Uncoupled Learning Dynamics with $O(\log T)$ Swap Regret in Multiplayer Games

Ioannis Anagnostides
Carnegie Mellon University
Pittsburgh, PA 15213
ianagnos@cs.cmu.edu

Gabriele Farina
Carnegie Mellon University
Pittsburgh, PA 15213
gfarina@cs.cmu.edu

Christian Kroer
Columbia University
New York, NY 10027
christian.kroer@columbia.edu

Chung-Wei Lee
University of Southern California
Los Angeles, CA 90007
leechung@@usc.edu

Haipeng Luo
University of Southern California
Los Angeles, CA 90007
haipengl@usc.edu

Tuomas Sandholm
Carnegie Mellon University
Strategy Robot, Inc.
Optimized Markets, Inc.
Strategic Machine, Inc.
Pittsburgh, PA 15213
sandholm@cs.cmu.edu

Abstract

In this paper we establish efficient and *uncoupled* learning dynamics so that, when employed by all players in a general-sum multiplayer game, the *swap regret* of each player after T repetitions of the game is bounded by $O(\log T)$, improving over the prior best bounds of $O(\log^4(T))$. At the same time, we guarantee optimal $O(\sqrt{T})$ swap regret in the adversarial regime as well. To obtain these results, our primary contribution is to show that when all players follow our dynamics with a *time-invariant* learning rate, the *second-order path lengths* of the dynamics up to time T are bounded by $O(\log T)$, a fundamental property which could have further implications beyond near-optimally bounding the (swap) regret. Our proposed learning dynamics combine in a novel way *optimistic* regularized learning with the use of *self-concordant barriers*. Further, our analysis is remarkably simple, bypassing the cumbersome framework of higher-order smoothness recently developed by Daskalakis, Fishelson, and Golowich (NeurIPS'21).

1 Introduction

Online learning and game theory share an intricately connected history tracing back to the inception of the modern *no-regret framework* with Robinson's analysis of *fictitious play* [Robinson, 1951] and Blackwell's *approachability theorem* [Blackwell, 1956]. Indeed, the no-regret framework addresses the fundamental question of how independent and decentralized agents can "learn" with only limited feedback from their environment, and has led to celebrated connections with game-theoretic equilibrium concepts [Hart and Mas-Colell, 2000, Foster and Vohra, 1997]. One of the remarkable features of these results is that the learning dynamics are fully *uncoupled* [Hart and Mas-Colell, 2000]: each player is completely agnostic to the utilities of the other players. Thus, there

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

is no communication between the players or any centralized authority dictating behavior throughout the game. Instead, the only "coordination device" is the common history of play. An additional desideratum, which is fundamentally tied to the no-regret framework, is what Daskalakis et al. [2011] refer to as *strong uncoupledness*:[1] players have no information whatsoever about the game (even their own utilities), and they only make decisions based on the utilities received as feedback throughout the repeated game.

In this context, it is well-known that there are broad families of no-regret learning algorithms that, after T repetitions, guarantee regret bounded by $O(\sqrt{T})$, and this bound is known to be insuperable in adversarial environments [Cesa-Bianchi and Lugosi, 2006]. However, this begs the question: *What if the player is not facing adversarial utilities, but instead is competing with other learning agents in a repeated game?* This question was first formulated and addressed by Daskalakis et al. [2011], who devised strongly uncoupled dynamics converging with a near-optimal rate of $O(\frac{\log T}{T})$ in zero-sum games, a substantial improvement over the $O(1/\sqrt{T})$ rate obtained via traditional approaches within the no-regret framework. Thereafter, there has been a considerable amount of effort in strengthening their result, leading to extensions along several important lines [Rakhlin and Sridharan, 2013, Syrgkanis et al., 2015, Chen and Peng, 2020, Farina et al., 2019, Daskalakis et al., 2021, Anagnostides et al., 2022a, Wei and Luo, 2018, Foster et al., 2016, Anagnostides et al., 2022b]. In particular, in a recent breakthrough result, Daskalakis et al. [2021] showed that when all players in a general game employ an *optimistic* variant of *multiplicative weights update (MWU)* (henceforth *OMWU*), the *external regret* of each player grows as $O(\log^4(T))$. That result was also subsequently extended to the substantially more challenging performance measure of *swap regret* [Anagnostides et al., 2022a]. Perhaps the main drawback of the latter results is the complexity of the analysis, relying on establishing a refined property for the dynamics they refer to as *higher-order smoothness*. Our primary contribution in this paper is to develop a novel and much simpler framework, which furthermore improves the prior state of the art $O(\log^4(T))$ regret bounds to $O(\log T)$ in general multiplayer games.

1.1 Overview of Our Contributions

Before we state our main result, let us first introduce some basic notation. We assume that each player $i \in [\![n]\!]$ selects at every iteration t of the repeated game a probability distribution (mixed strategy) over the set of available actions $\boldsymbol{x}_i^{(t)} \in \Delta(\mathcal{A}_i)$ (see Section 2 for further details). The following theorem is the primary contribution of our work.[2]

Theorem 1.1 (Precise Statement in Theorem 4.4)**.** *There exist strongly uncoupled no-swap-regret learning dynamics so that when employed by all players with learning rate $\eta = \Theta(1)$, the second-order path lengths of the dynamics up to any time $T \in \mathbb{N}$ are bounded by $O(\log T)$; that is,*

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 = O(\log T).$$

We are not aware of even an $o(T)$ bound for the second-order path lengths—under a *time-invariant* learning rate—prior to our work, except for very restricted classes of games such as zero-sum games. The dynamics of Theorem 1.1 combine: (i) the celebrated no-swap-regret template of Blum and Mansour [2007]; (ii) the *optimistic follow the regularizer leader (OFTRL)* algorithm of Syrgkanis et al. [2015]; and (iii) using a *self-concordant barrier* as a regularizer. The latter was introduced in online learning in the seminal work of Abernethy et al. [2008], where the authors obtained the first near-optimal and efficient online learning algorithm for linear bandit optimization; the way we leverage the *log-barrier* in the setting of no-regret learning in games is novel, and crucially leverages the *local norm* induced by the regularizer. The dynamics of Theorem 1.1 are also efficiently implementable (see Remark 4.7).

The implication of Theorem 1.1 is perhaps surprising in view of the inherent *cycling* aspect of no-regret learning in general games. Indeed, it is by now well-understood that any no-regret dynamics

[1] Daskalakis et al. [2011] also impose that players are only allowed to (privately) store only a constant number of observed utilities, an assumption also espoused in our work.

[2] For simplicity in the exposition, we use the $O(\cdot)$ notation in our introduction to suppress parameters that depend (polynomially) on the natural parameters of the game; precise statements are given in Section 4.

Reference	Algorithm	Swap Regret in Games	Adversarial Swap Regret
Blum and Mansour [2007]	*E.g.*, BM-MWU	—	$O(\sqrt{m\log(m)T})$
Chen and Peng [2020]	BM-OMWU	$O(\sqrt{n}(m\log(m))^{3/4})T^{1/4})$	$\widetilde{O}(\sqrt{mT})$
Anagnostides et al. [2022a]	BM-OMWU	$O(nm^4\log(m)\log^4(T))$	—
This paper	BM-OFTRL-LogBar	$O(nm^{5/2}\log(T))$	$O(\sqrt{m\log(m)T})$

Table 1: Prior results regarding the no-swap-regret algorithm of Blum and Mansour [2007] (BM). The second column indicates the algorithm internally employed by the "master" BM algorithm; our construction uses OFTRL with log-barrier regularization (Section 3). Further, m is the maximum number of actions available to each player. We point out that in the adversarial swap regret bound we have suppressed lower order factors in terms of T. We further remark that the near-optimal *internal regret* guarantee of Anagnostides et al. [2022a] in turn implies $O(nm\log(m)\log^4(T))$ swap regret for each individual player, but is obtained via the algorithm of Stoltz and Lugosi [2005].

will fail to converge—at least for certain games (*e.g.*, see [Milionis et al., 2022]). Nevertheless, Theorem 1.1 implies that players will change their strategies *arbitrarily slowly* as the game progresses. As such, players will observe utilities that exhibit very small variation over time, immediately implying near-optimal swap regret.

Corollary 1.2 (Precise Statement in Corollaries 4.5 and 4.6)**.** *There exist strongly uncoupled no-swap-regret learning dynamics so that when employed by all players, the individual swap regret of each player is bounded by $O(\log T)$. At the same time, when faced against adversarial utilities each player guarantees $O(\sqrt{T})$ swap regret.*

Corollary 1.2 improves over the prior best bounds of $O(\log^4(T))$ [Daskalakis et al., 2021, Anagnostides et al., 2022a]; a comparison with prior works regarding the algorithm of Blum and Mansour [2007] is given in Table 1. In fact, Corollary 1.2 yields, to our knowledge, the first no-regret guarantee in general games for uncoupled methods when players use a *time-invariant* learning rate, a feature that has been extensively motivated in prior works (see, *e.g.*, the discussion in [Bailey and Piliouras, 2019]). Corollary 1.2 also establishes near-optimality in the adversarial regime as well, a crucial desideratum in this line of work. Finally, swap regret is a powerful notion of hindsight rationality, trivially subsuming external regret. In particular, in light of well-established connections (see Theorem 2.3), we obtain the best known rate of convergence of $O(\frac{\log T}{T})$ to *correlated equilibria* in general games.

Corollary 1.3. *There exist strongly uncoupled learning dynamics so that, when employed by all players, the average correlated distribution of play after T repetitions of the game is an $O(\frac{\log T}{T})$-approximate correlated equilibrium.*

From a technical standpoint, our approach is conceptually remarkably simple and direct. Specifically, Theorem 1.1 is shown by first establishing the RVU bound—a fundamental property first identified in [Syrgkanis et al., 2015, Definition 3]—for *swap regret* in Theorem 4.3; the key ingredient is Lemma 4.2, which crucially leverages the local norm induced by the log-barrier regularizer over the simplex. Next, Theorem 1.1 follows directly by making a seemingly trivial observation: *swap regret is always nonnegative*. A related approach was recently employed in [Anagnostides et al., 2022c] for external regret, but only works for very restricted classes of games such as zero-sum. As such, we bypasses the cumbersome framework of higher-order smoothness recently introduced by Daskalakis et al. [2021].

1.2 Further Related Work

The first accelerated dynamics in general games were established by Syrgkanis et al. [2015]. In particular, they identified a broad class of no-regret learning dynamics—satisfying the so-called RVU property—for which the *sum* of the players' regrets is $O(1)$. On the other hand, they only obtained an $O(T^{1/4})$ bound for the individual external regret of each player. This is crucial given that the rate of convergence to *coarse* correlated equilibria is driven by the *maximum* of the external regrets. It is important to note that a bound for the sum of the external regrets does not necessarily translate to a bound for the maximum since *external regrets* can be negative. This is in stark contrast to swap regret

(Observation 2.1), a property crucially leveraged in our work. Furthermore, the $O(T^{1/4})$ bounds for the individual external regret in [Syrgkanis et al., 2015] were only recently extended to swap regret by Chen and Peng [2020]. The main challenge with swap regret—which is also the main focus of our paper—is that the underlying dynamics are much more complex, maintaining and aggregating over multiple independent external regret minimizers. In addition, the dynamics involve a fixed point operation—namely, the stationary distribution of a Markov chain—posing new challenges compared to the analysis of no-external-regret algorithms [Chen and Peng, 2020]. Finally, a very intriguing approach for obtaining near-optimal no-external-regret dynamics was recently introduced by Piliouras et al. [2021]. The main caveat of that result is that the dynamics they propose are *not* uncoupled, which has been a central desideratum in the line of work on no-regret learning in games. For this reason, the result in [Piliouras et al., 2021] is not directly comparable with the previous approaches.

2 Preliminaries

In this section we introduce the basic background on online optimization and learning in games. For a comprehensive treatment on the subject we refer the interested reader to the excellent book of Cesa-Bianchi and Lugosi [2006].

Conventions We denote by $\mathbb{N} = \{1, 2, \dots\}$ the set of natural numbers. We use the shorthand notation $[\![n]\!] \coloneqq \{1, 2, \dots, n\}$. Subscripts are typically used to indicate the player, or a parameter uniquely associated with a player (such as an action available to the player). On the other hand, superscripts are reserved almost exclusively for the (discrete) time index, which is represented via the variable t. Also, the r-th coordinate of a d-dimensional vector $\boldsymbol{x} \in \mathbb{R}^d$ is denoted by $\boldsymbol{x}[r]$. Finally, we let $\log(\cdot)$ be the natural logarithm.

2.1 Online Learning and Phi-Regret

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a nonempty convex and compact set of strategies, for some $d \in \mathbb{N}$. In the online learning framework the *learner* has to select at every iteration $t \in \mathbb{N}$ a *strategy* $\boldsymbol{x}^{(t)} \in \mathcal{X}$. Then, the environment—be it the "nature" or some "adversary"—returns a (linear) utility function $u^{(t)} : \mathcal{X} \ni \boldsymbol{x} \mapsto \langle \boldsymbol{x}, \boldsymbol{u}^{(t)} \rangle$, for some utility vector $\boldsymbol{u}^{(t)} \in \mathbb{R}^d$, so that the learner receives a utility of $\langle \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle$ at time t. In the *full information* model the learner receives as feedback the entire utility function, represented by $\boldsymbol{u}^{(t)}$. The canonical measure of performance in online learning is based on the notion of *regret*, or more generally, on *Phi-regret* [Greenwald and Jafari, 2003, Stoltz and Lugosi, 2007, Gordon et al., 2008]. Formally, for a set of transformations $\Phi : \mathcal{X} \to \mathcal{X}$, the Φ-*regret* of a regret minimization algorithm $\mathfrak{R}$ up to a *time horizon* $T \in \mathbb{N}$ is defined as

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

Naturally, a broader collection of transformations leads to a stronger notion of hindsight rationality; canonical instantiations of Phi-regret include:

(i) *External regret* (denoted by Reg): Φ includes only *constant transformations*;

(ii) *Swap regret* (denoted by $\mathrm{SwapReg}$): Φ includes *all* possible linear transformations.

As such, swap regret induces the more powerful notion of hindsight rationality. We point out that our main focus in this paper (Section 4) will be for the special case where $\mathcal{X}$ is the probability simplex. A crucial property of swap regret is that $\mathrm{SwapReg} \geq 0$, as formalized below.

Observation 2.1. *Fix any time horizon $T \in \mathbb{N}$. For any sequence of utilities $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(T)}$ and any sequence of strategies $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(T)}$ it holds that* $\mathrm{SwapReg}^T \geq 0$.

In proof, just consider the identity transformation $\Phi \ni \phi : \boldsymbol{x} \mapsto \boldsymbol{x}$ in (1). In contrast, this property does not necessarily hold for external regret.

Moreover, it will be convenient to model a regret minimization algorithm $\mathfrak{R}$ as a black box which interacts with its environment via the following two subroutines.

(i) $\mathfrak{R}$.NEXTSTRATEGY(): $\mathfrak{R}$ returns the next strategy of the learner;

(ii) $\mathfrak{R}$.OBSERVEUTILITY($\boldsymbol{u}$): $\mathfrak{R}$ receives as feedback from the environment a utility vector $\boldsymbol{u}$, and may adapt its internal state accordingly.

2.2 No-Regret Learning and Correlated Equilibria

A fundamental connection ensures that as long as all players employ no-swap-regret learning dynamics (in the sense that $\mathrm{SwapReg}^T = o(T)$), the average correlated distribution of play converges to the set of correlated equilibria [Hart and Mas-Colell, 2000, Foster and Vohra, 1997, Blum and Mansour, 2007]. Before we formalize this connection, let us first introduce some basic background on games.

Finite Games Let $[\![n]\!] := \{1, 2, \dots, n\}$ be the set of players, with $n \geq 2$. In a (finite) game, represented in *normal form*, each player $i \in [\![n]\!]$ has a finite set of actions $\mathcal{A}_i$; for notational simplicity, we will let $m_i := |\mathcal{A}_i| \geq 2$. For a given joint action profile $\boldsymbol{a} = (a_1, \dots, a_n) \in \bigtimes_{i=1}^n \mathcal{A}_i$, the (normalized) utility received by player i is given by some arbitrary function $u_i : \bigtimes_{i=1}^n \mathcal{A}_i \to [-1, 1]$. Players are allowed to randomize by selecting a (mixed) strategy $\boldsymbol{x}_i \in \Delta(\mathcal{A}_i) := \left\{\boldsymbol{x} \in \mathbb{R}_{\geq 0}^{|\mathcal{A}_i|} : \sum_{a_i \in \mathcal{A}_i} \boldsymbol{x}[a_i] = 1\right\}$; that is, a probability distribution over the available actions. For a joint strategy profile $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, player i receives an *expected utility* of $\mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{x}}[u_i(\boldsymbol{a})] = \sum_{\boldsymbol{a} \in \mathcal{A}} u_i(\boldsymbol{a}) \prod_{j \in [\![n]\!]} \boldsymbol{x}_j[a_j]$.

In the problem of no-regret learning in games, every player receives as feedback at time $t \in \mathbb{N}$ a utility vector $\boldsymbol{u}_i^{(t)} \in \mathbb{R}^{|\mathcal{A}_i|}$, so that $\boldsymbol{u}_i^{(t)}[a_i] := u_i(a_i; \boldsymbol{x}_{-i}^{(t)}) := \mathbb{E}_{\boldsymbol{a}_{-i} \sim \boldsymbol{x}_{-i}}[u_i(a_i, \boldsymbol{a}_{-i})]$, for any $a_i \in \mathcal{A}_i$; here, we used the notation $\boldsymbol{a}_{-i}$ to represent the joint action profile excluding i's component, and analogously for the notation $\boldsymbol{x}_{-i}$. *No other information is available to the player*. We are now ready to introduce the concept of a *correlated equilibrium* due to Aumann [1974].

Definition 2.2 (Correlated Equilibrium [Aumann, 1974])**.** A probability distribution $\boldsymbol{\mu}$ over $\bigtimes_{i=1}^n \mathcal{A}_i$ is an *ϵ-approximate correlated equilibrium*, for $\epsilon \geq 0$, if for any player $i \in [\![n]\!]$ and any swap function $\phi_i : \mathcal{A}_i \to \mathcal{A}_i$,

$$\mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\mu}}[u_i(\boldsymbol{a})] \geq \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\mu}}[u_i(\phi_i(a_i), \boldsymbol{a}_{-i})] - \epsilon.$$

Theorem 2.3 (Folklore)**.** *Suppose that each player $i \in [\![n]\!]$ employs a no-swap-regret algorithm such that the cumulative swap regret up to time $T \in \mathbb{N}$ is upper bounded by $\mathrm{SwapReg}_i^T$. Further, let $\boldsymbol{\mu}^{(t)} := \boldsymbol{x}_1^{(t)} \otimes \boldsymbol{x}_2^{(t)} \otimes \cdots \otimes \boldsymbol{x}_n^{(t)}$ be the product distribution at time $t \in [\![T]\!]$, and $\bar{\boldsymbol{\mu}} := \frac{1}{T} \sum_{t=1}^T \boldsymbol{\mu}^{(t)}$ be the average correlated distribution of play up to time T. Then, $\bar{\boldsymbol{\mu}}$ is a $\max_{i=1}^n \{\mathrm{SwapReg}_i^T / T\}$-approximate correlated equilibrium.*

Consequently, a central challenge for correlated equilibria is that the rate of convergence is driven by the *maximum* of the swap regrets; this is in contrast to, for example, the rate of convergence of the (utilitarian) social welfare in *smooth games*, which is driven by the *sum* of the players' external regrets [Syrgkanis et al., 2015, Roughgarden, 2015].

3 Optimistic Learning with Self-Concordant Barriers

Optimistic follow the regularizer leader (OFTRL) [Syrgkanis et al., 2015] is a *predictive* variant of the standard FTRL paradigm. Specifically, OFTRL maintains an internal *prediction* vector $\boldsymbol{m}^{(t)} \in \mathbb{R}^d$, and can be expressed with the following update rule for $t \in \mathbb{N}$.

$$\boldsymbol{x}^{(t)} := \arg\max_{\boldsymbol{x} \in \mathcal{X}} \left\{ \Phi^{(t)}(\boldsymbol{x}) := \eta \left\langle \boldsymbol{x}, \boldsymbol{m}^{(t)} + \sum_{\tau=1}^{t-1} \boldsymbol{u}^{(\tau)} \right\rangle - \mathcal{R}(\boldsymbol{x}) \right\}; \tag{OFTRL}$$

here, $\eta > 0$ serves as the *learning rate*, and $\mathcal{R}$ is the *regularizer*. For convenience, we also define $\boldsymbol{x}^{(0)} := \arg\min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{R}(\boldsymbol{x})$. Unless specified otherwise, (OFTRL) will be instantiated with $\boldsymbol{m}^{(t)} := \boldsymbol{u}^{(t-1)}$, for $t \in \mathbb{N}$. (For convenience in the analysis, and without any loss, we assume that players initially obtain the utilities corresponding to the other players' strategies at time $t = 0$.)

In [Syrgkanis et al., 2015] the regularizer $\mathcal{R}$ was assumed to be 1-strongly convex with respect to some (static) norm $\|\cdot\|$ on $\mathbb{R}^d$. On the other hand, we are introducing an important twist: $\mathcal{R}$ will

be a *self-concordant barrier function* over $\mathcal{X}$.[3] In this context, we first extend (in Appendix B) the so-called RVU bound established in [Syrgkanis et al., 2015] under self-concordant regularization. More precisely, we assume that $\mathcal{X}$ has nonempty interior $\mathrm{int}(\mathcal{X})$. Further, for $\boldsymbol{u} \in \mathbb{R}^d$ the primal local norm with respect to $\boldsymbol{x} \in \mathrm{int}(\mathcal{X})$ is defined as $\|\boldsymbol{u}\|_{\boldsymbol{x}} := \sqrt{\boldsymbol{u}^\top \nabla^2 \mathcal{R}(\boldsymbol{x}) \boldsymbol{u}}$, while the dual norm is defined as $\|\boldsymbol{u}\|_{*,\boldsymbol{x}} := \sqrt{\boldsymbol{u}^\top (\nabla^2 \mathcal{R}(\boldsymbol{x}))^{-1} \boldsymbol{u}}$, assuming that $\mathcal{R}$ nondegenerate—in the sense that its Hessian is positive definite. Finally, for the purpose of the analysis, we let $\boldsymbol{g}^{(t)}$ denote the *be the leader* sequence (see (BTL) in Appendix B); no attempt was made to optimize universal constants.

Theorem 3.1 (RVU for Self-Concordant Regularizers)**.** *Suppose that $\mathcal{R}$ is a nondegenerate self-concordant function for $\mathrm{int}(\mathcal{X})$. Moreover, let $\eta > 0$ be such that $\eta \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \leq \frac{1}{2}$ and $\eta \|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} \leq \frac{1}{2}$ for all $t \in [\![T]\!]$. Then, the regret $\mathrm{Reg}^T(\boldsymbol{x}^*)$ of* (OFTRL) *with respect to any comparator $\boldsymbol{x}^* \in \mathrm{int}(\mathcal{X})$ under any sequence of utilities $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(T)}$ can be bounded by*

$$\frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} + 2\eta \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|^2_{*,\boldsymbol{x}^{(t)}} - \frac{1}{4\eta} \sum_{t=1}^{T} \left(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|^2_{\boldsymbol{x}^{(t)}} + \|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}} \right).$$

Here, we also used the standard notation $\mathrm{Reg}^T(\boldsymbol{x}^*) := \sum_{t=1}^T \langle \boldsymbol{x}^* - \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle$. Next, we instantiate Theorem 3.1 using the *log-barrier* on the (probability) simplex: $\mathcal{R}(\boldsymbol{x}) = -\sum_{r=1}^d \log(\boldsymbol{x}[r])$. While the probability simplex has empty interior, there is a simple transformation on the relative interior $\mathrm{relint}(\Delta^d)$ that addresses that issue (see Appendix B).

Corollary 3.2 (RVU for Log-Barrier on the Simplex)**.** *Suppose that $\mathcal{R}$ is the log-barrier on the simplex and $\eta \leq \frac{1}{16}$. Then, the regret of* (OFTRL) *under any sequence of utilities $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(T)}$ can be bounded as*

$$\mathrm{Reg}^T(\boldsymbol{x}^*) \leq \frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} + 2\eta \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|^2_{*,\boldsymbol{x}^{(t)}} - \frac{1}{16\eta} \sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}},$$

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

We remark that a similar regret bound for *optimistic mirror descent* [Rakhlin and Sridharan, 2013] under log-barrier regularization was shown by [Wei and Luo, 2018, Theorem 7].

4 Main Result

In this section we sketch the proof of our main result, namely Theorem 1.1, leading to Corollaries 1.2 and 1.3; detailed proofs are deferred to Appendix C. In this context, we first employ the general template of Blum and Mansour [2007] for constructing a no-swap-regret minimizer $\mathfrak{R}_{swap}$ over the simplex. We proceed with a brief overview of their construction (summarized in Algorithm 1). In the sequel, we first perform the analysis from the perspective of a single player, without explicitly indicating so in our notation.

The Algorithm of Blum and Mansour Blum and Mansour [2007] construct a "master" regret minimization algorithm $\mathfrak{R}_{swap}$ by maintaining a separate and independent external regret minimizer $\mathfrak{R}_a$ for every action $a \in \mathcal{A}$. To compute the next strategy, $\mathfrak{R}_{swap}$ first obtains the strategy $\boldsymbol{x}_a^{(t)} \in \Delta(\mathcal{A})$ of $\mathfrak{R}_a$, for every $a \in \mathcal{A}$. Then, a (row) stochastic matrix $\mathbf{Q}^{(t)} \in \mathbb{S}^{|\mathcal{A}|}$ is constructed, so that the row associated with action $a \in \mathcal{A}$ is equal to the distribution $\boldsymbol{x}_a^{(t)}$, while $\mathfrak{R}_{swap}$ outputs as the next strategy $\boldsymbol{x}^{(t)} \in \Delta(\mathcal{A})$ any stationary distribution of $\mathbf{Q}^{(t)}$; that is, $(\mathbf{Q}^{(t)})^\top \boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t)}$. Next, upon observing a utility $\boldsymbol{u}^{(t)} \in \mathbb{R}^{|\mathcal{A}|}$, $\mathfrak{R}_{swap}$ forwards to each individual regret minimizer $\mathfrak{R}_a$ the utility $\boldsymbol{u}_a^{(t)} := \boldsymbol{u}^{(t)} \boldsymbol{x}^{(t)}[a] \in \mathbb{R}^{|\mathcal{A}|}$. This construction is summarized in Algorithm 1.

[3] To keep the exposition reasonably self-contained, we give an overview of self-concordant barriers in Appendix A.

Algorithm 1: Blum and Mansour [2007]

Input: A set of external regret minimizers $\{\mathfrak{R}_a\}_{a\in\mathcal{A}}$, each for the simplex $\Delta(\mathcal{A})$

```
function NextStrategy()
    Q^(t) ← 0 ∈ R^(|A|×|A|)
    for a ∈ A do
        Q^(t)[a, ·] ← R_a.NextStrategy()
    x^(t) ← StationaryDistribution(Q^(t))
    return x^(t)

function ObserveUtility(u^(t))
    for a ∈ A do
        R_a.ObserveUtility(x^(t)[a] u^(t))
```

Blum and Mansour [2007] showed that this algorithm guarantees *no-swap-regret* as long as each individual regret minimizer has sublinear *external regret*; this is formalized in the theorem below.

Theorem 4.1 (From External to Swap Regret [Blum and Mansour, 2007])**.** *Let* $\mathrm{SwapReg}^T$ *be the swap regret of* $\mathfrak{R}_{swap}$ *and* Reg_a^T *be the external regret of* $\mathfrak{R}_a$*, for each* $a \in \mathcal{A}$*, up to time* $T \in \mathbb{N}$*. Then,*

$$\mathrm{SwapReg}^T = \sum_{a\in\mathcal{A}} \mathrm{Reg}_a^T\,.$$

In this context, we will instantiate each individual regret minimizer $\mathfrak{R}_a$ with (OFTRL) under log-barrier regularization—and the same learning rate $\eta > 0$. We will refer to the resulting algorithm as `BM-OFTRL-LogBar`. A central ingredient in our proof of Theorem 1.1 is to establish that the resulting no-swap-regret algorithm $\mathfrak{R}_{swap}$ will enjoy an `RVU` bound, as stated in Theorem 4.3. To this end, we first apply Corollary 3.2 for each individual regret minimizer $\mathfrak{R}_a$, implying that $\mathrm{SwapReg}^T = \sum_{a\in\mathcal{A}} \mathrm{Reg}_a^T$ (by Theorem 4.1) is upper bounded as

$$\mathrm{SwapReg}^T \le \frac{2m^2 \log T}{\eta} + 2\eta \sum_{a\in\mathcal{A}} \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)}\boldsymbol{x}^{(t)}[a] - \boldsymbol{u}^{(t-1)}\boldsymbol{x}^{(t-1)}[a]\|^2_{*,\boldsymbol{x}_a^{(t)}} - \frac{1}{16\eta} \sum_{a\in\mathcal{A}} \sum_{t=1}^{T} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\|^2_{\boldsymbol{x}_a^{(t-1)}}. \tag{2}$$

The $\log T$ factor derives from the diameter of the log-barrier regularizer (see Theorem A.9), and appears to be unavoidable using our approach. Now the crux in establishing an `RVU` bound for $\mathfrak{R}_{swap}$ is to upper bound the last term in (2) in terms of the "movement" of the stationary distribution. This is exactly where the local norm induced by the log-barrier turns out to be crucial, leading to the following key technical ingredient.

Lemma 4.2. *Suppose that each regret minimizer* $\mathfrak{R}_a$ *employs* (OFTRL) *with log-barrier regularization and* $\eta \le \frac{1}{16}$*. Then, for any* $t \in \mathbb{N}$*,*

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2 \le 64|\mathcal{A}| \sum_{a\in\mathcal{A}} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\|^2_{\boldsymbol{x}_a^{(t-1)}}.$$

Intuitively, this lemma ensures that the "movement" of the stationary distribution is smooth in terms of the "movement" of each row of the transition matrix $\mathbf{Q}^{(t)}$. To show this, we use the Markov chain tree theorem (Theorem C.3), which provides a closed-form combinatorial formula for the stationary distribution of an ergodic Markov chain, along with the fact that the log-barrier regularizer guarantees "multiplicative stability" of the iterates (Corollary C.1). While similar in spirit results have been documented in the literature for dynamics akin to MWU [Candogan et al., 2013, Chen and Peng, 2020], our proof of Lemma 4.2 crucially hinges on the local norm induced by the log-barrier regularizer. Thus, we are now ready to derive an `RVU` bound for swap regret.

Theorem 4.3 (`RVU` Bound for Swap Regret)**.** *Suppose that each* $\mathfrak{R}_a$ *employs* (OFTRL) *with log-barrier regularization and* $\eta \le \frac{1}{128\sqrt{m}}$*. Then, for* $T \ge 2$*, the swap regret of* $\mathfrak{R}_{swap}$ *is bounded*

as

$$\mathrm{SwapReg}^T \leq \frac{2m^2 \log T}{\eta} + 4\eta \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|_\infty^2 - \frac{1}{2048 m\eta} \sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2.$$

This theorem follows directly from (2) and Lemma 4.2. So far we have focused on bounding the swap regret of each player when faced against arbitrary utilities. Next, we use Theorem 4.3 to establish a new fundamental property when all players employ the dynamics. Our proof crucially relies on the seemingly insignificant fact that $\mathrm{SwapReg}_i^T \geq 0$ (recall Observation 2.1).

Theorem 4.4 (Log-Bounded Second-Order Path Lengths)**.** *Suppose that each player $i \in [\![n]\!]$ employs* `BM-OFTRL-LogBar` *with $\eta = \frac{1}{128(n-1)\max_{j\in[\![n]\!]}\{\sqrt{m_j}\}}$. Then, for $T \geq 2$,*

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \leq 8192 \max_{i\in[\![n]\!]}\{m_i\} \sum_{i=1}^{n} m_i^2 \log T.$$

Proof. Consider any player $i \in [\![n]\!]$. Given that $|u_i(\boldsymbol{a})| \leq 1$, for any $\boldsymbol{a} \in \mathcal{A}$ (by the normalization assumption), we have that for any $t \in [\![T]\!]$,

$$\|\boldsymbol{u}_i^{(t)} - \boldsymbol{u}_i^{(t-1)}\|_\infty \leq \sum_{\boldsymbol{a}_{-i}\in\mathcal{A}_{-i}} \left| \prod_{j\neq i} \boldsymbol{x}_j^{(t)}[a_j] - \prod_{j\neq i} \boldsymbol{x}_j^{(t-1)}[a_j] \right| \leq \sum_{j\neq i} \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1,$$

where we used that the total variation distance between two product distributions is bounded by the sum of the total variations of each individual marginal distribution [Hoeffding and Wolfowitz, 1958]. Thus,

$$\left(\|\boldsymbol{u}_i^{(t)} - \boldsymbol{u}_i^{(t-1)}\|_\infty\right)^2 \leq \left(\sum_{j\neq i} \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1\right)^2 \leq (n-1) \sum_{j\neq i} \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1^2.$$

As a result, using Theorem 4.3 we conclude that $\sum_{i=1}^n \mathrm{SwapReg}_i^T$ can be upper bounded by

$$\begin{aligned}
&2\log T \sum_{i=1}^{n} \frac{m_i^2}{\eta} + 4\eta(n-1) \sum_{i=1}^{n} \sum_{j\neq i} \sum_{t=1}^{T} \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1^2 - \sum_{i=1}^{n} \frac{1}{2048 m_i \eta} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \\
=& 2\log T \sum_{i=1}^{n} \frac{m_i^2}{\eta} + \sum_{i=1}^{n} \left(4\eta(n-1)^2 - \frac{1}{2048 m_i \eta}\right) \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \\
\leq& 2\log T \sum_{i=1}^{n} \frac{m_i^2}{\eta} - \frac{1}{4096} \sum_{i=1}^{n} \frac{1}{m_i \eta} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2,
\end{aligned}$$

since $\eta = \frac{1}{128(n-1)\max_{j\in[\![n]\!]}\{\sqrt{m_j}\}}$. But, given that $0 \leq \sum_{i=1}^n \mathrm{SwapReg}_i^T$, we conclude that

$$\frac{1}{\max_{i\in[\![n]\!]}\{\sqrt{m_i}\}} \sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \leq 8192 \max_{i\in[\![n]\!]}\{\sqrt{m_i}\} \sum_{i=1}^{n} m_i^2 \log T.$$

□

We are not aware of even $o(T)$ bounds for the second-order path lengths in prior works (using a time-invariant learning rate), except in very restricted classes of games such as zero-sum and potential games [Anagnostides et al., 2022c]. An example of the implication of Theorem 4.4 in a variant of *Shapley's game* [Shapley, 1964, Daskalakis et al., 2010] is illustrated in Figure 1. Although the dynamics appear to cycle, and the Nash gap—the maximum of the best response gaps—is always large, the players are changing their (mixed) strategies with gradually diminishing speed; further discussion and experiments are included in Appendix D.

As an immediate consequence, combining Theorem 4.4 with Theorem 4.3 implies near-optimal individual swap regret.

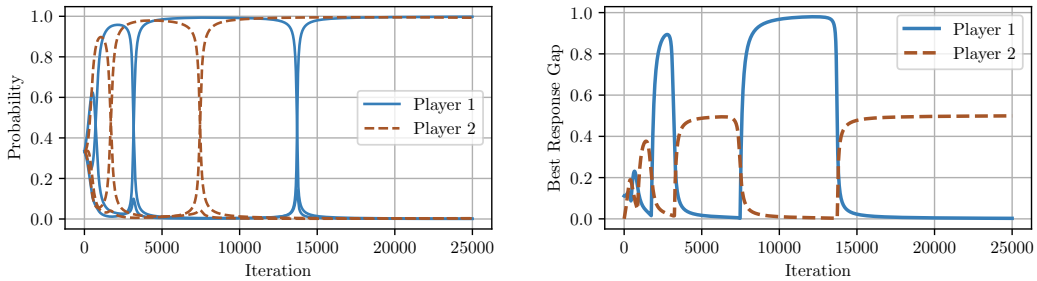

Figure 1: The trajectories of the BM-OFTRL-LogBar algorithm.

Corollary 4.5 (Near-Optimal Individual Swap Regret)**.** *Suppose that all players use* BM-OFTRL-LogBar *with* $\eta = \frac{1}{128(n-1)\max_{i\in[\![n]\!]}\{\sqrt{m_i}\}}$. *Then, the individual swap regret* $\mathrm{SwapReg}_i^T$ *up to time* $T \geq 2$ *of each player* $i \in [\![n]\!]$ *can be bounded as*

$$\mathrm{SwapReg}_i^T \leq 256 \max_{j\in[\![n]\!]}\{\sqrt{m_j}\}\left((n-1)m_i^2 + \sum_{j=1}^{n} m_j^2\right)\log T.$$

We point out that our distributed protocol makes the very mild assumption that each player knows an upper bound on the total number of players and the maximum number of actions in order to appropriately tune the learning rate. Further, as is the case with the result in [Daskalakis et al., 2021], the individual regret of each player predicted by Corollary 4.5 grows linearly with the number of players. This can be unsatisfactory in games with a large number of players—*i.e.*, $n \gg 1$. For this reason, in Theorem C.4 we refine and improve the guarantee of Corollary 4.5 in games where the utility of each player depends only on a small number of other players, and each player's actions only affect a small number of others players; no other constraint is imposed on the game. Understanding whether the linear dependence on n is necessary to obtain near-optimal (swap) regret is left as an interesting question for future work.

Finally, we adapt the learning dynamics so that each player enjoys at the same time near-optimal swap regret in the adversarial regime as well.

Corollary 4.6 (Adversarial Robustness)**.** *There exist dynamics such that when all players follow them the individual swap regret of each player grows as in Corollary 4.5. Moreover, when faced against adversarial utilities, such that* $\|\boldsymbol{u}_i^{(t)}\|_\infty \leq 1$ *for all* $t \in [\![T]\!]$, *the algorithm guarantees that*

$$\mathrm{SwapReg}_i^T \leq 256 \max_{j\in[\![n]\!]}\{\sqrt{m_j}\}\left((n-1)m_i^2 + \sum_{j=1}^{n} m_j^2\right)\log T + 2\sqrt{m_i \log m_i T} + 2.$$

Our adaptation is particularly natural: If all players follow the prescribed protocol, Theorem 4.4 implies that the observed utilities of each player i will be such that $\sum_{\tau=1}^{t}\|\boldsymbol{u}_i^{(\tau)} - \boldsymbol{u}_i^{(\tau-1)}\|_\infty = O(\log t)$. So, if at any time the player identifies that the previous condition was violated, it suffices to switch to a no-swap-regret minimizer (such as BM-MWU) tuned to face advarsarial losses—in which case it is crucial to use a vanishing learning rate $\eta = O(1/\sqrt{T})$.

Remark 4.7 (Numerical Precision). As is standard, we assumed that the iterates of (OFTRL) were computed exactly, without taking into account issues relating to numerical precision. To justify this, one can use *Damped Newton's method* in order to determine an ϵ-nearby point to the optimal in $O(\log\log(1/\epsilon))$ iterations [Nemirovski and Todd, 2008]. This would extend all the regret bounds with up to an $O(\epsilon T)$ error. So, with only $O(\log\log T)$ repetitions of Damped Newton's method (per iteration) the error in the regret bounds becomes $O(1)$, and all of our guarantees immediately extend; see [Farina et al., 2022, Appendix A.5] for an analogous extension under approximate iterates.

5 Discussion

Our main contribution in this paper was to establish a fundamental new property characterizing the trajectories of certain uncoupled no-regret learning dynamics, summarized in Theorem 1.1. This property directly guarantees the best known and near-optimal bound of $O(\log T)$ for the swap regret incurred by each player in a general multiplayer game. Investigating further consequences of Theorem 1.1 is an interesting direction for the future. We also believe that our framework could have new implications for learning in games with partial information; *e.g.*, see [Wei and Luo, 2018]. Another interesting avenue is to extend our scope to more general and combinatorial sets beyond the probability simplex, in order to (efficiently) encompass, for example, games in *extensive form*.

Further, our no-swap-regret learning dynamics have external regret trivially bounded according to Corollary 4.5. Consequently, our construction yields no-external-regret learning dynamics with a more favorable dependence on T compared to [Daskalakis et al., 2021] ($\log T$ compared to the $\log^4(T)$ of the latter), but with a worse dependence on the number of actions (polynomial rather than logarithmic). Our method also has higher per-iteration complexity. For these reasons, extending the scope of our framework beyond self-concordant regularization is an important direction for future research. Indeed, we conjecture that OMWU has *bounded* second-order path lengths, a property that would imply the first uncoupled learning dynamics with bounded regret, but establishing that likely requires new insights.

Acknowledgments

We are thankful to the anonymous NeurIPS reviewers for many helpful comments. Ioannis Anagnostides is grateful to Dimitris Achlioptas for helpful discussions. Christian Kroer is supported by the Office of Naval Research Young Investigator Program under grant N00014-22-1-2530. Haipeng Luo is supported by the National Science Foundation under grant IIS-1943607 (part of this work was done while he was visiting the Simons Institute for the Theory of Computing). Tuomas Sandholm is supported by the National Science Foundation under grants IIS-1901403 and CCF-1733556.

References

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54:296–301, 1951.

David Blackwell. An analog of the minmax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

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

Dean Foster and Rakesh Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2011.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Gabriele Farina, Christian Kroer, Noam Brown, and Tuomas Sandholm. Stable-predictive optimistic counterfactual regret minimization. In *International Conference on Machine Learning (ICML)*, 2019.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *CoRR*, abs/2108.06924, 2021.

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022*, pages 736–749. ACM, 2022a.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 2018.

Dylan J. Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 4727–4735, 2016.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Andrea Celli, and Tuomas Sandholm. Faster no-regret learning dynamics for extensive-form correlated and coarse correlated equilibria. In *EC '22: The 23rd ACM Conference on Economics and Computation, 2022*, pages 915–916. ACM, 2022b.

Avrim Blum and Yishay Mansour. From external to internal regret. *J. Mach. Learn. Res.*, 8: 1307–1324, 2007.

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

Jason Milionis, Christos Papadimitriou, Georgios Piliouras, and Kelly Spendlove. Nash, conley, and computation: Impossibility and incompleteness in game dynamics, 2022.

James P. Bailey and Georgios Piliouras. Fast and furious learning in zero-sum games: Vanishing regret with non-vanishing step sizes. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 12977–12987, 2019.

Gilles Stoltz and Gábor Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59 (1-2):125–159, 2005.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 536–581. PMLR, 2022c.

Georgios Piliouras, Ryann Sim, and Stratis Skoulakis. Optimal no-regret learning in general games: Bounded regret with unbounded step-sizes via clairvoyant mwu. *arXiv preprint arXiv:2111.14737*, 2021.

Amy Greenwald and Amir Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Conference on Learning Theory (COLT)*, Washington, D.C., 2003.

Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games Econ. Behav.*, 59(1):187–208, 2007.

Geoffrey J Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, pages 360–367. ACM, 2008.

Robert Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *J. ACM*, 62(5):32:1–32:42, 2015.

Ozan Candogan, Asuman E. Ozdaglar, and Pablo A. Parrilo. Dynamics in near-potential games. *Games Econ. Behav.*, 82:66–90, 2013.

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

Lloyd S Shapley. Some topics in two-person games. In M. Drescher, L. S. Shapley, and A. W. Tucker, editors, *Advances in Game Theory*. Princeton University Press, 1964.

Constantinos Daskalakis, Rafael M. Frongillo, Christos H. Papadimitriou, George Pierrakos, and Gregory Valiant. On learning algorithms for nash equilibria. In *Algorithmic Game Theory - Third International Symposium, SAGT 2010*, volume 6386 of *Lecture Notes in Computer Science*, pages 114–125. Springer, 2010.

Arkadi S Nemirovski and Michael J Todd. Interior-point methods for optimization. *Acta Numerica*, 17:191–234, 2008.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning for general convex games. *CoRR*, abs/2206.08742, 2022.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

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

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

David Avis, Gabriel D. Rosenberg, Rahul Savani, and Bernhard von Stengel. Enumeration of nash equilibria for two-player games. *Economic Theory*, 42(1):9–37, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix D
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]
(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Preliminaries on Self-Concordant Barriers

In this section we provide the necessary background on self-concordant barriers. For a more comprehensive overview on the theory of self-concordant barriers and their role in interior-point methods we refer to the book of Nesterov [2004], the lecture notes of Nemirovski [2004], as well as the survey of Nemirovski and Todd [2008]. We start this section by introducing the central concept of a *self-concordant function*.

A.1 Self-Concordant Functions

Definition A.1 (Self-Concordant Function)**.** Let $Q \subseteq \mathbb{R}^d$ be a nonempty open and convex set. A convex function $f : Q \to \mathbb{R}$ in $\mathcal{C}^3$ is called *self-concordant* on Q if it satisfies the following properties.

(i) (Barrier property) For every sequence $(\boldsymbol{x}_i \in Q)_{i=1}^{\infty}$ converging to a boundary point of Q as $i \to \infty$ it holds that $f(\boldsymbol{x}_i) \to \infty$;

(ii) (Differential inequality of self-concordance) f satisfies the inequality

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

for all $\boldsymbol{x} \in Q$ and $\boldsymbol{u} \in \mathbb{R}^d$.

In (3) we used the notation

$$D^k f(\boldsymbol{x})[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \coloneqq \left.\frac{\partial^k}{\partial s_1 \ldots \partial s_k}\right|_{s_1 = \cdots = s_k = 0} f(\boldsymbol{x} + s_1 \boldsymbol{u}_1 + \cdots + s_k \boldsymbol{u}_k)$$

to denote the k-th-order differential of f at point $\boldsymbol{x}$ along the directions $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$. Self-concordance, in the sense of Definition A.1, basically imposes a Lipschitz-continuity condition on the Hessian of f, but with respect to the *local norm* induced by the Hessian itself [Nemirovski, 2004]. One may allow (3) to hold with a multiplicative factor $M_f \geq 0$ on the right hand side, in which case f is said to be self-concordant with parameter M_f; unless explicitly specified otherwise, it will be assumed that $M_f = 1$. As a concrete example, we point out that the logarithmic barrier for the nonnegative ray, namely the univariate function $f : (0, +\infty) \ni x \mapsto -\log x$, is self-concordant (with parameter $M_f = 1$).

A crucial fact is that self-concordance is preserved under any linear perturbation, as can be verified directly from Definition A.1. We also point out a certain property which will be useful when composing different functions, and is also an immediate consequence of Definition A.1.

Lemma A.2 ([Nemirovski, 2004])**.** *Let f_i be self-concordant on $\operatorname{dom} f_i$, for all $i \in [\![k]\!]$. Then, assuming that $\operatorname{dom} f \coloneqq \cap_{i=1}^k \operatorname{dom} f_i \neq \emptyset$, the function $f(\boldsymbol{x}) \coloneqq \sum_{i=1}^k f_i(\boldsymbol{x})$ is self-concordant.*

A.2 Useful Inequalities

Let f be a self-concordant function. In the sequel we will tacitly assume that f is *nondegenerate*, in the sense that the Hessian $\nabla^2 f(\boldsymbol{x})$ is positive definite, for any $\boldsymbol{x} \in \operatorname{dom} f$. In this context, we define $\|\boldsymbol{u}\|_{f,\boldsymbol{x}} \coloneqq \sqrt{\boldsymbol{u}^\top \nabla^2 f(\boldsymbol{x}) \boldsymbol{u}}$ to be the *(primal) local norm* of direction u induced by f at point $\boldsymbol{x} \in \operatorname{dom} f$. (It is easy to verify that $\|\boldsymbol{u}\|_{f,\boldsymbol{x}}$ indeed satisfies the axioms of a norm.) To lighten our notation, we will oftentimes simply write $\|\boldsymbol{u}\|_{\boldsymbol{x}}$ when the underlying self-concordant function is clear from the context. The following inequality will be used to derive quadratic growth bounds with respect to the minimum of a self-concordant function.

Lemma A.3 ([Nesterov, 2004])**.** *Let f be a self-concordant function. Then, for any $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \operatorname{dom} f$,*

$$f(\widetilde{\boldsymbol{x}}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}} - \boldsymbol{x} \rangle + \omega\left(\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\|_{\boldsymbol{x}}\right),$$

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

It will be convenient to use a quadratic lower bound for $\omega(s)$, as implied by the following simple fact.

Fact A.4. *Let $\omega(s) = s - \log(1 + s)$. Then,*

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

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

Next, let us consider the optimization problem associated with the minimization of a self-concordant function, namely

$$\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \operatorname{dom} f\}, \tag{4}$$

for a self-concordant f. The *Newton Decrement* of f at point $\boldsymbol{x} \in \operatorname{dom} f$ is defined as

$$\lambda(\boldsymbol{x}, f) \coloneqq \|\nabla f(\boldsymbol{x})\|_{*,\boldsymbol{x}} = \sqrt{(\nabla f(\boldsymbol{x}))^\top (\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x})}.$$

The following result guarantees (existence and) uniqueness for the optimization problem (4).

Lemma A.5 ([Nesterov, 2004])**.** *Let f be a self-concordant function such that $\lambda(\boldsymbol{x}, f) < 1$, for some $\boldsymbol{x} \in \operatorname{dom} f$. Then, the optimization problem* (4) *has a unique solution.*

Assuming that $\mathcal{X}$ is a convex and compact set with nonempty interior, we will also use the following important fact.

Lemma A.6 ([Nemirovski, 2004])**.** *Let $f : \operatorname{int}(\mathcal{X}) \to \mathbb{R}$ be a self-concordant function with $\boldsymbol{x}^* \coloneqq \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$, and some $\boldsymbol{x} \in \operatorname{int}(\mathcal{X})$. Then, if $\lambda(\boldsymbol{x}, f) \leq \frac{1}{2}$,*

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\boldsymbol{x}} \leq 2\lambda(\boldsymbol{x}, f);$$
$$\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\boldsymbol{x}^*} \leq 2\lambda(\boldsymbol{x}, f).$$

A.3 Self-Concordant Barriers

Next, we introduce the concept of a *self-concordant barrier*.

Definition A.7 (Self-Concordant Barrier)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex and compact set with nonempty interior $\operatorname{int}(\mathcal{X})$ (domain). A function $f : \operatorname{int}(\mathcal{X}) \to \mathbb{R}$ is called a *θ-self-concordant barrier* for $\mathcal{X}$ if

(i) f is self-concordant on $\operatorname{int}(\mathcal{X})$; and

(ii) for all $\boldsymbol{x} \in \operatorname{int}(\mathcal{X})$ and $\boldsymbol{u} \in \mathbb{R}^d$,

$$|Df(\boldsymbol{x})[\boldsymbol{u}]| \leq \theta^{1/2} \left(D^2 f(\boldsymbol{x})[\boldsymbol{u}, \boldsymbol{u}]\right)^{1/2}. \tag{5}$$

We note that (5) imposes that f is Lipshitz continuous with parameter $\theta^{1/2}$, but with respect to the local Euclidean metric induced by the Hessian. As an example, it is immediate to see that the function $\mathcal{R}(x) \coloneqq -\log x$ is a 1-self-concordant barrier for the nonnegative ray. The following lemma will be useful when composing self-concordant barriers.

Lemma A.8 ([Nesterov, 2004])**.** *Let f_i be a θ_i-self-concordant barrier for the compact and convex domain $\mathcal{X}_i \subseteq \mathbb{R}^d$, for all $i \in [\![k]\!]$. If the set $\mathcal{X} \coloneqq \cap_{i \in [\![k]\!]} \mathcal{X}_i$ has nonempty interior, the function $f(\boldsymbol{x}) \coloneqq \sum_{i=1}^k f_i(\boldsymbol{x})$ is a $\left(\sum_{i=1}^k \theta_i\right)$-self-concordant barrier for $\mathcal{X}$.*

Minkowski Function Finally, we will require the fact that a self-concordant barrier does not grow overly quickly close to the boundary of $\mathcal{X}$. In particular, the growth is only *logarithmic* as a function of the inverse distance from the boundary. To formalize this, let us introduce the *Minkowski function* on $\mathcal{X}$, defined as follows.

$$\pi(\widetilde{\boldsymbol{x}}; \boldsymbol{x}) = \inf\left\{s \geq 0 : \boldsymbol{x} + s^{-1}(\widetilde{\boldsymbol{x}} - \boldsymbol{x}) \in \mathcal{X}\right\}.$$

We remark that $\pi(\widetilde{\boldsymbol{x}}; \boldsymbol{x}) \in [0, 1]$. When $\boldsymbol{x}$ is the "center" of $\mathcal{X}$, $\pi(\widetilde{\boldsymbol{x}}; \boldsymbol{x})$ can be thought of as the distance of $\widetilde{\boldsymbol{x}}$ from the boundary of $\mathcal{X}$. In this context, we will use the following theorem.

Theorem A.9. *For any θ-self-concordant barrier $\mathcal{R}$ on $\mathcal{X}$ and $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \operatorname{int}(\mathcal{X})$,*

$$\mathcal{R}(\widetilde{\boldsymbol{x}}) - \mathcal{R}(\boldsymbol{x}) \leq \theta \log\left(\frac{1}{1 - \pi(\widetilde{\boldsymbol{x}}; \boldsymbol{x})}\right).$$

B RVU Bounds under Self-Concordant Barriers

In this section we establish the RVU property [Syrgkanis et al., 2015] for (OFTRL) when the regularizer is a self-concordant function. The main result of this section is Theorem 3.1, while Corollary 3.2 is an instantiation on the probability simplex.

As usual, for the purpose of our analysis we consider the auxiliary *be the leader (BTL) sequence*, defined for $t \in \mathbb{N} \cup \{0\}$ as follows.

$$\boldsymbol{g}^{(t)} \coloneqq \arg\max_{\boldsymbol{g} \in \mathcal{X}} \left\{ \Psi^{(t)}(\boldsymbol{g}) \coloneqq \eta \left\langle \boldsymbol{g}, \sum_{\tau=1}^{t} \boldsymbol{u}^{(\tau)} \right\rangle - \mathcal{R}(\boldsymbol{g}) \right\}. \tag{BTL}$$

By convention, we have let $\boldsymbol{g}^{(0)} \coloneqq \arg\min_{\boldsymbol{g} \in \mathcal{X}} \mathcal{R}(\boldsymbol{g})$. We also remark that, as long as $\eta \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \leq \frac{1}{2}$ and $\eta \|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} \leq \frac{1}{2}$, for all $t \in [\![T]\!]$, both (BTL) and (OFTRL) are well-posed, as can be verified using Lemma A.5 (see Lemma B.2). For convenience, and without any loss of generality, in the sequel it is assumed that $\mathcal{R}$ is normalized so that $\min_{\boldsymbol{x}} \mathcal{R}(\boldsymbol{x}) = 0$. We are now ready to establish the following theorem.

Theorem B.1. *Suppose that $\mathcal{R}$ is a nondegenerate self-concordant function for $\operatorname{int}(\mathcal{X})$, and let $\eta > 0$ be such that $\eta \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \leq \frac{1}{2}$ and $\eta \|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} \leq \frac{1}{2}$, for all $t \in [\![T]\!]$. Then, the regret of* (OFTRL) *with respect to any $\boldsymbol{x}^* \in \operatorname{int}(\mathcal{X})$ and under any sequence of utilities $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(T)}$ can be bounded as*

$$\begin{aligned}\operatorname{Reg}^T(\boldsymbol{x}^*) \leq \frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} &+ \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}} \\ &- \frac{1}{\eta} \sum_{t=1}^{T} \left(\omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}) + \omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}) \right),\end{aligned}$$

where $\omega(\cdot)$ is defined as in Lemma A.3.

Proof. The proof proceeds similarly to [Syrgkanis et al., 2015, Theorem 19]. The first observation is that

$$\langle \boldsymbol{x}^* - \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle = \langle \boldsymbol{g}^{(t)} - \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)} \rangle + \langle \boldsymbol{g}^{(t)} - \boldsymbol{x}^{(t)}, \boldsymbol{m}^{(t)} \rangle + \langle \boldsymbol{x}^* - \boldsymbol{g}^{(t)}, \boldsymbol{u}^{(t)} \rangle.$$

Given that $\langle \boldsymbol{g}^{(t)} - \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)} \rangle \leq \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}$, by Hölder's inequality, it suffices to prove that for any $T \in \mathbb{N}$ and $\boldsymbol{x}^* \in \operatorname{int}(\mathcal{X})$,

$$\begin{aligned}\sum_{t=1}^{T} (\langle \boldsymbol{g}^{(t)} - \boldsymbol{x}^{(t)}, \boldsymbol{m}^{(t)} \rangle + \langle \boldsymbol{x}^* - \boldsymbol{g}^{(t)}, \boldsymbol{u}^{(t)} \rangle) &\leq \frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} \\ &- \frac{1}{\eta} \sum_{t=1}^{T} \left(\omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}) + \omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}) \right).\end{aligned} \tag{6}$$

We will establish this claim via induction. For convenience, we use as base for the induction the case where $T = 0$, in which case (6) holds trivially since $\mathcal{R}(\boldsymbol{x}^*) \geq 0$ for any $\boldsymbol{x}^* \in \operatorname{int}(\mathcal{X})$.[4] Now for the inductive step, assume that for some $T \in \{0, 1, \dots\}$,

$$\begin{aligned}\sum_{t=1}^{T} (\langle \boldsymbol{g}^{(t)} - \boldsymbol{x}^{(t)}, \boldsymbol{m}^{(t)} \rangle - \langle \boldsymbol{g}^{(t)}, \boldsymbol{u}^{(t)} \rangle) &\leq -\sum_{t=1}^{T} \langle \boldsymbol{x}^*, \boldsymbol{u}^{(t)} \rangle + \frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} \\ &- \frac{1}{\eta} \sum_{t=1}^{T} \left(\omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}) + \omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}) \right),\end{aligned} \tag{7}$$

[4] By convention, it is assumed that a sum over an empty set is 0.

for any $\boldsymbol{x}^* \in \operatorname{int}(\mathcal{X})$. We will prove the claim for $T+1$. Indeed, applying (7) for $\boldsymbol{x}^* = \boldsymbol{g}^{(T)}$ and adding on both sides the term $\langle \boldsymbol{g}^{(T+1)} - \boldsymbol{x}^{(T+1)}, \boldsymbol{m}^{(T+1)} \rangle - \langle \boldsymbol{g}^{(T+1)}, \boldsymbol{u}^{(T+1)} \rangle$ yields that

$$\begin{aligned}
&\sum_{t=1}^{T+1} \left(\langle \boldsymbol{g}^{(t)} - \boldsymbol{x}^{(t)}, \boldsymbol{m}^{(t)} \rangle - \langle \boldsymbol{g}^{(t)}, \boldsymbol{u}^{(t)} \rangle \right) \leq \\
&\quad - \left\langle \boldsymbol{g}^{(T)}, \sum_{t=1}^{T} \boldsymbol{u}^{(t)} \right\rangle + \frac{\mathcal{R}(\boldsymbol{g}^{(T)})}{\eta} + \langle \boldsymbol{g}^{(T+1)} - \boldsymbol{x}^{(T+1)}, \boldsymbol{m}^{(T+1)} \rangle - \langle \boldsymbol{g}^{(T+1)}, \boldsymbol{u}^{(T+1)} \rangle \\
&\quad - \frac{1}{\eta} \sum_{t=1}^{T} \left(\omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}) + \omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}) \right). \qquad (8)
\end{aligned}$$

Now, by the first-order optimality condition of the optimization problem associated with (BTL), we have that $\nabla \Psi^{(T)}(\boldsymbol{g}^{(T)}) = \mathbf{0}$. As a result, Lemma A.3 implies that

$$\begin{aligned}
&-\Psi^{(T)}(\boldsymbol{x}^{(T+1)}) + \Psi^{(T)}(\boldsymbol{g}^{(T)}) \geq \omega(\|\boldsymbol{x}^{(T+1)} - \boldsymbol{g}^{(T)}\|_{\boldsymbol{g}^{(T)}}) \implies \\
&- \left\langle \boldsymbol{x}^{(T+1)}, \sum_{t=1}^{T} \boldsymbol{u}^{(t)} \right\rangle + \frac{\mathcal{R}(\boldsymbol{x}^{(T+1)})}{\eta} + \left\langle \boldsymbol{g}^{(T)}, \sum_{t=1}^{T} \boldsymbol{u}^{(t)} \right\rangle - \frac{\mathcal{R}(\boldsymbol{g}^{(T)})}{\eta} \geq \frac{1}{\eta} \omega(\|\boldsymbol{x}^{(T+1)} - \boldsymbol{g}^{(T)}\|_{\boldsymbol{g}^{(T)}}), \qquad (9)
\end{aligned}$$

where we used the fact that $-\Psi^{(T)}$ is a self-concordant function, which in turn follows directly from the fact that linear perturbations do not affect self-concordance. Thus, plugging (9) to (8) yields that

$$\begin{aligned}
&\sum_{t=1}^{T+1} \left(\langle \boldsymbol{g}^{(t)} - \boldsymbol{x}^{(t)}, \boldsymbol{m}^{(t)} \rangle - \langle \boldsymbol{g}^{(t)}, \boldsymbol{u}^{(t)} \rangle \right) \leq \\
&\quad - \left\langle \boldsymbol{x}^{(T+1)}, \sum_{t=1}^{T} \boldsymbol{u}^{(t)} \right\rangle + \frac{\mathcal{R}(\boldsymbol{x}^{(T+1)})}{\eta} + \langle \boldsymbol{g}^{(T+1)} - \boldsymbol{x}^{(T+1)}, \boldsymbol{m}^{(T+1)} \rangle - \langle \boldsymbol{g}^{(T+1)}, \boldsymbol{u}^{(T+1)} \rangle \\
&\quad - \frac{1}{\eta} \sum_{t=1}^{T} \left(\omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}) + \omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}) \right) - \frac{1}{\eta} \omega(\|\boldsymbol{x}^{(T+1)} - \boldsymbol{g}^{(T)}\|_{\boldsymbol{g}^{(T)}}) \\
&\quad = - \left\langle \boldsymbol{x}^{(T+1)}, \boldsymbol{m}^{(T+1)} + \sum_{t=1}^{T} \boldsymbol{u}^{(t)} \right\rangle + \frac{\mathcal{R}(\boldsymbol{x}^{(T+1)})}{\eta} + \langle \boldsymbol{g}^{(T+1)}, \boldsymbol{m}^{(T+1)} \rangle - \langle \boldsymbol{g}^{(T+1)}, \boldsymbol{u}^{(T+1)} \rangle \\
&\quad - \frac{1}{\eta} \sum_{t=1}^{T} \left(\omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}) + \omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}) \right) - \frac{1}{\eta} \omega(\|\boldsymbol{x}^{(T+1)} - \boldsymbol{g}^{(T)}\|_{\boldsymbol{g}^{(T)}}). \qquad (10)
\end{aligned}$$

Similarly, by the first-order optimality condition of the optimization problem associated with (OFTRL), we have that $\nabla \Phi^{(T+1)}(\boldsymbol{x}^{(T+1)}) = \mathbf{0}$. Thus, by Lemma A.3 it follows that

$$-\Phi^{(T+1)}(\boldsymbol{g}^{(T+1)}) + \Phi^{(T+1)}(\boldsymbol{x}^{(T+1)}) \geq \omega(\|\boldsymbol{x}^{(T+1)} - \boldsymbol{g}^{(T+1)}\|_{\boldsymbol{x}^{(T+1)}}),$$

since $-\Phi^{(T+1)}$ is self-concordant. Plugging this bound to (10) implies that

$$\begin{aligned}
&\sum_{t=1}^{T+1}\left(\langle \boldsymbol{g}^{(t)}-\boldsymbol{x}^{(t)},\boldsymbol{m}^{(t)}\rangle-\langle \boldsymbol{g}^{(t)},\boldsymbol{u}^{(t)}\rangle\right)\leq\\
&\quad-\left\langle \boldsymbol{g}^{(T+1)},\boldsymbol{m}^{(T+1)}+\sum_{t=1}^{T}\boldsymbol{u}^{(t)}\right\rangle+\frac{\mathcal{R}(\boldsymbol{g}^{(T+1)})}{\eta}+\langle \boldsymbol{g}^{(T+1)},\boldsymbol{m}^{(T+1)}\rangle-\langle \boldsymbol{g}^{(T+1)},\boldsymbol{u}^{(T+1)}\rangle\\
&\quad-\frac{1}{\eta}\sum_{t=1}^{T+1}\left(\omega(\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}})+\omega(\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}})\right)\\
&\quad=-\left\langle \boldsymbol{g}^{(T+1)},\sum_{t=1}^{T+1}\boldsymbol{u}^{(t)}\right\rangle+\frac{\mathcal{R}(\boldsymbol{g}^{(T+1)})}{\eta}\\
&\quad-\frac{1}{\eta}\sum_{t=1}^{T+1}\left(\omega(\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}})+\omega(\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}})\right)\\
&\quad\leq-\left\langle \boldsymbol{x}^{*},\sum_{t=1}^{T+1}\boldsymbol{u}^{(t)}\right\rangle+\frac{\mathcal{R}(\boldsymbol{x}^{*})}{\eta}-\frac{1}{\eta}\sum_{t=1}^{T+1}\left(\omega(\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}})+\omega(\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}})\right),
\end{aligned}$$

for any $\boldsymbol{x}^*\in\operatorname{int}(\mathcal{X})$, where the last inequality follows since $\Psi^{(T+1)}(\boldsymbol{g}^{(T+1)})\geq\Psi^{(T+1)}(\boldsymbol{x}^*)$, for any $\boldsymbol{x}^*\in\operatorname{int}(\mathcal{X})$, by definition of $\boldsymbol{g}^{(T+1)}$. This establishes the inductive step, completing the proof of the theorem. □

Next, to cast Theorem B.1 in the form of an RVU bound (in the sense of [Syrgkanis et al., 2015]), we establish the stability of the iterates as formalized below.

Lemma B.2 (Stability)**.** *Let $\eta>0$ be such that $\eta\|\boldsymbol{u}^{(t)}-\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}}\leq\frac{1}{2}$ and $\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}}\leq\frac{1}{2}$, for all $t\in[\![T]\!]$. Then, for any $t\in[\![T]\!]$,*

$$\begin{aligned}
\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}&\leq 2\eta\|\boldsymbol{u}^{(t)}-\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}};\\
\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}&\leq 2\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}}.
\end{aligned}$$

Proof. Fix any $t\in[\![T]\!]$. We observe that $\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}=\|\boldsymbol{x}^{(t)}-\arg\min(-\Psi^{(t)})\|_{\boldsymbol{x}^{(t)}}$, by definition of (BTL). Further, we have that $\Psi^{(t)}(\boldsymbol{x})=\Phi^{(t)}(\boldsymbol{x})+\eta\langle\boldsymbol{x},\boldsymbol{u}^{(t)}-\boldsymbol{m}^{(t)}\rangle$, implying that $\nabla\Psi^{(t)}=\nabla\Phi^{(t)}+\eta(\boldsymbol{u}^{(t)}-\boldsymbol{m}^{(t)})$. By the first-order optimaility condition of the optimization problem associated with (OFTRL), it follows that $\nabla\Phi^{(t)}(\boldsymbol{x}^{(t)})=0$, in turn implying that $\nabla\Psi^{(t)}(\boldsymbol{x}^{(t)})=\eta(\boldsymbol{u}^{(t)}-\boldsymbol{m}^{(t)})$. As a result, we have shown that $\lambda(\boldsymbol{x}^{(t)},-\Psi^{(t)})=\|\nabla\Psi^{(t)}(\boldsymbol{x}^{(t)})\|_{*,\boldsymbol{x}^{(t)}}=\eta\|\boldsymbol{u}^{(t)}-\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}}\leq\frac{1}{2}$, by assumption. Thus, Lemma A.6 implies that

$$\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}=\|\boldsymbol{x}^{(t)}-\arg\min(-\Psi^{(t)})\|_{\boldsymbol{x}^{(t)}}\leq 2\lambda(\boldsymbol{x}^{(t)},-\Psi^{(t)})=2\eta\|\boldsymbol{u}^{(t)}-\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}},$$

concluding the first part of the claim. Similarly, we have that $\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}=\|\boldsymbol{g}^{(t-1)}-\arg\min(-\Phi^{(t)})\|_{\boldsymbol{g}^{(t-1)}}$, by definition of (OFTRL). Further, we observe that $\Phi^{(t)}(\boldsymbol{x})=\Psi^{(t-1)}(\boldsymbol{x})+\eta\langle\boldsymbol{x},\boldsymbol{m}^{(t)}\rangle$, implying that $\nabla\Phi^{(t)}=\nabla\Psi^{(t-1)}+\eta\boldsymbol{m}^{(t)}$. Moreover, by the first-order optimality condition of the optimization problem associated with (BTL), we have that $\nabla\Psi^{(t-1)}(\boldsymbol{g}^{(t-1)})=\boldsymbol{0}$. In turn, this implies that $\nabla\Phi^{(t)}(\boldsymbol{g}^{(t-1)})=\eta\boldsymbol{m}^{(t)}$. As a result, we have shown that $\lambda(\boldsymbol{g}^{(t-1)},-\Phi^{(t)})=\|\nabla\Phi^{(t)}(\boldsymbol{g}^{(t-1)})\|_{*,\boldsymbol{g}^{(t-1)}}=\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}}\leq\frac{1}{2}$, by assumption. Thus, Lemma A.6 implies that

$$\|\boldsymbol{x}^{(t)}-\boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}}=\|\boldsymbol{g}^{(t-1)}-\arg\min(-\Phi^{(t)})\|_{\boldsymbol{g}^{(t-1)}}\leq 2\lambda(\boldsymbol{g}^{(t-1)},-\Phi^{(t)})=2\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}}.$$

□

We are now ready to establish Theorem 3.1, the statement of which is recalled below.

Theorem 3.1 (RVU for Self-Concordant Regularizers)**.** *Suppose that $\mathcal{R}$ is a nondegenerate self-concordant function for $\operatorname{int}(\mathcal{X})$. Moreover, let $\eta > 0$ be such that $\eta\|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \leq \frac{1}{2}$ and $\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} \leq \frac{1}{2}$ for all $t \in [\![T]\!]$. Then, the regret $\operatorname{Reg}^T(\boldsymbol{x}^*)$ of* (OFTRL) *with respect to any comparator $\boldsymbol{x}^* \in \operatorname{int}(\mathcal{X})$ under any sequence of utilities $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(T)}$ can be bounded by*

$$\frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} + 2\eta\sum_{t=1}^{T}\|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|^2_{*,\boldsymbol{x}^{(t)}} - \frac{1}{4\eta}\sum_{t=1}^{T}\left(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|^2_{\boldsymbol{x}^{(t)}} + \|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}}\right).$$

Proof. First, combining Theorem B.1 with the fact that $\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}} \leq 2\eta\|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}}$ (by Lemma B.2) yields that $\operatorname{Reg}^T(\boldsymbol{x}^*)$ is upper bounded by

$$\frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} + 2\eta\sum_{t=1}^{T}\|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|^2_{*,\boldsymbol{x}^{(t)}} - \frac{1}{\eta}\sum_{t=1}^{T}\left(\omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}}) + \omega(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}})\right).$$

Further, it follows from Lemma B.2 that $\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|_{\boldsymbol{x}^{(t)}} \leq 1$ and $\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}} \leq 1$. Thus, Fact A.4 implies that $\operatorname{Reg}^T(\boldsymbol{x}^*)$ is upper bounded by

$$\frac{\mathcal{R}(\boldsymbol{x}^*)}{\eta} + 2\eta\sum_{t=1}^{T}\|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|^2_{*,\boldsymbol{x}^{(t)}} - \frac{1}{4\eta}\sum_{t=1}^{T}\left(\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|^2_{\boldsymbol{x}^{(t)}} + \|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}}\right).$$

□

For our purposes, it will be convenient to cast Theorem 3.1 in the following form, using the additional assumption that the Hessian $\nabla^2\mathcal{R}$ is stable.

Corollary B.3. *Suppose that $\mathcal{R}$ is a nondegenerate self-concordant function for $\operatorname{int}(\mathcal{X})$ such that $\nabla^2\mathcal{R}(\widetilde{\boldsymbol{x}}) \preceq 2\nabla^2\mathcal{R}(\boldsymbol{x})$ for any $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \operatorname{int}(\mathcal{X})$ with $\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\widetilde{\boldsymbol{x}}} \leq \frac{1}{4}$. Moreover, let $\eta > 0$ be such that $\eta\|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \leq \frac{1}{8}$ and $\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} \leq \frac{1}{2}$ for all $t \in [\![T]\!]$. Then, the regret of* (OFTRL) *under any sequence of utilities $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(T)}$ can be bounded as*

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

Proof. First, by Lemma B.2 we know that $\|\boldsymbol{x}^{(t-1)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{x}^{(t-1)}} \leq 2\eta\|\boldsymbol{u}^{(t-1)} - \boldsymbol{m}^{(t-1)}\|_{*,\boldsymbol{x}^{(t-1)}} \leq \frac{1}{4}$, for any $t \in \mathbb{N}$. Thus, by assumption, it follows that $\nabla^2\mathcal{R}(\boldsymbol{x}^{(t-1)}) \preceq 2\nabla^2\mathcal{R}(\boldsymbol{g}^{(t-1)})$, in turn implying that $\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}} \leq 2\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}}$. Further, the triangle inequality for the norm $\|\cdot\|_{\boldsymbol{x}^{(t-1)}}$ implies that

$$\begin{aligned}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}} &\leq 2\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}} + 2\|\boldsymbol{g}^{(t-1)} - \boldsymbol{x}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}} \\ &\leq 4\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}} + 4\|\boldsymbol{x}^{(t-1)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}},\end{aligned}$$

where we used Young's inequality in the first line, and the fact that $\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}} \leq 2\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}}$ in the second line. Thus, summing over all $t \in [\![T]\!]$ yields that

$$\begin{aligned}\sum_{t=1}^{T}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}} &\leq 4\sum_{t=1}^{T}\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}} + 4\sum_{t=1}^{T}\|\boldsymbol{x}^{(t-1)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{x}^{(t-1)}} \\ &\leq 4\sum_{t=1}^{T}\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|^2_{\boldsymbol{g}^{(t-1)}} + 4\sum_{t=1}^{T}\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t)}\|^2_{\boldsymbol{x}^{(t)}},\end{aligned}$$

since $\boldsymbol{x}^{(0)} = \boldsymbol{g}^{(0)}$. Finally, plugging this bound to Theorem 3.1 concludes the proof. □

Corollary B.4 (Stability of the Iterates)**.** *Suppose that $\mathcal{R}$ is a self-concordant function for $\operatorname{int}(\mathcal{X})$ such that $\nabla^2\mathcal{R}(\widetilde{\boldsymbol{x}}) \preceq 2\nabla^2\mathcal{R}(\boldsymbol{x})$ for any $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \operatorname{int}(\mathcal{X})$ with $\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\widetilde{\boldsymbol{x}}} \leq \frac{1}{4}$. Moreover, let $\eta > 0$ be such that $\eta\|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_{*,\boldsymbol{x}^{(t)}} \leq \frac{1}{8}$ and $\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} \leq \frac{1}{2}$ for all $t \in [\![T]\!]$. Then,*

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_{\boldsymbol{x}^{(t-1)}} \leq 4\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} + 2\eta\|\boldsymbol{u}^{(t-1)} - \boldsymbol{m}^{(t-1)}\|_{*,\boldsymbol{x}^{(t-1)}}.$$

Proof. Similarly to the proof of Corollary B.3, we obtain that

$$\begin{aligned}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_{\boldsymbol{x}^{(t-1)}} &\leq \|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{x}^{(t-1)}} + \|\boldsymbol{g}^{(t-1)} - \boldsymbol{x}^{(t-1)}\|_{\boldsymbol{x}^{(t-1)}} \\ &\leq 2\|\boldsymbol{x}^{(t)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{g}^{(t-1)}} + \|\boldsymbol{x}^{(t-1)} - \boldsymbol{g}^{(t-1)}\|_{\boldsymbol{x}^{(t-1)}} \\ &\leq 4\eta\|\boldsymbol{m}^{(t)}\|_{*,\boldsymbol{g}^{(t-1)}} + 2\eta\|\boldsymbol{u}^{(t-1)} - \boldsymbol{m}^{(t-1)}\|_{*,\boldsymbol{x}^{(t-1)}}.\end{aligned}$$

□

B.1 Log-Barrier Regularizer on the Simplex

Next, we instantiate our general RVU bound for the probability simplex. To this end, let us first point out that, leveraging Lemma A.8 (and Lemma A.2), we can construct a self-concordant barrier for any polytope defined by a set of inequalities $\mathbf{A}\boldsymbol{x} \geq \boldsymbol{b}$, for a matrix $\mathbf{A} \in \mathbb{R}^{k\times d}$ and a vector $\boldsymbol{b} \in \mathbb{R}^k$, as pointed out below.

Definition B.5 (Log-Barrier Regularizer for Polytopes)**.** Consider any polytope defined by a set of inequalities $\mathbf{A}\boldsymbol{x} \geq \boldsymbol{b}$, for a matrix $\mathbf{A} \in \mathbb{R}^{k\times d}$ and a vector $\boldsymbol{b} \in \mathbb{R}^k$. The *log-barrier* function $\mathcal{R}$ is defined as

$$\mathcal{R}(\boldsymbol{x}) \coloneqq -\sum_{r=1}^{k} \log(\mathbf{A}[r,:]\boldsymbol{x} - \boldsymbol{b}[r]). \tag{11}$$

Indeed, Lemma A.8 implies that $\mathcal{R}$ is a k-self-concordant barrier as it can be expressed as the sum of k 1-self-concordant barriers. Now let us focus on constructing a self-concordant barrier for the $(d-1)$-dimensional simplex $\Delta^d \coloneqq \left\{\boldsymbol{x} \in \mathbb{R}^d_{\geq 0} : \sum_{r=1}^d \boldsymbol{x}[r] = 1\right\}$. To address the fact that Δ^d has empty interior, we will restrict the problem to the domain $\Delta^\circ \coloneqq \left\{\boldsymbol{x} \in \mathbb{R}^{d-1}_{\geq 0} : \sum_{r=1}^{d-1} \boldsymbol{x}[r] \leq 1\right\}$. For notational convenience, we will also let $\boldsymbol{x}[d] = 1 - \sum_{r=1}^{d-1} \boldsymbol{x}[r]$. Thus, using the general *log-barrier* regularizer for polytopes given in (11), we arrive at the log-barrier regularizer for Δ°:

$$\mathcal{R}(\boldsymbol{x}) \coloneqq -\sum_{r=1}^{d-1} \log(\boldsymbol{x}[r]) - \log\left(1 - \sum_{r=1}^{d-1} \boldsymbol{x}[r]\right). \tag{12}$$

Naturally, $\mathcal{R}$ is a d-self-concordant barrier since it can be expressed as the sum of d 1-self-concordant barriers. It is important to stress that the regularizer given in (12) takes as input a $(d-1)$-dimensional vector. To reconcile this with the fact that the regret minimizer should receive a d-dimensional utility vector $\boldsymbol{u} \in \mathbb{R}^d$, in the sequel we will use a simple transformation of the observed utilities (while preserving the incurred regret). But first, let us also introduce an auxiliary regularizer for the purpose of our analysis; namely,

$$\widetilde{\mathcal{R}}(\boldsymbol{x}) \coloneqq -\sum_{r=1}^{d} \log \boldsymbol{x}[r]. \tag{13}$$

We are going to relate the local norm induced by the log-barrier (12) to that induced by the auxiliary regularizer (13). First, we characterize the primal local norm induced by $\mathcal{R}$ and $\widetilde{\mathcal{R}}$.

Claim B.6. *For any* $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \operatorname{int}(\Delta^\circ)$*,*

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

Proof. Let us first compute the Hessian of $\mathcal{R}$. A direct calculation gives that for $r \in [\![d-1]\!]$,

$$\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{x}[r]^2} = \frac{1}{(\boldsymbol{x}[r])^2} + \frac{1}{\left(1 - \sum_{r=1}^{d-1} \boldsymbol{x}[r]\right)^2} = \frac{1}{(\boldsymbol{x}[r])^2} + \frac{1}{(\boldsymbol{x}[d])^2},$$

where recall that $\boldsymbol{x}[d] = 1 - \sum_{r=1}^{d-1} \boldsymbol{x}[r]$ (by convention). Further, for $r' \neq r \in [\![d-1]\!]$ we have that

$$\frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{x}[r]\partial \boldsymbol{x}[r']} = \frac{\partial^2 \mathcal{R}}{\partial \boldsymbol{x}[r']\partial \boldsymbol{x}[r]} = \frac{1}{(\boldsymbol{x}[d])^2}.$$

Thus, the Hessian of $\mathcal{R}$ reads

$$\nabla^2 \mathcal{R} = \operatorname{diag}\left(\frac{1}{(\boldsymbol{x}[1])^2}, \ldots, \frac{1}{(\boldsymbol{x}[d-1])^2}\right) + \frac{1}{(\boldsymbol{x}[d])^2} \mathbf{1}_{d-1} \mathbf{1}_{d-1}^\top. \tag{14}$$

As a result,

$$\begin{aligned}
\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\mathcal{R},\boldsymbol{x}}^2 &= (\boldsymbol{x} - \widetilde{\boldsymbol{x}})^\top \operatorname{diag}\left(\frac{1}{(\boldsymbol{x}[1])^2}, \ldots, \frac{1}{(\boldsymbol{x}[d-1])^2}\right)(\boldsymbol{x} - \widetilde{\boldsymbol{x}}) + \frac{(\mathbf{1}_{d-1}^\top(\boldsymbol{x} - \widetilde{\boldsymbol{x}}))^2}{(\boldsymbol{x}[d])^2} \\
&= \sum_{r=1}^{d-1} \left(\frac{\boldsymbol{x}[r] - \widetilde{\boldsymbol{x}}[r]}{\boldsymbol{x}[r]}\right)^2 + \left(\frac{\sum_{r=1}^{d-1} \boldsymbol{x}[r] - \sum_{r=1}^{d-1} \widetilde{\boldsymbol{x}}[r]}{\boldsymbol{x}[d]}\right)^2 \\
&= \sum_{r=1}^{d-1} \left(\frac{\boldsymbol{x}[r] - \widetilde{\boldsymbol{x}}[r]}{\boldsymbol{x}[r]}\right)^2 + \left(\frac{\boldsymbol{x}[d] - \widetilde{\boldsymbol{x}}[d]}{\boldsymbol{x}[d]}\right)^2 \\
&= \sum_{r=1}^{d} \left(\frac{\boldsymbol{x}[r] - \widetilde{\boldsymbol{x}}[r]}{\boldsymbol{x}[r]}\right)^2.
\end{aligned}$$

□

Next, we characterize the dual norm induced by the regularizer $\mathcal{R}$. To this end, let us first explain how the regret minimizer over the domain Δ° should operate. Upon observing a utility vector $\boldsymbol{u} \in \mathbb{R}^d$, we construct the vector $\widetilde{\boldsymbol{u}} \in \mathbb{R}^{d-1}$ so that $\widetilde{\boldsymbol{u}}[r] = \boldsymbol{u}[r] - \boldsymbol{u}[d]$, for all $r \in [\![d-1]\!]$. It is easy to see that the regret incurred is preserved through this transformation.

Claim B.7. *For any $\widetilde{\boldsymbol{u}} \in \mathbb{R}^{d-1}$ and $\boldsymbol{x} \in \operatorname{int}(\Delta^\circ)$,*

$$\|\widetilde{\boldsymbol{u}}\|_{*,\mathcal{R},\boldsymbol{x}} = \|\boldsymbol{u} - c^* \mathbf{1}_d\|_{*,\widetilde{\mathcal{R}},\boldsymbol{x}},$$

where c^* is the scalar that minimizes the norm in the right hand side.

Proof. First, using the Sherman–Morrison formula we find that the inverse of the Hessian of $\mathcal{R}$ given in (14) can be expressed as

$$(\nabla^2 \mathcal{R})^{-1} = \operatorname{diag}(\widetilde{\boldsymbol{x}}[1], \ldots, \widetilde{\boldsymbol{x}}[d-1]) - \frac{1}{\sum_{r=1}^d (\boldsymbol{x}[r])^2} \widetilde{\boldsymbol{x}} \widetilde{\boldsymbol{x}}^\top,$$

where $\widetilde{\boldsymbol{x}} := ((\boldsymbol{x}[1])^2, \ldots, (\boldsymbol{x}[d-1])^2)$. Thus, by definition of $\widetilde{\boldsymbol{u}}$ we have that

$$\begin{aligned}
\|\widetilde{\boldsymbol{u}}\|_{*,\mathcal{R},\boldsymbol{x}} &= \sum_{r=1}^{d-1} (\boldsymbol{x}[r])^2 (\boldsymbol{u}[r] - \boldsymbol{u}[d])^2 - \frac{(\sum_{r=1}^{d-1} (\boldsymbol{x}[r])^2 (\boldsymbol{u}[r] - \boldsymbol{u}[d]))^2}{\sum_{r=1}^d (\boldsymbol{x}[r])^2} \\
&= \sum_{r=1}^{d} (\boldsymbol{x}[r])^2 (\boldsymbol{u}[r] - \boldsymbol{u}[d])^2 - \frac{(\sum_{r=1}^{d} (\boldsymbol{x}[r])^2 \boldsymbol{u}[r] - \boldsymbol{u}[d] \sum_{r=1}^{d} (\boldsymbol{x}[r])^2)^2}{\sum_{r=1}^d (\boldsymbol{x}[r])^2} \\
&= \sum_{r=1}^{d} (\boldsymbol{x}[r])^2 (\boldsymbol{u}[r] - \boldsymbol{u}[d])^2 - \frac{(\sum_{r=1}^{d} (\boldsymbol{x}[r])^2 \boldsymbol{u}[r])^2}{\sum_{r=1}^d (\boldsymbol{x}[r])^2} \\
&+ 2\boldsymbol{u}[d] \sum_{r=1}^{d} (\boldsymbol{x}[r])^2 \boldsymbol{u}[r] - (\boldsymbol{u}[d])^2 \sum_{r=1}^{d} (\boldsymbol{x}[r])^2 \\
&= \sum_{r=1}^{d} (\boldsymbol{x}[r] \boldsymbol{u}[r])^2 - \frac{(\sum_{r=1}^{d} (\boldsymbol{x}[r])^2 \boldsymbol{u}[r])^2}{\sum_{r=1}^d (\boldsymbol{x}[r])^2},
\end{aligned} \tag{15}$$

by simple algebraic calculations. Now let us define the scalar c^* as

$$c^* := \frac{\sum_{r=1}^d (\boldsymbol{x}[r])^2 \boldsymbol{u}[r]}{\sum_{r=1}^d (\boldsymbol{x}[r])^2}.$$

Then, continuing from (15),

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

$$= \sum_{r=1}^{r} (\boldsymbol{x}[r])^2 \left(\boldsymbol{u}[r] - c^*\right)^2 = \|\boldsymbol{u} - c^* \mathbf{1}_d\|_{*,\widetilde{\mathcal{R}},\boldsymbol{x}}. \tag{16}$$

But, it is easy to see that c^* is the minimizer of (16). This concludes the proof. □

An analogous argument shows that $\|\widetilde{\boldsymbol{u}}^{(t)} - \widetilde{\boldsymbol{u}}^{(t-1)}\|_{*,\mathcal{R},\boldsymbol{x}} = \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)} - c^* \mathbf{1}_d\|_{*,\widetilde{\mathcal{R}},\boldsymbol{x}} \leq \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|_{*,\widetilde{\mathcal{R}},\boldsymbol{x}}$. Finally, combining Claim B.6 and Claim B.7 with Theorem 3.1 and Corollary B.3 directly leads to the RVU bound of Corollary 3.2.

C Omitted Proofs from Section 4

In this section we provide the omitted proofs from Section 4. We start with the proof of Lemma 4.2. To this end, we first apply Corollary 3.2 for each individual regret minimizer $\mathfrak{R}_a$, leading to the following guarantee for $\eta \leq \frac{1}{16}$.

$$\mathrm{Reg}_a^T(\boldsymbol{x}_a^*) \leq \frac{\mathcal{R}(\boldsymbol{x}_a^*)}{\eta} + 2\eta \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} \boldsymbol{x}^{(t)}[a] - \boldsymbol{u}^{(t-1)} \boldsymbol{x}^{(t-1)}[a]\|_{*,\boldsymbol{x}_a^{(t)}}^2 - \frac{1}{16\eta} \sum_{t=1}^{T} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\|_{\boldsymbol{x}_a^{(t-1)}}^2, \tag{17}$$

for any $\boldsymbol{x}_a^* \in \mathrm{relint}(\Delta(\mathcal{A}))$; it is assumed that each regret minimizer $\mathfrak{R}_a$ is employing the same learning rate $\eta > 0$. Next, the triangle inequality along with Young's inequality imply that

$$\|\boldsymbol{u}^{(t)} \boldsymbol{x}^{(t)}[a] - \boldsymbol{u}^{(t-1)} \boldsymbol{x}^{(t-1)}[a]\|_{*,\boldsymbol{x}_a^{(t)}}^2 \leq 2(\boldsymbol{x}^{(t)}[a])^2 \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|_{*,\boldsymbol{x}_a^{(t)}}^2 + 2(\boldsymbol{x}^{(t)}[a] - \boldsymbol{x}^{(t-1)}[a])^2 \|\boldsymbol{u}^{(t-1)}\|_{*,\boldsymbol{x}_a^{(t)}}^2,$$

for any $a \in \mathcal{A}$. Summing this inequality over all $a \in \mathcal{A}$ yields that

$$\sum_{a \in \mathcal{A}} \|\boldsymbol{u}^{(t)} \boldsymbol{x}^{(t)}[a] - \boldsymbol{u}^{(t-1)} \boldsymbol{x}^{(t-1)}[a]\|_{*,\boldsymbol{x}_a^{(t)}}^2 \leq 2\|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|_{\infty}^2 + 2\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2. \tag{18}$$

Next, let us address the diameter term in (17). Let $\boldsymbol{x}_c \coloneqq \arg\min_{\boldsymbol{x}} \mathcal{R}(\boldsymbol{x})$, so that $\mathcal{R}(\boldsymbol{x}_c) = 0$. If $\pi(\boldsymbol{x}_a^*; \boldsymbol{x}_c) \leq 1 - \frac{1}{T}$, then, by Theorem A.9,

$$\mathcal{R}(\boldsymbol{x}_a^*) \leq |\mathcal{A}| \log \left(\frac{1}{1 - \pi(\boldsymbol{x}_a^*; \boldsymbol{x}_c)} \right) \leq m \log T,$$

where we used the notation $m \coloneqq |\mathcal{A}|$. Otherwise, we define $\widetilde{\boldsymbol{x}}_a^* \coloneqq (1 - 1/T)\boldsymbol{x}_a^* + (1/T)\boldsymbol{x}_c$, and we observe that

$$\mathrm{Reg}_a^T(\boldsymbol{x}_a^*) \leq \mathrm{Reg}_a^T(\widetilde{\boldsymbol{x}}_a^*) + \sum_{t=1}^{T} \langle \boldsymbol{x}_a^* - \widetilde{\boldsymbol{x}}_a^*, \boldsymbol{x}^{(t)}[a] \boldsymbol{u}^{(t)} \rangle \leq \mathrm{Reg}_a^T(\widetilde{\boldsymbol{x}}_a^*) + \frac{2}{T} \sum_{t=1}^{T} \boldsymbol{x}^{(t)}[a] \|\boldsymbol{u}^{(t)}\|_{\infty}.$$

Thus, from (17) we conclude that

$$\mathrm{Reg}_a^T \leq \frac{m \log T}{\eta} + \frac{2}{T} \sum_{t=1}^{T} \boldsymbol{x}^{(t)}[a] + 2\eta \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} \boldsymbol{x}^{(t)}[a] - \boldsymbol{u}^{(t-1)} \boldsymbol{x}^{(t-1)}[a]\|_{*,\boldsymbol{x}_a^{(t)}}^2 - \frac{1}{16\eta} \sum_{t=1}^{T} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\|_{\boldsymbol{x}_a^{(t-1)}}^2, \tag{19}$$

since $\|\boldsymbol{u}^{(t)}\|_{\infty} \leq 1$. Next, we will use the fact that the log-barrier regularizer guarantees *multiplicative stability*, in the following formal sense.

Corollary C.1 (Multiplicative Stability)**.** *In the setting of Corollary B.4, suppose that* $\|\boldsymbol{u}^{(t)}\|_\infty, \|\boldsymbol{m}^{(t)}\|_\infty \le 1$ *for all* $t \in [\![T]\!]$*. If* $\eta \le \frac{1}{16}$*, then for* $2 \le t \le T$*,*

$$\sqrt{\sum_{r=1}^{d} \left(1 - \frac{\boldsymbol{x}^{(t)}[r]}{\boldsymbol{x}^{(t-1)}[r]}\right)^2} \le 6\eta \|\boldsymbol{u}^{(t-1)}\|_\infty + 2\eta \|\boldsymbol{u}^{(t-2)}\|_\infty.$$

Proof. The claim follows directly from Corollary B.4 with $\boldsymbol{m}^{(t)} = \boldsymbol{u}^{(t-1)}$, using the fact that $\|\boldsymbol{u}\|_{*,\boldsymbol{x}^{(t)}} \le \|\boldsymbol{u}\|_\infty$. □

Now let

$$\mu_a^{(t)} \coloneqq \max_{a' \in \mathcal{A}} \left| 1 - \frac{\boldsymbol{x}_a^{(t)}[a']}{\boldsymbol{x}_a^{(t-1)}[a']} \right|,$$

for each $a \in \mathcal{A}$. Corollary C.1 implies that

$$\begin{aligned} \mu_a^{(t)} &\le 6\eta \|\boldsymbol{u}^{(t-1)} \boldsymbol{x}^{(t-1)}[a]\|_\infty + 2\eta \|\boldsymbol{u}^{(t-2)} \boldsymbol{x}^{(t-2)}[a]\|_\infty \\ &= 6\eta \boldsymbol{x}^{(t-1)}[a] \|\boldsymbol{u}^{(t-1)}\|_\infty + 2\eta \boldsymbol{x}^{(t-2)}[a] \|\boldsymbol{u}^{(t-2)}\|_\infty. \end{aligned}$$

Thus, summing over all $a \in \mathcal{A}$ yields that

$$\sum_{a \in \mathcal{A}} \mu_a^{(t)} \le 6\eta \sum_{a \in \mathcal{A}} \boldsymbol{x}^{(t-1)}[a] \|\boldsymbol{u}^{(t-1)}\|_\infty + 2\eta \sum_{a \in \mathcal{A}} \boldsymbol{x}^{(t-2)}[a] \|\boldsymbol{u}^{(t-2)}\|_\infty \le 8\eta, \tag{20}$$

for $t \ge 2$, where we used that $\boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(t-2)} \in \Delta(\mathcal{A})$, as well as the normalization assumption $\|\boldsymbol{u}\|_\infty \le 1$; it is also immediate to see that $\sum_{a \in \mathcal{A}} \mu_a^{(1)} \le 8\eta$.

For the proof of Lemma 4.2 we will require the Markov chain tree theorem. In particular, consider an m-node ergodic (*i.e.*, aperiodic and irreducible) Markov chain represented through a row-stochastic matrix $\mathbf{Q}$. The Markov chain tree theorem establishes a closed-form solution for the (unique) stationary distribution $\boldsymbol{\pi}$; that is, the vector $\boldsymbol{\pi} \in \Delta^m$ for which $\boldsymbol{\pi}^\top \mathbf{Q} = \boldsymbol{\pi}^\top$. To this end, we formalize the notion of a directed tree.

Definition C.2 (Directed Tree)**.** A directed graph $\mathcal{T} = (V, E)$ is a *directed tree* rooted at node a if (i) it containts no (directed) cycles; (ii) every node $V \setminus \{a\}$ has exactly one outgoing edge; and (iii) the root node a has no outgoing edges.

We will denote with $\mathbb{T}_a$ the set of all possible directed m-node trees rooted at node a. Finally, before we state the Markov chain tree theorem, we let Σ_a be defined as

$$\Sigma_a = \sum_{\mathcal{T} \in \mathbb{T}_a} \prod_{(u,v) \in E(\mathcal{T})} \mathbf{Q}[u, v]. \tag{21}$$

Theorem C.3 (Markov Chain Tree Theorem; *e.g.*, [Anantharam and Tsoucas, 1989])**.** *The stationary distribution* $\boldsymbol{\pi} \in \Delta^m$ *of an* m*-state ergodic markov chain with row-stochastic transition matrix* $\mathbf{Q}$ *is such that*

$$\boldsymbol{\pi}[a] = \frac{\Sigma_a}{\Sigma},$$

where $\Sigma \coloneqq \sum_a \Sigma_a$*, and each* Σ_a *is defined as in* (21)*.*

We are now ready to prove Lemma 4.2.

Lemma 4.2. *Suppose that each regret minimizer* $\mathfrak{R}_a$ *employs* (OFTRL) *with log-barrier regularization and* $\eta \le \frac{1}{16}$*. Then, for any* $t \in \mathbb{N}$*,*

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2 \le 64|\mathcal{A}| \sum_{a \in \mathcal{A}} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\|_{\boldsymbol{x}_a^{(t-1)}}^2.$$

Proof. Consider any $t \in \mathbb{N}$. From the Markov chain tree theorem (Theorem C.3) we know that

$$\boldsymbol{x}[a] = \frac{\Sigma_a}{\Sigma}, \quad \forall a \in \mathcal{A},$$

where $\Sigma_a := \sum_{\mathcal{T}\in\mathbb{T}_a}\prod_{(u,v)\in E(\mathcal{T})}\mathbf{Q}[u,v]$ and $\Sigma = \sum_{a\in\mathcal{A}}\Sigma_a$. Fix some action $a \in \mathcal{A}$ and a directed tree $\mathcal{T} \in \mathbb{T}_a$ rooted at node a. Then,

$$\prod_{(u,v)\in E(\mathcal{T})}\mathbf{Q}^{(t)}[u,v] = \prod_{(u,v)\in E(\mathcal{T})}\boldsymbol{x}_u^{(t)}[v] \le \prod_{(u,v)\in E(\mathcal{T})}(1+\mu_u^{(t)})\boldsymbol{x}_u^{(t-1)}[v],$$

where we used the fact that

$$\mu_u^{(t)} \ge \frac{\boldsymbol{x}_u^{(t)}[v]}{\boldsymbol{x}_u^{(t-1)}[v]} - 1 \implies \boldsymbol{x}_u^{(t)}[v] \le (1+\mu_u^{(t)})\boldsymbol{x}_u^{(t-1)}[v],$$

Thus,

$$\prod_{(u,v)\in E(\mathcal{T})}\mathbf{Q}^{(t)}[u,v] \le \prod_{u\neq a}(1+\mu_u^{(t)})\prod_{(u,v)\in E(\mathcal{T})}\mathbf{Q}^{(t-1)}[u,v],$$

where we used the fact that $\mathcal{T}$ is a directed tree rooted at a. Thus, summing over all $\mathcal{T} \in \mathbb{T}_a$ yields that

$$\begin{aligned}\Sigma_a^{(t)} = \sum_{\mathcal{T}\in\mathbb{T}_a}\prod_{(u,v)\in E(\mathcal{T})}\mathbf{Q}^{(t)}[u,v] &\le \Sigma_a^{(t-1)}\prod_{a'\in\mathcal{A}}(1+\mu_{a'}^{(t)}) \\ &\le \Sigma_a^{(t-1)}\exp\left\{\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}. \end{aligned} \tag{22}$$

This also implies that

$$\Sigma^{(t)} = \sum_{a\in\mathcal{A}}\Sigma_a^{(t)} \le \exp\left\{\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}\sum_{a\in\mathcal{A}}\Sigma_a^{(t-1)} = \exp\left\{\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}\Sigma^{(t-1)}. \tag{23}$$

Similarly,

$$\prod_{(u,v)\in E(\mathcal{T})}\mathbf{Q}^{(t)}[u,v] = \prod_{(u,v)\in E(\mathcal{T})}\boldsymbol{x}_u^{(t)}[v] \ge \prod_{(u,v)\in E(\mathcal{T})}(1-\mu_u^{(t)})\boldsymbol{x}_u^{(t-1)}[v],$$

where we used the fact that

$$\mu_u^{(t)} \ge 1 - \frac{\boldsymbol{x}_u^{(t)}[v]}{\boldsymbol{x}_u^{(t-1)}[v]} \implies \boldsymbol{x}_u^{(t)}[v] \ge (1-\mu_u^{(t)})\boldsymbol{x}_u^{(t-1)}[v].$$

Thus, summing over all $\mathcal{T} \in \mathbb{T}_a$ implies that

$$\begin{aligned}\Sigma_a^{(t)} = \sum_{\mathcal{T}\in\mathbb{T}_a}\prod_{(u,v)\in E(\mathcal{T})}\mathbf{Q}^{(t)}[u,v] &\ge \Sigma_a^{(t-1)}\prod_{a'\in\mathcal{A}}(1-\mu_{a'}^{(t)}) \\ &\ge \Sigma_a^{(t-1)}\exp\left\{-2\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}, \end{aligned} \tag{24}$$

where we used the inequality $1 - x \ge e^{-2x}$, for all $x \in [0, \frac{1}{2}]$, applicable since (by (20)) $\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)} \le 8\eta \le \frac{1}{2}$ for $\eta \le \frac{1}{16}$. This also implies that

$$\Sigma^{(t)} = \sum_{a\in\mathcal{A}}\Sigma_a^{(t)} \ge \exp\left\{-2\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}\sum_{a\in\mathcal{A}}\Sigma_a^{(t-1)} = \exp\left\{-2\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}\Sigma^{(t-1)}. \tag{25}$$

As a result, from (22) and (25) it follows that for any $a \in \mathcal{A}$,

$$\begin{aligned}\frac{\Sigma_a^{(t)}}{\Sigma^{(t)}} - \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}} \le \frac{\Sigma_a^{(t-1)}\exp\left\{\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}}{\Sigma^{(t-1)}\exp\left\{-2\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\}} - \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}} &= \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}}\left(\exp\left\{3\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right\} - 1\right) \\ &\le \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}}\left(8\sum_{a'\in\mathcal{A}}\mu_{a'}^{(t)}\right), \end{aligned}$$

where we used the inequality $e^x - 1 \leq \frac{8}{3}x$ for all $x \in [0, \frac{3}{2}]$, applicable since $\sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)} \leq \frac{1}{2}$. Similarly, (24) and (23) imply that for any $a \in \mathcal{A}$,

$$\frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}} - \frac{\Sigma_a^{(t)}}{\Sigma^{(t)}} \leq \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}} - \frac{\Sigma_a^{(t-1)} \exp\left\{-2\sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)}\right\}}{\Sigma^{(t-1)} \exp\left\{\sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)}\right\}} = \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}} \left(1 - \exp\left\{-3 \sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)}\right\}\right)$$
$$\leq \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}} \left(3 \sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)}\right).$$

As a result, we have established that

$$\left|\boldsymbol{x}^{(t)}[a] - \boldsymbol{x}^{(t-1)}[a]\right| = \left|\frac{\Sigma_a^{(t)}}{\Sigma^{(t)}} - \frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}}\right| \leq 8\frac{\Sigma_a^{(t-1)}}{\Sigma^{(t-1)}} \sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)} = 8\boldsymbol{x}^{(t-1)}[a] \sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)},$$

in turn implying that

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1 \leq 8 \left(\sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)}\right) \left(\sum_{a \in \mathcal{A}} \boldsymbol{x}^{(t-1)}[a]\right) = 8 \sum_{a' \in \mathcal{A}} \mu_{a'}^{(t)}, \tag{26}$$

since $\boldsymbol{x}^{(t-1)} \in \Delta(\mathcal{A})$. Thus,

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2 \leq 64 \left(\sum_{a \in \mathcal{A}} \mu_a^{(t)}\right)^2 \leq 64|\mathcal{A}| \sum_{a \in \mathcal{A}} \left(\mu_a^{(t)}\right)^2,$$

by Jensen's inequality. Finally,

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

and combining this bound with (26) concludes the proof. □

Theorem 4.3 (RVU Bound for Swap Regret)**.** *Suppose that each $\mathfrak{R}_a$ employs* (OFTRL) *with log-barrier regularization and $\eta \leq \frac{1}{128\sqrt{m}}$. Then, for $T \geq 2$, the swap regret of $\mathfrak{R}_{swap}$ is bounded as*

$$\mathrm{SwapReg}^T \leq \frac{2m^2 \log T}{\eta} + 4\eta \sum_{t=1}^T \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|_\infty^2 - \frac{1}{2048m\eta} \sum_{t=1}^T \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2.$$

Proof. Combining (19), (18), Theorem 4.1, and Lemma 4.2 implies that $\mathrm{SwapReg}_i^T$ is upper bounded by

$$\frac{2m^2 \log T}{\eta} + 4\eta \sum_{t=1}^T \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|_\infty^2 + 4\eta \sum_{t=1}^T \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2 - \frac{1}{1024m\eta} \sum_{t=1}^T \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2.$$

Further, for $\eta \leq \frac{1}{128\sqrt{m}}$ it follows that

$$4\eta\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_2^2 \leq 4\eta\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2 \leq \frac{1}{2048m\eta}\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2.$$

In turn, this implies that

$$\mathrm{SwapReg}_i^T \leq \frac{2m^2 \log T}{\eta} + 4\eta \sum_{t=1}^T \|\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t-1)}\|_\infty^2 - \frac{1}{2048m\eta} \sum_{t=1}^T \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1^2,$$

concluding the proof. □

Corollary 4.5 (Near-Optimal Individual Swap Regret)**.** *Suppose that all players use* `BM-OFTRL-LogBar` *with* $\eta = \frac{1}{128(n-1)\max_{i\in[\![n]\!]}\{\sqrt{m_i}\}}$*. Then, the individual swap regret* $\mathrm{SwapReg}_i^T$ *up to time* $T \geq 2$ *of each player* $i \in [\![n]\!]$ *can be bounded as*

$$\mathrm{SwapReg}_i^T \leq 256 \max_{j\in[\![n]\!]}\{\sqrt{m_j}\}\left((n-1)m_i^2 + \sum_{j=1}^n m_j^2\right)\log T.$$

Proof. By Theorem 4.3 and Theorem 4.4,

$$\begin{aligned}\mathrm{SwapReg}_i^T &\leq \frac{2m_i^2\log T}{\eta} + 4\eta(n-1)\sum_{j\neq i}\sum_{t=1}^T \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1^2 \\ &\leq \frac{2m_i^2\log T}{\eta} + 32768\eta(n-1)\max_{j\in[\![n]\!]}\{m_j\}\sum_{j=1}^n m_j^2\log T \\ &= 256\max_{j\in[\![n]\!]}\{\sqrt{m_j}\}\left((n-1)m_i^2 + \sum_{j=1}^n m_j^2\right)\log T.\end{aligned}$$

□

Corollary 4.6 (Adversarial Robustness)**.** *There exist dynamics such that when all players follow them the individual swap regret of each player grows as in Corollary 4.5. Moreover, when faced against adversarial utilities, such that* $\|\boldsymbol{u}_i^{(t)}\|_\infty \leq 1$ *for all* $t \in [\![T]\!]$*, the algorithm guarantees that*

$$\mathrm{SwapReg}_i^T \leq 256\max_{j\in[\![n]\!]}\{\sqrt{m_j}\}\left((n-1)m_i^2 + \sum_{j=1}^n m_j^2\right)\log T + 2\sqrt{m_i\log m_i T} + 2.$$

Proof. Each player $i \in [\![n]\!]$ initially follows the `BM-OFTRL-LogBar` dynamics with learning rate $\eta = \frac{1}{128(n-1)\max_{j\in[\![n]\!]}\{\sqrt{m_j}\}}$. Next, player i keeps track of the quantity $\sum_{\tau=1}^t \|\boldsymbol{u}_i^{(\tau)} - \boldsymbol{u}_i^{(\tau-1)}\|_\infty^2$. If for all $2 \leq t \leq T$ it holds that

$$\sum_{\tau=1}^t \|\boldsymbol{u}_i^{(\tau)} - \boldsymbol{u}_i^{(\tau-1)}\|_\infty^2 \leq 8192(n-1)\max_{j\in[\![n]\!]}\{m_j\}\sum_{j=1}^n m_j^2\log t, \tag{27}$$

then the swap regret of player $i \in [\![n]\!]$ enjoys the guarantee of Corollary 4.5, as follows directly from Theorem 4.3. In particular, (27) will hold as long as all players follow the prescribed dynamics, by virtue of Theorem 4.4. Otherwise, let $t \geq 2$ be the *first* iteration for which (27) is violated. The overall swap regret accumulated up to time $t-1$ is at most the guarantee of Corollary 4.5, as follows directly from Theorem 4.3, while the swap regret at time t is at most 2 since $\|\boldsymbol{u}_i^{(t)}\|_\infty \leq 1$. Next, the player switches to BM-MWU with learning rate $\eta = \sqrt{\frac{m_i\log m_i}{T}}$. Thereafter, the accumulated swap regret will be bounded by $2\sqrt{m_i\log m_i T}$. This completes the proof. □

Finally, we conclude this section with a refinement for games with a large number of players. In particular, we will assume that the utility of each player only depends on the actions of a small number of other players (Item 1), and that each player's actions only affect the utility of a small number of other players (Item 2). Understanding whether the linear dependence of Corollary 4.5 on n is necessary in general games is left as an interesting open question.

Theorem C.4 (Refinement for Large Games)**.** *Suppose that all players use* `BM-OFTRL-LogBar`*. Furthermore, assume that the utility of player* $i \in [\![n]\!]$ *depends on a subset of players* $\mathcal{N}_i \subseteq [\![n]\!]$*, so that*

1. $|\mathcal{N}_i| \leq c \leq n-1$*; and*
2. $\max_{i\in[\![n]\!]}|\{j \neq i : i \in \mathcal{N}_j\}| \leq c$.

Then, for $\eta = \frac{1}{128c \max_{j \in [\![n]\!]}\{\sqrt{m_j}\}}$,

$$\sum_{i=1}^{n} \mathrm{SwapReg}_i^T \leq 256c \max_{j \in [\![n]\!]}\{\sqrt{m_j}\} \sum_{j=1}^{n} m_j^2 \log T.$$

Moreover, for $\eta = \frac{1}{128\sqrt{cn} \max_{j \in [\![n]\!]}\{\sqrt{m_j}\}} \leq \frac{1}{128c \max_{j \in [\![n]\!]}\{\sqrt{m_j}\}}$,

$$\mathrm{SwapReg}_i^T \leq 256 \max_{j \in [\![n]\!]}\{\sqrt{m_j}\} \left(\sqrt{cn} m_i^2 + \sqrt{\frac{c}{n}} \sum_{j=1}^{n} m_j^2 \right) \log T.$$

In particular, if $m_i = m$ *for all* $i \in [\![n]\!]$,

$$\mathrm{SwapReg}_i^T \leq 512\sqrt{cn} m^{5/2} \log T.$$

Proof. The proof proceeds similarly to the proof of Theorem 4.4. First, we have that

$$\left(\|\boldsymbol{u}_i^{(t)} - \boldsymbol{u}_i^{(t-1)}\|_\infty \right)^2 \leq \left(\sum_{j \in \mathcal{N}_i} \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1 \right)^2 \leq c \sum_{j \neq i} \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1^2,$$

since $|\mathcal{N}_i| \leq c$. Thus, using Theorem 4.3, $\sum_{i=1}^{n} \mathrm{SwapReg}_i^T$ can be upper bounded by

$$\begin{aligned}
&2\frac{\log T}{\eta} \sum_{i=1}^{n} m_i^2 + 4\eta c \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} \sum_{t=1}^{T} \|\boldsymbol{x}_j^{(t)} - \boldsymbol{x}_j^{(t-1)}\|_1^2 - \sum_{i=1}^{n} \frac{1}{2048 m_i \eta} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \\
\leq &2\frac{\log T}{\eta} \sum_{i=1}^{n} m_i^2 + \sum_{i=1}^{n} \left(4\eta c^2 - \frac{1}{2048 m_i \eta} \right) \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \qquad (28) \\
\leq &2\frac{\log T}{\eta} \sum_{i=1}^{n} m_i^2 - \frac{1}{4096\eta} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2, \qquad (29)
\end{aligned}$$

where (28) uses the assumption that $|\{j \neq i : i \in \mathcal{N}_j\}| \leq c$, for any player $i \in [\![n]\!]$, and (29) follows since $\eta \leq \frac{1}{128c\sqrt{m_i}}$ for all $i \in [\![n]\!]$. As a result, for $\eta = \frac{1}{128c \max\{\sqrt{m_j}\}}$,

$$\sum_{i=1}^{n} \mathrm{SwapReg}_i^T \leq 256c \max_{j \in [\![n]\!]}\{\sqrt{m_j}\} \sum_{j=1}^{n} m_j^2 \log T.$$

Furthermore, given that $\sum_{i=1}^{n} \mathrm{SwapReg}_i^T \geq 0$,

$$\frac{1}{\max_{j \in [\![n]\!]}\{m_j\}} \sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \leq \sum_{i=1}^{n} \frac{1}{m_i} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \leq 8192 \sum_{i=1}^{n} m_i^2 \log T.$$

Thus, for $\eta = \frac{1}{128\sqrt{cn} \max_{j \in [\![n]\!]}\{\sqrt{m_j}\}} \leq \frac{1}{128c \max_{j \in [\![n]\!]}\{\sqrt{m_j}\}}$,

$$\mathrm{SwapReg}_i^T \leq 256\sqrt{cn} \max_{j \in [\![n]\!]}\{\sqrt{m_j}\} m_i^2 \log T + 256\sqrt{\frac{c}{n}} \max_{j \in [\![n]\!]}\{\sqrt{m_j}\} \sum_{j=1}^{n} m_j^2 \log T.$$

□

Hence, when c is a small constant this theorem implies an improvement of $\Theta(n)$ for the sum of the players' swap regrets, as well as an $\Theta(\sqrt{n})$ factor for each individual swap regret.

D Experiments

In this section we include additional experiments in order to corroborate some of our theoretical results. First, regarding Figure 1 in the main body, we considered a bimatrix (general-sum) game described with the following payoff matrices.

$$\mathbf{A} = \begin{pmatrix} 0 & 0.5 & 1.5 \\ 1.5 & 0 & 1 \\ 0.5 & 1.5 & 0 \end{pmatrix}; \qquad \mathbf{B} = \begin{pmatrix} 0 & 1.5 & 1 \\ 1 & 0 & 1.5 \\ 1.5 & 1 & 0 \end{pmatrix}. \tag{30}$$

This game is a slight variant of *Shapley's game* [Shapley, 1964], a general-sum two-player game used by Shapley in order to illustrate that fictitious play does not converge to Nash equilibria in general-sum games. Shapley's game is not suited to illustrate the cycling behavior of the dynamics in our case since it has a unique Nash equilibrium, occurring when both players play uniformly at random; as such, (OFTRL) is initialized at the equilibrium. On the other hand, the (unique) equilibrium of the game described in (30) occurs when $\boldsymbol{x}^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\boldsymbol{y}^* = (\frac{1}{4}, \frac{2}{5}, \frac{7}{20})$ [Avis et al., 2010]. As illustrated in Figure 1, `BM-OFTRL-LogBar` does not appear to converge to a Nash equilibrium—at least in a *last-iterate* sense. In contrast, we conjecture that the last iterate of (OFTRL) with log-barrier regularization converges to the set of Nash equilibria in zero-sum games, and this property seems plausible even under the BM construction.

Moreover, we conduct experiments on random 3×3 bimatrix (normal-form) general-sum games. Specifically, each entry of the payoff matrices is an independent random variable drawn from the uniform distribution in $[-1, 1]$. In Figure 2 we illustrate the swap regret of the `BM-OFTRL-LogBar` algorithm with a time-invariant learning rate $\eta = 0.1$.

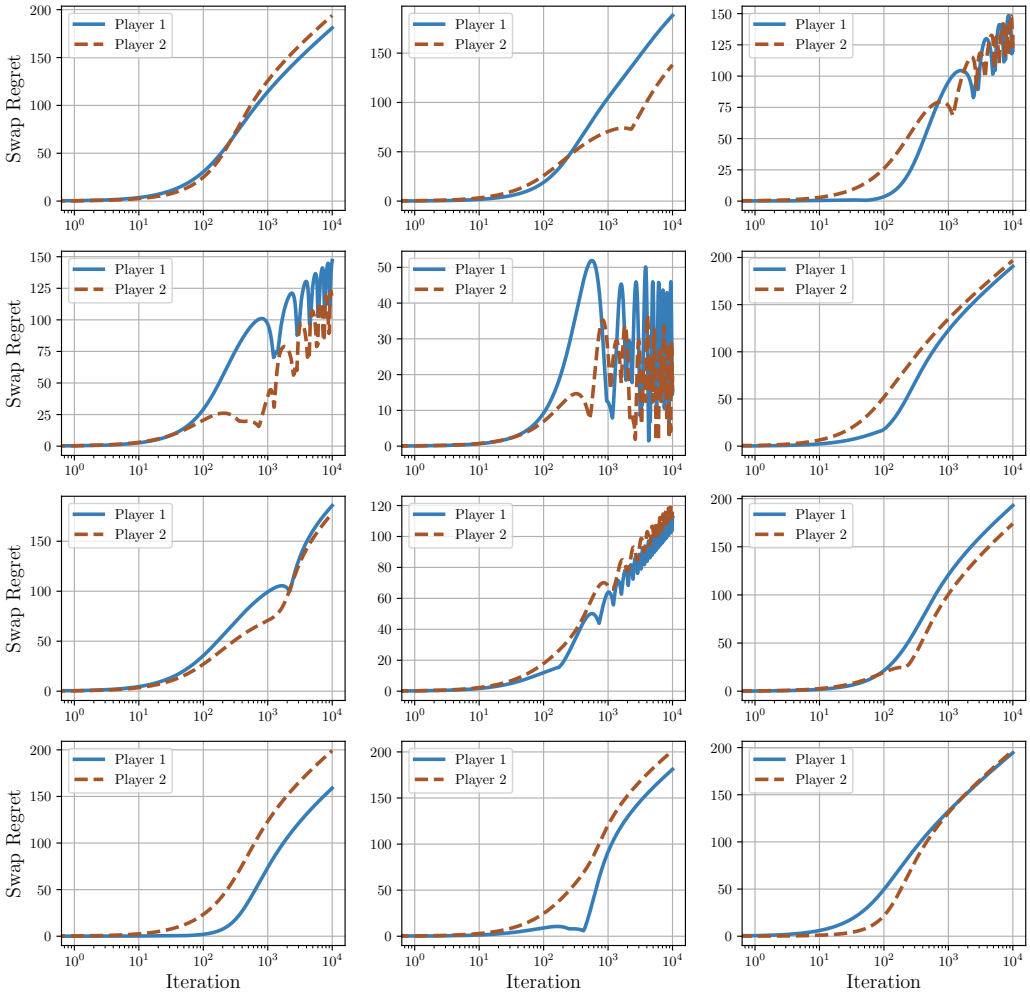

Figure 2: The swap regret experienced by each player for $T = 10^4$ iterations when both players employ `BM-OFTRL-LogBar` with $\eta = 0.1$. Each plot corresponds to a random 3×3 bimatrix game. The x-axis represents the iteration, in logarithmic scale, while the y-axis shows the swap regret experienced by each player at the given iteration. These results corroborate the $O(\log T)$ rates established in Corollary 4.5, showing that our analysis is essentially tight.