Meta-Learning in Games

Keegan Harris*[1], Ioannis Anagnostides*[1], Gabriele Farina[2], Mikhail Khodak[1], Zhiwei Steven Wu[1], and Tuomas Sandholm[1,3,4,5]

[1]Carnegie Mellon University
[2]FAIR, Meta AI
[3]Strategy Robot, Inc.
[4]Optimized Markets, Inc.
[5]Strategic Machine, Inc.
{keeganh,ianagnos,mkhodak,zhiweiw,sandholm}@cs.cmu.edu
gfarina@meta.com

Abstract

In the literature on game-theoretic equilibrium finding, focus has mainly been on solving a single game in isolation. In practice, however, strategic interactions—ranging from routing problems to online advertising auctions—evolve dynamically, thereby leading to many similar games to be solved. To address this gap, we introduce *meta-learning* for equilibrium finding and learning to play games. We establish the first meta-learning guarantees for a variety of fundamental and well-studied classes of games, including two-player zero-sum games, general-sum games, and Stackelberg games. In particular, we obtain rates of convergence to different game-theoretic equilibria that depend on natural notions of similarity between the sequence of games encountered, while at the same time recovering the known single-game guarantees when the sequence of games is arbitrary. Along the way, we prove a number of new results in the single-game regime through a simple and unified framework, which may be of independent interest. Finally, we evaluate our meta-learning algorithms on endgames faced by the poker agent *Libratus* against top human professionals. The experiments show that games with varying stack sizes can be solved significantly faster using our meta-learning techniques than by solving them separately, often by an order of magnitude.

*Equal contribution.

Contents

1 Introduction

Research on game-theoretic equilibrium computation has primarily focused on solving a single game in isolation. In practice, however, there are often many similar games which need to be solved. One use-case is the setting where one wants to find an equilibrium for each of multiple game variations—for example poker games where the players have various sizes of chip stacks. Another use-case is strategic interactions that evolve dynamically: in online advertising auctions, the advertiser's value for different keywords adapts based on current marketing trends [Nekipelov et al., 2015]; routing games—be it Internet routing or physical transportation—reshape depending on the topology and the cost functions of the underlying network [Hoefer et al., 2011]; and resource allocation problems [Johari and Tsitsiklis, 2004] vary based on the values of the goods/services. Successful agents in such complex decentralized environments must effectively learn how to incorporate past experience from previous strategic interactions in order to adapt their behavior to the current and future tasks.

Meta-learning, or *learning-to-learn* [Thrun and Pratt, 1998], is a common formalization for machine learning in dynamic single-agent environments. In the meta-learning framework, a learning agent faces a sequence of tasks, and the goal is to use knowledge gained from previous tasks in order to improve performance on the current task at hand. Despite rapid progress in this line of work, prior results have not been tailored to tackle multiagent settings. This begs the question: *Can players obtain provable performance improvements when meta-learning across a sequence of games?* We answer this question in the affirmative by introducing meta-learning for equilibrium finding and learning to play games, and providing the first performance guarantees in a number of fundamental multiagent settings.

1.1 Overview of Our Results

Our main contribution is to develop a general framework for establishing the first provable guarantees for meta-learning in games, leading to a comprehensive set of results in a variety of well-studied multiagent settings. In particular, our results encompass environments ranging from two-player zero-sum games with general constraint sets (and multiple extensions thereof), to general-sum games and Stackelberg games. See Table 1 for a summary of our results. Our refined guarantees are parameterized based on natural similarity metrics between the sequence of games. For example, in zero-sum games we obtain last-iterate rates that depend on the variance of the Nash equilibria (Theorem 3.2); in potential games based on the deviation of the potential functions (Theorem 3.4); and in Stackelberg games our regret bounds depend on the similarity of the leader's optimal commitment in hindsight (Theorem 3.8). All of these measures are algorithm-independent, and tie naturally to the underlying game-theoretic solution concepts.

Importantly, our algorithms are agnostic to how similar the games are, but are nonetheless specifically designed to adapt to the similarity. Our guarantees apply under a broad class of no-regret learning algorithms, such as *optimistic mirror descent (`OMD`)* [Chiang et al., 2012, Rakhlin and Sridharan, 2013b], with the important twist that each player employs an additional regret minimizer for meta-learning the parameterization of the base-learner; the latter component builds on the meta-learning framework of Khodak et al. [2019]. For example, in zero-sum games we leverage an initialization-dependent *RVU bound* [Syrgkanis et al., 2015] in order to meta-learn the initialization of `OMD` across the sequences of games, leading to per-game convergence rates to Nash equilibria that closely match our refined lower bound (Theorem 3.3). More broadly, in the worst-case—*i.e.*, when the sequence of games is arbitrary—we recover the near-optimal guarantees known for static games, but as the similarity metrics become more favorable we establish significant gains in terms of convergence to different notions of equilibria.

Along the way, we also obtain new insights and results even from a single-game perspective, including convergence rates of `OMD` and the *extra-gradient method* in Hölder continuous variational inequalities [Rakhlin and Sridharan, 2013a], and certain nonconvex-nonconcave problems such

as those considered by [Diakonikolas et al., 2021] and stochastic games. Further, our analysis is considerably simpler than prior techniques and unifies several prior results. Finally, in Section 4 we evaluate our techniques on a series of poker endgames faced by the poker agent *Libratus* [Brown and Sandholm, 2018] against top human professionals. The experiments show that our meta-learning algorithms offer significant gains compared to solving each game in isolation, often by an order of magnitude.

Table 1: A summary of our key theoretical results on meta-learning in games.

Class of games	**Specific problem**	**Key results**	**Location**
Zero-sum games	Bilinear saddle-point problems	Theorems 3.1 and 3.2	Section 3.1
	Hölder continuous VIs	Theorems C.17 and C.34	Appendices C.2 and C.6
	Lower bound	Theorem 3.3	Section 3.1
General-sum games	Potential games	Theorem 3.4	Section 3.2
	(Coarse) Correlated equilibria	Theorems D.7 and D.10	Appendices D.2 and D.3
	Approximately optimal welfare	Theorem 3.6	Section 3.2
Stackelberg games	Security games	Theorem 3.8	Section 3.3

1.2 Related Work

The study of online learning algorithms in games has engendered a prolific area of research, tracing back to the pioneering works of Robinson [1951] and Blackwell [1956]. While traditional analyses rely on black-box guarantees from the no-regret framework [Cesa-Bianchi and Lugosi, 2006], recent works have established exponential improvements over those guarantees when specific learning dynamics are in place (*e.g.*, [Daskalakis et al., 2015, Syrgkanis et al., 2015, Daskalakis et al., 2021]).

However, that line of work posits that the underlying game remains invariant. Yet, there is ample motivation for studying games that gradually change over time, such as online advertising [Nekipelov et al., 2015, Lykouris et al., 2016, Nisan and Noti, 2017] or congestion games [Hoefer et al., 2011, Bertrand et al., 2020, Meigs et al., 2017]. Indeed, a number of prior works study the performance of learning algorithms in time-varying zero-sum games [Zhang et al., 2022b, Fiez et al., 2021b, Duvocelle et al., 2022, Cardoso et al., 2019]; there, it is natural to espouse dynamic notions of regret [Yang et al., 2016, Zhao et al., 2020]. A work closely related to ours is the recent paper by Zhang et al. [2022b], which provides regret bounds in time-varying bilinear saddle-point problems parameterized by the similarity of the payoff matrices and the equilibria of those games. In contrast to our meta-learning setup, they study a more general setting in which the game can change arbitrarily from round-to-round. While our problem can be viewed a special type of a time-varying game in which the boundaries between different games are fixed and known, algorithms designed for generic time-varying games will not perform as well in our setting, as they do not utilize this extra information. As a result, we view these results as complementary to ours. For a more detailed discussion, see Appendix A. Another related direction consists of *warm starting* for solving zero-sum extensive-form games [Brown and Sandholm, 2016], which is typically employed in conjunction with abstraction-based techniques [Brown and Sandholm, 2014, 2015a, Kroer and Sandholm, 2018, Brown and Sandholm, 2015b]. Specifically, there one constructs a sequence of progressively finer abstractions for an underlying game, so that the equilibria of each game can assist the solution of the next one. Perhaps the cardinal difference with our setting is that in abstraction-based applications one is only interested in the performance in the ultimate game.

An emerging paradigm for modeling such considerations is meta-learning, which has gained increasing popularity in the machine learning community in recent years; for a highly incomplete set of pointers, we refer to [Balcan et al., 2015b, Al-Shedivat et al., 2018, Finn et al., 2017, 2019, Balcan et al., 2019, Li et al., 2017, Chen et al., 2022], and references therein. Our work constitutes the natural coalescence of meta-learning with the line of work on (decentralized) online learning in games. Although, as we pointed out earlier, learning in dynamic games has

already received considerable attention, we are the first (to our knowledge) to formulate and address such questions within the meta-learning framework; *c.f.*, see [Kayaalp et al., 2020, 2021, Li et al., 2022]. Finally, our methods may be viewed within the *algorithms with predictions* paradigm [Mitzenmacher and Vassilvitskii, 2020]: we speed up equilibrium computation by learning to predict equilibria across multiple games, with the task-similarity the measure of prediction quality. For further related work, see Appendix A.

2 Our Setup: Meta-Learning in Games

Notation We use boldface symbols to represent vectors and matrices. Subscripts are typically reserved to indicate the player, while superscripts usually correspond to the iteration or the index of the task. We let $\mathbb{N} \coloneqq \{1, 2, \dots, \}$ be the set of natural numbers. For $T \in \mathbb{N}$, we use the shorthand notation $[\![T]\!] \coloneqq \{1, 2, \dots, T\}$. For a nonempty convex and compact set $\mathcal{X}$, we denote by $\Omega_{\mathcal{X}}$ its ℓ_2-diameter: $\Omega_{\mathcal{X}} \coloneqq \max_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$. Finally, to lighten the exposition we use the $O(\cdot)$ notation to suppress factors that depend polynomially on the natural parameters of the problem.

The general setup We consider a setting wherein players interact in a sequence of T repeated games (or *tasks*), for some $\mathbb{N} \ni T \gg 1$. Each task itself consists of $m \in \mathbb{N}$ iterations. Any fixed task t corresponds to a multiplayer game $\mathcal{G}^{(t)}$ between a set $[\![n]\!]$ of players, with $n \geq 2$; it is assumed for simplicity in the exposition that n remains invariant across the games, but some of our results apply more broadly. Each player $k \in [\![n]\!]$ selects a strategy $\boldsymbol{x}_k$ from a convex and compact set of strategies $\mathcal{X}_k \subseteq \mathbb{R}^{d_k}$ with nonempty relative interior. For a given joint strategy profile $\boldsymbol{x} \coloneqq (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in \bigtimes_{k=1}^n \mathcal{X}_k$, there is a multilinear utility function $u_k : \boldsymbol{x} \mapsto \langle \boldsymbol{x}_k, \boldsymbol{u}_k(\boldsymbol{x}_{-k}) \rangle$ for each player k, where $\boldsymbol{x}_{-k} \coloneqq (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_n)$. We will also let $L > 0$ be a Lipschitz parameter of each game, in the sense that for any player $k \in [\![n]\!]$ and any two strategy profiles $\boldsymbol{x}_{-k}, \boldsymbol{x}'_{-k} \in \bigtimes_{k' \neq k} \mathcal{X}_{k'}$,

$$\|\boldsymbol{u}_k(\boldsymbol{x}_{-k}) - \boldsymbol{u}_k(\boldsymbol{x}'_{-k})\|_2 \leq L \|\boldsymbol{x}_{-k} - \boldsymbol{x}'_{-k}\|_2. \tag{1}$$

Here, we use the ℓ_2-norm for convenience in the analysis; (1) can be translated to any equivalent norm. Finally, for a joint strategy profile $\boldsymbol{x} \in \bigtimes_{k=1}^n \mathcal{X}_k$, the *social welfare* is defined as $\mathrm{SW}(\boldsymbol{x}) \coloneqq \sum_{k=1}^n u_k(\boldsymbol{x})$, so that $\mathrm{OPT} \coloneqq \max_{\boldsymbol{x} \in \bigtimes_{k=1}^n \mathcal{X}_k} \mathrm{SW}(\boldsymbol{x})$ denotes the optimal social welfare.

A concrete example encompassed by our setup is that of *extensive-form games*. More broadly, it captures general games with concave utilities [Rosen, 1965, Hsieh et al., 2021].

Online learning in games Learning proceeds in an online fashion as follows. At every iteration $i \in [\![m]\!]$ of some underlying game t, each player $k \in [\![n]\!]$ has to select a strategy $\boldsymbol{x}_k^{(t,i)} \in \mathcal{X}_k$. Then, in the full information setting, the player observes as feedback the utility corresponding to the other players' strategies at iteration i; namely, $\boldsymbol{u}_k^{(t,i)} \coloneqq \boldsymbol{u}_k(\boldsymbol{x}_{-k}^{(t,i)}) \in \mathbb{R}^{d_k}$. For convenience, we will assume that $\|\boldsymbol{u}_k(\boldsymbol{x}_{-k}^{(t,i)})\|_\infty \leq 1$. The canonical measure of performance in online learning is that of *external regret*, comparing the performance of the learner with that of the optimal fixed strategy in hindsight:

Definition 2.1 (Regret)**.** *Fix a player $k \in [\![n]\!]$ and some game $t \in [\![T]\!]$. The (external) regret of player k is defined as*

$$\mathrm{Reg}_k^{(t,m)} \coloneqq \max_{\mathring{\boldsymbol{x}}_k^{(t)} \in \mathcal{X}_k} \left\{ \sum_{i=1}^m \langle \mathring{\boldsymbol{x}}_k^{(t)}, \boldsymbol{u}_k^{(t,i)} \rangle \right\} - \langle \boldsymbol{x}_k^{(t,i)}, \boldsymbol{u}_k^{(t,i)} \rangle.$$

We will let $\mathring{\boldsymbol{x}}_k^{(t)}$ be an optimum-in-hindsight strategy for player k in game t; ties are broken arbitrarily, but according to a fixed rule (*e.g.*, lexicographically). In the meta-learning setting, our goal will be to optimize the average performance—typically measured in terms of convergence to different game-theoretic equilibria—across the sequence of games.

Optimistic mirror descent Suppose that $\mathcal{R}_k : \mathcal{X}_k \to \mathbb{R}$ is a 1-strongly convex regularizer with respect to a norm $\|\cdot\|$. We let $\mathcal{B}_{\mathcal{R}_k}(\boldsymbol{x}_k \,\|\, \boldsymbol{x}'_k) := \mathcal{R}_k(\boldsymbol{x}_k) - \mathcal{R}_k(\boldsymbol{x}'_k) - \langle \nabla \mathcal{R}_k(\boldsymbol{x}'_k), \boldsymbol{x}_k - \boldsymbol{x}'_k \rangle$ denote the *Bregman divergence* induced by $\mathcal{R}_k$, where $\boldsymbol{x}'_k$ is in the relative interior of $\mathcal{X}_k$. *Optimistic mirror descent (`OMD`)* [Chiang et al., 2012, Rakhlin and Sridharan, 2013b] is parameterized by a prediction $\boldsymbol{m}_k^{(t,i)} \in \mathbb{R}^{d_k}$ and a learning rate $\eta > 0$, and is defined at every iteration $i \in \mathbb{N}$ as follows.

$$\boldsymbol{x}_k^{(t,i)} := \arg\max_{\boldsymbol{x}_k \in \mathcal{X}_k} \left\{ \langle \boldsymbol{x}_k, \boldsymbol{m}_k^{(t,i)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}_k}(\boldsymbol{x}_k \,\|\, \hat{\boldsymbol{x}}_k^{(t,i-1)}) \right\},$$
$$\hat{\boldsymbol{x}}_k^{(t,i)} := \arg\max_{\hat{\boldsymbol{x}}_k \in \mathcal{X}_k} \left\{ \langle \hat{\boldsymbol{x}}_k, \boldsymbol{u}_k^{(t,i)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}_k}(\hat{\boldsymbol{x}}_k \,\|\, \hat{\boldsymbol{x}}_k^{(t,i-1)}) \rangle \right\}.$$

Further, $\hat{\boldsymbol{x}}_k^{(1,0)} := \arg\min_{\hat{\boldsymbol{x}}_k \in \mathcal{X}_k} \mathcal{R}_k(\hat{\boldsymbol{x}}_k) =: \boldsymbol{x}_k^{(1,0)}$, and $\boldsymbol{m}_k^{(t,0)} := \boldsymbol{u}_k(\boldsymbol{x}_{-k}^{(t,0)})$. Under Euclidean regularization, $\mathcal{R}_k(\boldsymbol{x}_k) := \frac{1}{2}\|\boldsymbol{x}_k\|_2^2$, we will refer to `OMD` as *optimistic gradient descent (`OGD`)*.

3 Meta-Learning How to Play Games

In this section, we present our main theoretical results: provable guarantees for online and decentralized meta-learning in games. We commence with zero-sum games in Section 3.1, and we then transition to general-sum games (Section 3.2) and Stackelberg (security) games (Section 3.3).

3.1 Zero-Sum Games

We first highlight our results for bilinear saddle-point problems (BSPPs), which take the form $\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{y}$, where $\mathbf{A} \in \mathbb{R}^{d_x \times d_y}$ is the payoff matrix of the game. A canonical application for this setting is on the solution of zero-sum imperfect-information extensive-form games [Romanovskii, 1962, Koller and Megiddo, 1992], as we explore in our experiments (Section 4). Next we describe a number of extensions to gradually more general settings, and we conclude with our lower bound (Theorem 3.3). The proofs from this subsection are included in Appendix C.

We first derive a refined meta-learning convergence guarantee for the average of the players' strategies. Below, we denote by $V_x^2 := \frac{1}{T} \min_{\boldsymbol{x} \in \mathcal{X}} \sum_{t=1}^T \|\mathring{\boldsymbol{x}}^{(t)} - \boldsymbol{x}\|_2^2$ the task similarity metric for player x, written in terms of the optimum-in-hindsight strategies; analogous notation is used for player y.

Theorem 3.1 (Informal; Detailed Version in Corollary C.2)**.** *Suppose that both players employ `OGD` with a suitable (fixed) learning rate and a meta-learning algorithm for the initialization. Then, the game-average duality gap of the players' average strategies is bounded by*

$$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{m} \left(\mathrm{Reg}_x^{(t,m)} + \mathrm{Reg}_y^{(t,m)} \right) \leq \frac{2L}{m} \left(V_x^2 + V_y^2 \right) + \frac{8L(1 + \log T)}{mT} \left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2 \right). \tag{2}$$

Here, the second term in the right-hand side of (2) becomes negligible for a large number of games T, while the first term depends on the task similarity measures. For any sequence of games, Theorem 3.1 nearly matches the lower bound in the single-task setting [Daskalakis et al., 2015], but our guarantee can be significantly better when $V_x^2, V_y^2 \ll 1$. To achieve this, the basic idea is to use—on top of `OGD`—a "meta" regret minimization algorithm that, for each player, learns a sequence of initializations by taking the average of the past optima-in-hindsight, which is equivalent to follow-the-leader (`FTL`) over the regret-upper-bounds of the within-task algorithm; see Algorithm 1 (in Appendix B) for pseudocode of the meta-version of `OGD` we consider. Similar results can be obtained more broadly for `OMD` (*c.f.*, see Appendices D.2 and D.3). We also obtain analogous refined bounds for the *individual* regret of each player (Corollary C.4).

One caveat of Theorem 3.1 is that the underlying task similarity measure could be algorithm-dependent, as the optimum-in-hindsight for each player could depend on the other player's

behavior. To address this, we show that if the meta-learner can initialize using *Nash equilibria (NE)* (recall Definition C.5) from previously seen games, the game-average last-iterate rates gracefully decrease with the similarity of the Nash equilibria of those games. More precisely, if $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y} =: \mathcal{Z}$, we let $V_{\mathrm{NE}}^2 := \frac{1}{T} \max_{\boldsymbol{z}^{(1,\star)},\dots,\boldsymbol{z}^{(T,\star)}} \min_{\boldsymbol{z}\in\mathcal{Z}} \sum_{t=1}^{T} \|\boldsymbol{z}^{(t,\star)} - \boldsymbol{z}\|_2^2$, where $\boldsymbol{z}^{(t,\star)}$ is any Nash equilibrium of the t-th game. As we point out in the sequel, we also obtain results under a more favorable notion of task similarity that does not depend on the worst sequence of NE.

Theorem 3.2 (Informal; Detailed Version in Theorem C.8)**.** *When both players employ* `OGD` *with a suitable parameterization, then*

$$\bar{m} \leq \frac{2V_{NE}^2}{\epsilon^2} + \frac{8(1+\log T)}{T\epsilon^2}\left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2\right)$$

iterations suffice to reach an $O(\epsilon)$-approximate Nash equilibrium of an average game.

Theorem 3.2 recovers the optimal $m^{-1/2}$ rates for `OGD` [Golowich et al., 2020a,b] under an arbitrary sequence of games, but offers substantial gains when the Nash equilibria of the games are close. For example, when they lie within a ball of ℓ_2-diameter $\sqrt{\delta(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2)}$, for some $\delta \in (0, 1]$, Theorem 3.2 improves upon the rate of `OGD` by at least a multiplicative factor of $1/\delta$ as $T \to \infty$. While *generic*—roughly speaking, randomly perturbed—zero-sum (normal-form) games have a unique Nash equilibrium [van Damme, 1987], the worst-case NE similarity metric used in Theorem 3.2 can be loose under multiplicity of equilibria. For that reason, in Appendix C.1.2 we further refine Theorem 3.2 using the most favorable sequence of Nash equilibria; this requires that players know each game after its termination, which is arguably a well-motivated assumption in some applications. We further remark that Theorem 3.2 can be cast in terms of the similarity $V_x^2 + V_y^2$, instead of V_{NE}^2, using the parameterization of Theorem 3.1. Finally, since the base-learner can be viewed as an algorithm with predictions—the number of iterations to compute an approximate NE is smaller if the Euclidean error of a prediction of it (the initialization) is small—Theorem 3.2 can also be viewed as *learning* these predictions [Khodak et al., 2022] by targeting that error measure.

Extensions Moving beyond bilinear saddle-point problems, we extend our results to gradually broader settings. First, in Appendix C.2 we apply our techniques to general variational inequality problems under a Lipschitz continuous operator for which the so-called *MVI property* [Mertikopoulos et al., 2019] holds. Thus, Theorems 3.1 and 3.2 are extended to settings such as smooth convex-concave games and zero-sum polymatrix (multiplayer) games [Cai et al., 2016]. Interestingly, extensions are possible even under the *weak MVI property* [Diakonikolas et al., 2021], which captures certain "structured" nonconvex-nonconcave games. In a similar vein, we also study the challenging setting of Shapley's stochastic games [Shapley, 1953] (Appendix C.5). There, we show that there exists a time-varying—instead of constant—but non-vanishing learning rate schedule for which `OGD` reaches minimax equilibria, thereby leading to similar extensions in the meta-learning setting. Next, we relax the underlying Lipschitz continuity assumption underpinning the previous results by instead imposing only α-Hölder continuity (recall Definition C.32). We show that in such settings `OGD` enjoys a rate of $m^{-\alpha/2}$ (Theorem C.34), which is to our knowledge a new result; in the special case where $\alpha = 1$, we recover the recently established $m^{-1/2}$ rates. Finally, while we have focused on the `OGD` algorithm, our techniques apply to other learning dynamics as well. For example, in Appendix C.7 we show that the extensively studied extra-gradient (`EG`) algorithm [Korpelevich, 1976] can be analyzed in a unifying way with `OGD`, thereby inheriting all of the aforementioned results under `OGD`; this significantly broadens the implications of [Mokhtari et al., 2020], which only applied in certain unconstrained problems. Perhaps surprisingly, although `EG` is *not* a no-regret algorithm, our analysis employs a regret-based framework using a suitable proxy for the regret (see Theorem C.35).

Lower bound We conclude this subsection with a lower bound, showing that our guarantee in Theorem 3.1 is essentially sharp under a broad range of our similarity measures. Our result significantly refines the single-game lower bound of Daskalakis et al. [2015] by constructing an appropriate distribution over sequences of zero-sum games.

Theorem 3.3 (Informal; Precise Version in Theorem C.39)**.** *For any $\epsilon > 0$, there exists a distribution over sequences of T zero-sum games, with a sufficiently large $T = T(\epsilon)$, such that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathrm{Reg}_x^{(t,m)} + \mathrm{Reg}_y^{(t,m)}] \geq \frac{1}{2}\left(V_x^2 + V_y^2\right) - \epsilon = \frac{1}{2}V_{NE}^2 - \epsilon.$$

3.2 General-Sum Games

In this subsection, we switch our attention to general-sum games. Here, unlike zero-sum games, no-regret learning algorithms are instead known to generally converge—in a time-average sense—to *correlated equilibrium* concepts, which are more permissive than the Nash equilibrium. Nevertheless, there are structured classes of general-sum games for which suitable dynamics do reach Nash equilibria; perhaps the most notable example being that of *potential games*. In this context, we first obtain meta-learning guarantees for potential games, parameterized by the similarity of the potential functions. Then, we derive meta-learning algorithms with improved guarantees for convergence to correlated and *coarse* correlated equilibria. Finally, we conclude this subsection with improved guarantees of convergence to near-optimal—in terms of social welfare—equilibria. Proofs from this subsection are included in Appendices B and D.

Potential games A potential game is endowed with the additional property of admitting a potential: a player-independent function that captures the player's benefit from unilaterally deviating from any given strategy profile (Definition D.2). In our meta-learning setting, we posit a sequence of potential games $(\Phi^{(t)})_{1\leq t\leq t}$, each described by its potential function. Unlike our approach in Section 3.1, a central challenge here is that the potential function is in general nonconcave/nonconvex, precluding standard regret minimization approaches. Instead, we find that by initializing at the previous last-iterate the dynamics still manage to adapt based on the similarity $V_\Delta \coloneqq \frac{1}{T}\sum_{t=1}^{T-1}\Delta(\Phi^{(t)}, \Phi^{(t+1)})$, where $\Delta(\Phi, \Phi') \coloneqq \max_{\boldsymbol{x}}(\Phi(\boldsymbol{x}) - \Phi'(\boldsymbol{x}))$, which captures the deviation of the potential functions. This initialization has the additional benefit of being agnostic to the boundaries of different tasks. Unlike our results in Section 3.1, the following guarantee applies even for vanilla (*i.e.*, non-optimistic) projected gradient descent (`GD`).

Theorem 3.4 (Informal; Detailed Version in Corollary D.5)**.** *For an average potential game, `GD` with suitable parameterization requires*

$$O\left(\frac{V_\Delta}{\epsilon^2} + \frac{\Phi_{max}}{\epsilon^2 T}\right)$$

iterations to reach an ϵ-approximate Nash equilibrium, where $\max_{\boldsymbol{x},t}|\Phi^{(t)}(\boldsymbol{x})| \leq \Phi_{max}$.

Theorem 3.4 matches the known rate of `GD` for potential games in the worst case, but offers substantial gains when the games are similar. For example, if $|\Phi^{(t)}(\boldsymbol{x}) - \Phi^{(t-1)}(\boldsymbol{x})| \leq \alpha$, for all $\boldsymbol{x} \in \bigtimes_{k=1}^{n}\mathcal{X}_k$ and $t \geq 2$, then $O(\alpha/\epsilon^2)$ iterations suffice to reach an ϵ-approximate NE on an average game, as $T \to +\infty$. Such a scenario may arise in, *e.g.*, a sequence of routing games if the cost functions for each edge change only slightly between games.

Convergence to correlated equilibria In contrast, for general games the best one can hope for is to obtain improved rates for convergence to correlated or coarse correlated equilibria [Hart

and Mas-Colell, 2000, Blum and Mansour, 2007]. It is important to stress that learning correlated equilibria is fundamentally different than learning Nash equilibria—which are product distributions. For example, for the former any initialization—which is inevitably a product distribution in the case of uncoupled dynamics—could fail to exploit the learning in the previous task (Proposition D.1): unlike Nash equilibria, correlated equilibria (in general) cannot be decomposed for each player, thereby making uncoupled methods unlikely to adapt to the similarity of CE. Instead, our task similarity metrics depend on the optima-in-hindsight for each player. Under this notion of task similarity, we obtain task-average guarantees for CCE by meta-learning the initialization (by running `FTL`) and the learning rate (by running the `EWOO` method of Hazan et al. [2007] over a sequence of regret upper bounds) of *optimistic hedge* [Daskalakis et al., 2021] (Theorem D.7)—`OMD` with entropic regularization. Similarly, to obtain guarantees for CE, we use the *no-swap-regret* construction of Blum and Mansour [2007] in conjuction with the logarithmic barrier [Anagnostides et al., 2022a] (Theorem D.10).

3.2.1 Social Welfare Guarantees

We conclude this subsection with meta-learning guarantees for converging to near-optimal equilibria (Theorem 3.6). Let us first recall the following central definition.

Definition 3.5 (Smooth games [Roughgarden, 2015])**.** *A game $\mathcal{G}$ is (λ, μ)-smooth, with $\lambda, \mu > 0$, if there exists a strategy profile $\boldsymbol{x}^\star \in \bigtimes_{k=1}^n \mathcal{X}_k$ such that for any $\boldsymbol{x} \in \bigtimes_{k=1}^n \mathcal{X}_k$,*

$$\sum_{k=1}^n u_k(\boldsymbol{x}_k^\star, \boldsymbol{x}_{-k}) \geq \lambda \mathrm{OPT} - \mu \mathrm{SW}(\boldsymbol{x}). \tag{3}$$

Smooth games capture a number of important applications, including network congestion games [Awerbuch et al., 2013, Christodoulou and Koutsoupias, 2005] and simultaneous auctions [Christodoulou et al., 2016, Roughgarden et al., 2017] (see Appendix B for additional examples); both of those settings are oftentimes non-static in real-world applications, thereby motivating our meta-learning considerations. In this context, we assume that there is a sequence of smooth games $(\mathcal{G}^{(t)})_{1 \leq t \leq T}$, each of which is $(\lambda^{(t)}, \mu^{(t)})$-smooth (Definition 3.5).

Theorem 3.6 (Informal; Detailed Version in Theorem B.11)**.** *If all players use `OGD` with suitable parameterization in a sequence of T games $(\mathcal{G}^{(t)})_{1 \leq t \leq T}$, each of which is $(\lambda^{(t)}, \mu^{(t)})$-smooth, then*

$$\frac{1}{mT} \sum_{t=1}^T \sum_{i=1}^m \mathrm{SW}(\boldsymbol{x}^{(t,i)}) \geq \frac{1}{T} \sum_{t=1}^T \frac{\lambda^{(t)}}{1 + \mu^{(t)}} \mathrm{OPT}^{(t)} - \frac{2L\sqrt{n-1}}{m} \sum_{k=1}^n V_k^2 - \widetilde{O}\left(\frac{1}{mT}\right), \tag{4}$$

where $\mathrm{OPT}^{(t)}$ is the optimal social welfare attainable at game $\mathcal{G}^{(t)}$ and $\widetilde{O}(\cdot)$ hides logarithmic terms.

The first term in the right-hand side of (4) is the average robust PoA in the sequence of games, while the third term vanishes as $T \to \infty$. The orchestrated learning dynamics reach approximately optimal equilibria much faster when the underlying task similarity is small; without meta-learning one would instead obtain the m^{-1} rate known from the work of Syrgkanis et al. [2015]. Theorem 3.6 is established by first providing a refined guarantee for the *sum of the players regrets* (Theorem B.3), and then translating that guarantee in terms of the social welfare using the smoothness condition for each game (Proposition B.10). Our guarantees are in fact more general, and apply for any suitable linear combination of players' utilities (see Corollary B.12).

3.3 Stackelberg (Security) Games

To conclude our theoretical results, we study meta-learning in repeated Stackelberg games. Following the convention of Balcan et al. [2015a], we present our results in terms of Stackelberg security games, although our results apply to general Stackelberg games as well (see [Balcan et al., 2015a, Section 8] for details on how such results extend).

Stackelberg security games A repeated Stackelberg security game is a sequential interaction between a defender and m attackers. In each round, the defender commits to a mixed strategy over d targets to protect, which induces a *coverage probability vector* $\boldsymbol{x} \in \Delta^d$ over targets. After having observed coverage probability vector, the attacker *best responds* by attacking some target $b(\boldsymbol{x}) \in [\![d]\!]$ in order to maximize their utility in expectation. Finally, the defender's utility is some function of their coverage probability vector $\boldsymbol{x}$ and the target attacked $b(\boldsymbol{x})$.

It is a well-known fact that no-regret learning in repeated Stackelberg games is not possible without any prior knowledge about the sequence of followers [Balcan et al., 2015a, Section 7], so we study the setting in which each attacker belongs to one of k possible *attacker types*. We allow sequence of attackers to be adversarially chosen from the k types, and assume the attacker's type is revealed to the leader after each round. We adapt the methodology of Balcan et al. [2015a] to our setting by meta-learning the initialization and learning rate of the multiplicative weights update (henceforth `MWU`) run over a finite (but exponentially-large) set of *extreme points* $\mathcal{E} \subset \Delta^d$.[1] Each point $\boldsymbol{x} \in \mathcal{E}$ corresponds to a leader mixed strategy, and $\mathcal{E}$ can be constructed in such a way that it will always contain a mixed strategy which is arbitrarily close to the optima-in-hindsight for each task.[2]

Our results are given in terms of guarantees on the task-average *Stackelberg regret*, which measures the difference in utility between the defender's deployed sequence of mixed strategies and the optima-in-hindsight, given that the attacker best responds.

Definition 3.7 (Stackelberg Regret)**.** *Denote attacker $f^{(t,i)}$'s best response to mixed strategy $\boldsymbol{x}$ as $b_{f^{(t,i)}}(\boldsymbol{x})$. The Stackelberg regret of the attacker in a repeated Stackelberg security game t is*

$$\text{StackReg}^{(t,m)}(\mathring{\boldsymbol{x}}^{(t)}) = \sum_{i=1}^{m} \langle \mathring{\boldsymbol{x}}^{(t)}, \boldsymbol{u}^{(t)}(b_{f^{(t,i)}}(\mathring{\boldsymbol{x}}^{(t)})) \rangle - \langle \boldsymbol{x}^{(t,i)}, \boldsymbol{u}^{(t)}(b_{f^{(t,i)}}(\boldsymbol{x}^{(t,i)})) \rangle.$$

In contrast to the standard notion of regret (Definition 2.1), Stackelberg regret takes into account the extra structure in the defender's utility in Stackelberg games; namely that it is a function of the defender's current mixed strategy (through the attacker's best response).

Theorem 3.8 (Informal; Detailed Version in Theorem E.1)**.** *Given a sequence of T repeated Stackelberg security games with d targets, k attacker types, and within-game time-horizon m, running `MWU` over the set of extreme points $\mathcal{E}$ as defined in Balcan et al. [2015a] with suitable initialization and sequence of learning rates achieves task-averaged expected Stackelberg regret*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\text{StackReg}^{(t,m)}] = O(\sqrt{H(\bar{\boldsymbol{y}})m}) + o_T(\mathsf{poly}(m, |\mathcal{E}|)),$$

where the sequence of attackers in each task can be adversarially chosen, the expectation is with respect to the randomness of `MWU`, $\bar{\boldsymbol{y}} := \frac{1}{T}\sum_{t=1}^{T} \mathring{\boldsymbol{y}}^{(t)}$, where $\mathring{\boldsymbol{y}}^{(t)}$ is the optimum-in-hindsight distribution over mixed strategies in $\mathcal{E}$ for game t, $H(\bar{\boldsymbol{y}})$ is the Shannon entropy of $\bar{\boldsymbol{y}}$, and $o_T(1)$ suppresses terms which decay with T.

$H(\bar{\boldsymbol{y}}) \leq \log|\mathcal{E}|$, so in the worst-case our algorithm asymptotically matches the $O(\sqrt{m \log|\mathcal{E}|})$ performance of the algorithm of Balcan et al. [2015a]. Entropy $H(\bar{\boldsymbol{y}})$ is small whenever the same small set of mixed strategies are optimal for the sequence of T Stackelberg games. For example, if in each task the adversary chooses from $s \ll k$ attacker types who are only interested in attacking $u \ll d$ targets (unbeknownst to the meta-learner), $H(\bar{\boldsymbol{y}}) = O(s^2 u \log(su))$. In Stackelberg security games $|\mathcal{E}| = O((2^d + kd^2)^d d^k)$, so $\log|\mathcal{E}| = O(d^2 k \log(dk))$. Finally, the distance between the set of optimal strategies does not matter, as $\bar{\boldsymbol{y}}$ is a categorical distribution over a discrete set of mixed strategies.

[1] This is likely unavoidable, as Li et al. [2016] show computing a Stackelberg strategy is strongly NP-Hard.

[2] For a precise definition of how to construct $\mathcal{E}$, we point the reader to [Balcan et al., 2015a, Section 4].

Game	Board	Pot	Sequences Pl. 1	Sequences Pl. 2	Decision Points Pl. 1	Decision Points Pl. 2	Payoff Matrix num. nonzeros
Endgame A	J♠ K♠ 5♣ Q♠ 7♦	3,700	18,789	19,237	6,710	6,870	14,718,298
Endgame B	4♠ 8♥ 10♣ 9♥ 2♠	500	46,875	47,381	16,304	16,480	62,748,525

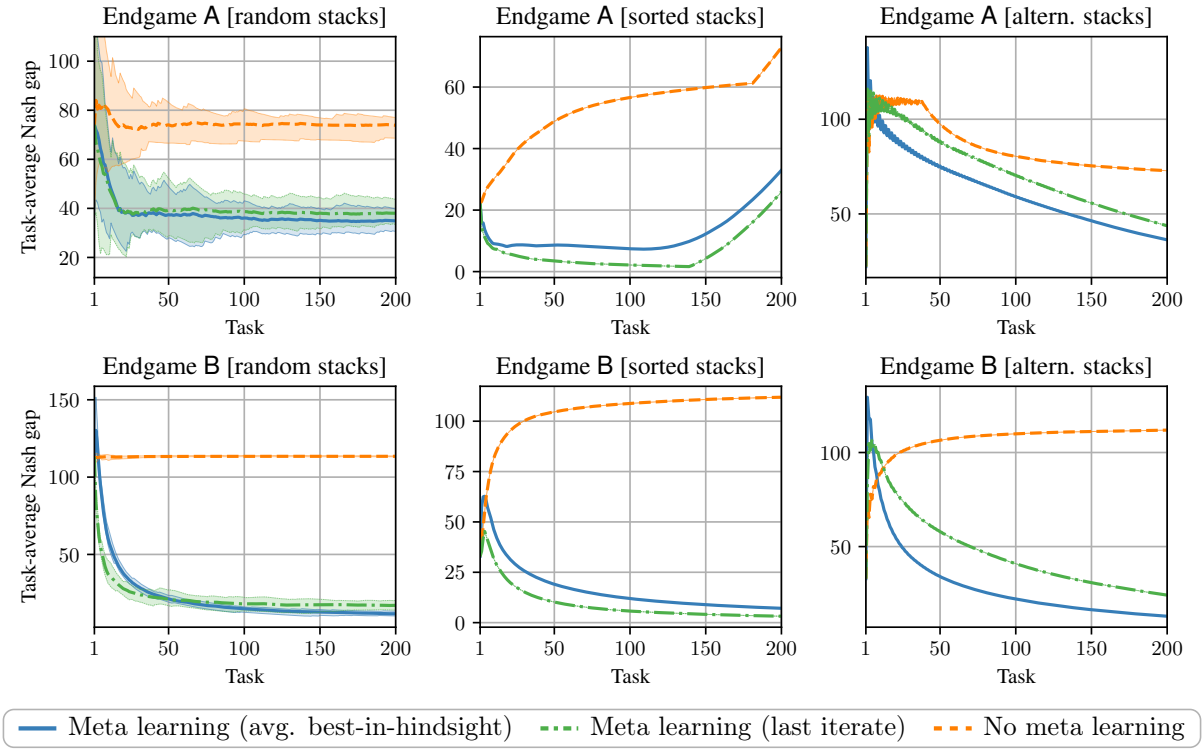

Figure 1: (Top) Parameters of each endgame. (Bottom) The task-averaged NE gap of the players' average strategies across 200 tasks, 2 endgames, and 3 different stack orderings. Both players use `OGD` with $\eta := 0.01$. For the random stacks, we repeated each experiment 10 times with different random seeds. The plots show the mean (thick line) as well as the minimum and maximum values. We see that across all task sequencing setups, meta-learning the initialization (using either technique) leads to up to an order of magnitude better performance compared to vanilla `OGD`. When stacks are sorted, initializing to the last iterate of the previous game obtains the best performance, whereas when stacks are alternated or random, initializing according to Theorem 3.1 performs best.

4 Experiments

In this section, we evaluate our meta-learning techniques in two River endgames that occurred in the *Brains vs AI* competition [Brown and Sandholm, 2018]. We use the two public endgames that were released by the authors,[3] denoted 'Endgame A' and 'Endgame B,' each corresponding to a zero-sum extensive-form game. For each of these endgames, we produced $T := 200$ individual tasks by varying the size of the stacks of each player according to three different *task sequencing setups*:[4]

1. (*random* stacks) In each task we select stack sizes for the players by sampling uniformly at random a multiple of 100 in the range $[1000, 20000]$.
2. (*sorted* stacks) Task $t \in \{1, \dots, 200\}$ corresponds to solving the endgame where the stack sizes are set to the amount $t \times 100$ for each player.
3. (*alternating* stacks) We sequence the stack amounts of the players as follows: in task 1, the

[3] Obtained from `https://github.com/Sandholm-Lab/LibratusEndgames`.

[4] While in the general meta-learning setup it is assumed that the number of tasks is large but per-task data is limited (*i.e.*, $T \gg m$), we found that setting $T := 200$ was already sufficient to see substantial benefits.

stacks are set to 100; in task 2 to 200,000; in task 3 to 200; in task 4 to 199,900; and so on.

For each endgame, we tested the performance when both players (1) employ `OGD` while meta-learning the initialization (Theorem 3.1), (2) employ `OGD` while setting the initialization equal to the last iterate of the previous task (see Remark B.8), and (3) use the vanilla initialization of `OGD`—*i.e.*, the players treat each task as a separate game. For each task, players run $m := 1000$ iterations. The ℓ_2 projection to the *sequence-form polytope* [Romanovskii, 1962, Koller and Megiddo, 1992]—the strategy set of each player in extensive-form games—required for the steps of `OGD` is implemented via an algorithm originally described by Gilpin et al. [2012], and further clarified in [Farina et al., 2022, Appendix B]. We tried different learning rates for the players selected from the set $\{0.1, 0.01, 0.001\}$. Figure 1 illustrates our results for $\eta := 0.01$, while the others are deferred to Appendix F. In the table at the top of Figure 1 we highlight several parameters of the endgames including the board configuration, the dimensions of the players' strategy sets—*i.e.*, the sequences—and the number of nonzero elements in each payoff matrix. Because of the scale of the games, we used the *Kronecker sparsification* algorithm of Farina and Sandholm [2022, Technique A] in order to accelerate the training.

5 Conclusions and Future Research

In this paper, we introduced the study of meta-learning in games. In particular, we considered many of the most central game classes—including zero-sum games, potential games, general-sum multi-player games, and Stackelberg security games—and obtained provable performance guarantees expressed in terms of natural measures of similarity between the games. Experiments on several sequences of poker endgames that were actually played in the *Brains vs AI* competition [Brown and Sandholm, 2018] show that meta-learning the initialization improves performance even by an order of magnitude.

Our results open the door to several exciting directions for future research, including meta-learning in other settings for which single-game results are known, such as general nonconvex-nonconcave min-max problems [Suggala and Netrapalli, 2020], the nonparametric regime [Daskalakis and Golowich, 2022], and partial feedback (such as bandit) models [Wei and Luo, 2018, Hsieh et al., 2022, Balcan et al., 2022, Osadchiy et al., 2022]. Another interesting, yet challenging, avenue for future research would be to consider strategy sets that can vary across tasks.

Acknowledgements

KH is supported by a NDSEG Fellowship. IA and TS are supported by NSF grants IIS-1901403 and CCF-1733556, and the ARO under award W911NF2010081. MK is supported by a Meta Research PhD Fellowship. ZSW is supported in part by the NSF grant FAI-1939606, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Meta Research Award, and a Mozilla Research Grant. The authors would like to thank Nina Balcan for helpful discussions throughout the course of the project. IA is grateful to Ioannis Panageas for insightful discussions regarding Appendix C.5.

References

Jacob D. Abernethy, Kevin A. Lai, Kfir Y. Levy, and Jun-Kun Wang. Faster rates for convex-concave games. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1595–1625. PMLR, 2018.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22:98:1–98:76, 2021.

Maruan Al-Shedivat, Trapit Bansal, Yura Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. *arXiv preprint arXiv:2204.11417*, 2022a.

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, 2022b.

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

Baruch Awerbuch, Yossi Azar, and Amir Epstein. The price of routing unsplittable flow. *SIAM J. Comput.*, 42(1):160–177, 2013.

Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 326–358. PMLR, 2021.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the sixteenth ACM conference on economics and computation*, pages 61–78, 2015a.

Maria-Florina Balcan, Avrim Blum, and Santosh S. Vempala. Efficient representations for lifelong learning and autoencoding. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 191–210. JMLR.org, 2015b.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 424–433. PMLR, 2019.

Maria-Florina Balcan, Keegan Harris, Mikhail Khodak, and Zhiwei Steven Wu. Meta-learning adversarial bandits. *arXiv preprint arXiv:2205.14128*, 2022.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.

Heinz H. Bauschke, Walaa M. Moursi, and Xianfu Wang. Generalized monotone operators and their averaged resolvents. *Math. Program.*, 189:55–74, 2021.

Nathalie Bertrand, Nicolas Markey, Suman Sadhukhan, and Ocan Sankur. Dynamic network congestion games. In *40th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2020*, volume 182 of *LIPIcs*, pages 40:1–40:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

Benjamin E. Birnbaum, Nikhil R. Devanur, and Lin Xiao. Distributed algorithms via gradient descent for fisher markets. In *Proceedings 12th ACM Conference on Electronic Commerce (EC-2011), 2011*, pages 127–136. ACM, 2011.

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

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *Conference on Learning Theory, 2-5 July 2022*, volume 178 of *Proceedings of Machine Learning Research*, pages 1716–1786. PMLR, 2022.

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

Noam Brown and Tuomas Sandholm. Regret transfer and parameter optimization. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014*, pages 594–601. AAAI Press, 2014.

Noam Brown and Tuomas Sandholm. Regret-based pruning in extensive-form games. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 1972–1980, 2015a.

Noam Brown and Tuomas Sandholm. Simultaneous abstraction and equilibrium finding in games. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 489–496. AAAI Press, 2015b.

Noam Brown and Tuomas Sandholm. Strategy-based warm starting for regret minimization in games. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 432–438. AAAI Press, 2016.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, 2019*, pages 1829–1836. AAAI Press, 2019.

Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos H. Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Math. Oper. Res.*, 41(2):648–655, 2016.

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

Adrian Rivera Cardoso, Jacob D. Abernethy, He Wang, and Huan Xu. Competing against nash equilibria in adversarially changing zero-sum games. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 921–930. PMLR, 2019.

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

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2-3):321–352, 2007.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player nash equilibria. *J. ACM*, 56(3):14:1–14:57, 2009.

Xi Chen, Christos H. Papadimitriou, and Binghui Peng. Memory bounds for continual learning. *CoRR*, abs/2204.10830, 2022.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT 2012 - The 25th Annual Conference on Learning Theory, 2012*, volume 23 of *JMLR Proceedings*, pages 6.1–6.20. JMLR.org, 2012.

George Christodoulou and Elias Koutsoupias. The price of anarchy of finite congestion games. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22-24, 2005*, pages 67–73. ACM, 2005.

George Christodoulou, Annamária Kovács, and Michael Schapira. Bayesian combinatorial auctions. *J. ACM*, 63(2):11:1–11:19, 2016.

Patrick L. Combettes and Teemu Pennanen. Proximal methods for cohypomonotone operators. *SIAM J. Control. Optim.*, 43(2):731–742, 2004.

Constantinos Daskalakis and Noah Golowich. Fast rates for nonparametric online learning: from realizability to learning in games. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022*, pages 846–859. ACM, 2022.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *SIAM J. Comput.*, 39(1):195–259, 2009.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. *Games Econ. Behav.*, 92:327–348, 2015.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 27604–27616, 2021.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021*, volume 130 of *Proceedings of Machine Learning Research*, pages 2746–2754. PMLR, 2021.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.

Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, and Dries Vermeulen. Multiagent online learning in time-varying games. *Mathematics of Operations Research*, 2022.

Gabriele Farina and Tuomas Sandholm. Fast payoff matrix sparsification techniques for structured extensive-form games. In *AAAI Conference on Artificial Intelligence*, 2022.

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.

Tanner Fiez, Lillian J. Ratliff, Eric Mazumdar, Evan Faulkner, and Adhyyan Narang. Global convergence to local minmax equilibrium in classes of nonconvex zero-sum games. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 29049–29063, 2021a.

Tanner Fiez, Ryann Sim, Stratis Skoulakis, Georgios Piliouras, and Lillian J. Ratliff. Online learning in periodic zero-sum games. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 10313–10325, 2021b.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017.

Chelsea Finn, Aravind Rajeswaran, Sham M. Kakade, and Sergey Levine. Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 1920–1930. PMLR, 2019.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Yuan Gao, Christian Kroer, and Donald Goldfarb. Increasing iterate averaging for solving saddle-point problems. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 7537–7544. AAAI Press, 2021.

Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 2147–2148. PMLR, 2021.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

Andrew Gilpin, Javier Peña, and Tuomas Sandholm. First-order algorithm with $\mathcal{O}(\ln(1/\epsilon))$ convergence for ϵ-equilibrium in two-person zero-sum games. *Math. Program.*, 133(1-2):279–298, 2012.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman E. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1758–1784. PMLR, 2020b.

Nika Haghtalab, Yanjun Han, Abhishek Shetty, and Kunhe Yang. Oracle-efficient online learning for beyond worst-case adversaries. *CoRR*, abs/2202.08549, 2022.

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

Sergiu Hart and David Schmeidler. Existence of correlated equilibria. *Mathematics of Operations Research*, 14(1):18–25, 1989.

Jason D. Hartline, Vasilis Syrgkanis, and Éva Tardos. No-regret learning in bayesian games. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 3061–3069, 2015.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: regret bounded by variation in costs. *Mach. Learn.*, 80(2-3):165–188, 2010.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *J. Mach. Learn. Res.*, 12: 1287–1311, 2011.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Martin Hoefer, Vahab S. Mirrokni, Heiko Röglin, and Shang-Hua Teng. Competitive routing over time. *Theor. Comput. Sci.*, 412(39):5420–5432, 2011.

Josef Hofbauer and William H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 6936–6946, 2019.

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 2388–2422. PMLR, 2021.

Yu-Guan Hsieh, Kimon Antonakopoulos, Volkan Cevher, and Panayotis Mertikopoulos. No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation. *CoRR*, abs/2206.06015, 2022.

Ramesh Johari and John N. Tsitsiklis. Efficiency loss in a network resource allocation game. *Mathematics of Operations Research*, 29(3):407–435, 2004.

Mert Kayaalp, Stefan Vlaski, and Ali H. Sayed. Dif-maml: Decentralized multi-agent meta-learning. *CoRR*, abs/2010.02870, 2020.

Mert Kayaalp, Stefan Vlaski, and Ali H Sayed. Distributed meta-learning with networked agents. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1361–1365. IEEE, 2021.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Mikhail Khodak, Maria-Florina Balcan, Ameet Talwalkar, and Sergei Vassilvitskii. Learning predictions for algorithms with predictions. In *Advances in Neural Information Processing Systems*, 2022. To appear.

Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games: extended abstract. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009*, pages 533–542. ACM, 2009.

Daphne Koller and Nimrod Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and Economic Behavior*, 4(4):528–552, 1992.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Christian Kroer and Tuomas Sandholm. A unified framework for extensive-form game abstraction with bounds. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 613–624, 2018.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

Shuangtong Li, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Learning to collaborate in decentralized learning of personalized models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9766–9775, 2022.

Yuqian Li, Vincent Conitzer, and Dmytro Korzhyk. Catcher-evader games. *arXiv preprint arXiv:1602.01896*, 2016.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few shot learning. *CoRR*, abs/1707.09835, 2017.

Brendan Lucier and Allan Borodin. Price of anarchy for greedy auctions. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010*, pages 537–553. SIAM, 2010.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: Adanormalhedge. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1286–1304. JMLR.org, 2015.

Thodoris Lykouris, Vasilis Syrgkanis, and Éva Tardos. Learning and efficiency in games with dynamic population. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016*, pages 120–129. SIAM, 2016.

H. Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011*, volume 15 of *JMLR Proceedings*, pages 525–533. JMLR.org, 2011.

Emily Meigs, Francesca Parise, and Asuman E. Ozdaglar. Learning dynamics in stochastic routing games. In *55th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2017*, pages 259–266. IEEE, 2017.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

Paul Milgrom and John Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, 58(6):1255–1277, 1990.

Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. In Tim Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 646–662. Cambridge University Press, 2020.

Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 1497–1507. PMLR, 2020.

Denis Nekipelov, Vasilis Syrgkanis, and Éva Tardos. Econometrics for learning agents. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC '15*, pages 1–18. ACM, 2015.

J. V. Neumann. A model of general economic equilibrium. *The Review of Economic Studies*, 13 (1):1–9, 1945.

Noam Nisan and Gali Noti. An experimental evaluation of regret-based econometrics. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*, pages 73–81. ACM, 2017.

Ilya Osadchiy, Kfir Y Levy, and Ron Meir. Online meta-learning in adversarial multi-armed bandits. *arXiv preprint arXiv:2205.15921*, 2022.

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

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT 2013 - The 26th Annual Conference on Learning Theory, 2013*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 993–1019. JMLR.org, 2013a.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013*, pages 3066–3074, 2013b.

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, 1951.

I. Romanovskii. Reduction of a game with complete memory to a matrix game. *Soviet Mathematics*, 3, 1962.

J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3):520–534, 1965.

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

Tim Roughgarden, Vasilis Syrgkanis, and Éva Tardos. The price of anarchy in auctions. *J. Artif. Intell. Res.*, 59:59–101, 2017.

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

Arun Sai Suggala and Praneeth Netrapalli. Follow the perturbed leader: Optimism and fast parallel algorithms for smooth minimax games. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

Vasilis Syrgkanis and Éva Tardos. Composable and efficient mechanisms. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, 2013*, pages 211–220. ACM, 2013.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 2989–2997, 2015.

Oskari Tammelin. Solving large imperfect information games using CFR+. *CoRR*, abs/1407.5042, 2014.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 1998.

Eric van Damme. *Stability and Perfection of Nash Equilibria*. Springer-Verlag, Berlin, Heidelberg, 1987.

Adrian Vetta. Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. In *43rd Symposium on Foundations of Computer Science (FOCS 2002)*, page 416. IEEE Computer Society, 2002.

Jun-Kun Wang, Jacob D. Abernethy, and Kfir Y. Levy. No-regret dynamics in the fenchel game: A unified framework for algorithmic convex optimization. *CoRR*, abs/2111.11309, 2021.

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.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

Andre Wibisono, Molei Tao, and Georgios Piliouras. Alternating mirror descent for constrained min-max games. *CoRR*, abs/2206.04160, 2022.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 449–457. JMLR.org, 2016.

Hugh Zhang, Adam Lerer, and Noam Brown. Equilibrium finding in normal-form games via greedy regret minimization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, 2022*, pages 9484–9492. AAAI Press, 2022a.

Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 26772–26808. PMLR, 2022b.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

A Additional Related Work

In this section, we provide an additional discussion on related work. Let us first compare in more detail our setting with that considered in [Zhang et al., 2022b]. Zhang et al. [2022b] study a setting which is more general than ours, thus making their algorithms applicable in the meta-learning setting we consider. Intuitively, their algorithms should not perform as well in the meta-learning setting, as they do not use knowledge of the game boundaries. More precisely, we begin by introducing the notion of dynamic Nash Equilibrium (NE) regret from Zhang et al. [2022b], and show that in the meta-learning setting it corresponds to an unnormalized version of the maximum task average regret with respect to both players.[5]

Definition A.1 (Dynamic NE regret, Zhang et al. [2022b])**.** *Given a sequence of two-player zero-sum games characterized by payoff matrices* $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(\tau)}$ *and player strategy sets* Δ^{d_x} *and* Δ^{d_y}*,*

$$\text{NE-Reg} := \left| \sum_{s=1}^{\tau} (\boldsymbol{x}^{(s)})^\top \mathbf{A}^{(s)} \boldsymbol{y}^{(s)} - \min_{\boldsymbol{x} \in \Delta^{d_x}} \max_{\boldsymbol{y} \in \Delta^{d_y}} \boldsymbol{x}^\top \mathbf{A}^{(s)} \boldsymbol{y} \right|.$$

In our meta-learning setting, $\tau = T \cdot m$, and $\mathbf{A}^{(t,i)} = \mathbf{A}^{(t)}, \forall i \in [\![m]\!], t \in [\![T]\!]$. Using this information, we can rewrite NE-Reg in our setting as

$$\text{NE-Reg} = \left| \sum_{t=1}^{T} \sum_{i=1}^{m} (\boldsymbol{x}^{(t,i)})^\top \mathbf{A}^{(t)} \boldsymbol{y}^{(t,i)} - \min_{\boldsymbol{x} \in \Delta^{d_x}} \max_{\boldsymbol{y} \in \Delta^{d_y}} \boldsymbol{x}^\top \mathbf{A}^{(t)} \boldsymbol{y} \right|.$$

Therefore,

$$\frac{1}{T} \text{NE-Reg} = \max \left\{ \frac{1}{T} \sum_{t=1}^{T} \text{Reg}_x^{(t,m)}, \frac{1}{T} \sum_{t=1}^{T} \text{Reg}_y^{(t,m)} \right\}.$$

Given this characterization, we now informally restate Theorem 6 of Zhang et al. [2022b], written in terms of task-average regret.

Theorem A.2 (Informal, Zhang et al. [2022b])**.** *When the x-player follows Algorithm 1 of Zhang et al. [2022b] and the y-player follows Algorithm 2 of Zhang et al. [2022b],*

$$\max \left\{ \frac{1}{T} \sum_{t=1}^{T} \text{Reg}_x^{(t,m)}, \frac{1}{T} \sum_{t=1}^{T} \text{Reg}_y^{(t,m)} \right\} \leq \widetilde{O}\left(\frac{1}{T} \min\{ \sqrt{(1+V)(1+P)} + P, 1 + W \} \right),$$

where $V \in \mathbb{R}_{\geq 0}$ *is a measure of the path-length variation of* $(\mathbf{A}^{(t)})_{1 \leq t \leq T}$*,* $P \in \mathbb{R}_{\geq 0}$ *is a measure of the variation of the corresponding Nash equilibrium strategies,* $W \in \mathbb{R}_{\geq 0}$ *is a measure of the variance of* $(\mathbf{A}^{(t)})_{1 \leq t \leq T}$*, and* $\widetilde{O}(\cdot)$ *hides logarithmic-in-*T *factors.*

It is easy to construct a sequence of games for which V, P, W are all $\Omega(T)$ (*e.g.*, consider an alternating sequence of two games with different payoff matrices and NE strategies for each player). Under such a setting, when players play according to the algorithms of Zhang et al. [2022b], their task average regret guarantee will actually *grow* with the number of tasks (albeit at a logarithmic rate). This is to be expected, as their algorithms are designed for a more general setting and therefore do not take the game boundaries into consideration. This is also in contrast to our results for two-player zero-sum games, where the parts of our bounds which explicitly depend on the number of games *decrease* as the number of games grows large (*e.g.*, Theorem 3.1).

[5] The results of Zhang et al. [2022b] are only applicable to two-player zero-sum games, so we will focus only on that setting in our comparison.

Broader context Moving beyond the line of work on learning in dynamic games, our results are related to the literature on algorithms with predictions; *e.g.*, see the excellent survey of Mitzenmacher and Vassilvitskii [2020], and the many references therein. More broadly, our setting can be viewed as a specific instance of online learning under more structured sequences, a topic that has received extensive attention in the literature [Block et al., 2022, Rakhlin and Sridharan, 2013a, Haghtalab et al., 2022, Cesa-Bianchi et al., 2007, Chiang et al., 2012, Hazan and Kale, 2010, 2011, Luo and Schapire, 2015]. Finally, our work provides a number of new insights on the last-iterate convergence of optimistic no-regret learning algorithms in a variety of important settings; that line of work was pioneered by Daskalakis et al. [2018], and has thereafter witnesses rapid progress (see [Mokhtari et al., 2020, Mertikopoulos et al., 2019, Golowich et al., 2020b,a, Wei et al., 2021, Azizian et al., 2021, Fiez et al., 2021a] for a highly incomplete list).

B Proofs from Section 3.2.1: Meta-Learning Approximately Optimal Equilibria

In this section, we study how meta-learning can improve the convergence rate of learning algorithms to approximately optimal equilibria, establishing the proofs omitted from Section 3.2.1. First, we provide some key preliminary results for meta-learning, which will be used throughout this paper beyond the proofs of Section 3.2.1, and in particular Appendix C. Then, equipped with those ingredients, we leverage the smoothness condition (Definition 3.5) to eventually arrive at the main theorem of Section 3.2.1 (Theorem 3.6), as well as extensions thereof.

B.1 Bounding the Social Regret

Here, we establish refined meta-learning bounds for the sum of the players' regrets—oftentimes referred to as *social regret*—under appropriate learning dynamics. A central ingredient of our analysis will be the so-called property of *regret bounded by variation in utilities (RVU)*, crystallized by Syrgkanis et al. [2015], a refined regret bound known to be satisfied by optimistic learning algorithms, such as `OMD` and *optimistic follow the regularized leader (`OFTRL`)* [Syrgkanis et al., 2015]. For our purposes, it will be crucial to use an RVU bound parameterized in terms of the initialization. Indeed, the regret guarantee below can be readily extracted from [Rakhlin and Sridharan, 2013b, Syrgkanis et al., 2015]; we remark that our assumptions are indeed compatible with the ones made in [Syrgkanis et al., 2015]. We further clarify that, in order to reduce the notational burden, when the underlying player or task are not necessary for our statement (as is below), we drop those dependencies from our notation.

Theorem B.1 (Initialization-Dependent RVU Bound [Syrgkanis et al., 2015])**.** *Suppose that we employ `OMD` with a 1-strongly convex regularizer $\mathcal{R}$ with respect to the norm $\|\cdot\|$. For any observed sequence of utilities $(\boldsymbol{u}^{(i)})_{1\le i\le m}$, the regret $\mathrm{Reg}^{(m)}$ of `OMD` up to time $m \in \mathbb{N}$ and initialized at $\boldsymbol{x}^{(0)} \in \mathcal{X}$ can be bounded as*

$$\mathrm{Reg}^{(m)} \le \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \,\|\, \boldsymbol{x}^{(0)}) + \eta \sum_{i=1}^{m} \|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|_*^2 - \frac{1}{8\eta}\sum_{i=1}^{m} \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i-1)}\|^2, \tag{5}$$

where $\mathring{\boldsymbol{x}} \in \mathcal{X}$ is an optimal strategy in hindsight, and $(\boldsymbol{x}^{(i)})_{1\le i\le m}$ is the sequence of strategies produced by `OMD`.

A similar regret bound applies for `OFTRL` as well [Syrgkanis et al., 2015], but with the important caveat that the first term in the right-hand side of (5) is not refined in terms of the initialization. That deficiency is a crucial impediment towards providing provable meta-learning guarantees under `OFTRL`, although under certain assumptions the mirror-descent viewpoint is known to be equivalent to the follow-the-regularider-leader one [McMahan, 2011]. Also, it is worth noting that

Theorem B.1 also applies under a broader class of prediction mechanisms—beyond the "one-step recency bias," which consists of $\boldsymbol{m}^{(i)} := \boldsymbol{u}^{(i-1)}$—without altering qualitatively the regret bound, as formalized by Syrgkanis et al. [2015]; such extensions will not be made precise here as they are direct.

Now, as it turns out, the RVU bound (Theorem B.1) is a powerful tool for obtaining $O(1)$ bounds for the social regret, as was first shown by Syrgkanis et al. [2015]. Below, we follow their approach to give an initialization-dependent bound for the social regret.

Corollary B.2. *Fix any $t \in [\![T]\!]$, and suppose that $L^{(t)}$ is the Lipschitz parameter of $\mathcal{G}^{(t)}$—in the sense of (1). If all players employ* `OGD` *with learning rate $\eta \leq \frac{1}{4L^{(t)}\sqrt{n-1}}$, then*

$$\sum_{k=1}^{n} \mathrm{Reg}_k^{(t,m)} \leq \frac{1}{\eta} \sum_{k=1}^{n} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) - \frac{1}{16\eta} \sum_{k=1}^{n} \sum_{i=1}^{m} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2.$$

Before we proceed with the proof, let us point out that we generally make the very mild assumption that players know the total number of players n and the Lipschitz parameter $L^{(t)}$ of each game $\mathcal{G}^{(t)}$—or, equivalently, reasonable upper bounds thereof, in order to tune the learning rate according to Corollary B.2. If that is not the case, one can still obtain similar guarantees using the by now standard "doubling trick" to estimate a suitable learning rate. Concretely, for a given learning rate, each player can compute at every iteration $i \in [\![m]\!]$ the growth of the second and the third term in (5) based on its "local" information. If at any iteration the sum of those terms is (strictly) positive, then we can simply halve the learning rate, and subsequently proceed by repeating the previous process; Corollary B.2 guarantees termination in a logarithmic (in the range of the parameters) number of repetitions, while only incurring a negligible increase in the social regret. While this protocol is not full uncoupled, given that some additional communication is required to notify the players to halve the learning rate, this is only a mild limitation that will not be addressed further in this work. The same discussion also applies to our results in the sequel.

Proof of Corollary B.2. By the Lipschitz continuity assumption (1), we have that for any player $k \in [\![n]\!]$ and iteration $i \in [\![m]\!]$,

$$\|\boldsymbol{u}_k(\boldsymbol{x}_{-k}^{(t,i)}) - \boldsymbol{u}_k(\boldsymbol{x}_{-k}^{(t,i-1)})\|_2 \leq L^{(t)} \|\boldsymbol{x}_{-k}^{(t,i)} - \boldsymbol{x}_{-k}^{(t,i-1)}\|_2.$$

Thus, combining with Theorem B.1 under $\|\cdot\| = \|\cdot\|_2$ (since we use `OGD`), we have that the regret $\mathrm{Reg}_k^{(t,m)}(\mathring{\boldsymbol{x}}_k^{(t)})$ of player $k \in [\![n]\!]$ is upper bounded by

$$\begin{aligned}
&\frac{1}{\eta} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) + \eta \sum_{i=1}^{m} \|\boldsymbol{u}_k^{(i)} - \boldsymbol{u}_k^{(i-1)}\|_2^2 - \frac{1}{8\eta} \sum_{i=1}^{m} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2 \\
&\leq \frac{1}{\eta} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) + \eta (L^{(t)})^2 \sum_{k' \neq k} \sum_{i=1}^{m} \|\boldsymbol{x}_{k'}^{(t,i)} - \boldsymbol{x}_{k'}^{(t,i-1)}\|_2^2 - \frac{1}{8\eta} \sum_{i=1}^{m} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2,
\end{aligned}$$

since $\|\boldsymbol{x}_{-k}^{(t,i)} - \boldsymbol{x}_{-k}^{(t,i-1)}\|_2^2 = \sum_{k' \neq k} \|\boldsymbol{x}_{k'}^{(t,i)} - \boldsymbol{x}_{k'}^{(t,i-1)}\|_2^2$ (Pythagorean theorem). In turn, this implies that the sum of the players' regrets $\sum_{k=1}^{n} \mathrm{Reg}_k^{(t,m)}(\mathring{\boldsymbol{x}}_k^{(t)})$ is upper bounded by

$$\begin{aligned}
&\frac{1}{\eta} \sum_{k=1}^{n} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) + \left(\eta (L^{(t)})^2 (n-1) - \frac{1}{8\eta} \right) \sum_{k=1}^{n} \sum_{i=1}^{m} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2 \\
&\leq \frac{1}{\eta} \sum_{k=1}^{n} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) - \frac{1}{16\eta} \sum_{k=1}^{n} \sum_{i=1}^{m} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2,
\end{aligned}$$

where the last bound follows since $\eta \leq \frac{1}{4L^{(t)}\sqrt{n-1}}$ (by assumption). This concludes the proof. □

ALGORITHM 1: Meta-`OGD`

Data:

- Number of players $n \in \mathbb{N} \setminus \{1\}$;
- Dimension of strategy set $d_k \in \mathbb{N}$ for each $k \in [\![n]\!]$; and
- Lipschitz constant $L > 0$.

Initialize $\boldsymbol{x}_k^{(1,0)} = \frac{1}{d_k}\mathbb{1}_{d_k} \in \mathcal{X}_k$ for each player $k \in [\![n]\!]$;

for *game* $t = 1, \dots, T$ **do**

Each player k runs `OGD` for m iterations using initialization $\boldsymbol{x}_k^{(t,0)}$ and learning rate $\eta = \frac{1}{4L\sqrt{n-1}}$;

Each player k computes the next initialization as $\boldsymbol{x}_k^{(t+1,0)} = \frac{1}{t}\sum_{s=1}^{t} \mathring{\boldsymbol{x}}_k^{(t)}$;

end

We remark that in the proof above we used the assumption that $\boldsymbol{m}_k^{(t,1)} := \boldsymbol{u}_k(\boldsymbol{x}_{-k}^{(t,0)})$, which is also needed in [Syrgkanis et al., 2015]—although it was not explicitly mentioned by the authors. That assumption can be circumvented using standard techniques; as such, it will be lifted in Appendix C.8.

Armed with Corollary B.2, we are now ready to state the key result of this subsection: a refined guarantee for the sum of the players' regrets, parameterized in terms of the similarity of the optimal in hindsight of each player. The meta-version of `OGD` that we consider is summarized in Algorithm 1.

Theorem B.3. *Suppose that each player $k \in [\![n]\!]$ employs `OGD` with learning rate $\eta = \frac{1}{4L\sqrt{n-1}}$, with $L := \max_{1\le t\le T} L^{(t)}$, and initialization $\boldsymbol{x}_k^{(t,0)} := \sum_{s<t} \mathring{\boldsymbol{x}}_k^{(s)}/(t-1)$, for $t \ge 2$. Then,*

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{n} \mathrm{Reg}_k^{(t,m)} \le 2L\sqrt{n-1}\sum_{k=1}^{n} V_k^2 + \frac{4L\sqrt{n-1}(1+\log T)}{T}\sum_{k=1}^{n}\Omega_{\mathcal{X}_k}^2. \tag{6}$$

In light of Corollary B.2, the main ingredient for establishing this theorem is a meta-algorithm that determines the initialization of `OGD`. As we shall see, the initialization seen in Theorem B.3, namely $\boldsymbol{x}_k^{(t,0)} := \sum_{s<t} \mathring{\boldsymbol{x}}_k^{(s)}/(t-1)$, is what comes out of *follow the leader* `FTL`—the "unregularized" version of `FTRL`. To justify this, we will need the following auxiliary results. We note that, for convenience, the guarantee below is stated in terms of an underlying sequence of losses, instead of utilities; naturally, those two viewpoints are equivalent.

Proposition B.4 ([Khodak et al., 2019])**.** *Let $\mathcal{R} : \mathcal{X} \to \mathbb{R}$ be a 1-strongly convex regularizer with respect to the norm $\|\cdot\|$. Then, for any sequence $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(T)} \in \mathcal{X}$, `FTL` run on the sequence of losses $\alpha^{(1)}\mathcal{B}_{\mathcal{R}}(\boldsymbol{x}^{(1)} \,\|\, \cdot), \dots, \alpha^{(T)}\mathcal{B}_{\mathcal{R}}(\boldsymbol{x}^{(T)} \,\|\, \cdot)$, where $\boldsymbol{\alpha} \in \mathbb{R}_{>0}^T$, has regret $\mathrm{Reg}^{(T)}$ bounded as*

$$\mathrm{Reg}^{(T)} \le 2C\Omega_{\mathcal{X}}\sum_{t=1}^{T}\frac{(\alpha^{(T)})^2 G^{(T)}}{\alpha^{(t)} + 2\sum_{s<t}\alpha^{(s)}},$$

where $C \in \mathbb{R}_{>0}$ is such that $\|\boldsymbol{x}\| \le C\|\boldsymbol{x}\|_2$ for any $\boldsymbol{x} \in \mathcal{X}$; $\Omega_{\mathcal{X}}$ is the ℓ_2-diameter of $\mathcal{X}$; and $G^{(t)}$ is the Lipschitz constant of the function $\mathcal{B}_{\mathcal{R}}(\boldsymbol{x}^{(t)} \,\|\, \cdot)$ with respect to $\|\cdot\|$.

In particular, in this section we will be using this proposition for the Bregman divergence induced by the Euclidean regularizer $\mathcal{R} : \boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|_2^2$, in which case the sequence of losses encountered by the `FTL` algorithm happens to be strongly convex; for that special case, logarithmic regret under `FTL` is well-known from earlier works [Cesa-Bianchi and Lugosi, 2006]. Proposition B.4 can then be further simplified so that $C = 1$ and $G^{(t)} \le \Omega_{\mathcal{X}}$, for any $t \in [\![T]\!]$, where we recall that $\Omega_{\mathcal{X}}$ is the ℓ_2-diameter of $\mathcal{X}$. To see this, we simply note that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$,

$$\left|\mathcal{B}_{\mathcal{R}}(\boldsymbol{x}^{(t)} \,\|\, \boldsymbol{x}) - \mathcal{B}_{\mathcal{R}}(\boldsymbol{x}^{(t)} \,\|\, \boldsymbol{x}')\right| = \left|\left\langle \boldsymbol{x} - \boldsymbol{x}', \frac{\boldsymbol{x}+\boldsymbol{x}'}{2} - \boldsymbol{x}^{(t)}\right\rangle\right| \le \Omega_{\mathcal{X}}\|\boldsymbol{x} - \boldsymbol{x}'\|_2,$$

since $\frac{\boldsymbol{x}+\boldsymbol{x}'}{2} \in \mathcal{X}$ (by convexity). Further, FTL takes a natural form, as implied by the following simple claim. We recall again that $\mathcal{X}$ it always assumed to be nonempty convex and compact.

Claim B.5. *Consider a sequence of points $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(T)} \in \mathcal{X} \subseteq \mathbb{R}^d$, for $d \in \mathbb{N}$. Then, the function*

$$\mathcal{X} \ni \boldsymbol{x} \mapsto \frac{1}{2} \sum_{t=1}^{T} \alpha^{(t)} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}\|_2^2,$$

with $\boldsymbol{\alpha} \in \mathbb{R}^T_{>0}$, attains its (unique) minimum at $\boldsymbol{x}^\star \coloneqq \sum_{t=1}^T \alpha^{(t)} \boldsymbol{x}^{(t)} / \sum_{t=1}^T \alpha^{(t)}$.

We point out that, by convexity, $\boldsymbol{x}^\star$ is indeed a feasible point in $\mathcal{X}$. Such a characterization (Claim B.5) is known to be extended beyond Euclidean regularization [Banerjee et al., 2005], which will be used in the proofs of Theorems D.7 and D.10. We are now ready to establish Theorem B.3.

Proof of Theorem B.3. First, by Corollary B.2, we have that for any task $t \in [\![T]\!]$,

$$\sum_{k=1}^{n} \mathrm{Reg}_k^{(t,m)} \leq 4L\sqrt{n-1} \sum_{k=1}^{n} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \parallel \boldsymbol{x}_k^{(t,0)}), \tag{7}$$

where we used the fact that $\eta \coloneqq \frac{1}{4L\sqrt{n-1}} \leq \frac{1}{4L^{(t)}\sqrt{n-1}}$ for each task $t \in [\![T]\!]$, by definition of $L \coloneqq \max_{1 \leq t \leq T} L^{(t)}$, thereby satisfying the precondition of Corollary B.2. Thus,

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{n} \mathrm{Reg}_k^{(t,m)} &\leq \frac{4L\sqrt{n-1}}{T} \sum_{k=1}^{n} \sum_{t=1}^{T} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \parallel \boldsymbol{x}_k^{(t,0)}) && (8)\\
&= \frac{2L\sqrt{n-1}}{T} \sum_{k=1}^{n} \sum_{t=1}^{T} \|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k^{(t,0)}\|_2^2 && (9)\\
&= \frac{2L\sqrt{n-1}}{T} \sum_{k=1}^{n} \min_{\boldsymbol{x}_k \in \mathcal{X}} \sum_{t=1}^{T} \|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k\|_2^2 \\
&\quad + \frac{2L\sqrt{n-1}}{T} \sum_{k=1}^{n} \left(\sum_{t=1}^{T} \|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k^{(t,0)}\|_2^2 - \min_{\boldsymbol{x}_k \in \mathcal{X}} \sum_{t=1}^{T} \|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k\|_2^2 \right) \\
&\leq \frac{2L\sqrt{n-1}}{T} \sum_{k=1}^{n} \min_{\boldsymbol{x}_k \in \mathcal{X}} \sum_{t=1}^{T} \|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k\|_2^2 + \frac{8L\sqrt{n-1}(1+\log T)}{T} \sum_{k=1}^{n} \Omega_{\mathcal{X}_k}^2 && (10)\\
&= \frac{2L\sqrt{n-1}}{T} \sum_{k=1}^{n} \sum_{t=1}^{T} \|\mathring{\boldsymbol{x}}_k^{(t)} - \bar{\boldsymbol{x}}_k\|_2^2 + \frac{8L\sqrt{n-1}(1+\log T)}{T} \sum_{k=1}^{n} \Omega_{\mathcal{X}_k}^2 && (11)\\
&= 2L\sqrt{n-1} \sum_{k=1}^{n} V_k^2 + \frac{8L\sqrt{n-1}(1+\log T)}{T} \sum_{k=1}^{n} \Omega_{\mathcal{X}_k}^2, && (12)
\end{aligned}$$

where

- (8) follows from (7);
- (9) uses the fact that OGD corresponds to OMD with $\mathcal{R}_k : \mathcal{X}_k \ni \boldsymbol{x}_k \mapsto \frac{1}{2}\|\boldsymbol{x}_k\|_2^2$, thereby implying that $\mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}^{(t)} \parallel \boldsymbol{x}_k^{(t,0)}) = \frac{1}{2}\|\mathring{\boldsymbol{x}}^{(t)} - \boldsymbol{x}_k^{(t,0)}\|_2^2$;

- (10) uses Proposition B.4 with $\|\cdot\| = \|\cdot\|_2$. In particular, the quantity

$$\frac{1}{2}\sum_{t=1}^{T}\left(\|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k^{(t,0)}\|_2^2 - \min_{\boldsymbol{x}_k\in\mathcal{X}}\|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k\|_2^2\right) \tag{13}$$

  can be recognized as the regret incurred by the FTL algorithm used by each player $k \in [\![n]\!]$, which in turn follows since we have initialized as $\boldsymbol{x}_k^{(t,0)} := \sum_{s<t}\mathring{\boldsymbol{x}}_k^{(s)}/(t-1)$, for $t \geq 2$, which by Claim B.5 is exactly the update of FTL under $\alpha^{(1)} = \cdots = \alpha^{(t)} = 1$. As a result, by Proposition B.4, the regret of FTL (13) for each player $k \in [\![n]\!]$ can be upper bounded by

$$2\Omega_{\mathcal{X}}^2\sum_{t=1}^{T}\frac{1}{2t-1} \leq 2\Omega_{\mathcal{X}}^2\sum_{t=1}^{T}\frac{1}{t} \leq 2\Omega_{\mathcal{X}}^2(1+\log T),$$

  where the last bound uses the well-known inequality that the t-th harmonic number $\mathcal{H}^{(t)}$ is upper bounded by $1+\log T$, where $\log(\cdot)$ here denotes the natural logarithm.

- (11) uses the fact that the function $\boldsymbol{x}_k \mapsto \sum_{t=1}^{T}\|\mathring{\boldsymbol{x}}^{(t)} - \boldsymbol{x}_k\|_2^2$ is minimized at $\bar{\boldsymbol{x}}_k := \sum_{t=1}^{T}\mathring{\boldsymbol{x}}_k^{(t)}/T$ (by Claim B.5); and

- (12) uses the notation $V_k^2 := \frac{1}{T}\sum_{t=1}^{T}\|\mathring{\boldsymbol{x}}_k^{(t)} - \bar{\boldsymbol{x}}_k\|_2^2$ for the similarity of the optimal in hindsight of player $k \in [\![n]\!]$.

□

Remark B.6. *For the sake of simplicity in the exposition, in Theorem B.3 we have assumed that players use the same learning rate for each task, but that guarantee can be further improved if players use different learning rates, as long as Corollary B.2 can be applied. Indeed, Proposition B.4 is versatile enough to capture the case where the Bregman divergences have different weights, though the induced expression in that case is rather convoluted.*

Remark B.7. *Theorem B.3 can be directly applied if the left-hand side of (6) is replaced by* $\frac{1}{T}\sum_{t=1}^{T}\alpha^{(t)}\sum_{k=1}^{n}\mathrm{Reg}_k^{(t,m)}$*, in which case the right-hand side of (6) ought to be multiplied by* $\max_{1\leq t\leq T}\alpha^{(t)}$*, for some vector* $\boldsymbol{\alpha} \in \mathbb{R}_{>0}^{T}$*. This simple fact will be exploited in the proof of Theorem 3.6.*

Remark B.8. *From (9), if each player* $k \in [\![n]\!]$ *initializes to* $\boldsymbol{x}_k^{(t,0)} = \mathring{\boldsymbol{x}}_k^{(t-1)}$*, i.e., the optima-in-hindsight of the previous game, then* $\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{n}\mathrm{Reg}_k^{(t,m)} \leq \frac{2L\sqrt{n-1}}{T}\sum_{t=1}^{T}\sum_{k=1}^{n}\|\mathring{\boldsymbol{x}}_k^{(t)} - \mathring{\boldsymbol{x}}_k^{(t-1)}\|_2^2$*, where by convention* $\mathring{\boldsymbol{x}}_k^{(0)} := \boldsymbol{x}_k^{(1,0)}$*. Similarly, if each player initializes to* $\boldsymbol{x}_k^{(t,0)} = \boldsymbol{x}_k^{(t-1,m)}$*, i.e., the last iterate of the previous game, then* $\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{n}\mathrm{Reg}_k^{(t,m)} \leq \frac{2L\sqrt{n-1}}{T}\sum_{t=1}^{T}\sum_{k=1}^{n}\|\mathring{\boldsymbol{x}}_k^{(t)} - \boldsymbol{x}_k^{(t-1,m)}\|_2^2$*, where by convention* $\boldsymbol{x}_k^{(0,m)} := \boldsymbol{x}_k^{(1,0)}$*. While the bound in Theorem B.3 might be generally better if there is no sequential structure to the games encountered, it may be desirable to instead use one of the other initializations if some form of sequential structure is known to exist, as illustrated in our experiments (Section 4).*

B.2 Implications for the Social Welfare

Having refined the guarantee in terms of the social regret (Theorem B.3), we can proceed with the proof of Theorem 3.6. Let us first further comment on Definition 3.5, which was introduced earlier in Section 3.2.1. To do so, we first give some basic background on normal-form (or strategic-form) games; this will also make our setup in Section 2 more concrete. In a normal-form game, each player $k \in [\![n]\!]$ has a finite and nonempty set of available actions $\mathcal{A}_k$. For a joint action profile $\boldsymbol{a} = (a_1, \ldots, a_n) \in \bigtimes_{k=1}^{n}\mathcal{A}_k$, there is a utility function $u_k : \boldsymbol{a} \mapsto u_k(\boldsymbol{a})$, assigning

a utility to each player k given $\boldsymbol{a}$. Players are allowed to randomize by selecting a probability distribution over their set of actions. Indeed, $\mathcal{X}_k \coloneqq \Delta(\mathcal{A}_k)$ corresponds to the strategy set of player $k \in [\![n]\!]$. In the setting described in Section 2, $u_k : \bigtimes_{k=1}^n \mathcal{X}_k \to \mathbb{R}$ is the mixed extension of the utility function: $u_k : \boldsymbol{x} \mapsto \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{x}}[u_k(\boldsymbol{a})]$. That is, players act so as to maximize their expected utility.

In this context, Roughgarden [2015] introduced his notion of smoothness with respect to pure strategies:

$$\sum_{k=1}^{n} u_k(\boldsymbol{a}_k^\star, \boldsymbol{a}_{-k}) \geq \lambda \mathrm{OPT} - \mu \mathrm{SW}(\boldsymbol{a}), \tag{14}$$

for any two joint action profiles $\boldsymbol{a}, \boldsymbol{a}^\star \in \bigtimes_{k=1}^n \mathcal{A}_k$; in fact, $\boldsymbol{a}^\star$ can be restricted to be an action profile that maximizes social welfare [Roughgarden, 2015]. Thus, Definition 3.5, stated in terms of mixed strategies, can be obtained by taking an expectation in (14) and restricting $\boldsymbol{x}^\star$ to be a pure strategy—namely, $\boldsymbol{a}^\star$.

For a (λ, μ)-smooth game $\mathcal{G}$, we recall that the *robust* price of anarchy is defined as $\rho \coloneqq \lambda/(1+\mu)$; it was coined "robust" by Roughgarden [2015] because—among others—it gives a guarantee for any coarse correlated equilibrium of $\mathcal{G}$—not just the Nash equilibria. We note that the robust price of anarchy might be different than the price of anarchy—the ratio between the worst Nash equilibrium (in terms of social welfare)—and the optimal state—hypothetically imposed by a benevolent dictator. The importance of Roughgarden's smoothness framework is that for many classes of games ρ is remarkably close to the actual price of anarchy [Roughgarden, 2015]. Indeed, we give in Table 2 a number of important settings for which ρ is close to 1, including games/mechanisms discussed by Hartline et al. [2015].

Table 2: Smoothness parameters for well-studied games/mechanisms in the literature. For some of those settings, smoothness—in the sense of Definition 3.5—was only subsequently crystallized by Roughgarden [2015] based on the earlier arguments.

Game/Mechanism	(λ, μ)	**Reference**
Simultaneous first-price auction with submodular bidders	$\left(1 - \frac{1}{e}, 1\right)$	[Syrgkanis and Tardos, 2013]
First-price multi-unit auction	$\left(1 - \frac{1}{e}, 1\right)$	[Syrgkanis and Tardos, 2013]
Simultaneous second-price auctions	$(1, 1)$	[Christodoulou et al., 2016]
Greedy combinatorial auction with d-complements	$\left(1 - \frac{1}{e}, d\right)$	[Lucier and Borodin, 2010]
Valid utility games	$(1, 1)$	[Vetta, 2002]
Congestion games with affine costs	$\left(\frac{5}{3}, \frac{1}{3}\right)$	[Christodoulou and Koutsoupias, 2005]

Now, for the sake of generality, let us treat the problem of maximizing welfare from a slightly broader standpoint. More precisely, for a vector $\boldsymbol{\alpha} \in \mathbb{R}^n_{>0}$, we define $\mathrm{SW}_{\boldsymbol{\alpha}}(\boldsymbol{x}) \coloneqq \sum_{k=1}^n \alpha_k u_k(\boldsymbol{x})$. That is, the utility of each player $k \in [\![n]\!]$ is weighted using a coefficient $\alpha_k > 0$; when $\alpha_1 = \cdots = \alpha_n = 1$, we recover the standard notion of (utilitarian) social welfare we introduced earlier. We also let $\mathrm{OPT}_{\boldsymbol{\alpha}} \coloneqq \max_{\boldsymbol{x} \in \mathcal{X}} \mathrm{SW}_{\boldsymbol{\alpha}}(\boldsymbol{x})$. In this context, we consider the following generalized notion of smoothness.

Definition B.9 (Extension of Definition 3.5)**.** *A game $\mathcal{G}$ is (λ, μ)-smooth with respect to $\boldsymbol{\alpha} \in \mathbb{R}^n_{>0}$, with $\lambda, \mu > 0$, if there exists a strategy profile $\boldsymbol{x}^\star \in \bigtimes_{k=1}^n \mathcal{X}_k$ such that for any $\boldsymbol{x} \in \bigtimes_{k=1}^n \mathcal{X}_k$,*

$$\sum_{k=1}^{n} \alpha_k u_k(\boldsymbol{x}_k^\star, \boldsymbol{x}_{-k}) \geq \lambda \mathrm{OPT}_{\boldsymbol{\alpha}} - \mu \mathrm{SW}_{\boldsymbol{\alpha}}(\boldsymbol{x}). \tag{15}$$

In the sequel, when $\boldsymbol{\alpha}$ will remain unspecified it will be implied that $\alpha_1 = \cdots = \alpha_n = 1$. Below we show a simple extension of an observation due to Roughgarden [2015].

Proposition B.10. *Suppose that each player $k \in [\![m]\!]$ incurs regret at most $\mathrm{Reg}_k^{(m)}$ in a (λ, μ)-smooth game $\mathcal{G}$ with respect to $\boldsymbol{\alpha} \in \mathbb{R}^n_{>0}$. Then,*

$$\frac{1}{m} \sum_{i=1}^{m} \mathrm{SW}_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(i)}) \geq \frac{\lambda}{1+\mu} \mathrm{OPT}_{\boldsymbol{\alpha}} - \frac{1}{1+\mu} \frac{1}{m} \sum_{k=1}^{n} \alpha_k \mathrm{Reg}_k^{(m)},$$

where $\lambda/(1+\mu)$ is the (robust) price of anarchy.

Proof. Suppose that $\boldsymbol{x}^\star \in \bigtimes_{k=1}^n \mathcal{X}_k$ is the strategy profile for which (15) is satisfied. Then, by definition, for any player $k \in [\![n]\!]$ and iteration $i \in [\![m]\!]$,

$$\alpha_k \mathrm{Reg}_k^{(m)} \geq \sum_{i=1}^{m} \alpha_k u_k(\boldsymbol{x}_k^\star, \boldsymbol{x}_{-k}^{(i)}) - \alpha_k \sum_{i=1}^{m} u_k(\boldsymbol{x}^{(i)}),$$

since $\alpha_k > 0$ (by assumption). So, summing over all players,

$$\sum_{k=1}^{n} \alpha_k \mathrm{Reg}_k^{(m)} \geq \lambda m \mathrm{OPT} - \mu \sum_{i=1}^{m} \mathrm{SW}_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(i)}) - \sum_{i=1}^{m} \mathrm{SW}_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(i)}),$$

and rearranging the last inequality concludes the proof. □

Next, we first focus on the case where $\alpha_1 = \cdots = \alpha_n = 1$, and we then give the more general result. In particular, below we show Theorem 3.6, the informal version of which was stated earlier in Section 3.2.1.

Theorem B.11 (Detailed Version of Theorem 3.6)**.** *If all players use* `OGD` *with learning rate $\eta = \frac{1}{4L\sqrt{n-1}}$, with $L \coloneqq \max_{1 \leq t \leq T} L^{(t)}$, and initialization $\boldsymbol{x}_k^{(t,0)} \coloneqq \sum_{s<t} \mathring{\boldsymbol{x}}_k^{(s)} / (t-1)$, for $t \geq 2$ and $k \in [\![n]\!]$, in a sequence of T games $(\mathcal{G}^{(t)})_{1 \leq t \leq T}$, each of which is $(\lambda^{(t)}, \mu^{(t)})$-smooth, then*

$$\frac{1}{mT} \sum_{t=1}^{T} \sum_{i=1}^{m} \mathrm{SW}(\boldsymbol{x}^{(t,i)}) \geq \frac{1}{T} \sum_{t=1}^{T} \frac{\lambda^{(t)}}{1+\mu^{(t)}} \mathrm{OPT}^{(t)} - \frac{2L\sqrt{n-1}}{m} \sum_{k=1}^{n} V_k^2 - \widetilde{O}\left(\frac{1}{mT}\right),$$

where $\boldsymbol{x}^{(t,i)} \coloneqq (\boldsymbol{x}_1^{(t,i)}, \ldots, \boldsymbol{x}_n^{(t,i)})$ is the strategy produced by the players at iteration i of task t, and $\mathrm{OPT}^{(t)}$ is the optimal social welfare attainable at game $\mathcal{G}^{(t)}$.

Proof. First, using Proposition B.10 we have that for any task $t \in [\![T]\!]$,

$$\frac{1}{m} \sum_{i=1}^{m} \mathrm{SW}(\boldsymbol{x}^{(t,i)}) \geq \frac{\lambda^{(t)}}{1+\mu^{(t)}} \mathrm{OPT}^{(t)} - \frac{1}{1+\mu^{(t)}} \frac{1}{m} \sum_{k=1}^{n} \mathrm{Reg}_k^{(t,m)}, \tag{16}$$

since each game $\mathcal{G}^{(t)}$ is assumed to be $(\lambda^{(t)}, \mu^{(t)})$-smooth with optimal social welfare $\mathrm{OPT}^{(t)}$. Hence, taking the average of (16) over all $t \in [\![T]\!]$ yields that

$$\frac{1}{mT} \sum_{t=1}^{T} \sum_{i=1}^{m} \mathrm{SW}(\boldsymbol{x}^{(t,i)}) \geq \frac{1}{T} \sum_{t=1}^{T} \frac{\lambda^{(t)}}{1+\mu^{(t)}} \mathrm{OPT}^{(t)} - \frac{1}{mT} \sum_{t=1}^{T} \frac{1}{1+\mu^{(t)}} \sum_{k=1}^{n} \mathrm{Reg}_k^{(t,m)}.$$

Finally, using Theorem B.3 along with the refinement discussed in Remark B.7—with $\alpha^{(t)} \coloneqq \frac{1}{1+\mu^{(t)}} \in [0,1]$—concludes the proof. □

For the more general result in which we allow any weight vector $\boldsymbol{\alpha} \in \mathbb{R}^n_{>0}$, it suffices to refine Corollary B.2 as follows.

Corollary B.12 (Extension of Corollary B.2)**.** *Fix any $t \in [\![T]\!]$, and suppose that $L^{(t)}$ is the Lipschitz parameter of $\mathcal{G}^{(t)}$—in the sense of (1). If all players employ* `OGD` *with learning rate*

$$\eta \leq \frac{1}{2\sqrt{2}L^{(t)}} \max_{k \in [\![n]\!]} \sqrt{\frac{\alpha_k}{\sum_{k' \neq k} \alpha_{k'}}}, \tag{17}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^n_{>0}$, then

$$\sum_{k=1}^{n} \alpha_k \operatorname{Reg}_k^{(t,m)} \leq \frac{1}{\eta} \sum_{k=1}^{n} \alpha_k \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}). \tag{18}$$

The proof is analogous to Corollary B.2, but we include it below for completeness. We remark that if $\alpha_k = 1$, for some player $k \in [\![n]\!]$, and the rest of the coefficients approach to 0, Corollary B.12 gives a vacuous guarantee: the learning rate required in (17) approaches to 0, thereby making the right-hand side of (18) unbounded. This is precisely the reason why obtaining near-optimal $\widetilde{O}(1)$ bounds for $\max_{k \in [\![n]\!]} \operatorname{Reg}_k^{(m)}$, instead of $\sum_{k=1}^n \operatorname{Reg}_k^{(m)}$, requires more refined techniques [Daskalakis et al., 2021].

Proof of Corollary B.12. Similarly to Corollary B.12, by $L^{(t)}$-Lipschitz continuity (1) and Theorem B.1, we have that $\alpha_k \operatorname{Reg}_k^{(t,m)}(\mathring{\boldsymbol{x}}_k^{(t)})$ can be upper bounded by

$$\frac{\alpha_k}{\eta} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) + \eta \alpha_k (L^{(t)})^2 \sum_{k' \neq k} \sum_{i=1}^{m} \|\boldsymbol{x}_{k'}^{(t,i)} - \boldsymbol{x}_{k'}^{(t,i-1)}\|_2^2 - \frac{\alpha_k}{8\eta} \sum_{i=1}^{m} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2,$$

for any player $k \in [\![k]\!]$, where we used the fact that $\alpha_k > 0$. As a result, we have that $\sum_{k=1}^n \alpha_k \operatorname{Reg}_k^{(t,m)}(\mathring{\boldsymbol{x}}_k^{(t)})$ can be in turn upper bounded by

$$\frac{1}{\eta} \sum_{k=1}^{n} \alpha_k \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) + \sum_{k=1}^{n} \left(\eta (L^{(t)})^2 \sum_{k' \neq k} \alpha_{k'} - \frac{\alpha_k}{8\eta} \right) \sum_{i=1}^{m} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2.$$

Combining this with the bound on the learning rate (17) completes the proof. □

C Proofs from Section 3.1: Meta-Learning in Zero-Sum Games

In this section, we primarily focus on meta-learning in two-player zero-sum games. We also give a number of gradual extensions and generalizations.

C.1 Bilinear Saddle-Point Problems

We begin by considering the bilinear saddle-point problem

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{y}, \tag{19}$$

where we recall that $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ is the nonempty convex and compact set of strategies of player x with $d_x \in \mathbb{N}$; $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ is the nonempty convex and compact set of strategies of player y with $d_y \in \mathbb{N}$; and $\mathbf{A} \in \mathbb{R}^{d_x \times d_y}$ is the coupling (payoff) matrix of the game. In what follows in the coming subsections, we will extend our results to more general settings.

We will first apply Theorem B.3 (proven in Appendix B) for the bilinear saddle-point problem (19). To do so, we first note that the Lipschitz continuity assumption (1) for a fixed task

t is satisfied with $L^{(t)} \coloneqq \|\mathbf{A}^{(t)}\|_2$, where $\|\mathbf{A}^{(t)}\|_2$ is the spectral norm of $\mathbf{A}^{(t)}$. Indeed, for player x it holds that $\boldsymbol{u}_x^{(t)}(\boldsymbol{y}) \coloneqq -\mathbf{A}^{(t)}\boldsymbol{y}$, thereby implying that $\|\boldsymbol{u}_x^{(t)}(\boldsymbol{y}) - \boldsymbol{u}_x^{(t)}(\boldsymbol{y}')\|_2 \leq \|\mathbf{A}^{(t)}\|_2\|\boldsymbol{y} - \boldsymbol{y}'\|_2$, for any $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$, by definition of the spectral norm $\mathbf{A}^{(t)}$; similar reasoning applies for the utility vectors observed by player y given that $\|(\mathbf{A}^{(t)})^\top\|_2 = \|\mathbf{A}^{(t)}\|_2$.

Theorem C.1 (Sum of Regrets in BSPPs)**.** *Suppose that both players employ* `OGD` *with initialization* $\boldsymbol{x}_k^{(t,0)} \coloneqq \sum_{s<t} \mathring{\boldsymbol{x}}_k^{(s)}/(t-1)$*, for* $t \geq 2$*, and learning rate* $\eta \coloneqq \frac{1}{4L}$*, where* $L \coloneqq \max_{t\in[T]} \|\mathbf{A}^{(t)}\|_2$*. Then, the average sum of the players' regrets over all tasks can be upper bounded by*

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathrm{Reg}_x^{(t,m)} + \mathrm{Reg}_y^{(t,m)}\right) \leq 2L\left(V_x^2 + V_y^2\right) + \frac{8L(1+\log T)}{T}\left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2\right).$$

Here, we recall that we use the notation V_y^2, V_y^2 for the task similarity in terms of the optimal in hindsight for player x and y, respectively. We will also obtain results that depend on the similarity of the Nash equilibria. In Theorem C.1 it is assumed that players know the Lipschitz constant of the game; such an assumption can be met by first rescaling the payoff matrix, so that L will be a universal constant. Alternatively, one can employ the doubling trick (see our discussion after Corollary B.2).

In light of the well-known connection that the sum of the players' regrets drives the rate of convergence of the average strategies to the set of Nash equilibria (*e.g.*, see [Freund and Schapire, 1997]), Theorem C.1 yields the following consequence.

Corollary C.2 (Average Duality Gap in BSPPs; Detailed Version of Theorem 3.1)**.** *In the setting of Theorem C.1, let* $\bar{\boldsymbol{x}}^{(t)}, \bar{\boldsymbol{y}}^{(t)}$ *be the average strategy of the two players, respectively, at each task* $t \in [\![T]\!]$*. Then,*

$$\frac{1}{T}\sum_{t=1}^{T}\textsc{DualGap}^{(t)}(\bar{\boldsymbol{x}}^{(t)}, \bar{\boldsymbol{y}}^{(t)}) \leq \frac{2L}{m}\left(V_x^2 + V_y^2\right) + \frac{8L(1+\log T)}{mT}\left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2\right).$$

Here, we used the notation

$$\textsc{DualGap}^{(t)} : \mathcal{X}\times\mathcal{Y} \ni \hat{\boldsymbol{x}}\times\hat{\boldsymbol{y}} \mapsto \max_{\boldsymbol{y}\in\mathcal{Y}} \hat{\boldsymbol{x}}^\top\mathbf{A}^{(t)}\boldsymbol{y} - \min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{x}^\top\mathbf{A}^{(t)}\hat{\boldsymbol{y}}.$$

Proof of Corollary C.2. The claim follows from Theorem C.1, using the fact that for any task $t \in [\![T]\!]$,

$$\begin{aligned}\frac{1}{m}\left(\mathrm{Reg}_x^{(t,m)} + \mathrm{Reg}_y^{(t,m)}\right) &= \frac{1}{m}\left(\min_{\boldsymbol{y}^\star\in\mathcal{Y}}\left\langle \boldsymbol{y}^\star, \sum_{i=1}^{m}(\mathbf{A}^{(t)})^\top\boldsymbol{x}^{(t,i)}\right\rangle - \max_{\boldsymbol{x}^\star\in\mathcal{X}}\left\langle \boldsymbol{x}^\star, \sum_{i=1}^{m}\mathbf{A}^{(t)}\boldsymbol{y}^{(t,i)}\right\rangle\right)\\ &= \textsc{DualGap}^{(t)}\left(\frac{\sum_{i=1}^{m}\boldsymbol{x}^{(t,m)}}{m}, \frac{\sum_{i=1}^{m}\boldsymbol{y}^{(t,m)}}{m}\right).\end{aligned}$$

□

This guarantee can significantly improve over the standard m^{-1} rates in zero-sum games (when each game is treated separately) [Daskalakis et al., 2015]. We next provide bounds in terms of the individual regret of each player using a simple observation in [Anagnostides et al., 2022b]. First, we connect the so-called second-order path lengths of the dynamics with the first term in the RVU bound (Theorem B.1):

Proposition C.3 (Bounded Second-Order Path Length)**.** *Fix any task* $t \in [\![T]\!]$ *associated with the bilinear saddle-point problem* (19) *under matrix* $\mathbf{A}^{(t)}$*. If both players employ* `OGD` *with learning rate* $\eta \leq \frac{1}{4\|\mathbf{A}^{(t)}\|_2}$*, then*

$$\sum_{i=1}^{m}\|\boldsymbol{x}^{(t,i)} - \boldsymbol{x}^{(t,i-1)}\|_2^2 + \sum_{i=1}^{m}\|\boldsymbol{y}^{(t,i)} - \boldsymbol{y}^{(t,i-1)}\|_2^2 \leq 8\left(\|\mathring{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t,0)}\|_2^2 + \|\mathring{\boldsymbol{y}}^{(t)} - \boldsymbol{y}^{(t,0)}\|_2^2\right).$$

Proof. The claim is immediate from Corollary B.2, using the fact that $\mathrm{Reg}_x^{(t,m)} + \mathrm{Reg}_y^{(t,m)} = \mathrm{Reg}_x^{(t,m)}(\mathring{\boldsymbol{x}}^{(t)}) + \mathrm{Reg}_y^{(t,m)}(\mathring{\boldsymbol{y}}^{(t)}) \geq 0$, by definition of $\mathring{\boldsymbol{x}}^{(t)} \in \mathcal{X}$ and $\mathring{\boldsymbol{y}}^{(t)} \in \mathcal{Y}$. □

When combined with the RVU bound, this proposition can be used to derive the following refined guarantee for the individual regret experienced by each player.

Corollary C.4 (Individual Per-Player Regret)**.** *In the setting of Theorem C.1, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T} \max\left\{\mathrm{Reg}_x^{(t,m)}, \mathrm{Reg}_y^{(t,m)}\right\} \leq 4L\left(V_x^2 + V_y^2\right) + \frac{16L(1+\log T)}{T}\left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2\right).$$

Proof. Let us first fix any task $t \in [\![T]\!]$. Combining Theorem B.1 and Proposition C.3, we have that

$$\max\left\{\mathrm{Reg}_x^{(t,m)}, \mathrm{Reg}_y^{(t,m)}\right\} \leq 4L\left(\|\mathring{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^{(t,0)}\|_2^2 + \|\mathring{\boldsymbol{y}}^{(t)} - \boldsymbol{y}^{(t,0)}\|_2^2\right).$$

Thus, the statement follows similarly to Theorem B.3. □

Analogous guarantees apply to games beyond two-player zero-sum, such as strategically zero-sum and zero-sum polymatrix games, using the same technique [Anagnostides et al., 2022b].

C.1.1 Last-Iterate Bounds

We next switch gears and focus on obtaining bounds on the number of iterations required so that the iterates of `OGD`—instead of the average iterates—reach approximate Nash equilibria, the definition of which is recalled below for general multiplayer games.

Definition C.5 (Approximate Nash Equilibria)**.** *A joint strategy profile $\boldsymbol{x}^\star \in \bigtimes_{k=1}^n \mathcal{X}_k$ is an ϵ-approximate Nash equilibrium, with $\epsilon \geq 0$, if for any player $k \in [\![n]\!]$ and any possible deviation $\boldsymbol{x}_k \in \mathcal{X}_k$,*

$$u_k(\boldsymbol{x}^\star) \geq u_k(\boldsymbol{x}_{-k}^\star, \boldsymbol{x}_k) - \epsilon.$$

Nash equilibria are known to exist under general assumptions [Rosen, 1965]. In this context, we will need the following refined RVU bound, which can be readily extracted from [Syrgkanis et al., 2015, Rakhlin and Sridharan, 2013b].

Theorem C.6 ([Syrgkanis et al., 2015, Rakhlin and Sridharan, 2013b])**.** *For any observed sequence of utilities $(\boldsymbol{u}^{(i)})_{1\leq i\leq m}$, the regret of `OMD` up to time $m \in \mathbb{N}$ can be bounded as*

$$\mathrm{Reg}^{(m)} \leq \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \,\|\, \boldsymbol{x}^{(0)}) + \eta\sum_{i=1}^{m}\|\boldsymbol{u}^{(i)} - \boldsymbol{u}^{(i-1)}\|_*^2 - \frac{1}{2\eta}\sum_{i=1}^{m}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2 - \frac{1}{2\eta}\sum_{i=1}^{m}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2,$$

where $\mathring{\boldsymbol{x}} \in \mathcal{X}$ is an optimal strategy in hindsight; $(\boldsymbol{x}^{(i)})_{1\leq i\leq m}$ is the primary sequence of strategies of `OMD`, while $(\hat{\boldsymbol{x}}^{(i)})_{1\leq i\leq m}$ is the secondary sequence of strategies of `OMD`.

Using this refined bound, the guarantee below follows analogously to Proposition C.3. For convenience, we will use the notation $\boldsymbol{z} \coloneqq (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y} =: \mathcal{Z}$ in order to concatenate the strategies of the two players. The primary importance of the following refinement is that the second-order path length bound is now parameterized in terms of *any Nash equilibrium* of the game.

Corollary C.7 (Refinement of Proposition C.3)**.** *In the setting of Proposition C.3,*

$$\sum_{i=1}^{m}\|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i)}\|_2^2 + \sum_{i=1}^{m}\|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i-1)}\|_2^2 \leq 2\|\boldsymbol{z}^{(t,\star)} - \boldsymbol{z}^{(t,0)}\|_2^2,$$

for any $\boldsymbol{z}^{(t,\star)} \in \mathcal{Z}^{(t,\star)}$—the set of Nash equilibria of the BSPP associated with $\mathbf{A}^{(t)}$.

Proof. Let $\mathrm{Reg}^{(t,m)}(\boldsymbol{z}) = \mathrm{Reg}_x^{(t,m)}(\boldsymbol{x}) + \mathrm{Reg}_y^{(t,m)}(\boldsymbol{y})$, where $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}$ and a fixed task $t \in [\![T]\!]$. We claim that $\mathrm{Reg}^{(t,m)}(\boldsymbol{z}^{(t,\star)}) \geq 0$ for any Nash equilibrium pair $\boldsymbol{z}^{(t,\star)} \in \mathcal{Z}^{(t,\star)}$. Indeed, for any step $i \in [\![m]\!]$,

$$(\boldsymbol{x}^{(t,i)})^\top \mathbf{A}^{(t)} \boldsymbol{y}^{(t,i)} - (\boldsymbol{x}^{(t,\star)})^\top \mathbf{A}^{(t)} \boldsymbol{y}^{(t,i)} + (\boldsymbol{x}^{(t,i)})^\top \mathbf{A}^{(t)} \boldsymbol{y}^{(t,\star)} - (\boldsymbol{x}^{(t,i)})^\top \mathbf{A}^{(t)} \boldsymbol{y}^{(t,i)} \geq 0. \tag{20}$$

Here, if $v^{(t)}$ is the value of the game associated with $\mathbf{A}^{(t)}$, which exists by the minimax theorem, we used the fact that $(\boldsymbol{x}^{(t,\star)})^\top \mathbf{A}^{(t)} \boldsymbol{y} \leq v^{(t)}$, for any equilibrium strategy $\boldsymbol{x}^{(t,\star)}$ for player x and any $\boldsymbol{y} \in \mathcal{Y}$, and similarly, $\boldsymbol{x}^\top \mathbf{A}^{(t)} \boldsymbol{y}^{(t,\star)} \geq v^{(t)}$, for any equilibrium strategy $\boldsymbol{y}^{(t,\star)}$ for player y and any $\boldsymbol{x} \in \mathcal{X}$. Thus, summing (20) for all $i \in [\![m]\!]$ verifies our claim: $\mathrm{Reg}^{(t,m)}(\boldsymbol{z}^{(t,\star)}) \geq 0$ for any $\boldsymbol{z}^{(t,\star)} \in \mathcal{Z}^{(t,\star)}$. Finally, the statement follows analogously to Proposition C.3. □

Next, we combine this guarantee with Proposition B.4 to obtain an iteration-complexity bound that depends on the worst-case similarity of the equilibria, defined as

$$V_{\mathrm{NE}}^2 \coloneqq \max_{\boldsymbol{z}^{(1,\star)},\ldots,\boldsymbol{z}^{(T,\star)}} \min_{\bar{\boldsymbol{z}} \in \mathcal{Z}} \sum_{t=1}^{T} \|\boldsymbol{z}^{(t,\star)} - \bar{\boldsymbol{z}}\|_2^2, \tag{21}$$

subject to the constraint that $\boldsymbol{z}^{(t,\star)} \in \mathcal{Z}^{(t,\star)}$ for each task $t \in [\![T]\!]$. We recall that the set of Nash equilibria in BSSPs is nonempty convex and compact, and so (21) is indeed well-defined. It is worth noting that in *generic zero-sum games*—roughly speaking, any zero-sum game perturbed with random noise—there is a unique Nash equilibrium [van Damme, 1987]. In that case, there are no equilibrium selection issues, and (21) happens to reduce to the smallest possible deviation of the Nash equilibria.

Theorem C.8 (Detailed Version of Theorem 3.2)**.** *Suppose that both players employ* `OGD` *with learning rate* $\eta \leq \frac{1}{4L}$*, where* $L \coloneqq \max_{1 \leq t \leq T} \|\mathbf{A}^{(t)}\|_2$*, and initialization* $\boldsymbol{z}^{(t,0)} \coloneqq \sum_{s<t} \boldsymbol{z}^{(t,\star)}/(t-1)$*, for any* $\boldsymbol{z}^{(t,\star)} \in \mathcal{Z}^{(t,\star)}$ *and* $t \geq 2$*. Then, for an average game* $t \in [\![T]\!]$

$$\left\lceil \frac{2V_{NE}^2}{\epsilon^2} + \frac{8(1+\log T)}{T\epsilon^2} \left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2\right) \right\rceil$$

iterations suffice to reach an $\epsilon \left(\frac{2\max\{\Omega_{\mathcal{X}}, \Omega_{\mathcal{Y}}\}}{\eta} + \|\mathbf{A}^{(t)}\|_2 \right)$*-approximate Nash equilibrium, where* V_{NE}^2 *is defined as in* (21)*.*

Proof. First, using Corollary C.7 for each game $t \in [\![T]\!]$,

$$\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{m} \|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i)}\|_2^2 + \sum_{i=1}^{m} \|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i-1)}\|_2^2 \right) \leq \frac{2}{T} \sum_{t=1}^{T} \|\boldsymbol{z}^{(t,\star)} - \boldsymbol{z}^{(t,0)}\|_2^2, \tag{22}$$

for any sequence of Nash equilibria $(\boldsymbol{z}^{(t,\star)})_{1 \leq t \leq T}$. Thus, analogously to Theorem B.3, the right-hand side of (22) can be in turn upper bounded by

$$2V_{\mathrm{NE}}^2 + \frac{8(1+\log T)}{T} \left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2\right).$$

As a result, for an average game

$$\left\lceil \frac{2V_{\mathrm{NE}}^2}{\epsilon^2} + \frac{8(1+\log T)}{T\epsilon^2} \left(\Omega_{\mathcal{X}}^2 + \Omega_{\mathcal{Y}}^2\right) \right\rceil$$

iterations suffice to reach an iterate such that $\|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i)}\|_2, \|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i-1)}\|_2 \leq \epsilon$, which in turn implies that $\boldsymbol{z}^{(t,i)}$ is an $\epsilon \left(\frac{2\max\{\Omega_{\mathcal{X}}, \Omega_{\mathcal{Y}}\}}{\eta} + \|\mathbf{A}^{(t)}\|_2 \right)$-approximate Nash equilibrium [Anagnostides et al., 2022b, Claim A.14]. □

Here, we have assumed that after the termination of each game the players obtain an exact Nash equilibrium of that game—in order to implement the initialization of Theorem C.8. If that is not the case, players have still learned an $O(1/m)$-approximate Nash equilibrium of the game after m iterations of learning. So, Theorem C.8 can be modified by considering the task similarity metric (21) with respect to $O(1/m)$-approximate Nash equilibrium (Definition C.5) (instead of exact Nash equilibria).

C.1.2 Improving the Task Similarity

The notion of task similarity used in Theorem C.8, namely (21), depends on the sequence of the worst Nash equilibria of the games. We can improve those guarantees under the assumption that the players observe the game $\mathcal{G}^{(t)}$ at the end of each task. In particular, each player can employ a meta-learning algorithm that observes as loss after task t the function

$$\boldsymbol{x}^{(t)} \mapsto \frac{1}{2} \min_{\boldsymbol{x}^{(t,\star)} \in \mathcal{X}^{(t,\star)}} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t,\star)}\|_2^2, \tag{23}$$

where $\mathcal{X}^{(t,\star)}$ is the set of Nash equilibria of game $\mathcal{G}^{(t)}$ projected to $\mathcal{X}$. The function (23) is easily seen to be convex, and its gradient can be computed if we have a (Euclidean) projection oracle for the set $\mathcal{X}^{(t,\star)}$. So, the meta-learning algorithm will perform at least as good as

$$\min_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{2} \sum_{t=1}^{T} \min_{\boldsymbol{x}^{(t,\star)} \in \mathcal{X}^{(t,\star)}} \|\boldsymbol{x} - \boldsymbol{x}^{(t,\star)}\|_2^2, \tag{24}$$

modulo an $o(T)$ additive term (from the regret bound). Similar reasoning applies for player y.

An illustrative example To demonstrate the difference between the notions of task similarity—based on Nash equilibria—we have considered so far, we study a simple sequence of games. In particular, we let

$$\mathcal{G} \coloneqq \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{G}' \coloneqq \begin{bmatrix} 1.1 & -1.1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{25}$$

be two zero-sum games in normal form described by their payoff matrix. Then, for $t \in [\![T]\!]$, we let

$$\mathcal{G}^{(t)} \coloneqq \begin{cases} \mathcal{G} & \text{if } t \mod 2 = 1, \\ \mathcal{G}' & \text{if } t \mod 2 = 0. \end{cases}$$

The point of this example is that the games $\mathcal{G}$ and $\mathcal{G}'$ have a common Nash equilibrium, but also different ones:

Claim C.9. *The strategy $((0,0,1),(0,0,1))$ is a Nash equilibrium in both $\mathcal{G}$ and $\mathcal{G}'$, defined in (25). On the other hand, the pair of strategies $((\frac{1}{2},\frac{1}{2},0),(\frac{1}{2},\frac{1}{2},0))$ is a Nash equilibrium for $\mathcal{G}$, but it is 0.05-far from being a Nash equilibrium in $\mathcal{G}'$. Conversely, the pair of strategies $((\frac{10}{21},\frac{11}{21},0),(\frac{1}{2},\frac{1}{2},0))$ is a Nash equilibrium for $\mathcal{G}'$, but it is $\frac{1}{21}$-far from being a Nash equilibrium for $\mathcal{G}$.*

Here, we say that a pair of strategies is α-far from being a Nash equilibrium if there exists a unilateral deviation with (additve) benefit at least $\alpha > 0$ for that player. This simple example shows that the task similarity based on (24) can be 0, while the improved task similarity (21) can be $\Omega(1)$.

C.1.3 Further Refinements

Moreover, we obtain further refinements when either the strategy set of each player corresponds to a probability simplex, or when the game is strongly convex-concave. In particular, let us treat the more general min-max optimization problem where $f(\boldsymbol{x}, \boldsymbol{y})$ is an L-smooth, convex-concave and differentiable function; a more general setting that encompasses such problems is discussed in more detail in Appendix C.2. We begin by noting the following general property of OMD. For convenience, we use a prediction based on the secondary sequence of OMD: $\boldsymbol{m}_x^{(i)} \coloneqq -\nabla_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}^{(i-1)}, \hat{\boldsymbol{y}}^{(i-1)})$ and $\boldsymbol{m}_y^{(i)} \coloneqq \nabla_{\boldsymbol{y}} f(\hat{\boldsymbol{x}}^{(i-1)}, \hat{\boldsymbol{y}}^{(i-1)})$.

Proposition C.10 (OMD Approaches the Set of NE)**.** *For any learning rate $\eta \leq \frac{1}{4L}$ and for any iteration $i \in [\![m]\!]$, OMD with $\boldsymbol{m}_x^{(i)} \coloneqq -\nabla_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}^{(i-1)}, \hat{\boldsymbol{y}}^{(i-1)})$ and $\boldsymbol{m}_y^{(i)} \coloneqq \nabla_{\boldsymbol{y}} f(\hat{\boldsymbol{x}}^{(i-1)}, \hat{\boldsymbol{y}}^{(i-1)})$ satisfies*

$$\mathcal{B}_{\mathcal{R}_x}(\boldsymbol{x}^\star \,\|\, \hat{\boldsymbol{x}}^{(i-1)}) + \mathcal{B}_{\mathcal{R}_y}(\boldsymbol{y}^\star \,\|\, \hat{\boldsymbol{y}}^{(i-1)}) - \mathcal{B}_{\mathcal{R}_x}(\boldsymbol{x}^\star \,\|\, \hat{\boldsymbol{x}}^{(i)}) - \mathcal{B}_{\mathcal{R}_y}(\boldsymbol{y}^\star \,\|\, \hat{\boldsymbol{y}}^{(i)}) \geq 0, \tag{26}$$

for any Nash equilibrium pair $(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \in \mathcal{X} \times \mathcal{Y}$. In particular, equality in (26) holds if and only $\hat{\boldsymbol{x}}^{(i)} = \hat{\boldsymbol{x}}^{(i-1)}$ and $\hat{\boldsymbol{y}}^{(i)} = \hat{\boldsymbol{y}}^{(i-1)}$.

Interestingly, this property follows by relying on the analysis of the RVU bound (Theorem B.1), but only for a single iteration of OMD.

Proof of Proposition C.10. First, for player x we have that for any $\boldsymbol{x}^\star \in \mathcal{X}$, the term $\langle \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle$ is upper bounded by

$$\frac{1}{\eta}\left(\mathcal{B}_{\mathcal{R}_x}(\boldsymbol{x}^\star \,\|\, \hat{\boldsymbol{x}}^{(i-1)}) - \mathcal{B}_{\mathcal{R}_x}(\boldsymbol{x}^\star \,\|\, \hat{\boldsymbol{x}}^{(i)})\right) + \langle \hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}, \nabla_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}^{(i-1)}, \hat{\boldsymbol{y}}^{(i-1)}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle$$
$$- \frac{1}{2\eta}\left(\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2 + \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2\right). \tag{27}$$

Similarly, we have that for any $\boldsymbol{y}^\star \in \mathcal{Y}$ the term $\langle \boldsymbol{y}^\star - \boldsymbol{y}^{(i)}, \nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle$ is upper bounded by

$$\frac{1}{\eta}\left(\mathcal{B}_{\mathcal{R}_y}(\boldsymbol{y}^\star \,\|\, \hat{\boldsymbol{y}}^{(i-1)}) - \mathcal{B}_{\mathcal{R}_y}(\boldsymbol{y}^\star \,\|\, \hat{\boldsymbol{y}}^{(i)})\right) + \langle \hat{\boldsymbol{y}}^{(i)} - \boldsymbol{y}^{(i)}, \nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) - \nabla_{\boldsymbol{y}} f(\hat{\boldsymbol{x}}^{(i-1)}, \hat{\boldsymbol{y}}^{(i-1)}) \rangle$$
$$- \frac{1}{2\eta}\left(\|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i-1)}\|^2 + \|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)}\|^2\right). \tag{28}$$

As a result, for $\eta \leq \frac{1}{4L}$ it follows that $\langle \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle + \langle \boldsymbol{y}^\star - \boldsymbol{y}^{(i)}, \nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle$ is upper bounded by

$$\frac{1}{\eta}\left(\mathcal{B}_{\mathcal{R}_x}(\boldsymbol{x}^\star \,\|\, \hat{\boldsymbol{x}}^{(i-1)}) - \mathcal{B}_{\mathcal{R}_x}(\boldsymbol{x}^\star \,\|\, \hat{\boldsymbol{x}}^{(i)})\right) + \frac{1}{\eta}\left(\mathcal{B}_{\mathcal{R}_y}(\boldsymbol{y}^\star \,\|\, \hat{\boldsymbol{y}}^{(i-1)}) - \mathcal{B}_{\mathcal{R}_y}(\boldsymbol{y}^\star \,\|\, \hat{\boldsymbol{y}}^{(i)})\right)$$
$$-\frac{1}{4\eta}\left(\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2 + \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2\right) - \frac{1}{4\eta}\left(\|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i-1)}\|^2 + \|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)}\|^2\right),$$

due to (27) and (28). Finally, the proof follows since $\langle \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle + \langle \boldsymbol{y}^\star - \boldsymbol{y}^{(i)}, \nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle \geq f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^\star) - f(\boldsymbol{x}^\star, \boldsymbol{y}^{(i)}) \geq 0$, by convexity-concavity of f along with the fact that $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is assumed to be a Nash equilibrium. □

While this proposition guarantees that OMD approaches the set of Nash equilibria—in the sense of (26), the improvement could be arbitrarily small. So, below we impose further structure on the problem.

First, we assume that the objective of player x is μ-strongly convex when the strategy of player y is fixed, and the objective of player y is μ-strongly concave when the strategy of player x is fixed. Under that assumption, (26) can be further strengthened in that the improvement is not only strict, but increases with the modulus of strong convexity.

Corollary C.11. *Let $\mu > 0$ be the modulus of strong convexity and strong concavity of $f(\cdot, \boldsymbol{y})$ and $f(\boldsymbol{x}, \cdot)$, respectively. Then, in the setting of Proposition C.10 with $\eta \leq \min\left\{\frac{1}{4L}, \frac{1}{2\mu}\right\}$,*

$$\left(\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 + \|\boldsymbol{y}^\star - \hat{\boldsymbol{y}}^{(i-1)}\|_2^2\right) \geq \left(1 + \frac{\mu}{2}\right)\left(\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}^{(i)}\|_2^2 + \|\boldsymbol{y}^\star - \hat{\boldsymbol{y}}^{(i)}\|_2^2\right), \tag{29}$$

for any Nash equilibrium pair $(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \in \mathcal{X} \times \mathcal{Y}$.

In proof, the key difference is that now $\langle \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\rangle \geq f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) - f(\boldsymbol{x}^\star, \boldsymbol{y}^{(i)}) + \frac{\mu}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}^{(i)}\|_2^2$, by μ-strong convexity of $f(\cdot, \boldsymbol{y}^{(i)})$. Further, $\frac{1}{4\eta}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|_2^2 + \frac{\mu}{2}\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^\star\|_2^2 \geq \frac{\mu}{4}\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^\star\|_2^2$, by Young's inequality and the fact that $\eta \leq \frac{1}{2\mu}$. Similar reasoning applies for player y, leading to Corollary C.11.

Using (29) inductively yields that

$$\left(\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}^{(m)}\|_2^2 + \|\boldsymbol{y}^\star - \hat{\boldsymbol{y}}^{(m)}\|_2^2\right) \leq \left(\frac{1}{1 + \mu/2}\right)^m \left(\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}^{(0)}\|_2^2 + \|\boldsymbol{y}^\star - \hat{\boldsymbol{y}}^{(0)}\|_2^2\right),$$

for any Nash equilibrium pair $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, in turn implying that $(\hat{\boldsymbol{x}}^{(m)}, \hat{\boldsymbol{y}}^{(m)})$ converges to the projection of $\mathcal{Z}^\star$ with a linear rate of $1/(1 + \mu/2) \in (0, 1)$. Perhaps surprisingly, linear rate is also achievable without any strong convexity assumptions in games with polyhedral sets [Wei et al., 2021, Theorem 8], although the rate there depends on condition number-like quantities and can be arbitrarily slow even in 2×2 games.

From a meta-learning standpoint, that property of `OMD`—converging to the projection to the set of Nash equilibria—is useful, and can partially address the bad example we show earlier in (25). In particular, as long as the number of iterations is large enough, the dynamics will project to the set of Nash equilibria of $\mathcal{G}$ and $\mathcal{G}'$ in tandem, gradually approaching the common Nash equilibrium. Yet, if the "angle" between those sets is small it would take a large number of tasks T so that the dynamics reach close to the common Nash equilibrium.

Remark C.12 (Alternating Updates)**.** *So far, we have studied the setting where both players update their strategies simultaneously—as in the definition of `OMD`. A different approach that has received extensive attention in the literature [Wibisono et al., 2022], not least due to its practical superiority [Tammelin, 2014], consists of performing the update rule in an alternating fashion. Interestingly, within the framework of optimistic mirror descent, alternation can be captured through the predictions: the first player uses the standard optimistic prediction (if any), but the player who updates second has a perfect prediction; that is, the prediction does not correspond to the previous strategy of the opponent, but the current one. All of our guarantees immediately apply under alternating updates using this simple observation.*

C.2 Beyond Bilinear Saddle-Point Problems: A VI Perspective

In this subsection, we extend our scope beyond the bilinear saddle-point problems of (19). In particular, we take a broader variational inequality (VI) perspective, which is commonly espoused in the context of min-max optimization.

Let us suppose that $F : \mathcal{Z} \to \mathcal{Z}$ is a singe-valued operator; many of our results below readily apply even if F is multi-valued, such as the subdifferential operator of a non-differentiable function. We begin by making some standard assumptions on the operator F below. We will then explain how the assumptions below immediately capture BSSPs. We remark that both of those assumptions will be weakened in the sequel: Lipschitz continuity (Assumption C.13) is relaxed in Appendix C.6 where we only assume Hölder continuity, while the MVI property (Assumption C.14) is relaxed in Appendix C.2.1 where we consider the so-called weak MVI property.

Assumption C.13. *F is L-Lipschitz continuous, in the sense that for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$, it holds that $\|F(\boldsymbol{z}) - F(\boldsymbol{z}')\|_2 \leq L\|\boldsymbol{z} - \boldsymbol{z}'\|_2$.*

Assumption C.14 (MVI Property [Mertikopoulos et al., 2019, Gidel et al., 2019])**.** *There exists a point $\boldsymbol{z}^\star \in \mathcal{Z}$ such that $\langle F(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{z}^\star \rangle \geq 0$ for any $\boldsymbol{z} \in \mathcal{Z}$.*

To relate those assumptions to the bilinear saddle-point problem we considered earlier in Appendix C.1, we take $F : \boldsymbol{z} \mapsto (\mathbf{A}\boldsymbol{y}, -\mathbf{A}^\top \boldsymbol{x})$. Then, given that $\langle \boldsymbol{z}, F(\boldsymbol{z}) \rangle = 0$ for any $\boldsymbol{z} \in \mathcal{Z}$, Assumption C.14 requests the existence of a point $(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \in \mathcal{Z}$ so that $\boldsymbol{x}^\top \mathbf{A} \boldsymbol{y}^\star - (\boldsymbol{x}^\star)^\top \mathbf{A} \boldsymbol{y} \geq 0$, for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, which is a well-known consequence of the minimax theorem; Assumption C.13 also follows immediately for BSPPs.

More broadly, Assumptions C.13 and C.14 induce a standard setup in min-max optimization, where $F \coloneqq (\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}))$ for certain differentiable and smooth objective functions $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Importantly, our techniques in Appendix C.1 can be readily applied to this more general setting. In particular, Corollary C.7 can be cast as follows.

Corollary C.15. *Consider an operator $F : \mathcal{Z} \to \mathcal{Z}$ that satisfies Assumptions C.13 and C.14. Then, under* `OGD` *with learning rate $\eta \leq \frac{1}{4L}$,*

$$\sum_{i=1}^{m} \|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i)}\|_2^2 + \sum_{i=1}^{m} \|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i-1)}\|_2^2 \leq 2\|\boldsymbol{z}^{(t,\star)} - \boldsymbol{z}^{(t,0)}\|_2^2,$$

where $\boldsymbol{z}^{(t,\star)} \in \mathcal{Z}$ is any point that satisfies the MVI property (Assumption C.14).

To analyze `OGD` in this more general setup, we define the regret up to time $m \in \mathbb{N}$ as

$$\mathrm{Reg}_{\mathcal{L}}^{(m)} \coloneqq \max_{\boldsymbol{z}^\star \in \mathcal{Z}} \left\{ \sum_{i=1}^{m} \langle \boldsymbol{z}^{(i)} - \boldsymbol{z}^\star, F(\boldsymbol{z}^{(i)}) \rangle \right\}. \tag{30}$$

When $F \coloneqq (\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}))$, (30) is precisely the sum of the "linearized" regrets incurred by the two players.

Proof of Corollary C.15. First, analogously to Theorem C.6, it follows that

$$\begin{aligned}\mathrm{Reg}_{\mathcal{L}}^{(m)}(\boldsymbol{z}^{(t,\star)}) \leq \frac{1}{2\eta}\|\boldsymbol{z}^{(t,\star)} - \boldsymbol{z}^{(t,0)}\|_2^2 + \eta \sum_{i=1}^{m} \|F(\boldsymbol{z}^{(t,i)}) - F(\boldsymbol{z}^{(t,i-1)})\|_2^2 \\ - \frac{1}{2\eta} \sum_{i=1}^{m} \left(\|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i)}\|_2^2 + \|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i-1)}\|_2^2 \right).\end{aligned}$$

Next, using the L-Lipschitz continuity of F (Assumption C.13), we have that for $\eta \leq \frac{1}{4L}$,

$$\mathrm{Reg}_{\mathcal{L}}^{(m)}(\boldsymbol{z}^{(t,\star)}) \leq \frac{1}{2\eta}\|\boldsymbol{z}^{(t,\star)} - \boldsymbol{z}^{(t,0)}\|_2^2 - \frac{1}{4\eta} \sum_{i=1}^{m} \left(\|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i)}\|_2^2 + \|\boldsymbol{z}^{(t,i)} - \hat{\boldsymbol{z}}^{(t,i-1)}\|_2^2 \right). \tag{31}$$

But, if a point $\boldsymbol{z}^{(t,\star)}$ satisfies the MVI property (Assumption C.14), it follows that

$$\langle \boldsymbol{z}^{(t,i)} - \boldsymbol{z}^{(t,\star)}, F(\boldsymbol{z}^{(t,i)}) \rangle \geq 0,$$

for any iteration $i \in [\![m]\!]$, and summing over all $i \in [\![m]\!]$ implies that $\mathrm{Reg}_{\mathcal{L}}^{(m)}(\boldsymbol{z}^{(t,\star)}) \geq 0$. Thus, the result follows immediately by rearranging (31). □

Before we proceed with the analog of Theorem C.8, let us first clarify our solution concept. We will say that a point $\boldsymbol{z}^\star \in \mathcal{Z}$ is an ϵ-approximate solution to the Stampacchia variational inequality (SVI) problem if

$$\langle \boldsymbol{z} - \boldsymbol{z}^\star, F(\boldsymbol{z}^\star) \rangle \geq -\epsilon, \quad \forall \boldsymbol{z} \in \mathcal{Z}. \tag{SVI}$$

We point out that when F is a monotone operator, meaning that $\langle F(\boldsymbol{z}) - F(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \rangle \geq 0$ for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$, then an ϵ-approximate solution to (SVI) also satisfies $\langle \boldsymbol{z} - \boldsymbol{z}^\star, F(\boldsymbol{z}) \rangle \geq -\epsilon$, for any $\boldsymbol{z} \in \mathcal{Z}$; a point $\boldsymbol{z}^\star$ satisfying the later property is referred to as an ϵ-approximate *weak* solution to the variational inequality problem. We will use the following property, which follows analogously to [Anagnostides et al., 2022b, Claim A.14].

Claim C.16. *Suppose that the sequences $(\boldsymbol{z}^{(i)})_{0 \leq i \leq m}$ and $(\hat{\boldsymbol{z}}^{(i)})_{0 \leq i \leq m}$ are updated using* OGD *with learning rate $\eta > 0$. If $\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2, \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2 \leq \epsilon$ for some iteration $i \in [\![m]\!]$, then $\boldsymbol{z}^{(i)} \in \mathcal{Z}$ is an $\epsilon \left(\frac{2\Omega_{\mathcal{Z}}}{\eta} + \|F(\boldsymbol{z}^{(i)})\|_2 \right)$-approximate solution to the (SVI) problem.*

Now in a complete analogy to Equation (21), we will use the following basic notion of task similarity.

$$V_{\mathrm{MVI}}^2 \coloneqq \sup_{\boldsymbol{z}^{(1,\star)}, \dots, \boldsymbol{z}^{(T,\star)}} \min_{\bar{\boldsymbol{z}} \in \mathcal{Z}} \sum_{t=1}^{T} \|\boldsymbol{z}^{(t,\star)} - \bar{\boldsymbol{z}}\|_2^2, \tag{32}$$

where $\boldsymbol{z}^{(1,\star)}, \dots, \boldsymbol{z}^{(T,\star)}$ are constrained to be in the (nonempty) set of points that satisfy the MVI property for the corresponding game.

Theorem C.17 (Extension of Theorem C.8)**.** *Consider an operator $F : \mathcal{Z} \to \mathcal{Z}$ that satisfies Assumptions C.13 and C.14. Suppose further that we employ* OGD *with learning rate $\eta \leq \frac{1}{4L}$, where $L \coloneqq \max_{1 \leq t \leq T} L^{(t)}$, and initialization $\boldsymbol{z}^{(t,0)} \coloneqq \sum_{s<t} \boldsymbol{z}^{(t,\star)}/(t-1)$, for any $\boldsymbol{z}^{(t,\star)} \in \mathcal{Z}^{(t,\star)}$ and $t \geq 2$, where $\emptyset \neq \mathcal{Z}^{(t,\star)} \subseteq \mathcal{Z}$ is the set of points that satisfy the MVI property (Assumption C.14). Then, for an average game $t \in [\![T]\!]$*

$$\left\lceil \frac{2V_{MVI}^2}{\epsilon^2} + \frac{8(1+\log T)}{T\epsilon^2} \Omega_{\mathcal{Z}}^2 \right\rceil$$

iterations suffice to reach an $\epsilon \left(\frac{2\Omega_{\mathcal{Z}}}{\eta} + \|F(\boldsymbol{z}^{(t,i)})\|_2 \right)$-approximate solution to the (SVI), where V_{MVI}^2 is defined as in (32).

It is also worth pointing out how to obtain in this setting an improved guarantee for the average sequence of OGD. Let $\boldsymbol{z}^\star \in \mathcal{Z}$. Assuming that F is a monotone operator, the average regret—in the sense of (30)—can be lower bounded as

$$\frac{1}{m} \mathrm{Reg}_{\mathcal{L}}^{(m)}(\boldsymbol{z}^\star) = \frac{1}{m} \sum_{i=1}^{m} \langle \boldsymbol{z}^{(i)} - \boldsymbol{z}^\star, F(\boldsymbol{z}^{(i)}) \rangle \geq \frac{1}{m} \sum_{i=1}^{m} \langle \boldsymbol{z}^{(i)} - \boldsymbol{z}^\star, F(\boldsymbol{z}^\star) \rangle = \langle \bar{\boldsymbol{z}} - \boldsymbol{z}^\star, F(\boldsymbol{z}^\star) \rangle,$$

where $\bar{\boldsymbol{z}} \coloneqq \sum_{i=1}^{m} \boldsymbol{z}^{(i)}/m$. Further, assuming that F is Lipchitz continuous (Assumption C.13), following the proof of Theorem C.17 we have that the average regret decays as $O(1/m)$ (assuming that $\mathcal{Z}$ is bounded). In other words, assuming monotonicity we have that

$$\langle \bar{\boldsymbol{z}} - \boldsymbol{z}, F(\boldsymbol{z}) \rangle \leq O\left(\frac{1}{m} \right), \quad \forall \boldsymbol{z} \in \mathcal{Z}.$$

That is, $\bar{\boldsymbol{z}}$ is an $O(1/m)$-approximate weak solution to the variational inequality problem.

C.2.1 Weak MVI Property

Nevertheless, there are important problems for which the MVI property fails. For that reason, Diakonikolas et al. [2021] introduced a weaker property, recalled below.

Assumption C.18 (Weak MVI Property)**.** *The operator $F : \mathcal{Z} \to \mathcal{Z}$ is said to satisfy the weak MVI property if there exists $\boldsymbol{z}^\star \in \mathcal{Z}$ such that*

$$\langle F(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{z}^\star \rangle \geq -\frac{\rho}{2}\|F(\boldsymbol{z})\|_2^2, \quad \forall \boldsymbol{z} \in \mathcal{Z}, \tag{33}$$

for a sufficiently small parameter $\rho > 0$.

We note that we use the ℓ_2-norm in the right-hand side of (33) for convenience, although that definition can be stated more broadly. This is related to a condition known in the literature as *cohypomonotonicity* [Bauschke et al., 2021, Combettes and Pennanen, 2004].

Definition C.19 (Cohypomonotonicity)**.** *We say that an operator $F : \mathcal{Z} \to \mathcal{Z}$ is ρ-cohypomonotone, with $\rho > 0$, if for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$,*

$$\langle \boldsymbol{z} - \boldsymbol{z}', F(\boldsymbol{z}) - F(\boldsymbol{z}') \rangle \geq -\rho\|F(\boldsymbol{z}) - F(\boldsymbol{z}')\|_2^2. \tag{34}$$

It is evident that (34) recovers the notion of monotonicity when $\rho = 0$, but cohypomonotonicity is a non-trivial extension of monotonicity. An operator that satisfies (34) is also sometimes referred to as $-\rho$-*comonotone*.

As in [Diakonikolas et al., 2021], here we focus on the unconstrained setting: $\mathcal{Z} := \mathbb{R}^d$, for some $d \in \mathbb{N}$. In that case, `OGD` can be simplified as

$$\begin{aligned}
\boldsymbol{z}^{(i)} &:= \mathrm{Proj}_{\mathcal{Z}}\left(\hat{\boldsymbol{z}}^{(i-1)} - \eta F(\boldsymbol{z}^{(i-1)})\right) = \hat{\boldsymbol{z}}^{(i-1)} - \eta F(\boldsymbol{z}^{(i-1)}),\\
\hat{\boldsymbol{z}}^{(i)} &:= \mathrm{Proj}_{\mathcal{Z}}\left(\hat{\boldsymbol{z}}^{(i-1)} - \eta F(\boldsymbol{z}^{(i)})\right) = \hat{\boldsymbol{z}}^{(i-1)} - \eta F(\boldsymbol{z}^{(i)}).
\end{aligned}$$

In this context, below we establish a bound on the number of iterations requires to make the norm of the operator at most ϵ. We point out that in the special case of min-max optimization, $F(\boldsymbol{z}) = (\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}))$, we have that $\|F(\boldsymbol{z})\|_2^2 = \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\|_2^2 + \|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})\|_2^2$, which is perhaps the most natural measure of convergence in unconstrained min-max optimization.

Theorem C.20 (`OGD` under the weak MVI)**.** *If $F : \mathcal{Z} \to \mathcal{Z}$ is L-Lipschitz continuous (Assumption C.13) and satisfies the weak MVI property with parameter $0 < \rho < \frac{1}{8L}$ (Assumption C.18), then for $2\rho < \eta < \frac{1}{4L}$, `OGD` satisfies*

$$\sum_{i=1}^{m-1} \|F(\boldsymbol{z}^{(i)})\|_2^2 \leq \frac{2}{\eta(\eta - 2\rho)}\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2 + \frac{2\rho}{\eta - 2\rho}\|F(\boldsymbol{z}^{(m)})\|_2^2,$$

where $\boldsymbol{z}^\star \in \mathcal{Z}$ is any point that satisfies the weak MVI property. In particular, there is an iterate $\boldsymbol{z}^{(i)} \in \mathcal{Z}$ such that

$$\|F(\boldsymbol{z}^{(i)})\|_2 \leq \frac{1}{\sqrt{m-1}}\sqrt{\frac{2}{\eta(\eta - 2\rho)}\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2 + \frac{2\rho}{\eta - 2\rho}\|F(\boldsymbol{z}^{(m)})\|_2^2}.$$

Proof. Similarly to the proof of Corollary C.15, it follows that

$$\mathrm{Reg}_{\mathcal{L}}^{(m)}(\boldsymbol{z}^\star) \leq \frac{1}{2\eta}\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2 - \frac{1}{4\eta}\sum_{i=1}^{m}\left(\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2^2 + \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^2\right). \tag{35}$$

But, the weak MVI property (Assumption C.18) implies that there exists $\boldsymbol{z}^\star \in \mathcal{Z}$ such that

$$\mathrm{Reg}_{\mathcal{L}}^{(m)}(\boldsymbol{z}^\star) \geq \sum_{i=1}^{m}\langle \boldsymbol{z}^{(i)} - \boldsymbol{z}^\star, F(\boldsymbol{z}^{(i)})\rangle \geq -\frac{\rho}{2}\sum_{i=1}^{m}\|F(\boldsymbol{z}^{(i)})\|_2^2.$$

Thus, combining with (35),

$$0 \leq \frac{1}{2\eta}\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2 + \frac{\rho}{2}\sum_{i=1}^{m}\|F(\boldsymbol{z}^{(i)})\|_2^2 - \frac{\eta}{4}\sum_{i=1}^{m}\|F(\boldsymbol{z}^{(i-1)})\|_2^2,$$

since $\boldsymbol{z}^{(i)} = \hat{\boldsymbol{z}}^{(i-1)} - \eta F(\boldsymbol{z}^{(i-1)})$. Thus,

$$\left(\frac{\eta}{4} - \frac{\rho}{2}\right)\sum_{i=1}^{m-1}\|F(\boldsymbol{z}^{(i)})\|_2^2 \leq \frac{1}{2\eta}\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2 + \frac{\rho}{2}\|F(\boldsymbol{z}^{(m)})\|_2^2.$$

The proof then follows by rearranging, given that $\eta > 2\rho$ (by assumption). □

In light of this theorem, the meta-learning version is completely analogous to Theorem C.17, and it is therefore omitted. An interesting question is to extend our analysis in Theorem C.20 in the constrained setting.

C.3 Weighting the Strategies

A common strategy used in order to compute faster (approximate) saddle-points consists of weighting the players' strategies in a non-uniform way (*e.g.*, [Zhang et al., 2022a, Brown and Sandholm, 2019, Gao et al., 2021]). In this subsection, we formalize the fact that the guarantees we established for the ergodic average (Corollary C.2) can still be applied under a broad class of weighting schemes. To establish this, we first provide a refined analysis of `OMD`, but with the twist that our guarantee will apply for a *weighted* notion of regret. Namely, for a sequence $\alpha^{(1)}, \dots, \alpha^{(m)} \in \mathbb{R}_{>0}$, we define *alpha regret* as

$$\boldsymbol{\alpha}\text{-Reg}^{(m)} := \max_{\mathring{\boldsymbol{x}} \in \mathcal{X}} \left\{ \sum_{i=1}^{m} \alpha^{(i)} \langle \mathring{\boldsymbol{x}} - \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} \rangle \right\}. \tag{36}$$

This generalized notion of regret has been proved useful, for example, in obtaining accelerates rates for the Frank-Wolfe algorithm in convex optimization [Abernethy et al., 2018]; see also [Wang et al., 2021]. We will insist on ensuring that $\sum_{i=1}^{m} \alpha^{(i)} = m$. Typical choices used in the literature include the so-called *linear averaging*, in which $\alpha^{(i)} := \frac{2i}{m+1}$, or *quadratic averaging*, in which $\alpha^{(i)} := \frac{6i^2}{(m+1)(2m+1)}$. Below, we show that `OMD` enjoys similar RVU-type guarantees (Theorem B.1) under a broad family of weighting sequences; this is perhaps unexpected given that `OMD` is designed to minimize (unweighted) regret.

Theorem C.21. *Consider any $\boldsymbol{\alpha} \in \mathbb{R}^m_{>0}$ with $\alpha^{(1)} \leq \cdots \leq \alpha^{(m)}$, and suppose that we use `OMD` with a 1-strongly convex regularizer $\mathcal{R}$ with respect to the norm $\|\cdot\|$. For any observed sequence of utilities $(\boldsymbol{u}^{(i)})_{1 \leq i \leq m}$, the alpha regret $\boldsymbol{\alpha}$-Reg$^{(m)}$ of `OMD` up to time $m \in \mathbb{N}$ and initialized at $\hat{\boldsymbol{x}}^{(0)} \in \mathcal{X}$ can be upper bounded by*

$$\frac{\alpha^{(1)}}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \,\|\, \hat{\boldsymbol{x}}^{(0)}) + \frac{(\alpha^{(m)} - \alpha^{(1)})}{\eta}\max_{1 \leq i \leq m-1}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \,\|\, \hat{\boldsymbol{x}}^{(i)}) + \eta\sum_{i=1}^{m}\alpha^{(i)}\|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|_*^2$$
$$-\frac{1}{2\eta}\sum_{i=1}^{m}\alpha^{(i)}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2 - \frac{1}{2\eta}\sum_{i=1}^{m}\alpha^{(i)}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2,$$

for any $\mathring{\boldsymbol{x}} \in \mathcal{X}$.

The proof of this theorem closely follows the argument due to Rakhlin and Sridharan [2013b], but it is included below for completeness.

Proof. Fix any $\boldsymbol{x}^\star \in \mathcal{X}$. First, by 1-strong convexity of the Bregman divergence (with respect to the first argument), we have that for any iteration $i \in [\![m]\!]$,

$$\langle \boldsymbol{x}^{(i)}, \boldsymbol{m}^{(i)} \rangle - \frac{1}{2\eta}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2 - \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{m}^{(i)} \rangle + \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\hat{\boldsymbol{x}}^{(i)} \parallel \hat{\boldsymbol{x}}^{(i-1)}) \geq \frac{1}{2\eta}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2. \quad (37)$$

Similarly, for any iteration $i \in [\![m]\!]$,

$$\langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i)} \rangle - \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\hat{\boldsymbol{x}}^{(i)} \parallel \hat{\boldsymbol{x}}^{(i-1)}) - \langle \mathring{\boldsymbol{x}}, \boldsymbol{u}^{(i)} \rangle + \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i-1)}) \geq \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i)}). \quad (38)$$

Further, for any step $i \in [\![m]\!]$ it holds that

$$\langle \mathring{\boldsymbol{x}} - \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} \rangle = \langle \mathring{\boldsymbol{x}}, \boldsymbol{u}^{(i)} \rangle - \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i)} \rangle - \langle \boldsymbol{x}^{(i)}, \boldsymbol{m}^{(i)} \rangle + \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{m}^{(i)} \rangle + \langle \hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)} \rangle.$$

Combining with (37) and (38),

$$\begin{aligned}\boldsymbol{\alpha}\text{-Reg}^{(m)}(\mathring{\boldsymbol{x}}) \leq & \frac{1}{\eta}\sum_{i=1}^{m} \alpha^{(i)} \left(\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i-1)}) - \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i)})\right) + \sum_{i=1}^{m} \alpha^{(i)} \langle \hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)} \rangle \\ & - \frac{1}{2\eta}\sum_{i=1}^{m} \alpha^{(i)} \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2 - \frac{1}{2\eta}\sum_{i=1}^{m} \alpha^{(i)} \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2. \quad (39)\end{aligned}$$

It is also easy to see that $\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\| \leq \eta \|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|_*$. So, it suffices to appropriately bound the first term in the right-hand side of (39). We have

$$\begin{aligned}\sum_{i=1}^{m} \alpha^{(i)} \left(\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i-1)}) - \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i)})\right) &= \alpha^{(1)}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(0)}) - \alpha^{(m)}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(m)}) \\ &\quad + \sum_{i=2}^{m} \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i-1)})(\alpha^{(i)} - \alpha^{(i-1)}),\end{aligned}$$

which in turn is upper bounded by

$$\alpha^{(1)}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(0)}) + (\alpha^{(m)} - \alpha^{(1)}) \max_{1 \leq i \leq m-1} \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \hat{\boldsymbol{x}}^{(i)}).$$

Here, we used the fact that $\alpha^{(m)} \geq 0$ and $\mathcal{B}_{\mathcal{R}}(\cdot \parallel \cdot) \geq 0$, as well as the assumption that $\alpha^{(i)} - \alpha^{(i-1)} \geq 0$ for any $2 \leq i \leq m$, which results in the telescopic sum above. Combing with (39) concludes the proof. □

This characterization has some interesting implications. First, it lifts many of the results of Syrgkanis et al. [2015] to any class of monotone weighting schemes (Theorem C.21), even though the underlying algorithm (`OMD`) remains the same. For example, we state the following immediate consequence of the sum of the players' regets.

Corollary C.22. *Suppose that all players employ `OGD` with learning rate $\eta \leq \frac{1}{4L\sqrt{n-1}}$.*

- *If $\alpha^{(i)} := \frac{2i}{m+1}$, then $\frac{1}{m}\sum_{k=1}^{n} \boldsymbol{\alpha}\text{-Reg}_k^{(m)} \leq 4L\sqrt{n-1}\sum_{k=1}^{n} \Omega_{\mathcal{X}_k}^2 \frac{1}{m}$.*
- *If $\alpha^{(i)} := \frac{6i^2}{(m+1)(2m+1)}$, then $\frac{1}{m}\sum_{k=1}^{n} \boldsymbol{\alpha}\text{-Reg}_k^{(m)} \leq 12L\sqrt{n-1}\sum_{k=1}^{n} \Omega_{\mathcal{X}_k}^2 \frac{m^2+1}{m(m+1)(2m+1)}$.*

This is important since, as we explained earlier, linear or quadratic averaging have shown to substantially improve performance in practice. From a meta-learning standpoint, it is possible that assigning different weights to different tasks could improve performance in practice, although this is not pursued in this paper.

C.4 Adaptive Regularization

In this subsection, we extend our scope to `OGD` [Chiang et al., 2012, Rakhlin and Sridharan, 2013b] under adaptive regularization, or *preconditioning*. That is, we will endow `OGD` with an *adaptive* regularizer, leading to `OptAdaGrad`, an extension of the popular `AdaGrad` algorithm [Duchi et al., 2011]. To this end, let $\mathbf{Q}^{(i)} \in \mathbb{R}^{d\times d}$ be a positive definite and symmetric matrix, for any $i \in [\![m]\!]$, that will serve as the *preconditioner*. We also let $\|\boldsymbol{x}\|_{\mathbf{Q}^{(i)}} \coloneqq \sqrt{\boldsymbol{x}^\top \mathbf{Q}^{(i)} \boldsymbol{x}}$ be the (Mahalanobis) norm induced by $\mathbf{Q}^{(i)}$, and $\mathcal{R}^{(i)}(\boldsymbol{x}) \coloneqq \frac{1}{2}\|\boldsymbol{x}\|^2_{\mathbf{Q}^{(i)}}$ be the associated *regularizer*. If we denote by $\mathrm{Proj}^{\mathbf{Q}^{(t)}}_{\mathcal{X}}(\boldsymbol{x}) \coloneqq \arg\min_{\hat{\boldsymbol{x}}\in\mathcal{X}} \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2_{\mathbf{Q}^{(i)}}$, `OptAdaGrad` can be expressed via the following update rule for all $i \in \mathbb{N}$.

$$\begin{aligned} \boldsymbol{x}^{(i)} &\coloneqq \mathrm{Proj}^{\mathbf{Q}^{(i)}}_{\mathcal{X}}\left(\hat{\boldsymbol{x}}^{(i-1)} + (\mathbf{Q}^{(i)})^{-1}\boldsymbol{m}^{(i)}\right), \\ \hat{\boldsymbol{x}}^{(i)} &\coloneqq \mathrm{Proj}^{\mathbf{Q}^{(i)}}_{\mathcal{X}}\left(\hat{\boldsymbol{x}}^{(i-1)} + (\mathbf{Q}^{(i)})^{-1}\boldsymbol{u}^{(i)}\right). \end{aligned} \tag{OptAdaGrad}$$

Further, we define $\boldsymbol{x}^{(0)} = \hat{\boldsymbol{x}}^{(0)} \coloneqq \min_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_2^2$. When $\mathbf{Q}^{(t)} = \frac{1}{\eta}\mathbf{I}_{d\times d}$, for some *learning rate* $\eta > 0$, `OptAdaGrad` reduces to `OGD`. While this algorithm can be impractical in high dimensions, if we restrict the preconditioner $\mathbf{Q}^{(i)}$ to be, for example, a diagonal matrix, the inverse can be computed in linear time. A central theme in `AdaGrad` [Duchi et al., 2011] is that the preconditioner $\mathbf{Q}^{(i)}$ has to vary slowly over time. In our setting, we formalize that requirement in the following condition.

Condition C.23. *Fix some time horizon $m \geq 2$. The sequence of preconditioners $(\mathbf{Q}^{(i)})_{1\leq i\leq m}$ is such that $\sum_{i=1}^{m-1} \|\mathbf{Q}^{(i+1)} - \mathbf{Q}^{(i)}\|_2 \leq \sigma(m)$, for some function $\sigma(m) = o_m(m)$.*

Before we proceed with the regret analysis of `OptAdaGrad`, let us make some observations. First, $\mathcal{R}^{(i)}$ is 1-strongly convex with respect to the norm $\|\cdot\|_{\mathbf{Q}^{(i)}}$ since

$$\mathcal{R}^{(i)}(\boldsymbol{x}) \geq \mathcal{R}^{(i)}(\boldsymbol{x}') + \langle \nabla\mathcal{R}^{(i)}(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\rangle + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2_{\mathbf{Q}^{(i)}},$$

for any $i \in [\![m]\!]$ and $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. In this setup, it will be convenient to express `OptAdaGrad` in the following form for $i \in [\![m]\!]$:

$$\begin{aligned} \boldsymbol{x}^{(i)} &\coloneqq \arg\max_{\boldsymbol{x}\in\mathcal{X}} \left\{ \Phi^{(i)} \coloneqq \langle \boldsymbol{x}, \boldsymbol{m}^{(i)}\rangle - \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}^{(i-1)})^\top \mathbf{Q}^{(i)}(\boldsymbol{x} - \hat{\boldsymbol{x}}^{(i-1)}) \right\}, \\ \hat{\boldsymbol{x}}^{(i)} &\coloneqq \arg\max_{\hat{\boldsymbol{x}}\in\mathcal{X}} \left\{ \Psi^{(i)} \coloneqq \langle \hat{\boldsymbol{x}}, \boldsymbol{u}^{(i)}\rangle - \frac{1}{2}(\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)})^\top \mathbf{Q}^{(i)}(\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}) \right\}. \end{aligned} \tag{40}$$

Theorem C.24. *Suppose that Condition C.23 holds for some function $\sigma(m)$. Then, the regret of `OptAdaGrad` up to time $m \in \mathbb{N}$ can be bounded as*

$$\mathrm{Reg}^{(m)}(\mathring{\boldsymbol{x}}) \leq \frac{1}{2}\|\mathring{\boldsymbol{x}} - \boldsymbol{x}^{(0)}\|^2_{\mathbf{Q}^{(1)}} + \frac{\Omega^2_{\mathcal{X}}}{2}\sigma(m) + \sum_{i=1}^{m} \|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|^2_{*,\mathbf{Q}^{(i)}} - \frac{1}{2}\Sigma^{(m)}, \tag{41}$$

where

$$\Sigma^{(m)} \coloneqq \sum_{i=1}^{m} \left(\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} + \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}} \right).$$

This regret bound differs from the analysis of `AdaGrad` [Duchi et al., 2011] in that we incorporate optimism. Theorem C.24 also goes beyond the analysis of [Rakhlin and Sridharan, 2013b] in that the regularizer varies over time, while we recover the guarantee of `OGD` in the special case where $\sigma(m) = 0$.

Proof. First, given that $\mathcal{R}^{(i)}$ is 1-strongly convex with respect to the norm $\|\cdot\|_{\mathbf{Q}^{(i)}}$, it follows that both $\Phi^{(i)}$ and $\Psi^{(i)}$ (recall the definition in (40)) are also 1-strongly convex with respect to $\|\cdot\|_{\mathbf{Q}^{(i)}}$. In turn, this implies that

$$\langle \boldsymbol{x}^{(i)}, \boldsymbol{m}^{(i)} \rangle - \frac{1}{2}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} - \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{m}^{(i)} \rangle + \frac{1}{2}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} \geq \frac{1}{2}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}}; \quad (42)$$

and

$$\langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i)} \rangle - \frac{1}{2}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} - \langle \mathring{\boldsymbol{x}}, \boldsymbol{u}^{(i)} \rangle + \frac{1}{2}\|\boldsymbol{x}^* - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} \geq \frac{1}{2}\|\hat{\boldsymbol{x}}^{(i)} - \mathring{\boldsymbol{x}}\|^2_{\mathbf{Q}^{(i)}}, \quad (43)$$

for any $\mathring{\boldsymbol{x}} \in \mathcal{X}$. Furthermore, we have that

$$\langle \mathring{\boldsymbol{x}} - \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} \rangle = \langle \mathring{\boldsymbol{x}}, \boldsymbol{u}^{(i)} \rangle - \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i)} \rangle - \langle \boldsymbol{x}^{(i)}, \boldsymbol{m}^{(i)} \rangle + \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{m}^{(i)} \rangle + \langle \hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)} \rangle.$$

As a result, combining this identity with (42) and (43) we get that for any $\mathring{\boldsymbol{x}} \in \mathcal{X}$,

$$\begin{aligned}\mathrm{Reg}^{(m)}(\mathring{\boldsymbol{x}}) \leq \frac{1}{2}\sum_{i=1}^{m}\left(\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} - \|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}}\right) + \sum_{i=1}^{m}\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_{\mathbf{Q}^{(i)}}\|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|_{*,\mathbf{Q}^{(i)}} \\ - \frac{1}{2}\sum_{i=1}^{m}\left(\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} + \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}}\right), \quad (44)\end{aligned}$$

where we also used that $\langle \hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)} \rangle \leq \|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_{\mathbf{Q}^{(i)}}\|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|_{*,\mathbf{Q}^{(i)}}$ (by Cauchy-Schwarz inequality). Next, to further bound (44) we will show the following simple claims.

Claim C.25. *It holds that*

$$\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_{\mathbf{Q}^{(i)}} \leq \|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|_{*,\mathbf{Q}^{(i)}}.$$

Proof. By 1-strong convexity of $\Phi^{(i)}$ with respect to $\|\cdot\|_{\mathbf{Q}^{(i)}}$,

$$\langle \boldsymbol{x}^{(i)}, \boldsymbol{m}^{(i)} \rangle - \frac{1}{2}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} - \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{m}^{(i)} \rangle + \frac{1}{2}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} \geq \frac{1}{2}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}}; \quad (45)$$

Similarly,

$$\langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i)} \rangle - \frac{1}{2}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} - \langle \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i)} \rangle + \frac{1}{2}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} \geq \frac{1}{2}\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|^2_{\mathbf{Q}^{(i)}}. \quad (46)$$

Hence, summing (45) and (46) yields that

$$\langle \boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{m}^{(i)} - \boldsymbol{u}^{(i)} \rangle \geq \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}} \implies \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|_{\mathbf{Q}^{(i)}} \leq \|\boldsymbol{u}^{(i)} - \boldsymbol{m}^{(i)}\|_{*,\mathbf{Q}^{(i)}},$$

by Cauchy-Schwarz inequality. □

Claim C.26. *Suppose that* $\sum_{i=1}^{m-1}\|\mathbf{Q}^{(i+1)} - \mathbf{Q}^{(i)}\|_2 \leq \sigma(m)$. *Then,*

$$\frac{1}{2}\sum_{i=1}^{m}\left(\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} - \|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}}\right) \leq \frac{1}{2}\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(0)}\|^2_{\mathbf{Q}^{(1)}} + \frac{\Omega^2_{\mathcal{X}}}{2}\sigma(m).$$

Proof. First, we observe that

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{m}\left(\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}\|^2_{\mathbf{Q}^{(i)}} - \|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}\|^2_{\mathbf{Q}^{(i)}}\right) \leq \frac{1}{2}\sum_{i=2}^{m}(\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)})^\top(\mathbf{Q}^{(i)} - \mathbf{Q}^{(i-1)})(\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}) \\ + \frac{1}{2}\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(0)}\|^2_{\mathbf{Q}^{(1)}} - \frac{1}{2}\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(m)}\|^2_{\mathbf{Q}^{(m)}}.\end{aligned}$$

The first term on the right-hand side can be further bounded as

$$\begin{aligned}(\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)})^\top(\mathbf{Q}^{(i)} - \mathbf{Q}^{(i-1)})(\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}) &\leq \|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}\|_2\|(\mathbf{Q}^{(i)} - \mathbf{Q}^{(i-1)})(\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)})\|_2 \\ &\leq \|\mathbf{Q}^{(i)} - \mathbf{Q}^{(i-1)}\|_2\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 \\ &\leq \Omega_{\mathcal{X}}^2\|\mathbf{Q}^{(i)} - \mathbf{Q}^{(i-1)}\|_2.\end{aligned}$$

As a result, we conclude that

$$\frac{1}{2}\sum_{i=1}^{m}\left(\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i-1)}\|_{\mathbf{Q}^{(i)}}^2 - \|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}\|_{\mathbf{Q}^{(i)}}^2\right) \leq \frac{1}{2}\|\mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(0)}\|_{\mathbf{Q}^{(1)}}^2 + \frac{\Omega_{\mathcal{X}}^2}{2}\sigma(m).$$

□

Finally, combining Claim C.25 and Claim C.26 with (44) completes the proof. □

An important advantage of this guarantee is that the term in (41) that captures the misprediction error depends on the sequence of preconditioners, which could be potentially selected in a dynamics way to minimize that term.

C.5 Stochastic Games

In this subsection, we study another settings that has received considerable attention, especially in recent years; namely, stochastic games [Shapley, 1953]. In particular, we will extend our previous results in (infinite-horizon) two-player zero-sum stochastic games, but with an important caveat that will be explained in the sequel. Such games are not covered by our previous techniques, as it will become clear shortly.

For the sake of simplicity in the exposition, we will focus on a simple stochastic games that—in a certain sense—captures the difficulty of the problem, known as Von Neumann's ratio game [Neumann, 1945], defined as

$$V(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \frac{\boldsymbol{x}^\top\mathbf{R}\boldsymbol{y}}{\boldsymbol{x}^\top\mathbf{S}\boldsymbol{y}}, \tag{47}$$

where $\boldsymbol{x} \in \Delta^{d_x}$; $\boldsymbol{y} \in \Delta^{d_y}$; and $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{d_x \times d_y}$. It is further assumed that for any $\boldsymbol{x} \in \Delta^{d_x}$ and $\boldsymbol{y} \in \Delta^{d_y}$ it holds that $\boldsymbol{x}^\top\mathbf{S}\boldsymbol{y} \geq \zeta > 0$, which ensures that the objective (47) is indeed well-defined. The ratio game (47) can be interpreted as a single-state stochastic game, in which the immediate reward under actions $(a_x, a_y) \in \mathcal{A}_x \times \mathcal{A}_y$ is given by $\mathbf{R}[a_x, a_y]$, while $\mathbf{S}[a_x, a_y]$ represents the probability of stopping at reach round. We point out that the subsequent analysis can be extended to general infinite-horizon (discounted) two-player zero-sum stochastic games by appropriately performing the analysis for each state.

The ratio game (47) evidently captures bilinear saddle-point problems (studied in Appendix C.1) in the special case when $\boldsymbol{x}^\top\mathbf{S}\boldsymbol{y} = 1$ for any $(\boldsymbol{x}, \boldsymbol{y}) \in \Delta^{d_x} \times \Delta^{d_y}$, but the objective (47) is in general nonconvex-nonconcave. In fact, as pointed out in [Daskalakis et al., 2020, Proposition 2], the MVI property (Assumption C.14) could fail in such games. It is subsequently an open problem to understand whether `OGD` converges in such games [Daskalakis et al., 2020, Open Problem 1]. Here, we will show that there exists a learning rate schedule, time-varying but non-vanishing, that ensures that `OGD` reaches a minimax equilibrium of (47); the main caveat is that we have not been able to identify such a learning rate schedule in an algorithmically useful way.

Let us first state some useful properties. Two-player zero-sum (discounted) stochastic games, and in particular Von Neumann's ratio game (47), admit a minimax theorem, as was shown in the pioneering work of Shapley [1953]:

Fact C.27 ([Shapley, 1953, Neumann, 1945]). *Let $V : \Delta^{d_x} \times \Delta^{d_y} \to \mathbb{R}$ be defined as in (47). Then,*

$$\min_{\boldsymbol{x}^\star \in \mathcal{X}} \max_{\boldsymbol{y}^\star \in \mathcal{Y}} V(\boldsymbol{x}^\star, \boldsymbol{y}^\star) = \max_{\boldsymbol{y}^\star \in \mathcal{Y}} \min_{\boldsymbol{x}^\star \in \mathcal{X}} V(\boldsymbol{x}^\star, \boldsymbol{y}^\star).$$

The second useful property we will require is the so-called *gradient dominance property* (see [Agarwal et al., 2021] and references therein), which is formalized below.

Fact C.28 (Gradient Dominance [Agarwal et al., 2021]). *Let $V : \Delta^{d_x} \times \Delta^{d_y} \to \mathbb{R}$ be defined as in (47). There is a parameter $C \in \mathbb{R}_{>0}$ such that for any fixed $\boldsymbol{y} \in \Delta^{d_y}$,*

$$V(\boldsymbol{x}, \boldsymbol{y}) - \min_{\boldsymbol{x}^\star \in \mathcal{X}} V(\boldsymbol{x}^\star, \boldsymbol{y}) \leq C \max_{\boldsymbol{x}^\star \in \mathcal{X}} \langle \boldsymbol{x} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}} V(\boldsymbol{x}, \boldsymbol{y}) \rangle. \tag{48}$$

Furthermore, an analogous relation holds for player y.

Establishing (48) for the ratio game is immediate, but that property holds generally—under relatively mild assumptions on the underlying stochastic game. The importance of Fact C.28 it that it guarantees that approximate stationary points of gradient-descent-type algorithms are also globally optimal for each player—even though the optimization problem faced by each player is nonconvex (nonconcave). In light of the fact that the MVI property (Assumption C.14) fails even for the ratio game [Daskalakis et al., 2020], when $F := (\nabla_{\boldsymbol{x}} V(\boldsymbol{x}, \boldsymbol{y}), -\nabla_{\boldsymbol{y}} V(\boldsymbol{x}, \boldsymbol{y}))$, we introduce a more general condition below.

Property C.29 (Generalized MVI for Min-Max Problems). *Consider the operator $F := (\nabla_{\boldsymbol{x}} V(\boldsymbol{x}, \boldsymbol{y}), -\nabla_{\boldsymbol{y}} V(\boldsymbol{x}, \boldsymbol{y}))$. For any sequence $\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(m)} \in \mathcal{Z}$, there exists $(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \in \mathcal{X} \times \mathcal{Y}$ and sequences of weights $\alpha_x^{(1)}, \dots, \alpha_x^{(m)} \in \mathbb{R}_{>0}$ and $\alpha_y^{(1)}, \dots, \alpha_y^{(m)} \in \mathbb{R}_{>0}$, with $0 < \ell \leq \alpha_x^{(i)}, \alpha_y^{(i)} \leq h$, for each $i \in [\![m]\!]$, such that*

$$\sum_{i=1}^{m} \alpha_x^{(i)} \langle \nabla_{\boldsymbol{x}} V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}), \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star \rangle + \sum_{i=1}^{m} \alpha_y^{(i)} \langle \nabla_{\boldsymbol{y}} V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}), \boldsymbol{y}^\star - \boldsymbol{y}^{(i)} \rangle \geq 0. \tag{49}$$

Further, for any $2 \leq i \leq m$, there exists $W \in \mathbb{R}_{>0}$ such that

$$\frac{\alpha_x^{(i)}}{\alpha_x^{(i-1)}}, \frac{\alpha_y^{(i)}}{\alpha_y^{(i-1)}} \leq 1 + W \|\boldsymbol{z}^{(i)} - \boldsymbol{z}^{(i-1)}\|_2. \tag{50}$$

It is evident that if the MVI property holds, (49) and (50) also follow with all the coefficients being equal to 1, but, unlike the MVI property, as we shall see Property C.29 it satisfied for stochastic games. The sequence of weights in (49) is designed to take into account the distribution shift, a central challenge in reinforcement learning. We now observe that Property C.29 is satisfied for the ratio game.

Proposition C.30. *Let $F := (\nabla_{\boldsymbol{x}} V(\boldsymbol{x}, \boldsymbol{y}), -\nabla_{\boldsymbol{y}} V(\boldsymbol{x}, \boldsymbol{y}))$, where V is the ratio game defined in (47). Then, Property C.29 is satisfied with $\ell = \zeta / S_{max}$ and $h = S_{max}/\zeta$, where $S_{max} := \max_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}} \boldsymbol{x}^\top \mathbf{S} \boldsymbol{y}$ and $\min_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}} \boldsymbol{x}^\top \mathbf{S} \boldsymbol{y} \geq \zeta$.*

Before we proceed with the proof of Proposition C.30, we remark that Property C.29—and subsequently Proposition C.30—can be directly extended to multi-state stochastic games as well.

Proof of Proposition C.30. Consider any sequence $\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(m)} \in \mathcal{Z}$, for some $m \in \mathbb{N}$. We first see that for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$,

$$\nabla_{\boldsymbol{x}} V(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathbf{R}\boldsymbol{y}(\boldsymbol{x}^\top \mathbf{S} \boldsymbol{y}) - \mathbf{S}\boldsymbol{y}(\boldsymbol{x}^\top \mathbf{R} \boldsymbol{y})}{(\boldsymbol{x}^\top \mathbf{S} \boldsymbol{y})^2},$$

and similarly,

$$\nabla_{\boldsymbol{y}} V(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathbf{R}^\top \boldsymbol{x}(\boldsymbol{x}^\top \mathbf{S}\boldsymbol{y}) - \mathbf{S}^\top \boldsymbol{x}(\boldsymbol{x}^\top \mathbf{R}\boldsymbol{y})}{(\boldsymbol{x}^\top \mathbf{S}\boldsymbol{y})^2}.$$

As a result, for any $\boldsymbol{x}^\star \in \mathcal{X}$,

$$\begin{aligned}
\sum_{i=1}^m \left(V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) - V(\boldsymbol{x}^\star, \boldsymbol{y}^{(i)}) \right) &= \sum_{i=1}^m \left(\frac{(\boldsymbol{x}^{(i)})^\top \mathbf{R}\boldsymbol{y}^{(i)}}{(\boldsymbol{x}^{(i)})^\top \mathbf{S}\boldsymbol{y}^{(i)}} - \frac{(\boldsymbol{x}^\star)^\top \mathbf{R}\boldsymbol{y}^{(i)}}{(\boldsymbol{x}^\star)^\top \mathbf{S}\boldsymbol{y}^{(i)}} \right) \\
&= \sum_{i=1}^m \left(\frac{(\boldsymbol{x}^{(i)})^\top \mathbf{R}\boldsymbol{y}^{(i)} (\boldsymbol{x}^\star)^\top \mathbf{S}\boldsymbol{y}^{(i)} - (\boldsymbol{x}^{(i)})^\top \mathbf{S}\boldsymbol{y}^{(i)} (\boldsymbol{x}^\star)^\top \mathbf{R}\boldsymbol{y}^{(i)}}{(\boldsymbol{x}^{(i)})^\top \mathbf{S}\boldsymbol{y}^{(i)} (\boldsymbol{x}^\star)^\top \mathbf{S}\boldsymbol{y}^{(i)}} \right) \\
&= \sum_{i=1}^m \frac{(\boldsymbol{x}^{(i)})^\top \mathbf{S}\boldsymbol{y}^{(i)}}{(\boldsymbol{x}^\star)^\top \mathbf{S}\boldsymbol{y}^{(i)}} \langle \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}} V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle \\
&= \sum_{i=1}^m \alpha_x^{(i)} \langle \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}} V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle, \qquad (51)
\end{aligned}$$

where

$$\alpha_x^{(i)}(\boldsymbol{x}^\star) \coloneqq \frac{(\boldsymbol{x}^{(i)})^\top \mathbf{S}\boldsymbol{y}^{(i)}}{(\boldsymbol{x}^\star)^\top \mathbf{S}\boldsymbol{y}^{(i)}}. \qquad (52)$$

Similarly, for any $\boldsymbol{y}^\star \in \mathcal{Y}$,

$$\sum_{i=1}^m \left(V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^\star) - V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \right) = \sum_{i=1}^m \alpha_y^{(i)} \langle \boldsymbol{y}^\star - \boldsymbol{y}^{(i)}, \nabla_{\boldsymbol{y}} V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \rangle, \qquad (53)$$

where

$$\alpha_y^{(i)}(\boldsymbol{y}^\star) \coloneqq \frac{(\boldsymbol{x}^{(i)})^\top \mathbf{S}\boldsymbol{y}^{(i)}}{(\boldsymbol{x}^{(i)})^\top \mathbf{S}\boldsymbol{y}^\star}. \qquad (54)$$

But, optimizing over $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, the sum of the left-hand sides of (51) and (53) can be lower bounded by

$$\begin{aligned}
&\max_{\boldsymbol{y}^\star \in \mathcal{Y}} \left\{ \sum_{i=1}^m V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^\star) \right\} - \min_{\boldsymbol{x}^\star \in \mathcal{X}} \left\{ \sum_{i=1}^m V(\boldsymbol{x}^\star, \boldsymbol{y}^{(i)}) \right\} \\
&\qquad \geq m \max_{\boldsymbol{y}^\star \in \mathcal{Y}} \min_{\boldsymbol{x}^\star \in \mathcal{X}} V(\boldsymbol{x}^\star, \boldsymbol{y}^\star) - m \min_{\boldsymbol{x}^\star \in \mathcal{X}} \max_{\boldsymbol{y}^\star \in \mathcal{Y}} V(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \geq 0,
\end{aligned}$$

by Fact C.27, which in turn implies that Property C.29 is satisfied with coefficients as defined in (52) and (54). Indeed, bounding those coefficients is immediate, yielding (49), while (50) also follows directly from (52) and (54). □

Importantly, for a min-max problem that enjoys Property C.29, one can employ `OGD`, but with learning rate that adapts to the sequence of weights associated with Property C.29. In particular, we analyze the following variant of `OGD`, defined for any iteration $i \in \mathbb{N}$ as

$$\begin{aligned}
\boldsymbol{x}^{(i)} &\coloneqq \mathrm{Proj}_{\mathcal{X}} \left(\hat{\boldsymbol{x}}^{(i-1)} - \eta \alpha_x^{(i-1)} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i-1)}, \boldsymbol{y}^{(i-1)}) \right), \\
\hat{\boldsymbol{x}}^{(i)} &\coloneqq \mathrm{Proj}_{\mathcal{X}} \left(\hat{\boldsymbol{x}}^{(i-1)} - \eta \alpha_x^{(i)} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \right),
\end{aligned}$$

where $\alpha_x^{(i-1)}, \alpha_x^{(i)}$ are defined as in (52). A similar update rule holds for player y.

Theorem C.31. *Consider the ratio game $V(\boldsymbol{x}, \boldsymbol{y})$ defined in (47). If both players employ `OGD` with a suitable learning rate schedule, $O(1/\epsilon^2)$ iterations suffices to reach a point $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \in \mathcal{X} \times \mathcal{Y}$ such that*

$$V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) - \min_{\boldsymbol{x} \in \mathcal{X}} V(\boldsymbol{x}, \boldsymbol{y}^{(i)}) \leq \epsilon \quad \textit{and} \quad \max_{\boldsymbol{y} \in \mathcal{Y}} V(\boldsymbol{x}^{(i)}, \boldsymbol{y}) - V(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \leq \epsilon.$$

Sketch of Proof. First, for any iteration $i \in [\![m]\!]$,

$$\alpha_x^{(i-1)}\langle \boldsymbol{x}^{(i)}, \nabla_{\boldsymbol{x}}^{(i-1)}\rangle - \frac{1}{2\eta}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 - \alpha_x^{(i-1)}\langle \hat{\boldsymbol{x}}^{(i)}, \nabla_{\boldsymbol{x}}^{(i-1)}\rangle + \frac{1}{2\eta}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 \geq \frac{1}{2\eta}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|_2^2,$$

where we used the shorthand notation $\nabla_{\boldsymbol{x}}^{(i)} \coloneqq \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$. In turn, this implies that

$$\alpha_x^{(i)}\langle \boldsymbol{x}^{(i)}, \nabla_{\boldsymbol{x}}^{(i-1)}\rangle - \frac{\alpha_x^{(i)}}{2\alpha_x^{(i-1)}\eta}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 - \alpha_x^{(i)}\langle \hat{\boldsymbol{x}}^{(i)}, \nabla_{\boldsymbol{x}}^{(i)}\rangle + \frac{\alpha_x^{(i)}}{2\alpha_x^{(i-1)}\eta}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 \geq \frac{\alpha_x^{(i)}}{2\alpha_x^{(i-1)}\eta}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|_2^2. \tag{55}$$

Further, for any iteration $i \in [\![m]\!]$,

$$\alpha_x^{(i)}\langle \hat{\boldsymbol{x}}^{(i)}, \nabla_{\boldsymbol{x}}^{(i)}\rangle - \frac{1}{2\eta}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 - \alpha_x^{(i)}\langle \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}}^{(i)}\rangle + \frac{1}{2\eta}\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 \geq \frac{1}{2\eta}\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}^{(i)}\|_2^2. \tag{56}$$

Moreover, we have

$$\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|_2 \leq \eta\alpha_x^{(i-1)}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i-1)}, \boldsymbol{y}^{(i-1)})\|_2, \tag{57}$$

since the Euclidean projection operator $\mathrm{Proj}(\cdot)$ is nonexpansive with respect to $\|\cdot\|_2$. Similarly,

$$\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2 \leq \eta\alpha_x^{(i)}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\|_2. \tag{58}$$

Thus, combining (57) and (58) with (50) of Property C.29 implies that for a sufficiently small learning rate $\eta > 0$, it holds that $\frac{\alpha_x^{(i)}}{\alpha_x^{(i-1)}} \leq 1 + \gamma$, for a sufficiently small universal constant γ. Now when combining (55) and (56) there is a mismatch term of the form

$$\frac{\alpha_x^{(i)}}{2\alpha_x^{(i-1)}\eta}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 - \frac{1}{2\eta}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2 \leq \frac{\gamma}{2\eta}\|\hat{\boldsymbol{x}}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2^2.$$

Thus, when $\frac{\alpha_x^{(i)}}{\alpha_x^{(i-1)}} \geq 1$ this mismatch term will be subsumed by the other terms in (55). As a result, we can obtain an RVU bound for the weighted regret $\sum_{i=1}^m \alpha_x^{(i)}\langle \boldsymbol{x}^{(i)} - \boldsymbol{x}^\star, \nabla_{\boldsymbol{x}}^{(i)}\rangle$, and similar reasoning applies for player y. As a result, analogously to the proof of Theorem C.17, we conclude that $O(1/\epsilon^2)$ suffice so that $\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|_2, \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i-1)}\|_2, \|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)}\|_2, \|\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)}\|_2 \leq \epsilon$ (using Proposition C.30). Finally, this implies that $\max_{x\in\mathcal{X}}\langle \boldsymbol{x}^{(i)} - \boldsymbol{x}, \nabla_{\boldsymbol{x}}^{(i)}\rangle = O(\epsilon)$ and $\max_{\boldsymbol{y}\in\mathcal{Y}}\langle \boldsymbol{y} - \boldsymbol{y}^{(i)}, \nabla_{\boldsymbol{y}}^{(i)}\rangle = O(\epsilon)$, and the gradient dominance property (Fact C.28) yields the statement. □

Finally, it is immediate to derive the meta-learning version of Theorem C.31, parameterized in terms of the similarity of the minimax equilibria of the underlying stochastic games.

C.6 Hölder Smooth Games

So far our analysis has (either implicitly or explicitly) required some form of Lipschitz continuity in order to appropriate massage the RVU bound. In this subsection, we will relax that condition. In particular, we study variational inequality problems for which the operator is Hölder continuous, in the following precise sense.

Definition C.32. *Consider an operator $F : \mathcal{Z} \to \mathcal{Z}$. We say that F is α-Hölder continuous, with $\alpha \in (0, 1]$, if for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$,*

$$\|F(\boldsymbol{z}) - F(\boldsymbol{z}')\|_2 \leq H\|\boldsymbol{z} - \boldsymbol{z}'\|_2^{\alpha}, \tag{59}$$

for some parameter $H > 0$.

Naturally, (59) reduces to Lipschitz continuity of F when $\alpha = 1$. This class of problems was also studied in the seminal work of Rakhlin and Sridharan [2013b]. Indeed, the following argument partially overlaps with [Rakhlin and Sridharan, 2013b, Lemma 3]. For convenience, below we use a prediction based on the secondary sequences of `OGD`.

Proposition C.33 (Refinement of Corollary C.7 under Hölder Smoothness)**.** *Suppose that $F : \mathcal{Z} \to \mathcal{Z}$ is α-Hölder smooth, with $\alpha \in (0, 1)$, and satisfies the MVI property (Assumption C.14). Then, for `OGD` with prediction $\boldsymbol{m}^{(i)} := F(\hat{\boldsymbol{x}}^{(i-1)})$ and learning rate $\eta > 0$ set as*

$$\eta(m) := \left(\frac{\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2}{mH^{\frac{2}{1-\alpha}} g(\alpha)}\right)^{\frac{1-\alpha}{2}}, \tag{60}$$

where $g(\alpha) = (1 + \alpha)(2 + 2\alpha)^{\frac{1-\alpha}{1+\alpha}}$, it holds that

$$\sum_{i=1}^{m} \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2^2 + \sum_{i=1}^{m} \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^2 \leq 2\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2,$$

where $\boldsymbol{z}^\star \in \mathcal{Z}$ is any point that satisfies the MVI property (Assumption C.14).

In the argument below we do not make any attempt to optimize factors that depend (solely) on α—the modulus of Hölder continuity (Definition C.32).

Proof. First, analogously to the proof of Corollary C.7, we have that for any $\boldsymbol{z}^\star \in \mathcal{Z}^\star$,

$$\begin{aligned} 0 \leq \frac{1}{2\eta}\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2 + H\sum_{i=1}^{m} \langle \boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}, F(\boldsymbol{z}^{(i)}) - F(\hat{\boldsymbol{z}}^{(i-1)})\rangle \\ - \frac{1}{2\eta}\sum_{i=1}^{m} \left(\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2^2 + \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^2\right). \end{aligned}$$

Further, by Cauchy-Schwarz inequality we have that $\langle \boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}, F(\boldsymbol{z}^{(i)}) - F(\hat{\boldsymbol{z}}^{(i-1)})\rangle \leq \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2 \|F(\boldsymbol{z}^{(i)}) - F(\hat{\boldsymbol{z}}^{(i-1)})\|_2$, in turn implying that

$$\begin{aligned} 0 \leq \frac{1}{2\eta}\|\boldsymbol{z}^\star - \boldsymbol{z}^{(0)}\|_2^2 + H\sum_{i=1}^{m} \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2^{1+\alpha} + H\sum_{i=1}^{m} \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^{1+\alpha} \\ - \frac{1}{2\eta}\sum_{i=1}^{m} \left(\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2^2 + \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^2\right), \end{aligned} \tag{61}$$

where we used the fact that $\|F(\boldsymbol{z}^{(i)}) - F(\hat{\boldsymbol{z}}^{(i-1)})\|_2 \leq H\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^{\alpha}$ (by α-Hölder smoothness), and Young's inequality, $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$ for any $a, b \in \mathbb{R}_{\geq 0}$ and $p, q \in \mathbb{R}_{>0}$ such that $\frac{1}{p} + \frac{1}{q} = 1$, implying that $\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2 \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^{\alpha} \leq \frac{1}{1+\alpha}\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2^{1+\alpha} + \frac{\alpha}{1+\alpha}\|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^{1+\alpha} \leq \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i)}\|_2^{1+\alpha} + \|\boldsymbol{z}^{(i)} - \hat{\boldsymbol{z}}^{(i-1)}\|_2^{1+\alpha}$ (for $p = 1 + \alpha$ and $q = \frac{1+\alpha}{\alpha}$). Let us now bound the middle

terms in the right-hand side of (61). We have

$$H\sum_{i=1}^{m}\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i)}\|^{1+\alpha}=\sum_{i=1}^{m}H\left((2+2\alpha)\eta\right)^{\frac{1+\alpha}{2}}\left(\frac{\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i)}\|}{\sqrt{(2+2\alpha)\eta}}\right)^{1+\alpha}$$

$$\leq\left(\sum_{i=1}^{m}H^{\frac{2}{1-\alpha}}\left((2+2\alpha)\eta\right)^{\frac{1+\alpha}{1-\alpha}}\right)^{\frac{1-\alpha}{2}}\left(\sum_{i=1}^{m}\frac{\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i)}\|_2^2}{(2+2\alpha)\eta}\right)^{\frac{1+\alpha}{2}} \tag{62}$$

$$\leq\frac{1-\alpha}{2}mH^{\frac{2}{1-\alpha}}\left((2+2\alpha)\eta\right)^{\frac{1+\alpha}{1-\alpha}}+\frac{1}{4\eta}\sum_{i=1}^{m}\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i)}\|_2^2, \tag{63}$$

where (62) uses Hölder's inequality with conjugate powers $(1-\alpha)/2$ and $(1+\alpha)/2$, and (63) uses the (weighted) AM-GM inequality. Similarly,

$$H\sum_{i=1}^{m}\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i-1)}\|^{1+\alpha}\leq\frac{1-\alpha}{2}mH^{\frac{2}{1-\alpha}}\left((2+2\alpha)\eta\right)^{\frac{1+\alpha}{1-\alpha}}+\frac{1}{4\eta}\sum_{i=1}^{m}\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i-1)}\|_2^2.$$

We now optimize the following function

$$\frac{1}{2\eta}\|\boldsymbol{z}^\star-\boldsymbol{z}^{(0)}\|_2^2+(1-\alpha)mH^{\frac{2}{1-\alpha}}(2+2\alpha)^{\frac{1+\alpha}{1-\alpha}}\eta^{\frac{1+\alpha}{1-\alpha}} \tag{64}$$

in terms of the learning rate $\eta>0$. In particular, if we let $g(\alpha):=(1+\alpha)(2+2\alpha)^{\frac{1-\alpha}{1+\alpha}}$, a simple calculation yields that the optimal value for η is

$$\eta(m):=\left(\frac{\|\boldsymbol{z}^\star-\boldsymbol{z}^{(0)}\|_2^2}{2mH^{\frac{2}{1-\alpha}}g(\alpha)}\right)^{\frac{1-\alpha}{2}},$$

which gives a value to (64) equal to

$$\left(\frac{H\|\boldsymbol{z}^\star-\boldsymbol{z}^{(0)}\|^{1+\alpha}}{2^{\frac{1+\alpha}{2}}(g(\alpha))^{\frac{\alpha-1}{2}}}\right)m^{\frac{1-\alpha}{2}}.$$

Plugging this bound to (61) and rearranging concludes the proof. □

Above, we have assumed access to a point $\boldsymbol{z}^\star\in\mathcal{Z}$ that satisfies the MVI property in order to optimally tune the learning rate. That assumption can be sidestepped similarly to our approach in Appendix D.2. While Proposition C.33 still guarantees that the second-order path lengths are bounded ((60)), the difference with our previous bounds under Lipschitz continuity is that the learning rate η has to decay with the time horizon m. This is reflected on a worse bound on the number of iterations required to reach an approximate solution to the VI problem, as we formalize below. To our knowledge, the following guarantee is the first of its kind.

Theorem C.34 (Rates for Hölder Smooth Functions)**.** *Consider an α-Hölder continuous operator $F:\mathcal{Z}\to\mathcal{Z}$ that satisfies the MVI property (Assumption C.14). Then, after $O(m)$ iterations of* `OGD` *with the learning rate of Proposition C.33 there is an iterate that is an $m^{-\alpha/2}$-approximate strong solution to the VI problem.*

Proof. By Proposition C.33, after m iterations there is an iterate $i\in[\![m]\!]$ such that $\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i)}\|_2,\|\boldsymbol{z}^{(i)}-\hat{\boldsymbol{z}}^{(i-1)}\|_2\leq\frac{2\|\boldsymbol{z}^\star-\boldsymbol{z}^{(0)}\|_2}{\sqrt{m}}$. Thus, the statement follows from Claim C.16. □

In the special case where $\alpha=1$, the guarantee above recovers the recently established $m^{-1/2}$ rates of `OGD` under Lipschitz continuity, which is known to be tight [Golowich et al., 2020a,b]. An interesting question is to derive lower bounds under the weaker Hölder continuity assumption studied in the present subsection. Finally, obtaining a meta-learning version of Theorem C.34 follows immediately from our previous techniques. We caution that one has to also meta-learn the learning rate in this case as Proposition C.33 assumes knowledge of $\boldsymbol{z}^\star\in\mathcal{Z}$; this is analogous to our approach in Theorem D.7, presented in Appendix D.2.

C.7 The Extra-Gradient Method

In this subsection, we extend our results to the *extra-gradient* method. Starting from the seminal work of Korpelevich [1976], that method—and variants thereof—has received tremendous attention in the literature; we refer to the excellent discussion of Hsieh et al. [2019] for further pointers. We will consider a generalized version of the extra-gradient method, defined with the following update rule for $i \in \mathbb{N}$.

$$\begin{aligned} \hat{\boldsymbol{x}}^{(i)} &\coloneqq \arg\max_{\hat{\boldsymbol{x}} \in \mathcal{X}} \left\{ \langle \hat{\boldsymbol{x}}, \boldsymbol{u}^{(i-1)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\hat{\boldsymbol{x}} \,\|\, \boldsymbol{x}^{(i-1)}) \right\}, \\ \boldsymbol{x}^{(i)} &\coloneqq \arg\max_{\boldsymbol{x} \in \mathcal{X}} \left\{ \langle \boldsymbol{x}, \hat{\boldsymbol{u}}^{(i)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\boldsymbol{x} \,\|\, \boldsymbol{x}^{(i-1)}) \right\}. \end{aligned} \tag{65}$$

The initialization is defined as in `OGD`. For brevity, we will refer to this algorithm as `EG`. We remark that the update rule presented in (65) is more general than the standard extra-gradient method, which corresponds to (65) with Euclidean regularization. While similar to `OGD`, `EG` has one central difference: the update of $\boldsymbol{x}^{(i)}$ uses an auxiliary feedback $\hat{\boldsymbol{u}}^{(i)}$. For this reason, unlike `OGD`, `EG` does not fit into the traditional online learning framework. In fact, any non-trivial modification of `EG` fails to provide meaningful regret guarantees when faced against an adversarially selected sequence of utilities [Golowich et al., 2020a]. That additional feedback $\hat{\boldsymbol{u}}^{(i)}$ is also unsatisfactory since it requires two gradient-oracle calls per iteration—unlike single-call variants, such as `OGD` [Hsieh et al., 2019].

In this context, the main purpose of this subsection is to show that `EG` can be analyzed in the same framework as `OGD`, which will ensure that the guarantees we have provided so far for `OGD` can be translated to `EG` as well. In particular, although `EG` does not lie in the no-regret framework, we will show that it still admits an RVU-type bound; this leads to a unifying analysis of `OGD` and `EG`, and recovers several well-known results from the literature in a much simpler fashion. The first key idea is to consider the regret incurred by the secondary sequence of `EG` $(\hat{\boldsymbol{x}}^{(i)})_{1 \le i \le m}$, and with respect to the auxiliary sequence of utilities $(\hat{\boldsymbol{u}}^{(i)})_{1 \le i \le m}$:

$$\hat{\mathrm{Reg}}^{(m)}(\mathring{\boldsymbol{x}}) \coloneqq \sum_{i=1}^{m} \langle \mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)} \rangle. \tag{66}$$

Below, we show that `EG` also admits an RVU-type bound, but with respect to the notion of regret introduced in (66).

Theorem C.35 (RVU-type Bound for Extra-Gradient)**.** *For the extra-gradient method (65) it holds that $\sum_{i=1}^{m} \langle \mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)} \rangle$ can be upper bounded, for any $\mathring{\boldsymbol{x}} \in \mathcal{X}$, as*

$$\frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \,\|\, \boldsymbol{x}^{(0)}) + \eta \sum_{i=1}^{m} \|\hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)}\|_*^2 - \frac{1}{2\eta} \sum_{i=1}^{m} \left(\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|^2 + \|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i-1)}\|^2 \right).$$

Proof. Fix any $i \in [\![m]\!]$. By 1-strong convexity of $\mathcal{R}$ with respect to $\|\cdot\|$,

$$\langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i-1)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\hat{\boldsymbol{x}}^{(i)} \,\|\, \boldsymbol{x}^{(i-1)}) - \langle \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i-1)} \rangle + \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\boldsymbol{x}^{(i)} \,\|\, \boldsymbol{x}^{(i-1)}) \ge \frac{1}{2\eta} \|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|^2, \tag{67}$$

and for any $\mathring{\boldsymbol{x}} \in \mathcal{X}$,

$$\langle \boldsymbol{x}^{(i)}, \hat{\boldsymbol{u}}^{(i)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\boldsymbol{x}^{(i)} \,\|\, \boldsymbol{x}^{(i-1)}) - \langle \mathring{\boldsymbol{x}}, \hat{\boldsymbol{u}}^{(i)} \rangle + \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \,\|\, \boldsymbol{x}^{(i-1)}) \ge \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \,\|\, \boldsymbol{x}^{(i)}). \tag{68}$$

Moreover, for any $i \in [\![m]\!]$,

$$\langle \mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)} \rangle = \langle \mathring{\boldsymbol{x}}, \hat{\boldsymbol{u}}^{(i)} \rangle - \langle \boldsymbol{x}^{(i)}, \hat{\boldsymbol{u}}^{(i)} \rangle + \langle \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i-1)} \rangle - \langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i-1)} \rangle + \langle \boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)} \rangle.$$

Combining with (67) and (68) yields that $\sum_{i=1}^{m}\langle \mathring{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)}\rangle$ is upper bounded by

$$\frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \boldsymbol{x}^{(0)}) + \sum_{i=1}^{m}\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|\|\hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)}\|_* - \frac{1}{2\eta}\sum_{i=1}^{m}\left(\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|^2 + \|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i-1)}\|^2\right), \tag{69}$$

where we further used Cauchy-Schwarz inequality for the dual pair of norms $(\|\cdot\|, \|\cdot\|_*)$ to obtain that $\langle \boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)}\rangle \le \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|\|\hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)}\|_*$, as well as the telescopic summation

$$\frac{1}{\eta}\sum_{i=1}^{m}\left(\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \boldsymbol{x}^{(i-1)}) - \mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \boldsymbol{x}^{(i)})\right) = \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \boldsymbol{x}^{(0)}) - \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \boldsymbol{x}^{(m)}) \le \frac{1}{\eta}\mathcal{B}_{\mathcal{R}}(\mathring{\boldsymbol{x}} \parallel \boldsymbol{x}^{(0)}),$$

since $\mathcal{B}_{\mathcal{R}}(\cdot \parallel \cdot) \ge 0$. Finally, we will need the following stability bound.

Claim C.36. *For any $i \in [\![m]\!]$,*

$$\|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\| \le \eta\|\hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)}\|_*.$$

Proof of Claim C.36. By replacing $\mathring{\boldsymbol{x}} \coloneqq \hat{\boldsymbol{x}}^{(i)}$ in (68) and summing with (67),

$$\langle \hat{\boldsymbol{x}}^{(i)}, \boldsymbol{u}^{(i-1)}\rangle - \langle \boldsymbol{x}^{(i)}, \boldsymbol{u}^{(i-1)}\rangle + \langle \boldsymbol{x}^{(i)}, \hat{\boldsymbol{u}}^{(i)}\rangle - \langle \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)}\rangle \ge \frac{1}{\eta}\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|^2,$$

in turn implying that

$$\langle \boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}, \hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)}\rangle \ge \frac{1}{\eta}\|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|^2 \implies \|\hat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\| \le \eta\|\hat{\boldsymbol{u}}^{(i)} - \boldsymbol{u}^{(i-1)}\|_*,$$

where the last bound follows from the Cauchy-Schwarz inequality. □

Finally, the statement follows by combining Claim C.36 with (69). □

This regret bound allows us to automatically inherit the many and important consequences of the RVU bound for `EG` as well. For example, assuming Lipschitz continuity for the underlying operator F, Theorem C.35 guarantees that the sum of the regrets of the secondary sequences—in the sense of (66)—will be bounded by $O(1)$; assuming that F is also monotone, that implies that the average secondary sequences yields an $O(1/m)$-approximate (weak) solution to the VI problem. Further, assuming the MVI property, it holds that the sum of the regrets of the secondary sequences is always nonnegative, thereby implying bounded second-order path lengths for `EG` (as in Corollary C.7); in turn, that gives that $O(1/\epsilon^2)$ iterations suffice to reach an ϵ-approximate (strong) solution to the VI problem. The implications we have derived earlier for `OGD` in the meta-learning setting also follow directly.

C.8 Lower Bounds

Finally, this subsection derives lower bounds on the players' regrets in the meta-learning setting. We begin by first recalling the lower bound for a single zero-sum game (in normal-form) due to Daskalakis et al. [2015] (Proposition C.37). We then extend their idea to the meta-learning setting.

The basic idea is to consider the family of games for which the only nonzero entries of the $d \times d$ payoff matrix are at a single row. In particular, for an index $r \in [\![d]\!]$, we let

$$\mathbf{A}_r[a_x, a_y] = \begin{cases} 1 & \text{if } a_x = r, \\ 0 & \text{otherwise.} \end{cases} \tag{70}$$

We then consider the family of games described by $\{\mathbf{A}_r\}_{1\le r\le d}$. We suppose for convenience that the entries of each matrix correspond to the utility of the row player. Now in any zero-sum game from that class, the column player has no impact on the game since each payoff matrix (70) remains invariant under changes in the columns. Further, the row player is clearly just searching for the index r, which indicates the row with entries all-ones. Since the row player has no knowledge about the payoff matrix in the beginning of the game, there is a high probability that the row player will incur constant regret in the first iteration of the game. Conditioned on that event, this means that $\mathrm{Reg}_x^{(m)} + \mathrm{Reg}_y^{(m)} = \Omega(1)$, for any $m \in \mathbb{N}$, given that $\mathrm{Reg}_y^{(m)} = 0$.

Proposition C.37 ([Daskalakis et al., 2015])**.** *There exists a class of zero-sum games for which* $\mathbb{E}[\mathrm{Reg}_x^{(m)} + \mathrm{Reg}_y^{(m)}] = \Omega(1)$*, where the expectation is with respect to the agents' randomization.*

The construction is the meta-learning setting Now let us imagine that one repeats the previous construction for a sequence of zero-sum games derived from the class in (70). If every such zero-sum game is selected uniformly at random, then Proposition C.37 still applies since there is no additional structure that can be exploited by the meta-learner. In this context, the key idea to refine that lower bound in the meta-learning setting is to assume that there is some prior distribution $\boldsymbol{p} \in \Delta^d$ over indexes, which is in fact assumed to be known by the agents. Then, the sequence of games is produced by selecting a zero-sum game $\mathbf{A}^{(t)}$ from the family $\{\mathbf{A}_r\}_{1\le r\le d}$ according to distribution $\boldsymbol{p}$; those random draws in the course of the t games are independent of each other. The following observation is a direct extension of Proposition C.37.

Proposition C.38. *Consider a sequence of zero-sum games* $(\mathbf{A}^{(t)})_{1\le t\le T}$ *such that each matrix* $\mathbf{A}^{(t)}$ *is produced by selecting from the set* $\{\mathbf{A}_r\}_{1\le r\le d}$ (70) *with probability according to* $\boldsymbol{p}$*. Then,* $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathrm{Reg}_x^{(t,m)} + \mathrm{Reg}_y^{(t,m)}] \ge 1 - \max_{1\le r\le d}\boldsymbol{p}[r]$.

Now, for any $\epsilon > 0$ and $\delta \in (0,1)$, if $T = \mathsf{poly}(d, 1/\epsilon, \log(1/\delta))$ is sufficiently large, standard arguments imply that the empirical frequency $\hat{\boldsymbol{p}} \in \Delta^d$ is ϵ-close to $\boldsymbol{p}$ in terms of total variation distance from $\boldsymbol{p}$, meaning that $\frac{1}{2}\sum_{r=1}^{d}|\boldsymbol{p}[r] - \hat{\boldsymbol{p}}[r]| \le \epsilon$, with probability at least $1-\delta$. Moreover, from the perspective of player x, both the Nash equilibrium and the optimal in hindsight strategy is obvious: for a game $\mathbf{A}_r$, for some $r \in [\![d]\!]$, select (with probability 1) the action indexed by r; the perspective of player y is of no consequence since player y does not incur any regret. Further, let $r^\star \in [\![d]\!]$ be amongst the most frequent indeces of $\boldsymbol{p}$, that is $r^\star \in \arg\max_{1\le r\le d}\boldsymbol{p}[r]$, and let $\boldsymbol{\pi}_{r^\star} \in \Delta^d$ be such that $\boldsymbol{\pi}_{r^\star}[r^\star] = 1$. Thus, conditioned on the event that $\hat{\boldsymbol{p}}$ is ϵ-close to $\boldsymbol{p}$ in terms of total variation distance,

$$V_x^2 := \frac{1}{T}\min_{\boldsymbol{x}\in\mathcal{X}}\sum_{t=1}^{T}\|\mathring{\boldsymbol{x}}^{(t)} - \boldsymbol{x}\|_2^2 \le \frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{\pi}_{r^{(t)}} - \boldsymbol{\pi}_{r^\star}\|_2^2 \le \frac{1}{T}\sum_{t=1}^{T}\max_{r\neq r'}\|\boldsymbol{\pi}_r - \boldsymbol{\pi}_{r'}\|_2^2\mathbb{1}\{r^\star \neq r^{(t)}\}$$
$$= 2\sum_{r\neq r^\star}\hat{\boldsymbol{p}}[r] \le 2(1 - \max_{1\le r\le d}\boldsymbol{p}[r]) + 4\epsilon,$$

where $r^{(t)}$ is the index of the unique nonzero row of the payoff matrix $\mathbf{A}^{(t)}$. Combining this with Proposition C.38, we are now ready to establish our lower bound. Below, we point out that V_y^2 is 0 since any possible strategy for player y yields the same utility (recall our tie-breaking convention for $\mathring{\boldsymbol{y}}^{(t)}$ made after Definition 2.1), while we make a similar convention for V_{NE}^2 to avoid trivialities.

Theorem C.39 (Precise Verison of Theorem 3.3)**.** *Consider a sequence of zero-sum games* $(\mathbf{A}^{(t)})_{1\le t\le T}$ *such that each matrix* $\mathbf{A}^{(t)}$ *is produced by selecting from the set* $\{\mathbf{A}_r\}_{1\le r\le d}$ (70) *with probability according to* $\boldsymbol{p}$*. Then, for any* $\epsilon > 0$ *and a sufficiently large* $T = \mathsf{poly}(d, 1/\epsilon)$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathrm{Reg}_x^{(t,m)} + \mathrm{Reg}_y^{(t,m)}] \ge \frac{1}{2}\left(V_x^2 + V_y^2\right) - \epsilon = \frac{1}{2}V_{NE}^2 - \epsilon.$$

D Proofs from Section 3.2: Meta-Learning in General-Sum Games

In this section, we switch our attention to general-sum games. Unlike zero-sum games, where computing a Nash equilibrium can be phrased as a linear program, there are inherent computational barriers for finding Nash equilibria in general-sum games [Chen et al., 2009, Daskalakis et al., 2009]. Instead, learning algorithms are known to converge to different notions of *correlated equilibria* [Hart and Mas-Colell, 2000, Blum and Mansour, 2007], which are more permissive than the Nash equilibrium [Aumann, 1974].

In our meta-learning context, it is natural to ask whether one can obtain refined regret bounds that are parameterized based on the similarity of the correlated equilibria of the games. However, that appears hard to achieve using uncoupled learning algorithms. Indeed, suppose that after a sufficient number of iterations the agents have converged—in terms of the average product distribution of play—to a, say, coarse correlated equilibrium (CCE) of the underlying game. Then, under uncoupled methods, any initialization at the next game will inevitably produce a product distribution, which could be far from the previous CCE. In other words, we may fail to appropriately leverage the learning of the previous task since the initialization alone is not enough to encode the previous CCE. The separation we are alluding to is summarized below.

Proposition D.1 (Separation between Zero-Sum and General-Sum)**.** *Consider the two-player games $\mathcal{G}$ and $\mathcal{G}'$ with maximum pairwise difference [Candogan et al., 2013] $d(\mathcal{G}, \mathcal{G}') \leq \epsilon$, for a sufficiently small $\epsilon \leq 1/\max\{\mathsf{poly}(\mathcal{G}'), \mathsf{poly}(\mathcal{G})\}$.*

- *If $\mathcal{G}, \mathcal{G}'$ are both zero-sum, then no-regret learning in $\mathcal{G}$ allows the players to have $O(\epsilon)$ regret from the first iteration in $\mathcal{G}'$.*
- *On the other hand, if the games are general-sum, then even if the players have 0 regret in $\mathcal{G}$, finding an initialization with $O(\epsilon)$ regret from the first iteration in $\mathcal{G}'$ is* PPAD*-hard, even if the games are known.*

One way to bypass the aforedescribed difficulties is by incorporating some form of centralization into the learning process. Indeed, computing a correlated equilibrium can be phrased as a zero-sum game between a "mediator," which will serve as the coordinating party, and the players (see, *e.g.*, [Hart and Schmeidler, 1989]). As such, one can leverage our previous results for zero-sum games even in general games, where it is more likely to obtain bounds that depend on the similarity of the correlated equilibria. Nonetheless, investigating this further is not in our scope since it deviates substantially from the online learning paradigm we follow in this paper.

Instead, in Appendices D.2 and D.3 we primarily focus on obtaining refined convergence bounds to correlated and coarse correlated equilibria that depend on the task similarity of the optimal in hindsight. But first, we remark that learning algorithms are known to lead to Nash equilibria in some "structued" general-sum games. Perhaps the most prominent example is that of potential games [Kleinberg et al., 2009, Hofbauer and Sandholm, 2002, Candogan et al., 2013], which is pursued in Appendix D.1 below. Two other interesting directions for which obtaining meta-learning guarantees is left for future work are supermodular games [Milgrom and Roberts, 1990] and games possessing *strict Nash equilibria* [Giannou et al., 2021].

D.1 Potential Games

Here we study meta-learning on potential games given in strategic form (the strategic-form representation was introduced in the beginning of Appendix B.2); we suspect that similar results apply more generally to Markov potential games [Leonardos et al., 2022]. We first recall the definition of a potential game we will use throughout this subsection.

Definition D.2 (Potential Game)**.** *A strategic-form game $\mathcal{G}$ is potential if there exists a function $\Phi : \bigtimes_{k=1}^{n} \mathcal{X}_k \to \mathbb{R}$ such that for any player $i \in [\![n]\!]$, joint strategy $\boldsymbol{x} \in \bigtimes_{k=1}^{n} \mathcal{X}_k$ and action $a_k \in \mathcal{A}_k$,*

$$\frac{\partial \Phi(\boldsymbol{x})}{\partial \boldsymbol{x}_k[a_k]} = u_k(a_k, \boldsymbol{x}_{-k}).$$

We let $\Phi_{\max}$ be an upper bound on the function $|\Phi(\boldsymbol{x})|$. In potential games, unlike zero-sum games, variants of mirror descent, such as gradient descent (`GD`), are known to reach Nash equilibria. In this context, we assume that players face a sequence of potential games $(\Phi^{(t)})_{1 \le t \le T}$, described by their potential functions, and the goal will be to obtain parameterized regret bounds that depend on the similarity of the potential functions. In particular, we will use the following result [Anagnostides et al., 2022b, Corollary 4.4], which gives an initialization-dependent bound on the second-order path length of `GD`.

Proposition D.3 ([Anagnostides et al., 2022b])**.** *Suppose that all players employ `GD` in a potential game with a sufficiently small learning rate $\eta > 0$. Then,*

$$\frac{1}{2\eta} \sum_{i=1}^{m} \sum_{k=1}^{n} \|\boldsymbol{x}_k^{(t,i)} - \boldsymbol{x}_k^{(t,i-1)}\|_2^2 \le \Phi^{(t)}(\boldsymbol{x}^{(t,m)}) - \Phi^{(t)}(\boldsymbol{x}^{(t,0)}), \tag{71}$$

where $\Phi^{(t)} : \bigtimes_{k=1}^{n} \mathcal{X}_k \to \mathbb{R}$ is the potential function of the t-th game, and $\boldsymbol{x}^{(t,i)} = (\boldsymbol{x}_1^{(t,i)}, \dots, \boldsymbol{x}_n^{(t,i)})$ is the players' joint strategy at iteration $i \in [\![m]\!]$.

Now the main challenge here is that the potential function is in general nonconcave (and nonconvex). Thus, employing a meta-regret minimizer for learning the initilization appears to be computationally prohibitive. While there are interesting settings in which the potential function is concave (see Appendix D.1.1), such as linear Fisher markets [Birnbaum et al., 2011], we will bypass the need for concavity in the general setting by using a rather simpler initialization. In particular, let us denote by $\Delta(\Phi, \Phi') \coloneqq \max_{\boldsymbol{x}}(\Phi(\boldsymbol{x}) - \Phi'(\boldsymbol{x}))$ the difference functional for two functions $\Phi, \Phi' : \bigtimes_{k=1}^{n} \mathcal{X}_k \to \mathbb{R}$. Then, we will use the following notion of similarity for the sequence of encountered potential games.

$$V_\Delta \coloneqq \frac{1}{T} \sum_{t=1}^{T-1} \Delta(\Phi^{(t)}, \Phi^{(t+1)}). \tag{72}$$

We are now ready to derive a refined bound for the second-order path lengths of `GD` in terms of (72). Similar results apply more broadly for other variants of mirror descent.

Corollary D.4. *Suppose that all players employ `GD` in a potential game with a sufficiently small learning rate $\eta > 0$ and initialization $\boldsymbol{x}_k^{(t,0)} \coloneqq \boldsymbol{x}_k^{(t-1,m)}$, for all $k \in [\![n]\!]$ and $t \ge 2$. Then,*

$$\frac{1}{2\eta T} \sum_{t=1}^{T} \sum_{i=1}^{m} \|\boldsymbol{x}^{(t,i)} - \boldsymbol{x}^{(t,i-1)}\|_2^2 \le \frac{2\Phi_{max}}{T} + V_\Delta,$$

where V_Δ is defined as in (72), and $\boldsymbol{x}^{(t,i)} = (\boldsymbol{x}_1^{(t,i)}, \dots, \boldsymbol{x}_n^{(t,i)})$ is the joint strategy at task $t \in [\![t]\!]$ and iteration $i \in [\![m]\!]$.

Proof. By applying Proposition D.3 for each task $t \in [\![T]\!]$, we have

$$\begin{aligned}
\frac{1}{2\eta}\sum_{t=1}^{T}\sum_{i=1}^{m}\|\boldsymbol{x}^{(t,i)}-\boldsymbol{x}^{(t,i-1)}\|_2^2 &\leq \sum_{t=1}^{T}\left(\Phi^{(t)}(\boldsymbol{x}^{(t,m)})-\Phi^{(t)}(\boldsymbol{x}^{(t,0)})\right)\\
&\leq 2\Phi_{\max}+\sum_{t=1}^{T-1}\left(\Phi^{(t)}(\boldsymbol{x}^{(t,m)})-\Phi^{(t+1)}(\boldsymbol{x}^{(t+1,0)})\right)\\
&\leq 2\Phi_{\max}+\sum_{t=1}^{T-1}\left(\Phi^{(t)}(\boldsymbol{x}^{(t,m)})-\Phi^{(t+1)}(\boldsymbol{x}^{(t,m)})\right) \qquad (73)\\
&\leq 2\Phi_{\max}+\sum_{t=1}^{T-1}\Delta(\Phi^{(t)},\Phi^{(t+1)}), \qquad (74)
\end{aligned}$$

where (73) uses the initialization $\boldsymbol{x}^{(t+1,0)} \coloneqq \boldsymbol{x}^{(t,m)}$ and the definition of $\Phi_{\max}$, and (74) follows from the definition of $\Delta(\cdot,\cdot)$. The statement then follows by averaging (74) over the T tasks and using (72). □

Thus, GD under the initialization of Corollary D.4 enjoys the meta-learning guarantee provided below. There are also consequences for the players' regrets, which we omit since they follow directly.

Corollary D.5 (Detailed Version of Theorem 3.4)**.** *Under the assumptions of Corollary D.4, for an average potential game*

$$m \coloneqq \left\lceil \frac{4\eta\Phi_{max}}{\epsilon^2 T}+\frac{2\eta V_{\Delta}}{\epsilon^2}\right\rceil$$

iterations suffice to reach an

$$\epsilon\left(\frac{\max_{k\in[\![n]\!]}\Omega_k}{\eta}+\sqrt{\max_{k\in[\![n]\!]}\mathcal{A}_k}\right)-\textit{approximate Nash equilibrium.}$$

Proof. By Corollary D.4, for an average task $t \in [\![T]\!]$ it follows that $\left\lceil \frac{4\eta\Phi_{\max}}{\epsilon^2 T}+\frac{2\eta V_{\Delta}}{\epsilon^2}\right\rceil$ iterations suffice so that $\|\boldsymbol{x}_k^{(t,i)}-\boldsymbol{x}_k^{(t,i-1)}\|_2 \leq \epsilon$, for some $i \in [\![m]\!]$, for all $k \in [\![n]\!]$. Then, the statement follows by [Anagnostides et al., 2022b, Claim B.7]. □

D.1.1 Refinements under Concave Potentials

Moreover, as we pointed out earlier, there are important settings for which the potential function is concave, such as linear Fisher markets [Birnbaum et al., 2011]; such settings are not necessarily expressed in strategic form, but are indeed readily amenable to our techniques. Then, one can use a meta-algorithm for learning the initialization that receives after the termination of each task $t \in [\![T]\!]$ the cost function

$$(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n) =: \boldsymbol{x} \mapsto \Phi^{(t)}(\boldsymbol{x}^{(t,m)})-\Phi^{(t)}(\boldsymbol{x}).$$

In particular, here it is assumed that $\Phi^{(t)}$ is also available after the end of the task—a rather stringent assumption compared to what is required in Corollary D.5. By concavity, the meta-learner will incur $o(T)$ regret, thereby leading to an $o_T(1)$ overhead in the per-task performance, which becomes negligible as $T \to +\infty$. Concavity also guarantees that the meta-learner can be implemented efficiently. Moreover, by virtue of Definition D.2, gradient-descent-type algorithms on the potential function can be in fact implemented in a full decentralized way by having each

player perform a local update based on the observed utility. As a result, this leads to a task similarity of the form

$$\min_{\boldsymbol{x}^\star} \sum_{t=1}^{T} \left(\Phi^{(t)}(\boldsymbol{x}^{(t,m)}) - \Phi^{(t)}(\boldsymbol{x}^\star) \right),$$

where $\boldsymbol{x}^\star \in \bigtimes_{k=1}^{n} \mathcal{X}_k$.

D.2 Coarse Correlated Equilibria

In this subsection, we study meta-learning for coarse correlated equilibria. We will use optimistic Hedge (`OptHedge`) as our base learner. In particular, we will use the following guarantee [Daskalakis et al., 2021, Theorem 3.1].

Theorem D.6 ([Daskalakis et al., 2021])**.** *There exist universal constants $C, C' > 0$ so that when all players employ `OptHedge` with learning rate $\eta < \frac{1}{Cn\log^4(m)}$, the regret of each player $k \in [\![n]\!]$ is bounded by*

$$\mathrm{Reg}_k^{(m)}(\mathring{\boldsymbol{x}}_k) \leq \frac{D_{KL}(\mathring{\boldsymbol{x}}_k \,\|\, \boldsymbol{x}_k^{(0)})}{\eta} + \eta C' \log^5(m).$$

We remark that [Daskalakis et al., 2021, Theorem 3.1] was stated slightly differently, but the statement we include here (Theorem D.6) follows readily from their analysis. Since $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ is non-Lipschitz near the (relative) boundary of the strategy set, we initialize `OptHedge` α-away from the boundary in each task. Since the optimal learning rate will generally depend on the task similarity, we meta-learn the learning rate by running exponentially-weighted online optimization (`EWOO`) [Hazan et al., 2007] over a sequence of strongly convex regret-upper-bounds $U_1, \ldots, U_T$ (to be specified at a later point in the proof), similar to previous work in meta-learning (*i.e.*, Khodak et al. [2019], Osadchiy et al. [2022]). Specifically, we set the learning rate in task t by

$$\eta^{(t)} = \frac{\int_\rho^{\sqrt{D^2+\rho^2 D^2}} \eta \exp(-\beta \sum_{s<t} U_s(\eta)) d\eta}{\int_\rho^{\sqrt{D^2+\rho^2 D^2}} \exp(-\beta \sum_{s<t} U_s(\eta)) d\eta},$$

where $\beta = \frac{2}{D}\min\{1, \frac{\rho^2}{D^2}\}$, $\frac{D_{\mathrm{KL}}(\tilde{\tilde{\boldsymbol{y}}}^{(t)} \,\|\, \boldsymbol{y}^{(t,0)})}{C' \log^5(m)} \leq D^2, \forall t$, and $\rho > 0$.

Since our regret-upper-bounds are of the form $C' \log^5(m) \left(\eta + \frac{D_{\mathrm{KL}}(\tilde{\tilde{\boldsymbol{x}}}^{(t)} \,\|\, \boldsymbol{x}^{(t,0)})}{C' \log^5(m)\eta} \right)$, they are non-smooth near zero and not strongly convex if $\tilde{\tilde{\boldsymbol{x}}}^{(t)} = \boldsymbol{x}^{(t,0)}$. Therefore, we run `EWOO` over the regularized sequence $U_1, \ldots, U_T$, where

$$U_t(\eta) = C' \log^5(m) \left(\eta + \frac{\frac{D_{\mathrm{KL}}(\tilde{\tilde{\boldsymbol{x}}}^{(t)} \,\|\, \boldsymbol{x}^{(t,0)})}{C' \log^5(m)} + D^2\rho^2}{\eta} \right).$$

We are now ready to present our main result for convergence to CCE.

Theorem D.7. *Let $\tilde{\boldsymbol{x}} := (1-\alpha)\boldsymbol{x} + \alpha \frac{1}{d}\mathbb{1}_d$. When all players employ `OptHedge` with $\boldsymbol{x}_k^{(t,0)} = \frac{1}{t-1}\sum_{s<t} \tilde{\tilde{\boldsymbol{x}}}^{(s)}$, learning rate given by `EWOO` [Hazan et al., 2007] with suitably chosen hyperparameters, and $\alpha = \frac{1}{\sqrt{Tm}}$, there exist universal constants $C, C' > 0$ so that the average regret of each*

player $k \in [\![n]\!]$ over a sequence of T repeated games is bounded by

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{Reg}_k^{(t,m)} &\le 2\sqrt{\frac{m}{T}} \\ &+ \sqrt{C'\log^5(m)\log(d\sqrt{mT})}\left(\min\left\{\frac{\sqrt{\log(d\sqrt{mT})}}{\eta^*\sqrt{TC'\log^5(m)}}, \frac{1}{T^{1/4}}\right\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right) \\ &+ \min_{0<\eta\le\bar{\eta}}\left\{\eta C'\log^5(m) + \frac{V_k^2}{\eta} + \frac{8d\sqrt{m}(1+\log T)}{\eta\sqrt{T}}\right\},\end{aligned}$$

where $\bar{\eta} = \frac{1}{Cn\log^4(m)}$ *and* $V_k^2 \coloneqq \frac{1}{T}\sum_{t=1}^{T} D_{KL}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \bar{\boldsymbol{x}}_k)$ *for* $\bar{\boldsymbol{x}}_k \coloneqq \frac{1}{T}\sum_{t=1}^{T}\mathring{\boldsymbol{x}}_k^{(t)}$.

Proof. By Theorem D.6,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{Reg}_k^{(t,m)} &\le 2\alpha m + \frac{1}{T}\sum_{t=1}^{T}\left(\eta^{(t)}C'\log^5(m) + \frac{1}{\eta^{(t)}}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)})\right) \\ &= 2\alpha m + \Delta_U + \frac{1}{T}\min_{0<\eta\le\bar{\eta}}\left\{\sum_{t=1}^{T}\left(\eta C'\log^5(m) + \frac{1}{\eta}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)})\right)\right\},\end{aligned}$$

where $\Delta_U \coloneqq \frac{1}{T}\sum_{t=1}^{T}\eta^{(t)}C'\log^5(m) + \frac{1}{\eta^{(t)}}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)}) - \frac{1}{T}\min_{0<\eta\le\bar{\eta}}\sum_{t=1}^{T}\eta C'\log^5(m)$ $+\frac{1}{\eta}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)})$.

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{Reg}_k^{(t,m)} &\le 2\alpha m + \Delta_U + \frac{1}{T}\min_{0<\eta\le\bar{\eta}}\left\{\sum_{t=1}^{T}\left(\eta C'\log^5(m)\right) + \min_{\boldsymbol{x}_k\in\Delta^d}\sum_{t=1}^{T}\left(\frac{1}{\eta}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k)\right)\right. \\ &\quad \left. + \sum_{t=1}^{T}\left(\frac{1}{\eta}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k^{(t,0)})\right) - \min_{\boldsymbol{x}_k\in\Delta^d}\sum_{t=1}^{T}\left(\frac{1}{\eta}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k)\right)\right\} \\ &\le 2\alpha m + \Delta_U + \frac{1}{T}\min_{0<\eta\le\bar{\eta}}\left\{\sum_{t=1}^{T}\left(\eta C'\log^5(m)\right)\right. \\ &\quad \left. + \min_{\boldsymbol{x}_k\in\Delta^d}\sum_{t=1}^{T}\left(\frac{1}{\eta}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \boldsymbol{x}_k)\right) + \frac{8d(1+\log T)}{\eta\alpha}\right\} \\ &= 2\alpha m + \Delta_U + \frac{1}{T}\min_{0<\eta\le\bar{\eta}}\left\{\sum_{t=1}^{T}\left(\eta C'\log^5(m)\right)\right. \\ &\quad \left. + \sum_{t=1}^{T}\left(\frac{1}{\eta}D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{x}}}_k^{(t)} \,\|\, \tilde{\bar{\boldsymbol{x}}}_k)\right) + \frac{8d(1+\log T)}{\eta\alpha}\right\} \\ &\le 2\alpha m + \Delta_U + \min_{0<\eta\le\bar{\eta}}\left\{\eta C'\log^5(m)\right. \\ &\quad \left. + \frac{1}{\eta T}\sum_{t=1}^{T}\left(D_{\mathrm{KL}}(\mathring{\boldsymbol{x}}_k^{(t)} \,\|\, \bar{\boldsymbol{x}}_k)\right) + \frac{8d(1+\log T)}{\eta\alpha T}\right\},\end{aligned}$$

where the second inequality follows from Balcan et al. [2022, Lemma A.1] with $S = \frac{d}{\alpha}$ and $K = 1$, the inequality follows from the fact that FTL over a sequence of Bregman divergences reduces to the average [Banerjee et al., 2005], and the third inequality follows from the joint convexity of KL divergence.

Next, we bound Δ_U using the regret guarantees of `EWOO`. In particular, we apply Khodak et al. [2019, Corollary C.2]:

Corollary D.8 (Khodak et al. [2019]). *Let $\{U^{(t)} : \mathbb{R}_+ \to \mathbb{R}\}_{t\geq 1}$ be a sequence of functions of form $U^{(t)}(\eta) = \left(\frac{(B^{(t)})^2}{\eta} + \eta\right)\gamma^{(t)}$ for any positive scalars $\gamma^{(1)}, \ldots, \gamma^{(T)} \in \mathbb{R}_+$ and adversarially chosen $B_t \in [0, D]$. Then the $\epsilon -$ `EWOO` algorithm, for which $\epsilon > 0$ uses the actions of `EWOO` run on the functions $\tilde{U}_t(\eta) = \left(\frac{(B^{(t)})^2+\epsilon^2}{\eta} + \eta\right)\gamma^{(t)}$ over the domain $[\epsilon, \sqrt{D^2+\epsilon^2}]$ to determine $\eta^{(t)}$ achieves regret*

$$\min\left\{\frac{\epsilon^2}{\eta^*}, \epsilon\right\}\sum_{t=1}^{T}\gamma^{(t)} + \frac{D\gamma_{max}}{2}\max\left\{\frac{D^2}{\epsilon^2}, 1\right\}(1+\log(T+1))$$

for all $\eta^* > 0$.

Applying Corollary D.8 with $\epsilon = \rho D$, $\gamma^{(t)} = C'\log^5(m)$, $\forall t \in [\![T]\!]$, $D = \frac{\sqrt{\log(d/\alpha)}}{\sqrt{C'\log^5(m)}}$, and $\rho = \frac{1}{T^{1/4}}$ we see that

$$\begin{aligned}\Delta_U &\leq DC'\log^5(m)\left(\min\{\frac{D}{\eta^*\sqrt{T}}, \frac{1}{T^{1/4}}\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right)\\ &= \sqrt{C'\log^5(m)\log(d/\alpha)}\left(\min\left\{\frac{\sqrt{\log(d/\alpha)}}{\eta^*\sqrt{TC'\log^5(m)}}, \frac{1}{T^{1/4}}\right\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right).\end{aligned}$$

Plugging in our upper bound for Δ_U into our task average regret-upper-bound, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{Reg}_k^{(t,m)} \leq\ & 2\alpha m\\ &+ \sqrt{C'\log^5(m)\log(d/\alpha)}\left(\min\left\{\frac{\sqrt{\log(d/\alpha)}}{\eta^*\sqrt{TC'\log^5(m)}}, \frac{1}{T^{1/4}}\right\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right)\\ &+ \min_{0<\eta\leq\bar{\eta}}\left\{\eta C'\log^5(m) + \frac{1}{\eta T}\sum_{t=1}^{T}\left(D_{\mathrm{KL}}(\mathring{\boldsymbol{x}}_k^{(t)} \parallel \bar{\boldsymbol{x}}_k)\right) + \frac{8d(1+\log T)}{\eta\alpha T}\right\}.\end{aligned}$$

By setting $\alpha = \frac{1}{\sqrt{mT}}$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{Reg}_k^{(t,m)} \leq\ & 2\sqrt{\frac{m}{T}}\\ &+ \sqrt{C'\log^5(m)\log(d\sqrt{mT})}\left(\min\left\{\frac{\sqrt{\log(d\sqrt{mT})}}{\eta^*\sqrt{TC'\log^5(m)}}, \frac{1}{T^{1/4}}\right\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right)\\ &+ \min_{0<\eta\leq\bar{\eta}}\left\{\eta C'\log^5(m) + \frac{V_k^2}{\eta} + \frac{8d\sqrt{m}(1+\log T)}{\eta\sqrt{T}}\right\},\end{aligned}$$

where $V_k^2 \coloneqq \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}(\mathring{\boldsymbol{x}}_k^{(t)} \parallel \bar{\boldsymbol{x}}_k)$. □

It is worth noting that one can obviate meta-learning the learning rate using an uncoupled version of *clairvoyant hedge* [Piliouras et al., 2021], albeit with the need to appropriately average the iterates.

D.3 Correlated Equilibria

Finally, we conclude this section with a refined meta-learning guarantee for converging to *correlated equilibria*, an important strengthening of coarse correlated equilibria. As in the previous subsection, we focus on normal-form games. As it turns out [Blum and Mansour, 2007], converging to the set of correlated equilibria requires minimizing a stronger notion of regret, referred to as *swap regret*:

$$\mathrm{SwapReg}_k^{(m)} \coloneqq \max_{\phi^\star \in \Phi} \left\{ \sum_{i=1}^{m} \langle \phi^\star(\boldsymbol{x}_k^{(i)}) - \boldsymbol{x}_k^{(i)}, \boldsymbol{u}_k^{(t)} \rangle \right\}, \tag{75}$$

where Φ includes all possible linear functions $\phi : \Delta(\mathcal{A}_k) \to \Delta(\mathcal{A}_k)$. Swap regret (75) is clearly a stronger notion of hindsight rationality, since Φ includes all constants transformations. To minimize swap regret, Blum and Mansour [2007] gave a general reduction to the problem of minimizing external regret, which we briefly describe for the sake of completeness.

The algorithm of Blum and Mansour Each player k maintains a separate regret minimizer for each available action $a \in \mathcal{A}_k$. The next strategy is computed by obtaining the next strategy of each of those individual regret minimizers, and then determining any stationary distribution $\boldsymbol{x}_k^{(i)}$ of the induced Markov chain—wherein each state corresponds to an action and the transition probabilities are given by the startegies of the individual regret minimizers. Finally, upon receiving as feedback a utility vector $\boldsymbol{u}_k^{(i)} \in \mathbb{R}^{d_k}$, it suffices to forward to the regret minimizer associated with action a the utility $\boldsymbol{u}_a^{(i)} \coloneqq \boldsymbol{x}_k^{(i)}[a]\boldsymbol{u}_k^{(i)}$, for all $a \in \mathcal{A}_k$.

In our meta-learning setting, we will use as base learner `OMD` with the log-barrier as regularizer. Recent work [Anagnostides et al., 2022a] established that this leads to an RVU bound for swap regret. This will also bypass the need to meta-learn the learning rate, as we had to do in Theorem D.7. We point out that the RVU bound for swap regret in [Anagnostides et al., 2022a] can be readily expressed in terms of the initialization (*e.g.*, see [Wei and Luo, 2018]), leading to the following conclusion.

Theorem D.9 ([Anagnostides et al., 2022a])**.** *Suppose that each player $k \in [\![n]\!]$ employs the no-swap-regret algorithm of Blum and Mansour [2007] instantiated with `OMD` under a logarithmic regularizer $\mathcal{R}_k : \boldsymbol{x}_k \mapsto -\sum_{a \in \mathcal{A}_k} \log \boldsymbol{x}_k[a]$. There exists a sufficiently small learning rate $\eta > 0$ such that*

$$\sum_{k=1}^{n} \mathrm{SwapReg}_k^{(m)} \leq \sum_{k=1}^{n} \sum_{a \in \mathcal{A}_k} \mathrm{Reg}_a^{(m)}(\mathring{\boldsymbol{x}}_a) \leq \frac{1}{\eta} \sum_{k=1}^{n} \sum_{a \in \mathcal{A}_k} \mathcal{B}_{\mathcal{R}_k}(\mathring{\boldsymbol{x}}_a \,\|\, \boldsymbol{x}_a^{(0)}). \tag{76}$$

In (76) we denote by $\mathrm{Reg}_a^{(m)}$ the *external* regret up to time $m \in \mathbb{N}$ of the regret minimizer associated with action $a \in \mathcal{A}_k$ of player $k \in [\![n]\!]$. We are now ready to show our refinement for swap regret in the meta-learning setting.

Theorem D.10. *Let $\tilde{\boldsymbol{x}}_k \coloneqq (1-\alpha)\boldsymbol{x}_k + \alpha \frac{1}{d} \mathbb{1}_{d_k}$, and suppose that each player $k \in [\![n]\!]$ employs the no-swap-regret algorithm of Blum and Mansour [2007] instantiated with `OMD` under a logarithmic regularizer $\mathcal{R}_k : \boldsymbol{x}_k \mapsto -\sum_{a \in \mathcal{A}_k} \log \boldsymbol{x}_k[a]$ with a sufficiently small learning rate $\eta > 0$ and $\boldsymbol{x}_a^{(t,0)} = \frac{1}{t-1} \sum_{s<t} \tilde{\mathring{\boldsymbol{x}}}_a^{(s)}$, for all $a \in \mathcal{A}_k$ and $t \geq 2$. Then, for $\alpha \coloneqq (mT)^{-1/3}$,*

$$\frac{1}{T} \sum_{t=1}^{T} \mathrm{SwapReg}_k^{(t,m)} \leq \frac{1}{\eta} \sum_{k=1}^{n} \sum_{a \in \mathcal{A}_k} \tilde{V}_{k,a}^2 + \frac{m^{2/3}}{T^{1/3}} \left(2n + \sum_{k=1}^{n} \frac{8 d_k^3 (1 + \log T)}{\eta} \right),$$

where $\tilde{V}_{k,a}^2 \coloneqq \frac{1}{T} \sum_{t=1}^{T} \mathcal{B}_{\mathcal{R}_k}(\tilde{\mathring{\boldsymbol{x}}}_a^{(t)} \,\|\, \bar{\boldsymbol{x}}_a)$ for $\bar{\boldsymbol{x}}_a = \frac{1}{T} \sum_{t=1}^{T} \tilde{\mathring{\boldsymbol{x}}}_a^{(t)}$ is the task similarity for the individual regret minimizer of player $k \in [\![n]\!]$ associated with action $a \in \mathcal{A}_k$.

Proof. By Theorem D.9,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{n}\mathrm{SwapReg}_k^{(t,m)} \le 2\alpha mn + \frac{1}{\eta T}\sum_{t=1}^{T}\sum_{k=1}^{n}\sum_{a\in\mathcal{A}_k}\mathcal{B}_{\mathcal{R}_k}(\tilde{\mathring{\boldsymbol{x}}}_a^{(t)} \,\|\, \boldsymbol{x}_a^{(t,0)}) \tag{77}$$

$$\le 2\alpha mn + \frac{1}{\eta}\sum_{k=1}^{n}\sum_{a\in\mathcal{A}_k}\frac{1}{T}\sum_{t=1}^{T}\mathcal{B}_{\mathcal{R}_k}(\tilde{\mathring{\boldsymbol{x}}}_a^{(t)} \,\|\, \bar{\boldsymbol{x}}_a) + \frac{1}{T}\sum_{k=1}^{n}\frac{8d_k^3(1+\log T)}{\eta\alpha^2}, \tag{78}$$

where (77) follows from Theorem D.9 and the fact that for any player $k \in [\![n]\!]$,

$$\begin{aligned}\sum_{a\in\mathcal{A}_k}\mathrm{Reg}_a^{(t,m)}(\mathring{\boldsymbol{x}}_a^{(t)}) &\le \sum_{a\in\mathcal{A}_k}\mathrm{Reg}_a^{(t,m)}(\tilde{\mathring{\boldsymbol{x}}}_a^{(t)}) + \alpha\sum_{i=1}^{m}\sum_{a\in\mathcal{A}_k}\|\boldsymbol{u}_a^{(t,i)}\|_\infty\|\mathring{\boldsymbol{x}}_a^{(t)} - \tilde{\mathring{\boldsymbol{x}}}_a^{(t)}\|_1\\ &\le \sum_{a\in\mathcal{A}_k}\mathrm{Reg}_a^{(t,m)}(\tilde{\mathring{\boldsymbol{x}}}_a^{(t)}) + 2\alpha\sum_{i=1}^{m}\sum_{a\in\mathcal{A}_k}\boldsymbol{x}_k^{(t,i)}[a]\\ &\le \sum_{a\in\mathcal{A}_k}\mathrm{Reg}_a^{(t,m)}(\tilde{\mathring{\boldsymbol{x}}}_a^{(t)}) + 2\alpha m,\end{aligned}$$

since $\boldsymbol{u}_a^{(t,i)} = \boldsymbol{x}_k^{(t,i)}[a]\boldsymbol{u}_k^{(t,i)}$ and $\|\boldsymbol{u}_k^{(t,i)}\|_\infty \le 1$ (by assumption), and (78) uses [Balcan et al., 2022, Lemma A.1] with $K = 1$ and $S = \frac{d_k^2}{\alpha^2}$. Finally, the statement follows by taking $\alpha := (mT)^{-1/3}$ and observing that $\sum_{k'=1}^{n}\mathrm{SwapReg}_{k'}^{(t,m)} \ge \mathrm{SwapReg}_k^{(t,m)}$, for any player $k \in [\![n]\!]$ and task $t \in [\![T]\!]$, since swap regret is always nonnegative.[6] □

E Proofs from Section 3.3: Meta-Learning in Stackelberg Games

Theorem E.1. *Given a sequence of T repeated Stackelberg security games with d targets, k attacker types, and within-game time-horizon m, running* `MWU` *over the set of extreme points $\mathcal{E}$ as defined in Balcan et al. with initialization $\boldsymbol{y}^{(t,0)} = \frac{1}{t-1}\sum_{s<t}\mathring{\boldsymbol{y}}^{(t)}$ and learning rate given by β-EWOO [Hazan et al., 2007] with suitably chosen hyperparameters achieves expected task-averaged Stackelberg regret*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathrm{StackReg}^{(t,m)}] &\le 4\sqrt{\frac{m}{T}}\\ &\quad + \sqrt{m\log(|\mathcal{E}|\sqrt{mT})}\left(\min\left\{\frac{\sqrt{\log(|\mathcal{E}|\sqrt{mT})}}{\eta^*\sqrt{Tm}}, \frac{1}{T^{1/4}}\right\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right)\\ &\quad + \min_{0<\eta\le\bar{\eta}}\left\{\eta m + \frac{H(\bar{\boldsymbol{y}})}{\eta} + \frac{8|\mathcal{E}|\sqrt{m}(\log T+1)}{\eta\sqrt{T}}\right\},\end{aligned}$$

where the sequence of attackers in each task can be adversarially chosen.

Proof. First, we have

$$\begin{aligned}\mathrm{StackReg}^{(t,m)} &= \sum_{i=1}^{m}\langle\mathring{\boldsymbol{x}}^{(t)}, \boldsymbol{u}^{(t)}(b_{f^{(t,i)}}(\mathring{\boldsymbol{x}}^{(t)}))\rangle - \langle\boldsymbol{x}^{(t,i)}, \boldsymbol{u}^{(t)}(b_{f^{(t,i)}}(\boldsymbol{x}^{(t,i)}))\rangle\\ &\le \max_{\boldsymbol{x}^{(t)}\in\mathcal{E}}\sum_{i=1}^{m}\langle\boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)}(b_{f^{(t,i)}}(\boldsymbol{x}^{(t)}))\rangle - \langle\boldsymbol{x}^{(t,i)}, \boldsymbol{u}^{(t)}(b_{f^{(t,i)}}(\boldsymbol{x}^{(t,i)}))\rangle + 2\gamma m,\end{aligned}$$

[6] Better rates may be possible by meta-learning the optimal boundary offset α as in Balcan et al. [2022, Theorem 5], *i.e.*, by running multiplicative weights over a discretized grid of boundary offset values.

where the inequality follows from Lemma 4.3 of Balcan et al. [2015a]. Let $\boldsymbol{y}^{(t,i)} \in \Delta^{|\mathcal{E}|}$ denote the distribution over actions the algorithm plays in game t at time i and $\boldsymbol{u}^{(t,i)} = [\langle \boldsymbol{x}, \boldsymbol{u}^{(t)}(b_{f^{(t,i)}}(\boldsymbol{x}))\rangle]_{\boldsymbol{x}\in\mathcal{E}} \in [-1,1]^{|\mathcal{E}|}$. Then,

$$\begin{aligned}
\mathbb{E}[\text{StackReg}^{(t,m)}] &\leq \sum_{i=1}^{m} \left(\langle \mathring{\boldsymbol{y}}^{(t)}, \boldsymbol{u}^{(t,i)} \rangle - \langle \boldsymbol{y}^{(t,i)}, \boldsymbol{u}^{(t,i)} \rangle \right) + 2\gamma m \\
&= \sum_{i=1}^{m} \left(\langle \mathring{\boldsymbol{y}}^{(t)}, \boldsymbol{u}^{(t,i)} \rangle - \langle \tilde{\mathring{\boldsymbol{y}}}^{(t)}, \boldsymbol{u}^{(t,i)} \rangle + \langle \tilde{\mathring{\boldsymbol{y}}}^{(t)}, \boldsymbol{u}^{(t,i)} \rangle - \langle \boldsymbol{y}^{(t,i)}, \boldsymbol{u}^{(t,i)} \rangle \right) + 2\gamma m \\
&\leq 2\gamma m + 2\alpha m + \eta^{(t)} m + \frac{1}{\eta^{(t)}} D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)}),
\end{aligned}$$

where $\tilde{\mathring{\boldsymbol{y}}}^{(t)} := (1-\alpha)\mathring{\boldsymbol{y}}^{(t)} + \frac{\alpha}{|\mathcal{E}|}\mathbb{1}_{|\mathcal{E}|}$. As was the case in Theorem D.7, it is necessary to initialize α-away from the boundary of the strategy space and meta-learn the learning rate using `EWOO` [Hazan et al., 2007]. We refer the reader to the proof of Theorem D.7 and references therein for more details about `EWOO`.

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\text{StackReg}^{(t,m)}] &\leq 2\alpha m + 2\gamma m + \frac{1}{T}\sum_{t=1}^{T} \left(\eta^{(t)} m + \frac{D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)})}{\eta^{(t)}} \right) \\
&= 2\alpha m + 2\gamma m + \frac{1}{T} \min_{\eta>0} \left\{ \sum_{t=1}^{T} \left(\eta m + \frac{D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)})}{\eta} \right) \right\} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T} \left(\eta^{(t)} m + \frac{D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)})}{\eta^{(t)}} \right) \\
&\quad - \frac{1}{T} \min_{\eta>0} \left\{ \sum_{t=1}^{T} \left(\eta m + \frac{D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)})}{\eta} \right) \right\} \\
&= 2\alpha m + 2\gamma m + \Delta_U + \frac{1}{T} \min_{\eta>0} \left\{ \sum_{t=1}^{T} \left(\eta m + \frac{D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)})}{\eta} \right) \right\},
\end{aligned}$$

where $\Delta_U := \frac{1}{T}\sum_{t=1}^{T} \eta^{(t)} m + \frac{D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)})}{\eta^{(t)}} - \frac{1}{T}\min_{0<\eta\leq\bar{\eta}} \sum_{t=1}^{T} \eta m + \frac{D_{\mathrm{KL}}(\tilde{\mathring{\boldsymbol{y}}}^{(t)} \parallel \boldsymbol{y}^{(t,0)})}{\eta}$.

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\text{StackReg}^{(t,m)}] &\le 2\alpha m + 2\gamma m + \Delta_U + \min_{\eta>0}\left\{\eta m + \frac{1}{T}\min_{\boldsymbol{y}\in\Delta^{|\mathcal{E}|}}\sum_{t=1}^{T}\frac{D_{\mathrm{KL}}(\tilde{\hat{\boldsymbol{y}}}^{(t)} \,\|\, \boldsymbol{y})}{\eta}\right.\\
&\quad \left. + \frac{1}{T}\sum_{t=1}^{T}\frac{D_{\mathrm{KL}}(\tilde{\hat{\boldsymbol{y}}}^{(t)} \,\|\, \boldsymbol{y}^{(t,0)})}{\eta} - \frac{1}{T}\min_{\boldsymbol{y}\in\Delta^{|\mathcal{E}|}}\sum_{t=1}^{T}\frac{D_{\mathrm{KL}}(\tilde{\hat{\boldsymbol{y}}}^{(t)} \,\|\, \boldsymbol{y})}{\eta}\right\}\\
&\le 2\alpha m + 2\gamma m + \Delta_U\\
&\quad + \min_{\eta>0}\left\{\eta m + \frac{1}{T}\min_{\boldsymbol{y}\in\Delta^{|\mathcal{E}|}}\sum_{t=1}^{T}\left(\frac{D_{\mathrm{KL}}(\tilde{\hat{\boldsymbol{y}}}^{(t)} \,\|\, \boldsymbol{y})}{\eta}\right) + \frac{8|\mathcal{E}|(\log T+1)}{\eta\alpha T}\right\}\\
&= 2\alpha m + 2\gamma m + \Delta_U\\
&\quad + \min_{\eta>0}\left\{\eta m + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{D_{\mathrm{KL}}(\tilde{\hat{\boldsymbol{y}}}^{(t)} \,\|\, \tilde{\bar{\boldsymbol{y}}})}{\eta}\right) + \frac{8|\mathcal{E}|(\log T+1)}{\eta\alpha T}\right\}\\
&\le 2\alpha m + 2\gamma m + \Delta_U\\
&\quad + \min_{\eta>0}\left\{\eta m + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{D_{\mathrm{KL}}(\mathring{\boldsymbol{y}}^{(t)} \,\|\, \bar{\boldsymbol{y}})}{\eta}\right) + \frac{8|\mathcal{E}|(\log T+1)}{\eta\alpha T}\right\}\\
&= 2\alpha m + 2\gamma m + \Delta_U + \min_{\eta>0}\left\{\eta m + \frac{H(\bar{\boldsymbol{y}})}{\eta} + \frac{8|\mathcal{E}|(\log T+1)}{\eta\alpha T}\right\},
\end{aligned}
$$

where the second inequality follows from Lemma A.1 of Balcan et al. [2022] with $S = \frac{|\mathcal{E}|}{\alpha}$ and $K = 1$ and the first equality follows from the fact that `FTL` over a sequence of Bregman divergences reduces to the average [Banerjee et al., 2005].

Applying Corollary D.8 with $\epsilon = \rho D$, $\gamma^{(t)} = m$, $\forall t \in [\![T]\!]$, $D = \frac{\sqrt{\log(|\mathcal{E}|/\alpha)}}{\sqrt{m}}$, and $\rho = \frac{1}{T^{1/4}}$ we see that

$$
\begin{aligned}
\Delta_U &\le Dm\left(\min\{\frac{D}{\eta^*\sqrt{T}}, \frac{1}{T^{1/4}}\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right)\\
&= \sqrt{m\log(|\mathcal{E}|/\alpha)}\left(\min\left\{\frac{\sqrt{\log(|\mathcal{E}|/\alpha)}}{\eta^*\sqrt{Tm}}, \frac{1}{T^{1/4}}\right\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\text{StackReg}^{(t,m)}] &\le 2\alpha m + 2\gamma m\\
&\quad + \sqrt{m\log(|\mathcal{E}|/\alpha)}\left(\min\left\{\frac{\sqrt{\log(|\mathcal{E}|/\alpha)}}{\eta^*\sqrt{Tm}}, \frac{1}{T^{1/4}}\right\} + \frac{1+\log(T+1)}{2\sqrt{T}}\right)\\
&\quad + \min_{0<\eta\le\bar{\eta}}\left\{\eta m + \frac{H(\bar{\boldsymbol{y}})}{\eta} + \frac{8|\mathcal{E}|(\log(T)+1)}{\eta\alpha T}\right\}.
\end{aligned}
$$

Setting $\alpha = \gamma = \frac{1}{\sqrt{mT}}$ completes the proof. □

F Further Experimental Results

In this section, we present some additional experimental results we omitted earlier from Section 4. In particular, Figures 2 and 3 illustrate the task-averaged Nash equilibrium gap of `OGD` under different values for the learning rate.

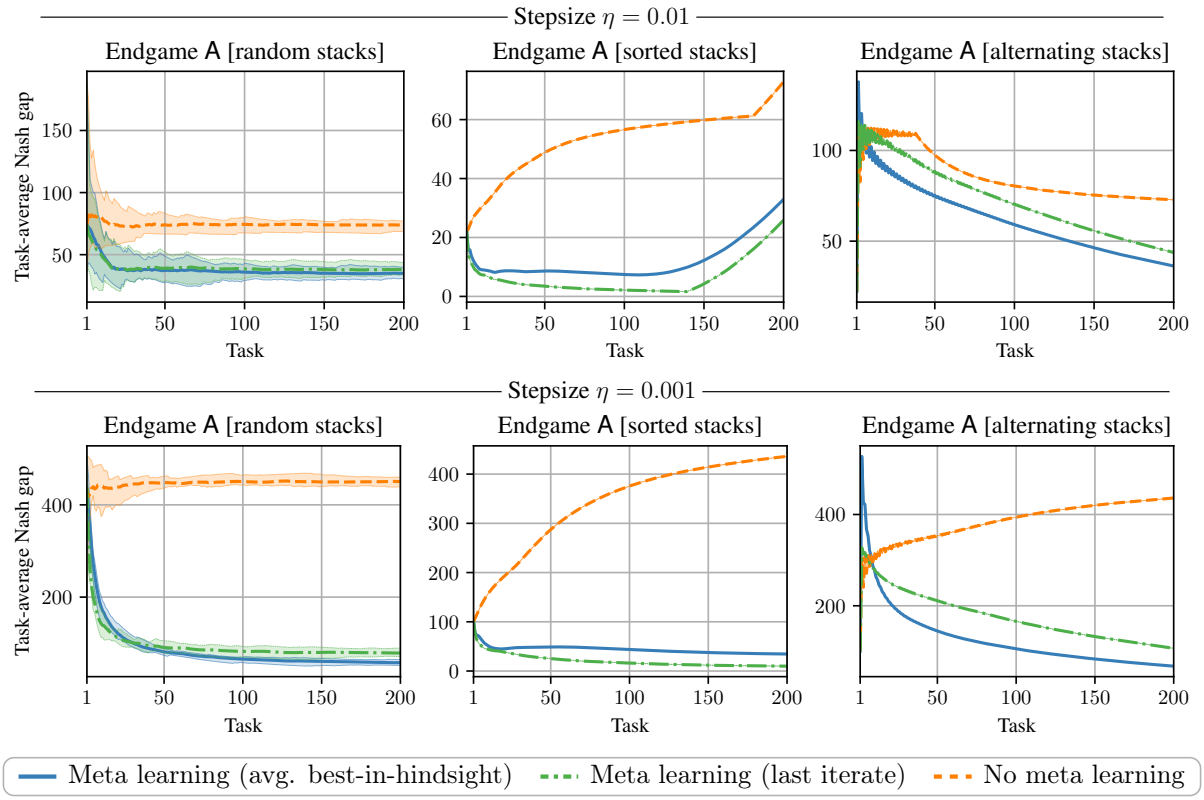

Figure 2: Vanilla OGD versus our meta-learning versions of OGD for different values of the learning rate η in Endgame A. Results for $\eta = 0.1$ are not included, as it was too large of a learning rate for any of the methods to converge in this endgame.

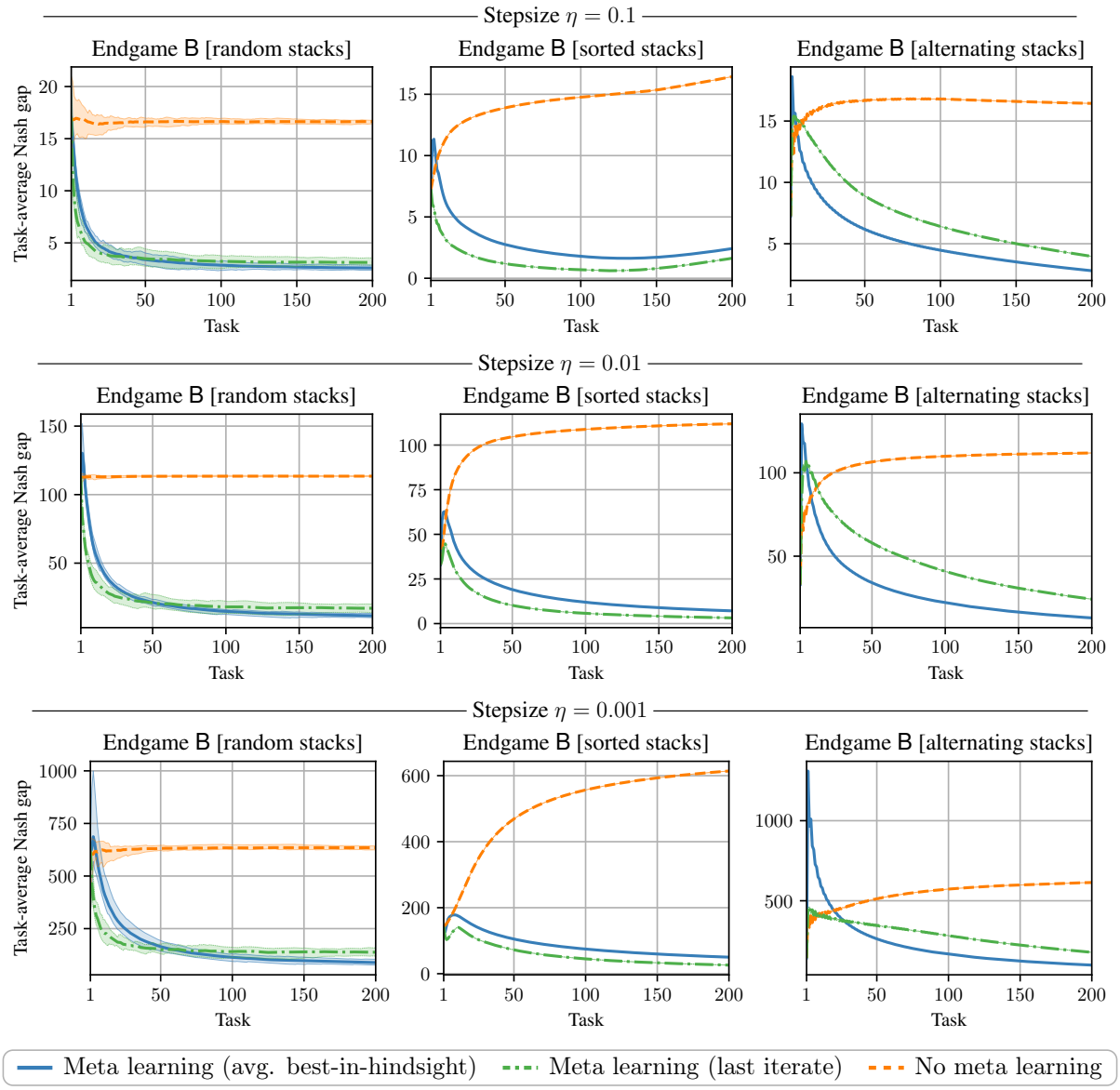

Figure 3: Vanilla `OGD` versus our meta-learning versions of `OGD` for different values of the learning rate η in Endgame B.