related to its inability to guarantee *adaptive* regret in the sense defined in Hazan & Seshadhri (2009), who also point out the inherent non-adaptivity of this algorithm.

The claim below illustrates an example of an MDP with a small constant change in the dynamics leading to FTRL incurring linear regret. Essentially, this is a simple example where FTRL fails to achieve adaptive regret, embedded in a non-stationary MDP. We remark that while the MDP in our construction makes a single, abrupt shift in the dynamics, the lower bound does not stem from the abruptness of the change. Rather, this choice is only for simplicity; the construction may be generalized to the case where the per episode drift must be bounded by e.g., $1/\sqrt{T}$, by augmenting the construction with a "shift period" of $\sqrt{T}$ episodes. In addition, it is not hard to show the construction can be generalized to subsets of the action simplex — this is to say that the lower bound also does not stem from lack of exploration that can be solved by truncating the simplex, as we have done in the OMD case. Finally, we remark that the same lower bound remains valid also when considering OFTRL; 1-step recency bias does not make the algorithm sufficiently adaptive for the example in question.

We refer in the statement to a symmetric regularizer, meaning one that is insensitive to permutations of the input coordinates. This assumption is only for simplicity; it can be relaxed by generalizing the instance appearing in the lower bound to a mixture of two instances with action roles reversed, and observing that on at least one of them FTRL must incur linear regret.

Claim 1. There exists a non-stationary MDP $M = (S, \{a, b\}, \{P_t\}_{t=1}^T, \ell)$, such that $\sum_{t=2}^T \|P_t - P_{t-1}\|_1 \leq 1$, but policy optimization with FTRL over the action simplex $\Delta_{\{a,b\}}$, any symmetric regularizer, and any step size incurs regret of $\Omega(T)$.

Proof. Let $S = \{s_0, s_1, s_2, L_0, L_1\}$ denote the state space, and consider the non-stationary MDP $M = (S, \{a, b\}, \{P_t\}_{t=1}^T, \ell)$, where the immediate loss function is independent of the action and is specified by $\ell(s_i) = 0\ \forall i$, $\ell(L_0) = 0$, and $\ell(L_1) = 1$. Further, assume that

- for $t \leq T/3$, $P_t(s_1|s_0, \cdot) = 1$, $P_t(L_1|s_2, a) = 1$, and $P_t(L_0|s_2, b) = 1$, (see Figure 1)
- for $t > T/3$, $P_t(s_2|s_0, \cdot) = 1$, $P_t(L_1|s_2, b) = 1$, and $P_t(L_0|s_2, a) = 1$ (see Figure 2).

Consider running policy optimization with FTRL over $\mathcal{X} = \Delta_{\{a,b\}}$, and a symmetric regularizer $R\colon \mathcal{X} \to \mathbb{R}$ for T episodes.

First, observe that the optimal policy in hindsight $\pi_\star$ selects action a with probability 1 in state s_2; $\pi_\star(a|s_2) = 1$. Note that the actions chosen in the rest of the states do not affect the loss, and therefore need not be specified. Thus, in the first $T/3$ episodes, $\pi_\star$ loses nothing since state s_2 is never reached, and in the remaining $2T/3$ episodes it loses nothing on account of selecting an action which leads to L_0. This establishes that $\sum_{t=1}^T V^{\pi_\star}(s_0; P_t) = 0$.

On the other hand, for $t > T/3$, the FTRL objective on episode t at state s_2, is given by

$$\pi_{t+1}(\cdot|s_2) \leftarrow \arg\min_{x \in \mathcal{X}} \left\{ \eta \left\langle \sum_{j=1}^{t} \hat{\ell}_t, x \right\rangle + R(x) \right\},$$

where $\hat{\ell}_t(\cdot) \coloneqq Q_t(s_2, \cdot)$. Thus,

$$\begin{cases} \hat{\ell}_t(a) = 1, \hat{\ell}_t(b) = 0 & t \le T/3, \\ \hat{\ell}_t(a) = 0, \hat{\ell}_t(b) = 1 & t > T/3, \end{cases}$$

which leads us to conclude that in all rounds $t \le 2T/3$, the action b seems favorable according to the minimization objective. This implies that for all $t \le 2T/3$, $\pi_t(b|s_2) \ge 1/2$. Note that we use here the fact that the regularizer and decision set are symmetric, and treat all coordinates equally. Now,

$$\sum_{t=1}^{T} V^{\pi_t}(s_0; P_t) - V^{\pi_\star}(s_0; P_t) \ge \sum_{t=T/3}^{2T/3} V^{\pi_t}(s_0; P_t) \ge \frac{T}{6},$$

as claimed. □

H. Auxiliary Lemmas

Lemma H.1. *Let $k \in \mathbb{N}$, and consider the truncated simplex $\Delta_k^\gamma \subseteq \Delta_k$ (see Equation (5)). It holds that:*

1. *For log-barrier regularizer $R\colon \Delta_k^\gamma \to \mathbb{R}$ (see Equation (4)), we have $D_R(x, x') \le \frac{3}{\gamma}$ for all $x, x' \in \Delta_k^\gamma$.*
2. *If $0 < \gamma \le 1/2k$, for all $x \in \Delta_k$, there exists $x^\gamma \in \Delta_k^\gamma$ such that $\|x - x^\gamma\|_1 \le 2\gamma k$.*

Proof. See below.

- We have, for any $x, x' \in \Delta_{\mathcal{A}_i}^\gamma$;

$$D_R(x, x') = \sum_{a \in \mathcal{A}_i} \log \frac{x(a)}{x'(a)} + \frac{x(a) - x'(a)}{x'(a)} \le \log \frac{1}{\gamma} + \frac{2}{\gamma} \le \frac{3}{\gamma}.$$

- Let $I = \{i \in [k] \mid x(a) \le \gamma\}$. Then $|I| \le k - 1$, otherwise $\sum_{i=1}^k x(i) \le 1/2$. Now, set $\tilde{x}^\gamma(i) = \gamma$, for $i \in I$, and $\tilde{x}^\gamma(i) = x(i)$ for $i \notin I$. We have

$$\sum_{i=1}^{k} \tilde{x}^\gamma(i) = 1 + \delta, \quad \text{where } \delta \le \gamma|I|,$$

and $\|x - \tilde{x}^\gamma\|_1 \le (k-1)\gamma$. Now, subtract from the largest coordinate value $\tilde{x}^\gamma(i_{\max})$ the excess weight δ. In the event that $\tilde{x}^\gamma(i_{\max}) \le \gamma + \delta$, subtract to γ, and continue iteratively to the second largest etc. This process must terminate before reaching coordinates in I, since $\sum_{i=1}^k x(i) = 1$. Now, $\|x^\gamma\|_1 = 1$, and

$$\|x - x^\gamma\|_1 \le (k-1)\gamma + \delta \le (2k-1)\gamma.$$

□

Lemma H.2. *Let p and q be any two product distributions over $X_1 \times \cdots \times X_m$, i.e., $p(x_1, \ldots, x_m) = \prod_{i=1}^m p_i(x_i)$, and $q(x_1, \ldots, x_m) = \prod_{i=1}^m q_i(x_i)$. Then*

$$\|p - q\|_1 \le \sum_{i=1}^{m} \|p_i - q_i\|_1 .$$

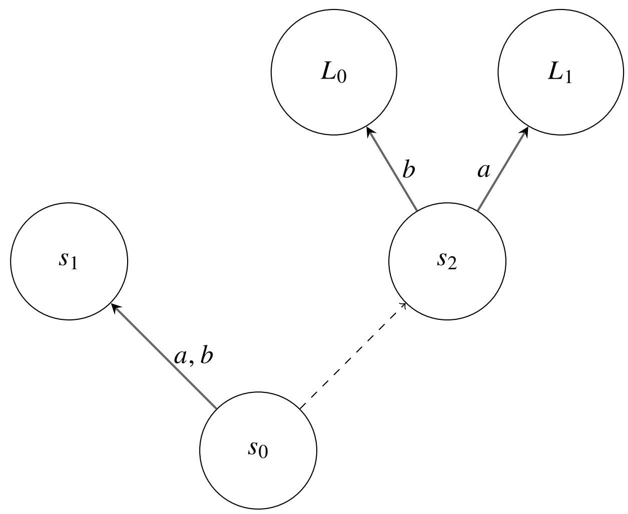

Figure 1. MDP at $t = 1, \ldots, T/3$

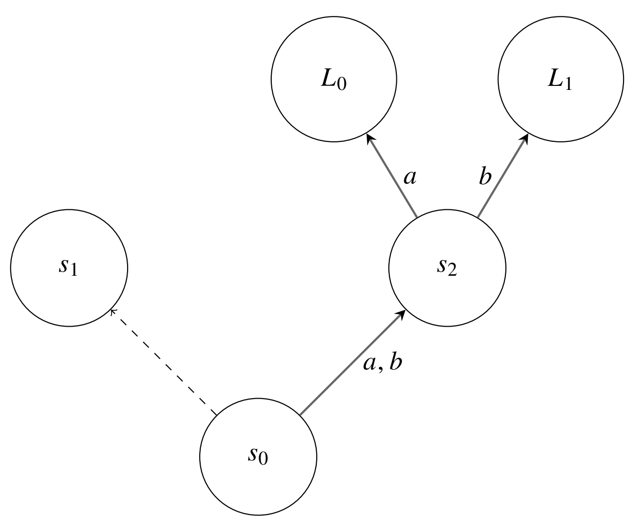

Figure 2. MDP at $t = T/3 + 1, \ldots, T$

Proof. We have;

$$
\begin{aligned}
\|p-q\|_1 &= \sum_{x_1\in X_1}\cdots\sum_{x_m\in X_m}\left|p_m(x_m)\prod_{i=1}^{m-1}p_i(x_i) - q_m(x_m)\prod_{i=1}^{m-1}q_i(x_i)\right| \\
&\le \sum_{x_1\in X_1}\cdots\sum_{x_m\in X_m}\prod_{i=1}^{m-1}p_i(x_i)\Big|p_m(x_m)-q_m(x_m)\Big| \\
&\quad + \sum_{x_1\in X_1}\cdots\sum_{x_m\in X_m} q_m(x_m)\left|\prod_{i=1}^{m-1}p_i(x_i) - \prod_{i=1}^{m-1}q_i(x_i)\right| \\
&= \|p_m-q_m\|_1\sum_{x_1\in X_1}\cdots\sum_{x_{m-1}\in X_{m-1}}\prod_{i=1}^{m-1}p_i(x_i) \\
&\quad + \left\|\prod_{i=1}^{m-1}p_i - \prod_{i=1}^{m-1}q_i\right\|_1\sum_{x_m\in X_m}q_m(x_m) \\
&= \|p_m-q_m\|_1 + \left\|\prod_{i=1}^{m-1}p_i - \prod_{i=1}^{m-1}q_i\right\|_1,
\end{aligned}
$$

and the claim follows by induction. □

I. Markov Games with Independent Transition Function

In this section we consider a variant of Markov Games for which each agent has its own state and the transition is affected only by the agent's own action. Formally, each agent has its own set of states $\mathcal{S}^i$. Further, P is the transition kernel, where given the state at time h, $s\in\mathcal{S}^i_h$, and the agent's action $a\in\mathcal{A}_i$, $P(\cdot\mid s,a)\in\Delta_{\mathcal{S}^i_{h+1}}$ is the probability distribution over the next state. The loss function at time h depends on the states and actions at time h of all agents: $\ell^i_h : (\bigtimes_{i\in[m]}\mathcal{S}^i_h)\times\mathcal{A}\to[0,1]$ The policy of player i, depends on its individual state. That is, $\pi^i(\cdot\mid\cdot):\mathcal{A}_i\times\mathcal{S}^i\to[0,1]$, is a function such that $\pi^i(a\mid s)$ gives the probability of player i to take action a in state s. Similar to before, denote the expected loss function of agent i at time t given action a and state $s\in\mathcal{S}^i_h$ by $\ell^i_t(s,a)=\mathbb{E}[\ell^i(\mathbf{s},\mathbf{a})\mid\boldsymbol{\pi}_t, s^i_h=s]$ where $\boldsymbol{\pi}_t$ is the joint policy of the agents and $\mathbf{s}=(s^1_h,\ldots,s^m_h)$ is the vector of the agents' states at time h. Similar to before, we denote the value and action-value functions of a policy $\pi\in\mathcal{S}\to\Delta_{\mathcal{A}_i}$ by

$$
V^{i,\pi}_t(s)=\mathbb{E}\left[\sum_{h'=h}^{H}\ell^i_t(\mathbf{s}_{h'},\mathbf{a}_{h'})\mid\boldsymbol{\pi}^{-i}_t, s_h=s\right]\ ;\ Q^{i,\pi}_t(s,a)=\mathbb{E}\left[\sum_{h'=h}^{H}\ell^i_t(\mathbf{s}_{h'},\mathbf{a}_{h'})\mid\boldsymbol{\pi}^{-i}_t, s_h=s\right],
$$

where $s\in\mathcal{S}^i_h$ and $a\in\mathcal{A}_i$. We note that we sometimes use the shorthand $V^i_t(\cdot)$ for $V^{i,\pi^i_t}_t(\cdot)$ and $Q^i_t(\cdot,\cdot)$ for $Q^{i,\pi^i_t}_t(\cdot,\cdot)$.

In the setting of individual state transitions, it is possible to achieve much better regret bounds than in Markov games; specifically, we show that using Algorithm 1 each player can obtain $O(\log T)$ individual swap regret. This possibility stems from the fact that in contrast to Markov games, the MDPs each player experiences throughout the episodes remain constant, and hence it is possible to obtain a regret bound which depends on the sum of second order path lengths of the players' policies rather than on the first order path lengths (see Theorem 3.4 for the corresponding result for Markov games).

Theorem I.1. *In the independent transition function setting, assume that every player i adopts Algorithm 1 with log-barrier regularization (Equation (4)) and $\gamma\le 1/2A_i$, and that $\|\hat{Q}_t-Q_t\|_\infty\le\varepsilon$ for all t. Then, assuming $H\ge 2$, the swap-regret of player i is bounded as*

$$
\mathrm{Swap}\mathfrak{R}^i_T \le \frac{A^2\log\frac{1}{\gamma}}{\eta} + 24\eta H^4 Am\sum_{j=1}^{m}\sum_{t=1}^{T}\left\|\pi^j_{t+1}-\pi^j_t\right\|^2_{\infty,1} + \varepsilon HT + 8\eta\varepsilon^2 T
$$

Proof. As opposed to the standard Markov game setting, the occupancy measure of the benchmark policy remains stationary over time: $q_t^{i,\star} = q^{i,\star}$. Therefore, much like in the proof of Theorem 3.4,

$$\begin{aligned}\mathfrak{R}_T^i(\pi_\star^i) &\le \varepsilon HT + \sum_{s\in\mathcal{S}} q^{i,\star}(s) \sum_{t=1}^T \left\langle \hat{Q}_t^i(s,\cdot), \pi_t^i(\cdot \mid s) - x_\star^{i,s} \right\rangle \\ &\le \varepsilon HT + \sum_{s\in\mathcal{S}} q^{i,\star}(s) \sum_{a\in\mathcal{A}_i} \sum_{t=1}^T \left\langle g_t^{i,s,a}, x_t^{i,s,a} - x_\star^{i,s} \right\rangle.\end{aligned}$$

From Lemma 3.6,

$$\begin{aligned}\sum_{t=1}^T \left\langle g_t^{i,s,a}, x_t^{i,s,a} - x_\star^{i,s} \right\rangle &\le \frac{A\log\frac{1}{\gamma}}{\eta} + 4\sum_{t=1}^T \eta \left\| g_t^{i,s,a} - g_{t-1}^{i,s,a} \right\|_{*,x_t^{i,s,a}}^2 \\ &\le \frac{A\log\frac{1}{\gamma}}{\eta} + 24\eta H^4 m \sum_{j=1}^m \sum_{t=1}^T \left\| \pi_{t+1}^j - \pi_t^j \right\|_{\infty,1}^2 + 8\eta\varepsilon^2 \sum_{t=1}^T \pi_{t+1}^j(a \mid s).\end{aligned}$$

Combining the last two displays completes the proof. □

Theorem I.2. *If each player uses Algorithm 1 with log-barrier regularization (Equation (4)) and $\eta = \frac{1}{96H^2 m\sqrt{SA}}$ then the following path length bound holds on the jointly generated policy sequence;*

$$\sum_{t=1}^T \sum_{i=1}^m \left\| \pi_{t+1}^i - \pi_t^i \right\|_{\infty,1}^2 \le 768SA^3 m \log\frac{1}{\gamma} + \frac{4\varepsilon^2 T}{mH^4}.$$

The proof follows by the exact same arguments in the proof of Theorem 3.5. Combining Theorems I.1 and I.2 gives us the following corollary:

Corollary I.3. *In the independent transition function setting with full information (i.e., $\varepsilon = 0$), assume that every player i adopts Algorithm 1 with log-barrier regularization (Equation (4)), $\eta = \frac{1}{96H^2 m\sqrt{SA}}$ and $\gamma = 1/T$. Then, assuming $H \ge 2$ and $T \ge 2A$, the swap-regret of player i is bounded as*

$$\mathrm{Swap}\mathfrak{R}_T^i \le 288 H^2 S^{3/2} A^{7/2} m \log T.$$