CONVERGENCE IN MIN-MAX OPTIMIZATION

A Dissertation
Presented to
The Academic Faculty

By

Kevin A. Lai

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Algorithms, Combinatorics, and Optimization

Georgia Institute of Technology

May 2020

CONVERGENCE IN MIN-MAX OPTIMIZATION

Approved by:

Dr. Jacob Abernethy, Advisor
School of Computer Science
Georgia Institute of Technology

Dr. Rachel Cummings
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jamie Morgenstern
School of Computer Science and Engineering
University of Washington

Dr. Sebastian Pokutta
Institute of Mathematics
Technische Universität Berlin

Dr. Mohit Singh
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: March 12, 2020

To my parents

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor Jake Abernethy for his guidance and support throughout my PhD. Jake was always thoughtful and kind, and it was truly a pleasure working with him. I would also like to thank Rachel Cummings for her advice and help both during my PhD and with my job applications. I was fortunate to have Satyen Kale as a mentor at Google, and I am grateful for his help with my job applications as well. I am grateful to Jake Abernethy, Rachel Cummings, Jamie Morgenstern, Sebastian Pokutta, and Mohit Singh for being part of my thesis committee, with particular thanks to Rachel Cummings for being the reader for my thesis. I am also grateful to my collaborators throughout my PhD: Jake Abernethy, Brian Bullins, Rachel Cummings David Durfee, Sara Krehbiel, Kfir Levy, Richard Peng, Anup Rao, Saurabh Sawlani, Tao Tantipongpipat, Santosh Vempala, Jun-Kun Wang, and Andre Wibisono.

I was lucky to have a great cohort at Georgia Tech, and I am glad I went through the PhD with Digvijay Boob, Matthew Fahrbach, Anna Kirkpatrick, Samantha Petti, Samira Samadi, Saurabh Sawlani, and Peng Zhang. I am also grateful for the guidance of ACO students who were senior to me when I started my PhD: Prateek Bhakta, Sarah Cannon, Ben Cousins, David Durfee, and Sadra Yazdanbod. I am grateful for the constant support of my girlfriend Dantong Zhu. Finally, I would like to thank my parents for their unwavering support and belief in me.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

Min-max optimization is a classic problem with applications in constrained optimization, robust optimization, and game theory. This dissertation covers new convergence rate results in min-max optimization. We show that the classic fictitious play dynamic with lexicographic tiebreaking converges quickly for diagonal payoff matrices, partly answering a conjecture by Karlin from 1959. We also show that linear last-iterate convergence rates are possible for the HAMILTONIAN GRADIENT DESCENT algorithm for the class of "sufficiently bilinear" min-max problems. Finally, we explore higher-order methods for min-max optimization and monotone variational inequalities, showing improved iteration complexity compared to first-order methods such as Mirror Prox.

CHAPTER 1
INTRODUCTION

Game dynamics have been central to many exciting recent developments in machine learning. In some cases, game-playing is an inherent part of the problem, as in Deepmind's Alphastar program for playing Starcraft [Vin+19]. In other cases, game dynamics are used as a tool to train complex systems, as in Generative Adversarial Networks (GANs) [Goo+14]. These applications often involve finding a *Nash Equilibrium* in a zero-sum game, which is equivalent to *min-max optimization*.

This work addresses several open questions related to solving min-max optimization problems. We begin in Section 1.1 by describing some of the many settings where min-max problems arise. We then summarize the results of this thesis in Section 1.2.

1.1 Examples of min-max problems

Min-max problems typically take the following form:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y) \tag{1.1}$$

where $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ maps from constraint sets $\mathcal{X}$ and $\mathcal{Y}$ to some real number. Solving (1.1) is equivalent to finding the Nash Equilibrium in a zero-sum game, and this perspective has been fruitful in developing algorithms for optimization problems. One of the earliest applications of min-max optimization was in solving linear programs, where duality of linear programs can be viewed as a consequence of the min-max theorem [Dan51; Adl13]. This perspective has led to the development of *primal-dual* algorithms for solving linear programs. In addition to explicitly motivating algorithms, the min-max optimization perspective has also provided useful interpretations of existing algorithms. For instance, the Boosting

algorithm of [FS96] can be viewed as a game between a player that selects distributions and a player that chooses a weak oracle. Another recent line of work [AW17; ALLW18a; WA18] explored convex optimization through a min-max formulation called the *Fenchel game*. Given an optimization problem over a convex function f, we can write:

$$\min_x f(x) = \min_x \max_y \langle y, x \rangle - f^*(y)$$

where f^* is the Fenchel conjugate of f. As shown in [AW17; ALLW18a; WA18], popular convex optimization algorithms such as Nesterov's accelerated gradient descent [Nes83] and Frank-Wolfe [FW56] can be written in terms of certain no-regret update rules for the x and y players in the Fenchel game. This representation also provides new methods for proving convergence rates of these algorithms.

When (1.1) is viewed as a two-player game, the solution to (1.1) can be thought of as a point x^* that is robust to all possible plays of the y-player. This notion of robustness has been useful in many domains. For example, one can write a constrained optimization problem as an equivalent augmented Lagrangian as follows:

$$\min_{\substack{x \in \mathcal{X} \\ \text{s.t. } \forall i \in [n],\, h_i(x)=0}} f(x) = \min_{x \in \mathcal{X}} \max_{\lambda} f(x) - \sum_{i=1}^{n} \lambda_i h_i(x). \tag{1.2}$$

Then the resulting min-max problem can be viewed as a two-player zero-sum game in which the λ-player wants to find indices where $h_i(x) \neq 0$ to maximize her reward, while the x-player wants to find a point that is robust to the constraint player's actions. [FS96] show how this perspective can be useful when the number of constraints is large, as one can find an approximate min-max with suboptimality that scales only log arithmically with the number of constraints. This approach has been applied to domains such as differential privacy and fairness [HRU13; Aga+18]. Min-max problems also arise naturally in the context of adversarial robustness, in which one wants to guarantee accuracy bounds for a classifier, such as a neural network, in the face of inputs that undergo small adversarial perturbations.

For instance, [Mad+18] describe adversarial robustness as solving the following min-max problem:

$$\min_{\theta} \rho(\theta), \text{where } \mathbb{E}_{(x,y)\sim D}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right]$$

where D is a distribution over true example-label pairs, $\mathcal{S}$ is some space of perturbations (e.g. an ℓ_∞ ball), and $L(\theta, x, y)$ is a loss function for a classifier θ on an input x with true label y.

1.2 Summary of Contributions

In this section, we summarize the contributions in this dissertation, which relate to algorithms for solving min-max problems. Just as gradient descent is ubiquitous in vanilla optimization problems, the analogous *gradient descent/ascent* (GDA) dynamic is a popular algorithm for min-max optimization. GDA is an instance of a no-regret algorithm, which is a broad and well-studied class of algorithms that has provable guarantees for min-max settings. While no-regret algorithms have seen extensive use, there are many cases in which such algorithms are undesirable and where the no-regret analysis no longer provides provable guarantees. One example is GAN training, in which GDA and other no-regret algorithms can provably lead to cycling and non-convergence. Given these limitations, it is natural to ask whether one can solve min-max problems without using the no-regret framework.

This dissertation focuses on several new results that go beyond the no-regret framework. In Chapter 3, we explore the classic *fictitious play* (FP) dynamic of [Bro49] for solving Nash Equilibria in zero-sum matrix games. FP is a simple dynamic that does not fall under the no-regret framework. Nearly 70 years ago it was shown by Robinson [Rob51] that FP does converge to the Nash Equilibrium, although the rate she proved was exponential in the total number of actions of the players. In 1959, Karlin [Kar59] conjectured that FP converges at the more natural rate of $O(1/\sqrt{k})$. However, Daskalakis and Pan [DP14] disproved a version

of this conjecture in 2014, showing that an exponentially-slow rate can occur, although their result relied on adversarial tie-breaking. We show that Karlin's conjecture is indeed correct in two major instances if you appropriately handle ties. First, we show that if the game matrix is diagonal and ties are broken lexicographically, then FP converges at a $O(1/\sqrt{k})$ rate, and we also show a matching lower bound under this tie-breaking assumption. Our second result shows that FP converges at a rate of $O(1/\sqrt{k})$ when the players' decision sets are smooth, and $\tilde{O}(1/k)$ under an additional assumption. In this last case, we also show that a modification of FP, known as Optimistic FP, converges at a rate of $O(1/k)$. This chapter is based on joint work with Andre Wibisono and Jacob Abernethy [ALW19a].

In Chapter 4, we focus on *last-iterate convergence* guarantees, motivated by nonconvex min-max problems in which iterate averaging is undesirable, such as the GAN setting. While the no-regret framework gives average-iterate convergence results in convex-concave problems, it says virtually nothing about the last-iterates of no-regret dynamics. In fact, one can show that a broad class of no-regret algorithms provable diverge or cycle even in simple convex-concave games [MPP18], and previous work on global last-iterate convergence rates has been limited to the bilinear and convex-strongly concave settings. We show that the HAMILTONIAN GRADIENT DESCENT (HGD) algorithm achieves linear convergence in a variety of more general settings, including convex-concave problems that satisfy a novel sufficiently bilinear condition. We also prove convergence rates for stochastic HGD and for some parameter settings of the Consensus Optimization algorithm of [MNG17]. This chapter is based on joint work with Andre Wibisono and Jacob Abernethy [ALW19b].

In Chapter 5, we provide higher-order methods for solving constrained convex-concave min-max problems and monotone variational inequalities with higher-order smoothness. No-regret algorithms are typically first-order, and lower bounds prevent first-order algorithms from achieving better than $\Omega(1/k)$ iteration complexity. We are able to improve upon the iteration complexity of first-order methods by using higher-order methods. In the min-max setting, we give an algorithm HIGHERORDERMIRRORPROX that achieves an iteration

complexity of $O(1/k^{\frac{p+1}{2}})$ when given access to an oracle for minimizing a p^{th} order Taylor expansion and when the p^{th}-order derivatives are Lipschitz continuous. We give analogous rates for the weak monotone variational inequality problem. For $p > 2$, our results improve on the iteration complexity of the first-order Mirror Prox method of [Nem04] and the second-order method of [MS12]. We further instantiate our entire algorithm in the unconstrained $p = 2$ case. This chapter is based on joint work with Brian Bullins [BL19].

1.3 Notation and basic definitions

We now review some basic notation and definitions. We go over some more background on game theory and common approaches for solving min-max problems in Chapter 2.

We use $[n]$ to denote the set $\{1, ..., n\}$. I_n denotes the $n \times n$ identity matrix. We let e_i denote the i^{th} elementary basis vector. For a vector v, we let $v(i)$ denote the i^{th} entry of v. Let $\Delta_n = \{x \in \mathbb{R}^n \colon x_i \geq 0, \sum_{i=1}^n x_i = 1\}$ be the $(n-1)$-dimensional simplex. We use ∇ to denote the Jacobian operator. We use $||\cdot||$ to denote an arbitrary norm and $||\cdot||_*$ to denote its dual norm. We use $||\cdot||_2$ to denote the Euclidean norm for vectors and the operator norm for matrices. For a symmetric matrix A, we will use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the smallest and largest eigenvalues of A. For a general real matrix A, $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote the smallest and largest singular values of A.

We use $D : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ to denote a Bregman divergence over a distance generating function $d : \mathcal{Z} \to \mathbb{R}$ that is 1-strongly convex with respect to some norm $||\cdot||$. Recall that the definition of a Bregman divergence is as follows:

$$D(u, v) = d(u) - d(v) - \langle \nabla d(v), u - v \rangle \tag{1.3}$$

for all $u, v \in \mathcal{Z}$.

Definition 1.3.1. *A critical point of $f : \mathbb{R}^d \to \mathbb{R}$ is a point $x \in \mathbb{R}^d$ such that $\nabla f(x) = 0$.*

Definition 1.3.2 (Convexity / Strong convexity)**.** *Let $\mu \geq 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is*

μ-strongly convex if for any $u, v \in \mathbb{R}^d$, $f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle + \frac{\mu}{2} ||u - v||$. When f is twice-differentiable, f is μ-strongly-convex iff for all $x \in \mathbb{R}^d$, $\nabla^2 f(x) \succeq \mu I$. If $\mu = 0$ in either of the above definitions, f is called convex.

Definition 1.3.3 (Monotone / Strongly monotone)**.** *Let $\mu \geq 0$. A vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ is μ-strongly monotone if for any $x, y \in \mathbb{R}^d$, $\langle x - y, v(x) - v(y) \rangle \geq \mu ||x - y||^2$. If $\mu = 0$, v is called monotone.*

Given a min-max optimization objective $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we will often consider x and y to be components of one vector $z = (x\ ,\ y)$. We will use subscripts to denote iterate indices. Following [Bal+18], we use

$$\xi = (\nabla_x g, -\nabla_y g) \tag{1.4}$$

to denote the signed vector of partial derivatives.

CHAPTER 2

BACKGROUND

In this chapter, we review some important background on game theory and min-max optimization. A two-player zero-sum game is defined by an objective $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that in every round of the game, the x and y player choose $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively and then the x player pays $g(x, y)$ to the y player. As such, the x player would like to minimize $g(x, y)$ and the y player would like to maximize $g(x, y)$. The x player would like to solve the min-max problem (1.1), as doing so will guarantee for herself the *minimax* value of the game, which she achieves by playing the *minimax point* $x^* = \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y)$. Likewise, the y player wants to play the *maximin point* $y^* = \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} g(x, y)$, which guarantees her the *maximin* value of the game $\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} g(x, y)$.

The pair of points (x^*, y^*) forms a *Nash Equilibrium*, i.e. it satisfies the following inequality for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$g(x^*, y) \leq g(x^*, y^*) \leq g(x, y^*) \tag{2.1}$$

We will also call such points *min-max solutions* or *saddle point* solutions. We can see that solving the minimax and maximin problems for each player respectively is equivalent to finding the Nash Equilibrium of the zero-sum game.

One popular class of zero-sum games is the class of *convex-concave* games, where g is a continuous function that is convex in its first argument and concave in its second argument and $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets. One of the most fundamental results in game theory is Von Neumann's celebrated min-max theorem [Neu28], which holds for convex-concave

games, and which states the following:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y) = v = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} g(x, y) \tag{2.2}$$

where v is a real value we will call the *value* of the game.

2.1 Approximately solving a game

We seek algorithms to find approximate Nash Equilibria or approximate min-maxes in convex-concave zero-sum games. One classic and natural solution concept is the *duality gap* $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$:

$$\psi(x, y) = \max_{\hat{y} \in \mathcal{Y}} g(x, \hat{y}) - \min_{\hat{x} \in \mathcal{X}} g(\hat{x}, y) \tag{2.3}$$

The duality gap is implicitly defined in terms of a min-max objective g, but we leave it implicit because the relevant g will be clear from context. We can see that:

$$\psi(x, y) \geq \max_{\hat{y} \in \mathcal{Y}} g(x, \hat{y}) - \min_{\hat{x} \in \mathcal{X}} \max_{\hat{y} \in \mathcal{Y}} g(\hat{x}, \hat{y})$$

$$\text{and } \psi(x, y) \geq \max_{\hat{y} \in \mathcal{Y}} \min_{\hat{x} \in \mathcal{X}} g(\hat{x}, \hat{y}) - \min_{\hat{x} \in \mathcal{X}} g(\hat{x}, y)$$

From (2.2), we can then see that if $\psi(x, y) \leq \epsilon$, then $g(x, y)$ is within ϵ of v, so both players achieve within ϵ of their optimum payoff value.

One of the oldest algorithms for finding a min-max is the *fictitious play* (FP) algorithm proposed by Brown in 1949 [Bro49; Bro51]. In 1951, Robinson proved that FP converges to a min-max at a rate of $O(1/k^{\frac{1}{2n-2}})$. FP applies to the *matrix game* case where $g(x, y) = x^\top A y$ for some matrix $A \in \mathbb{R}^{n \times m}$ and where $\mathcal{X}$ and $\mathcal{Y}$ are probability simplices. Later advances [Bla56; Han57; FS99] showed a general method for finding min-max points in convex-concave games using *no-regret* online learning algorithms, which we describe in the next section.

2.1.1 Using no-regret algorithms to solve games

The online learning setting takes place over a series of K rounds, where in each round k, the learner plays some iterate z_k from a convex set $\mathcal{Z}$ and receives a convex payoff function $\ell_k : \mathcal{Z} \to \mathbb{R}$. The goal is to minimize the *regret*, defined as:

$$\text{Regret}_K = \sum_{k=1}^{K} \ell_k(z_k) - \min_{z \in \mathcal{Z}} \sum_{k=1}^{K} \ell_k(z) \tag{2.4}$$

Essentially, the regret measures how well the algorithm performs compared to the single best point within $\mathcal{Z}$. A *no-regret* algorithm is one such that the *average regret* $\frac{\text{Regret}_K}{K}$ goes to 0 as K goes to infinity. An important feature of the online learning framework is that the loss functions ℓ_k may be chosen completely adversarially, which means that algorithms with no-regret algorithms are in some sense robust.

One classic application of no-regret algorithms is to find approximate Nash Equilibria in convex-concave zero-sum games [Bla56; Han57; FS99]. To do so, we use the following procedure:

Algorithm 1 No-regret algorithms for solving a game

Input: $K > 0$

for $k = 1$ **to** K **do**

x_k is selected according to no-regret algorithm OAlg^x

y_k is selected according to (possibly different) no-regret algorithm OAlg^y

x-player receives loss function $\ell_k^x(\cdot) = g(\cdot, y_k)$

y-player receives loss function $\ell_k^y(\cdot) = -g(x_k, \cdot)$

end for

Let $(\bar{x}_K, \bar{y}_K) = (\frac{1}{K}\sum_{k=1}^{K} x_k, \frac{1}{K}\sum_{k=1}^{K} y_k)$

return $(\bar{x}, \bar{y})$

From this procedure, we can prove the following classic theorem:

Theorem 2.1.1. *Suppose we run Algorithm 1 with algorithms OAlg^x and OAlg^y whose regret after K rounds is bounded by* Regret_K^x *and* Regret_K^y *respectively. Then the output of Algorithm 1 satisfies* $\psi(\bar{x}, \bar{y}) \leq \frac{\text{Regret}_K^x + \text{Regret}_K^y}{K}$.

Proof. By the regret bound of OAlg^x, we have:

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} g(x_k, y_k) &\leq \min_{x \in \mathcal{X}} \frac{1}{K}\sum_{k=1}^{K} g(x, y_k) + \frac{\text{Regret}_K^x}{K} \\ &\leq \min_{x \in \mathcal{X}} g(x, \bar{y}_K) + \frac{\text{Regret}_K^x}{K}\end{aligned}$$

where the second inequality follows by Jensen's inequality and the fact that g is concave in y. Likewise, we can use the regret bound of OAlg^y to show:

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} g(x_k, y_k) &\geq \max_{y \in \mathcal{Y}} \frac{1}{K}\sum_{k=1}^{K} g(x_k, y) - \frac{\text{Regret}_K^y}{K} \\ &\geq \max_{y \in \mathcal{Y}} g(\bar{x}_K, y) - \frac{\text{Regret}_K^y}{K}\end{aligned}$$

Putting these two together, we get:

$$\max_{y \in \mathcal{Y}} g(\bar{x}_K, y) - \min_{x \in \mathcal{X}} g(x, \bar{y}_K) \leq \frac{\text{Regret}_K^y}{K} + \frac{\text{Regret}_K^x}{K}$$

□

CHAPTER 3

FAST CONVERGENCE OF FICTITIOUS PLAY

In this chapter, we consider algorithms for finding Nash Equilibria in zero-sum games where $g(x, y) = x^\top Ay$. Fictitious play (FP) is one of the simplest and most natural dynamics for such games. Originally proposed by [Bro49], FP is still popular today — a variant of it was used in training Deepmind's AlphaStar [Vin+19]. The FP dynamic imagines that each player considers the empirical distribution of the actions of the other player and selects their action as the best response to this statistic. Nearly 70 years ago it was shown by Robinson [Rob51] that FP does converge to the Nash Equilibrium, although the rate she proved was exponential in the total number of actions of the players. In 1959, Karlin [Kar59] conjectured that FP converges at the more natural rate of $O(1/\sqrt{k})$. However, Daskalakis and Pan [DP14] disproved a version of this conjecture in 2014, showing that an exponentially-slow rate can occur, although their result relied on adversarial tie-breaking. In this chapter, we show that Karlin's conjecture is indeed correct in two major instances if you appropriately handle ties. First, we show that if the game matrix is diagonal and ties are broken lexicographically, then FP converges at a $O(1/\sqrt{k})$ rate, and we also show a matching lower bound under this tie-breaking assumption. Our second result shows that FP converges at a rate of $O(1/\sqrt{k})$ when the players' decision sets are smooth, and $\tilde{O}(1/k)$ under an additional assumption. In this last case, we also show that a modification of FP, known as Optimistic FP, converges at a rate of $O(1/k)$.

3.1 Introduction

The FP dynamic of [Bro49] imagines that each player considers the empirical distribution of the actions of the other player and selects their action as the best response to this statistic. Mathematically speaking, we can define state variables x_k, y_k at each iteration k and update

according to the rule

$$\begin{aligned} x_{k+1} &= x_k + \arg\min_{x \in \Delta_n} \; x^\top A y_k \\ y_{k+1} &= y_k + \arg\max_{y \in \Delta_m} \; x_k^\top A y. \end{aligned} \tag{3.1}$$

Despite its simplicity, there still remain unanswered questions regarding the FP dynamic. Julia Robinson [Rob51] proved in the 1950s that the duality gap of the scaled state variables $(\hat{x}_k, \hat{y}_k) = (\frac{1}{k}x_k, \frac{1}{k}y_k)$ is bounded by $O(1/k^{\frac{1}{n+m-2}})$. Robinson's result utilized a recursive argument that introduced a $\frac{1}{k}$ factor for each available action of the players, and she did not address whether this was a tight rate. In what is often known as *Karlin's Conjecture* from 1959, Samuel Karlin [Kar59] suggested that the true rate may be significantly faster, perhaps on the order of $O\left(\frac{1}{\sqrt{k}}\right)$. This remained an open question for decades, but was seemingly put to rest in 2014 by Daskalakis and Pan [DP14] who were able to produce an instance of a game and a FP dynamic for which the convergence rate was indeed exponential in the number of actions, matching the bound of Robinson. Their lower bound construction follows along the same lines as the upper bound of Robinson, recursively generating harder instances as more actions are given to the players.

We address the issue of ties in two different ways. We first consider the convergence of a well-defined version FP with *lexicographic tie-breaking*, where the $\arg\min$ and $\arg\max$ functions break ties by selecting the winner with the smallest index. We show that this version of FP has a rate of $O\left(\frac{1}{\sqrt{k}}\right)$ for a class of payoff matrices which includes the matrix used in the lower bound of Daskalakis and Pan. We further provide a lower bound of $\Omega\left(\frac{1}{\sqrt{k}}\right)$ for one such matrix in the class, yet we leave open the question of whether the $O\left(\frac{1}{\sqrt{k}}\right)$ upper bound is true for any arbitrary payoff matrix. Second, as the issue of ties is in part due to the fact that the decision sets Δ_n and Δ_m are polytopes with flat boundaries, we consider a scenario where the decision sets are instead slightly round bodies. In this setting, we are able to establish that the convergence rate is guaranteed to be $O\left(\frac{1}{\sqrt{k}}\right)$, and in some cases is $\tilde{O}\left(\frac{1}{k}\right)$. We also show that a modification of FP known as *Optimistic* FP,

converges at a rate of $O\left(\frac{1}{k}\right)$.

3.2 Related work

We now give a brief overview of prior work on fictitious play, game dynamics, and what results exists for convergence to equilibrium.

The original formulation of FP was by Brown [Bro49; Bro51], where he mentions both discrete and continuous time dynamics. Robinson [Rob51] proved the slow convergence rate of $O(k^{-\frac{1}{m+n-2}})$ for FP in discrete time, under arbitrary tie-breaking. Karlin [Kar59] later conjectured that the convergence rate was $O(k^{-\frac{1}{2}})$. Danskin [Dan81] simplified and extended Robinson's result to when the min and max have errors. Daskalakis and Pan [DP14] constructed a counter-example for Karlin's strong conjecture using carefully designed adversarial tie-breaking rules, showing that FP for a zero-sum game on the $n \times n$ identity matrix has a worst-case convergence rate of $\Omega(k^{-\frac{1}{n}})$.

FP has also been studied for more general games. Miyasawa [Miy61] showed convergence of FP for non-zero-sum 2×2 two-player games. Shapley [Sha64] showed FP does not converge in a certain 3×3 non-zero-sum-game. Monderer and Sela [MS96] later constructed 2×2 non-zero-sum game for which FP does not converge. Brandt et al. [BFH10] show that it will take exponentially long for the iterates of FP (as opposed to the scaled iterates) to reach an equilibrium for several classes of games.

Much work has also been done on continuous-time FP. Harris [Har98] proved that a continuous-time FP dynamic with t as the time parameter has a convergence rate of $O(t^{-1})$ for any two-person zero-sum game. Ostrovski and van Strien [OS11] studied the piecewise-linear Hamiltonian flows generated by fictitious play algorithms and the combinatorics of the trajectories for 3×3 games. Ostrovski and van Strien [OS14] studied the convergence and trajectories of FP in continuous time for 3×3 games. Swenson and Kar [SK17] showed exponential convergence rate for continuous-time FP for "regular" games.

Finally, the FP dynamic is closely related to dynamics where both players use no-regret

algorithms to choose their actions, and a lot of work has been done trying to understand these dynamics as well. In particular, under the FP dynamic, both players update their actions using the Follow-The-Leader algorithm. Hofbauer and Sandholm [HS02] studied stochastic fictitious play and showed global convergence of an algorithm now known as Follow-The-Perturbed-Leader. Swenson et al. [SKXL17] studied robustness of fictitious play under perturbations. Bailey and Piliouras [BP19b] showed that network Follow-The-Regularized-Leader (FTRL) is Hamiltonian flow. Bailey and Piliouras [BP19a] showed $O(k^{-\frac{1}{2}})$ regret for fixed step-size FTRL with a quadratic regularizer for 2×2 zero-sum games. Finally, Bailey et al. [BGP19] showed finite regret for alternating FTRL with a quadratic regularizer.

3.3 Preliminaries

We now provide some precise definitions for games, dynamics, and convergence. Along the way, we lay out our main results and describe them in the context of other work. The techniques are described in greater detail in Section 3.5 and beyond.

Notation For the remainder of the chapter, we assume we are working with square payoff matrices $A \in \mathbb{R}^{n \times n}$, and the decision set for the row and column players are $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^n$, respectively. For a matrix A, let $A_{\min}$ be the minimum diagonal entry of A and let $A_{\max}$ be the maximum diagonal entry of A. The $\tilde{O}$ and $\tilde{\Theta}$ notation hides factors that are logarithmic in the number of iterations k.

Note that in the matrix game setting, the duality gap can be written as follows:

$$\psi(x, y) = \max_{\tilde{y} \in \mathcal{Y}} x^\top A\tilde{y} - \min_{\tilde{x} \in \mathcal{X}} \tilde{x}^\top Ay. \tag{3.2}$$

While ψ is defined on all of $\mathbb{R}^n \times \mathbb{R}^n$, it holds that $\psi(x, y) \geq 0$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We will also consider a slight generalization of matrix games where $\mathcal{X}$ and $\mathcal{Y}$ need not be simplices.

3.3.1 The Fictitious Play dynamic

The *fictitious play* (FP) dynamic involves a sequence of state variables $x_k, y_k \in \mathbb{R}^n$ which evolve for a series of iterations (or rounds) $k = 1, 2, \ldots$. The initial iterates x_0 and y_0 are classically initialized at $\mathbf{0}$, but we will also allow initializations in $\mathcal{X} \times \mathcal{Y}$. We define the recursive update

$$a_k := \arg\min_{x \in \mathcal{X}} x^\top A y_k \qquad b_k := \arg\max_{y \in \mathcal{Y}} x_k^\top A y \tag{3.3}$$

$$x_{k+1} := x_k + a_k \qquad y_{k+1} := y_k + b_k \tag{3.4}$$

Concretely, at each iteration $k \geq 1$ each player plays the action that is the best response to the long-term distribution of their opponent's actions. It is convenient to consider the *scaled history* of each player's state, as this is appropriately normalized:

$$\hat{x}_k := \frac{x_k}{k} \quad \text{and} \quad \hat{y}_k := \frac{y_k}{k}$$

Note that $a_k \in \mathcal{X}$, $b_k \in \mathcal{Y}$, so $\hat{x}_k \in \mathcal{X}$ and $\hat{y}_k \in \mathcal{Y}$ for $k \geq 1$. Note that we can evaluate ψ on either $(\hat{x}_k, \hat{y}_k)$ or (x_k, y_k), and while it makes less sense to refer to it as the "duality gap" in the former case we will use the terminology in both cases.

For the remainder of the chapter, we will focus on evaluating the normalized duality gap $\psi(\hat{x}_k, \hat{y}_k)$ as $k \to \infty$, and to determine at what rate $\psi(\hat{x}_k, \hat{y}_k)$ converges to 0. For convenience, our proofs will often do this by showing the equivalent claim that $\psi(x_k, y_k) = o(k)$.

Following the discussion of tie-breaking earlier, we need to address the case when the $\arg\min$ or $\arg\max$ in (3.3) is non-unique.

Assumption 3.3.1. *Ties in the* $\arg\min$ *and* $\arg\max$ *in the FP dynamic are broken according to lexicographic order. That is, the* $\arg\min$ *and* $\arg\max$ *in the FP dynamic are always unique.*

3.3.2 Fictitious play as skew-gradient flow

We will now characterize the fictitious play dynamic as a *discrete-time skew-gradient flow*.

Recall the *support function* $\phi_{\mathcal{X}} \colon \mathbb{R}^n \to \mathbb{R}$ of a set $\mathcal{X} \subseteq \mathbb{R}^n$ is given by

$$\phi_{\mathcal{X}}(\theta) := \max_{x \in \mathcal{X}} \theta^\top x.$$

We can express the duality gap in terms of the support functions of the decision sets $\mathcal{X}, \mathcal{Y}$:

$$\psi(x, y) = \phi_{\mathcal{Y}}(A^\top x) + \phi_{\mathcal{X}}(-Ay).$$

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{2n}$, and let $S \in \mathbb{R}^{2n \times 2n}$ denote the skew-symmetric matrix $S = \begin{pmatrix} 0 & -A \\ A^\top & 0 \end{pmatrix}$. Then for $z = (x, y)$, we can write the duality gap as the support function of the skewed input:

$$\psi(x, y) = \psi(z) = \phi_{\mathcal{Z}}(Sz)$$

since indeed $\phi_{\mathcal{Z}}(Sz) = \phi_{\mathcal{X} \times \mathcal{Y}}(-Ay, A^\top x) = \phi_{\mathcal{X}}(-Ay) + \phi_{\mathcal{Y}}(A^\top x)$.

Recall the gradient of the support function is the following maximizer:

$$\nabla \phi_{\mathcal{X}}(\theta) = \arg\max_{x \in \mathcal{X}} \theta^\top x.$$

In general when $\phi_{\mathcal{X}}$ is not differentiable, the set of subgradients corresponds to the $\arg\max$ above. We can write fictitious play as the $\epsilon = 1$ case of

$$x_{k+1} = x_k + \epsilon \nabla \phi_{\mathcal{X}}(-Ay_k) \tag{3.5}$$

$$y_{k+1} = y_k + \epsilon \nabla \phi_{\mathcal{Y}}(A^\top x_k). \tag{3.6}$$

As $\epsilon \to 0$, the above converges to the continuous-time dynamic

$$\dot{X}_t = \nabla \phi_{\mathcal{X}}(-AY_t)$$
$$\dot{Y}_t = \nabla \phi_{\mathcal{Y}}(A^\top X_t)$$

where $\dot{X}_t = \frac{d}{dt}X_t$ and $\dot{Y}_t = \frac{d}{dt}Y_t$. Let us write $Z_t = (X_t, Y_t)$, so $SZ_t = (-AY_t, A^\top X_t)$. Then

$$\dot{Z}_t = \nabla \phi_{\mathcal{Z}}(SZ_t). \tag{3.7}$$

Note the gradient of $\psi(z) = \phi_{\mathcal{Z}}(Sz)$ is $\nabla\psi(z) = S^\top \nabla \phi_{\mathcal{Z}}(Sz)$. If S is invertible, then we can write the above as a *skew-gradient flow*: $\dot{Z}_t = (S^\top)^{-1}\nabla\psi(Z_t)$, which preserves the duality gap since $(S^\top)^{-1}$ is skew-symmetric. However, even when S is not invertible, the flow (3.7) always preserves the duality gap:

$$\frac{d}{dt}\psi(Z_t) = \nabla\psi(Z_t)^\top \dot{Z}_t = \nabla\phi_{\mathcal{Z}}(SZ_t)^\top S \nabla \phi_{\mathcal{Z}}(SZ_t) = 0.$$

Therefore, for the scaled history $\hat{Z}_t = \frac{Z_t}{t}$, the duality gap decreases at an $O(t^{-1})$ rate:

$$\psi(\hat{Z}_t) = \frac{\psi(Z_t)}{t} = \frac{\psi(Z_1)}{t} = \Theta(t^{-1}).$$

In the above, the first equality is because support function is homogeneous.

In discrete time, the forward method for discretizing the dynamic (3.7) is

$$z_{k+1} = z_k + \epsilon \nabla \phi_{\mathcal{Z}}(Sz_k), \tag{3.8}$$

which is (3.5) for $z_k = (x_k, y_k)$. Since ψ is a convex function, the forward method increases

ψ. Indeed, by Jensen's inequality and since S is skew-symmetric,

$$\psi(z_{k+1}) - \psi(z_k) \geq \nabla\psi(z_k)^\top (z_{k+1} - z_k) = \epsilon \nabla\phi_{\mathcal{Z}}(Sz_k)^\top S \nabla\phi_{\mathcal{Z}}(Sz_k) = 0.$$

This is similar to [BP19b] when the regularizer is the indicator function of the domain.

3.4 Main Results

In light of the preliminary material above, we can now give a birds-eye view of the work in this chapter. The formal results will be laid out in full detail in the following sections.

3.4.1 Fast convergence for diagonal matrices

Our first core result is to show that Karlin's conjecture is indeed true for the class of diagonal matrices, as long as the natural Assumption 3.3.1 holds true. This class is an important special case, as it includes the identity matrix used by the lower bound of Daskalakis and Pan [DP14]. This shows that the slow-converging construction is obliterated by Assumption 3.3.1.

Theorem (informal)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a diagonal matrix with a strictly positive*[1] *diagonal. Then the FP dynamic* (3.4)*, under Assumption 3.3.1, guarantees* $\psi(\frac{1}{k}x_k, \frac{1}{k}y_k) = O\left(\sqrt{\frac{A_{\max}^3}{A_{\min}}} k^{-1/2}\right)$.

Our result greatly expands the class of games for which the FP dynamic has been shown to converge quickly to equilibrium. Previously, the FP dynamic was only known to achieve a $O(k^{-1/2})$ convergence rate for 2×2 matrices. Also of note is that our convergence rate is independent of the dimension n. The main proof of Theorem 3.6.15 is in Section 3.6, with minor proofs being deferred to Section 3.7.

Our proof of this result relies on three main properties. We first note that in the diagonal case under Assumption 3.3.1, the dynamic alternates between two distinct phases, which we

[1]The requirement that the diagonal be strictly positive is without loss of generality, as we discuss in Section 3.6.

call *sync* and *split* phases. We use the term *sync-split pair* to denote a pair of consecutive phases consisting of a sync phase followed by a split phase. Second, we show that the duality gap can only increase by a constant amount over the course of each sync-split pair. Finally, we define a potential function that allows us to show that the duration of each sync-split pair is proportional to the duality gap at the start of the sync-split pair.

From these properties, we can derive the rate. To get some intuition, we can consider the case when round 1 is the first round of a sync-split pair and the duality gap always increases by a constant c during each sync-split pair. That is, the duality gap at the start of the τ^{th} sync-split pair is $(\tau - 1)c$. Then by the end of the t^{th} sync-split pair, the total duality gap will be tc. Meanwhile, it will take $\sum_{j=1}^{t}(j-1)c = \Theta(t^2 c)$ rounds to complete these t sync-split pairs. So we can see that the duality gap grows as the square root of the number of rounds.

We also prove the following lower bound:

Theorem (informal)**.** *Let A be the $n \times n$ identity matrix. Then the FP dynamic* (3.4)*, under Assumption 3.3.1, satisfies $\psi(\frac{1}{k}x_k, \frac{1}{k}y_k) = \Omega\left(\sqrt{\frac{1}{n}}k^{-1/2}\right)$.*

While analogous lower bounds existed for the 2×2 case, to our knowledge, no lower bound has been proven for the FP dynamic under Assumption 3.3.1 for settings in more than two dimensions. This lower bound shows that the dependence on k in Theorem 3.6.15 is tight. The dependence on n is likely suboptimal, and we leave improving that dependence to future work. We prove Theorem 3.6.23 in Section 3.6.3. The proof is structured similarly to the proof of the upper bound, as the characterization of the FP dynamic in that proof is actually quite tight.

3.4.2 Faster convergence in the smooth case

Our second set of results focuses on the generalization of the FP dynamic that we introduced in Section 3.3.2. We observed that the FP dynamic can be viewed through the lens of a skew-gradient flow, where the pair of $z_k = (x_k, y_k)$ is updated as $z_{k+1} = z_k + \nabla\phi_{\mathcal{Z}}(Sz_k)$, where $\phi_{\mathcal{Z}}$ is the support function on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and S is an appropriately chosen skew-symmetric

matrix. This perspective is helpful as it allows us to reason about the convergence of the dynamic through properties of $\phi_{\mathcal{Z}}$.

What we show in Section 3.5 is that when $\phi_{\mathcal{Z}}$ is *smooth*, the FP dynamic is well-behaved and easy to control. Of course, this does not apply to the case when $\mathcal{Z} = \Delta_n \times \Delta_n$, which is the standard FP setting, since the sharp edges of the probability simplex lead to non-smoothness of the support function. But if we consider a "slightly rounder" body $\mathcal{Z}$—a relaxed version of $\Delta_n \times \Delta_n$, for example—then we can obtain convergence rates in line with Karlin's conjecture.

Theorem (informal)**.** *Let $\mathcal{Z}$ be such that $\phi_{\mathcal{Z}}$ is twice differentiable everywhere but at the origin. Consider the dynamic on $z_k = (x_k, y_k)$ described above. Then*

$$\psi\left(\frac{1}{k} z_k\right) = \begin{cases} O\left(\frac{\log k}{k}\right) & \text{when } 0 \notin S\mathcal{Z} \\ O\left(\frac{1}{\sqrt{k}}\right) & \text{when } 0 \in (S\mathcal{Z})^\circ \end{cases}$$

We also show that our bound in the $0 \in (S\mathcal{Z})^\circ$ case is tight when ϕ is orthogonally strongly convex. We note that requiring $\phi_{\mathcal{Z}}$ to be a smooth function is another way to avoid the tie-breaking issue. A tie in the $\arg\min$ or $\arg\max$ occurs when $\phi_{\mathcal{Z}}$ is non-differentiable and hence the subgradient set is non-unique.

We also show similar convergence rates for an *optimistic* version of FP, defined as $z_{k+1} = z_k + \nabla\phi(Sz_{k+\frac{1}{2}})$, where $z_{k+\frac{1}{2}} = z_k + \nabla\phi(Sz_k)$.

Theorem (informal)**.** *Let $\mathcal{Z}$ be such that $\phi_{\mathcal{Z}}$ is twice differentiable everywhere but at the origin, and assume $0 \notin S\mathcal{Z}$. Consider the optimistic FP dynamic on $z_k = (x_k, y_k)$ described above. Then $\phi(\frac{1}{k} Sz_k) = O(k^{-1})$.*

3.5 Analysis of fictitious play in the smooth case

In this section, we outline our results for fictitious play over smooth constraint sets. Let $\mathcal{Z}$ be a nonempty, compact, convex set in $\mathbb{R}^m$. We consider the optimization problem:

$$\min_{z \in \mathcal{Z}} \phi(Sz)$$

where $S = -S^\top \in \mathbb{R}^{m \times m}$ is a skew-symmetric matrix and $\phi \colon \mathbb{R}^n \to \mathbb{R}$ is the support function of $\mathcal{Z}$. Analogous to von Neumann's minimax theorem, the minimum value is always 0; note that this holds even without any smoothness assumption on ϕ.

Theorem 3.5.1. *Suppose $\mathcal{Z} \subset \mathbb{R}^m$ is nonempty, compact, and convex. Suppose $S^\top = -S$. Then*

$$\min_{z \in \mathcal{Z}} \phi(Sz) = 0. \tag{3.9}$$

Note that $\phi(0) = 0$, and ϕ is positively homogeneous: $\phi(t\theta) = t\phi(\theta)$ for all $t \geq 0$, $\theta \in \mathbb{R}^m$. So at $\theta = 0$ the function ϕ has a cone structure and it is not differentiable. But away from 0, ϕ can be differentiable. In this section we make the following assumption.

Assumption 3.5.2. *The support function $\phi(\theta)$ is twice-differentiable at all $\theta \neq 0$.*

The positive homogeneity of ϕ implies the gradient is scale-invariant: $\nabla\phi(t\theta) = \nabla\phi(\theta)$, and the Hessian is inversely proportional to the input: $\nabla^2\phi(t\theta) = \frac{1}{t}\nabla^2\phi(\theta)$ for all $t > 0$, $\theta \in \mathbb{R}^n \setminus \{0\}$.

We note this smoothness assumption does not hold for the original fictitious play algorithm (in which $\mathcal{Z} = \Delta_n \times \Delta_n$). However, in general we can arbitrarily approximate any convex set with a smooth set (i.e., one with a smooth support function). Here we show in the smooth case, the behavior of fictitious play is different from the behavior on the simplex.

We study the **forward method** (*fictitious play*), which starts from an arbitrary $z_1 \in \mathcal{Z}$

and for $k \geq 1$ maintains

$$z_{k+1} = z_k + \nabla\phi(Sz_k).$$

Note that $z_k \in k\mathcal{Z}$. We define the scaled history $\hat{z}_k = \frac{z_k}{k} \in \mathcal{Z}$.

As noted in Section 3.3.2, the forward method in fact increases the support function; indeed, since ϕ is convex, by Jensen's inequality

$$\phi(Sz_{k+1}) \geq \phi(Sz_k) + \nabla\phi(Sz_k)^\top S\nabla\phi(Sz_k) = \phi(Sz_k).$$

We will bound how much $\phi(Sz_k)$ grows along the forward method. We present the analysis in two cases: In Section 3.5.1 we consider $0 \notin S\mathcal{Z}$ (as in the original fictitious play) and show $\phi(Sz_k) = O(\log k)$. In Section 3.5.2 we consider $0 \in S\mathcal{Z}$ and show $\phi(Sz_k) = O(\sqrt{k})$; furthermore, we show a matching lower bound under a notion of orthogonal strong convexity. In Section 3.5.3 we propose an optimistic variant of the forward method and show $\phi(Sz_k) = O(1)$ in the first case.

3.5.1 Case 1: $0 \notin S\mathcal{Z}$

Suppose $0 \notin S\mathcal{Z}$ (so the minimum is achieved on a ray). Assume ϕ is twice-differentiable. Let

$$d = \min_{z\in\mathcal{Z}} \|Sz\| > 0, \quad D = \max_{z\in\mathcal{Z}} \|Sz\| < \infty, \quad L = \sup_{\|\theta\|=1} \|\nabla^2\phi(\theta)\| < \infty. \tag{3.10}$$

Note along the forward method we have $z_k \in k\mathcal{Z}$, so $\|z_k\| = \Theta(k)$. This implies that the support function only increases by $O(1/k)$ in each step of the forward method.

Lemma 3.5.3. *Assume $0 \notin S\mathcal{Z}$ and Assumption 3.5.2. For each $k \geq 1$, the forward method*

satisfies:

$$\phi(Sz_{k+1}) \leq \phi(Sz_k) + \frac{LD^2}{2dk}.$$

By iterating, we have the following bound on the support function along the forward method.

Theorem 3.5.4. *Assume $0 \notin S\mathcal{Z}$ and Assumption 3.5.2. For each $k \geq 2$, the forward method satisfies:*

$$\phi(Sz_k) \leq \phi(Sz_1) + \frac{LD^2}{2d}\left(1 + \log(k-1)\right) = O(\log k).$$

Furthermore, recall along the forward method the support function increases: $\phi(Sz_k) \geq \phi(Sz_1)$. Therefore, we have $\Omega(k^{-1}) \leq \phi(S\hat{z}_k) \leq O(k^{-1}\log k)$ for the scaled history $\hat{z}_k = \frac{z_k}{k}$.

Note that this is different from the $\Omega(k^{-\frac{1}{2}})$ behavior for the original fictitious play on the simplex.

3.5.2 Case 2: $0 \in (S\mathcal{Z})^\circ$

Suppose $0 \in (S\mathcal{Z})^\circ$, which means $0 \in S\mathcal{Z}$ and $0 \notin \partial(S\mathcal{Z}) = S\partial\mathcal{Z}$ (so the minimizer is $z^* = 0$). We have $\phi(\theta) > 0$ for all $\theta \in \mathbb{R}^n \setminus \{0\}$. Assume ϕ is twice-differentiable. Let

$$m = \min_{\|\theta\|=1} \phi(\theta) > 0, \quad R = \max_{z \in \partial\mathcal{Z}} \|z\| < \infty, \quad L = \sup_{\|\theta\|=1} \|\nabla^2 \phi(\theta)\| < \infty.$$

Note that for all $\theta \in \mathbb{R}^n$ we have

$$m\|\theta\| \leq \phi(\theta) = \theta^\top \nabla\phi(\theta) \leq R\|\theta\|.$$

In this case we can show the forward method increases the support function by an

amount inversely proportional to its current value.

Lemma 3.5.5. *Assume $0 \in (S\mathcal{Z})^\circ$ and Assumption 3.5.2. For each $k \geq 1$, the forward method satisfies:*

$$\phi(Sz_{k+1}) \leq \phi(Sz_k) + \frac{L'}{\phi(Sz_k)}$$

where $L' = LR^4\|A\|^2/m$.

By iterating, we get the following bound on the support function along the forward method.

Theorem 3.5.6. *Assume $0 \in (S\mathcal{Z})^\circ$ and Assumption 3.5.2. For each $k \geq 1$, the forward method satisfies:*

$$\phi(Sz_k) \leq \sqrt{\phi(Sz_1)^2 + (k-1)L''} = O(\sqrt{k})$$

where $L'' = \frac{L'^2}{\phi(Sz_1)^2} + 2L'$ and $L' = LR^4\|A\|^2/m$. That is, $\phi(S\hat{z}_k) \leq O(k^{-\frac{1}{2}})$ for the scaled history $\hat{z}_k = \frac{z_k}{k}$.

Under orthogonal strong convexity, we can show this rate is tight.

Lower bound under orthogonal strong convexity

Since a support function ϕ is positively homogeneous ($\phi(t\theta) = t\phi(\theta)$), the Hessian is singular along its input: $\nabla^2\phi(\theta)\theta = 0$. But orthogonal to the input, ϕ can have some curvature.

Definition 3.5.7. *We say ϕ is α-***orthogonally strongly convex** *if ϕ is twice-differentiable and α-strongly convex along directions orthogonal to the input:*

$$\alpha := \inf_{\|\theta\|=1} \inf_{\substack{\|v\|=1 \\ v^\top\theta=0}} v^\top \nabla^2\phi(\theta)v > 0. \tag{3.11}$$

In this section we make the following assumption.

Assumption 3.5.8. *The support function ϕ is α-orthogonally strongly convex for some $\alpha > 0$.*

Under orthogonal strong convexity, we can prove a matching lower bound. Let

$$m = \min_{\|\theta\|=1} \phi(\theta) > 0, \quad R_0 = \min_{z \in \partial \mathcal{Z}} \|Sz\| > 0, \quad R = \max_{z \in \partial \mathcal{Z}} \|z\| < \infty.$$

Lemma 3.5.9. *Assume $0 \in (S\mathcal{Z})^\circ$ and Assumption 3.5.8. For each $k \geq 1$, the forward method satisfies:*

$$\phi(Sz_{k+1}) \geq \phi(Sz_k) + \frac{C}{\phi(Sz_k)}$$

where $C = \frac{\alpha m^3 R_0}{16 R^2} \min\{\|Sz_1\|, R_0\}$.

By iterating, we have the following lower bound.

Theorem 3.5.10. *Assume $0 \in (S\mathcal{Z})^\circ$ and Assumption 3.5.8. For each $k \geq 1$, the forward method satisfies:*

$$\phi(Sz_k) \geq \sqrt{\phi(Sz_1)^2 + 2C(k-1)} = \Omega(\sqrt{k})$$

where $C = \frac{\alpha m^3 R_0}{16 R^2} \min\{\|Sz_1\|, R_0\}$. Therefore $\phi(S\hat{z}_k) = \Theta(k^{-\frac{1}{2}})$ for the scaled history $\hat{z}_k = \frac{z_k}{k}$.

An example where ϕ is orthogonally strongly convex is when $\mathcal{Z}$ is an ellipsoid (or any ℓ_p-ball, $p > 1$). In this case we indeed get a $\Theta(k^{-\frac{1}{2}})$ rate.

Example 3.5.11. *Let $\mathcal{Z} = \{z \in \mathbb{R}^n : z^\top B^{-1} z \leq 1\}$ where $B = B^\top \succ 0$. Then $\phi(\theta) = \sqrt{\theta^\top B \theta} = \|\theta\|_B$ and $\nabla \phi(\theta) = B\theta / \|\theta\|_B$. The forward method becomes*

$$z_{k+1} = z_k + \frac{BSz_k}{\|Sz_k\|_B}.$$

Then $\|Sz_{k+1}\|_B^2 = \|Sz_k\|_B^2 + \frac{\|SBSz_k\|_B^2}{\|Sz_k\|_B^2}$, *so* $\phi(Sz_k) = \|Sz_k\|_B = \Theta(\sqrt{k})$ *and* $\phi(S\hat{z}_k) = \Theta(k^{-1/2})$.

3.5.3 Faster convergence in via optimism

We study the following **optimistic forward method**:[2]

$$z_{k+1} = z_k + \nabla\phi(Sz_{k+\frac{1}{2}}) \quad \text{where } z_{k+\frac{1}{2}} = z_k + \nabla\phi(Sz_k).$$

In this section we assume ϕ is twice-differentiable (Assumption 3.5.2). We also assume $0 \notin S\mathcal{Z}$ (which is the case in the original fictitious play). We recall the definitions of d, D, L from (3.10).

Lemma 3.5.12. *Assume* $0 \notin S\mathcal{Z}$ *and Assumption 3.5.2. For* $k \geq 1$*, the optimistic forward method satisfies:*

$$\phi(Sz_{k+1}) \leq \phi(Sz_k) + \frac{L^2D^2\|S\|}{d^2k(k+1)}.$$

Since $\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$, we can write the above as $\phi(Sz_{k+1}) + \frac{L^2D^2\|S\|}{d^2(k+1)} \leq \phi(Sz_k) + \frac{L^2D^2\|S\|}{d^2k}$. That is,

$$E_k = \phi(Sz_k) + \frac{L^2D^2\|S\|}{d^2k}$$

is a *Lyapunov function*, which means it decreases along the optimistic forward method. This implies the following bound. In particular, as $k \to \infty$, we see the support function is finite.

Theorem 3.5.13. *Assume* $0 \notin S\mathcal{Z}$ *and Assumption 3.5.2. For* $k \geq 1$*, the optimistic forward*

[2]We note the above is in extra-gradient form. There is another optimistic form: $z_{k+1} = z_k + 2\nabla\phi(Sz_k) - \nabla\phi(Sz_{k-1})$. We study the extra-gradient form above for simplicity; similar results can also be established for the optimistic form.

method satisfies:

$$\phi(Sz_k) \leq \phi(Sz_1) + \frac{L^2 D^2 \|S\|}{d^2}.$$

Therefore, $\phi(S\hat{z}_k) = O(k^{-1})$ *for the scaled history* $\hat{z}_k = \frac{z_k}{k}$.

3.6 Fast Convergence of Fictitious Play for Diagonal Payoff Matrices

In this section, we deal with the case when $\mathcal{X} = \mathcal{Y} = \Delta_n$. We define some new notation that will aid in this analysis. Let $p_k(i) = e_i^\top A y_k$ and $q_k(j) = x_k^\top A e_j$, so the fictitious play dynamic can be written as:

$$x_{k+1} = x_k + e_{\arg\min_i \, p_k(i)}$$

$$y_{k+1} = y_k + e_{\arg\max_j \, q_k(j)}$$

Let $p_k^* = \min_{i \in [n]} p_k(i)$ and $q_k^* = \max_{j \in [n]} q_k(j)$. Then $\psi(x_k, y_k) = q_k^* - p_k^*$.

Now we define the *gap vectors* and *total gap vector*:

Definition 3.6.1. *The* gap vectors *for a given round* k *are vectors* $u_k \in \mathbb{R}^n$ *and* $v_k \in \mathbb{R}^n$ *such that* $u_k(i) = p_k(i) - p_k^*$ *and* $v_k(j) = q_k^* - q_k(j)$.

Definition 3.6.2. *The* total gap vector *for a given round* k *is a vector* $w_k \in \mathbb{R}^n$ *such that* $w_k(i) = A_{ii}^{-1}(u_k(i) + v_k(i))$.

We see that the i^{th} entry of u and v tracks how far the i^{th} action is from being the optimal action for the x and y players respectively. Note that at least one entry of u_k and one entry of v_k is 0, corresponding to the best action for the x and y players respectively. Moreover, u and v are always nonnegative, which implies that w is always nonnegative.

It will be useful to define the following states of the dynamic:

Definition 3.6.3 (Sync and split rounds)**.** *Suppose in some round* k*, the row and column players both play action* i*. Then round* k *is called a* sync round*, and in particular it is a*

sync(i,i) *round. We will also say that round* k*'s* type *is "sync*(i,i)*". Suppose in some round* k'*, the row player plays action* j *and the column player plays action* i *such that* $i \neq j$*. Then round* k' *is called a* split round*, and in particular it is a* split(j,i) *round. We will also say that round* k*'s* type *is "split*(j,i)*". So a round's type will either be "sync*(i,i)*" for some* $i \in [n]$ *or "split*(j,i)*" for some* $i,j \in [n]$*.*

Definition 3.6.4. *Let a* phase *denote a maximal consecutive block of rounds of a particular type. In particular, suppose:*

1. *rounds* k *to* $k+s$ *are all of some type, call it type* α*;*
2. *if* $k \geq 2$*, round* $k-1$ *is not type* α*;*
3. *round* $k+s+1$ *is not type* α*.*

Then rounds k *to* $k+s$ *constitute a* phase*. Moreover, if rounds* k *to* $k+s$ *are all sync*(i,i) *rounds, then they constitute a* sync phase *and in particular a* sync(i,i) phase*. Likewise, if rounds* k *to* $k+s$ *are all split*(j,i) *rounds, then they constitute a* split phase *and in particular a* split(j,i) phase*.*

Round #	$k-1$	k	$k+1$	$k+2$	$k+3$
Row player action	i	i	i	i	j
Column player action	ℓ	i	i	i	i
Round type	split(i,ℓ)	sync(i,i)	sync(i,i)	sync(i,i)	split(j,i)
		$\underbrace{\qquad\qquad\qquad\qquad}$			
		sync(i,i) phase			

Figure 3.1: Illustration of Definitions 3.6.3 and 3.6.4. Rounds k to $k+2$ form a sync(i,i) phase.

Definition 3.6.5. *Suppose rounds* k *to* $k+k'-1$ *form a sync*(i,i) *phase and rounds* $k+k'$ *to* $k+s-1$ *form a complete split*(j,i) *phase Then we call rounds* k *to* $k+s-1$ *a* sync-split pair *and in particular a* sync-split$(i \to j)$ pair*.*

If we look at the trajectory of the FP dynamic, namely (x_k, y_k) for $k \in \{0, 1, 2, ...\}$, we will encounter a countable number of sync and split phases. Suppose that the sync phases start in rounds $\{s_1, s_2, ...\}$ where $s_{t_1} < s_{t_2}$ for $t_1 < t_2$. Then we say the τ^{th} sync phase of the trajectory is the sync phase starting in round s_τ. We will use the indices i, j, ℓ to denote generic actions in $\{1, ..., n\}$ unless otherwise specified. We will generally use k to specify a generic round of the FP dynamic where $k \geq 1$.

In the rest of this section, we will assume that Assumption 3.3.1 holds, which motivates the following definition:

Definition 3.6.6. *Let the* tiebreak order *of the fictitious play dynamic be a pair of permutations $(\sigma_x, \sigma_y) \in S_n \times S_n$ such that when breaking ties between a set of indices $\mathcal{I}$, the x player chooses the index $r_x = \arg\min_{\ell \in \mathcal{I}} \sigma_x(\ell)$ and the y player chooses the index $r_y = \arg\min_{\ell \in \mathcal{I}} \sigma_y(\ell)$.*

In the rest of this section, we will also assume that A is a diagonal matrix with positive diagonal, so we omit this from the lemma statements for notational clarity. Note that if all diagonal entries of A are negative, we can simply reverse the roles of x and y and play on the matrix $-A$. Moreover, if A has positive and non-positive diagonal entries, then the Nash Equilibria will not have full support because any equilibrium strategy for the x player will not use the rows with positive diagonal entries and any equilibrium strategy for the y player will not use the columns with non-positive diagonal entries.

3.6.1 Important properties of the FP dynamic

In this section, we characterize some key properties of the FP dynamic. We start by showing that the dynamic alternates between sync and split phases:

Lemma 3.6.7. *Suppose round k is a sync(i, i) phase. Then this phase will end in some round $k + s$ for finite s, and round $k + s + 1$ will be a split(j, i) round for some $j \neq i$. Likewise, if round k is a split(j, i) phase, then this phase will end in some round $k + s$*

for finite s, and round $k+s+1$ will be a sync(j,j) round. Thus, the dynamic alternates between sync and split phases, and the dynamic will proceed through an unbounded number of sync and split phases. Moreover, for any $t \geq 1$, if the t^{th} sync phase of the FP dynamic is a sync(i_t, i_t) phase and the $(t+1)^{th}$ sync phase is a sync(i_{t+1}, i_{t+1}) phases, then $i_t \neq i_{t+1}$.

Next, we characterize how the duality gap and w change over the course of sync and split phases with Lemmas 3.6.8 and 3.6.9. From these lemmas, we can see that the duality gap only increases by at most $A_{\max}$ during each sync and split phase and that each entry of w increases by an amount proportional to the increases in the duality gap.

Lemma 3.6.8. *Suppose rounds k to $k+s$ are sync(i,i) rounds for $s \geq 0$ and round $k+s+1$ is a split(j,i) round. Let $\epsilon = A_{ii} - u_{k+s-1}(j)$. Then*

1. $0 \leq \epsilon \leq A_{ii}$
2. $w_{k+s}(\ell) = w_{k-1}(\ell) + A_{\ell\ell}^{-1}\epsilon$ *for all* ℓ.
3. $\epsilon = \psi(x_{k+s}, y_{k+s}) - \psi(x_{k-1}, y_{k-1})$

Lemma 3.6.9. *Suppose rounds k to $k+s$ are split(j,i) rounds for $s \geq 0$ and round $k+s+1$ is a sync(j,j) round. Let $\epsilon = A_{jj} - v_{k+s-1}(j)$. Then,*

1. $0 \leq \epsilon \leq A_{jj}$
2. $w_{k+s}(\ell) = w_{k-1}(\ell) + A_{\ell\ell}^{-1}\epsilon$ *for* $\ell \notin \{i,j\}$
3. $\epsilon = \psi(x_{k+s}, y_{k+s}) - \psi(x_{k-1}, y_{k-1})$
4. $w_{k+s}(j) = 0$ *and* $w_{k+s}(i) = w_{k-1}(i) + w_{k-1}(j) + (A_{ii}^{-1} + A_{jj}^{-1})\epsilon$

Using Lemmas 3.6.8 and 3.6.9, we can prove the following lemma, which shows that over the course of a sync-split phase, w changes in a very precise way. At the start of the sync-split pair, w has $n-1$ non-zero values. At the end of the sync-split pair, each of these values has increased by an amount proportional to the increase in the duality gap, and the value in the j^{th} coordinate has moved to the i^{th} coordinate.

Lemma 3.6.10. *Suppose rounds k to $k+s-1$ form a sync-split$(i \to j)$ pair. Let $\epsilon = \psi(x_{k+s-1}, y_{k+s-1}) - \psi(x_{k-1}, y_{k-1})$. Then $w_{k+s-1}(\ell) \geq w_{k-1}(\ell) + \frac{\epsilon}{A_{\max}}$ for $\ell \notin \{i, j\}$ and $w_{k+s-1}(j) = w_{k-1}(i) = 0$ and $w_{k+s-1}(i) \geq w_{k-1}(j) + \frac{2\epsilon}{A_{\max}}$.*

From Lemma 3.6.10, we can inductively prove the following corollary, which describes how w evolves over the course of a series of consecutive sync-split pairs.

Corollary 3.6.11. *Let $t \geq 0$. Suppose we play $t+1$ consecutive sync-split pairs starting in rounds $s_1, ..., s_{t+1}$ respectively, and let round s_{t+1} be a sync(i_{t+1}, i_{t+1}) round. Let $\epsilon_j = \psi(x_{s_{j+1}-1}, y_{s_{j+1}-1}) - \psi(x_{s_j-1}, y_{s_j-1})$. Then $w_{s_{t+1}-1}(\ell) \geq \sum_{j=1}^{t} \frac{\epsilon_j}{A_{\max}}$ for $\ell \neq i_{t+1}$.*

Finally, the following lemma shows that the length of a sync-split pair is lower bounded by an entry of w.

Lemma 3.6.12. *Suppose rounds k to $k+s-1$ form a sync-split$(i \to j)$ pair. Then $s \geq \frac{A_{\min} w_{k-1}(j)}{A_{\max}}$.*

3.6.2 Proof of main theorem

Using the results in the previous section, we can prove our main lemma:

Lemma 3.6.13. *Let A be an $n \times n$ diagonal matrix with positive diagonal, and let Assumption 3.3.1 hold. Suppose we initialize the fictitious play dynamic at some $(x_0, y_0) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ such that round 1 is a sync round. Then for any $k \geq 1$ such that round $k+1$ is the first round of a sync phase, we have $\psi(x_k, y_k) - \psi(x_0, y_0) \leq 2\sqrt{\frac{A_{\max}^3}{A_{\min}}}\sqrt{k}$.*

Proof of Lemma 3.6.13. Let $\delta = \psi(x_k, y_k) - \psi(x_0, y_0)$. Let the sync phase starting in round $k+1$ be the $(t+1)^{th}$ sync phase of the FP trajectory. Note that $t \geq 1$ because the first sync phase starts in round 1. Let the τ^{th} sync phase be a sync(i_τ, i_τ) phase, and let s_τ be the round in which the τ^{th} sync phase starts. By assumption, the dynamic starts in a sync phase and round $k+1$ is the first round of a new sync phase, so t sync-split pairs will have

completed by the end of round k. Then we have:

$$k = \sum_{j=1}^{t}((s_{j+1}-1)-(s_j-1)) \geq \sum_{j=1}^{t} \frac{A_{\min} w_{s_j-1}(i_{j+1})}{A_{\max}}$$

where the inequality comes from Lemma 3.6.12.

Note that round s_j is a sync(i_j, i_j) round and $i_j \neq i_{j+1}$ by Lemma 3.6.7. Let $\epsilon_j = \psi(x_{s_{j+1}-1}, y_{s_{j+1}-1}) - \psi(x_{s_j-1}, y_{s_j-1})$. By Corollary 3.6.11, we have $w_{s_j}(i_{j+1}) \geq \sum_{\ell=1}^{j} \frac{\epsilon_l}{A_{\max}}$. Then we have:

$$k \geq \frac{A_{\min}}{A_{\max}^2} \sum_{j=1}^{t} \sum_{\ell=1}^{j} \epsilon_l = \frac{A_{\min}}{A_{\max}^2} \sum_{j=1}^{t} (t-j+1)\epsilon_j$$

Note that $\sum_{j=1}^{t} \epsilon_j = \delta$ and for all $j \in [t]$, $0 \leq \epsilon_j \leq A_{\max}$ and $0 \leq \delta \leq tA_{\max}$. So either $\delta < 2A_{\max}$ or by Lemma 3.6.14 we have $k \geq \frac{A_{\min}\delta^2}{4A_{\max}^3}$. Overall, this shows that

$$\delta \leq \max\left\{2\sqrt{\frac{A_{\max}^3}{A_{\min}}}\sqrt{k}, 2A_{\max}\right\}$$

But for $k \geq 1$, the first term in the max will dominate, so we have proved that $\psi(x_k, y_k) - \psi(x_0, y_0) \leq 2\sqrt{\frac{A_{\max}^3}{A_{\min}}}\sqrt{k}$. □

Lemma 3.6.14. *Let* $\mathcal{H} = \{h \in \mathbb{R}^t | \sum_{j=1}^{t} h_j = \delta$ *and* $\forall j, 0 \leq h_j \leq A_{\max}\}$ *where* $0 \leq \delta \leq tA_{\max}$. *Assume* $\delta \geq 2A_{\max}$. *Then*

$$\min_{h \in \mathcal{H}} \sum_{j=1}^{t} (t-j+1)h_j \geq \frac{\delta^2}{4A_{\max}}$$

Using Lemma 3.6.13, it is straightforward to show our main theorem:

Theorem 3.6.15. *Let* A *be an* $n \times n$ *diagonal matrix with positive diagonal, and let Assumption 3.3.1 hold. Suppose we initialize the fictitious play dynamic at some* $(x_0, y_0) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$. *Then for any* $k \geq 9A_{\max}$, *we have* $\psi(\frac{1}{k}x_k, \frac{1}{k}y_k) - \psi(\frac{1}{k}x_0, \frac{1}{k}y_0) \leq 4\sqrt{\frac{A_{\max}^3}{A_{\min}}}k^{-1/2}$.

Proof. For $j \geq 1$, let s_j be the round in which the j^{th} sync phase of the FP trajectory starts. First note that if $k < s_2 - 1$, then the dynamic will have completed at most two split phases and one sync phase, so by Lemmas 3.6.8 and 3.6.9, the duality gap will be at most $3A_{\max}$. Now consider the case when $k \geq s_2 - 1$. Let t be such that $s_t - 1 \leq k < s_{t+1} - 1$. Then we can use Lemma 3.6.13 with the FP dynamic initialized at (x_{s_1-1}, y_{s_1-1}) to get $\psi(x_{s_t-1}, y_{s_t-1}) - \psi(x_{s_1-1}, y_{s_1-1}) \leq 2\sqrt{\frac{A_{\max}^3}{A_{\min}}}\sqrt{s_t - 1 - s_1}$. There will be at most one split phase before round $s_1 - 1$ and at most one sync phase and one split phase and one sync phase in rounds s_t to k. Thus, by Lemmas 3.6.8 and 3.6.9, the duality gap can increase by at most $3A_{\max}$ over the course of those rounds. Then we have:

$$\psi(x_k, y_k) - \psi(x_0, y_0) \leq 2\sqrt{\frac{A_{\max}^3}{A_{\min}}}\sqrt{s_t - 1 - s_1} + 3A_{\max} \leq 2\sqrt{\frac{A_{\max}^3}{A_{\min}}}\sqrt{k} + 3A_{\max}$$

Note that this is obviously bigger than the upper bound on the duality gap in the $k < s_2 - 1$ case. We get the final bound by noting that the $\sqrt{k}$ term dominates for $k \geq 9A_{\max}$. □

3.6.3 Proof of lower bound

In this section, we prove our lower bound. We start with the following lemmas which show that the duality gap increases by either $\epsilon \in \{0, 1\}$ in the last round of a sync or split phase. Moreover, the value of ϵ depends on the tiebreak order.

Lemma 3.6.16. *Let Assumption 3.3.1 hold and let (σ_x, σ_y) be the tiebreak order for the FP dynamic. Let $A = I_n$. Let round k be a sync(i, i) round, and let round $k + 1$ be a split(j, i) round. Let $\epsilon = \psi(x_k, y_k) - \psi(x_{k-1}, y_{k-1})$. Then if $\sigma_x(i) < \sigma_x(j)$, we have $\epsilon = 1$, while if $\sigma_x(i) < \sigma_x(j)$, we have $\epsilon = 0$.*

Lemma 3.6.17. *Let Assumption 3.3.1 hold and let (σ_x, σ_y) be the tiebreak order for the FP dynamic. Let $A = I_n$. Let round k be a split(j, i) round, and let round $k + 1$ be a split(j, j) round. Let $\epsilon = \psi(x_k, y_k) - \psi(x_{k-1}, y_{k-1})$. Then if $\sigma_y(i) < \sigma_y(j)$, we have $\epsilon = 0$, while if $\sigma_y(i) > \sigma_y(j)$, we have $\epsilon = 1$.*

Next we prove the following lemma, which shows that the duality gap will only increase under the settings in Lemmas 3.6.16 and 3.6.17 and that the duality gap is non-decreasing.

Lemma 3.6.18. *Let Assumption 3.3.1 hold and let $A = I_n$. Then $\psi(x_k, y_k)$ is integral for all $k \geq 0$, and only increases in the settings described by Lemmas 3.6.16 and 3.6.17. In particular, if $\psi(x_k, y_k) > \psi(x_{k-1}, y_{k-1})$ for $k \geq 1$, then $\psi(x_k, y_k) = \psi(x_{k-1}, y_{k-1}) + 1$. Moreover, for $k \geq 1$, we have $\psi(x_k, y_k) \geq \psi(x_{k-1}, y_{k-1})$ (i.e. the duality gap is non-decreasing).*

We can also write the following lemma and corollary, analogous to Lemma 3.6.10 and Corollary 3.6.20.

Lemma 3.6.19. *Let Assumption 3.3.1 hold and let $A = I_n$. Suppose rounds k to $k + s - 1$ form a sync-split$(i \to j)$ pair. Let $\epsilon = \psi(x_{k+s}, y_{k+s}) - \psi(x_{k-1}, y_{k-1})$. Then $w_{k+s-1}(\ell) \leq w_{k-1}(\ell) + 2\epsilon$ for all $\ell \in [n]$.*

Corollary 3.6.20. *Let Assumption 3.3.1 hold and let $A = I_n$. Let $t \geq 0$. Let the first $t + 1$ sync phases of the FP trajectory start in rounds $s_1, ..., s_{t+1}$ respectively. Let $\epsilon_j = \psi(x_{s_{j+1}-1}, y_{s_{j+1}-1}) - \psi(x_{s_j-1}, y_{s_j-1})$. Then $w_{s_{t+1}-1}(\ell) - w_{s_1-1}(\ell) \leq 2\sum_{j=1}^{t} \epsilon_j$ for all $\ell \in [n]$. That is, $w_{s_{t+1}-1}(\ell) - w_{s_1-1}(\ell) \leq 2(\psi(x_{s_{t+1}-1}, y_{s_{t+1}-1}) - \psi(x_{s_1-1}, y_{s_1-1}))$ for all $\ell \in [n]$.*

Proof of Corollary 3.6.20. Note that each sync phase will be followed by a split phase, so by Lemma 3.6.7, rounds s_j to $s_{j+1} - 1$ form a sync-split pair. Then we have for all $\ell \in [n]$:

$$w_{s_{t+1}-1}(\ell) - w_{s_1-1}(\ell) = \sum_{j=1}^{t} (w_{s_{j+1}-1}(\ell) - w_{s_j-1}(\ell)) \leq 2\sum_{j=1}^{t} \epsilon_j$$

where the inequality follows from Lemma 3.6.19. □

Finally, we prove the following lemma, which can be proved similarly to Lemma 3.6.12.

Lemma 3.6.21. *Let Assumption 3.3.1 hold and let $A = I_n$. Suppose round k is the first round of a sync-split$(i \to j)$ pair and round $k + s - 1$ is the last round of the sync-split pair. Then $s \leq 4\psi(x_{k-1}, y_{k-1}) + 3$.*

This allows us to prove our main lemma, which gives an upper bound on the number of rounds before the duality gap increases.

Lemma 3.6.22. *Let Assumption 3.3.1 hold and let $A = I_n$. Let $k \geq 0$ and let $\delta = \psi(x_k, y_k)$. Then $\psi(x_{k+\tau}, y_{k+\tau}) \geq \delta + 1$ for $\tau \geq (4\delta + 3)(n + 1)$.*

Proof. Let (σ_x, σ_y) be the tiebreak order for the FP dynamic. We will upper bound the number of rounds before the duality gap increases. Let round $k + s$ be the earliest round in which the duality gap is larger than δ. By Lemma 3.6.18, $\psi(x_{k+s}, x_{k+s}) = \delta + 1$. Rounds k to $k + s - 1$ will consist of an alternating sequence of sync and split phases by Lemma 3.6.7. Let $t \geq 0$ be the number of sync phases we start in between rounds k and $k + s - 1$ inclusive and let the τ^{th} such sync phase be a sync(i_τ, i_τ) phase. Since the duality gap is always non-decreasing by Lemma 3.6.18, it cannot increase during rounds k to $k + s - 1$. Then by Lemma 3.6.16, we must have $\sigma_x(i_1) < \sigma_x(i_2) < ... < \sigma_x(i_t)$. Thus, $t \leq n$. Since sync and split phases alternate by Lemma 3.6.7, we can have at most $n + 1$ split phases during rounds k to $k + s - 1$. Each split phase will be part of a sync-split pair starting in some round τ such that $\psi(x_{\tau-1}, y_{\tau-1}) \leq \delta$, and these sync-split pairs also include all sync phases that start during rounds k to $k + s - 1$. Then by Lemma 3.6.21, each sync-split pair will take at most $4\delta + 3$ rounds. Thus, we have that when the duality gap is δ, it will increase to $\delta + 1$ in at most $(4\delta + 3)(n + 1)$ rounds, i.e. $s \leq (4\delta + 3)(n + 1)$. As mentioned earlier, the duality gap is non-decreasing by Lemma 3.6.18, which gives the result. □

Using Lemma 3.6.22, we can prove our main theorem.

Theorem 3.6.23. *Let Assumption 3.3.1 hold. Then the fictitious play dynamic on the $n \times n$ identity matrix initialized at $x_0 = y_0 = 0$ satisfies $\psi(\frac{1}{k}x_k, \frac{1}{k}y_k) = \Omega\left(\sqrt{\frac{1}{n}}k^{-1/2}\right)$ for $k \geq 60(n + 1)$.*

Proof of Theorem 3.6.23. For any nonnegative integer δ, let r_δ be the earliest round τ in which $\psi(x_\tau, y_\tau) = \delta$ (by Lemma 3.6.22, all r_δ are finite and well-defined). Let $B_k = \psi(x_k, y_k) - \psi(x_0, y_0)$ for all $k \geq 0$. Then by Lemma 3.6.22, B_k must satisfy the following:

$$k \leq \sum_{\delta=1}^{B_k+1} (r_\delta - r_{\delta-1}) \leq \sum_{\delta=1}^{B_k+1} (4\delta + 3)(n+1) \tag{3.12}$$

$$\iff k \leq (2B_k^2 + 9B_k + 7)(n+1) \tag{3.13}$$

If $k \geq 60(n+1)$, we must have $B_k \geq 3$, so $9B_k + 7 \leq 4B_k^2$, which gives $B_k \geq \sqrt{\frac{k}{6(n+1)}}$. □

3.7 Proofs for Section 3.6

3.7.1 Proofs of Lemmas 3.6.7-3.6.12

Lemma 3.7.1. *Suppose Assumption 3.3.1 holds. Suppose we are in a sync(i,i) phase in round k. Then this sync phase will end after some finite number of rounds s and in round $k+s+1$, we enter a split(j,i) phase for some $j \in [n]$.*

Proof. Since round k is a sync(i,i) round, we must have $p_{k-1}(i) = p^*_{k-1}$ and $q_{k-1}(i) = q^*_{k-1}$. During the sync(i,i) phase, the row player only plays action i, so $q(i)$ increases while the other entries of q stay the same. This means the column player will never switch actions in the round after a sync(i,i) round because $q(i)$ remains the maximum entry of q and is unique by Assumption 3.3.1. On the other hand, $p(i)$ increases in each round of a sync phase because $A_{ii} > 0$, while the other entries of p stay the same. Since the entries of p are finite, the row player will eventually switch in some round $k+s+1$, and in that round the column player will not switch, so we enter a split phase. □

Lemma 3.7.2. *Suppose Assumption 3.3.1 holds. Suppose we are in a split(j,i) phase in round k. Then this split phase will end after some finite number of rounds s and in round $k+s+1$, we enter a sync(j,j) phase.*

Proof. Since round k is a split(j, i) round, we must have $p_{k-1}(j) = p^*_{k-1}$ and $q_{k-1}(i) = q^*_{k-1}$. During this phase, the column player only plays action i, so $p(i)$ increases while the other entries of p stay the same. This means that the row player will never switch actions in the round after a split(j, i) round because $p(j)$ remains the minimum entry of p and is unique by Assumption 3.3.1. On the other hand, $q(j)$ increases in each round of the split phase because $A_{jj} > 0$, while all other entries of q remain the same. Since the entries of q are finite, the column player will eventually switch to action j, and in that round the row player will not switch, so we enter a sync phase. □

Proof of Lemma 3.6.7. The first two claims follow by Lemma 3.7.1 and Lemma 3.7.2. Since the dynamic will begin in a sync or split phase by definition, the dynamic must alternate between sync and split phases. Since each sync and split phase is finite, we will go through an unbounded number of sync and split phases. Moreover, we can see that the next sync phase after a sync(i, i) phase will be a sync(j, j) phase for $j \neq i$, proving the last claim. □

Proof of Lemma 3.6.8. Since the column player just plays i during the sync phase, the maximum entry of q is $q(i)$ for rounds $[k-1, k+s]$. Moreover, since the row player just plays action i, $q(i)$ is the only entry of q that changes, and $q(i)$ increases by A_{ii} in each round. Thus,

1. $q^*_{k+s} = q^*_{k-1} + (s+1)A_{ii}$
2. $v_{k+s}(i) = v_{k-1}(i) = 0$
3. $v_{k+s}(\ell) = q^*_{k+s} - q_{k+s}(\ell) = q^*_{k-1} + (s+1)A_{ii} - q_{k-1}(\ell) = v_k(\ell) + (s+1)A_{ii}$ for $\ell \neq i$

Since the row player just plays action i, $p(i)$ is the minimum entry of p for rounds $[k-1, k+s-1]$. Since the column player just plays action i, $p(i)$ increases by A_{ii} in each round, and $p(i)$ is the only entry of p that changes. However, in round $k+s$, the minimum

entry of p_{k+s} must be $p_{k+s}(j)$. So

$$\begin{aligned} p^*_{k-1} = p_{k-1}(i) = p_{k+s-1}(i) - sA_{ii} &= p_{k+s-1}(j) + (p_{k+s-1}(i) - p_{k+s-1}(j)) - sA_{ii} \\ &= p_{k+s}(j) - u_{k+s-1}(j) - sA_{ii} \\ &= p^*_{k+s} - u_{k+s-1}(j) - sA_{ii} \\ &= p^*_{k+s} - (s+1)A_{ii} + \epsilon \end{aligned}$$

where the last equality follows because $\epsilon = A_{ii} - u_{k+s-1}(j)$. Since $p_{k+s-1}(i) \leq p_{k+s-1}(j) \leq p_{k+s}(i) = p_{k+s-1}(i) + A_{ii}$, we know $A_{ii} \geq u_{k+s-1}(j)$, so $\epsilon \in [0, A_{ii}]$.

Thus,

1. $p^*_{k+s} = p^*_{k-1} + (s+1)A_{ii} - \epsilon$
2. $u_{k+s}(\ell) = u_{k-1}(\ell) - (s+1)A_{ii} + \epsilon$ for $\ell \neq i$. In particular, $u_{k+s}(j) = u_{k-1}(j) - (s+1)A_{ii} + \epsilon = 0$.
3. $u_{k+s}(i) = u_{k-1}(i) + \epsilon = \epsilon$

Putting together the above, we have: $w_{k+s}(\ell) = v_{k+s}(\ell) + u_{k+s}(\ell) = w_{k-1}(\ell) + A^{-1}_{\ell\ell}\epsilon$ for all ℓ.

Also, note that $\psi(x_{k+s}, y_{k+s}) = q^*_{k+s} - p^*_{k+s} = q^*_{k-1} - p^*_{k-1} + \epsilon = \psi(x_{k-1}, y_{k-1}) + \epsilon$. □

Proof of Lemma 3.6.9. Since the row player only plays action j, we have $p^*_\tau = p_\tau(j)$ for all $\tau \in [k-1, k+s]$. Since the column player only plays action i in these rounds, $p(i)$ is the only entry of p that increases in each round, and it increases by A_{ii} in each round. Thus, all entries of u are non-decreasing in rounds $[k-1, k+s]$, which in turn means $p^*_{k+s} = p_{k-1}(j)$. Thus,

1. $p^*_{k+s} = p^*_{k-1}$
2. $u_{k+s}(\ell) = u_{k-1}(\ell)$ for $\ell \neq j$. In particular, $u_{k+s}(j) = u_{k-1}(j) = 0$.

3. $u_{k+s}(i) = u_{k-1}(i) + (s+1)A_{ii}$

In rounds k to $k+s$, the row player only plays action j, $q(j)$ is the only entry of q that changes in these rounds, and it increases by A_{jj} in each round. Since the column player plays action i in these rounds, the maximum entry of q_τ is $q_\tau(i)$ for $\tau \in [k-1, k+s-1]$, which means that $q^*_{k+s-1} = q_{k+s-1}(i) = q_{k-1}(i) = q^*_{k-1}$. In round $k+s$, the maximum entry of q becomes $q(j)$. So

$$\begin{aligned} q^*_{k+s} = q_{k+s}(j) = q_{k+s-1}(j) + A_{jj} &= q^*_{k+s-1} + (q_{k+s-1}(j) - q^*_{k+s-1}) + A_{jj} \\ &= q^*_{k+s-1} - v_{k+s-1}(j) + A_{jj} \\ &= q^*_{k+s-1} - \epsilon \end{aligned}$$

where we used $\epsilon = A_{jj} - v_{k+s-1}(j)$. Since $q_{k+s-1}(j) \le q_{k+s-1}(i) \le q_{k+s}(j) = q_{k+s-1}(i) + A_{jj}$, we know $A_{jj} \ge v_{k+s-1}(j)$, so $\epsilon \in [0, A_{jj}]$. Thus, we have:

1. $q^*_{k+s} = q^*_{k-1} + \epsilon$
2. $v_{k+s}(\ell) = v_{k-1}(\ell) + \epsilon$ for $\ell \neq j$. In particular, $v_{k+s}(i) = v_{k-1}(i) + \epsilon = \epsilon$.
3. $v_{k+s}(j) = v_{k-1}(j) - (s+1)A_{jj} + \epsilon = 0$, so $s+1 = A^{-1}_{jj}(v_{k-1}(j) + \epsilon)$

Putting the above together, we see that $w_{k+s}(\ell) = w_{k-1}(\ell) + A^{-1}_{\ell\ell}\epsilon$ for $\ell \notin \{i, j\}$. Moreover, we see that $w_{k+s}(j) = 0$. Also,

$$\begin{aligned} w_{k+s}(i) = A^{-1}_{ii}(u_{k+s}(i) + v_{k+s}(i)) &= A^{-1}_{ii}(u_{k-1}(i) + (s+1)A_{ii} + v_{k-1}(i) + \epsilon) \\ &= A^{-1}_{ii}(u_{k-1}(i) + A_{ii}A^{-1}_{jj}(v_{k-1}(j) + \epsilon) + v_{k-1}(i) + \epsilon) \\ &= w_{k-1}(i) + A^{-1}_{jj}v_{k-1}(j) + (A^{-1}_{ii} + A^{-1}_{jj})\epsilon \\ &= w_{k-1}(i) + w_{k-1}(j) + (A^{-1}_{ii} + A^{-1}_{jj})\epsilon \end{aligned}$$

where the last equality follows because $u_{k-1}(j) = 0$. Finally, note that $\psi(x_{k+s}, y_{k+s}) = q^*_{k+s} - p^*_{k+s} = q^*_{k-1} - p^*_{k-1} + \epsilon = \psi(x_{k-1}, y_{k-1}) + \epsilon$. □

Proof of Lemma 3.6.12. Let round $k + k'$ be the first round of the split phase in the sync-split pair. We know that in round $k + k'$, the row player plays action j, so $u_{k+k'-1}(j) = 0$. Meanwhile, in each round of the sync phase, both players play action i, which causes $p(i)$ and $q(i)$ to increase by A_{ii}. Therefore, for $\tau \in [k, k + k' - 1]$, we have $u_\tau(j) = u_{\tau-1}(j) - \min\{A_{ii}, u_{\tau-1}(j)\}$ and $v_\tau(j) = v_{\tau-1}(j) + A_{ii}$. So overall, $k' \geq u_{k-1}(j)/A_{ii}$. Moreover, $v_{k+k'-1}(j) = v_{k-1}(j) + k'A_{ii}$.

In each round of the split phase, the row player plays action j, which causes $q(j)$ to increase by A_{jj} in each round, so for $\tau \in [k + k', k + s - 1]$, we have $v_\tau(j) = v_{\tau-1}(j) - \min\{A_{jj}, v_{\tau-1}(j)\}$ and $v_{k+s-1}(j) = 0$. Thus, $s - k' \geq v_{k+k'-1}(j)/A_{jj}$ rounds. We know that $v_{k+k'-1}(j) = v_{k-1}(j) + k'A_{ii} \geq v_{k-1}(j)$ because $A_{ii} > 0$, so overall we have:

$$\begin{aligned} s \geq \frac{u_{k-1}(j)}{A_{ii}} + \frac{v_{k-1}(j)}{A_{jj}} &= A_{jj}^{-1}\left(\frac{A_{jj}u_{k-1}(j)}{A_{ii}} + v_{k-1}(j)\right) \\ &\geq \frac{A_{\min}}{A_{\max}}A_{jj}^{-1}(u_{k-1}(j) + v_{k-1}(j)) = \frac{A_{\min}}{A_{\max}}w_{k-1}(j) \end{aligned}$$

□

To prove Lemma 3.6.10, we will need the following useful lemma. We will use this lemma in our lower bound proof as well.

Lemma 3.7.3. *Suppose rounds k to $k + s - 1$ form a sync-split$(i \to j)$ pair and let round $k + k'$ be the last round of the sync phase for this sync-split pair. Let $\epsilon_1 = \psi(x_{k+k'}, y_{k+k'}) - \psi(x_{k-1}, y_{k-1})$ and $\epsilon_2 = \psi(x_{k+s-1}, y_{k+s-1}) - \psi(x_{k+k'}, y_{k+k'})$. Then we have:*

1. $w_{k+s-1}(\ell) = w_{k-1}(\ell) + A_{\ell\ell}^{-1}(\epsilon_1 + \epsilon_2)$ *for* $\ell \neq \{i, j\}$
2. $w_{k+s-1}(j) = w_{k-1}(i) = 0$
3. $w_{k+s-1}(i) = w_{k-1}(j) + (A_{ii}^{-1} + A_{jj}^{-1})(\epsilon_1 + \epsilon_2)$

Proof of Lemma 3.7.3. This follows from the characterizations in Lemmas 3.6.8 and 3.6.9.

Since the last round of the sync phase is round $k + k'$, Lemma 3.6.8 gives:

$$w_{k+k'}(\ell) = w_{k-1}(\ell) + A_{\ell\ell}^{-1}\epsilon_1 \text{ for } \ell \in [n] \tag{3.14}$$

Next, by Lemma 3.6.9, we see that

$$w_{k+s-1}(i) = w_{k+k'}(i) + w_{k+k'}(j) + (A_{ii}^{-1} + A_{jj}^{-1})\epsilon_2 \tag{3.15}$$

$$w_{k+s-1}(\ell) = w_{k+k'}(\ell) + A_{\ell\ell}^{-1}\epsilon_2 \text{ for } \ell \notin \{i, j\} \tag{3.16}$$

Combining (3.14) and (3.16) immediately gives the first claim of the lemma. Next, observe that $w_{k-1}(i) = u_{k-1}(i) + v_{k-1}(i) = 0$ since round k is a sync(i, i) round. Likewise $w_{k+s-1}(j) = u_{k+s-1}(j) + v_{k+s-1}(j) = 0$ since round $k + s$ is a sync(j, j) round. This gives the second claim of the lemma. Finally, we have

$$\begin{aligned} w_{k+s-1}(i) &= w_{k+k'}(i) + w_{k+k'}(j) + (A_{ii}^{-1} + A_{jj}^{-1})\epsilon_2 \\ &= w_{k-1}(i) + A_{ii}^{-1}\epsilon_1 + w_{k-1}(j) + A_{jj}^{-1}\epsilon_1 + (A_{ii}^{-1} + A_{jj}^{-1})\epsilon_2 \\ &= w_{k-1}(j) + (A_{ii}^{-1} + A_{jj}^{-1})(\epsilon_1 + \epsilon_2) \end{aligned}$$

□

Proof of Lemma 3.6.10. This follows immediately from Lemma 3.7.3 by noting that $\epsilon = \epsilon_1 + \epsilon_2$. □

Proof of Corollary 3.6.11. For $t = 0$, this is trivially true because entries of w are always non-negative. Now assume the statement is true for $t \leq \tau - 1$ and suppose we play $\tau + 1$ consecutive sync-split pairs starting in rounds $s_1, ..., s_{\tau+1}$. Without loss of generality, let the τ^{th} sync-split pair be a sync-split$(i \to j)$ pair. By the inductive hypothesis, $w_{s_\tau - 1}(\ell) \geq \sum_{j=1}^{\tau-1} \frac{\epsilon_j}{A_{\max}}$ for $\ell \neq i$. Then by Lemma 3.6.10:

1. $w_{s_{\tau+1}-1}(\ell) \geq w_{s_\tau - 1}(\ell) + \frac{\epsilon_\tau}{A_{\max}}$ for $\ell \notin \{i, j\}$

2. $w_{s_{\tau+1}-1}(j) = w_{s_\tau - 1}(i) = 0$

3. $w_{s_{\tau+1}-1}(i) \geq w_{s_\tau - 1}(j) + \frac{2\epsilon}{A_{\max}}$.

Thus, all non-zero entries of $w_{s_{\tau+1}-1}$ are at least $\sum_{j=1}^{\tau} \frac{\epsilon_j}{A_{\max}}$. □

3.7.2 Proof of Lemma 3.6.14

Proof of Lemma 3.6.14. To prove Lemma 3.6.14, we will use Lemma 3.7.4 with $\alpha = A_{\max}$, $\beta = \delta$, and $c_j = (t - j + 1)$ for all $j \in [t]$. We plug in the resulting h^* and note that it is non-negative and its last $\lfloor \delta / A_{\max} \rfloor$ entries are $A_{\max}$, which gives:

$$\begin{aligned}
\min_{h \in \mathcal{H}} \sum_{j=1}^{t} (t - j + 1) h_j = \sum_{j=1}^{t} (t - j + 1) h_j^* &\geq \sum_{j=1}^{\lfloor \delta / A_{\max} \rfloor} j A_{\max} \\
&\geq \frac{A_{\max}(\delta / A_{\max})(\delta / A_{\max} - 1)}{2} \\
&\geq \frac{\delta^2}{4 A_{\max}}
\end{aligned}$$

where we used $\delta / A_{\max} \geq 2$ for the last inequality. □

Lemma 3.7.4. *Let $\mathcal{H} = \{h \in \mathbb{R}^t | \sum_{j=1}^{t} h_j = \beta$ and $\forall j, 0 \leq h_j \leq \alpha\}$ for $0 \leq \beta \leq t\alpha$. Let $g(h) = \sum_{j=1}^{t} c_j h_j$ for $0 < c_t < c_{t-1} < ... < c_1$. Let $h^* \in \mathbb{R}^t$ be the following vector:*

$$h^* = (\underbrace{0, 0, ..., 0, 0}_{t - \lfloor \beta/\alpha \rfloor - 1 \text{ entries}}, \beta - \alpha \cdot \lfloor \beta / \alpha \rfloor, \underbrace{\alpha, \alpha, ..., \alpha}_{\lfloor \beta/\alpha \rfloor \text{ entries}}).$$

In other words, h^* has $\lceil \beta / \alpha \rceil$ non-zeros and entries as follows:

1. *The last $\lfloor \beta / \alpha \rfloor$ entries of h^* are α. Namely, $h_j^* = \alpha$ for $j \in \{t - \lfloor \beta/\alpha \rfloor + 1, t - \lfloor \beta/\alpha \rfloor + 2, ..., t\}$.*

2. *$h^*_{t - \lfloor \beta/\alpha \rfloor} = \beta - \alpha \cdot \lfloor \beta / \alpha \rfloor$*

3. *$h_j^* = 0$ for $j \in \{1, ..., t - \lfloor \beta/\alpha \rfloor - 1\}$*

Then $h^* = \arg\min_{h \in \mathcal{H}} g(h)$.

Proof of Lemma 3.7.4. Since g is linear and $\mathcal{H}$ is a bounded, non-empty $(t-1)$-dimensional polytope, any solution $h \in \arg\min_{h \in \mathcal{H}} g(h)$ must be tight for at least $t-1$ constraints. That is, $t-1$ coordinates of h must be either 0 or α. Due to the sum constraint, it is clear that $\lfloor \beta/\alpha \rfloor$ entries of h must be α and $t - \lfloor \beta/\alpha \rfloor - 1$ entries must be 0. Moreover, the remaining entry of h must have value $\beta - \alpha \cdot \lfloor \beta/\alpha \rfloor$. In other words, the solution must be a vector h such that $h_j = h^*_{\sigma(j)}$ for some permutation σ. Then it suffices to show that only permutations that sort h^* in non-decreasing order minimize $f(\sigma) = \sum_{j=1}^{t} h^*_{\sigma(j)} c_j$. Note that c_j is sorted in decreasing order. Consider some permutation $\hat{\sigma}$ that doesn't sort h^* in non-decreasing order. Then for some $i < j$, we have $h^*_{\hat{\sigma}(i)} > h^*_{\hat{\sigma}(j)}$. Consider a permutation σ' such that $\sigma'(\ell) = \hat{\sigma}(\ell)$ for $\ell \notin \{i, j\}$ and $\sigma'(i) = \hat{\sigma}(j)$ and $\sigma'(j) = \hat{\sigma}(i)$. Then $f(\hat{\sigma}) - f(\sigma') = (h^*_{\hat{\sigma}(j)} - h^*_{\hat{\sigma}(i)})(c_j - c_i) > 0$ since $i < j$. We have shown that $\hat{\sigma}$ cannot be the minimizer of f, so the minimizer must sort h in non-decreasing order, as is the case for the identity permutation of h^*. □

3.7.3 Proof of Lemma 3.6.16

Proof of Lemma 3.6.16. Since all $A_{ii} = 1$, all entries of p must be integral. Note that $p^*_{k-1} = p_{k-1}(i)$ and $p^*_k = p_k(j)$. Since round k is a sync(i, i) round, $p_k(i) = p_{k-1}(i) + 1$. Moreover, since $q(i)$ is the maximum entry of q for rounds $k-1$ and k, we have $q^*_k = q^*_{k-1} + 1$.

If $\sigma_x(j) < \sigma_x(i)$, then we must have $p_{k-1}(i) < p_{k-1}(j)$ otherwise the x player would have played j due to the tiebreak order. This implies $p_k(i) = p_k(j)$, which means $p^*_k = p^*_{k-1} + 1$. So $\epsilon = q^*_k - p^*_k - (q^*_{k-1} - p^*_{k-1}) = 1$ in this case.

If $\sigma_x(j) > \sigma_x(i)$, then $p_k(i) > p_k(j)$. Due to the integrality of p, we have $p_k(i) = p_k(j) + 1$, which implies $p_{k-1}(j) = p_{k-1}(i)$. Thus, $p^*_k = p^*_{k-1}$. So $\epsilon = q^*_k - p^*_k - (q^*_{k-1} - p^*_{k-1}) = 0$ in this case. □

3.7.4 Proof of Lemma 3.6.17

Proof of Lemma 3.6.17. Since all $A_{ii} = 1$, all entries of q must be integral. Note that $q^*_{k-1} = q_{k-1}(i)$ and $q^*_k = q_k(j)$. Since round k is a split(j, i) round, $q_k(j) = q_{k-1}(j) + 1$. Moreover, since $p(j)$ is the maximum entry of p for rounds $k-1$ and k, we have $p^*_k = p^*_{k-1}$.

If $\sigma_y(j) < \sigma_y(i)$, then we must have $q_{k-1}(i) > q_{k-1}(j)$ otherwise the y player would have played j due to the tiebreak order. This implies $q_k(i) = q_k(j)$, which means $q^*_k = q^*_{k-1}$. So $\epsilon = q^*_k - p^*_k - (q^*_{k-1} - p^*_{k-1}) = 0$ in this case.

If $\sigma_y(j) > \sigma_y(i)$, then $q_k(i) < q_k(j)$. Due to the integrality of q, we have $q_k(j) = q_k(i) + 1$, which implies $q_{k-1}(j) = q_{k-1}(i)$. Thus, $q^*_k = q^*_{k-1} + 1$. So $\epsilon = q^*_k - p^*_k - (q^*_{k-1} - p^*_{k-1}) = 1$ in this case. □

3.7.5 Proof of Lemma 3.6.18

We first need to prove the following lemma:

Lemma 3.7.5. *Suppose rounds k and $k+1$ are both sync(i, i) rounds or both split(j, i) rounds. Then $\psi(x_k, y_k) = \psi(x_{k-1}, y_{k-1})$.*

Proof. If rounds k and $k+1$ are both of the same type, then $p^*_{k-1} = p^*_k$ and $q^*_{k-1} = q^*_k$ because both players play the same action in rounds k and $k+1$. Then we have $\psi(x_k, y_k) = q^*_k - p^*_k = q^*_{k-1} - p^*_{k-1} = \psi(x_{k-1}, y_{k-1})$. □

Proof of Lemma 3.6.18. The first claim follows because $\psi(x_k, y_k) = q^*_k - p^*_k$ and p and q are integral for $A = I_n$. For any given pair of rounds, Lemma 3.6.7 implies that either the rounds are of the same type or we encounter the settings of Lemmas 3.6.16 and 3.6.17. If the rounds are the same type, then by Lemma 3.7.5, the duality gap is unchanged. Thus, in all cases the duality gap can never increase and it only increases in the settings described by Lemmas 3.6.16 and 3.6.17. □

3.7.6 Proof of Lemma 3.6.21

We first need the following lemma, which is analogous to Lemma 3.6.12.

Lemma 3.7.6. *Let Assumption 3.3.1 hold and let A be a diagonal matrix with positive diagonal. Suppose rounds k to $k+s-1$ form a sync-split$(i \to j)$ pair. Then $s \leq \frac{2A_{\max}}{A_{\min}} w_{k-1}(j) + \frac{1}{A_{\min}} + 2$.*

Proof. Let round $k+k'$ be the first round of the split phase in the sync-split pair. We know that in round $k+k'$, the row player plays action j, so $u_{k+k'-1}(j) = 0$. Meanwhile, in each round of the sync phase, both players play action i, which causes $p(i)$ and $q(i)$ to increase by A_{ii}. Therefore, for $\tau \in [k, k+k'-1]$, we have $u_\tau(j) = u_{\tau-1}(j) - \min\{A_{ii}, u_{\tau-1}(j)\}$ and $v_\tau(j) = v_{\tau-1} + A_{ii}$. So overall, $k' \leq u_{k-1}(j)/A_{ii} + 1$. Moreover, $v_{k+k'}(j) = v_{k-1}(j) + k'A_{ii}$.

In each round of the split phase, the row player plays action j, which causes $q(j)$ to increase by 1 in each round, so for $\tau \in [k+k', k+s-1]$, we have $v_\tau(j) = v_{\tau-1}(j) - \min\{A_{jj}, v(j)\}$ and until $v_{k+s-1}(j) = 0$. Thus, $s - k' \leq v_{k+k'}(j)/A_{jj} + 1$. So overall we have:

$$\begin{aligned}
s \leq \frac{u_{k-1}(j)}{A_{ii}} + 1 + \frac{v_{k+k'}(j)}{A_{jj}} + 1 &= \frac{u_{k-1}(j)}{A_{ii}} + \frac{v_{k-1}(j) + k'A_{ii}}{A_{jj}} + 2 \\
&\leq \frac{u_{k-1}(j)}{A_{ii}} + \frac{v_{k-1}(j) + (\frac{u_{k-1}(j)}{A_{ii}} + 1)A_{ii}}{A_{jj}} + 2 \\
&\leq A_{jj}^{-1} \left(\left(\frac{A_{jj}}{A_{ii}} + 1 \right) u_{k-1}(j) + v_{k-1}(j) + 1 \right) + 2 \\
&\leq \frac{2A_{\max}}{A_{\min}} w_{k-1}(j) + \frac{1}{A_{\min}} + 2
\end{aligned}$$

□

Proof of Lemma 3.6.21. By Lemma 3.7.6, a sync-split$(i \to j)$ pair will last for at most $2w_{k-1}(j) + 3$ rounds. By Corollary 3.6.20, $w_{k-1}(j) \leq 2\psi(x_{k-1}, y_{k-1})$, which gives the result. □

3.8 Proofs for Section 3.5

Proof of Theorem 3.5.1. For all $z \in \mathcal{Z}$, we have $\phi(Sz) = \max_{w \in \mathcal{Z}} w^\top Az \geq z^\top Az = 0$. Therefore, $\min_{z \in \mathcal{Z}} \phi(Sz) \geq 0$. We will show there exists $z^* \in \mathcal{Z}$ such that $\phi(Sz^*) = 0$.

Define the set-valued map $T \colon \mathcal{Z} \to 2^{\mathcal{Z}}$ by

$$T(z) = \partial\phi(Sz) = \arg\max_{\tilde{z} \in \mathcal{Z}} \tilde{z}^\top Sz.$$

Note that $T(z)$ is a nonempty, closed, and convex set for each $z \in \mathcal{Z}$. We will show T is a closed map. Then since $\mathcal{Z}$ is compact and convex, by Kakutani's fixed point theorem, there exists a fixed point $z^* \in \mathcal{Z}$ of T, so $z^* \in T(z^*) = \partial\phi(Sz^*) = \arg\max_{\tilde{z} \in \mathcal{Z}} \tilde{z}^\top Sz^*$. Therefore, z^* satisfies $\phi(Sz^*) = (z^*)^\top Sz^* = 0$, as desired.

We now show T is a closed map, that is, if $z_n \in \mathcal{Z}$ and $w_n \in T(z_n)$ such that $\lim_{n\to\infty} z_n = z$ and $\lim_{n\to\infty} w_n = w$, then $w \in T(z)$. Note that $w_n \in T(z_n) = \partial\phi(Sz_n) = \arg\max_{\tilde{z} \in \mathcal{Z}} \tilde{z}^\top Sz_n$ means $\phi(Sz_n) = w_n^\top Sz_n$. Since ϕ is a continuous function, and $w_n \to w$, $z_n \to z$, we have $\phi(Sz) = \lim_{n\to\infty} \phi(Sz_n) = \lim_{n\to\infty} w_n^\top Sz_n = w^\top Sz$. Therefore, $w \in \arg\max_{\tilde{z} \in \mathcal{Z}} \tilde{z}^\top Sz = \partial\phi(Sz) = T(z)$, as desired. □

Proof of Lemma 3.5.3. Let $v_k = z_{k+1} - z_k = \nabla\phi(Sz_k)$. By Taylor's formula, we can write

$$\phi(Sz_{k+1}) = \phi(Sz_k) + \nabla\phi(Sz_k)^\top Sv_k + \int_0^1 (1-t)\, v_k^\top S^\top \nabla^2\phi(Sz_{k,t}) Sv_k \, dt \tag{3.17}$$

where $z_{k,t} = z_k + tv_k$ for $0 \leq t \leq 1$. Note that

$$\nabla\phi(Sz_k)^\top Sv_k = \nabla\phi(Sz_k)^\top S\nabla\phi(Sz_k) = 0.$$

Furthermore, since $z_{k,t} \in (k+t)\mathcal{Z}$, we have $\|Sz_{k,t}\| \geq (k+t)d \geq kd$, so $\nabla^2\phi(Sz_{k,t}) \preceq$

$\frac{L}{\|Sz_{k,t}\|} I \preceq \frac{L}{kd} I$. Therefore,

$$\int_0^1 (1-t)\, v_k^\top S^\top \nabla^2 \phi(Sz_{k,t}) S v_k \, dt \le \int_0^1 (1-t)\, \frac{L\|Sv_k\|^2}{kd} dt \le \frac{LD^2}{2kd}.$$

Then from (3.17) we have $\phi(Sz_{k+1}) \le \phi(Sz_k) + \frac{LD^2}{2kd}$, as desired. □

Proof of Theorem 3.5.4. From Lemma 3.5.3, we have

$$\begin{aligned}
\phi(Sz_k) &\le \phi(Sz_1) + \frac{LD^2}{2d} \sum_{\ell=1}^{k-1} \frac{1}{\ell} \\
&\le \phi(Sz_1) + \frac{LD^2}{2d} \left(1 + \int_1^{k-1} \frac{1}{t}\, dt\right) \\
&= \phi(Sz_1) + \frac{LD^2}{2d} \left(1 + \log(k-1)\right).
\end{aligned}$$

□

3.8.1 Auxiliary results for $0 \in (S\mathcal{Z})^\circ$

We will use the following auxiliary result.

Lemma 3.8.1. *Suppose* $0 \in (S\mathcal{Z})^\circ$. *Assume Assumption 3.5.2. For all* $\theta \in \mathbb{R}^n \setminus \{0\}$,

$$|\cos(\theta, S\nabla\phi(\theta))| \le \sqrt{1 - \frac{m^2}{R^2}}.$$

Proof. Note that $\phi(\theta) = \theta^\top \nabla\phi(\theta)$. Then for all $\theta \in \mathbb{R}^n \setminus \{0\}$ with $\hat{\theta} = \theta/\|\theta\|$,

$$\cos\angle(\theta, \nabla\phi(\theta)) = \frac{\theta^\top \nabla\phi(\theta)}{\|\theta\|\|\nabla\phi(\theta)\|} = \frac{\phi(\theta)}{\|\theta\|\|\nabla\phi(\theta)\|} = \frac{\phi(\hat{\theta})}{\|\nabla\phi(\theta)\|} \ge \frac{m}{R}.$$

Furthermore, since $S^\top = -S$,

$$\cos(\nabla\phi(\theta), S\nabla\phi(\theta)) = \frac{\nabla\phi(\theta)^\top S \nabla\phi(\theta)}{\|\nabla\phi(\theta)\|\|S\nabla\phi(\theta)\|} = 0.$$

Then by Lemma 3.8.4,

$$|\cos\angle(\theta, S\nabla\phi(\theta))| \le \sqrt{1 - \frac{m^2}{R^2}}$$

as desired. □

Lemma 3.8.2. *Assume Assumption 3.5.2. Let* $L = \sup_{\|\theta\|=1} \|\nabla^2\phi(\theta)\| < \infty$. *For all* $\theta, v \in \mathbb{R}^n \setminus \{0\}$ *with* $|\cos\angle(\theta, v)| \le C < 1$, *we have*

$$\langle \nabla\phi(\theta + v) - \nabla\phi(\theta), v\rangle \le \frac{L}{\sqrt{1 - C^2}} \frac{\|v\|^2}{\|\theta\|}.$$

Proof. Let $c \equiv \cos\angle(\theta, v)$ and $r = \|v\|/\|\theta\|$. For $0 \le t \le 1$, let $\theta_t = \theta + tv$, so

$$\begin{aligned}
\|\theta_t\|^2 &= \|\theta\|^2 + 2tv^\top\theta + t^2\|v\|^2 \\
&= \|\theta\|^2 + 2ct\|v\|\|\theta\| + t^2\|v\|^2 \\
&= \|\theta\|^2(1 + 2ctr + t^2r^2) \\
&\ge \|\theta\|^2(1 - 2Ctr + t^2r^2) \\
&= \|\theta\|^2(1 - C^2 + (C - tr)^2) \\
&\ge \|\theta\|^2(1 - C^2). \qquad (3.18)
\end{aligned}$$

Let $\hat\theta_t = \theta_t/\|\theta_t\|$. We can write $\nabla\phi(\theta + v) - \nabla\phi(\theta) = \int_0^1 \nabla^2\phi(\theta_t)v\, dt$, so

$$\begin{aligned}
\langle \nabla\phi(\theta + v) - \nabla\phi(\theta), v\rangle &= \int_0^1 \langle v, \nabla^2\phi(\theta_t)v\rangle\, dt \\
&= \int_0^1 \frac{\langle v, \nabla^2\phi(\hat\theta_t)v\rangle}{\|\theta_t\|}\, dt \\
&\le \int_0^1 \frac{L\|v\|^2}{\|\theta_t\|}\, dt \\
&\overset{(3.18)}{\le} \frac{L}{\sqrt{1 - C^2}} \frac{\|v\|^2}{\|\theta\|}
\end{aligned}$$

as desired. □

Proof of Lemma 3.5.5. Let $v_k = z_{k+1} - z_k = \nabla\phi(Sz_k)$ so $\|v_k\| \le R$. By Lemma 3.8.1,

$$|\cos\angle(Sz_k, Sv_k)| \le C := \sqrt{1 - \frac{m^2}{R^2}}.$$

Then by Jensen's inequality and Lemma 3.8.2,

$$\begin{aligned}
\phi(Sz_{k+1}) - \phi(Sz_k) &\le \langle \nabla\phi(Sz_{k+1}),\, S(z_{k+1} - z_k)\rangle \\
&= \langle \nabla\phi(Sz_{k+1}) - \nabla\phi(Sz_k),\, Sz_{k+1} - Sz_k\rangle \\
&\le \frac{L}{\sqrt{1-C^2}} \frac{\|Sv_k\|^2}{\|Sz_k\|} \\
&= \frac{LR}{m} \frac{\|Sv_k\|^2}{\|Sz_k\|} \\
&\le \frac{LR}{m} \frac{\|S\|^2 R^2}{\|Sz_k\|} \\
&\le \frac{LR}{m} \frac{\|S\|^2 R^3}{\phi(Sz_k)}.
\end{aligned}$$

Therefore,

$$\phi(Sz_{k+1}) \le \phi(Sz_k) + \frac{L'}{\phi(Sz_k)}$$

where $L' := LR^4\|S\|^2/m$. □

Proof of Theorem 3.5.6. By Lemma 3.5.5,

$$\begin{aligned}
\phi(Sz_{k+1})^2 &\le \phi(Sz_k)^2 + \frac{L'^2}{\phi(Sz_k)^2} + 2L' \\
&\le \phi(Sz_k)^2 + \frac{L'^2}{\phi(Sz_1)^2} + 2L'
\end{aligned}$$

where the last inequality holds since $\phi(Sz_k) \geq \phi(Sz_1)$. Therefore,

$$\phi(Sz_k)^2 \leq \phi(Sz_1)^2 + (k-1)L'' = O(k)$$

where $L'' := \frac{L'^2}{\phi(\theta_1)^2} + 2L'$, so

$$\phi(Sz_k) \leq \sqrt{\phi(Sz_1)^2 + (k-1)L''} = O(\sqrt{k})$$

as desired. □

3.8.2 Auxiliary results for strong convexity

Lemma 3.8.3. *Suppose ϕ is α-orthogonally strongly convex. For all $\theta, v \in \mathbb{R}^n \setminus \{0\}$,*

$$v^\top \nabla^2 \phi(\theta) v \geq \alpha \frac{\|v\|^2}{\|\theta\|} \sin^2 \angle(\theta, v)$$

where $\angle(\theta, v)$ is the angle between θ and v (from the origin).

Proof. Let $\hat{\theta} = \theta/\|\theta\|$ and $\hat{v} = v/\|v\|$. Note that by the homogeneity property of ϕ,

$$v^\top \nabla^2 \phi(\theta) v = \frac{\|v\|^2}{\|\theta\|} \hat{v}^\top \nabla^2 \phi(\hat{\theta}) \hat{v}.$$

Let $c \equiv \cos \angle(\theta, v) = \hat{\theta}^\top \hat{v}$. Let $v_\perp = \hat{v} - (\hat{\theta}^\top \hat{v})\hat{\theta} = \hat{v} - c\hat{\theta}$ denote the component of $\hat{v}$ orthogonal to $\hat{\theta}$, and note that $\|v_\perp\|^2 = 1 - c^2 = \sin^2 \angle(\theta, v)$. Then since $\nabla^2 \phi(\hat{\theta})\hat{\theta} = 0$ and using the definition of α-orthogonal strong convexity, we have

$$\hat{v}^\top \nabla^2 \phi(\hat{\theta}) \hat{v} = (v_\perp + c\hat{\theta})^\top \nabla^2 \phi(\hat{\theta})(v_\perp + c\hat{\theta}) = v_\perp^\top \nabla^2 \phi(\hat{\theta}) v_\perp \geq \alpha \|v_\perp\|^2 = \sin^2 \angle(\theta, v)$$

as desired. □

Lemma 3.8.4. *Let $u, v, w \in \mathbb{R}^n \setminus \{0\}$ with $\cos \angle(u, v) = 0$ and $|\cos \angle(v, w)| \geq c \geq 0$.*

Then

$$\sin^2 \angle(u, w) \geq c^2.$$

Proof. Without loss of generality assume $\|u\| = \|v\| = \|w\| = 1$. We choose a coordinate system such that $u = (1, 0, \ldots, 0)$, $v = (0, 1, 0, \ldots, 0)$, and $w = (x, y, z)$ for some $x, y \in \mathbb{R}$, $z \in \mathbb{R}^{n-2}$, with $x^2 + y^2 + \|z\|^2 = \|w\|^2 = 1$. Then $\cos \angle(u, v) = u^\top v = 0$ and $\cos \angle(v, w) = v^\top w = y$, so we assume $|y| \geq c \geq 0$. $c(v, w) = v^\top w = y$ and $c(u, w) = u^\top w = x$. Furthermore, $\cos \angle(u, w) = u^\top w = x$, and

$$\sin^2 \angle(u, w) = 1 - x^2 = y^2 + \|z\|^2 \geq y^2 \geq c^2$$

as desired. □

Lemma 3.8.5. *Let $\theta, v \in \mathbb{R}^n \setminus \{0\}$, let $r = \|v\|/\|\theta\|$, and $c = \cos \angle(\theta, v)$. Then for all $0 \leq t \leq 1$,*

$$c(\theta + tv,\ v) = \frac{c + tr}{\sqrt{1 + 2ctr + t^2 r^2}}.$$

Proof. Let $c = c(\theta, v)$ and $r = \|v\|/\|\theta\|$. For $0 \leq t \leq 1$, let $\theta_t = \theta + tv$, so

$$\|\theta_t\|^2 = \|\theta\|^2(1 + 2ctr + t^2 r^2).$$

Then

$$c(\theta_t, v) = \frac{\theta_t^\top v}{\|\theta_t\|\|v\|} = \frac{\theta^\top v + t\|v\|^2}{\|\theta\|\|v\|\sqrt{1 + 2ctr + t^2 r^2}} = \frac{c + tr}{\sqrt{1 + 2ctr + t^2 r^2}}.$$

□

Proof of Lemma 3.5.9. Let $v_k = Sz_{k+1} - Sz_k = \nabla\phi(Sz_k) \in \mathcal{Z}$, so $\|Sv_k\| \geq R_0$. Let

$c \equiv \cos(Sz_k, Sv_k)$, so $|c| \leq \sqrt{1 - \frac{m^2}{R^2}}$ by Lemma 3.8.1, and let $r = \|Sv_k\|/\|Sz_k\|$. For $0 \leq t \leq 1$, let $z_{k,t} = z_k + tv_k$, so

$$\|Sz_{k,t}\| = \|Sz_k\|\sqrt{1 + 2ctr + t^2r^2}. \tag{3.19}$$

Let $c_t = \cos(Sz_{k,t}, Sv_k)$, so by Lemma 3.8.5,

$$c_t = \frac{c + tr}{\sqrt{1 + 2ctr + t^2r^2}}. \tag{3.20}$$

Let $\widehat{Sz_{k,t}} = Sz_{k,t}/\|Sz_{k,t}\|$ and $\widehat{Sv_k} = Sv_k/\|Sv_k\|$. Then by Lemma 3.8.3,

$$\begin{aligned}
\phi(Sz_{k+1}) - \phi(Sz_k) &= \int_0^1 (1-t)\,(Sv_k)^\top \nabla^2\phi(Sz_{k,t})Sv_k\,dt \\
&= \int_0^1 (1-t)\,\frac{\|Sv_k\|^2}{\|Sz_{k,t}\|}\widehat{Sv_k}^\top \nabla^2\phi(\widehat{Sz_{k,t}})\widehat{Sv_k}\,dt \\
&\overset{(3.19)}{=} \frac{\|Sv_k\|^2}{\|Sz_k\|}\int_0^1 (1-t)\,\frac{\widehat{Sv_k}^\top \nabla^2\phi(\widehat{Sz_{k,t}})\widehat{Sv_k}}{\sqrt{1 + 2ctr + t^2r^2}}\,dt \\
&\geq \frac{\|Sv_k\|^2}{\|Sz_k\|}\int_0^1 (1-t)\,\frac{\alpha(1-c_t^2)}{\sqrt{1 + 2ctr + t^2r^2}}\,dt \\
&\overset{(3.20)}{=} \frac{\|Sv_k\|^2}{\|Sz_k\|}\int_0^1 (1-t)\,\frac{\alpha(1-c^2)}{(1 + 2ctr + t^2r^2)^{3/2}}\,dt \\
&\geq \frac{\|Sv_k\|^2}{\|Sz_k\|}\,\frac{\alpha m^2}{R^2}\int_0^1 \frac{1-t}{(1 + 2ctr + t^2r^2)^{3/2}}\,dt.
\end{aligned}$$

We consider two cases:

- Suppose $r \leq 1$. Then since $1 + 2ctr + t^2r^2 \leq 1 + 2 + 1 = 4$,

$$\int_0^1 \frac{1-t}{(1 + 2ctr + t^2r^2)^{3/2}}\,dt \geq \frac{1}{8}\int_0^1 (1-t)\,dt = \frac{1}{16}.$$

Therefore, in this case

$$\begin{aligned}\phi(Sz_{k+1}) - \phi(Sz_k) &\geq \frac{\alpha m^2}{16R^2}\frac{\|Sv_k\|^2}{\|Sz_k\|} \\ &\geq \frac{\alpha m^2 R_0^2}{16R^2}\frac{1}{\|Sz_k\|} \\ &\geq \frac{\alpha m^3 R_0^2}{16R^2}\frac{1}{\phi(Sz_k)}.\end{aligned}$$

- Suppose $r > 1$. Then for $0 \leq t \leq \frac{1}{2r}$, $1 + 2ctr + t^2r^2 \leq 1 + 1 + \frac{1}{4} = \frac{9}{4}$, so

$$\begin{aligned}\int_0^1 \frac{1-t}{(1+2ctr+t^2r^2)^{3/2}}\,dt &\geq \int_0^{\frac{1}{2r}} \frac{1-t}{(1+2ctr+t^2r^2)^{3/2}}\,dt \\ &\geq \int_0^{\frac{1}{2r}} \frac{1-\frac{1}{2r}}{(3/2)^3}\,dt \\ &= \frac{1}{2r}\left(1-\frac{1}{2r}\right)\frac{8}{27} \\ &\geq \frac{2}{27r} \\ &\geq \frac{1}{16r}.\end{aligned}$$

Therefore, in this case

$$\begin{aligned}\phi(Sz_{k+1}) - \phi(Sz_k) &\geq \frac{\alpha m^2}{R^2}\frac{\|Sv_k\|^2}{\|Sz_k\|}\frac{1}{16r} \\ &= \frac{\alpha m^2}{16R^2}\|Sv_k\| \\ &\geq \frac{\alpha m^2 R_0}{16R^2} \\ &\geq \frac{\alpha m^2 R_0}{16R^2}\frac{\phi(Sz_1)}{\phi(Sz_k)} \\ &\geq \frac{\alpha m^3 R_0\|Sz_1\|}{16R^2}\frac{1}{\phi(Sz_k)}.\end{aligned}$$

In both cases above we have

$$\phi(Sz_{k+1}) \geq \phi(Sz_k) + \frac{C}{\phi(Sz_k)}$$

with $C = \frac{\alpha m^3 R_0}{16R^2} \min\{\|\theta_1\|,\ R_0\}$, as desired. □

Proof of Theorem 3.5.10. By Lemma 3.5.9,

$$\phi(Sz_{k+1})^2 \geq \phi(Sz_k)^2 + 2C + \frac{C^2}{\phi(Sz_k)^2} \geq \phi(Sz_k)^2 + 2C.$$

Therefore,

$$\phi(Sz_k)^2 \geq \phi(Sz_1)^2 + 2C(k-1),$$

so

$$\phi(Sz_k) \geq \sqrt{\phi(Sz_1)^2 + 2C(k-1)} = \Omega(\sqrt{k}).$$

□

Proof of Lemma 3.5.12. Since ϕ is convex, by Jensen's inequality,

$$\begin{aligned}
\phi(Sz_{k+1}) - \phi(Sz_k) &\leq \nabla\phi(Sz_{k+1})^\top S(z_{k+1} - z_k) \\
&= \nabla\phi(Sz_{k+1})^\top S\nabla\phi(Sz_{k+\frac{1}{2}}) \\
&= \left(\nabla\phi(Sz_{k+1}) - \nabla\phi(Sz_{k+\frac{1}{2}})\right)^\top S\nabla\phi(Sz_{k+\frac{1}{2}}) \\
&\leq \|\nabla\phi(Sz_{k+1}) - \nabla\phi(Sz_{k+\frac{1}{2}})\|\,\|S\nabla\phi(Sz_{k+\frac{1}{2}})\| \\
&\leq \|\nabla\phi(Sz_{k+1}) - \nabla\phi(Sz_{k+\frac{1}{2}})\|\,D. \qquad (3.21)
\end{aligned}$$

By Taylor's formula, we can write

$$\begin{aligned}\nabla\phi(Sz_{k+1}) - \nabla\phi(Sz_{k+\frac{1}{2}}) &= \int_0^1 \nabla^2\phi(S\tilde{z}_{k,t})S(z_{k+1} - z_{k+\frac{1}{2}})\,dt \\ &= \int_0^1 \nabla^2\phi(S\tilde{z}_{k,t})S(\nabla\phi(Sz_{k+\frac{1}{2}}) - \nabla\phi(Sz_k))\,dt\end{aligned}$$

where for $0 \le t \le 1$, $\tilde{z}_{k,t} = (1-t)z_{k+\frac{1}{2}} + tz_{k+1} = z_k + (1-t)\nabla\phi(Sz_k) + t\nabla\phi(Sz_{k+\frac{1}{2}})$. Note that $\tilde{z}_{k,t} \in (k+1)\mathcal{Z}$, so $\|S\tilde{z}_{k,t}\| \ge d(k+1)$, which implies $\nabla^2\phi(S\tilde{z}_{k,t}) \preceq \frac{L}{d(k+1)}I$, and thus

$$\|\nabla\phi(Sz_{k+1}) - \nabla\phi(Sz_{k+\frac{1}{2}})\| \le \frac{L}{d(k+1)}\,\|S\|\,\|\nabla\phi(Sz_{k+\frac{1}{2}}) - \nabla\phi(Sz_k)\|. \tag{3.22}$$

Again by Taylor's formula, we can write

$$\begin{aligned}\nabla\phi(Sz_{k+\frac{1}{2}}) - \nabla\phi(Sz_k) &= \int_0^1 \nabla^2\phi(Sz_{k,t})S(z_{k+\frac{1}{2}} - z_k)\,dt \\ &= \int_0^1 \nabla^2\phi(Sz_{k,t})S\nabla\phi(Sz_k)\,dt\end{aligned}$$

where for $0 \le t \le 1$, $z_{k,t} = (1-t)z_k + tz_{k+\frac{1}{2}} = z_k + t\nabla\phi(Sz_k)$. Note that $z_{k,t} \in (k+t)\mathcal{Z}$, so $\|Sz_{k,t}\| \ge d(k+t) \ge dk$, which implies $\nabla^2\phi(Sz_{k,t}) \preceq \frac{L}{dk}I$, and thus

$$\|\nabla\phi(Sz_{k+\frac{1}{2}}) - \nabla\phi(Sz_k)\| \le \frac{L}{dk}\|S\nabla\phi(Sz_k)\| \le \frac{LD}{dk}.$$

Plugging this to (3.22) and back to (3.21), we conclude $\phi(Sz_{k+1}) \le \phi(Sz_k) + \frac{L^2D^2\|S\|}{d^2k(k+1)}$, as desired. □

Theorem of Lemma 3.5.13. By Lemma 3.5.12 and writing $\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$, we have

$$\begin{aligned}
\phi(Sz_k) &\leq \phi(Sz_1) + \frac{L^2 D^2 \|S\|}{d^2} \sum_{\ell=1}^{k-1} \left(\frac{1}{\ell} - \frac{1}{\ell+1} \right) \\
&= \phi(Sz_1) + \frac{L^2 D^2 \|S\|}{d^2} \left(1 - \frac{1}{k} \right) \\
&\leq \phi(Sz_1) + \frac{L^2 D^2 \|S\|}{d^2}.
\end{aligned}$$

□

CHAPTER 4

LAST-ITERATE CONVERGENCE RATES FOR MIN-MAX OPTIMIZATION

In this chapter, we focus on *last-iterate* convergence of algorithms in smooth unconstrained min-max optimization problems. While classic work in convex-concave min-max optimization relies on average-iterate convergence results, the emergence of nonconvex applications such as training Generative Adversarial Networks has led to renewed interest in last-iterate convergence guarantees. Proving last-iterate convergence is challenging because many natural algorithms, such as Gradient Descent/Ascent, provably diverge or cycle even in simple convex-concave min-max settings, and previous work on global last-iterate convergence rates has been limited to the bilinear and convex-strongly concave settings. We show that the HAMILTONIAN GRADIENT DESCENT (HGD) algorithm achieves linear convergence in a variety of more general settings, including convex-concave problems that satisfy a novel sufficiently bilinear condition. We also prove convergence rates for stochastic HGD and for some parameter settings of the Consensus Optimization algorithm of [MNG17].

4.1 Introduction

Last-iterate convergence guarantees for min-max problems have been challenging to prove since standard analysis of no-regret algorithms says essentially nothing about last-iterate convergence. Widely used no-regret algorithms, such as Gradient Descent/Ascent (GDA), fail to converge even in the simple *bilinear* setting where $g(x, y) = x^\top Cy$ for some arbitrary matrix C. GDA provably cycles in continuous time and diverges in discrete time (see for example [DISZ18; MGN18]). In fact, the full range of Follow-The-Regularized-Leader (FTRL) algorithms provably do not converge in zero-sum games with interior equilibria [MPP18]. This occurs because the iterates of the FTRL algorithms exhibit cyclic behavior, a phenomenon commonly observed when training GANs in practice as well.

Much of the recent research on last-iterate convergence in min-max problems has focused on *asymptotic* or *local* convergence [Mer+19; MNG17; DP18; Bal+18; Let+19; MJS19]. While these results are certainly useful, one would ideally like to prove *global non-asymptotic* last-iterate convergence rates. Provable global convergence rates allow for quantitative comparison of different algorithms and can aid in choosing learning rates and architectures to ensure fast convergence in practice. Yet despite the extensive amount of literature on convergence rates for convex optimization, very few global last-iterate convergence rates have been proved for min-max problems. Prior work on global last-iterate convergence rates has been limited to the bilinear or convex-strongly concave settings [Tse95; LS19; DH19; MOP19]. In particular, the following basic question is still open:

"What global last-iterate convergence rates are achievable for convex-concave min-max problems?"

Understanding global last-iterate rates in the convex-concave setting is an important stepping stone towards provable last-iterate rates in the nonconvex-nonconcave setting. Motivated by this, we prove new linear last-iterate convergence rates in the convex-concave setting for an algorithm called HAMILTONIAN GRADIENT DESCENT (HGD) under weaker assumptions compared to previous results. HGD is gradient descent on the squared norm of the gradient, and it has been mentioned in [MNG17; Bal+18]. Our results are the first to show non-asymptotic convergence of an efficient algorithm in settings that not linear or strongly convex in either input. In particular, we introduce a novel "sufficiently bilinear" condition on the second-order derivatives of the objective g and show that this condition is sufficient for HGD to achieve linear convergence in convex-concave settings. The "sufficiently bilinear" condition appears to be a new sufficient condition for linear convergence rates that is distinct from previously known conditions such as the Polyak-Łojasiewicz (PL) condition or pure bilinearity. Our analysis relies on showing that the squared norm of the gradient satisfies the PL condition in various settings. As a corollary of this result, we can leverage [KNS16] to show that a stochastic version of HGD will have a last-iterate convergence rate of $O(1/\sqrt{k})$

in the "sufficiently bilinear" setting, albeit with an additional smoothness assumption. On the practical side, while vanilla HGD has issues training GANs in practice, [MNG17] show that a related algorithm known as Consensus Optimization (CO) can effectively train GANs in a variety of settings, including on CIFAR-10 and celebA. We show that CO can be viewed as a perturbation of HGD, which implies that for some parameter settings, CO converges at the same rate as HGD.

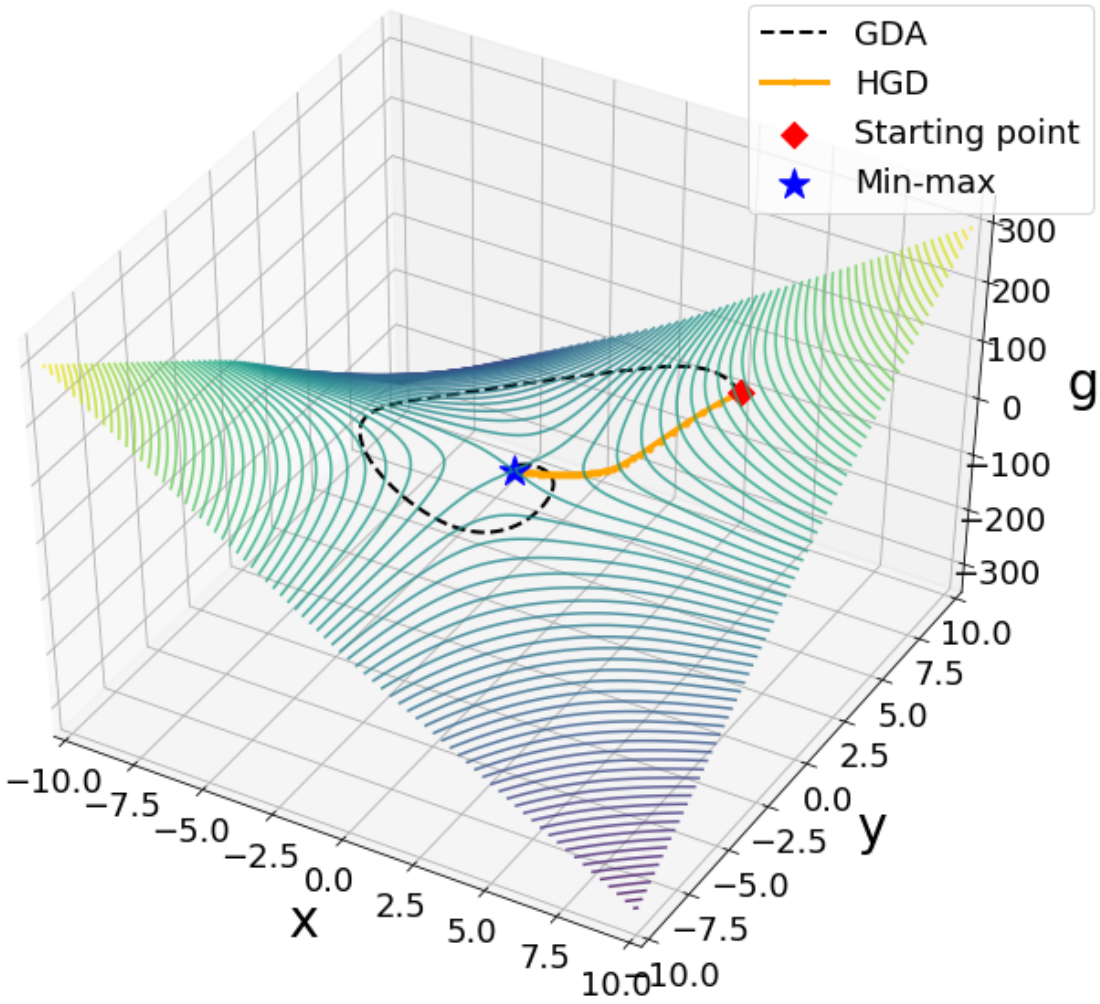


Figure 4.1: HGD converges quickly, while GDA spirals. This nonconvex-nonconcave objective is defined in Section 4.14.

We begin in Section 4.2 with background material and notation, including some of our key assumptions. In Section 4.4, we discuss Hamiltonian Gradient Descent (HGD), and

we present our linear convergence rates for HGD in various settings. In Section 4.5, we present some of the key technical components used to prove our results from Section 4.4. We present our results for Stochastic HGD in Section 4.6, and we present our results for Consensus Optimization in Section 4.7. The details of our proofs are in Section 4.13.

4.2 Preliminaries

Notation Recall that we use ξ to denote the gradient descent ascent vector field, as defined in (1.4). Under this notation, the Gradient Descent/Ascent (GDA) update can be written as $z_{k+1} = z_k - \eta\xi(z_k)$.

For notational convenience, we will use J to denote the Jacobian of ξ, i.e.

$$J \equiv \nabla\xi = \begin{pmatrix} \nabla^2_{xx} g & \nabla^2_{xy} g \\ -\nabla^2_{yx} g & -\nabla^2_{yy} g \end{pmatrix}.$$

Note that unlike the Hessian in standard optimization, J is not symmetric, due to the negative sign in ξ. When clear from the context, we often omit dependence on x when writing $\xi, J, g, \mathcal{H}$, and other functions. Note that ξ, J, and $\mathcal{H}$ are defined for a given objective g – we omit this dependence as well for notational clarity. We will always assume g is sufficiently differentiable whenever we take derivatives. In particular, we assume third-order differentiability in Section 4.4.

We will also use the following non-standard definitions for notational convenience:

Definition 4.2.1 (Higher-order Lipschitz)**.** *A function $g : \mathbb{R}^n \to \mathbb{R}$ is (L_2, L_3)-Lipschitz if for all $z \in \mathbb{R}^n$, we have $||\nabla\xi(z)|| \leq L_2$ and $||\nabla J(z)|| \leq L_3$.*

Definition 4.2.2 (Smoothness at a point)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth at a point z if $||\nabla^2 f(z)|| \leq L$.*

Notions of convergence in min-max problems Since we are in the unconstrained setting, the normal notion of duality gap (2.3) is unsuitable. As such, we use a different notion

of convergence, as we now describe. Our rates will apply to min-max problems where g satisfies the following assumption:

Assumption 4.2.3. *All critical points of the objective g are global min-maxes (i.e. they satisfy (2.1)).*

In other words, we prove convergence rates to min-maxes in settings where convergence to critical points is necessary and sufficient for convergence to min-maxes. This assumption is true for convex-concave settings, but also holds for some nonconvex-nonconcave settings, as we discuss in Section 4.10. This assumption allows us to measure the convergence of our algorithms to *ϵ-approximate critical points*, defined as follows:

Definition 4.2.4. *Let $\epsilon \geq 0$. A point $z \in \mathbb{R}^n \times \mathbb{R}^n$ is an ϵ-approximate critical point if $||\xi(z)|| \leq \epsilon$.*

Convergence to approximate critical points is a common goal in standard convex and nonconvex optimization (see for example [AZH16; GL16; CHDS17; Aga+17]), as it is a necessary condition for convergence to local or global minima, and it is a natural measure of convergence since the value of g at a given point gives no information about how close we are to a min-max. Our main convergence rate results focus on this first-order notion of convergence, which is sufficient given Assumption 4.2.3. We discuss notions of second-order convergence and ways to adapt our results to the general nonconvex setting in Section 4.3.

4.3 Related work

Asymptotic and local convergence In standard nonconvex optimization, a common goal is to find second-order local minima, which are approximate critical points where $\nabla^2 f$ is approximately positive definite. Likewise, a common goal in nonconvex min-max optimization is to find approximate critical points where an analogous second-order condition holds, namely that $\nabla^2_{xx} g(x)$ is approximately positive definite and $\nabla^2_{yy} g(x)$ is approximately negative definite. Critical points where this second-order condition holds are called *local*

min-maxes. When Assumption 4.2.3 holds, all critical points are *global* min-maxes, but in more general settings, we may encounter critical points that do not satisfy these conditions. Critical points may be local min-mins or max-mins or indefinite points. A number of recent papers have proposed dynamics for nonconvex min-max optimization, showing local stability or local asymptotic convergence results [MNG17; Mer+19; DP18; Bal+18; Let+19; MJS19]. The key guarantee that these papers generally give is that their algorithms will be stable at local min-maxes and unstable at some set of undesirable critical points (such as local max-mins). This essentially amounts to a guarantee that in the convex-concave setting, their algorithms will converge asymptotically and in the strictly concave-strictly convex setting (i.e. where there is only an undesirable *max-min*), their algorithms will diverge asymptotically. This type of local stability is essentially the best one can ask for in the general nonconvex setting, and we show how to give similar guarantees for our algorithm in Section 4.8.

Non-asymptotic convergence rates Work on global non-asymptotic last-iterate convergence rates has been limited to very restrictive settings. A classic result by [Roc76] shows a linear convergence rate for the proximal point method in the bilinear and strongly convex-strongly concave cases. Another classic result, by [Tse95], shows a linear convergence rate for the extragradient algorithm in the bilinear case. [LS19] show that a number of algorithms achieve a linear convergence rate in the bilinear case, including Optimistic Mirror Descent (OMD) and Consensus Optimization (CO). They also show that GDA obtains a linear convergence rate in the strongly convex-strongly concave case. [MOP19] show that OMD and EG obtain a linear rate for the strongly convex-strongly concave case, in addition to proving similar results for generalized versions of both algorithms. [DH19] show that GDA achieves a linear convergence rate for a convex-strongly concave setting with a full column rank linear interaction term.[1] Finally, concurrent work by [AMLJG19]

[1]Specifically, they assume $g(x, y) = f(x) + y^T Ax - h(y)$, where f is smooth and convex, h is smooth and strongly convex, and A has full column rank. We make a brief comparison of our work to that of [DH19]

shows global linear convergence rates for various algorithms in a very similar setting to ours. Other concurrent work by [Azi+20] provides convergence rates for an accelerated variant of Consensus Optimization.

Non-uniform average-iterate convergence A number of recent works have studied the convergence of non-uniform averages of iterates. Iterate averaging can lend stability to an algorithm or improve performance if the algorithm cycles around the solution. On the other hand, uniform averages can suffer from worse performance in nonconvex settings if early iterates are far from optimal. Non-uniform averaging is a way to achieve the stability benefits of iterate averaging while potentially speeding up convergence compared to uniform averaging. In this way, one can view non-uniform averaging as an interpolation between average-iterate and last-iterate algorithms.

One popular non-uniform averaging scheme is the exponential moving average (EMA). For an algorithm with iterates $z^{(0)}, ..., z^{(T)}$, the EMA at iterate t is defined recursively as

$$z_{EMA}^{(t)} = \beta z_{EMA}^{(t-1)} + (1-\beta) z_{EMA}^{(t-1)}$$

where $z_{EMA}^{(0)} = z^{(0)}$ and $\beta < 1$. A typical value for β is 0.999. [Yaz+19] and [GBVLJ19] show that uniform and EMA schemes can improve GAN performance on a variety of datasets. [MGN18] and [KALL18] use EMA to evaluate the GAN models they train, showing the effectiveness of EMA in practice.

In terms of theoretical results, [Kro19] studies saddle point problems of the form

$$\min_x \max_y f(x) + g(x) + \langle Kx, y \rangle - h^*(y),$$

where f is a smooth convex function, g and h are convex functions with easily computable prox-mappings, and K is some linear operator. They show that for certain algorithms,

for the convex-strongly concave setting in Section 4.9.

linear averaging and quadratic averaging schemes are provably at least as good as the uniform average scheme in terms of iterate complexity. [ALLW18b] show how linear and exponential averaging schemes can be used to achieve faster convergence rates in some specific convex-concave games.

Overall, while non-uniform averaging is appealing for a variety of reasons, there is currently no theoretical explanation for why it outperforms uniform averages or why it would converge at all in many settings. In fact, one natural way to show convergence for an EMA scheme would be to show last-iterate convergence.

4.4 Hamiltonian Gradient Descent

Our main algorithm for finding saddle points of $g(x, y)$ is called HAMILTONIAN GRADIENT DESCENT (HGD). HGD consists of performing gradient descent on a particular objective function $\mathcal{H}$ that we refer to as the *Hamiltonian*, following the terminology of [Bal+18].[2] If we let $\xi := \left(\frac{\partial g}{\partial x}, -\frac{\partial g}{\partial y}\right)$ be the vector of (appropriately-signed) partial derivatives, then the Hamiltonian is:

$$\mathcal{H}(z) := \tfrac{1}{2}\|\xi(z)\|^2 = \tfrac{1}{2}\left(\|\tfrac{\partial g}{\partial x}(z)\|^2 + \|\tfrac{\partial g}{\partial y}(z)\|^2\right).$$

Since a critical point occurs when $\xi(z) = 0$, we can find a (approximate) critical point by finding a (approximate) minimizer of $\mathcal{H}$. Moreover, under Assumption 4.2.3, finding a critical point is equivalent to finding a saddle point. This motivates the HGD update procedure on $z_k = (x_k, y_k)$ with step-size $\eta > 0$:

$$z_{k+1} = z_k - \eta \nabla \mathcal{H}(z_k), \tag{4.1}$$

HGD has been mentioned in [MNG17; Bal+18], and it strongly resembles the Consensus

[2] We note that the function $\mathcal{H}$ is not the Hamiltonian as in the sense of classical physics, as we do not use the symplectic structure in our analysis, but rather we only perform gradient descent on $\mathcal{H}$.

Optimization (CO) approach of [MNG17]. The HGD update requires a Hessian-vector product because $\nabla\mathcal{H} = \xi^\top J$, making HGD a second-order iterative scheme. However, Hessian-vector products are cheap to compute when the objective is defined by a neural net, taking only two gradient oracle calls [Pea94]. This makes the Hessian-vector product oracle a theoretically appealing primitive, and it has been used widely in the nonconvex optimization literature. Since Hessian-vector product oracles are feasible to compute for GANs, many recent algorithms for local min-max nonconvex optimization have also utilized Hessian-vector products [MNG17; Bal+18; ADLH19; Let+19; MJS19].

To the best of our knowledge, previous work on last-iterate convergence rates has only focused on how algorithms perform in three particular cases: (a) when the objective g is bilinear, (b) when g is strongly convex-strongly concave, and (c) when g is convex-strongly concave [Tse95; LS19; DH19; MOP19]. The existence of methods with provable finite-time guarantees for settings beyond the aforementioned has remained an open problem. This work is the first to show that an efficient algorithm, namely HGD, can achieve non-asymptotic convergence in settings that are not strongly convex or linear in either player.

4.4.1 Convergence Rates for HGD

We now state our main theorems for this chapter, which show convergence to critical points. When Assumption 4.2.3 holds, we get convergence to min-maxes. All of our main results will use the following multi-part assumption:

Assumption 4.4.1. *Let $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.*

1. *Assume a critical point for g exists.*
2. *Assume g is (L_2, L_3)-Lipschitz.*

Our first theorem shows that HGD converges for the strongly convex-strongly concave case. Although simple, this result will help us demonstrate our analysis techniques.

Theorem 4.4.2. *Let Assumption 4.4.1 hold and let $g(x, y)$ be α-strongly convex in x and α-strongly concave in y. Let $L_{\mathcal{H}}^2 = ||\xi(z_0)|| \cdot L_3 + L_2^2$. Then HGD with step-size $\eta = 1/L_{\mathcal{H}}^2$ starting from some $z_0 \in \mathbb{R}^n \times \mathbb{R}^n$ will have the following convergence rate:*

$$||\xi(z_k)|| \leq \left(1 - \frac{\alpha^2}{L_{\mathcal{H}}^2}\right)^{k/2} ||\xi(z_0)||. \tag{4.2}$$

Next, we show that HGD converges when g is linear in one of its arguments and the cross-derivative is full rank. This setting allows a slightly tighter analysis compared to Theorem 4.4.4.

Theorem 4.4.3. *Let Assumption 4.4.1 hold and let $g(x, y)$ be L-smooth in x and linear in y, and assume the cross derivative $\nabla_{xy}^2 g$ is full rank with all singular values at least $\gamma > 0$ for all $z \in \mathbb{R}^n \times \mathbb{R}^n$. Let $L_{\mathcal{H}}^2 = ||\xi(z_0)|| \cdot L_3 + L_2^2$. Then HGD with step-size $\eta = 1/L_{\mathcal{H}}^2$ starting from some $z_0 \in \mathbb{R}^n \times \mathbb{R}^n$ will have the following convergence rate:*

$$||\xi(z_k)|| \leq \left(1 - \frac{\gamma^4}{(2\gamma^2 + L^2)L_{\mathcal{H}}^2}\right)^{k/2} ||\xi(z_0)||. \tag{4.3}$$

Finally, we show our main result, which requires smoothness in both players and a large, well-conditioned cross-derivative.

Theorem 4.4.4. *Let Assumption 4.4.1 hold and let g be L-smooth in x and L-smooth in y. Let $\mu^2 = \min_{x,y} \lambda_{\min}((\nabla_{yy}^2 g(x, y))^2)$ and $\rho^2 = \min_{x,y} \lambda_{\min}((\nabla_{xx}^2 g(x, y))^2)$, and assume the cross derivative $\nabla_{xy}^2 g$ is full rank with all singular values lower bounded by $\gamma > 0$ and upper bounded by Γ for all $z \in \mathbb{R}^n \times \mathbb{R}^n$. Moreover, let the following "sufficiently bilinear" condition hold:*

$$(\gamma^2 + \rho^2)(\mu^2 + \gamma^2) - 4L^2\Gamma^2 > 0. \tag{4.4}$$

Let $L_{\mathcal{H}}^2 = ||\xi(z_0)|| \cdot L_3 + L_2^2$. Then HGD with step-size $\eta = 1/L_{\mathcal{H}}^2$ starting from some

$z_0 \in \mathbb{R}^n \times \mathbb{R}^n$ *will satisfy*

$$||\xi(z_k)|| \leq \left(1 - \frac{(\gamma^2+\rho^2)(\gamma^2+\mu^2)-4L^2\Gamma^2}{(2\gamma^2+\rho^2+\mu^2)L_{\mathcal{H}}^2}\right)^{k/2} ||\xi(z_0)||\,. \tag{4.5}$$

As discussed above, Theorem 4.4.4 provides the first last-iterate convergence rate for min-max problems that are not strongly convex or linear in either input. For example, the objective $g(x, y) = f(x) + 3Lx^\top y - h(y)$, where f and h are L-smooth convex functions, satisfies the assumptions of Theorem 4.4.4 and is not strongly convex or linear in either input. We discuss a simple example that is not convex-concave in Section 4.10. We also show how our results can be applied to specific settings, such as the Dirac-GAN, in Section 4.12.

The "sufficiently bilinear" condition (4.4) is in some sense necessary for our linear convergence rate since linear convergence is impossible in general for convex-concave settings, due to lower bounds on convex optimization [AH18; ASS17]. We give some explanations for this condition in the following section. In simple experiments for HGD on convex-concave and nonconvex-nonconcave objectives, the convergence rate speeds up when there is a larger bilinear component, as expected from our theoretical results. We show these experiments in Section 4.14.

4.4.2 Explanation of "sufficiently bilinear" condition

In this section, we explain the "sufficiently bilinear" condition (4.4). Suppose our objective is $g(x, y) = \hat{g}(x, y) + cx^\top y$ for a smooth function $\hat{g}$. Then for sufficiently large values of c (i.e. g has a large enough bilinear term), we see that g satisfies (4.4). To see this, note that if we have $\gamma^4 > 4L^2\Gamma^2$, then condition (4.4) holds. Let γ' and Γ' be lower and upper bounds on the singular values of $\nabla^2_{xy}\hat{g}$. Then it suffices to have $(\gamma' + c)^4 > 4L^2(\Gamma' + c)^2$, which is true for $c = 3\max\{L, \Gamma'\}$ (i.e. $c = O(L)$ suffices).

This condition is analogous to the case when we use GDA on the objective $g(x, y) = \hat{g}(x, y) + c\,||x||^2 - c\,||y||^2$ for L-smooth convex-concave $\hat{g}$. According to [LS19], GDA

will converge at a rate of roughly $\frac{\tilde{L}^2}{c^2}\log(1/\epsilon)$ for $\tilde{L}$-smooth and c-strongly convex-strongly concave objectives.[3] For $c = 0$, GDA will diverge in the worst case. For $c = o(L)$, we get linear convergence, but it will be slow because $\frac{L+c}{c}$ is large (this can be thought of as a large condition number). Finally, for $c = \Omega(L)$, we get fast linear convergence, since $\frac{L+c}{c} = O(1)$. Thus, to get fast linear convergence it suffices to make the problem "sufficiently strongly convex-strongly concave" (or "sufficiently strongly monotone").

Theorem 4.4.4 and condition (4.4) show that there exists another class of settings where we can achieve linear rates in the min-max setting. In our case, if we have an objective $g(x, y) = \hat{g}(x, y) + cx^\top y$ for a smooth function $\hat{g}$, we will get linear convergence if $\|\nabla^2_{xy}\hat{g}\| \leq \delta L$ and $c \geq 3(1 + \delta)L$, which ensures that the problem is "sufficiently bilinear." Intuitively, it makes sense that the "sufficiently bilinear" setting allows a linear rate because the pure bilinear setting allows a linear rate.

Another way to understand condition (4.4) is that it is a sufficient condition for the existence of a unique critical point in a general class of settings, as we show in the following lemma, which we prove in Section 4.11.

Lemma 4.4.5. *Let $g(x, y) = f(x) + cx^\top y - h(y)$ where f and h are L-smooth. Moreover, assume that $\nabla^2 f(x)$ and $\nabla^2 h(y)$ each have a 0 eigenvalue for some x and y. If (4.4) holds, then g has a unique critical point.*

4.5 Proof sketches for HGD convergence rate results

In this section, we go over the key components of the proofs for our convergence rates from Section 4.4.1. Recall that the intuition behind HGD was that critical points (where $\xi(z) = 0$) are global minima of $\mathcal{H} = \frac{1}{2}||\xi||^2$. On the other hand, there is no guarantee that $\mathcal{H}$ is a convex potential function, and a priori, one would not assume gradient descent on this potential would find a critical point. Nonetheless, we are able to show that in a variety of settings, $\mathcal{H}$ satisfies the *PL condition*, which allows HGD to have linear convergence.

[3] The actual rate is $\frac{\beta}{c}\log(1/\epsilon)$, for some parameter β that is at least $(L + c)^2$.

Proving this requires proving properties about the singular values of $J \equiv \nabla\xi$.

4.5.1 The Polyak-Łojasiewicz condition for the Hamiltonian

We begin by recalling the definition of the PL condition.

Definition 4.5.1 (Polyak-Łojasiewicz (PL) condition [Pol63; Loj63])**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the PL condition with parameter $\alpha > 0$ if for all $x \in \mathbb{R}^n$, $\frac{1}{2}||\nabla f(x)||^2 \geq \alpha(f(x) - \min_{x^* \in \mathbb{R}^n} f(x^*))$.*

The PL condition is well-known to be the weakest condition necessary to obtain linear convergence rate for gradient methods; see for example [KNS16]. We will show that $\mathcal{H}$ satisfies the PL condition, which allows us to use the following slightly modified form of a classic theorem.

Theorem 4.5.2 (Linear rate under PL [Pol63; Loj63])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the PL condition with parameter α and let $z^* \in \arg\min_{z \in \mathbb{R}^n} f(z)$. Suppose we run gradient descent from $z_0 \in \mathbb{R}^n$ with step-size $\frac{1}{L}$ and suppose that f is L-smooth at each z_k. Then we have: $f(z_k) - f(z^*) \leq (1 - \frac{\alpha}{L})^k (f(z_0) - f(z^*))$.*

Proof. Using a second-order Taylor expansion, we get:

$$\begin{aligned} f(z_{k+1}) \leq f(z_k) - \langle \nabla f(z_k), z_{k+1} - z_k \rangle \\ + \frac{L}{2} \nabla^2 f(z_k) \, ||z_{k+1} - z_k||^2 \end{aligned}$$

Using the update rule for gradient descent and using the fact that $||\nabla^2 f(z_k)|| \leq L$ gives:

$$f(z_{k+1}) \leq f(z_k) - \frac{1}{2L} ||\nabla f(z_k)||^2 \tag{4.6}$$

Subtracting $f(x^*)$ from both sides of (4.6) and applying the PL condition gives:

$$f(z_{k+1}) - f(x^*) \leq f(z_k) - f(z^*) \tag{4.7}$$

$$- \frac{\alpha}{L}(f(z_k) - f(x^*)) \tag{4.8}$$

$$= \left(1 - \frac{\alpha}{L}\right)(f(z_k) - f(z^*)) \tag{4.9}$$

Applying the last line recursively gives the result. □

To show that $\mathcal{H}$ satisfies the PL condition, we will use the following key lemma:

Lemma 4.5.3. *Let Assumption 4.4.1 hold and assume we have a twice differentiable $g(x, y)$ with associated $\xi, \mathcal{H}, J$. Let $c > 0$. If $JJ^\top \succeq \alpha I$ for every x, then $\mathcal{H}$ satisfies the PL condition with parameter α.*

Proof. Consider the squared norm of the gradient of the Hamiltonian:

$$\tfrac{1}{2}\|\nabla\mathcal{H}\|^2 = \frac{1}{2}\|J^\top\xi\|^2 = \frac{1}{2}\langle\xi, (JJ^\top)\xi\rangle \geq \frac{\alpha}{2}\,||\xi||^2 = \alpha\mathcal{H}.$$

By Assumption 4.4.1, we are guaranteed that g has a critical point. The proof is finished by noting that $\mathcal{H}(z) = 0$ when z is a critical point. □

To use Lemma 4.5.3, we will need control over the eigenvalues of $JJ^\top$, which we achieve with the following linear algebraic lemmas. We provide their proofs in Section 4.13.

Lemma 4.5.4. *Let $H = \begin{pmatrix} M_1 & B \\ -B^\top & -M_2 \end{pmatrix}$ and let $\epsilon \geq 0$. If $M_1 \succ \epsilon I$ and $M_2 \prec -\epsilon I$, then for all eigenvalues λ of $HH^\top$, we have $\lambda > \epsilon^2$.*

Lemma 4.5.5. *Let $H = \begin{pmatrix} A & C \\ -C^\top & 0 \end{pmatrix}$, where C is square and full rank. Then if λ is an eigenvalue of $HH^\top$, then we must have $\lambda \geq \frac{\sigma_{\min}^4(C)}{2\sigma_{\min}^2(C) + ||A||^2}$.*

Finally, to use Theorem 4.5.2, we will also need to show that $\mathcal{H}$ is smooth at all z_k, which holds when g is (L_2, L_3)-Lipschitz.

Lemma 4.5.6. *Consider any $g(x, y)$ which is (L_2, L_3)-Lipschitz for constants $L_2, L_3 > 0$. Suppose we run HGD initialized at some z_0 and with $\eta = 1/L_{\mathcal{H}}^2$. Then for all z_k encountered during HGD, we have that $\mathcal{H}(z_k)$ is $(||\xi(z_0)|| \cdot L_3 + L_2^2)$-smooth.*

Proof. Note that $\nabla\mathcal{H} = \xi^\top J$ and $\nabla^2\mathcal{H} = \xi^\top \nabla J + J^\top J$. Then we have:

$$\begin{aligned}\left|\left|\nabla^2\mathcal{H}\right|\right| = \left|\left|\xi^\top \nabla J + J^\top J\right|\right| &\leq ||\nabla J|| \cdot ||\xi|| + \left|\left|J^\top J\right|\right| \\ &\leq ||\xi|| \cdot L_3 + L_2^2\end{aligned}$$

Thus, it suffices to show that $||\xi(z_k)|| \leq ||\xi(z_0)||$ for all $k \geq 0$. Suppose we take a gradient descent step on $\mathcal{H}$ with parameter $\eta = 1/L_{\mathcal{H}}^2$ from some point z to some point z', and let $L_{\mathcal{H}}^2$ be such that $||\nabla^2\mathcal{H}(z)|| \leq L_{\mathcal{H}}^2$. Then by (4.6), we must have $\mathcal{H}(z') \leq \mathcal{H}(z)$. Then we always have that $\mathcal{H}(z_{k+1}) \leq \mathcal{H}(z_k)$ for $k \geq 0$, which implies that $||\xi(z_k)|| \leq ||\xi(z_0)||$ for all $k \geq 0$. This completes the proof. □

4.5.2 Proof sketches for Theorems 4.4.2, 4.4.3, and 4.4.4

We now proceed to sketch the proofs of our main theorems using the techniques we have described. The following lemma shows it suffices to prove the PL condition for $\mathcal{H}$ for the various settings of our theorems:

Lemma 4.5.7. *Given $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, suppose $\mathcal{H}$ satisfies the PL condition with parameter α^2. Suppose we use HGD starting from some $z_0 \in \mathbb{R}^n \times \mathbb{R}^n$ with step-size $\eta = 1/L_{\mathcal{H}}^2$ and suppose $\mathcal{H}$ is $L_{\mathcal{H}}^2$-smooth at all z_k visited by HGD. Then we have the following:*

$$||\xi(z_k)|| \leq \left(1 - \tfrac{\alpha^2}{L_{\mathcal{H}}^2}\right)^{k/2} ||\xi(z_0)||.$$

Proof. Since $\mathcal{H}$ satisfies the PL condition with parameter α^2 and $\mathcal{H}$ is $L_{\mathcal{H}}^2$-smooth at all z_k, we know by Theorem 4.5.2 that gradient descent on $\mathcal{H}$ with step-size $1/L_{\mathcal{H}}^2$ converges at a rate of $\mathcal{H}(z_k) \leq (1 - \frac{\alpha^2}{L_{\mathcal{H}}^2})^k \mathcal{H}(z_0)$. Substituting in for $\mathcal{H}$ gives the lemma. □

It remains to show that $\mathcal{H}$ satisfies the PL condition in the settings of Theorems 4.4.2 to 4.4.4. First, we show the result for the strongly convex-strongly concave setting of Theorem 4.4.2.

Lemma 4.5.8 (PL for the strongly convex-strongly concave setting)**.** *Let g be c-strongly convex in x and c-strongly concave in y. Then $\mathcal{H}$ satisfies the PL condition with parameter $\alpha = c^2$.*

Proof. We apply Lemma 4.5.4 with $H = J$. Since g is c-strongly-convex in x and c-strongly concave in y we have $M_1 = \nabla^2_{xx} g \succ cI$ and $M_2 = -\nabla^2_{yy} g \succ cI$. Then the magnitude of the eigenvalues of J is at least c. Thus, $JJ^\top \succeq c^2 I$, so by Lemma 4.5.3, $\mathcal{H}$ satisfies the PL condition with parameter c^2. □

Next, we show that $\mathcal{H}$ satisfies the PL condition for the nonconvex-linear setting of Theorem 4.4.3. We prove this lemma in Section 4.13.3 by using Lemma 4.5.5.

Lemma 4.5.9 (PL for the smooth nonconvex-linear setting)**.** *Let g be L-smooth in x and linear in y. Moreover, for all $z \in \mathbb{R}^n \times \mathbb{R}^n$, let $\nabla^2_{xy} g(x, y)$ be full rank and square with $\sigma_{\min}(\nabla^2_{xy} g(x, y)) \geq \gamma$. Then $\mathcal{H}$ satisfies the PL condition with parameter $\alpha = \frac{\gamma^4}{2\gamma^2 + L^2}$.*

Finally, we prove that $\mathcal{H}$ satisfies the PL condition in the nonconvex-nonconvex setting of Theorem 4.4.4. The proof for Lemma 4.5.10 is in Section 4.13.4, and it uses Lemma 4.13.2, which is similar to Lemma 4.5.5.

Lemma 4.5.10 (PL for the smooth nonconvex-nonconvex setting)**.** *Let g be L-smooth in x and L-smooth in y. Also, let $\nabla^2_{xy} g$ be full rank and let all of its singular values be lower bounded by γ and upper bounded by Γ for all $z \in \mathbb{R}^n \times \mathbb{R}^n$. Let*

$$\rho^2 = \min_{x,y} \lambda_{\min}((\nabla^2_{xx} g(x, y))^2)$$

$$\text{and } \mu^2 = \min_{x,y} \lambda_{\min}((\nabla^2_{yy} g(x, y))^2)$$

Assume the following condition holds:

$$(\gamma^2 + \rho^2)(\gamma^2 + \mu^2) - 4L^2\Gamma^2 > 0.$$

Then $\mathcal{H}$ satisfies the PL condition with parameter $\alpha = \frac{(\gamma^2+\rho^2)(\gamma^2+\mu^2)-4L^2\Gamma^2}{2\gamma^2+\rho^2+\mu^2}$.

Combining Lemmas 4.5.8 to 4.5.10 with Lemma 4.5.7 yields Theorems 4.4.2 to 4.4.4.

4.6 Extension to Stochastic HGD

Our results above also imply rates for stochastic HGD, where the gradient $\nabla\mathcal{H}$ in (4.1), is replaced by a stochastic estimator v of $\nabla\mathcal{H}$ such that $\mathbb{E}[v] = \nabla\mathcal{H}$. Since we show that $\mathcal{H}$ satisfies the PL condition with parameter α in different settings, we can use Theorem 4 in [KNS16] to show that stochastic HGD converges at a $O(1/\sqrt{k})$ rate in the settings of Theorems 4.4.2 to 4.4.4, including the "sufficiently bilinear" setting. However, we need to explicitly assume that $\mathcal{H}$ is $L_{\mathcal{H}}^2$-smooth at each iterate of the algorithm. While this assumption may seem strong, it will be satisfied as long as the iterates of the algorithm remain in some bounded region.

Theorem 4.6.1. *Let $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Assume a critical point for g exists and suppose $\mathcal{H}$ satisfies the PL condition with parameter α^2. Suppose we use the update $z_{k+1} = z_k - \eta_k v(z_k)$, where v is a stochastic estimate of $\nabla\mathcal{H}$ such that $\mathrm{E}[v] = \nabla\mathcal{H}$ and $\mathrm{E}[\|v(z_k)\|^2] \leq C^2$ for all z_k. Moreover, assume that $\mathcal{H}$ is $L_{\mathcal{H}}$ smooth at all z_k. Then if we use $\eta_k = \frac{2k+1}{2\alpha^2(k+1)^2}$, we have the following convergence rate: $\mathrm{E}[\|\xi(z_k)\|] \leq \sqrt{\frac{L_{\mathcal{H}}^2 C^2}{k\alpha^4}}$.*

To prove Theorem 4.6.1, we need the following theorem from [KNS16].[4]

Theorem 4.6.2 ([KNS16])**.** *Assume that f has a non-empty solution set and satisfies the PL condition with parameter α. Let f^* be the minimum value of f. Let v be a stochastic estimate of ∇f such that $\mathbb{E}[v] = \nabla f$. Assume $E[\|v(z_k)\|^2] \leq C^2$ for all z_k and some C.*

[4]The actual theorem in [KNS16] is stated in a slightly different way, but it is equivalent to our presentation.

Suppose we use the SGD update $z_{k+1} = z_k - \eta_k v(z_k)$ with $\eta_k = \frac{2k+1}{2\alpha(k+1)^2}$, and suppose f is L-smooth at each z_k. Then we get a convergence rate of

$$\mathbb{E}[f(z_k) - f^*] \leq \frac{LC^2}{2k\alpha^2} \tag{4.10}$$

If instead we use a constant $\eta_k = \eta < \frac{1}{2\alpha}$, then we obtain a linear convergence rate up to a solution level that is proportional to η,

$$\mathbb{E}[f(z_k) - f^*] \leq (1 - 2\alpha\eta)^k [f(z_0) - f^*] + \frac{LC^2\eta}{4\alpha}$$

We now show how to use Theorem 4.6.2 to prove Theorem 4.6.1.

Proof of Theorem 4.6.1. If $\mathcal{H}$ satisfies the PL condition with parameter α^2, then we can apply Theorem 4.6.2 to the stochastic variant of HGD. since $\mathcal{H}^* = 0$, we get

$$\mathrm{E}\left[\frac{1}{2}\|\xi(z_k)\|^2\right] \leq \frac{L_{\mathcal{H}}^2 C^2}{2k\alpha^4} \tag{4.11}$$

The theorem follows from Jensen's inequality, which implies that

$$\mathbb{E}\left[\|\xi(z_k)\|\right] \leq \sqrt{\mathrm{E}\left[\|\xi(z_k)\|^2\right]}.$$

□

4.7 Extension to Consensus Optimization

The Consensus Optimization (CO) algorithm of [MNG17] is as follows:

$$z_{k+1} = z_k - \eta(\xi(z_k) + \gamma\nabla\mathcal{H}(z_k)) \tag{4.12}$$

where $\gamma > 0$. This is essentially a weighted combination of GDA and HGD. [MNG17] remark that while HGD has poor performance on nonconvex problems in practice, CO can effectively train GANs in a variety of settings, including on CIFAR-10 and celebA. While they frame CO as GDA with a small modification, they actually set $\gamma = 10$ for several of their experiments, which suggests that one can also view CO as a modified form of HGD.

Using this perspective, we prove Theorem 4.7.1, which implies that we get linear convergence of CO in the same settings as Theorems 4.4.2 to 4.4.4 provided that γ is sufficiently large (i.e. the HGD update is large compared to the GDA update). Previously, [LS19] proved that CO achieves linear convergence in the bilinear setting, so our result greatly expands the settings where CO has provable non-asymptotic convergence.

Theorem 4.7.1. *Let Assumption 4.4.1 hold. Let g be L_g smooth and suppose $\mathcal{H}$ satisfies the PL condition with parameter α^2. Let $L_{\mathcal{H}}^2 = ||\xi(z_0)|| \cdot L_3 + L_2^2$. Then if we update some $z_0 \in \mathbb{R}^n \times \mathbb{R}^n$ using the CO update (4.12) with step-size $\eta = \frac{\alpha^2}{4L_{\mathcal{H}}^2 L_g}$ and $\gamma = \frac{4L_g}{\alpha^2}$, we get the following convergence:*

$$||\xi(z_k)|| \leq \left(1 - \tfrac{\alpha^2}{4L_{\mathcal{H}}^2}\right)^k ||\xi(z_0)||. \tag{4.13}$$

We also show that CO converges in practice on some simple examples in Section 4.14.

The key technical component to proving Theorem 4.7.1 is showing that HGD still performs well even with small arbitrary perturbations, as we show in the following lemma:

Lemma 4.7.2. *Let $z_{k+1} = z_k - \eta \nabla \mathcal{H}(z_k) + \eta_v v^{(k)}$ where $v^{(k)}$ is some arbitrary vector such that $\left|\left|v^{(k)}\right|\right| = ||\xi(z_k)||$. Let g be L_g-smooth and suppose $\mathcal{H}$ satisfies the PL condition with parameter α. Let $\eta = \frac{1}{L_{\mathcal{H}}^2}$ and let $\eta_v = \frac{\alpha^2}{4L_{\mathcal{H}}^2 L_g}$. Then we get the following convergence:*

$$||\xi(z_k)|| \leq \left(1 - \tfrac{\alpha^2}{4L_{\mathcal{H}}^2}\right)^k ||\xi(z_0)||. \tag{4.14}$$

From Lemma 4.7.2, it is simple to prove Theorem 4.7.1.

Proof of Theorem 4.7.1. Note that the CO update (4.12) with $\gamma = \frac{4L_g}{\alpha^2}$ is exactly the update in Lemma 4.7.2 with $v^{(k)} = -\xi(z_k)$, so we get the desired convergence rate. □

Our result treats GDA as an adversarial perturbation even though this is not the case, which suggests that this analysis may be improved. It would be nice if one could directly apply the PL-based analysis that we used for HGD, but this does not seem to work for CO because CO is not an instance of gradient descent on some proxy objective.

Finally, we prove Lemma 4.7.2.

Proof of Lemma 4.7.2. Let $z_{k+1/2} = z_k - \eta \nabla \mathcal{H}(z_k)$, so $z_{k+1} = z_{k+1/2} + \eta_v v^{(k)}$. From (4.9) in the proof of Theorem 4.5.2 with $\eta = \frac{1}{L_{\mathcal{H}}^2}$, we get

$$\left|\left|\xi(z_{k+1/2})\right|\right| \leq \left(1 - \frac{\alpha^2}{L_{\mathcal{H}}^2}\right)^{1/2} ||\xi(z_k)|| \leq (1 - \frac{\alpha^2}{2L_{\mathcal{H}}^2}) \left|\left|\xi(z_k)\right|\right|. \tag{4.15}$$

Next, note that the triangle inequality and smoothness of g imply:

$$\begin{aligned} ||\xi(z_{k+1})|| &\leq \left|\left|\xi(z_{k+1/2})\right|\right| + \left|\left|\xi(z_{k+1}) - \xi(z_{k+1/2})\right|\right| \\ &\leq \left|\left|\xi(z_{k+1/2})\right|\right| x + L_g \left|\left|z_{k+1} - z_{k+1/2}\right|\right| \\ &= \left|\left|\xi(z_{k+1/2})\right|\right| + L_g \left|\left|\eta_v v\right|\right| \end{aligned}$$

Using the above result and $\left|\left|v^{(k)}\right|\right| = ||\xi(z_k)||$, we get:

$$||\xi(z_{k+1})|| \leq \left(1 - \frac{\alpha^2}{2L_{\mathcal{H}}^2} + L_g \eta_v\right) ||\xi(z_k)|| \tag{4.16}$$

Setting $\eta_v = \frac{\alpha^2}{4L_{\mathcal{H}}^2 L_g}$ gives the result. □

Note that for this result, we assume g is L_g smooth in x and y jointly, whereas in other parts of the paper we assume g is smooth in x or y separately. If g is L-smooth in x and L-smooth in y and $\left|\left|\nabla_{xy}^2 g(x,y)\right|\right| \leq L_c$ for all x, y, then g will be $L + L_c$ smooth.

4.8 Nonconvex extensions for HGD

While the naive version of HGD will try to converge to all critical points, we can modify HGD slightly to achieve second-order stability guarantees as in various related work such as [Bal+18; Let+19]. In particular, we consider modifying HGD so that there is some scalar α in front of the $\nabla\mathcal{H}$ term as follows:

$$z_{k+1} = z_k - \eta\alpha\nabla\mathcal{H}(z_k) \tag{4.17}$$

We now present two ways to choose α. Our first method is inspired by the Simplectic Gradient Adjustment algorithm of [Bal+18], which is as follows:

$$z_{k+1} = z_k - \eta(\xi(z_k) - \lambda A^\top \xi(z_k)) \tag{4.18}$$

where A is the antisymmetric part of J and $\lambda = \text{sgn}\left(\langle \xi, J\rangle \left\langle A^\top \xi, J\right\rangle\right)$. [Bal+18] show that λ is positive when in a strictly convex-strictly concave region and negative in a strictly concave-strictly convex region. Thus, if we choose $\alpha = \lambda = \text{sgn}\left(\langle \xi, J\rangle \left\langle A^\top \xi, J\right\rangle\right)$, we can ensure that the modified HGD will exhibit local stability around strict min-maxes and local instability around strict max-mins. This follows simply because we will do gradient *descent* on $\mathcal{H}$ in the first case and gradient *ascent* on $\mathcal{H}$ in the second case.

Another way to choose α involves using an approximate eigenvalue computation on $\nabla^2_{xx}g$ and $\nabla^2_{yy}g$ to detect whether $\nabla^2_{xx}g$ is positive semidefinite and $\nabla^2_{yy}g$ is negative semidefinite (which would mean we are in a convex-concave region). We set $\alpha = 1$ if we are in a convex-concave region and -1 otherwise, which will guarantee local stability around min-maxes and local instability around other critical points. This approximate eigenvector computation can be done using a logarithmic number of Hessian-vector products.

4.9 Comparison of Theorem 4.4.4 to [DH19]

In this section, we compare our results in Theorem 4.4.4 to those of [DH19]. [DH19] prove a rate for GDA when g is L-smooth and convex in x and L-smooth and μ-strongly concave in y and $\nabla^2_{xy} g$ is some fixed matrix A. The specific setting they consider is to find the unconstrained min-max for a function $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ defined as $g(x, y) = f(x) + y^\top Ax - h(y)$ where f is convex and smooth, h is strongly-convex and smooth, and $A \in \mathbb{R}^{d_2 \times d_1}$ has rank d_1 (i.e. A has full column rank).

Their rate uses the potential function $P_t = \lambda a_t + b_t$, where we have:

$$\lambda = \frac{2L\Gamma(L + \frac{\Gamma^2}{\mu})}{\mu\gamma^2} \tag{4.19}$$

$$a_k = ||z_k - x^*|| \tag{4.20}$$

$$b_k = \left|\left|y^{(k)} - y^*\right|\right| \tag{4.21}$$

where (x^*, y^*) is the min-max for the objective. Their rate (Theorem 3.1 in [DH19]) is

$$P_{k+1} \leq \left(1 - c\frac{\mu^2\gamma^4}{L^3\Gamma^2(L + \frac{\Gamma^2}{\mu})}\right)^k P_k \tag{4.22}$$

for some constant $c > 0$. To translate this rate into bounds on $||\xi||$, we can use the smoothness of g in both of its arguments to note that $\left|\left|\frac{\partial g}{\partial x}(x, y)\right|\right| = \left|\left|\frac{\partial g}{\partial x}(x, y) - \frac{\partial g}{\partial x}(x_1^*, x_2^*)\right|\right| \leq L\,||z_k - x^*||$ and likewise for y. So the rate on P_k translates into a rate on $||\xi||$ with some additional factor in front.

Their rate and our rate are incomparable – neither is strictly better. For instance when $\gamma = \Gamma$ is much larger than all other quantities, their rates simplify to $\left(1 - O\left(\frac{\mu^3}{L^3}\right)\right)^k$, while ours go to $\left(1 - O\left(\frac{\gamma^2}{L_{\mathcal{H}}^2}\right)\right)^{k/2}$. While our convergence rate requires the sufficiently bilinear condition (4.4) to hold, we do not require convexity in x or concavity in y. Moreover, we allow $\nabla^2_{xy} g$ to change as long as the bounds on the singular values hold whereas [DH19]

require $\nabla^2_{xy} g$ to be a fixed matrix.

4.10 Nonconvex-nonconcave setting where Assumption 4.2.3 and the conditions for Theorem 4.4.4 hold

In this section we give a concrete example of a nonconvex-nonconcave setting where Assumption 4.2.3 and the conditions for Theorem 4.4.4 hold. We choose this example for simplicity, but one can easily come up with other more complicated examples.

For our example, we define the following function:

$$F(x) = \begin{cases} -3(x + \frac{\pi}{2}) & \text{for } x \leq -\frac{\pi}{2} \\ -3\cos x & \text{for } -\frac{\pi}{2} < x \leq \frac{\pi}{2} \\ -\cos x + 2x - \pi & \text{for } x > \frac{\pi}{2} \end{cases} \tag{4.23}$$

The first and second derivatives of F are as follows:

$$F'(x) = \begin{cases} -3 & \text{for } x \leq -\frac{\pi}{2} \\ 3\sin x & \text{for } -\frac{\pi}{2} < x \leq \frac{\pi}{2} \\ \sin x + 2 & \text{for } x > \frac{\pi}{2} \end{cases} \tag{4.24}$$

$$F''(x) = \begin{cases} 0 & \text{for } x \leq -\frac{\pi}{2} \\ 3\cos x & \text{for } -\frac{\pi}{2} < x \leq \frac{\pi}{2} \\ \cos x & \text{for } x > \frac{\pi}{2} \end{cases} \tag{4.25}$$

From Figure 4.2, we can see that this function is neither convex nor concave. We note that although this function is not thrice differentiable, which is technically required to prove smoothness of $\mathcal{H}$ in our result, we can instead show this smoothness for the iterates of the

algorithm by showing it is true for all points on the line between z_k and z_{k+1} for $k \geq 0$.

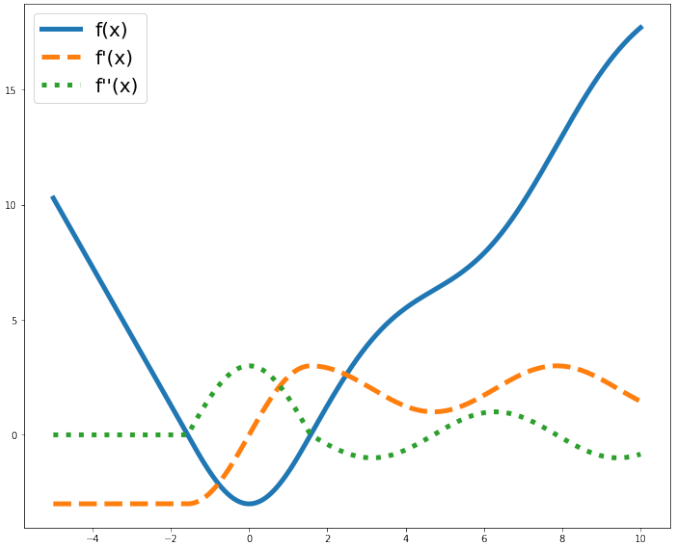


Figure 4.2: Plot of nonconvex function $F(x)$ defined in (4.23), as well as its first and second derivatives

Our objective will be $g(x, g_2) = F(x) + 4x^\top y - F(y)$. Note that $L = 3$ because $F''(x) \leq 3$ for all x. Also, $\gamma = \Gamma = 4$ since $\nabla^2_{xy} g = 4I$.

First, we show that g satisfies Assumption 4.4.1. We see that g has a critical point at $(0, 0)$. Moreover, g is (L_2, L_3)-Lipschitz for any finite-sized region of $\mathbb{R}^2$. Thus, if we assume our algorithm stays within a ball of some radius R, the (L_2, L_3)-Lipschitz assumption will be satisfied. Since our algorithm does not diverge and indeed converges at a linear rate to the min-max, this assumption is fairly mild.

Next, we show that g satisfies condition (4.4). Condition (4.4) requires $\gamma^4 > 4L^2\Gamma^2$ for g. We see that this holds because $\gamma^4 = 4^4 = 256$ and $4L\Gamma^2 = 4 * 3 * 4^2 = 192$.

Therefore, the assumptions of Theorem 4.4.4 are satisfied.

We can also show that this objective satisfies Assumption 4.2.3, so we get convergence to the min-max of g. We will show that g has only one critical point (at $(0, 0)$) and that this critical point is a min-max. We first give a "proof by picture" below, showing a plot of g in Figure 4.3, along with plots of $g(\cdot, 0)$ and $g(0, \cdot)$ showing that $(0, 0)$ is indeed a min-max.

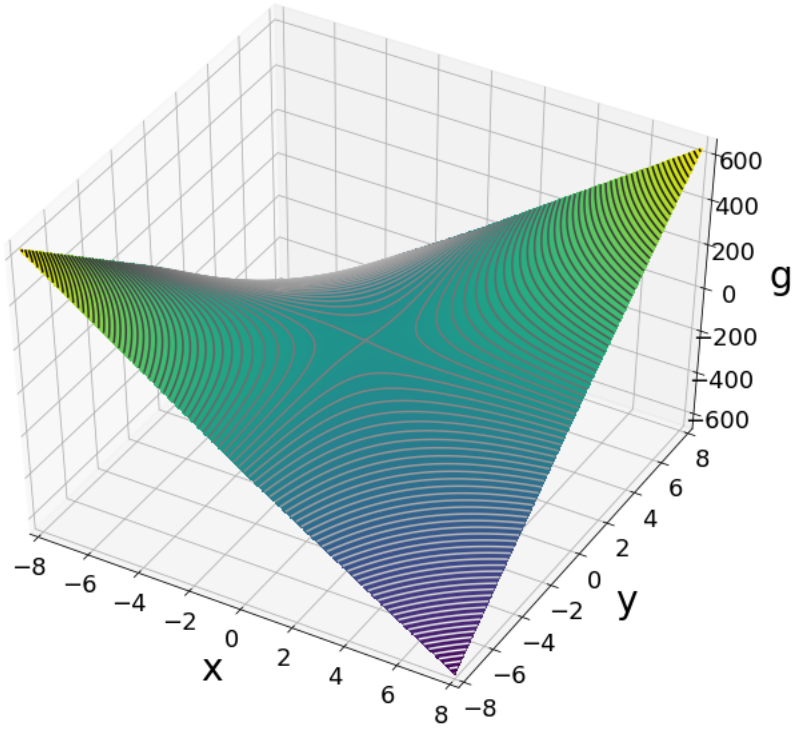


Figure 4.3: Plot of nonconvex-nonconcave $g(x, y) = F(x) + 4x^\top y - F(y)$

We can also formally show that $(0, 0)$ is the unique critical point of g and that it is a min-max. We prove this for completeness, although the calculations more or less amount to a simple case analysis. Let us look at the derivatives of g with respect to x and y:

$$\frac{\partial g}{\partial x}(x, y) = \begin{cases} -3 + 4y & \text{for } x \leq -\frac{\pi}{2} \\ 3\sin x + 4y & \text{for } -\frac{\pi}{2} < x \leq \frac{\pi}{2} \\ \sin x + 2 + 4y & \text{for } x > \frac{\pi}{2} \end{cases} \tag{4.26}$$

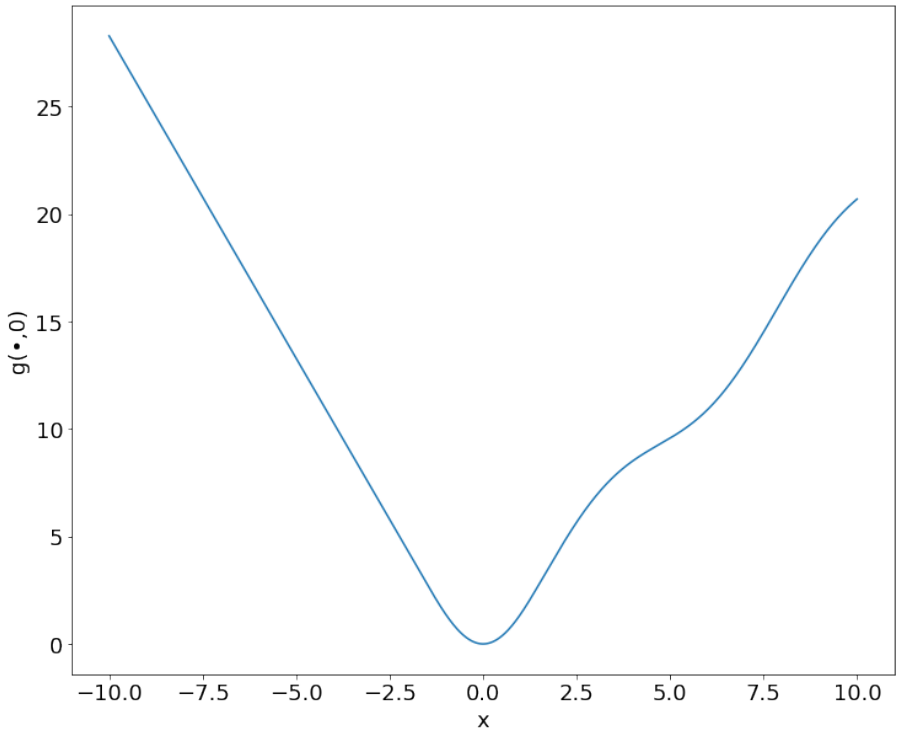


Figure 4.4: Plot of $g(\cdot, 0)$. We can see that there is only one min and it occurs at $x = 0$.

$$\frac{\partial g}{\partial y}(x,y) = \begin{cases} 3 + 4x & \text{for } y \leq -\frac{\pi}{2} \\ -3\sin y + 4x & \text{for } -\frac{\pi}{2} < y \leq \frac{\pi}{2} \\ -\sin y + 2 + 4x & \text{for } y > \frac{\pi}{2} \end{cases} \tag{4.27}$$

Observe that if $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ then critical points of g must satisfy $3\sin x + 4y = 0$, which implies that $y \in [-\frac{3}{4}, \frac{3}{4}]$. Likewise, if $y \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, then critical points of g must have $x \in [-\frac{3}{4}, \frac{3}{4}]$. We show that this implies that g only has critical points where x and y are both in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

Suppose g had a critical point such that $x \leq -\frac{\pi}{2}$. Then this critical point must satisfy $y = \frac{3}{4}$. But from our observation above, if a critical point has $y = \frac{3}{4}$, then x must lie in $[-\frac{3}{4}, \frac{3}{4}]$, which contradicts $x \leq -\frac{\pi}{2}$.

Next, suppose g had a critical point such that $x > \frac{\pi}{2}$. Then this critical point must satisfy $y = -\frac{1}{4}(\sin x + 2)$, which implies that $y \in [-\frac{3}{4}, \frac{3}{4}]$. But then by the observation above, x

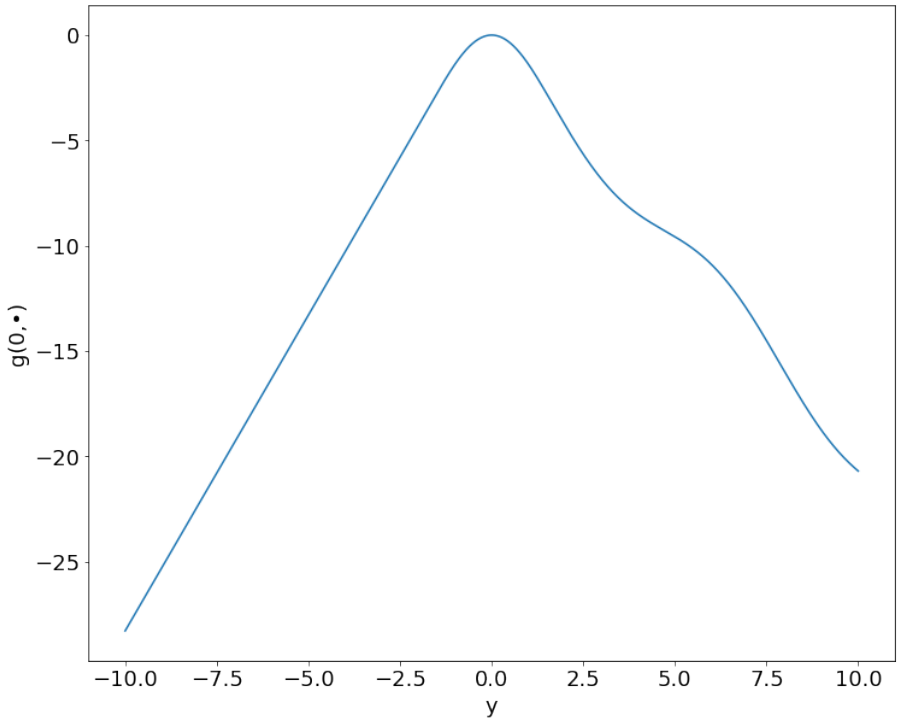


Figure 4.5: Plot of $g(0, y)$. We can see that there is only one max and it occurs at $y = 0$.

must lie in $[-\frac{3}{4}, \frac{3}{4}]$, which contradicts $x > \frac{\pi}{2}$.

From this we see that any critical point of g must have $x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. We can make analogous arguments to show that any critical point of g must have $y \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

From this, we can conclude that all critical points of g must satisfy the following:

$$3 \sin x + 4y = 0 \tag{4.28}$$

$$-3 \sin y + 4x = 0 \tag{4.29}$$

These equations imply the following:

$$x = \frac{3}{4}\sin y \tag{4.30}$$

$$y = -\frac{3}{4}\sin x \tag{4.31}$$

$$\Rightarrow x = \frac{3}{4}\sin\left(-\frac{3}{4}\sin y\right) \tag{4.32}$$

$$\Rightarrow y = -\frac{3}{4}\sin\left(\frac{3}{4}\sin y\right) \tag{4.33}$$

That is, for all critical points of g, x must be a fixed point of $h_1(x) = \frac{3}{4}\sin\left(-\frac{3}{4}\sin x\right)$ and y must be a fixed point of $h_2(x) = -\frac{3}{4}\sin\left(\frac{3}{4}\sin x\right)$. Since $|h_1'(x)| < 1$ and $|h_2'(x)| < 1$ always, h_1 and h_2 are contractive maps, so they have only one fixed point each. Thus, g will only have one critical point, namely the point (x, y) such that x is the unique fixed point of h_1 and y is the unique fixed point of h_2.

Finally, we can observe that $(0, 0)$ is a critical point of g, so it must be the unique critical point of g. One can also see that this is a min-max by looking at the second derivatives of F in (4.25).

4.11 Proof of Lemma 4.4.5

To prove Lemma 4.4.5, we will use the following lemma:

Lemma 4.11.1. *Let $g(x, y) = f(x) + cx^\top y - h(y)$ where f and h are L-smooth. Then if $c > L$, g has a unique critical point.*

Proof of Lemma 4.4.5. Condition (4.4) is as follows:

$$(\gamma^2 + \rho^2)(\mu^2 + \gamma^2) - 4L^2\Gamma^2 > 0. \tag{4.34}$$

Note that in our setting, $\gamma = \Gamma = c$. Next, observe that if $\nabla^2 f(x)$ and $\nabla^2 h(y)$ each have a 0

eigenvalue for some x and y, condition (4.4) reduces to:

$$c > 2L. \tag{4.35}$$

Then by Lemma 4.11.1, we see that g must have a unique critical point. □

Next, we prove Lemma 4.11.1.

Proof of Lemma 4.11.1. Suppose our objective is $g(x, y) = f(x) + cx^\top y - h(y)$ where f and h are both L-smooth convex functions. Critical points of g must satisfy the following:

$$\nabla f(x) + cy = 0 \tag{4.36}$$

$$-\nabla h(y) + cx = 0 \tag{4.37}$$

$$\Rightarrow x = \frac{1}{c}\nabla h(y) \tag{4.38}$$

$$\Rightarrow y = -\frac{1}{c}\nabla f\left(\frac{1}{c}\nabla h(y)\right) \tag{4.39}$$

In other words, y must be a fixed point of $F(z) = -\frac{1}{c}\nabla f(\frac{1}{c}\nabla h(z))$. The function F will have a unique fixed point if it is a contractive map. We now show that for $c > L$, this is the case.

$$||F(u) - F(v)|| = \left|\left|\frac{1}{c}\nabla f\left(\frac{1}{c}\nabla h(u)\right) - \frac{1}{c}\nabla f\left(\frac{1}{c}\nabla h(v)\right)\right|\right| \tag{4.40}$$

$$\leq \frac{L}{c} \cdot \left|\left|\frac{1}{c}\nabla h(u) - \frac{1}{c}\nabla h(v)\right|\right| \tag{4.41}$$

$$\leq \frac{L^2}{c^2}\,||u - v|| < ||u - v|| \tag{4.42}$$

where the inequalities follow from smoothness of f and h. An analogous property can be shown by solving for x instead. Thus, if $c > L$, then g will have a unique fixed point.

Condition (4.4) is thus a sufficient condition for the existence of a unique critical point for the class of objectives above. □

4.12 Applications

In this section, we discuss how our results can be applied to various settings. One simple setting is the Dirac-GAN from [MGN18], where $g(x, y) = \min_x \max_y f(x^\top y) - f(0)$ for some function f whose derivative is always non-zero. When $f(t) = t$, the Dirac-GAN is just a bilinear game, so HGD will converge globally to the Nash Equilibrium (NE) of this Dirac-GAN, as shown in [Bal+18]. Our results prove global convergence rates for HGD on the Dirac-GAN even when a small smooth convex regularizer is added for the discriminator or subtracted for the generator. Moreover, Lemma 2.2 of [MGN18] shows that the diagonal blocks of the Jacobian are 0 at the NE for arbitrary f with non-zero derivative. As such, HGD will achieve the convergence rates in this chapter in a region around the NE for the Dirac-GAN for arbitrary f with non-zero derivative even when a small smooth convex regularizer is added for either player.

[DH19] list several applications where the min-max formulation is relevant, such as in ERM problems with a linear classifier. Given a data matrix A, the ERM problem involves solving $\min_x \ell(Ax) + f(x)$ for some smooth, convex loss ℓ and smooth, convex regularizer f. This problem has the saddle point formulation $\min_x \max_y y^\top Ax - \ell^*(y) + f(x)$. According to [DH19], this formulation can be advantageous when it allows a finite-sum structure, reduces communication complexity in a distributed setting, or allows some sparsity structure to be exploited. Our results show that linear rates are possible for this problem if A is square, well-conditioned, and sufficiently large compared to ℓ and f.

4.13 Proofs for Section 4.5

In this section, we prove our main results about the convergence of HGD, starting with some key technical lemmas.

4.13.1 Proof of Lemma 4.5.4

Proof. Note that $HH^\top = \begin{pmatrix} M_1^2 + BB^T & -M_1B - BM_2 \\ -(M_1B + BM_2)^T & M_2^2 + B^TB \end{pmatrix} = \begin{pmatrix} M_1 & -B \\ -B^T & M_2 \end{pmatrix}^2$.

Now let $Z = \begin{pmatrix} M_1 & -B \\ -B^T & M_2 \end{pmatrix}$. It suffices to show that for any eigenvalue δ of Z, $|\delta| \leq \epsilon$. For the sake of contradiction, let v be an eigenvalue of Z with eigenvalue δ such that $|\delta| \leq \epsilon$. Let $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$. Since $Zv = \delta v$ for $|\delta| \leq \epsilon$ and $M_1 \succ \epsilon I$ and $M_2 \prec -\epsilon I$, we must have $v_1 \neq 0$ and $v_2 \neq 0$. Then we have:

$$\begin{pmatrix} M_1v_1 - Bv_2 \\ M_2v_2 - B^\top v_1 \end{pmatrix} = \delta \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{4.43}$$

This implies

$$(M_1 - \delta I)v_1 = Bv_2 \tag{4.44}$$

$$(M_2 - \delta I)v_2 = B^\top v_1 \tag{4.45}$$

Let $\hat{M}_1 = M_1 - \delta I$ and let $\hat{M}_2 = M_2 - \delta I$. Note that $\hat{M}_1 \succ 0$ and $\hat{M}_2 \prec 0$. Then we can write $v_1 = \hat{M}_1^{-1}Bv_2$. Further, we can substitute into (4.45) to get

$$\hat{M}_2v_2 = B^\top \hat{M}_1^{-1}Bv_2 \tag{4.46}$$

$$\iff -\hat{M}_2^{-1}B^\top \hat{M}_1^{-1}Bv_2 = -v_2 \tag{4.47}$$

In other words, v_2 is an eigenvector of $-\hat{M}_2^{-1}B^\top \hat{M}_1^{-1}B$ with eigenvalue -1. Let $A = -\hat{M}_2^{-1}$ and $T = B^\top \hat{M}_1^{-1}B$. Note that A is positive definite and T is PSD. Then we have:

$$AT = A^{1/2}(A^{1/2}TA^{1/2})A^{-1/2} \tag{4.48}$$

Since $A^{1/2}TA^{1/2}$ is PSD, and AT is similar to $A^{1/2}TA^{1/2}$, we must have that all of the eigenvalues of AT are nonnegative. This contradicts that v_2 is an eigenvector of AT with eigenvalue -1.

Thus, all eigenvalues of Z must have magnitude greater than ϵ. □

4.13.2 Proof of Lemma 4.5.5

Proof. Suppose λ is an eigenvalue of $HH^\top$ with eigenvector $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$. WLOG, suppose $\lambda < \sigma^2_{\min}(C)$. Since v is an eigenvector, we have:

$$\begin{pmatrix} A^2 + CC^\top & -AC \\ -C^\top A & C^\top C \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{4.49}$$

Thus, we have:

$$(A^2 + CC^\top - \lambda I)v_1 - ACv_2 = 0 \tag{4.50}$$

$$-C^\top Av_1 + (C^\top C - \lambda I)v_2 = 0 \tag{4.51}$$

Since $\lambda < \sigma^2_{\min}(C)$, we have that $C^\top C - \lambda I$ is invertible, so we can write $v_2 = (C^\top C - \lambda I)^{-1}C^\top Av_1$ from the (4.51). Plugging this into (4.50) gives:

$$(A^2 + CC^\top - \lambda I - AC(C^\top C - \lambda I)^{-1}C^\top A)v_1 = 0 \tag{4.52}$$

$$(A(I - C(C^\top C - \lambda I)^{-1}C^\top)A + CC^\top - \lambda I)v_1 = 0 \tag{4.53}$$

Write the SVD of C as $C = U\Sigma V^\top$. Then we have:

$$C(C^\top C - \lambda I)^{-1}C^\top = U\Sigma V^\top (V\Sigma U^\top U\Sigma V^\top - \lambda I)^{-1} V\Sigma U^\top \tag{4.54}$$

$$= U\Sigma V^\top (V(\Sigma^2 - \lambda I)V^\top)^{-1} V\Sigma U^\top \tag{4.55}$$

$$= U\Sigma V^\top V^{-T}(\Sigma^2 - \lambda I)^{-1} V^{-1} V\Sigma U^\top \tag{4.56}$$

$$= U\Sigma^2(\Sigma^2 - \lambda I)^{-1} U^\top \tag{4.57}$$

$$= UDU^\top \tag{4.58}$$

where the second line follows because $VV^\top = I$ when C is full rank and where D is a diagonal matrix such that $D_{ii} = \frac{\sigma_i^2(C)}{\sigma_i^2(C) - \lambda}$.

Let $M = I - D$, so M is diagonal with $M_{ii} = \frac{-\lambda}{\sigma_i^2(C) - \lambda}$. Then (4.53) becomes:

$$(AMA + CC^\top - \lambda I)v_1 = 0 \tag{4.59}$$

This means $T = AMA + CC^\top - \lambda I$ has a 0 eigenvalue. A simple lower bound for the eigenvalues of T is

$$\lambda_{\min}(T) \geq -||A||^2 \frac{\lambda}{\sigma_{\min}^2 - \lambda} + \sigma_{\min}^2(C) - \lambda \tag{4.60}$$

We will show that if $\lambda < \delta$, where $\delta = \sigma_{\min}^2(C) + \frac{||A||^2}{2} - \sqrt{(\sigma_{\min}^2 + \frac{||A||^2}{2})^2 - \sigma_{\min}^4}$, then $\lambda_{\min}(T) > 0$, which is a contradiction. It suffices to show the following inequality:

$$-||A||^2 \frac{\lambda}{\sigma_{\min}^2 - \lambda} + \sigma_{\min}^2(C) - \lambda > 0 \tag{4.61}$$

$$\iff \sigma_{\min}^2(C) - \lambda > ||A||^2 \frac{\lambda}{\sigma_{\min}^2 - \lambda} \tag{4.62}$$

$$\iff (\sigma_{\min}^2(C) - \lambda)^2 > ||A||^2 \lambda \tag{4.63}$$

$$\iff \lambda^2 - (2\sigma_{\min}^2(C) + ||A||^2)\lambda + \sigma_{\min}^4(C) > 0 \tag{4.64}$$

(4.64) has zeros at the following values:

$$\sigma_{\min}^2(C) + \frac{||A||^2}{2} \pm \sqrt{\left(\sigma_{\min}^2 + \frac{||A||^2}{2}\right)^2 - \sigma_{\min}^4(C)} \tag{4.65}$$

Since (4.64) is a convex parabola, if λ is less than both zeros, we will have proved (4.64). This is clearly true if $\lambda < \delta$.

As a last step, we can give a slightly nicer form of δ, using Lemma 4.13.1. Letting $x = \sigma_{\min}^2(C) + \frac{||A||^2}{2}$ and $c = \sigma_{\min}^4(C)$, we have $\delta > \frac{\sigma_{\min}^4(C)}{2\sigma_{\min}^2(C)+||A||^2}$. So to reiterate, if $\lambda < \frac{\sigma_{\min}^4(C)}{2\sigma_{\min}^2(C)+||A||^2} < \delta$, then (4.64) holds, so $T \succ 0$, which contradicts (4.59). □

Lemma 4.13.1. *For $x \in (0,1)$ and $c \in (0, x^2)$, we have:*

$$x - \sqrt{x^2 - c} > \frac{c}{2x}$$

Proof.

$$x - \sqrt{x^2 - c} = x - x\sqrt{1 - \frac{c}{x^2}} > x - x\left(1 - \frac{c}{2x^2}\right) = \frac{c}{2x}$$

□

4.13.3 Proof of Lemma 4.5.9

Proof. Let $C(x,y) = \nabla_{xy}^2 g(x,y)$. For all $z \in \mathbb{R}^n \times \mathbb{R}^n$, $C(x,y)$ is square and full rank by assumption, so we can apply Lemma 4.5.5 with $H = J$ at each point $z \in \mathbb{R}^n \times \mathbb{R}^n$, which gives $\lambda(JJ^\top) \geq \frac{\sigma_{\min}^4(C(x,y))}{2\sigma_{\min}^2(C(x,y))+||\nabla_{xx}^2 g(x,y)||^2}$. We have $||\nabla_{xx}^2 g(x,y)|| \leq L$ since g is smooth in x. Also, $\sigma_{\min}^2(C(x,y)) \geq \gamma$. Then we have that $JJ^\top \succeq \frac{\gamma^4}{2\gamma^2+L^2} I$, so by Lemma 4.5.3, $\mathcal{H}$ satisfies the PL condition with parameter $\frac{\gamma^4}{2\gamma^2+L^2}$. □

4.13.4 Proof of Lemma 4.5.10

To prove Lemma 4.5.10, we use the following lemma:

Lemma 4.13.2. *Let* $H = \begin{pmatrix} A & C \\ -C^\top & -B \end{pmatrix}$*, where* C *is square and full rank. Moreover, let* $c = (\sigma_{\min}^2(C) + \lambda_{\min}(A^2))(\lambda_{\min}(B^2) + \sigma_{\min}^2(C)) - \sigma_{\max}^2(C)(||A|| + ||B||)^2$ *and assume* $c > 0$*. Then if* λ *is an eigenvalue of* $HH^\top = \begin{pmatrix} A^2 + CC^\top & -AC - CB \\ -C^\top A - BC^\top & B^2 + C^\top C \end{pmatrix}$*, we must have*

$$\lambda \geq \frac{(\sigma_{\min}^2(C) + \lambda_{\min}(A^2))(\lambda_{\min}(B^2) + \sigma_{\min}^2(C)) - \sigma_{\max}^2(C)(||A|| + ||B||)^2}{(2\sigma_{\min}^2(C) + \lambda_{\min}(A^2) + \lambda_{\min}(B^2))^2}.$$

Proof of Lemma 4.13.2. This proof resembles that of Lemma 4.5.5. Let $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ be an eigenvector of $HH^\top$ with eigenvalue λ. Expanding $HH^\top v = \lambda v$, we have:

$$(A^2 + CC^\top - \lambda I)v_1 - (AC + CB)v_2 = 0 \tag{4.66}$$

$$-(C^\top A + BC^\top)v_1 + \underbrace{(B^2 + C^\top C - \lambda I)}_{M} v_2 = 0 \tag{4.67}$$

$$\Rightarrow v_2 = M^{-1}(C^\top A + BC^\top)v_1 \tag{4.68}$$

$$\Rightarrow (-(AC + CB)M^{-1}(C^\top A + BC^\top) + A^2 + CC^\top - \lambda I)v_1 = 0 \tag{4.69}$$

where M is invertible because $C^\top C$ is positive definite and WLOG, we may assume that $\lambda < \lambda_{\min}(C^\top C) = \sigma_{\min}^2(C)$. We will show that if the assumptions in the statement of the lemma hold, then we get a contradiction if λ is below some positive threshold. In particular, we show that the following inequality holds for small enough λ (this inequality contradicts

(4.69)):

$$\begin{aligned}
&\sigma_{\min}^2(C) - \lambda + \lambda_{\min}(A^2) > \sigma_{\max}^2(C)(||A|| + ||B||)^2 \left|\left|M^{-1}\right|\right| \\
\Leftarrow\ &\sigma_{\min}^2(C) - \lambda + \lambda_{\min}(A^2) > \frac{\sigma_{\max}^2(C)}{\lambda_{\min}(B^2) + \sigma_{\min}^2(C) - \lambda}(||A|| + ||B||)^2 \\
\iff\ &\lambda^2 - (2\sigma_{\min}^2(C) + \lambda_{\min}(A^2) + \lambda_{\min}(B^2))\lambda + \\
&(\sigma_{\min}^2(C) + \lambda_{\min}(A^2))(\lambda_{\min}(B^2) + \sigma_{\min}^2(C)) - \sigma_{\max}^2(C)(||A|| + ||B||)^2 > 0
\end{aligned}$$

Letting $b = 2\sigma_{\min}^2(C) + \lambda_{\min}(A^2) + \lambda_{\min}(B^2)$, we can solve for the zeros of the above equation:

$$\lambda = \frac{b \pm \sqrt{b^2 - 4c}}{2} \tag{4.70}$$

Note that we have $c > 0$ by assumption, so this equation has only positive roots. Note also that $b^2 > 4c$, so the roots will not be imaginary. Then we see that if $\lambda < \delta = \frac{b - \sqrt{b^2 - 4c}}{2}$, we get a contradiction. Using Lemma 4.13.1, we see that $\delta > \frac{c}{b}$. So we've proven that $\lambda < \frac{c}{b}$ gives a contradiction, so we must have $\lambda \geq \frac{c}{b}$, i.e.

$$\lambda \geq \frac{(\sigma_{\min}^2(C) + \lambda_{\min}(A^2))(\lambda_{\min}(B^2) + \sigma_{\min}^2(C)) - \sigma_{\max}^2(C)(||A|| + ||B||)^2}{2\sigma_{\min}^2(C) + \lambda_{\min}(A^2) + \lambda_{\min}(B^2)}.$$

□

Proof of Lemma 4.5.10. The proof is very similar to that of Lemma 4.5.9. Let $C(x, y) = \nabla_{xy}^2 g(x, y)$. For all $z \in \mathbb{R}^n \times \mathbb{R}^n$, $C(x, y)$ is square and full rank with bounds on its singular values by assumption. Moreover, (4.4) holds, so we can apply Lemma 4.13.2 with $H = J$ at each point $z \in \mathbb{R}^n \times \mathbb{R}^n$. Using the fact that g is smooth in x and y, this gives

$$\lambda(JJ^\top) \geq \frac{(\sigma_{\min}^2(C(x, y)) + \lambda_{\min}(A^2))(\sigma_{\min}^2(C(x, y)) + \mu^2) - 4L^2\sigma_{\max}^2(C(x, y))}{2\sigma_{\min}^2(C(x, y)) + \lambda_{\min}(A^2) + \mu^2}.$$

Using the bounds on the singular values of $C(x, y)$, we have that

$$JJ^\top \succeq \frac{(\gamma^2 + \lambda_{\min}(A^2))(\gamma^2 + \mu^2) - 4L^2\Gamma^2}{2\gamma^2 + \lambda_{\min}(A^2) + \mu^2} I,$$

so by Lemma 4.5.3, $\mathcal{H}$ satisfies the PL condition with parameter $\frac{(\gamma^2+\lambda_{\min}(A^2))(\gamma^2+\mu^2)-4L^2\Gamma^2}{2\gamma^2+\lambda_{\min}(A^2)+\mu^2}$.

□

4.14 Experiments

In this section, we present some experimental results showing how GDA, HGD, and CO perform on a convex-concave objective and a nonconvex-nonconcave objective. For our CO plots, γ refers to the γ parameter in the CO algorithm. All of our experiments are initialized at $(5, 5)$. The step-size η for HGD and GDA is always 0.01, while the step-size η for CO with $\gamma = \{0.1, 1, 10\}$ is $\{0.1, 0.01, 0.001\}$ respectively to account for the fact that increasing γ increases the effective step-size, so the η parameter needs to be decreased accordingly. The experiments were all run on a standard 2017 Macbook Pro.

The main takeaways from the experiments are that CO with low γ will not converge if there is a large bilinear term, while CO with high γ and HGD all converge for small and large bilinear terms. When the bilinear term is large, CO with high γ and HGD both will converge in fewer iterations (for the same step-size). We did not optimize for step-size, so it is possible this effect may change if the optimal step-size is chosen for each setting.

4.14.1 Convex-concave objective

The convex-concave objective we use is $g(x, y) = f(x) + cxy - f(y)$ where $f(x) = \log(1 + e^x)$. We show a plot of f in Figure 4.6.

When $c = 3$, GDA converges, and when $c = 10$, GDA diverges. We note that HGD and CO (for large enough γ) tend to converge faster when c is larger.

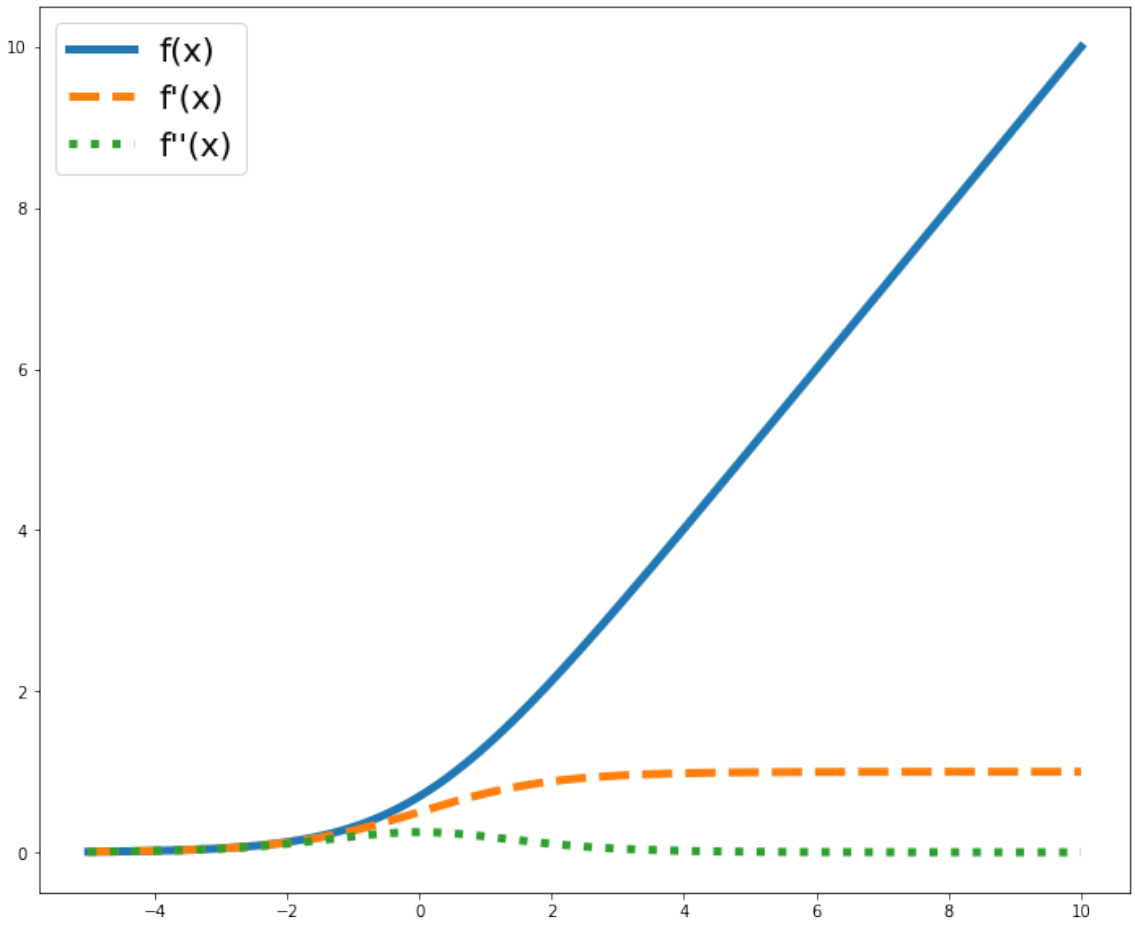
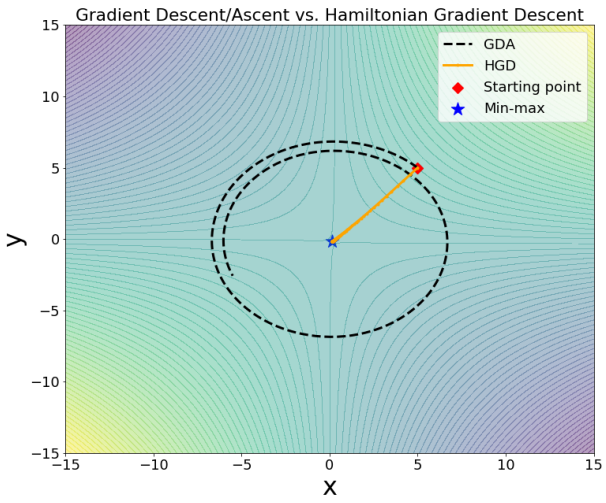


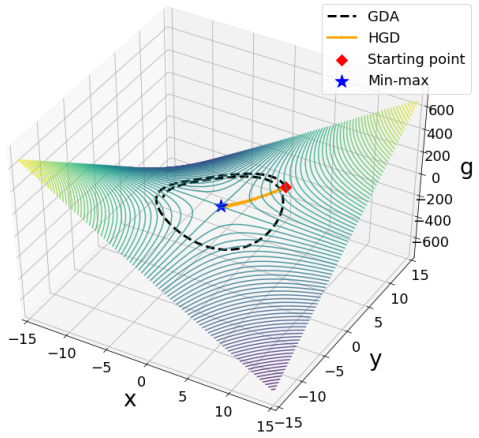
Figure 4.6: Plot of $f(x) = \log(1 + e^x)$ with its first and second derivatives. This is a convex, smooth function

GDA converges ($c = 3$)

These plots show g when $c = 3$, so GDA converges, as does CO with $\gamma = 0.1$.

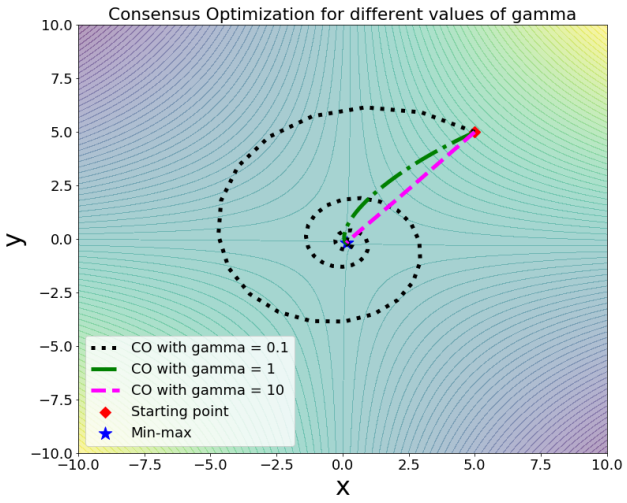


(a)

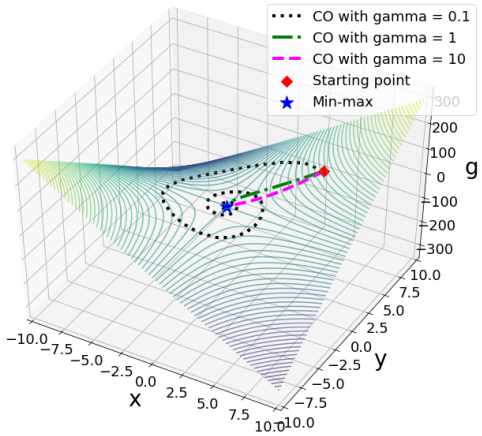


(b)

Figure 4.7: GDA vs. HGD for 300 iterations for $g(x, y) = f(x) + cxy - f(y)$ where $f(x) = \log(1 + e^x)$ and $c = 3$. GDA slowly circles towards the min-max, and HGD goes directly to the min-max.

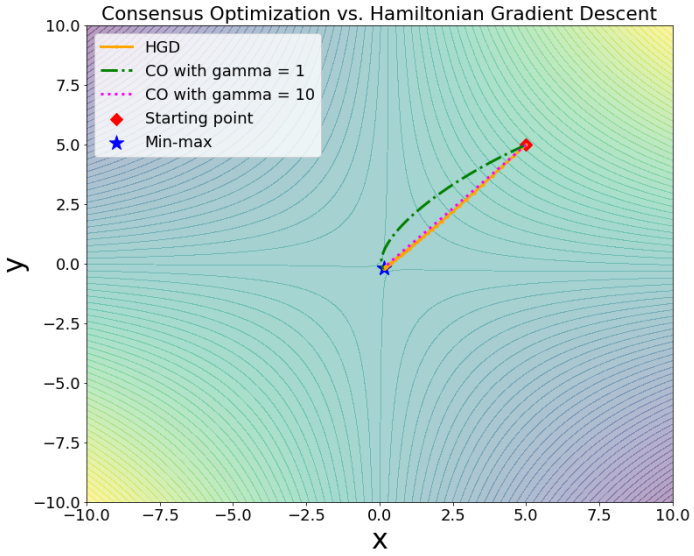


(a)

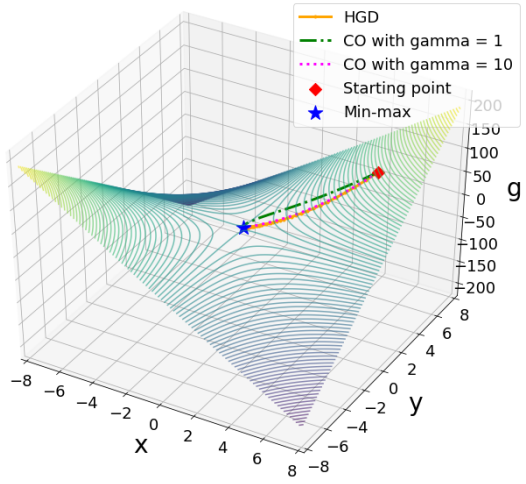


(b)

Figure 4.8: CO for 100 iterations with different values of γ for $g(x, y) = f(x) + cxy - f(y)$ where $f(x) = \log(1+e^x)$ and $c = 3$. The $\gamma = 0.1$ curve slowly circles towards the min-max, while the other curves go directly to the min-max.



(a)

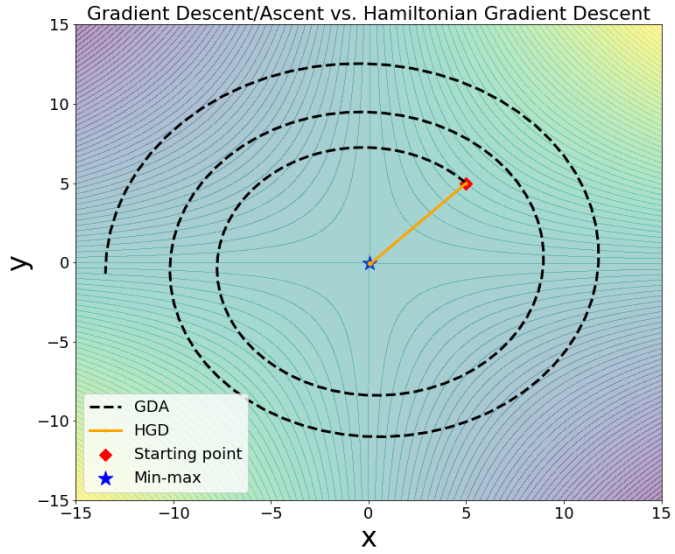


(b)

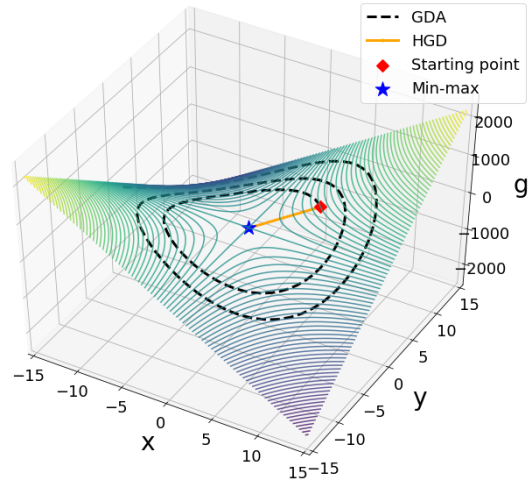
Figure 4.9: HGD vs. CO for 100 iterations for $g(x, y) = f(x) + cxy - f(y)$ where $f(x) = \log(1 + e^x)$ and $c = 3$ with different values of γ.

GDA diverges ($c = 10$)

These plots show g when $c = 10$, so GDA diverges, as does CO with $\gamma = 0.1$. Note that in this case, CO with $\gamma \geq 1$ and HGD both require very few iterations (typically about 2) to reach the min-max.

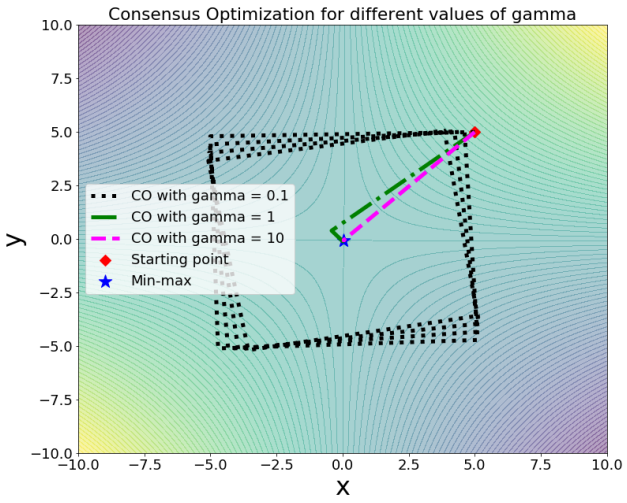


(a)

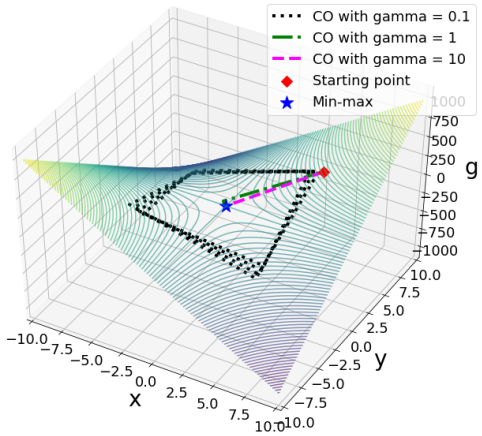


(b)

Figure 4.10: GDA vs. HGD for 150 iterations for $g(x, y) = f(x) + cxy - f(y)$ where $f(x) = \log(1 + e^x)$ and $c = 10$. GDA slowly circles away from the min-max, while HGD goes directly to the min-max.

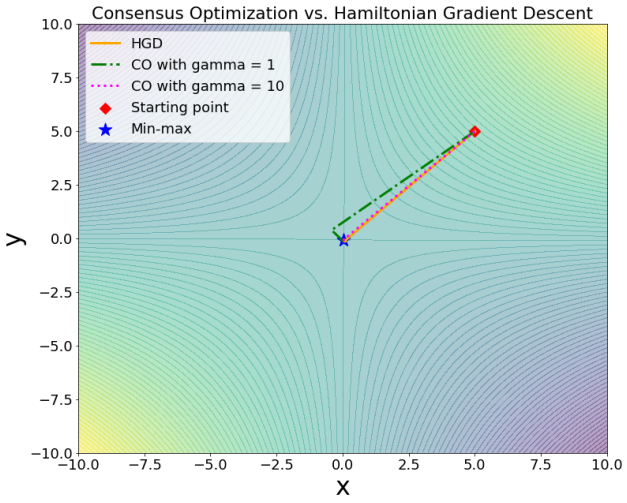


(a)

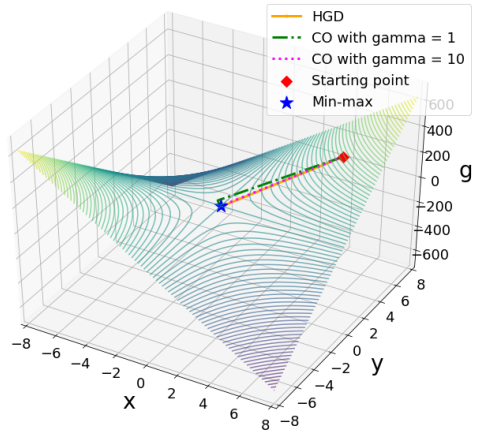


(b)

Figure 4.11: CO for 15 iterations with different values of γ for $g(x, y) = f(x) + cxy - f(y)$ where $f(x) = \log(1 + e^x)$ and $c = 10$. The $\gamma = 0.1$ curve makes a cyclic pattern around the min-max, while the other curves go directly to the min-max.



(a)



(b)

Figure 4.12: HGD vs. CO for 15 iterations with different values of γ for $g(x, y) = f(x) + cxy - f(y)$ where $f(x) = \log(1 + e^x)$ and $c = 10$.

4.14.2 Nonconvex-nonconcave objective

The nonconvex-nonconcave objective we use is $g(x, y) = F(x) + cxy - F(y)$ where F is defined as in (4.23) in Section 4.10.

$$F(x) = \begin{cases} -3(x + \frac{\pi}{2}) & \text{for } x \leq -\frac{\pi}{2} \\ -3\cos x & \text{for } -\frac{\pi}{2} < x \leq \frac{\pi}{2} \\ -\cos x + 2x - \pi & \text{for } x > \frac{\pi}{2} \end{cases} \tag{4.71}$$

We show a plot of F in Figure 4.13.

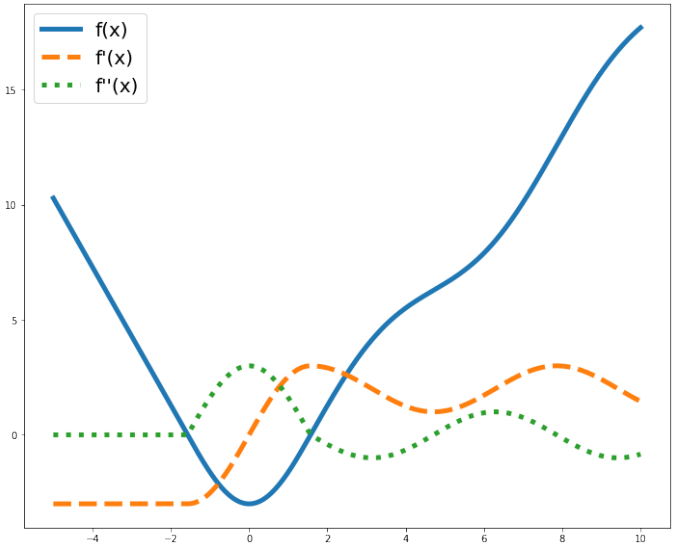
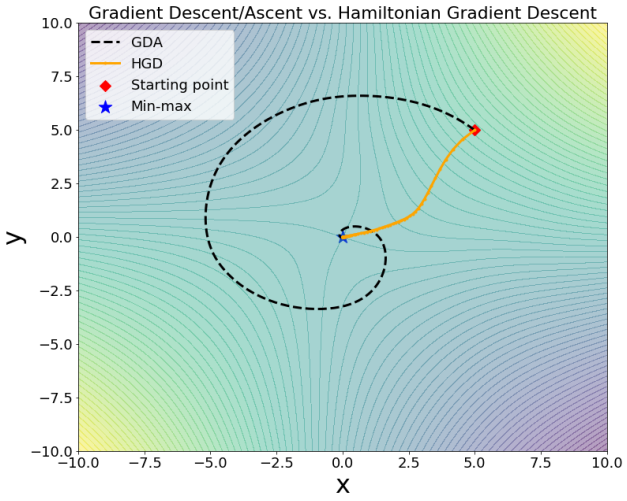


Figure 4.13: Plot of nonconvex function $F(x)$ defined in (4.23), as well as its first and second derivatives

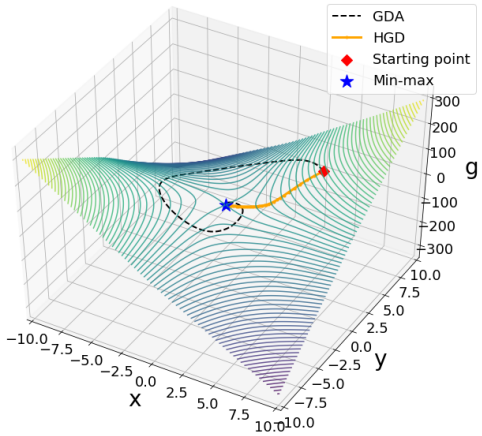
As in the convex-concave case, when $c = 3$, GDA converges, and when $c = 10$, GDA diverges. Again, HGD and CO (for large enough γ) tend to converge faster when c is larger.

GDA converges ($c = 3$)

These plots show g when $c = 3$, so GDA converges, as does CO with $\gamma = 0.1$.

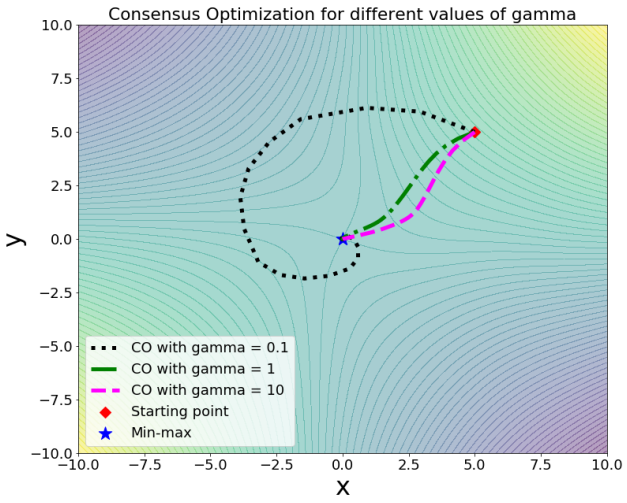


(a)

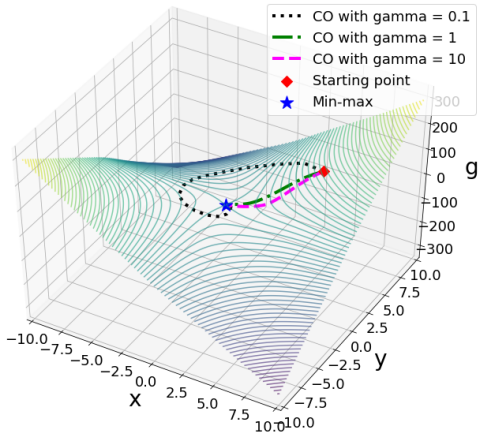


(b)

Figure 4.14: GDA vs. HGD for 300 iterations for $g(x, y) = F(x) + cxy - F(y)$ where $F(x)$ is defined in (4.71) and $c = 3$. GDA slowly circles towards the min-max, and HGD goes more directly to the min-max.

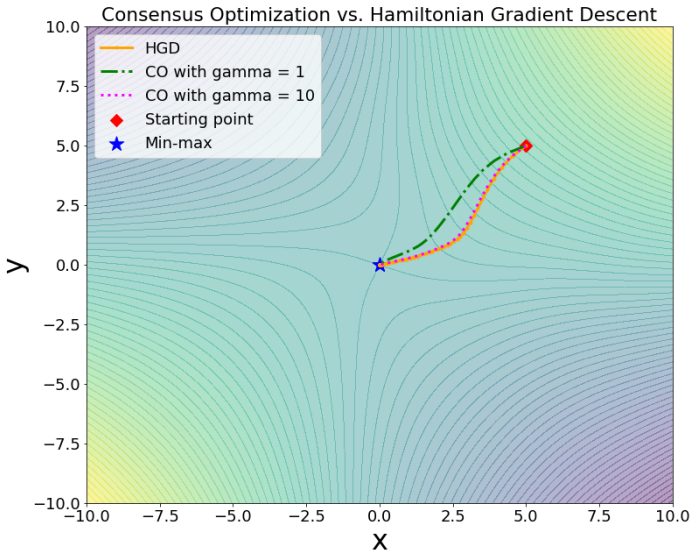


(a)

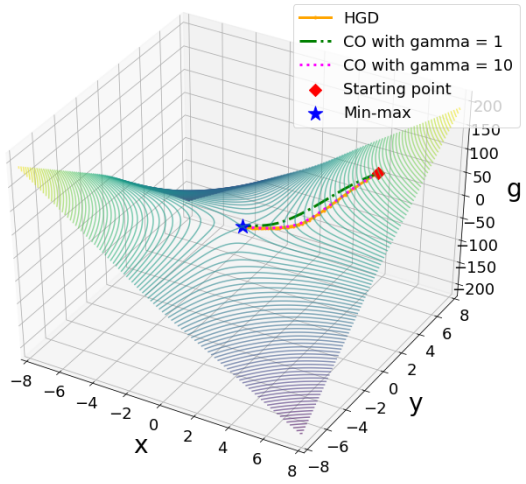


(b)

Figure 4.15: CO for 100 iterations with different values of γ for $g(x, y) = F(x)+cxy-F(y)$ where $F(x)$ is defined in (4.71) and $c = 3$. The $\gamma = 0.1$ curve slowly circles towards the min-max, while the other curves go more directly to the min-max.



(a)

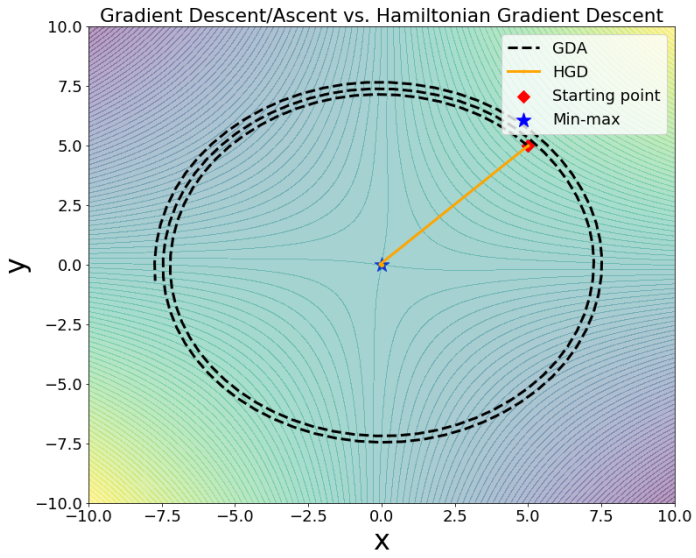


(b)

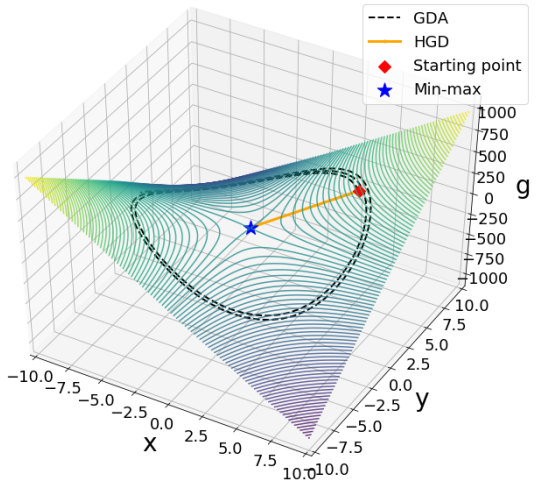
Figure 4.16: HGD vs. CO for 100 iterations for $g(x, y) = F(x) + cxy - F(y)$ where $F(x)$ is defined in (4.71) and $c = 3$ with different values of γ.

GDA diverges ($c = 10$)

These plots show g when $c = 10$, so GDA diverges, as does CO with $\gamma = 0.1$. Note that in this case, CO with $\gamma \geq 1$ and HGD both require very few iterations (typically about 2) to reach the min-max.

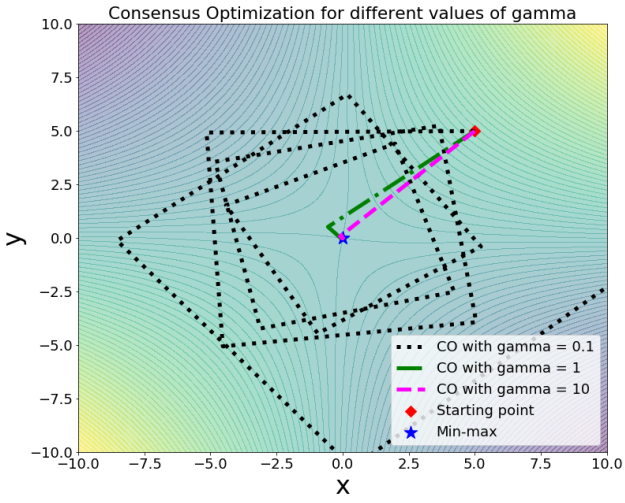


(a)

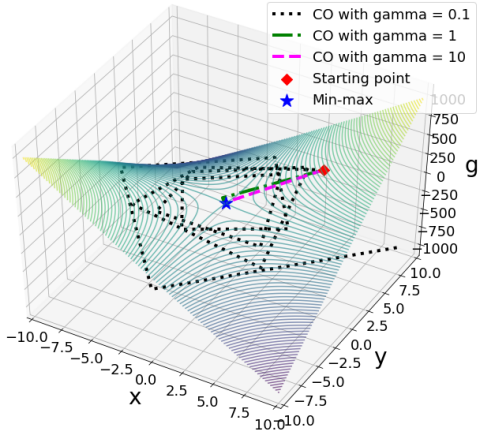


(b)

Figure 4.17: GDA vs. HGD for 150 iterations for $g(x, y) = F(x) + cxy - F(y)$ where $F(x)$ is defined in (4.71) and $c = 10$. GDA slowly circles away from the min-max, while HGD goes directly to the min-max.

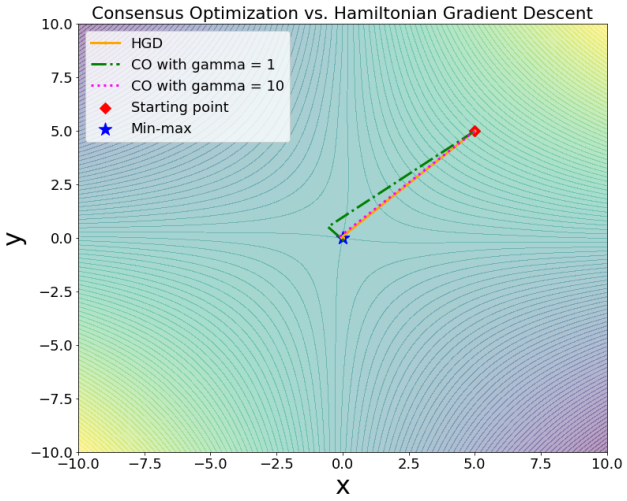


(a)

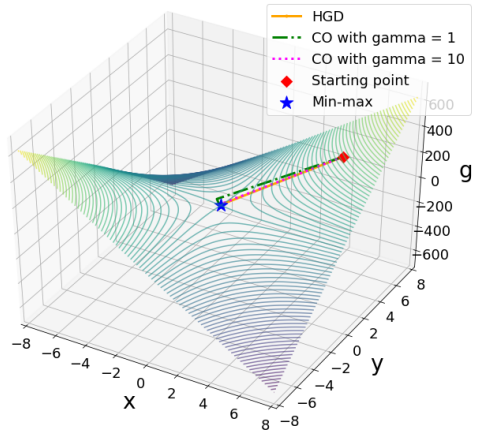


(b)

Figure 4.18: CO for 15 iterations with different values of γ for $g(x, y) = F(x) + cxy - F(y)$ where $F(x)$ is defined in (4.71) and $c = 10$. The $\gamma = 0.1$ curve makes an erratic cycle around the min-max, slowly diverging, while the other curves go directly to the min-max.



(a)



(b)

Figure 4.19: HGD vs. CO for 15 iterations with different values of γ for $g(x, y) = F(x) + cxy - F(y)$ where $F(x)$ is defined in (4.71) and $c = 10$.

4.14.3 Effect of bilinear term on HGD convergence in nonconvex-nonconvex objective

In this section, we look at the convergence of HGD for the same objective as discussed in the previous section, namely $g(x, y) = F(x) + cxy - F(y)$ where F is defined as in (4.23)

in Section 4.10.

$$F(x) = \begin{cases} -3(x + \frac{\pi}{2}) & \text{for } x \leq -\frac{\pi}{2} \\ -3\cos x & \text{for } -\frac{\pi}{2} < x \leq \frac{\pi}{2} \\ -\cos x + 2x - \pi & \text{for } x > \frac{\pi}{2} \end{cases} \tag{4.72}$$

In this case, we will vary c to show that HGD converges faster for higher c and will not converge for sufficiently low c.

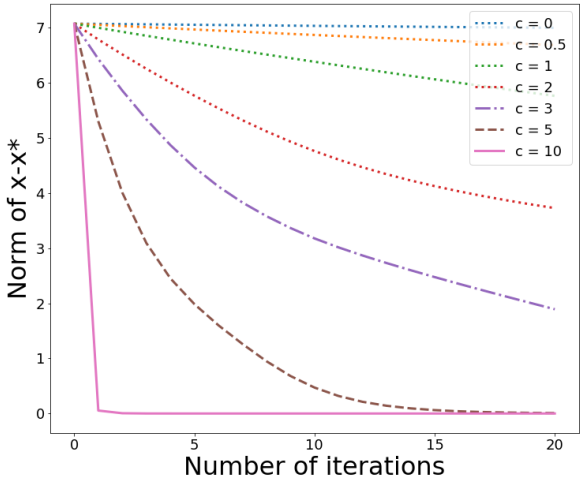


Figure 4.20: Distance to minmax for HGD iterates for different values of c in the objective $g(x, y) = F(x) + cxy - F(y)$ where $F(x)$ is defined in (4.71).

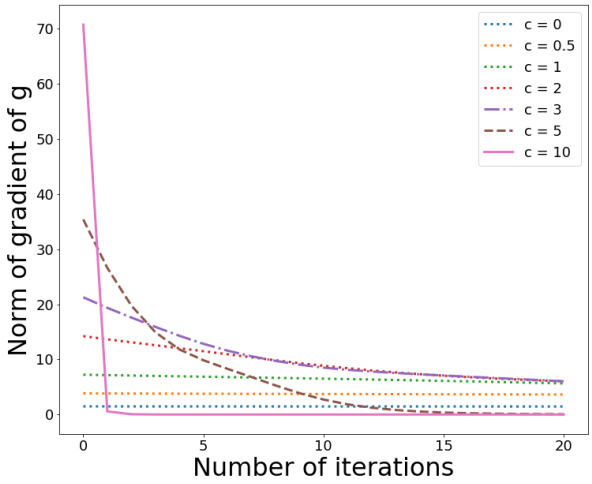


Figure 4.21: Gradient norm for HGD iterates for different values of c in the objective $g(x, y) = F(x) + cxy - F(y)$ where $F(x)$ is defined in (4.71). Since all runs are initialized at $(5, 5)$, when c is increased, the initial gradient norm also increases. Nonetheless, HGD still converges faster for the cases with higher c.

CHAPTER 5

HIGHER-ORDER METHODS FOR CONVEX-CONCAVE MIN-MAX OPTIMIZATION AND MONOTONE VARIATIONAL INEQUALITIES

In this chapter, we consider solving convex-concave min-max problems as well as a more general class of problems known as monotone variational inequalities. We will show an algorithm called HIGHERORDERMIRRORPROX that achieves an iteration complexity of $O(1/k^{\frac{p+1}{2}})$ when given access to an oracle for minimizing a p^{th} order Taylor expansion and when the p^{th}-order derivatives are Lipschitz continuous. We also give analogous rates for the weak monotone variational inequality problem. For $p > 2$, our results improve on the iteration complexity of the first-order Mirror Prox method of [Nem04] and the second-order method of [MS12]. We further instantiate our entire algorithm in the unconstrained $p = 2$ case.

5.1 Introduction

Monotone variational inequalities (MVIs) are a well-studied class of problems that are very related to convex-concave min-max problems [Min+62; KS80; Nem04]. In an MVI, we are given a *monotone* operator $F : \mathcal{Z} \to \mathbb{R}^n$ over a convex set $\mathcal{Z} \subseteq \mathbb{R}^n$, and the goal is to find a point $z^* \in \mathcal{Z}$ such that

$$\forall z \in \mathcal{Z}, \langle F(z), z^* - z \rangle \leq 0. \tag{5.1}$$

Such a point is called a solution to a weak (Minty) MVI [Kom99].

The Mirror Prox (MP) algorithm of [Nem04] is a popular method for solving both (5.1) (when F is Lipschitz continuous) and (1.1) (when g is smooth). MP is a generalization of the extragradient algorithm of [Kor76], and it converges in $O(1/k)$ iterations, which is tight

for *first-order methods* (FOMs) [NY83]. Given that MP achieves the optimal performance for FOMs, there is a natural question of whether one can improve on the iteration complexity by using *higher-order methods* (HOMs), which tend to converge in fewer iterations but at the expense of higher cost per iteration. HOMs use higher-order derivatives of the objective function and generally require *higher-order smoothness*, namely that the higher derivatives of the objective be Lipschitz continuous.

In vanilla optimization, while FOMs such as gradient descent are the gold standard for optimization algorithms, HOMs are useful in a variety of different settings. Newton's method is one of the most well-known HOMs, and it is a central component of path-following interior-point methods [NN94]. In cases when the higher-order update is efficiently computable, HOMs can achieve faster overall running times than FOMs. For example, HOMs have been used to find approximate local minima in nonconvex optimization faster than gradient descent [Aga+17; CDHS18]. While second-order methods are the most common type of HOM, there has also been significant recent work on HOMs beyond second-order methods [AH18; ASS18; Gas+18; JWZ18; Bub+18; Bul18].

HOMs have seen much less study in the context of MVIs and min-max problems. [MS12] use a second-order method with an *implicit* update that achieves improved iteration complexity of $O(1/k^{\frac{3}{2}})$ for problems with second-order smoothness. Their method uses the Hybrid Proximal Extragradient (HPE) framework established in [MS10] and requires access to an oracle for a second-order constrained optimization problem. However, it was unknown whether one could achieve further improved iteration complexity in the presence of third-order smoothness and beyond.

In this chapter, our main contribution is a higher-order method HIGHERORDERMIRRORPROX for approximately solving MVIs and convex-concave min-max problems that achieves an iteration complexity of $O(1/k^{\frac{p+1}{2}})$ for problems with p^{th}-order smoothness. To our knowledge, this is the first result showing that improved convergence rates are possible for problems with third-order smoothness and beyond. Our algorithm requires access to an

oracle for minimizing a p^{th}-order Taylor expansion and uses a higher-order implicit update that can be thought of as a generalization of Mirror Prox. Since the implicit update may be difficult to compute in the constrained case, we show how to instantiate our algorithm in the second-order unconstrained case, giving overall running time bounds in that setting.

We begin by reviewing definitions, notions of convergence, and related work in Section 5.2. Then we summarize our main results and our algorithm in Section 5.3. In Section 5.4, we present the proof of our main result. We then show how to fully instantiate our algorithm in the unconstrained $p = 2$ case in Section 5.5.

5.2 Preliminaries

We will use MVI($F, \mathcal{Z}$) to denote the MVI given in (5.1) over a vector field $F : \mathcal{Z} \to \mathbb{R}^n$ and convex constraint set $\mathcal{Z} \subseteq \mathbb{R}^n$. Unless otherwise specified, we will use z^* to signify a solution to MVI($F, \mathcal{Z}$). Throughout the chapter, we will use γ_k to represent positive weights, and we let $\Gamma_K \stackrel{\text{def}}{=} \sum_{k=1}^{K} \gamma_k$.

For notational convenience, we assume our algorithms have access to a monotone operator F. This is the usual assumption in MVIs, but it will also allow us to solve min-max problems, as we now show. For min-max problems (1.1), recall that we defined the gradient descent-ascent field of g:

$$\xi(x, y) \stackrel{\text{def}}{=} \begin{pmatrix} \nabla_x g(x, y) \\ -\nabla_y g(x, y) \end{pmatrix} \tag{5.2}$$

Letting $z = \begin{pmatrix} x \\ y \end{pmatrix}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, we can say ξ maps $\mathcal{Z}$ to $\mathbb{R}^n$ with only a slight abuse of notation. It is then easy to show that ξ is monotone when g is convex-concave. So to apply our algorithms to min-max settings, we simply apply them on ξ.

Our algorithms will require the following assumption:

Definition 5.2.1. *A vector field $F : \mathcal{Z} \to \mathbb{R}^n$ is p^{th}-order L_p smooth w.r.t. $||\cdot||$ if, for all $u, v \in \mathcal{Z}$,*

$$\left|\left|\nabla^{p-1}F(u) - \nabla^{p-1}F(v)\right|\right|_* \leq L_p \left|\left|u - v\right|\right|,$$

where we define

$$\left|\left|\nabla^{p-1}F(u) - \nabla^{p-1}F(v)\right|\right|_* \stackrel{def}{=} \max_{h:||h||\leq 1} \left|\nabla^{p-1}F(u)[h]^{p-1} - \nabla^{p-1}F(v)[h]^{p-1}\right|.$$

Remark 5.2.2. *Our definition of p^{th}-order smoothness as a property of the $(p-1)^{th}$ derivative of F is motivated by the min-max setting (1.1), where ξ is already expressed in terms of the gradient of g. If ξ is p^{th} order smooth, this is a statement about the Lipschitz continuity of p^{th} order derivatives of g.*

Another key component of our algorithms is the *p^{th}-order Taylor expansion* of F at u evaluated at v:

$$\mathcal{T}_p(v; u) = \sum_{i=0}^{p} \nabla^{(i)}F(u)[v-u]^i \tag{5.3}$$

While $\mathcal{T}$ depends on F, we leave this implicit to lighten notation, as the relevant F will always be obvious from context.

Remark 5.2.3. *To be consistent with Remark 5.2.2, when we refer to "p^{th}-order methods," we will be referring to methods that use a $(p-1)^{th}$-order Taylor expansion of F and which typically require p^{th}-order smoothness. Again, this indexing makes sense in the context of min-max problems, where a p^{th}-order method uses a Taylor expansion involving p^{th}-order derivatives of g.*

A well-studied consequence of Definition 5.2.1 is the following:

Fact 5.2.4. *Let $u, v \in \mathcal{Z}$, and let $F : \mathcal{Z} \to \mathbb{R}^n$ be p^{th}-order L_p smooth. Then,*

$$\left|\left|F(v) - \mathcal{T}_{p-1}(v; u)\right|\right|_* \leq \frac{L_p}{p!} \left|\left|v - u\right|\right|^p. \tag{5.4}$$

Finally, our algorithms will all require the following assumption:

Assumption 5.2.5. *There exists a solution $x^* \in \mathcal{X}$ to the weak variational inequality MVI$(F, \mathcal{X})$, namely x^* is a point that satisfies (5.1).*

Assumption 5.2.5 always holds when $\mathcal{Z}$ is a compact convex set and F is continuous on $\mathcal{Z}$ [KS80].

5.2.1 Notions of convergence for variational inequalities

The main solution concept for (5.1) that we consider is an *ϵ-approximate weak solution* to MVI($F, \mathcal{Z}$), namely a point z^* such that:

$$\forall z \in \mathcal{Z}, \langle F(z), z^* - z \rangle \leq \epsilon. \tag{5.5}$$

Our main bounds will be of the form:

$$\forall z \in \mathcal{Z}, \frac{1}{\Gamma_K} \sum_{k=1}^{K} \gamma_k \langle F(z_k), z_k - z \rangle \leq \epsilon, \tag{5.6}$$

where z_k are iterates produced by our algorithm and γ_k are positive constants. We now show conditions under which a guarantee of the form (5.6) gives ϵ-approximate weak solutions.

Lemma 5.2.6. *Let $F : \mathcal{Z} \to \mathbb{R}^n$, let $z_k \in \mathcal{Z}$ for $k \in [K]$ be monotone, and let $\gamma_k > 0$. Let $\bar{z}_k = \frac{1}{\Gamma_K} \sum_{k=1}^{K} \gamma_k z_k$. Assume (5.6) holds. Then $\bar{z}_k$ is an ϵ-approximate weak solution to MVI$(F, \mathcal{Z})$.*

Proof. By monotonicity, we have:

$$\langle F(z_k), z_k - z \rangle \geq \langle F(z), z_k - z \rangle$$

Therefore,

$$\sum_{k=1}^{K} \gamma_k \langle F(z_k), z_k - z \rangle \geq \sum_{k=1}^{K} \gamma_k \langle F(z), z_k - z \rangle = \Gamma_K \langle F(z), \bar{z}_k - z \rangle$$

Then $\bar{z}$ is an ϵ-approximate solution to the weak MVI problem. □

5.2.2 Solving convex-concave min-max problems with variational inequalities

The classic notion of convergence for (1.1) is the duality gap, which we defined in (2.3). In this section, we will sometimes write $\psi_{\mathcal{X}\times\mathcal{Y}}$ to specify the sets over which the max and min are taken:

$$\psi_{\mathcal{X}\times\mathcal{Y}}(x, y) = \max_{\hat{y}\in\mathcal{Y}} g(x, \hat{y}) - \min_{\hat{x}\in\mathcal{X}} g(\hat{x}, y) \tag{5.7}$$

We will now show how to prove bounds on the duality gap given a bound like in (5.6), using the following lemma:

Lemma 5.2.7. *Let $F : \mathcal{Z} \to \mathbb{R}^n$, let $z_k \in \mathcal{Z}$ for $k \in [K]$, and let $\gamma_k > 0$. Let $\bar{z}_k = \frac{1}{\Gamma_K} \sum_{k=1}^{K} \gamma_k z_k$. Assume (5.6) holds. If F is the gradient descent-ascent field for a convex-concave problem (as in (5.2)), then $\psi_{\mathcal{X}\times\mathcal{Y}}(\bar{z}_k) \leq \epsilon$.*

Proof. When F is the gradient descent-ascent field for a convex-concave problem, we have:

$$\begin{aligned}
\langle F(z_k), z_k - z \rangle &= (\langle \nabla_x g(x_k, y_k), x_k - x \rangle + \langle -\nabla_y g(x_k, y_k), y_k - y \rangle) \\
&\geq g(x_k, y_k) - g(x, y_k) + g(x_k, y) - g(x_k, y_k) \\
&= g(x_k, y) - g(x, y_k)
\end{aligned}$$

Overall, then we have:

$$\sum_{k=1}^{K} \gamma_k \langle F(z_k), z_k - z \rangle \geq \sum_{k=1}^{K} \gamma_k (g(x_k, y) - g(x, y_k)) \geq \Gamma_K (g(\bar{x}_k, y) - g(x, \bar{y}_k))$$
$$\geq \Gamma_K \cdot \psi_{\mathcal{X} \times \mathcal{Y}}(\bar{x}_k, \bar{y}_k)$$

□

5.2.3 Related work

Monotone variational inequalities The weak MVI (5.1) is a classic and well-studied optimization problem [Min+62; Kom99; Nem04; MS10]. It is closely related to the strong MVI problem [Sta70], where the goal is to find a $z^* \in \mathcal{Z}$ such that

$$\forall z \in \mathcal{Z}, \langle F(z^*), z^* - z \rangle \leq 0. \tag{5.8}$$

When F is continuous and single-valued, any solution to the weak MVI (5.1) is a solution to the strong MVI.

Our algorithm is based on the Mirror Prox (MP) algorithm of [Nem04], which is a generalization of the extragradient method of [Kor76]. MP is a first-order method that achieves $O(1/k)$ iteration complexity, which is tight [NY83]. [MS10] prove convergence rates for MP in the unconstrained case by formulating MP as an instance of what they call a Hybrid Proximal Extragradient (HPE) algorithm. [MS12] provide a second-order algorithm to solve (5.1) in settings with second-order smoothness. That algorithm achieves an $O(1/k^{\frac{3}{2}})$ iteration complexity, and its analysis goes through the HPE framework from [MS10].

Min-max optimization Many convex-concave min-max optimization problems are either solved with MP or first-order no-regret algorithms. [OX18] show a lower bound of $\Omega(1/k)$ for first-order methods in constrained smooth convex-concave saddle point problems, even

in the simple case when $g(x, y) = f(x) + \langle Ax - b, y\rangle - h(y)$ for convex f and h. A number of recent works have also applied second-order methods to unconstrained smooth min-max problems, where the second-order information is often accessed through Hessian-vector products [Bal+18; GM18; Let+19; ADLH19; ALW19b; SA19].

Higher-order methods for convex optimization Higher-order methods have a long history of use in solving convex optimization problems. Assuming Lipschitz continuity of the Hessian, [Nes08] provided an accelerated variant of the cubic regularization method [NP06], which was further generalized by [Bae09] under p^{th}-order smoothness assumptions. The rate in [Nes08] was later improved by [MS13], and since then several works concerning lower bounds in this setting [AH18; ASS18] have shown that this rate is essentially tight (up to logarithmic factors) when the Hessian is Lipschitz continuous. Recently, several works have shown that the lower bound is also essentially tight for $p > 2$ [Gas+18; JWZ18; Bub+18; Bul18], leading to advances in related problems, such as ℓ_∞ regression [BL19] and parallel non-smooth convex optimization [Bub+19].

5.3 Main results

Our main result is a new higher-order method HIGHERORDERMIRRORPROX (Algorithm 2) for solving MVIs and convex-concave min-max problems with higher-order smoothness. We prove the following convergence rate:

Theorem 5.3.1. *Suppose $F : \mathcal{Z} \to \mathbb{R}^n$ is p^{th}-order L_p-smooth. Let $R \overset{def}{=} \max_{z \in \mathcal{Z}} D(z, z_1)$. Moreover, let $\epsilon = \frac{16 L_p}{p!} \left(\frac{R}{k}\right)^{\frac{p+1}{2}}$. Then for $\bar{z}_K$ as output by Algorithm 2:*

1. *If F is monotone, then $\bar{z}_k$ is an ϵ-approximate solution to the weak MVI problem.*
2. *If F is the gradient descent-ascent field for a convex-concave problem over $\mathcal{X}$ and $\mathcal{Y}$, then $\psi_{\mathcal{X}\times\mathcal{Y}}(\bar{z}_k) \leq \epsilon$.*

Our result matches the rate of [MS12] when $p = 2$ and gives improved convergence rates for higher p. To our knowledge, this is the first algorithm to achieve improved iteration complexity in the presence of higher-order smoothness. We compare our algorithm to that of [MS12] in more detail in Section 5.3.3.

As in other higher-order algorithms [Gas+18; JWZ18; Bub+18], each iteration of our algorithm requires access to an oracle for solving a minimization over a p^{th} order Taylor series. This oracle may be difficult to compute, particularly in the constrained setting. We can also consider running our algorithm in the unconstrained setting, which requires a slightly weaker unconstrained minimization oracle rather than a constrained minimization oracle. We discuss how to interpret our bounds in the unconstrained setting in Section 5.3.1.

Finally, we show how to instantiate our method in the second-order unconstrained case, giving the following running time bounds:

Theorem 5.3.2 (Main theorem, $p = 2$ (Informal))**.** *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is sufficiently smooth, and let $\{(\hat{z}_k, \gamma_k)\}_{k\in[K]}$ be the output of* HIGHERORDERMIRRORPROX $(p = 2)$ + BINARYSEARCH$_\gamma$ *(Algorithm 3). Then, for $\Gamma_K \stackrel{def}{=} \sum_{k=1}^{K} \gamma_k$, the iterates $\{\hat{z}_k\}_{k\in[K]}$ satisfy, for all $z \in \mathbb{R}^n$,*

$$\frac{1}{\Gamma_K}\sum_{k=1}^{K}\langle \gamma_k F(\hat{z}_k), \hat{z}_k - z\rangle \leq 8L_2\left(\frac{\max\{D(z, z_1), 1\}}{K}\right)^{\frac{3}{2}}, \tag{5.9}$$

with per-iteration cost dominated by $\tilde{O}(1)$ matrix inversions.[1]

5.3.1 Interpreting our results in the unconstrained setting

In the unconstrained setting, the standard solution concepts for MVIs and min-max problems can be vacuous in general. For example, for $g(x, y) = x^\top y$ and the associated vector field ξ, all approximate solutions to the min-max problem / MVI are exact solutions. However, the bounds we prove are still meaningful. In the MVI case, our guarantee can be interpreted

[1]Here we use the $\tilde{O}(\cdot)$ notation to suppress logarithmic factors.

Algorithm 2 HIGHERORDERMIRRORPROX

Input: $z_1 \in \mathcal{Z}$, $p \geq 1$, $0 < \epsilon < 1$, $K > 0$,
for $k = 1$ **to** K **do**
Determine γ_k, $\hat{z}_k$ such that:

$$\hat{z}_k = \arg\min_{z \in \mathcal{Z}} \left\{ \gamma_k \langle \mathcal{T}_p(\hat{z}_k; z_k), z - z_k \rangle + D(z, z_k) \right\}, \text{ and} \tag{5.10}$$

$$\frac{p!}{32 L_p \, ||\hat{z}_k - z_k||^{p-1}} \leq \gamma_k \leq \frac{p!}{16 L_p \, ||\hat{z}_k - z_k||^{p-1}} \tag{5.11}$$

$$z_{k+1} = \arg\min_{z \in \mathcal{Z}} \left\{ \langle \gamma_k F(\hat{z}_k), z - \hat{z}_k \rangle + D(z, z_k) \right\} \tag{5.12}$$

end for
Define $\Gamma_K \stackrel{\text{def}}{=} \sum_{k=1}^{K} \gamma_k$
return $\bar{z}_K \stackrel{\text{def}}{=} \frac{1}{\Gamma_K} \sum_{k=1}^{K} \gamma_k \hat{z}_k$

as stating that for all z such that $D(z, z_1) \leq R$, we have $\langle F(z), \bar{z}_k - z \rangle \leq O(R/k^{\frac{p+1}{2}})$ as long as $D(z^*, z_1) \leq R$. Likewise, for min-max problems, if $\mathcal{Z}'$ is a convex set containing z^*, then we can say that $\psi_{\mathcal{Z}'}(\bar{z}_k) \leq O(R/k^{\frac{p+1}{2}})$, where $R \geq \max_{z \in \mathcal{Z}'} D(z, z_1)$.

5.3.2 Explanation of our algorithm

Our algorithm is inspired by the Mirror Prox (MP) algorithm of [Nem04], defined as follows:

$$\hat{z}_k = \arg\min_{z \in \mathcal{Z}} \left\{ \langle \gamma_k F(z_k), z - z_k \rangle + D(z, z_k) \right\} \tag{5.13}$$

$$z_{k+1} = \arg\min_{z \in \mathcal{Z}} \left\{ \langle \gamma_k F(\hat{z}_k), z - \hat{z}_k \rangle + D(z, z_k) \right\} \tag{5.14}$$

where D is a Bregman divergence. [Nem04] motivates MP with a "conceptual prox method", which is given as follows:

$$z_{k+1} = \arg\min_{z \in \mathcal{Z}} \{ \langle \gamma_{k+1} F(z_{k+1}), z - z_{k+1} \rangle + D(z, z_k) \}. \tag{5.15}$$

This is an *implicit* method, as computing z_{k+1} requires solving the equation above for a given step-size γ_{k+1}. However, this method has good iteration complexity. [Nem04] shows that if

one could run (5.15) exactly, then the γ-averaged iterate $z_T = \frac{1}{\Gamma_K} \sum_{k=1}^{K} \gamma_k z_k$ converges at a rate of $O(1/\Gamma_K)$. Thus, if one could implement (5.15) with large step-sizes, one could achieve faster iteration complexity.

It turns out that as long as one approximates (5.15) with small error, one can achieve a similar convergence rate. The MP algorithm with constant γ_k does just that, leading to a $O(1/k)$ convergence rate. While one would like to increase the step-size in MP to improve the convergence rate, this approach does not work because MP with large step-sizes will no longer approximate (5.15) with small error.

In our algorithm, we replace the first-order minimization in MP (5.13) with a p^{th}-order minimization (5.10). We also simultaneously choose a particular step-size. This can be viewed as approximating (5.15) with large step-sizes while using the higher-order minimization to ensure that our algorithm is still a "good" approximation of (5.15).

5.3.3 Comparison to [MS12]

[MS12] give a second-order algorithm for solving (5.1) with iteration complexity $O(1/k^{\frac{3}{2}})$ in the presence of second-order smoothness. Like our algorithm, their algorithm also heavily relies on the idea of approximating a proximal point method with a large step-size. In fact, their algorithm is very similar to our algorithm in the second-order case. However, our analysis is rather different and arguably simpler. While their analysis goes through the Hybrid Proximal Extragradient framework of [MS10], our analysis relies on a natural extension of the Mirror Prox analysis. Finally, [MS12] only deal with the Euclidean setting, whereas we allow arbitrary norms.

While [MS12] do not explicitly instantiate their second-order oracle, they mention that their oracle reduces to solving a strongly monotone variational inequality, which can then be solving using a variety of approaches, including interior point methods. In the $p = 2$ case, our oracle can be similarly instantiated.

5.4 Higher-Order Mirror Prox Guarantees

In this section, we prove our main result of the convergence guarantees provided by Algorithm 2.

Lemma 5.4.1. *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is p^{th}-order L_p-smooth and let $\Gamma_K \overset{def}{=} \sum_{k=1}^{K} \gamma_k$. Then, the iterates $\{\hat{z}_k\}_{k\in[K]}$ generated by Algorithm 2 satisfy, for all $z \in \mathcal{Z}$,*

$$\frac{1}{\Gamma_K}\sum_{k=1}^{K}\langle \gamma_k F(\hat{z}_k), \hat{z}_k - z\rangle \leq \frac{16L_p}{p!}\left(\frac{D(z,z_1)}{K}\right)^{\frac{p+1}{2}}. \tag{5.16}$$

Theorem 5.4.2. *Suppose $F : \mathcal{Z} \to \mathbb{R}^n$ is p^{th}-order L_p-smooth. Let $R \overset{def}{=} \max_{z\in\mathcal{Z}} D(z, z_1)$. Moreover, let $\epsilon = \frac{16L_p}{p!}\left(\frac{R}{k}\right)^{\frac{p+1}{2}}$. Then for $\bar{z}_k$ as output by Algorithm 2:*

1. *If F is monotone, then $\bar{z}_k$ is an ϵ-approximate solution to the weak MVI problem.*
2. *If F is the gradient descent-ascent field for a convex-concave problem over $\mathcal{X}$ and $\mathcal{Y}$, then $\psi_{\mathcal{X}\times\mathcal{Y}}(\bar{z}_k) \leq \epsilon$.*

Theorem 5.4.2 follows immediately from Lemmas 5.2.6, 5.2.7, and 5.4.1. To prove Lemma 5.4.1, we will need to establish our main technical result (Lemma 5.4.3), which we prove in Section 5.4.1 and whose proof proceeds in a similar manner to the Mirror Prox analysis [Nem04; Tse08].

Lemma 5.4.3. *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is p^{th}-order L_p-smooth. Then, $\{\gamma_k, \hat{z}_k, z_{k+1}\}_{k\in[K]}$ as generated by Algorithm 2 satisfy, for all $z \in \mathcal{Z}$,*

$$\sum_{k=1}^{K}\langle \gamma_k F(\hat{z}_k), \hat{z}_k - z\rangle + \frac{1}{4}\sum_{k=1}^{K}||\hat{z}_k - z_k||^2 + \frac{1}{4}\sum_{k=1}^{K}||z_{k+1} - \hat{z}_k||^2 \leq D(z,z_1) - D(z,z_{k+1}). \tag{5.17}$$

We will also need the following technical lemma:

Lemma 5.4.4. *Let $R, a_k \geq 0$ for all $k \in [K]$, and let $\sum_{k=1}^{K} a^2 \leq R$. Then $\sum_{k=1}^{K} a^{-p} \geq \frac{k^{\frac{p}{2}+1}}{R^{\frac{p}{2}}}$.*

We prove Lemma 5.4.4 in Section 5.6.1. We now have the necessary tools to prove Lemma 5.4.1.

Proof of Lemma 5.4.1. Using Lemma 5.4.3, we can divide both sides of (5.17) by Γ_K, and so using the non-negativity of $\|\cdot\|$ and the Bregman divergence, we get:

$$\frac{1}{\Gamma_K}\sum_{k=1}^{K}\langle\gamma_k F(\hat{z}_k), \hat{z}_k - z\rangle \leq \frac{D(z, z_1)}{\Gamma_K}.$$

We simply need to lower bound $\frac{1}{\Gamma_K}$ in order to prove our convergence rate result. By Assumption 5.2.5, we know that there exists a solution z^* to MVI($F, \mathcal{Z}$), which means that for all $k \in [K]$, we have $\langle\gamma_k F(\hat{z}_k), \hat{z}_k - z^*\rangle \geq 0$. We can combine this with Lemma 5.4.3 to get that $\frac{1}{4}\sum_{k=1}^{K}||\hat{z}_k - z_k|| \leq D(z^*, z_1)$. Since $\gamma_k \geq \frac{p!}{32L_p||\hat{z}_k - z_k||^{p-1}}$, we can apply Lemma 5.4.4 by setting $a_k = ||\hat{z}_k - z_k||$ and $R = D(z^*, z_1)$, which gives the result. □

5.4.1 Proof of main technical result (Lemma 5.4.3)

Before proving Lemma 5.4.3, we state a useful lemma concerning the updates (5.10) and (5.12) in Algorithm 2.

Lemma 5.4.5 ([Tse08])**.** *Let $\phi(\cdot)$ be a convex function, let $z \in \mathcal{Z}$, and let*

$$z_+ = \arg\min_x \{\phi(x) + D(x, z)\}. \tag{5.18}$$

Then, for all $x \in \mathcal{Z}$,

$$\phi(x) + D(x, z) \geq \phi(z_+) + D(z_+, z) + D(x, z_+). \tag{5.19}$$

We now prove Lemma 5.4.3, which is our main technical result.

Proof of Lemma 5.4.3. By Lemma 5.4.5, along with the algorithm's determination of $\hat{z}_k$,

we have that for all $z \in \mathcal{Z}$,

$$\gamma_k \langle \mathcal{T}_{p-1}(\hat{z}_k; z_k), \hat{z}_k - z \rangle \leq D(z, z_k) - D(z, \hat{z}_k) - D(\hat{z}_k, z_k) \tag{5.20}$$

Using Lemma 5.4.5 again with the choice of z_{k+1}, it follows that for all $z \in \mathcal{Z}$,

$$\gamma_k \langle F(\hat{z}_k), z_{k+1} - z \rangle \leq D(z, z_k) - D(z, z_{k+1}) - D(z_{k+1}, z_k). \tag{5.21}$$

We may now observe that

$$\begin{aligned}
\gamma_k \langle F(\hat{z}_k), \hat{z}_k - z \rangle &= \gamma_k \langle F(\hat{z}_k), \hat{z}_k - z_{k+1} \rangle + \gamma_k \langle F(\hat{z}_k), z_{k+1} - z \rangle \\
&= \gamma_k \langle F(\hat{z}_k) - \mathcal{T}_{p-1}(\hat{z}_k; z_k), \hat{z}_k - z_{k+1} \rangle + \gamma_k \langle \mathcal{T}_{p-1}(\hat{z}_k; z_k), \hat{z}_k - z_{k+1} \rangle \\
&\quad + \gamma_k \langle F(\hat{z}_k), z_{k+1} - z \rangle \\
&\leq \gamma_k \langle F(\hat{z}_k) - \mathcal{T}_{p-1}(\hat{z}_k; z_k), \hat{z}_k - z_{k+1} \rangle - D(z_{k+1}, \hat{z}_k) - D(\hat{z}_k, z_k) \\
&\quad + D(z, z_k) - D(z, z_{k+1}),
\end{aligned}$$

where the final inequality follows from (5.20) and (5.21). Now by Hölder's inequality, using eq. (5.4), and the 1-strong convexity of $d(\cdot)$ w.r.t. $||\cdot||$, it follows that

$$\begin{aligned}
\gamma_k \langle F(\hat{z}_k), \hat{z}_k - z \rangle &\leq \gamma_k \left|\left|F(\hat{z}_k) - \mathcal{T}_{p-1}(\hat{z}_k; z_k)\right|\right|_* \cdot ||\hat{z}_k - z_{k+1}|| - D(z_{k+1}, \hat{z}_k) - D(\hat{z}_k, z_k) \\
&\qquad + D(z, z_k) - D(z, z_{k+1}) \\
&\leq \frac{\gamma_k L_p}{p!} ||\hat{z}_k - z_k||^p \cdot ||\hat{z}_k - z_{k+1}|| - D(z_{k+1}, \hat{z}_k) - D(\hat{z}_k, z_k) + D(z, z_k) - D(z, z_{k+1}) \\
&\leq \frac{\gamma_k L_p}{p!} ||\hat{z}_k - z_k||^p \cdot ||\hat{z}_k - z_{k+1}|| - \frac{1}{2} ||z_{k+1} - \hat{z}_k||^2 - \frac{1}{2} ||\hat{z}_k - z_k||^2 \\
&\quad + D(z, z_k) - D(z, z_{k+1}).
\end{aligned}$$

Finally, by our guarantee from Algorithm 2 that $\gamma_k \leq \frac{p!}{16 L_p ||\hat{z}_k - z_k||^{p-1}}$, and using the fact

that $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$ for $a, b \geq 0$, it follows that

$$\gamma_k \langle F(\hat{z}_k), \hat{z}_k - z \rangle + \frac{1}{4} ||\hat{z}_k - z_k||^2 + \frac{1}{4} ||z_{k+1} - \hat{z}_k||^2 \leq D(z, z_k) - D(z, z_{k+1}). \quad (5.22)$$

Summing over $k = 1, \ldots, K$ gives the result. □

5.5 Instantiating HIGHERORDERMIRRORPROX (for $p = 2$)

In this section, we provide an efficient implementation of HIGHERORDERMIRRORPROXfor the case where F is second-order smooth. In particular, we consider the unconstrained problem (i.e., $\mathcal{Z} = \mathbb{R}^n$) with the Bregman divergence chosen as $D(u, v) = \frac{1}{2} ||u - v||_2^2$. First, for technical reasons, we require the following assumption:

Assumption 5.5.1. *During the execution of Algorithm 3, for all $k \geq 1$, $\gamma > 0$, we assume that $(\mathbf{I} + \gamma \nabla F(z_k))$ is invertible and $\sigma_{\min}(\gamma^{-1}\mathbf{I} + \nabla F(z_k)) \geq \sigma_{\min}(\nabla F(z_k))$.*

As we discuss further in Section 5.10, this always holds for convex-concave min-max problems. We then arrive at the following result for this setting:

Theorem 5.5.2 (Main theorem, $p = 2$)**.** *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is first-order L_1-smooth, second-order L_2-smooth, and Assumption 5.5.1 holds. Let z^* be a solution to MVI($F, \mathbb{R}^n$), let $K > 0$, and let $\{(\hat{z}_k, \gamma_k)\}_{k \in [K]}$ be the output of* HIGHERORDERMIRRORPROX *($p = 2$) + BINARYSEARCH$_\gamma$ (Algorithm 3). Further assume that, for all k, $\sigma_{\min}(\nabla F(z_k)) \geq \mu$. Then, for $\Gamma_K \stackrel{def}{=} \sum_{k=1}^{K} \gamma_k$, the iterates $\{\hat{z}_k\}_{k \in [K]}$ satisfy, for all $z \in \mathbb{R}^n$,*

$$\frac{1}{\Gamma_K} \sum_{k=1}^{K} \langle \gamma_k F(\hat{z}_k), \hat{z}_k - z \rangle \leq 8 L_2 \left(\frac{\max\{D(z, z_1), 1\}}{K} \right)^{\frac{3}{2}}. \quad (5.23)$$

In addition, the computational cost of each iteration of Algorithm 3 is dominated by a total of $O\left(\log\left(\frac{L_1 ||z_1 - z^*||_2 K}{\mu}\right)\right)$ matrix inversions.

The proof of Theorem 5.5.2 can be found in Section 5.7.1, and we provide a sketch below.

Sketch. In the second-order unconstrained case, we can compute the implicitly defined update (5.10) for a given γ_k by setting $\hat{z}_k \stackrel{\text{def}}{=} z_k - \gamma_k(\mathbf{I} + \gamma_k \nabla F(z_k))^{-1} F(z_k)$. Thus, it suffices to find a γ_k that satisfies (5.11). We show that either we can find such a γ_k or we find a suitable γ_+ and γ_- for our binary search algorithm. Finally, Lemma 5.5.3 shows that our binary search algorithm BINARYSEARCH$_\gamma$ finds appropriate γ_k in $O\left(\log\left(\frac{L_1 ||z_1 - z^*||_2 K}{\mu}\right)\right)$ matrix inversions. A key part of proving Lemma 5.5.3 is showing that a certain function $q(\gamma)$ (defined in (5.25)) has bounded derivative (Lemma 5.5.4). □

Algorithm 3 HIGHERORDERMIRRORPROX ($p = 2$) + BINARYSEARCH$_\gamma$

Input: $z_1 \in \mathbb{R}^n$, $0 < \epsilon < 1$, $K > 0$
for $k = 1$ **to** K **do**
  Set $\gamma_- = \frac{\sigma_{\min}(\nabla F(z_k))}{12 ||F(z_k)||_2}$, $\gamma_+ = k^{\frac{3}{2}}$
  if $\gamma_+ < \frac{1}{8||\hat{z}_k(\gamma_+) - z_k||_2}$ **then**
    $\gamma_k \leftarrow \gamma_+$
  else if $\gamma_- \geq \gamma_+$ **then**
    $\gamma_k \leftarrow \gamma_-$
  else
    $\gamma_k \leftarrow$ BINARYSEARCH$_\gamma(z_k, \epsilon, \gamma_-, \gamma_+)$
  end if
  $\hat{z}_k \stackrel{\text{def}}{=} z_k - \gamma_k(\mathbf{I} + \gamma_k \nabla F(z_k))^{-1} F(z_k)$
  $z_{k+1} = \arg\min_z \{\langle \gamma_k F(\hat{z}_k), z - \hat{z}_k \rangle + D(z, z_k)\}$
end for
Define $\Gamma_K \stackrel{\text{def}}{=} \sum_{k=1}^{K} \gamma_k$
return $\bar{z}_K \stackrel{\text{def}}{=} \frac{1}{\Gamma_K} \sum_{k=1}^{K} \gamma_k \hat{z}_k$

5.5.1 Binary search

The following lemmas show the correctness of the main binary search procedure. We prove these lemmas in Sections 5.7.2 and 5.7.3.

Lemma 5.5.3. *Suppose γ_-, γ_+ are such that $\gamma_- \leq \frac{1}{12||\hat{z}_k(\gamma_-)-z_k||_2}$ and $\gamma_+ \geq \frac{1}{12||\hat{z}_k(\gamma_+)-z_k||_2}$, where $\hat{z}_k(\gamma) = z_k - \gamma(\mathbf{I} + \gamma\nabla F(z_k))^{-1}F(z_k)$, Then,* BINARYSEARCH$_\gamma$ *(Algorithm 4) outputs $\bar{\gamma}$ such that*

$$\frac{1}{16\,||\hat{z}_k(\bar{\gamma}) - z_k||_2} \leq \bar{\gamma} \leq \frac{1}{8\,||\hat{z}_k(\bar{\gamma}) - z_k||_2} \tag{5.24}$$

after $N = O\left(\log\left(\frac{\bar{C}K}{\delta}\right)\right)$ iterations of the binary search procedure, where δ, $\bar{C}$ are as defined in the algorithm.

Lemma 5.5.4. *Let $q : \mathbb{R} \mapsto \mathbb{R}$ be defined as*

$$q(\gamma) \overset{def}{=} \frac{1}{12\gamma\,||(\mathbf{I} + \gamma\nabla F(z_k))^{-1}F(z_k)||} = \frac{1}{12\,||(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1}F(z_k)||}, \tag{5.25}$$

and let $\delta \overset{def}{=} \frac{\sigma_{\min}(\nabla F(z_k))}{12||F(z_k)||}$. Then, for all $\gamma \geq \delta$, we have

$$\left|\frac{d}{d\gamma}q(\gamma)\right| \leq C, \quad \textit{for} \quad C \overset{def}{=} \frac{1}{\delta^2}\left(\frac{\frac{1}{\delta} + ||\nabla F(z_k)||}{12\sigma_{\min}(\nabla F(z_k))\,||F(z_k)||}\right)^3. \tag{5.26}$$

Algorithm 4 BINARYSEARCH$_\gamma$

Input: z_k, $0 < \epsilon < 1$, γ_-^{init}, γ_+^{init}.
Initialize $\gamma_- \leftarrow \gamma_-^{\text{init}}$, $\gamma_+ \leftarrow \gamma_+^{\text{init}}$, $\bar{\gamma} \leftarrow \frac{\gamma_- + \gamma_+}{2}$
Set $\delta = \frac{\sigma_{\min}(\nabla F(z_k))}{12||F(z_k)||}$, $C = \frac{1}{\delta^2}\left(\frac{\frac{1}{\delta} + ||\nabla F(z_k)||}{12\sigma_{\min}(\nabla F(z_k))||F(z_k)||}\right)^3$, $\bar{C} = \max\{C, 1\}$, $N = O(\log(\frac{\bar{C}K}{\delta}))$.
Define $\hat{z}_k(\gamma) \overset{\text{def}}{=} z_k - \gamma(\mathbf{I} + \gamma\nabla F(z_k))^{-1}F(z_k)$
for $k = 0$ **to** $N - 1$ **do**
  $D = \frac{1}{12||\hat{z}_k(\bar{\gamma}) - z_k||}$
  if $\bar{\gamma} \leq D$ **then**
    $\gamma_- \leftarrow \bar{\gamma}$
  else
    $\gamma_+ \leftarrow \bar{\gamma}$
  end if
  $\bar{\gamma} = \frac{\gamma_- + \gamma_+}{2}$
end for
return $\bar{\gamma} \leftarrow \gamma_+$

5.6 Proofs from Section 5.4

5.6.1 Proof of Lemma 5.4.4

Proof of Lemma 5.4.4. We use the following power means:

$$M_1(x) = \frac{\sum_{k=1}^K x_k}{K}$$
$$M_{-2/p}(x) = \left(\frac{\sum_{k=1}^K x_k^{-2/p}}{K}\right)^{-p/2}$$

By the power mean inequality, we have $M_{1/p}(x) \geq M_{-2/p}(x)$, so letting $x_k = \frac{1}{a_k^p}$ gives:

$$\frac{\sum_{k=1}^K \frac{1}{a_k^p}}{K} \geq \left(\frac{K}{\sum_{k=1}^K a_k^2}\right)^{p/2} \geq \left(\frac{K}{R}\right)^{p/2}$$
$$\Rightarrow \sum_{k=1}^K \frac{1}{a_k^p} \geq \frac{K^{1+p/2}}{R^{p/2}}$$

□

5.6.2 Proof of Lemma 5.4.5

Proof of Lemma 5.4.5. By the optimality condition for z_+, we know that for all $x \in \mathcal{Z}$,

$$\phi(x) + \langle \nabla_x D(z_+, z), x - z_+ \rangle \geq \phi(z_+). \tag{5.27}$$

Rearranging and adding $D(x, z)$ to both sides gives us

$$\begin{aligned}\phi(x) + D(x, z) &\geq \phi(z_+) + D(x, z) - \langle \nabla_x D(z_+, z), x - z_+ \rangle \\ &= \phi(z_+) + D(x, z) + D(x, z_+) + D(z_+, z) - D(x, z) \\ &= \phi(z_+) + D(x, z_+) + D(z_+, z),\end{aligned}$$

where the first equality comes from the Bregman three-point property, i.e.,

$$\langle \nabla d(w) - \nabla d(v), u - v \rangle = D(u, v) + D(v, w) - D(u, w), \text{ for all } u, v, w \in \mathcal{Z}. \quad (5.28)$$

□

5.7 Proofs from Section 5.5

5.7.1 Proof of Theorem 5.5.2

Proof of Theorem 5.5.2. We will first show that the choices of γ_+ and γ_- are valid binary search bounds whenever BINARYSEARCH$_\gamma$ is called by Algorithm 3, i.e., that $\gamma_+ \geq \frac{1}{12||\hat{z}_k(\gamma_+) - z_k||_2}$ and $\gamma_- \leq \frac{1}{12||\hat{z}_k(\gamma_-) - z_k||_2}$. We begin with our choice of $\gamma_+ = k^{\frac{3}{2}}$. Suppose that, for some iteration t, it is the case that $\gamma_+ < \frac{1}{8||\hat{z}_k(\gamma_+) - z_k||_2}$. If so, then the algorithm sets $\gamma_k \leftarrow \gamma_+$, which means that $\Gamma_K \geq \gamma_+ = k^{\frac{3}{2}}$. Therefore, since we know that

$$\frac{1}{\Gamma_K} \sum_{k=1}^{K} \langle \gamma_k F(\hat{z}_k), \hat{z}_k - z \rangle \leq 8L_2 \frac{D(z, z_1)}{\Gamma_K}, \quad (5.29)$$

it follows that

$$\frac{1}{\Gamma_K} \sum_{k=1}^{K} \langle \gamma_k F(\hat{z}_k), \hat{z}_k - z \rangle \leq 8L_2 \frac{D(z, z_1)}{K^{\frac{3}{2}}} \leq 8L_2 \frac{D(z, z_1)}{K^{\frac{3}{2}}} \leq 8L_2 \left(\frac{\max\{D(z, z_1), 1\}}{K} \right)^{\frac{3}{2}}, \quad (5.30)$$

and so we would be done. In addition, supposing it is the case that $\gamma_- \geq \gamma_+$ (at which point, the algorithm sets $\gamma_k \leftarrow \gamma_-$), we again reach this conclusion by the same reasoning. For ensuring the validity of γ_-, note that by (5.37), it follows that $\gamma_- = \delta \leq \frac{1}{12||\hat{z}_k(\delta) - z_k||_2}$.

Having established the validity of the binary search bounds in the case that the search routine is in fact called, we now move on to show how we may explicitly instantiate the implicitly defined update in (5.10). Namely, in this setting the key conditions (5.10) and

(5.11) that must simultaneously hold can be equivalently expressed as

$$\hat{z}_k = \arg\min_{z \in \mathbb{R}^n} \left\{ \gamma_k \langle F(z_k) + \nabla F(z_k)(\hat{z}_k - z_k), z - z_k \rangle + \frac{1}{2} ||z - z_k||^2 \right\}, \text{ and} \tag{5.31}$$

$$\frac{1}{16 L_1 ||\hat{z}_k - z_k||_2} \leq \gamma_k \leq \frac{1}{8 L_1 ||\hat{z}_k - z_k||_2}. \tag{5.32}$$

From (5.31), it follows by first-order optimality conditions that $\gamma_k(F(z_k) + \nabla F(z_k)(\hat{z}_k - z_k)) + \hat{z}_k - z_k = 0$, and so rearranging gives us

$$(\mathbf{I} + \gamma_k \nabla F(z_k))\hat{z}_k = (\mathbf{I} + \gamma_k \nabla F(z_k)) z_k - \gamma_k F(z_k).$$

Since we assume that $(\mathbf{I} + \gamma_k \nabla F(z_k))$ is invertible, it follows that

$$\hat{z}_k = z_k - \gamma_k (\mathbf{I} + \gamma_k \nabla F(z_k))^{-1} F(z_k), \tag{5.33}$$

which is precisely the update that occurs in Algorithm 3. All that remains is to ensure that we may determine γ_k such that (5.32) holds, which follows from the output of BINARYSEARCH$_\gamma$ as a consequence of Lemma 5.5.3. Finally, since the iteration complexity of BINARYSEARCH$_\gamma$ is bounded by

$$N = O\left(\log\left(\frac{\bar{C}K}{\delta}\right)\right) = O\left(\log\left(\frac{K ||F(z_k)||_2}{\sigma_{\min}(\nabla F(z_k))}\right)\right) \leq O\left(\log\left(\frac{L_1 ||z_1 - z^*||_2 K}{\mu}\right)\right), \tag{5.34}$$

where the final inequality follows from Lemma 5.8.1, which bounds $||F(z_k)||$, along with our assumption that, for all k, $\sigma_{\min}(\nabla F(z_k)) \geq \mu$, and each iteration of BINARYSEARCH$_\gamma$ requires $O\left(\log\left(\frac{L_1 ||z_1 - z^*||_2 K}{\mu}\right)\right)$ matrix inversions, which results in the total complexity in the theorem. □

5.7.2 Proof of Lemma 5.5.3

Proof of Lemma 5.5.3. By assumption, we have that γ_- and γ_+ are initialized to be valid search bounds, i.e., $\gamma_- \leq \frac{1}{12||\hat{z}_k(\gamma_-)-z_k||}$ and $\gamma_+ \geq \frac{1}{12||\hat{z}_k(\gamma_+)-z_k||}$. By Lemma 5.5.4 and letting $\bar{C} \stackrel{\text{def}}{=} \max\{C, 1\}$, we know that, for all $x, y \geq \gamma_-$,

$$|q(y) - q(x)| \leq \bar{C} \cdot |y - x| \tag{5.35}$$

After $N = O\left(\log\left(\frac{\bar{C}K}{\delta}\right)\right)$ iterations of the binary search procedure we know that

$$|\gamma_+ - \gamma_-| \leq \frac{\delta}{100\bar{C}} \leq \frac{\delta}{100}, \tag{5.36}$$

and so taken together with (5.35), we have

$$\begin{aligned}\gamma_+ &\leq \gamma_- + \frac{\delta}{100} \leq q(\gamma_-) + \frac{\delta}{100} \leq q(\gamma_+) + C\,|\gamma_+ - \gamma_-| + \frac{\delta}{100} \leq q(\gamma_+) + \frac{2\delta}{100} \\ &\leq \frac{3}{2} q(\gamma_+) = \frac{1}{8\,||\hat{z}_k(\gamma_+) - z_k||}.\end{aligned}$$

Here, the last inequality follows from the fact that, for $\gamma > 0$,

$$\begin{aligned}q(\gamma) = \frac{1}{12\,||(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1} F(z_k)||} &\geq \frac{1}{12\,||(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1}|| \cdot ||F(z_k)||} \\ &\geq \frac{1}{12\,||\nabla F(z_k)^{-1}|| \cdot ||F(z_k)||} \\ &= \frac{\sigma_{\min}(\nabla F(z_k))}{12\,||F(z_k)||} \\ &= \delta.\end{aligned} \tag{5.37}$$

Thus, it follows that

$$\frac{1}{16\,||\hat{z}_k(\bar{\gamma}) - z_k||} \leq \bar{\gamma} \leq \frac{1}{8\,||\hat{z}_k(\bar{\gamma}) - z_k||} \tag{5.38}$$

for $\bar{\gamma} = \gamma_+$, as determined by Algorithm 4.

□

5.7.3 Proof of Lemma 5.5.4

Proof of Lemma 5.5.4. We begin by rewriting $q(\gamma)$ as

$$\begin{aligned} q(\gamma) &= \frac{1}{12}\left(F(z_k)^\top(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1^\top}(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1}F(z_k)\right)^{-1/2} \\ &= \frac{1}{12}\left(F(z_k)^\top(\gamma^{-1}\mathbf{I} + \nabla F(z_k)^\top)^{-1}(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1}F(z_k)\right)^{-1/2} \end{aligned}$$

Now, let $M_1(\gamma) \stackrel{\text{def}}{=} (\gamma^{-1}\mathbf{I} + \nabla F(z_k)^\top)^{-1}$ and $M_2(\gamma) \stackrel{\text{def}}{=} (\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1}$. By standard matrix calculus, we may observe that

$$\frac{d}{d\gamma}M_1(\gamma) = \frac{1}{\gamma^2}(\gamma^{-1}\mathbf{I} + \nabla F(z_k)^\top)^{-2} \text{ and } \frac{d}{d\gamma}M_2(\gamma) = \frac{1}{\gamma^2}(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-2}. \quad (5.39)$$

It follows that

$$\begin{aligned} \frac{d}{d\gamma}q(\gamma) &= -\frac{1}{2}q(\gamma)^3 \cdot \Bigg(F(z_k)^\top(\gamma^{-1}\mathbf{I} + \nabla F(z_k)^\top)^{-1}\left(\frac{d}{d\gamma}M_2(\gamma)\right)F(z_k) \\ &\qquad + F(z_k)^\top\left(\frac{d}{d\gamma}M_1(\gamma)\right)(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1}F(z_k)\Bigg) \\ &= -\frac{1}{2\gamma^2}q(\gamma)^3 \cdot \Bigg(F(z_k)^\top(\gamma^{-1}\mathbf{I} + \nabla F(z_k)^\top)^{-1}(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-2}F(z_k) \\ &\qquad + F(z_k)^\top(\gamma^{-1}\mathbf{I} + \nabla F(z_k)^\top)^{-2}(\gamma^{-1}\mathbf{I} + \nabla F(z_k))^{-1}F(z_k)\Bigg). \end{aligned}$$

Now, by standard norm inequalities, we have

$$\begin{aligned}\left|\frac{d}{d\gamma}q(\gamma)\right| &\leq \frac{1}{2\gamma^2}\left|q(\gamma)\right|^3\Big(\left|\left|(\gamma^{-1}\mathbf{I}+\nabla F(z_k)^\top)^{-1}\right|\right|\cdot\left|\left|(\gamma^{-1}\mathbf{I}+\nabla F(z_k))^{-1}\right|\right|^2 \\ &\qquad + \left|\left|(\gamma^{-1}\mathbf{I}+\nabla F(z_k)^\top)^{-1}\right|\right|^2\cdot\left|\left|(\gamma^{-1}\mathbf{I}+\nabla F(z_k))^{-1}\right|\right|\Big)\left|\left|F(z_k)\right|\right|^2 \\ &= \frac{1}{\gamma^2}\left|q(\gamma)\right|^3\cdot\left|\left|(\gamma^{-1}\mathbf{I}+\nabla F(z_k))^{-1}\right|\right|^3\cdot\left|\left|F(z_k)\right|\right|^2.\end{aligned}$$

Note that for all $\gamma \geq \delta$,

$$|q(\gamma)| = \frac{1}{12\left|\left|(\gamma^{-1}\mathbf{I}+\nabla F(z_k))^{-1}F(z_k)\right|\right|} \leq \frac{\gamma^{-1}+||\nabla F(z_k)||}{||F(z_k)||} \leq \frac{\frac{1}{\delta}+||\nabla F(z_k)||}{12\left|\left|F(z_k)\right|\right|} \tag{5.40}$$

and

$$\left|\left|(\gamma^{-1}\mathbf{I}+\nabla F(z_k))^{-1}\right|\right| = \frac{1}{\sigma_{\min}(\gamma^{-1}\mathbf{I}+\nabla F(z_k))} \leq \frac{1}{\sigma_{\min}(\nabla F(z_k))}, \tag{5.41}$$

where the final inequality follows by Assumption 5.5.1. Taken together, this gives us that

$$\left|\frac{d}{d\gamma}q(\gamma)\right| \leq \frac{1}{\delta^2}\left(\frac{\frac{1}{\delta}+||\nabla F(z_k)||}{12\sigma_{\min}(\nabla F(z_k))\left|\left|F(z_k)\right|\right|}\right)^3, \tag{5.42}$$

and so the lemma follows. □

5.8 Proof of Lemma 5.8.1

Lemma 5.8.1. *Assume F is first-order L_1 smooth and $D(u,v) = \frac{1}{2}\left|\left|u-v\right|\right|^2$.*

$$||F(z_k)|| \leq 4\sqrt{k}L_1\left|\left|z_1 - z^*\right|\right| \tag{5.43}$$

To prove Lemma 5.8.1, we need the following lemma, which we prove in Section 5.8.1.

Lemma 5.8.2. *Suppose $F:\mathbb{R}^n\to\mathbb{R}^n$ is p^{th}-order L_p-smooth. Let $\{z_k\}_{k=1}^K$ be the iterates*

generated by Algorithm 2 and let z^* be a solution to MVI($F, \mathcal{Z}$). Then for any $v \in \mathcal{Z}$,

$$||z_k - v||^2 \leq 2k \left(8D(z_1, z^*) + ||z_1 - v||^2\right) \tag{5.44}$$

Proof of Lemma 5.8.1. By Assumption 5.2.5, we know there exists a z^* such that (5.1) holds. By Lemma 5.9.1, any such z^* is also a solution to (5.8), namely:

$$\forall z \in \mathbb{R}^n, \langle F(z^*), z^* - z\rangle \leq 0 \tag{5.45}$$

Since we are in the unconstrained setting, this implies that $F(z^*) = 0$. Then we have:

$$||F(z_k)|| = ||F(z_k) - F(z^*)|| \leq L_1 ||z_k - z^*|| \tag{5.46}$$

where the inequality follows by the L_1 smoothness of F. By Lemma 5.8.2, we have

$$||z_k - z^*|| \leq \sqrt{2k \left(4 ||z_1 - z^*||^2 + ||z_1 - z^*||^2\right)} \leq 4\sqrt{k} ||z_1 - z^*|| \tag{5.47}$$

Combining this with (5.46) gives the result. □

5.8.1 Proof of Lemma 5.8.2

We will need the following two lemmas to prove Lemma 5.8.2:

Lemma 5.8.3. *Let $a_i > 0$ for $i \in [n]$. Then,*

$$\left(\sum_{i=1}^{n} a_i\right)^2 \leq n \sum_{i=1}^{n} a_i^2 \tag{5.48}$$

Proof. Let a be the vector of a_i's. We define the following power means:

$$M_1(a) = \frac{\sum_{i=1}^n a_i}{n} \tag{5.49}$$

$$M_2(a) = \left(\frac{\sum_{i=1}^n a_i^2}{n}\right)^{1/2} \tag{5.50}$$

By the power mean inequality, we have $M_1(a) \leq M_2(a)$, which gives the result. □

Lemma 5.8.4. *Let z^* be the solution to MVI$(F, \mathcal{Z})$. Then for the iterates z_k of Algorithm 2 initialized at z_1, we have:*

$$\frac{1}{8}\sum_{k=1}^K ||z_{k+1} - z_k||^2 \leq D(z^*, z_1). \tag{5.51}$$

Proof. This follows from two simple observations. First, note that:

$$\sum_{k=1}^K ||z_{k+1} - z_k||^2 \leq \sum_{k=1}^K (2\,||z_{k+1} - \hat{z}_k||^2 + 2\,||\hat{z}_k - z_k||^2). \tag{5.52}$$

Now, by Assumption 5.2.5, we know that each term of $\sum_{k=1}^K \langle \gamma_k F(\hat{z}_k), \hat{z}_k - z\rangle$ is non-negative for some $z^* \in \mathcal{Z}$, namely the solution to MVI$(F, \mathcal{Z})$. Combining this with Lemma 5.4.3 and (5.52) gives the result. □

Proof of Lemma 5.8.2. By the triangle inequality, we have:

$$||z_k - v||^2 \leq \left(\sum_{\tau=1}^k ||z_\tau - z_{\tau+1}|| + ||z_1 - v||\right)^2 \tag{5.53}$$

$$\leq (k+1)\left(\sum_{\tau=1}^k ||z_\tau - z_{\tau+1}||^2 + ||z_1 - v||^2\right) \tag{5.54}$$

where the second inequality follows from using Lemma 5.8.3 with $a_i = ||z_i - z_{i+1}||$ for $i \in [k]$ and $a_{k+1} = ||z_1 - v||$. We then apply Lemma 5.8.4 to (5.54) to get the result, using the fact that $k + 1 \leq 2k$. □

5.9 Equivalence of exact solutions to weak and strong MVIs

Lemma 5.9.1 ([KS80])**.** *For continuous $F : \mathbb{R}^n \to \mathbb{R}^n$, any solution of (5.1) is a solution to (5.8).*

Proof. Let z^* be a solution to (5.1). Let $z = z^* + t(v - z^*)$ for an arbitrary $v \in \mathcal{Z}$ and for $t > 0$. Then:

$$\langle F(z^* + t(v - z^*)), -t(v - z^*)\rangle \leq 0 \tag{5.55}$$

$$\iff \langle F(z^* + t(v - z^*)), z^* - v)\rangle \leq 0 \tag{5.56}$$

Taking the limit of (5.56) as t goes to 0 gives:

$$\langle F(z^*), z^* - v\rangle \leq 0 \tag{5.57}$$

Thus, z^* is a solution to (5.8). □

5.10 Invertibility concerns

While the general setting of Algorithm 3 assumes $(\mathbf{I} + \nabla F(z_k))$ is invertible, it turns out that for convex-concave games, this assumption is not necessary. In particular, the Jacobian of the vector field (5.2) is

$$\nabla F(x, y) = \begin{bmatrix} \nabla^2_{xx}\phi(x, y) & \nabla^2_{xy}\phi(x, y) \\ -\nabla^2_{yx}\phi(x, y) & -\nabla^2_{yy}\phi(x, y) \end{bmatrix}. \tag{5.58}$$

Note that there is a natural decomposition of $\nabla F(x, y)$ as a sum of a symmetric and an anti-symmetric matrix, namely

$$\nabla F(x,y) = \begin{bmatrix} \nabla^2_{xx}\phi(x,y) & \mathbf{0} \\ \mathbf{0} & -\nabla^2_{yy}\phi(x,y) \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \nabla^2_{xy}\phi(x,y) \\ -\nabla^2_{yx}\phi(x,y) & \mathbf{0} \end{bmatrix}. \quad (5.59)$$

The following is a useful lemma about the real part of eigenvalues of matrices, based on such a symmetric-asymmetric decomposition.

Lemma 5.10.1. *Let M be a real matrix such that $M = S + A$, where S is a symmetric real matrix and A is an antisymmetric real matrix. If S is nonsingular, then M is nonsingular. Likewise, if S is positive definite (or PSD), then the real part of eigenvalues of M are positive (or non-negative).*

Proof of Lemma 5.10.1. Let v be an eigenvector of M with eigenvalue λ (these may both be complex). Let $v = v_r + iv_i$ and $\lambda = \lambda_r + i\lambda_i$ be the decompositions of v and λ into real and imaginary parts.

$$\begin{aligned} \lambda\,||v|| = v^*Mv &= v^*Sv + v^*Av \\ &= (v_r - iv_i)^\top S(v_r + iv_i) + (v_r - iv_i)^\top A(v_r + iv_i) \\ &= v_r^\top Sv_r + v_i^\top Sv_i + i(v_r^\top Sv_i - v_i^\top Sv_r) + v_r^\top Av_r + v_i^\top Av_i \\ &\quad + i(v_r^\top Av_i - v_i^\top Av_r) \end{aligned}$$

Since $x^\top Ax = 0$ for any antisymmetric matrix A, we have that $\lambda_r = \frac{1}{||v||}(v_r^\top Sv_r + v_i^\top Sv_i)$, which implies the conclusions of the lemma. To see the fact about antisymmetric matrices, observe:

$$x^\top Ax = x^\top A^\top x = -x^\top Ax \iff 2x^\top Ax = 0$$

□

By convexity and concavity of $\phi(x, y)$ in x and y, respectively, we know that the symmetric part of (5.59) is PSD for all $z \in \mathcal{Z}$. It follows that, for all t, $(\mathbf{I} + \nabla F(z_k))$ is positive definite, and therefore invertible. It may additionally be seen in this setting that $\sigma_{\min}(\gamma^{-1}\mathbf{I} + \nabla F(z_k)) \geq \sigma_{\min}(\nabla F(z_k))$.

REFERENCES

[ALLW18a] J. Abernethy, K. A. Lai, K. Levy, and J.-K. Wang. "Faster Rates for Convex-Concave Games". In: *Conference on Learning Theory (COLT)* (2018).

[ALLW18b] J. Abernethy, K. A. Lai, K. Y. Levy, and J.-K. Wang. "Faster rates for convex-concave games". In: CONFERENCE ON LEARNING THEORY (COLT) (2018).

[ALW19a] J. Abernethy, K. A. Lai, and A. Wibisono. "Fictitious Play: Convergence, Smoothness, and Optimism". In: (2019). URL: http://arxiv.org/abs/1911.08418.

[ALW19b] J. Abernethy, K. A. Lai, and A. Wibisono. "Last-iterate convergence rates for min-max optimization". In: (2019). URL: https://arxiv.org/abs/1906.02027.

[AW17] J. Abernethy and J.-K. Wang. "Frank-Wolfe and Equilibrium Computation". In: *Annual Conference on Neural Information Processing Systems (NIPS)* (2017).

[Adl13] I. Adler. "The equivalence of linear programs and zero-sum games". In: *International Journal of Game Theory* 42.1 (2013), pp. 165–177.

[ADLH19] L. Adolphs, H. Daneshmand, A. Lucchi, and T. Hofmann. "Local Saddle Point Optimization: A Curvature Exploitation Approach". In: *Artificial Intelligence and Statistics (AISTATS)*. 2019.

[Aga+18] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. "A Reductions Approach to Fair Classification". In: *International Conference on Machine Learning (ICML)*. 2018, pp. 60–69.

[AH18] N. Agarwal and E. Hazan. "Lower bounds for higher-order convex optimization". In: *Conference on Learning Theory (COLT)*. 2018.

[Aga+17] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. "Finding approximate local minima faster than gradient descent". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2017, pp. 1195–1199.

[AZH16] Z. Allen-Zhu and E. Hazan. "Variance reduction for faster non-convex optimization". In: *International Conference on Machine Learning (ICML)*. 2016, pp. 699–707.

[ASS17] Y. Arjevani, O. Shamir, and R. Shiff. “Oracle complexity of second-order methods for smooth convex optimization”. In: *Mathematical Programming* (2017), pp. 1–34.

[ASS18] Y. Arjevani, O. Shamir, and R. Shiff. “Oracle complexity of second-order methods for smooth convex optimization”. In: *Mathematical Programming* (2018), pp. 1–34.

[AMLJG19] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. “A Tight and Unified Analysis of Extragradient for a Whole Spectrum of Differentiable Games”. In: *arXiv preprint arXiv:1906.05945* (2019).

[Azi+20] W. Azizian, D. Scieur, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. “Accelerating Smooth Games by Manipulating Spectral Shapes”. In: *arXiv preprint arXiv:2001.00602* (2020).

[Bae09] M. Baes. “Estimate sequence methods: extensions and approximations”. In: (2009).

[BGP19] J. P. Bailey, G. Gidel, and G. Piliouras. “Finite Regret and Cycles with Fixed Step-Size via Alternating Gradient Descent-Ascent”. In: *arXiv preprint arXiv:1907.04392* (2019).

[BP19a] J. P. Bailey and G. Piliouras. “Fast and Furious Learning in Zero-Sum Games: Vanishing Regret with Non-Vanishing Step Sizes”. In: *Neural Information Processing Systems, NeurIPS 2019, Vancouver, Canada*. 2019.

[BP19b] J. P. Bailey and G. Piliouras. “Multi-agent learning in network zero-sum games is a Hamiltonian system”. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2019, pp. 233–241.

[Bal+18] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. “The Mechanics of n-Player Differentiable Games”. In: *International Conference on Machine Learning (ICML)*. 2018.

[Bla56] D. Blackwell. “An analog of the minimax theorem for vector payoffs”. In: *Pacific Journal of Mathematics* 6.1 (1956), pp. 1–8.

[BFH10] F. Brandt, F. Fischer, and P. Harrenstein. “On the rate of convergence of fictitious play”. In: *International Symposium on Algorithmic Game Theory*. Springer. 2010, pp. 102–113.

[Bro49] G. W. Brown. *Some notes on computation of games solutions*. Tech. rep. RAND CORP Santa Monica, CA, 1949.

[Bro51] G. W. Brown. "Iterative solution of games by fictitious play". In: *Activity analysis of production and allocation* 13.1 (1951), pp. 374–376.

[Bub+18] S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford. "Near-optimal method for highly smooth convex optimization". In: *arXiv preprint arXiv:1812.08026* (2018).

[Bub+19] S. Bubeck, Q. Jiang, Y.-T. Lee, Y. Li, and A. Sidford. "Complexity of Highly Parallel Non-Smooth Convex Optimization". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13900–13909.

[Bul18] B. Bullins. "Fast minimization of structured convex quartics". In: *arXiv preprint arXiv:1812.10349* (2018).

[BL19] B. Bullins and K. A. Lai. "Higher-order methods for min-max optimization". In: (2019).

[CDHS18] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "Accelerated methods for nonconvex optimization". In: *SIAM Journal on Optimization* 28.2 (2018), pp. 1751–1772.

[CHDS17] Y. Carmon, O. Hinder, J. C. Duchi, and A. Sidford. ""Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions". In: *International Conference on Machine Learning (ICML)*. 2017.

[Dan81] J. M. Danskin. "Fictitious play for continuous games revisited". In: *International Journal of Game Theory* 10.3 (Sept. 1981), pp. 147–154. URL: https://doi.org/10.1007/BF01755961.

[Dan51] G. B. Dantzig. "A proof of the equivalence of the programming problem and the game problem". In: *Activity Analysis of Production and Allocation* (1951). Ed. by T. Koopmans, pp. 330–335.

[DISZ18] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. "Training GANs with Optimism". In: *International Conference on Learning Representations (ICLR)*. 2018.

[DP14] C. Daskalakis and Q. Pan. "A counter-example to Karlin's strong conjecture for fictitious play". In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE. 2014, pp. 11–20.

[DP18] C. Daskalakis and I. Panageas. “The Limit Points of (Optimistic) Gradient Descent in Min-Max Optimization”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 9255–9265.

[DH19] S. S. Du and W. Hu. “Linear Convergence of the Primal-Dual Gradient Method for Convex-Concave Saddle Point Problems without Strong Convexity”. In: *Artificial Intelligence and Statistics (AISTATS)*. 2019.

[FW56] M. Frank and P. Wolfe. “An algorithm for quadratic programming”. In: *Naval research logistics quarterly* 3.1-2 (1956), pp. 95–110.

[FS96] Y. Freund and R. E. Schapire. “Game theory, on-line prediction and boosting”. In: *Conference on Learning Theory (COLT)*. 1996, pp. 325–332.

[FS99] Y. Freund and R. E. Schapire. “Adaptive Game Playing Using Multiplicative Weights”. In: *Games and Economic Behavior* 29.1-2 (Oct. 1999), pp. 79–103.

[Gas+18] A. Gasnikov, P. Dvurechensky, E. Gorbunov, D. Kovalev, A. Mohhamed, E. Chernousova, and C. A. Uribe. “The global rate of convergence for optimal tensor methods in smooth convex optimization”. In: *arXiv preprint arXiv:1809.00382 (v10)* (2018).

[GM18] I. Gemp and S. Mahadevan. “Global convergence to the equilibrium of gans using variational inequalities”. In: *arXiv preprint arXiv:1808.01531* (2018).

[GL16] S. Ghadimi and G. Lan. “Accelerated gradient methods for nonconvex nonlinear and stochastic programming”. In: *Mathematical Programming* 156.1-2 (2016), pp. 59–99.

[GBVLJ19] G. Gidel, H. Berard, P. Vincent, and S. Lacoste-Julien. “A Variational Inequality Perspective on Generative Adversarial Nets”. In: *International Conference on Learning Representations (ICLR)* (2019).

[Goo+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2014, pp. 2672–2680.

[Han57] J. Hannan. “Approximation to Bayes risk in repeated play”. In: *Contributions to the Theory of Games* 3 (1957), pp. 97–139.

[Har98] C. Harris. “On the rate of convergence of continuous-time fictitious play”. In: *Games and Economic Behavior* 22.2 (1998), pp. 238–259.

[HS02] J. Hofbauer and W. H. Sandholm. “On the global convergence of stochastic fictitious play”. In: *Econometrica* 70.6 (2002), pp. 2265–2294.

[HRU13] J. Hsu, A. Roth, and J. Ullman. “Differential privacy for the analyst via private equilibrium computation”. In: *Symposium on Theory of Computing (STOC)*. 2013, pp. 341–350.

[JWZ18] B. Jiang, H. Wang, and S. Zhang. “An optimal high-order tensor method for convex optimization”. In: *arXiv preprint arXiv:1812.06557* (2018).

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. “Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2016, pp. 795–811.

[Kar59] S. Karlin. *Mathematical Methods and Theory in Games, Programming, and Economics*. Addison-Wesley, 1959.

[KALL18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. “Progressive growing of gans for improved quality, stability, and variation”. In: *International Conference on Learning Representations (ICLR)* (2018).

[KS80] D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*. Vol. 31. Siam, 1980.

[Kom99] S. Komlósi. “On the Stampacchia and Minty variational inequalities”. In: *Generalized Convexity and Optimization for Economic and Financial Decisions* (1999), pp. 231–260.

[Kor76] G Korpelevich. “The extragradient method for finding saddle points and other problems”. In: *Ekonomika i Matematicheskie Metody v. 12* (1976), pp. 747–756.

[Kro19] C. Kroer. “First-Order Methods with Increasing Iterate Averaging for Solving Saddle-Point Problems”. In: *arXiv preprint arXiv:1903.10646* (2019).

[Let+19] A. Letcher, J. Foerster, D. Balduzzi, T. Rocktäschel, and S. Whiteson. “Stable Opponent Shaping in Differentiable Games”. In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=SyGjjsC5tQ.

[LS19] T. Liang and J. Stokes. “Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks”. In: *Artificial Intelligence and Statistics (AISTATS)* (2019).

[Loj63] Lojasiewicz. “A topological property of real analytic subsets (in French)”. In: *Coll. du CNRS, Les équations aux dérivées partielles* (1963), 8789.

[Mad+18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. “Towards Deep Learning Models Resistant to Adversarial Attacks”. In: (2018).

[MJS19] E. V. Mazumdar, M. I. Jordan, and S. S. Sastry. “On Finding Local Nash Equilibria (and Only Local Nash Equilibria) in Zero-Sum Games”. In: *arXiv preprint arXiv:1901.00838* (2019).

[MPP18] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. “Cycles in adversarial regularized learning”. In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2018, pp. 2703–2717.

[Mer+19] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. “Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[MGN18] L. Mescheder, A. Geiger, and S. Nowozin. “Which training methods for GANs do actually Converge?” In: *International Conference on Machine Learning (ICML)*. 2018, pp. 3478–3487.

[MNG17] L. Mescheder, S. Nowozin, and A. Geiger. “The numerics of GANs”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 1825–1835.

[Min+62] G. J. Minty et al. “Monotone (nonlinear) operators in Hilbert space”. In: *Duke Mathematical Journal* 29.3 (1962), pp. 341–346.

[Miy61] K. Miyasawa. *On the convergence of the learning process in a 2 x 2 non-zero-sum two-person game*. Tech. rep. Princeton University, 1961.

[MOP19] A. Mokhtari, A. Ozdaglar, and S. Pattathil. “A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach”. In: *arXiv preprint arXiv:1901.08511* (2019).

[MS96] D. Monderer and A. Sela. “A 2×2 game without the fictitious play property”. In: *Games and Economic Behavior* 14.1 (1996), pp. 144–148.

[MS12] R. D. Monteiro and B. F. Svaiter. “Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems”. In: *SIAM Journal on Optimization* 22.3 (2012), pp. 914–935.

[MS10] R. D. Monteiro and B. F. Svaiter. “On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean”. In: *SIAM Journal on Optimization* 20.6 (2010), pp. 2755–2787.

[MS13] R. D. Monteiro and B. F. Svaiter. “An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods”. In: *SIAM Journal on Optimization* 23.2 (2013), pp. 1092–1125.

[Nem04] A. Nemirovski. “Prox-method with rate of convergence O (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems”. In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251.

[NY83] A. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in Optimization*. J. Wiley & Sons, 1983.

[Nes08] Y. Nesterov. “Accelerating the cubic regularization of Newtons method on convex problems”. In: *Mathematical Programming* 112.1 (2008), pp. 159–181.

[Nes83] Y. E. Nesterov. “A method for solving the convex programming problem with convergence rate $O(1/k^2)$”. In: *Soviet Mathematics Doklady*. Vol. 269. 1983, pp. 543–547.

[NN94] Y. Nesterov and A. Nemirovski. *Interior-point polynomial algorithms in convex programming*. Vol. 13. Siam, 1994.

[NP06] Y. Nesterov and B. T. Polyak. “Cubic regularization of Newton method and its global performance”. In: *Mathematical Programming* 108.1 (2006), pp. 177–205.

[Neu28] J. von Neumann. “Zur theorie der gesellschaftsspiele”. In: *Mathematische annalen* 100.1 (1928), pp. 295–320.

[OS11] G. Ostrovski and S. van Strien. “Piecewise linear Hamiltonian flows associated to zero-sum games: Transition combinatorics and questions on ergodicity”. In: *Regular and Chaotic Dynamics* 16.1 (Feb. 2011), pp. 128–153.

[OS14] G. Ostrovski and S. van Strien. “Payoff performance of fictitious play”. In: *Journal of Dynamics and Games* 1.4 (2014), pp. 621–638.

[OX18] Y. Ouyang and Y. Xu. “Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems”. In: *arXiv preprint arXiv:1808.02901* (2018).

[Pea94] B. A. Pearlmutter. “Fast exact multiplication by the Hessian”. In: *Neural computation* 6.1 (1994), pp. 147–160.

[Pol63] B. T. Polyak. “Gradient methods for minimizing functionals (in Russian)”. In: *Zh. Vychisl. Mat. Mat. Fiz.* (1963), 643653.

[Rob51] J. Robinson. “An iterative method of solving a game”. In: *Annals of mathematics* (1951), pp. 296–301.

[Roc76] R. T. Rockafellar. “Monotone operators and the proximal point algorithm”. In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898.

[SA19] F. Schäfer and A. Anandkumar. “Competitive gradient descent”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 7623–7633.

[Sha64] L. S. Shapley. “Some topics in two-person games”. In: *Advances in Game Theory* (1964), p. 1.

[Sta70] G. Stampacchia. “Variational inequalities”. In: *Congrès international des mathématiciens*. 1970.

[SK17] B. Swenson and S. Kar. “On the exponential rate of convergence of fictitious play in potential games”. In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2017, pp. 275–279.

[SKXL17] B. Swenson, S. Kar, J. Xavier, and D. S. Leslie. “Robustness properties in fictitious-play-type algorithms”. In: *SIAM Journal on Control and Optimization* 55.5 (2017), pp. 3295–3318.

[Tse95] P. Tseng. “On linear convergence of iterative methods for the variational inequality problem”. In: *Journal of Computational and Applied Mathematics* 60.1-2 (1995), pp. 237–252.

[Tse08] P. Tseng. “On accelerated proximal gradient methods for convex-concave optimization”. In: (2008).

[Vin+19] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. In: *Nature* (2019), pp. 1–5.

[WA18] J.-K. Wang and J. D. Abernethy. “Acceleration through optimistic no-regret dynamics”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 3824–3834.

[Yaz+19] Y. Yazıcı, C.-S. Foo, S. Winkler, K.-H. Yap, G. Piliouras, and V. Chandrasekhar. “The unusual effectiveness of averaging in GAN training”. In: *International Conference on Learning Representations (ICLR)* (2019).