# Near-optimal Local Convergence of Alternating Gradient Descent-Ascent for Minimax Optimization

**Guodong Zhang**
University of Toronto

**Yuanhao Wang**
Princeton University

**Laurent Lessard**
Northeastern University

**Roger Grosse**
University of Toronto

## Abstract

Smooth minimax games often proceed by simultaneous or alternating gradient updates. Although algorithms with alternating updates are commonly used in practice, the majority of existing theoretical analyses focus on simultaneous algorithms for convenience of analysis. In this paper, we study alternating gradient descent-ascent (Alt-GDA) in minimax games and show that Alt-GDA is superior to its simultaneous counterpart (Sim-GDA) in many settings. We prove that Alt-GDA achieves a near-optimal local convergence rate for strongly convex-strongly concave (SCSC) problems while Sim-GDA converges at a much slower rate. To our knowledge, this is the *first* result of any setting showing that Alt-GDA converges faster than Sim-GDA by more than a constant. We further adapt the theory of integral quadratic constraints (IQC) and show that Alt-GDA attains the same rate *globally* for a subclass of SCSC minimax problems. Empirically, we demonstrate that alternating updates speed up GAN training significantly and the use of optimism only helps for simultaneous algorithms.

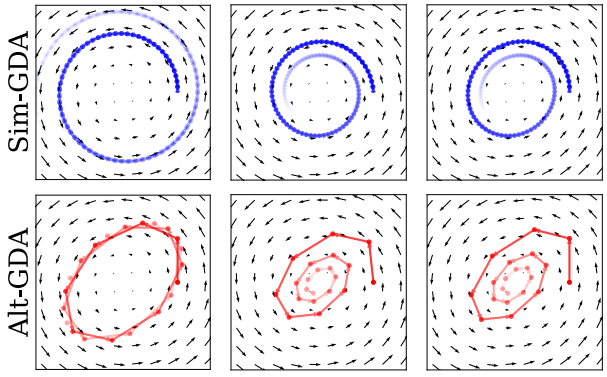



**Figure 1: Left:** a simple bilinear game with function $f(x, y) = 10xy$; **Middle:** a SCSC minimax game with $f(x, y) = 0.5x^2 + 10xy - 0.5y^2$. **Right:** a minimax game that is not strongly-convex in $x$ with $f(x, y) = 10xy - y^2$; Color from dark to light indicates the direction of the trajectory. **Key Observation:** Alt-GDA converges much faster than Sim-GDA when they both converge.

## 1 INTRODUCTION

Since the seminal work of von Neumann (von Neumann, 1928), minimax optimization in the form of $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ has been a major focus of research in mathematics, economics and computer science (von Neumann and Morgenstern, 1944; Başar and Olsder, 1998; Roughgarden, 2010). Recently, minimax optimization has gained tremendous attention in machine learning as it offers a flexible paradigm that goes beyond ordinary loss function minimization. In particular, there is an increasing set of models that can be formulated as minimax problems, including (but not limited to) generative adversarial networks (Goodfellow et al., 2014; Arjovsky et al., 2017), adversarial training (Madry et al., 2018), robust optimization (Ben-Tal et al., 2009) and primal-dual reinforcement learning (Du et al., 2017; Yang et al., 2020c).

The most natural and frequently used method for solving minimax problems is a generalization of gradient descent known as gradient descent-ascent (GDA), with either simultaneous or alternating updates of the two players, referred to as Sim-GDA and Alt-GDA, respectively, throughout the sequel. Unlike gradient descent, which converges to a local minimum for minimization problems under a broad range of conditions (Lee et al., 2016, 2017), it is known that GDA with constant step-sizes can fail to converge for general smooth functions (Mescheder et al., 2017), even for unconstrained bilinear games (Gidel et al., 2019b; Bailey and Piliouras, 2018). Even when it does converge, GDA may

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. 

exhibit rotational behaviors (Mescheder et al., 2017; Letcher et al., 2019; Schaefer and Anandkumar, 2019) and hence converge slowly (see Figure 1). To combat these issues, several algorithms have been introduced specifically for smooth minimax games, including consensus optimization (Mescheder et al., 2017), symplectic gradient adjustment (Letcher et al., 2019), negative momentum (NM) (Gidel et al., 2019b; Zhang and Wang, 2021), optimistic gradient descent-ascent (OGDA) (Popov, 1980; Rakhlin and Sridharan, 2013; Daskalakis et al., 2018; Mertikopoulos et al., 2019) and extra-gradient (EG) (Korpelevich, 1976).

In theory, many of these algorithms enjoy improved convergence rates compared to GDA. In particular, both OGDA and EG are near-optimal for SCSC minimax problems (Mokhtari et al., 2020b). However, in practice, GDA and its adaptive variants are still the go-to algorithms for many applications (e.g., GAN optimization and offline policy evaluation (Yang et al., 2020c)). Here, the catch is that the overwhelming majority of existing theoretical analyses focus on simultaneous algorithms where players update their strategies at the same time, as simultaneous updates are easier to analyze and can often be formulated as solving a variational inequality problem (Harker and Pang, 1990; Gidel et al., 2019a; Zhang et al., 2021). This is in stark contrast to our common practice where alternating algorithms are actually used. Nonetheless, our understanding of alternating algorithms in minimax optimization is severely limited to simple bilinear games. Despite it being a very natural question to ask, the convergence properties of Alt-GDA for SCSC minimax games and many other settings remain largely unknown. The key difficulty is that every iteration of an alternating algorithm is a composition of two half updates, which greatly complicates analysis.

**Our contributions.** In this paper, we take a step towards understanding Alt-GDA and closing the gap between theory and practice. We first revisit the convergence properties of Alt-GDA in bilinear games for completeness. We then discuss our main contributions on proving near-optimal convergence rates of Alt-GDA. In more detail:

1. We prove that, for SCSC minimax games[1], Alt-GDA achieves an iteration complexity of $\mathcal{O}(\kappa)$ locally ($\kappa$ is the condition number), which is quadratically better than the $\mathcal{O}(\kappa^2)$ bound for Sim-GDA and even matches EG/OGDA. Importantly, the complexity bound for Alt-GDA in this setting is near-optimal as it matches the coarse lower bound in (Azizian et al., 2020b, Corollary 1).
2. We further prove that both Sim-GDA and Alt-GDA attain linear convergence when the minimax problem has only strong concavity in $\mathbf{y}$ but no strong convexity in $\mathbf{x}$ by assuming non-singularity of the coupling matrix.
3. We show that Alt-GDA can converge with the same rate $\mathcal{O}(\kappa)$ globally for a class of SCSC minimax games with a bilinear coupling term. This is done by using theory of IQC to automatically search for a Lyapunov function.
4. Lastly, we validate our theory on quadratic minimax games. Empirically, we demonstrate that alternating updates could speed up GAN training dramatically (which matches the existing results in Goodfellow et al. (2014); Radford et al. (2015)) and perform on par with optimistic updates though GAN objective is generally nonconvex-nonconcave.

[1]The SCSC setting is fundamental. Via reduction (Lin et al., 2020; Yang et al., 2020b), an efficient algorithm for this setting implies efficient algorithms for other settings, including strongly convex-concave, convex-concave, and non-convex-concave settings.

## 2 PRELIMINARIES

**Notation**. In this paper, scalars are denoted by lower-case letters (e.g., $\lambda$), vectors by lower-case bold letters (e.g., $\mathbf{x}$), matrices by upper-case bold letters (e.g., $\mathbf{J}$). The spectrum of a square matrix $\mathbf{A}$ is denoted by $\mathrm{Sp}(\mathbf{A})$, and a generic eigenvalue by $\lambda$. We respectively note $\sigma_{\min}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$ the smallest and the largest positive singular values of $\mathbf{A}$. For matrix inequality $\mathbf{A} \succeq \mathbf{B}$, we mean $\mathbf{A} - \mathbf{B}$ is positive semi-definite. We use $\Re$ and $\Im$ to denote the real part and imaginary part of a complex scalar respectively. We use $\mathbb{R}$ and $\mathbb{C}$ to denote the set of real numbers and complex numbers, respectively. We use $\rho(\mathbf{A}) = \lim_{t\to\infty} \|\mathbf{A}^t\|^{1/t}$ to denote the spectral radius of matrix $\mathbf{A}$. $\mathcal{O}$, $\Omega$ and $\Theta$ are standard asymptotic notations.

### 2.1 Two-player Minimax Games

We begin by presenting the fundamental two-player zero-sum game that we will consider in the sequel. To be specific, our problem of interest is the following unconstrained minimax optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x}, \mathbf{y}). \tag{1}$$

We are usually interested in finding a *Nash equilibrium* (von Neumann and Morgenstern, 1944): a set of parameters from which no player can (unilaterally) improve its objective function. In this work, we focus on the case of $f$ being a convex-concave and smooth function. Here we state the assumption formally.

**Assumption 1.** *The function $f$ is continuously differentiable and $L$-smooth in $\mathbf{x}$ and $\mathbf{y}$. Furthermore, we assume $f$ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$.*

For completeness, we state the definition of smooth function. We note that the smoothness assumption is standard for convergence analysis in the literature.

**Definition 1.** *A differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ is $L$-smooth if it has $L$-Lipschitz gradient on $\mathbb{R}^d$, i.e., for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$, we have $\|\nabla\phi(\mathbf{x}_1) - \nabla\phi(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$.*

One of the nice properties of working with convex-concave problems is that there often exists at least one *global* Nash equilibrium $(\mathbf{x}^*, \mathbf{y}^*)$ such that for any $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ we have

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq f(\mathbf{x}, \mathbf{y}^*).$$

### 2.2 Gradient Descent-Ascent Family

We now present two algorithms we will discuss in this paper, Sim-GDA and Alt-GDA. The de-facto standard algorithm for finding Nash equilibria of general smooth two-player minimax games is simultaneous gradient descent-ascent (Sim-GDA) which is a direct generalization of gradient descent to minimax games. In particular, it updates both players $\mathbf{x}$ and $\mathbf{y}$ simultaneously:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \\ \mathbf{y}_{t+1} &= \mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \end{aligned} \tag{2}$$

where $\eta$ is the step size[2]. Succinctly, Sim-GDA updates (2) can be defined as the repeated application of a nonlinear operator in the form of $\mathbf{z}_{t+1} = F_\eta^{\mathrm{Sim}}(\mathbf{z}_t) \triangleq \mathbf{z}_t - \eta V(\mathbf{z}_t)$ with $\mathbf{z} = [\mathbf{x}^\top, \mathbf{y}^\top]^\top$ and $V(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})^\top, -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})^\top]^\top$, the gradient vector field. By contrast, Alt-GDA takes advantage of the fact that the iterates $\mathbf{x}_{t+1}$ and $\mathbf{y}_{t+1}$ are computed sequentially:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \\ \mathbf{y}_{t+1} &= \mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t). \end{aligned} \tag{3}$$

Similarly, we can write the updates as $\mathbf{z}_{t+1} = F_\eta^{\mathrm{Alt}}(\mathbf{z}_t)$.

### 2.3 Local Convergence Rates

We stress that local convergence analysis has been widely adopted in smooth game optimization (see e.g., Gidel et al. (2019b); Wang et al. (2019); Azizian et al. (2020b); Zhang and Wang (2021); Liang and Stokes (2019); Fiez and Ratliff (2021)). Under certain conditions on a fixed point operator $F$, linear convergence is guaranteed in a neighborhood around a fixed point $\mathbf{z}^*$ (i.e. local convergence).

[2]Using separate step sizes for two players does not improve the worst-case convergence rate.

**Theorem 1** ((Bertsekas, 1997, Proposition 4.4.1))**.** *For a continuously differentiable nonlinear operator $F$ with the fixed point $\mathbf{z}^*$, if the spectral radius $\rho_F \triangleq \rho(\nabla F(\mathbf{z}^*)) < 1$, then for any $\mathbf{z}_0$ in a neighborhood of $\mathbf{z}^*$, the iterates of $\mathbf{z}_t$ converge to $\mathbf{z}^*$ with a linear rate of $\mathcal{O}((\rho_F + \epsilon)^t)$ for any $\epsilon > 0$.*

With this theorem, one can obtain local convergence rate of an algorithm by just computing the spectral radius of $\nabla F_\eta(\mathbf{z}^*)$, which is a constant matrix depending on $\eta$ in our setting. In the paper, we focus on the *worst-case* convergence rate which is defined (up to a $\epsilon$ difference) as follows:

$$\min_{\eta} \max_{F \in \mathcal{M}} \rho(\nabla F_\eta(\mathbf{z}^*)), \tag{4}$$

where the inner maximization is over all possible instances within the whole problem class $\mathcal{M}$.

### 2.4 Revisiting Alt-GDA for Bilinear Games

In this section, we revisit the unconstrained bilinear games (Gidel et al., 2019b; Daskalakis and Panageas, 2018; Liang and Stokes, 2019; Mokhtari et al., 2020a) for which Sim-GDA diverges with any finite step size. Formally, the bilinear game is given by

$$\min_{\mathbf{x} \in \mathbb{R}^m} \max_{\mathbf{y} \in \mathbb{R}^n} \mathbf{x}^\top \mathbf{B} \mathbf{y}, \tag{5}$$

where we ignore the linear terms without loss of generality. Here, the Nash equilibrium is $(\mathbf{x}^*, \mathbf{y}^*)$ satisfying $\mathbf{B}^\top \mathbf{x}^* = \mathbf{0}$ and $\mathbf{B}\mathbf{y}^* = \mathbf{0}$. To measure convergence, one could monitor the distance to the equilibrium:

$$\Delta_t = \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \|\mathbf{y}_t - \mathbf{y}^*\|_2^2. \tag{6}$$

We aim to understand the difference between the dynamics of simultaneous and alternating methods. Practitioners have been widely using the latter instead of the former when optimizing GANs despite the rich optimization literature on simultaneous methods.

For Sim-GDA, the eigenvalues of $\mathbf{I} - \nabla F_\eta^{\mathrm{Sim}}$ are all pure imaginary. As a result, we have the spectral radius as $\rho(\nabla F_\eta^{\mathrm{Sim}}) = 1 + \eta^2 \sigma_{\max}^2(\mathbf{B})$. Therefore, we have

**Theorem 2** (Gidel et al. (2019b))**.** *For any $\eta > 0$, the iterates of Sim-GDA diverges as*

$$\Delta_t \in \Omega\left(\Delta_0 (1 + \eta^2 \sigma_{max}^2(\mathbf{B}))^t\right)$$

This theorem states that the iterates of Sim-GDA diverge linearly for any positive constant step-size $\eta$. By contrast, the iterates of Alt-GDA stay bounded due to the sequential update rule which significantly shifts the eigenvalues of the Jacobian. Specifically, the eigenvalues of $\nabla F_\eta^{\mathrm{Alt}}$ are roots of the polynomial $(x-1)^2 + \eta^2 \lambda x$ with $\lambda \in \mathrm{Sp}(\mathbf{B}^\top \mathbf{B})$. As a consequence, the spectral radius of $\nabla F_\eta^{\mathrm{Alt}}$ is upper bounded by 1 for some $\eta$ and hence the iterates of Alt-GDA stays bounded.

**Theorem 3.** *For any $0 < \eta \leq \frac{2}{\sigma_{max}(\mathbf{B})}$, the iterates of Alt-GDA stay bounded*

$$\Delta_t \in \mathcal{O}(\Delta_0)$$

Similar results can be found in the literature (see e.g., Gidel et al. (2019b); Zhang and Yu (2020)). In addition, one can show that for bilinear games, Alt-GDA is a symplectic integrator applied on the continuous dynamics (Bailey et al., 2020), which preserves energy and volume.

## 3 NEAR-OPTIMAL LOCAL CONVERGENCE IN SCSC SETTING

Bilinear games, as discussed previously, are somewhat simplistic in that they obey a conservation law and can be easily solved by performing gradient descent on the Hamiltonian (Letcher et al., 2019; Azizian et al., 2020b). In this section, we consider a different class of games whose Jacobian has both symmetric and antisymmetric components, and are therefore arguably harder to solve. In particular, we assume $f(\mathbf{x}, \mathbf{y})$ is SCSC and smooth, which implies

$$\mu_{\mathbf{x}}\mathbf{I} \preceq \nabla^2_{\mathbf{x}} f \preceq L_{\mathbf{x}}\mathbf{I},\ \mu_{\mathbf{y}}\mathbf{I} \preceq -\nabla^2_{\mathbf{y}} f \preceq L_{\mathbf{y}}\mathbf{I},$$
$$\|\nabla^2_{\mathbf{xy}} f\|_2 \leq L_{\mathbf{xy}}.$$

We let $L \triangleq \max\{L_{\mathbf{x}}, L_{\mathbf{y}}, L_{\mathbf{xy}}\}$ and $\mu \triangleq \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}$ and define the condition number $\kappa \triangleq L/\mu$. Accordingly, one can define $\kappa_{\mathbf{x}} \triangleq L/\mu_{\mathbf{x}}$ and $\kappa_{\mathbf{y}} \triangleq L/\mu_{\mathbf{y}}$. We now briefly summarize some known results about convergence of Sim-GDA in this setting. The worst-case convergence rate (4) of Sim-GDA reduces to $\rho(\mathbf{I} - \eta\nabla V(\mathbf{z}^*))$, which is equivalent to

$$\min_{\eta} \max_{\lambda \in \mathcal{K}} |1 - \eta\lambda|, \tag{7}$$

where $\mathcal{K}$ is the support of the eigenvalues of the Jacobian of the gradient vector field $V$. It can be shown that $\mathcal{K} = \left\{\lambda \in \mathbb{C} : |\lambda| \leq \sqrt{2}L, \Re\lambda \geq \mu > 0\right\}$ (see Appendix A.2). This set is the intersection between a circle and a halfplane (Azizian et al., 2020b). Eqn. (7) leaves open the choice of $\eta$, and it is known that the presence of large imaginary eigenvalues of the Jacobian forces a small value of $\eta$, thereby limiting the rate of convergence (Mescheder et al., 2017). We summarize the result below:

**Theorem 4.** *With the step size $\eta = \frac{\mu}{2L^2}$, we have $\rho(\nabla F_\eta^{\text{Sim}}(\mathbf{z}^*)) < 1 - \frac{1}{4\kappa^2}$. Hence, Sim-GDA converges locally at a linear rate $\mathcal{O}\left(\left(1 - \frac{1}{4\kappa^2}\right)^t\right)$.*

This theorem suggests that Sim-GDA converges to the equilibrium linearly with an iteration complexity of $\mathcal{O}(\kappa^2)$, which is known to be tight (Azizian et al., 2020b) but much slower than the $\mathcal{O}(\kappa)$ iteration complexity of extra-gradient (EG) or optimistic gradient-descent-ascent (OGDA) (Gidel et al., 2019a; Mokhtari et al., 2020a; Azizian et al., 2020a; Zhang et al., 2021).

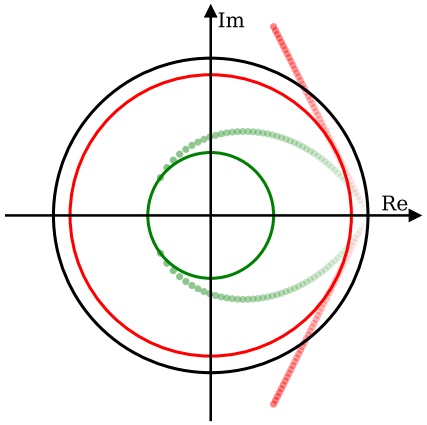



**Figure 2:** Eigenvalues of $\nabla F_\eta^{\text{Sim}}$ and $\nabla F_\eta^{\text{Alt}}$ for a minimax problem with the function $f(x, y) = 0.3x^2 + 1.2xy - 0.3y^2$. For a fixed step-size $\eta$, eigenvalues of Sim-GDA are represented with **red dots** and eigenvalues of Alt-GDA are **green dots**. Their trajectories as $\eta$ sweeps in $[0, 1]$ are shown from light colors to dark colors. Convergence circles for Sim-GDA are in **red**, Alt-GDA in **green**, and unit circle in **black**. The convergence circles are optimized over all step-sizes. Alternating updates help as its convergence circle (green) is smaller, due to the fact that it allows us to use much larger step-sizes. Figure inspired by Gidel et al. (2019b).

To understand why, we note the maximization over $\lambda$ in (7) is attained by $\lambda = \mu + \sqrt{2L^2 - \mu^2}i$, which has a large imaginary component. It is easy to show (see e.g., Mescheder et al. (2017)) that the largest feasible step size in (7) is inversely proportional to $(|\lambda|/\Re(\lambda))^2$. Hence, the step size has to be extremely small in the presence of eigenvalues with large imaginary parts, which in turn, leads to slow convergence. In a nutshell, the culprits of slow convergence in Sim-GDA are eigenvalues of the Jacobian of the associated vector field $V$ with large imaginary parts. We stress that eigenvalues with large imaginary components contribute to a strong "rotational force". To improve convergence, many algorithms have been introduced to suppress the rotational force, including EG, OGDA, and NM. Indeed, all three of these algorithms improve the convergence rate by some margin in theory. Nevertheless, GDA (or its adaptive variant) is still the go-to algorithm in practice.

We believe the reason these alternative algorithms haven't been adopted widely is that practical algorithms for cases such as GANs are typically based on Alt-GDA rather than Sim-GDA. Surprisingly, despite the popularity of Alt-GDA, its convergence properties haven't been analyzed in this setting. While it is per-

haps intuitive that Alt-GDA should perform better than Sim-GDA due to its use of fresher gradient information, we show that, in fact, Alt-GDA achieves a quadratic speedup over Sim-GDA locally and matches the convergence rate of EG and OGDA.

Local convergence analyses of Sim-GDA, EG and OGDA are based on matrix spectral calculation, and in principle one can apply this to Alt-GDA as well. However, bounding the spectral radius is much harder for Alt-GDA, since the algorithm involves two half steps, and the spectral radius of the matrix product can't be bounded straightforwardly in terms of the spectral radii of the two factors. This is likely why the convergence rate of Alt-GDA remained unknown. By treating complex eigenvalues differently and adopting a fined-grained analysis, we arrive at the following bounds for the eigenvalues of $\nabla F_\eta^{\text{Alt}}(\mathbf{z}^*)$:

**Theorem 5.** *With the step size $\eta \le \frac{1}{2L}$, the eigenvalues of $\nabla F_\eta^{\text{Alt}}(\mathbf{z}^*)$ satisfy*

$$\text{if real: } |\lambda| \le \max\{1-\eta\mu_{\mathbf{x}}, 1-\eta\mu_{\mathbf{y}}\},$$
$$\text{if complex: } |\lambda| \le \sqrt{(1-\eta\mu_{\mathbf{x}})(1-\eta\mu_{\mathbf{y}})}.$$

**Remark 1.** *In stark contrast to Sim-GDA, for which the complex eigenvalues of $\nabla F_\eta^{\text{Sim}}$ can have magnitude as large as $\sqrt{1-2\eta\mu+2\eta^2L^2}$, the complex eigenvalues of $\nabla F_\eta^{\text{Alt}}$ are much smaller in magnitude and are even smaller than the real eigenvalues as shown in Theorem 5. As a result, we are allowed to use a larger step size, which gives an improved convergence rate (see Figure 2 for details).*

Following immediately from Theorem 5, we have the following Corollary.

**Corollary 1.** *With $\eta = \frac{1}{2L}$, we have $\rho(\nabla F_\eta^{\text{Alt}}(\mathbf{z}^*)) \le 1 - \frac{1}{2\kappa}$. Hence by Theorem 1, Alt-GDA converges locally at a linear rate $\mathcal{O}\left(\left(1-\frac{1}{2\kappa}+\epsilon\right)^t\right)$ with $\epsilon > 0$ an arbitrarily small constant.*

In particular, this corollary suggests that the iteration complexity of Alt-GDA matches the coarse lower iteration complexity bound[3] $\Omega(\kappa)$ (Azizian et al., 2020b, Corollary 1) up to a constant, implying Alt-GDA is near-optimal (at least locally). This is the *first* time that one can rigorously show the Alt-GDA converges faster than Sim-GDA by more than a constant, let alone quadratically faster.

Furthermore, it implies that the convergence rate of Alt-GDA is no worse than its rate for pure cooperative games with $\mathbf{B} \triangleq \nabla^2_{\mathbf{xy}} f = \mathbf{0}$. Put differently, the adversarial component (the existence of coupling matrix $\mathbf{B}$) does *not* make the optimization any harder for Alt-GDA. We remark that this is *not* true for Sim-GDA because in that case, the coupling matrix $\mathbf{B}$ introduces complex eigenvalues with large imaginary parts, which slow down convergence.

[3]The fine-grained bound (Zhang et al., 2019b) is $\Omega(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$. One could achieve this bound by using a accelerated proximal point framework (Yang et al., 2020b) with Alt-GDA in the inner-loop.

## 4 ACCELERATION WITHOUT STRONG CONVEXITY

We have shown that Alt-GDA achieves a near-optimal local convergence rate for SCSC minimax games. In this section, we further consider the case that has only strong concavity in the player $\mathbf{y}$ but *no* strong convexity in $\mathbf{x}$. In particular, it is equivalent to assuming

$$\mathbf{0} \preceq \nabla^2_{\mathbf{x}} f \preceq L_{\mathbf{x}}\mathbf{I},\ \mu_{\mathbf{y}}\mathbf{I} \preceq -\nabla^2_{\mathbf{y}} f \preceq L_{\mathbf{y}}\mathbf{I},$$
$$\|\nabla^2_{\mathbf{xy}} f\|_2 \le L_{\mathbf{xy}}.$$

This setting was investigated in empirical policy evaluation where no strong convex regularization is applied on the primal variables (Du et al., 2017). They showed that the non-singularity of the coupling matrix $\mathbf{B} \triangleq \nabla^2_{\mathbf{xy}} f(\mathbf{x}^*, \mathbf{y}^*)$ can help achieve linear convergence for Sim-GDA. Technically, the coupling matrix $\mathbf{B}$ has to be full-row rank (i.e., $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top) > 0$) and we simply assume $\mu_{\mathbf{xy}} \triangleq \sigma_{\min}(\mathbf{B}) > 0$. Then for Sim-GDA, we have the eigenvalues of its Jacobian as follows:

**Theorem 6.** *Let $\eta \le \frac{1}{L}$, the eigenvalues of $\nabla F_\eta^{\text{Sim}}(\mathbf{z}^*)$ satisfy the following bound*

$$\text{if real: } |\lambda| \le \max\left\{1-\tfrac{\eta}{L}\mu_{\mathbf{xy}}^2, 1-\eta\mu_{\mathbf{y}}\right\},$$
$$\text{if complex: } |\lambda| \le \sqrt{1-\eta\mu_{\mathbf{y}}+2\eta^2L^2}.$$

To be noted, our eigenvalue bounds in Theorem 6 are slightly different from that in Du et al. (2017) as they allow step size separation for player $\mathbf{x}$ and $\mathbf{y}$. As a result, we get the following local convergence rate by optimizing over the step-size $\eta$.

**Corollary 2.** *With $\eta = \frac{\mu_{\mathbf{y}}}{4L^2}$, we have $\rho(\nabla F_\eta^{\text{Sim}}(\mathbf{z}^*)) < 1 - \frac{1}{16\max\{\kappa_{\mathbf{y}}\kappa^2_{\mathbf{xy}}, \kappa^2_{\mathbf{y}}\}}$. Hence, Sim-GDA converges locally at a linear rate $\mathcal{O}\left(\left(1-\frac{1}{16\max\{\kappa_{\mathbf{y}}\kappa^2_{\mathbf{xy}}, \kappa^2_{\mathbf{y}}\}}\right)^t\right)$.*

This corollary suggests that the convergence rate of Sim-GDA could match the rate in Theorem 4 if the coupling matrix is well-conditioned (i.e., $\kappa_{\mathbf{xy}} \approx 1$), albeit the absence of strong convexity in $\mathbf{x}$. Naturally, this begs the question: whether we can derive similar results for Alt-GDA that improves upon the rate bound of Sim-GDA. We answer this question in the affirmative. In particular, we have the following bounds for the eigenvalues of $\nabla F_\eta^{\text{Alt}}$.

**Theorem 7.** *Let* $\eta \leq \frac{1}{2L}$, *the eigenvalues of* $\nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*)$ *satisfy the following bound*

$$\text{if real: } |\lambda| \leq \max\{1-\eta^2\mu_{\mathbf{xy}}^2, 1-\eta\mu_{\mathbf{y}}\},$$
$$\text{if complex: } |\lambda| \leq \sqrt{1-\eta\mu_{\mathbf{y}}}.$$

Compare the eigenvalue bound of Alt-GDA to Sim-GDA, one may notice that the main difference is the complex eigenvalues. Similar to the SCSC setting, the complex eigenvalues of Alt-GDA are much smaller in magnitude, thus allowing us to use larger step sizes. Consequently, we have a better convergence rate for Alt-GDA (see Figure 1 for detailed comparisons).

**Corollary 3.** *Choosing* $\eta = \frac{1}{2L}$, *we have we have* $\rho(\nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*)) \leq 1 - \frac{1}{4\max\{\kappa_{\mathbf{xy}}^2, \kappa_{\mathbf{y}}\}}$. *Hence, Alt-GDA converges locally at a linear rate* $\mathcal{O}((1-\frac{1}{4\max\{\kappa_{\mathbf{xy}}^2,\kappa_{\mathbf{y}}\}}+\epsilon)^t)$ *with* $\epsilon > 0$ *a small constant.*

Compared with the bound of Sim-GDA in Theorem 2, one can see that Alt-GDA converges much faster than Sim-GDA, especially when $\kappa_{\mathbf{y}}$ is large.

## 5 GLOBAL CONVERGENCE FOR BILINEARLY-COUPLED MINIMAX GAMES

So far, we derived local convergence rates of Alt-GDA in different settings. Nonetheless, the global convergence results remain largely unknown. Unlike local convergence analysis, we have to switch to Lyapunov theory for global convergence analysis. Finding a right Lyapunov function for Alt-GDA turns out to be extremely hard and we resort to integral quadratic constraints (IQC) theory (Lessard et al., 2016; Zhang et al., 2021) for a computer-aided proof[4]. Basically, we view the algorithm as an interconnected dynamical system with nonlinear feedback (i.e., the gradient) and model the nonlinear feedback with quadratic constraints[5]. Then it allows us to *automatically* search for a quadratic Lyapunov function for certifying the worst-case convergence rate by solving a semi-definite program (SDP). Due to space constraints, we refer the reader to Appendix B for all the details.

In particular, we analyze Alt-GDA for bilinearly-coupled minimax games with the following form:

$$\min_{\mathbf{x}\in\mathbb{R}^m}\max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x}) + \mathbf{x}^\top\mathbf{B}\mathbf{y} - g(\mathbf{y}), \tag{8}$$

where we assume both $f$ and $g$ are $\mu$-strongly-convex and $L$-smooth, $\|\mathbf{B}\|_2 \leq L$. This problem is a special case of the minimax games that is amenable to IQC analysis. This problem has been studied extensively Chambolle and Pock (2011); Du and Hu (2019); Xie et al. (2021). However, the convergence properties of Alt-GDA again remain unknown.

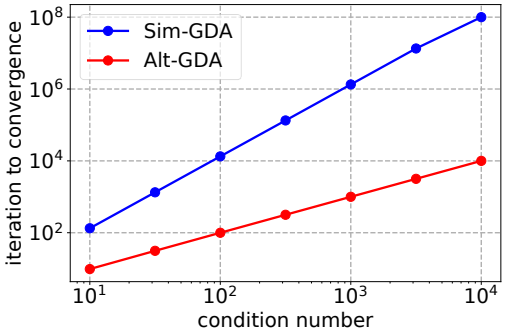



**Figure 3:** Certified iteration complexities of Sim-GDA and Alt-GDA for blinearly-coupled minimax games. **Observations:** (1) Sim-GDA converges with an iteration complexity of $\mathcal{O}(\kappa^2)$; (1) Alt-GDA achieves an improved rate of $\mathcal{O}(\kappa)$, which matches the local rate.

Using the IQC framework, we are able to search for the best possible convergence rate of Alt-GDA for every given condition number $\kappa$ by solving a SDP. However, the size of the SDP is proportional to $m$ and $n$. This can be problematic in cases where $m$ (or $n$) is large because it can be computationally costly to solve large SDPs. Fortunately, we prove that the high dimensional problem isn't any harder than than the case of $m = n = 1$, so we can reduce the problem to a SDP with $m = n = 1$, which is easy to solve.

**Theorem 8.** *Using the IQC framework to analyze the convergence rate of Alt-GDA on problem* (8), *we can simply assume* $m = n = 1$ *if* $\mathbf{B}$ *is diagonal. Let* $\rho_{m,n}$ *be the IQC-certified rate for problem* (8) *with* $\mathbf{x} \in \mathbb{R}^m$ *and* $\mathbf{y} \in \mathbb{R}^n$, *then we have* $\rho_{m,n} \leq \rho_{1,1}$.

In Figure 3, we plot the IQC-certified iteration complexity as a function of condition number. We observe that the bound for Alt-GDA does improve upon that of Sim-GDA, especially when the condition number is large. In particular, the complexity of Alt-GDA scales linearly with the condition number, suggesting its iteration complexity is $\mathcal{O}(\kappa)$. This implies that Alt-GDA does accelerate the convergence globally for this class of problem.

## 6 RELATED WORK

The discussion of simultaneous and alternating updates in iterative algorithms dates back to the Jacobi and Gauss–Seidel methods in numerical linear algebra (Saad, 2003). The Jacobi method makes simultaneous updates and is therefore naturally amenable

[4]See the blog by Adrien Taylor for more details about computer-aided analyses (https://francisbach.com/computer-aided-analyses/).

[5]Both convexity and smoothness can be characterized tightly with quadratic constraints.

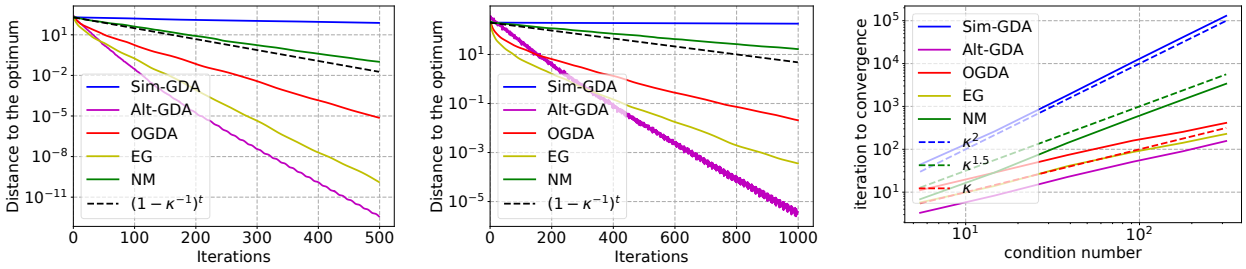



**Figure 4: Left:** Distance to the optimum as a function of training iterations. Alt-GDA accelerates Sim-GDA significantly on this particular quadratic minimax problem. In particular, its convergence rate is better than the worst-case rate of $1-\kappa^{-1}$ with $\kappa = \max\{|\lambda_i|\}/\min\{\Re(\lambda_i)\}$ ($\kappa \approx 50$). Moreover, Alt-GDA outperforms OGDA and EG by a visible margin albeit the fact that they all have the $\mathcal{O}(\kappa)$ complexity bound. **Middle:** A hard problem with entries of $\mathbf{B}$ sampled from $\mathcal{N}(0,1)$ ($\kappa \approx 300$). **Right:** Iterations to convergence for the all 5 algorithms with tuned hyperparameters. The iteration complexity of Alt-GDA scales linearly with the condition number while that of Sim-GDA scales quadratically.

to parallelization. On the other hand, the Gauss-Seidel method updates sequentially, so that each update leverages fresh information, and therefore is typically more stable and converges in fewer iterations. In minimax optimization, there is an analogous trade-off between simultaneous and alternating updates.

The discussion of alternating algorithms is lacking and is largely limited to simple bilinear games. Gidel et al. (2019b) showed that Alt-GDA stays bounded and negative momentum with alternating updates converges linearly in bilinear games. Later, Bailey et al. (2020) extended the analysis of Alt-GDA to no-regret online learning, albeit just for simple bilinear games. Zhang and Yu (2020) provided some evidence that alternating versions of many popular algorithms outperform their simultaneous counterpart in bilinear games. Very recently, Yang et al. (2020a) established the global convergence of Alt-GDA in a subclass of nonconvex-nonconcave objectives satisfying a so-called two-sided Polyak-Łojasiewicz inequality. Xu et al. (2020); Boţ and Böhm (2020) proved convergence rates for alternating (proximal) GDA for nonconvex-concave minimax problems. However, it remains unclear whether these alternating methods improve the convergence compared to their simultaneous counterparts in the above two settings.

By contrast, there is a large body of work on simultaneous methods in minimax optimization. For the strongly convex-strongly concave case, Tseng (1995) and Nesterov and Scrimali (2011) proved that their algorithms find an $\epsilon$-saddle point with a gradient complexity of $\mathcal{O}(\kappa \ln(1/\epsilon))$ using a variational inequality approach. Using a different approach, Gidel et al. (2019a) and Mokhtari et al. (2020a) derived the same convergence results for OGDA. Particularly, Mokhtari et al. (2020a) gave a unified analysis of OGDA and EG from the perspective of proximal point methods. Later, Zhang et al. (2021) provided a unified and automated framework for analyzing various first-order methods using the theory of integral quadratic constraints from control theory. Very recently, Ibrahim et al. (2020); Zhang et al. (2019b) established fine-grained lower complexity bounds among all the first-order algorithms in this setting, and these bounds were later achieved by Lin et al. (2020); Wang and Li (2020). For the convex-concave setting, it is known that the optimal rate of convergence for first-order methods is $\mathcal{O}(1/T)$, and this rate is achieved by both the EG and OGDA algorithms (Nemirovski, 2004; Tseng, 2008; Hsieh et al., 2019; Mokhtari et al., 2020b) for the averaged (ergodic) iterates. Later, (Golowich et al., 2020b,a) derived a $\mathcal{O}(1/\sqrt{T})$ bound for the last iterate of EG and OGDA.

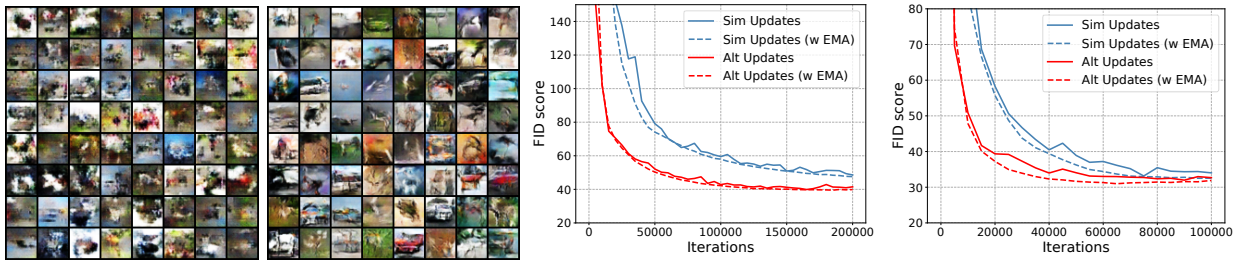
# 7 EXPERIMENTS

## 7.1 Quadratic Minimax Games

In this section, we compare the performance of Alt-GDA with Sim-GDA along with other three popular algorithms (EG, OGDA and NM) so as to verify our theoretical results on the convergence rate of Alt-GDA. In particular, we focus on the following quadratic minimax problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^d} f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{B}\mathbf{y} - \frac{1}{2}\mathbf{y}^\top\mathbf{C}\mathbf{y} \quad (9)$$

where we set the dimension $d = 100$. We note both linear regression (Du and Hu, 2019) and robust least squares (Yang et al., 2020a) problems admit this minimax formulation. The matrices $\mathbf{A}$ and $\mathbf{C}$ are set to have eigenvalues $\{\frac{1}{i}\}_{i=1}^d$. For matrix $\mathbf{B}$, we set it to be a random matrix with entries sampling from a Gaussian distribution (either $\mathcal{N}(0, 0.01)$ or $\mathcal{N}(0,1)$). In the case of $\mathbf{B}$ sampled from $\mathcal{N}(0,1)$, the resulting gradient vector field has a strong rotational force since the off-diagonal blocks of its Jacobian dominates (see (10) in the Appendix). For all algorithms, the iterates start with $\mathbf{x}_0 = \mathbf{1}$ and $\mathbf{y}_0 = \mathbf{1}$. Figure 4 shows that the distance to the optimum of Sim-GDA, Alt-GDA, OGDA,




**(a)** Sim-GDA **(b)** Alt-GDA **(c)** Trained with SGD **(d)** Trained with AMSGrad

**Figure 5:** Comparisons between simultaneous and alternating algorithms on GAN training. **(a)** We train the DCGAN model on CIFAR-10 with simultaneous SGD . The samples are saved at iteration 30000. **(b)** We train the DCGAN model on CIFAR10 with alternating SGD. Again, the samples are saved at iteration 30000. **(c)** The curves of FID scores on CIFAR-10 with SGD. The dash lines are computed at exponential moving averaged models. **(d)** The curves of FID scores on CIFAR-10 with AMSGrad. In all settings, alternating version of the algorithms converges much faster than their simultaneous counterparts and achieve better FID scores in the end.

EG and NM[6] versus the number of iterations for this problem. For all methods, we tune their hyperparameters by grid-search. We notice that all methods converge linearly to the optimum. As expected, Alt-GDA performs significantly better than Sim-GDA and yields a convergence rate that is better than its worst-case rate (black dashed line). Moreover, we find that Alt-GDA outperforms OGDA and EG by a visible margin. This is surprising, in that OGDA and EG take another memory buffer for accelerating the convergence.

Furthermore, we study how the convergence rates (or iteration complexities) scale with the condition numbers. To this end, we randomly sample[7] matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and compute the condition number by $\kappa = \frac{\max\{|\lambda_i|\}}{\min\{\Re(\lambda_i)\}}$ where $\lambda_i$ are eigenvalues of the Jacobian $\mathbf{J}$ of the gradient vector field in (10). Once we have all these three matrices, we can compute the spectral radius $\rho$ of all algorithms with tuned step-sizes and momentum value. We plot $-1/\log(\rho)$ versus the condition number $\kappa$ in Figure 4 (right) to get a sense of how the relative iteration complexity scales as a function of condition number. We find that the iteration complexity of Alt-GDA scales linearly with the condition number, matching our prediction in Corollary 1. On the other hand, Sim-GDA takes roughly $\kappa^2$ iterations to convergence, as predicted in Theorem 4. In addition, Alt-GDA is slightly better than OGDA and EG as its curve is below that of OGDA and EG, albeit with the same slope.

[6]We implemented the simultaneous version of negative momentum (NM). For alternating NM, the optimal damping value of NM is roughly zero, making it the same algorithm as Alt-GDA.

[7]We let the eigenvalues of matrices $\mathbf{A}$ and $\mathbf{C}$ be $\{\frac{1}{n_i}\}_{i=1}^{d}$ where $n_i$ are evenly spaced from 1 to $N$, where $N$ is in $[\sqrt{10}, 10^3]$. We sample all entries of $\mathbf{B}$ from standard normal distribution $\mathcal{N}(0,1)$ and then normalize it.

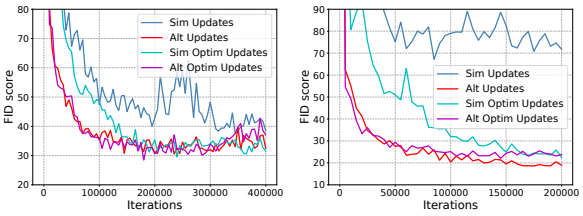
### 7.2 Generative Adversarial Networks

In this section, we investigate the effect of alternating updates on training generative adversarial networks. The purpose of this section is to show that the insights gained from our analyses carry over to GAN training despite the fact that the GAN objective is generally nonconvex-nonconcave. In addition, we note that while GAN training is a stochastic problem, stochastic problems are sometimes in a curvature-dominated regime where the convergence behavior resembles that of the deterministic problems (Zhang et al., 2019a).

We first compare alternating algorithms with their simultaneous counterparts on CIFAR10 (Krizhevsky et al., 2009) image generation task with the WGAN-GP (Gulrajani et al., 2017) objective and a DCGAN (Radford et al., 2015) architecture. In particular, we choose SGD and AMSGrad (Reddi et al., 2018) as our base optimizers. For more implementation details, please see Appendix D. We evaluate all algorithms with Fréchet Inception Distance[8] (FID) (Heusel et al., 2017). Figure 5c and 5d summarize our results. With SGD as our optimizer[9], we observe that alternating SGD not only converges faster, but also converges to a better point with lower FID score. Although both alternating version and simultaneous version of AMSGrad converges to models with similar FID scores, the alternating version again converges with many fewer iterations, matching our prediction. In addition, we generate samples from trained Generators at iteration 30000 with SGD optimizer (see Figure 5a and 5b). It is easy to see that the model trained with alternating updates generates better samples given the same

[8]Inception score (Salimans et al., 2016) is also a popular metric, however it was shown by Chavdarova et al. (2021) that it is less consistent with the sample quality, so we instead use FID score here.

[9]We also include the results with exponential moving average (EMA).




**(a)** SGD **(b)** AMSGrad

**Figure 6:** **(a)** ResNet model trained with SGD on CIFAR-10. **(b)** ResNet model trained with AMSGrad on CIFAR-10. Alternating algorithms dominate simultaneous ones. More interestingly, the use of optimism does *not* help for alternating algorithms.

compute budget.

Furthermore, we compare simultaneous methods and alternating methods on a deep ResNet (Miyato et al., 2018). We also include optimistic updates (Daskalakis et al., 2018) in the training, which is the key component of OGDA[10]. We report all results in Figure 6. The first observation is that alternating algorithms take fewer iterations to converge regardless of whether optimism is used, and sometimes converge to models with better FID scores (similar to DCGAN results). Second, we observe that the use of optimism only helps for simultaneous algorithms, suggesting that alternating updates and optimistic updates play similar roles in improving GAN training. This could be explained by our theoretical results that Alt-GDA enjoys a similar convergence rate to OGDA.

## 8 CONCLUSION

In this paper, we take an important step towards understanding alternating algorithms in minimax optimization by analyzing Alt-GDA in three distinct settings. In particular, we show theoretically that Alt-GDA outperforms its simultaneous counterpart by a big margin in all three settings. Unexpectedly, Alt-GDA achieves a near-optimal convergence rate locally for strongly convex-strongly concave smooth minimax games, matching the known coarse lower bound. Moreover, the acceleration effect of Alt-GDA remains when the minimax problem has only strong concavity in the dual variables.

Our numerical simulations on toy quadratic games verified our claims. Further, we demonstrate empirically that alternating updates could significantly speed up GAN training though GAN objective is generally nonconvex-nonconcave. More interestingly, we show that the use of optimism only helps for simultaneous algorithms. We believe that the default use of alternating update rule in GAN training was an important reason for its success.

[10]See Appendix D for detailed update rule.

## References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873, 2020a.

Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1705–1715, 2020b.

James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 321–338, 2018.

James P Bailey, Gauthier Gidel, and Georgios Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Conference on Learning Theory*, pages 391–407. PMLR, 2020.

Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Tatjana Chavdarova, Matteo Pagliardini, Sebastian U Stich, François Fleuret, and Martin Jaggi. Taming GANs with lookahead-minmax. In *International Conference on Learning Representations*, 2021.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

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

Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 196–205, 2019.

Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058, 2017.

Tanner Fiez and Lillian J Ratliff. Local convergence analysis of gradient descent ascent with finite timescale separation. In *International Conference on Learning Representations*, 2021.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. 2019a.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019b.

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

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3):161–220, 1990.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pages 6938–6948, 2019.

Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

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

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob N Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *J. Mach. Learn. Res.*, 20:84–1, 2019.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915, 2019.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2738–2779, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial

attacks. In *International Conference on Learning Representations*, 2018.

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

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507, 2020a.

Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

Arkadi 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. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete & Continuous Dynamical Systems-A*, 31(4):1383, 2011.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

Tim Salimans, Ian J Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

Florian Schaefer and Anima Anandkumar. Competitive gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

John von Neumann and Oskar Morgenstern. Theory of games and economic behavior. 1944.

Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. In *Advances in Neural Information Processing Systems*, 2020.

Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In *International Conference on Learning Representations*, 2019.

Guangzeng Xie, Yuze Han, and Zhihua Zhang. Dippa: An improved method for bilinear saddle point problems. *arXiv preprint arXiv:2103.08270*, 2021.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv preprint arXiv:2006.02032*, 2020.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, 2020a.

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33, 2020b.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33, 2020c.

Guodong Zhang and Yuanhao Wang. On the suboptimality of negative momentum for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. *Advances in neural information processing systems*, 32:8196–8207, 2019a.

Guodong Zhang, Xuchan Bao, Laurent Lessard, and Roger Grosse. A unified analysis of first-order methods for smooth games via integral quadratic constraints. *Journal of Machine Learning Research*, 22: 1–39, 2021.

Guojun Zhang and Yaoliang Yu. Convergence of gradient methods on bilinear zero-sum games. In *International Conference on Learning Representations*, 2020.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019b.

# A Technical Proofs

For notational convenience, we define the gradient vector field of minimax games $V(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})^\top, -\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})^\top]^\top$ and its associated Jacobian matrix at Nash equilibrium $\mathbf{z}^*$:

$$\mathbf{J} = \begin{bmatrix} \nabla^2_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}^*,\mathbf{y}^*) \\ -\nabla^2_{\mathbf{yx}} f(\mathbf{x}^*,\mathbf{y}^*) & -\nabla^2_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*) \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^\top & \mathbf{C} \end{bmatrix} \tag{10}$$

## A.1 Proof of Theorem 3

For the bilinear games, the spectral radius of $\nabla F_\eta^{\text{Alt}}$ is easy to bound. In particular, we have

$$\nabla F_\eta^{\text{Alt}} = \begin{bmatrix} \mathbf{I} & -\eta\mathbf{B} \\ \eta\mathbf{B}^\top & \mathbf{I} - \eta^2\mathbf{B}^\top\mathbf{B} \end{bmatrix}. \tag{11}$$

Here, we define the SVD decomposition of $\mathbf{B}$ as $\mathbf{B} = \mathbf{U}\hat{\mathbf{B}}\mathbf{V}^\top$ where $\hat{\mathbf{B}}$ is diagonal, hence we have $\rho(\nabla F_\eta^{\text{Alt}})$ equivalent to the spectral radius of the following matrix

$$\begin{bmatrix} \mathbf{I} & -\eta\hat{\mathbf{B}} \\ \eta\hat{\mathbf{B}}^\top & \mathbf{I} - \eta^2\hat{\mathbf{B}}^\top\hat{\mathbf{B}} \end{bmatrix}, \tag{12}$$

whose eigenvalues satisfying $(x-1)^2 + \eta^2\lambda x = 0,\ \lambda \in \text{Sp}(\mathbf{B}^\top\mathbf{B})$. As long as $|2-\eta^2\lambda| \le 2$, we have the roots of this polynomial satisfying $|x| = 1$. Therefore, we have $\rho(\nabla F_\eta^{\text{Alt}}) = 1$ as long as $\eta \le \frac{2}{\sigma_{\max}(\mathbf{B})}$. In the meantime, it can shown that $\rho(\nabla F_\eta^{\text{Alt}})$ is diagonalizable, so by (Gidel et al., 2019b, Lemma 3), the iterates of Alt-GDA stay bounded. This finishes the proof.

## A.2 Proof of Theorem 4

To prove Theorem 4, we first claim that all eigenvalues of $\mathbf{J}$ in (10) fall within the following set:

$$\mathcal{K} = \left\{ \lambda \in \mathbb{C} : |\lambda| \le \sqrt{2}L, \Re\lambda \ge \mu > 0 \right\}. \tag{13}$$

We first prove $\Re\lambda \ge \mu$. Let $\lambda \triangleq a + bi$ be a complex eigenvalue of $\mathbf{J}$ such that $\mathbf{J}\mathbf{v} = \lambda\mathbf{v}$. In general, the eigenvector $\mathbf{v}$ is a complex vector and we let $\mathbf{v} = \mathbf{u} + \mathbf{w}i$. Then, one can show

$$\Re(\lambda) = \frac{\mathbf{u}^\top\mathbf{J}\mathbf{u} + \mathbf{w}^\top\mathbf{J}\mathbf{w}}{\mathbf{u}^\top\mathbf{u} + \mathbf{w}^\top\mathbf{w}} = \frac{\mathbf{u}_1^\top\mathbf{A}\mathbf{u}_1 + \mathbf{u}_2^\top\mathbf{C}\mathbf{u}_2 + \mathbf{w}_1^\top\mathbf{A}\mathbf{w}_1 + \mathbf{w}_2^\top\mathbf{C}\mathbf{w}_2}{\mathbf{u}^\top\mathbf{u} + \mathbf{w}^\top\mathbf{w}} \tag{14}$$

where $\mathbf{u}_1 \in \mathbb{R}^m$ is the first half of the vector $\mathbf{u}$ and $\mathbf{u}_2 \in \mathbb{R}^n$ is the second half (the same for $\mathbf{w}_1, \mathbf{w}_2$). As we know that $\mathbf{A} \succeq \mu_{\mathbf{x}}\mathbf{I} \succeq \mu\mathbf{I}$ and $\mathbf{C} \succeq \mu_{\mathbf{y}}\mathbf{I} \succeq \mu\mathbf{I}$, we have

$$\Re(\lambda) \ge \frac{\mu\mathbf{u}_1^\top\mathbf{u}_1 + \mu\mathbf{u}_2^\top\mathbf{u}_2 + \mu\mathbf{w}_1^\top\mathbf{w}_1 + \mu\mathbf{w}_2^\top\mathbf{w}_2}{\mathbf{u}^\top\mathbf{u} + \mathbf{w}^\top\mathbf{w}} = \mu \tag{15}$$

We next prove $|\lambda| \le \sqrt{2}L$. To this end, it suffices to show $\lambda_{\max}(\mathbf{J}^\top\mathbf{J}) \le 2L^2$. Recall the definition of $\mathbf{J}$ in (10), we have

$$\mathbf{J}^\top\mathbf{J} = \begin{bmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^\top & \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^2 + \mathbf{B}\mathbf{B}^\top & \mathbf{A}\mathbf{B} - \mathbf{B}\mathbf{C} \\ \mathbf{B}^\top\mathbf{A} - \mathbf{C}\mathbf{B}^\top & \mathbf{B}^\top\mathbf{B} + \mathbf{C}^2 \end{bmatrix}. \tag{16}$$

Hence,

$$\lambda_{\max}\left(\mathbf{J}^\top\mathbf{J}\right) = \max_{\|\mathbf{v}\|=1} \mathbf{v}^\top\mathbf{J}^\top\mathbf{J}\mathbf{v} = \max_{\|\mathbf{v}\|=1} \mathbf{v}_1^\top(\mathbf{A}^2 + \mathbf{B}\mathbf{B}^\top)\mathbf{v}_1 + \mathbf{v}_2^\top(\mathbf{C}^2 + \mathbf{B}^\top\mathbf{B})\mathbf{v}_2 \tag{17}$$

Because we assume $\mathbf{A} \preceq L_{\mathbf{x}}\mathbf{I} \preceq L\mathbf{I}$, $\mathbf{C} \preceq L_{\mathbf{y}}\mathbf{I} \preceq L\mathbf{I}$ and $\|\mathbf{B}\|_2 \le L_{\mathbf{xy}} \le L$, we have

$$\lambda_{\max}\left(\mathbf{J}^\top\mathbf{J}\right) \le \max_{\|\mathbf{v}\|=1} 2L^2(\mathbf{v}_1^\top\mathbf{v}_1 + \mathbf{v}_2^\top\mathbf{v}_2) = 2L^2. \tag{18}$$

Therefore, we get $|\lambda| \le \sqrt{2}L$. Now the convergence rate bound of Sim-GDA reduces to the following problem:

$$\min_\eta \max_{\lambda\in\mathcal{K}} |1 - \eta\lambda| = \min_\eta \max_{\lambda\in\mathcal{K}} \sqrt{(1 - \eta\Re(\lambda))^2 + \eta^2\Im(\lambda)^2} \tag{19}$$

where the maximum modulus is achieved at the point $\lambda = \mu + \sqrt{2L^2 - \mu^2}i$. Hence, we have

$$\rho(\nabla F_\eta^{\mathrm{Sim}}(\mathbf{z}^*)) \le \min_\eta \max_{\lambda \in \mathcal{K}} |1 - \eta\lambda| = \min_\eta \sqrt{1 - 2\eta\mu + 2\eta^2 L^2} = \sqrt{1 - \frac{\mu^2}{2L^2}}. \tag{20}$$

By invoking Theorem 1, we finish our proof.

### A.3 Proof of Theorem 5

Recall the updates of Alt-GDA:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t),\ \mathbf{y}_{t+1} = \mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t).$$

By definition, we have the Jacobian matrix of Alt-GDA updates in the following form:

$$\nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*) = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \eta\mathbf{B}^\top & \mathbf{I} - \eta\mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{I} - \eta\mathbf{A} & -\eta\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \eta\mathbf{A} & -\eta\mathbf{B} \\ \eta\mathbf{B}^\top(\mathbf{I} - \eta\mathbf{A}) & \mathbf{I} - \eta\mathbf{C} - \eta^2\mathbf{B}^\top\mathbf{B} \end{bmatrix}. \tag{21}$$

Without loss of generality, we assume matrices $\mathbf{A}$ and $\mathbf{C}$ to be diagonal with eigenvalues $[\alpha]_i$ and $[\beta]_j$. We have the characteristic polynomial:

$$\begin{aligned} \mathrm{Det}(\lambda\mathbf{I} - \nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*)) = \mathrm{Det}(\lambda\mathbf{I} - (\mathbf{I} - \eta\mathbf{A})) \\ \mathrm{Det}(\lambda\mathbf{I} - (\mathbf{I} - \eta\mathbf{C} - \eta^2\mathbf{B}^\top\mathbf{B} - \eta^2\mathbf{B}^\top(\mathbf{I} - \eta\mathbf{A})(\lambda\mathbf{I} - \mathbf{I} + \eta\mathbf{A})^{-1}\mathbf{B})), \end{aligned} \tag{22}$$

where we used properties of the Schur complement. For $\mathrm{Det}(\lambda\mathbf{I} - \nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*))$ to be zero, $\lambda$ has to one of the eigenvalues of the following matrix:

$$\mathbf{M} \triangleq \mathbf{I} - \eta\mathbf{C} - \eta^2\mathbf{B}^\top \mathrm{Diag}\left(\frac{\lambda}{\lambda - 1 + \eta\alpha_i}\right)\mathbf{B}. \tag{23}$$

**In the case of $\lambda$ being real**, it is easy to show that for any $\eta \le \frac{1}{2L}$, we have

$$|\lambda| \le \max\{1 - \eta\alpha_{\min}, 1 - \eta\beta_{\min}\}. \tag{24}$$

We prove that by contradiction. Suppose $\lambda > \max\{1 - \eta\alpha_{\min}, 1 - \eta\beta_{\min}\} \ge 0$, then we have

$$\lambda_{\max}(\mathbf{M}) \le 1 - \eta\lambda_{\min}(\mathbf{C}) = 1 - \eta\beta_{\min}. \tag{25}$$

That is because the term $\eta^2\mathbf{B}^\top \mathrm{Diag}\left(\frac{\lambda}{\lambda - 1 + \eta\alpha_i}\right)\mathbf{B}$ is positive semi-definite when $\lambda > \max\{1 - \eta\alpha_{\min}, 1 - \eta\beta_{\min}\} \ge 0$. Since we know $\lambda$ is one of the eigenvalue of $\mathbf{M}$ and hence it has to be smaller than $1 - \eta\beta_{\min}$, contradiction.

Suppose $\lambda < -\max\{1 - \eta\alpha_{\min}, 1 - \eta\beta_{\min}\} \le 0$, we have

$$\mathbf{M} \succeq \mathbf{I} - \eta\mathbf{C} - \eta^2\mathbf{B}^\top\mathbf{B}. \tag{26}$$

For bilinear games where $\mathbf{C} = \mathbf{0}$, we have $\mathbf{M} \succeq \mathbf{0}$ and hence $\lambda$ is impossible to be smaller than $-1$. On the other hand, since we have $\eta \le \frac{1}{2L}$, we know $\mathbf{M} \succeq \mathbf{0}$, contradiction again. Therefore, we proved that $|\lambda| \le \max\{1 - \eta\alpha_{\min}, 1 - \eta\beta_{\min}\}$.

**In the case of $\lambda$ being complex**, we let $\lambda = a + bi$ with $b \ne 0$ and $\mathbf{v}$ be the eigenvector associated with $\lambda$ such that $\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$. Then we have the following identities:

$$\begin{aligned} \mathbf{v}^{\mathsf{H}}(\mathbf{M} + \mathbf{M}^{\mathsf{H}})\mathbf{v} = 2\Re(\lambda) = 2a, \\ \mathbf{v}^{\mathsf{H}}(\mathbf{M} - \mathbf{M}^{\mathsf{H}})\mathbf{v} = 2\Im(\lambda)i = 2bi. \end{aligned} \tag{27}$$

Plugging the value of $\mathbf{M}$, we have

$$a = \frac{1}{2}\mathbf{v}^{\mathsf{H}}(\mathbf{M} + \mathbf{M}^{\mathsf{H}})\mathbf{v} = \sum_j (1 - \eta\beta_j)|\mathbf{v}_j|^2 - \eta^2 \sum_j |(\mathbf{B}\mathbf{v})_j|^2 \frac{a^2 - a(1 - \eta\alpha_j) + b^2}{(a - 1 + \eta\alpha_j)^2 + b^2} \tag{28}$$

and

$$bi = \frac{1}{2}\mathbf{v}^{\mathsf{H}}(\mathbf{M} - \mathbf{M}^{\mathsf{H}})\mathbf{v} = \eta^2 \sum_j |(\mathbf{B}\mathbf{v})_j|^2 \frac{(1-\eta\alpha_j)bi}{(a-1+\eta\alpha_j)^2 + b^2}. \tag{29}$$

From (29), one can get

$$\eta^2 \sum_j |(\mathbf{B}\mathbf{v})_j|^2 \frac{(1-\eta\alpha_j)}{(a-1+\eta\alpha_j)^2 + b^2} = 1. \tag{30}$$

Next, combining (28) and (30), we have

$$\sum_j (1-\eta\beta_j)|\mathbf{v}_j|^2 - \eta^2 |(\mathbf{B}\mathbf{v})_j|^2 \frac{a^2+b^2}{\Delta_j} = 0, \tag{31}$$

where $\Delta_j = (a-1+\eta\alpha_j)^2 + b^2$. It follows from (31) and (30) that

$$\begin{aligned} 1 &= \eta^2 \sum_j |(\mathbf{B}\mathbf{v})_j|^2 \frac{(1-\eta\alpha_j)}{\Delta_j} \le \eta^2 \sum_j |(\mathbf{B}\mathbf{v})_j|^2 \frac{(1-\eta\alpha_{\min})}{\Delta_j} \\ &= \eta^2 \frac{1-\eta\alpha_{\min}}{a^2+b^2} \sum_j |(\mathbf{B}\mathbf{v})_j|^2 \frac{a^2+b^2}{\Delta_j} \\ &\overset{(31)}{=} \frac{1-\eta\alpha_{\min}}{a^2+b^2} \sum_j (1-\eta\beta_j)|\mathbf{v}_j|^2 \le \frac{(1-\eta\alpha_{\min})(1-\eta\beta_{\min})}{a^2+b^2} \end{aligned}. \tag{32}$$

As a result,

$$|\lambda|^2 = a^2 + b^2 \le (1-\eta\alpha_{\min})(1-\eta\beta_{\min}). \tag{33}$$

### A.4 Proof of Theorem 6

Recall that the Jacobian matrix $\nabla F_\eta^{\mathrm{Sim}}(\mathbf{z}^*)$ of Sim-GDA:

$$\nabla F_\eta^{\mathrm{Sim}}(\mathbf{z}^*) = \begin{bmatrix} \mathbf{I} - \eta\mathbf{A} & -\eta\mathbf{B} \\ \eta\mathbf{B}^\top & \mathbf{I} - \eta\mathbf{C} \end{bmatrix}. \tag{34}$$

To bound its spectral radius, we first compute its characteristic polynomial and simplify it with the Schur complement.

$$\mathrm{Det}\left(\lambda\mathbf{I} - \nabla F_\eta^{\mathrm{Sim}}(\mathbf{z}^*)\right) = \mathrm{Det}(\lambda\mathbf{I} - (\mathbf{I}-\eta\mathbf{C}))\mathrm{Det}(\lambda\mathbf{I} - (\mathbf{I} - \eta\mathbf{A} - \eta^2\mathbf{B}(\lambda\mathbf{I} - (\mathbf{I}-\eta\mathbf{C}))^{-1}\mathbf{B}^\top)) \tag{35}$$

**In the case of $\lambda$ being real**, for $\eta \le \frac{1}{L}$, one can prove that $\lambda$ is within the range $(0, 1)$ by contradiction argument. Without loss of generality, we assume matrices $\mathbf{A}$ and $\mathbf{C}$ to be diagonal with eigenvalues $[\alpha]_i$ and $[\beta]_j$. Let assume $\lambda > 1 - \eta\beta_{\min}$, then we claim that $\lambda \le 1 - \frac{\eta}{L}\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$. The key is (by $\lambda < 1$)

$$(\lambda\mathbf{I} - (\mathbf{I}-\eta\mathbf{C}))^{-1} \succeq (\eta\mathbf{C})^{-1} \succeq \frac{1}{\eta L_{\mathbf{y}}}\mathbf{I} \succeq \frac{1}{\eta L}\mathbf{I} \tag{36}$$

Therefore, we have a upper bound for $\mathbf{M} \triangleq \mathbf{I} - \eta\mathbf{A} - \eta^2\mathbf{B}(\lambda\mathbf{I} - (\mathbf{I}-\eta\mathbf{C}))^{-1}\mathbf{B}^\top)$

$$\mathbf{M} \preceq \mathbf{I} - \frac{\eta}{L}\mathbf{B}\mathbf{B}^\top \preceq \left(1 - \frac{\eta}{L}\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right)\mathbf{I}. \tag{37}$$

Hence, $\lambda$, as one of the eigenvalues of $\mathbf{M}$, has to be smaller than $1 - \frac{\eta}{L}\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$. In summary, we proved that

$$0 < \lambda \le \max\left\{1 - \eta\beta_{\min}, 1 - \frac{\eta}{L}\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right\} \tag{38}$$

**In the case of $\lambda$ being complex**, we claim that $\Re(\lambda) \le 1 - \frac{\eta}{2}\mu_{\mathbf{y}}$. Let $\lambda = a + bi$ with $b \ne 0$ and $\mathbf{v}$ be the eigenvector associated with $\lambda$ such that $\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$. Then we have the following identities:

$$\begin{aligned} \mathbf{v}^{\mathsf{H}}(\mathbf{M} + \mathbf{M}^{\mathsf{H}})\mathbf{v} &= 2\Re(\lambda) = 2a \\ \mathbf{v}^{\mathsf{H}}(\mathbf{M} - \mathbf{M}^{\mathsf{H}})\mathbf{v} &= 2\Im(\lambda)i = 2bi \end{aligned} \tag{39}$$

Plugging the value of $\mathbf{M}$, we have

$$a = \frac{1}{2}\mathbf{v}^{\mathsf{H}}(\mathbf{M}+\mathbf{M}^{\mathsf{H}})\mathbf{v} = \sum_j (1-\eta\alpha_j)|\mathbf{v}_j|^2 - \eta^2 \sum_j |(\mathbf{B}^\top \mathbf{v})_j|^2 \frac{a-(1-\eta\beta_j)}{(a-1+\eta\beta_j)^2+b^2} \tag{40}$$

and

$$bi = \frac{1}{2}\mathbf{v}^{\mathsf{H}}(\mathbf{M}-\mathbf{M}^{\mathsf{H}})\mathbf{v} = \eta^2 \sum_j |(\mathbf{B}^\top \mathbf{v})_j|^2 \frac{bi}{(a-1+\eta\beta_j)^2+b^2} \tag{41}$$

From (41), we get

$$\eta^2 \sum_j |(\mathbf{B}^\top \mathbf{v})_j|^2 \frac{1}{(a-1+\eta\beta_j)^2+b^2} = 1 \tag{42}$$

Next, combining (41) and (40), we have

$$\begin{aligned} 2a-1 &= \sum_j (1-\eta\alpha_j)|\mathbf{v}_j|^2 - \eta^2 \sum_j |(\mathbf{B}^\top \mathbf{v})_j|^2 \frac{\eta\beta_j}{(a-1+\eta\beta_j)^2+b^2} \\ &\le \sum_j (1-\eta\alpha_{\min})|\mathbf{v}_j|^2 - \eta^2 \sum_j |(\mathbf{B}^\top \mathbf{v})_j|^2 \frac{\eta\beta_{\min}}{(a-1+\eta\beta_j)^2+b^2} \\ &\overset{(41)}{=} 1-\eta\beta_{\min} \end{aligned} \tag{43}$$

Therefore, we proved our claim that $\Re(\lambda) = a \le 1 - \frac{\eta}{2}\beta_{\min}$.

Then, we could prove $\mathbf{J} = \frac{1}{\eta}(\mathbf{I} - \nabla F_\eta^{\mathrm{Sim}}(\mathbf{z}^*))$ has the operator norm $\|\mathbf{J}\| \le \sqrt{2}L$ by the same argument in (18). Consequently, we have $|\Im(\lambda)| \le \eta\sqrt{2L^2 - \frac{1}{4}\mu_{\mathbf{y}}^2}$. Follows immediately, we have

$$|\lambda| = \sqrt{\Re(\lambda)^2 + \Im(\lambda)^2} \le (1-\frac{\eta}{2}\mu_{\mathbf{y}})^2 + \eta^2(2L^2 - \frac{1}{4}\mu_{\mathbf{y}}^2) = 1 - \eta\mu_{\mathbf{y}} + 2\eta^2 L^2 \tag{44}$$

### A.5 Proof of Theorem 7

As in the proof of Theorem 5, we analyze the eigenvalues of $\nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*)$. Recall $\nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*)$ has the following form:

$$\nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*) = \begin{bmatrix} \mathbf{I}-\eta\mathbf{A} & -\eta\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \eta\mathbf{B}^\top & \mathbf{I}-\eta\mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{I}-\eta\mathbf{A}-\eta^2\mathbf{B}\mathbf{B}^\top & -\eta\mathbf{B}(\mathbf{I}-\eta\mathbf{C}) \\ \eta\mathbf{B}^\top & \mathbf{I}-\eta\mathbf{C} \end{bmatrix} \tag{45}$$

Notice that this matrix in (45) is slightly different from the one defined in (21), but they have the same eigenspectrum. Without loss of generality, we assume matrices $\mathbf{A}$ and $\mathbf{C}$ to be diagonal with eigenvalues $[\alpha]_i$ and $[\beta]_j$. We then have the characteristic polynomial:

$$\begin{aligned} \mathrm{Det}(\lambda\mathbf{I} - \nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*)) &= \mathrm{Det}(\lambda\mathbf{I} - (\mathbf{I}-\eta\mathbf{C})) \\ &\quad \mathrm{Det}(\lambda\mathbf{I} - (\mathbf{I}-\eta\mathbf{A}-\eta^2\mathbf{B}\mathbf{B}^\top - \eta^2\mathbf{B}(\mathbf{I}-\eta\mathbf{C})(\lambda\mathbf{I}-\mathbf{I}+\eta\mathbf{C})^{-1}\mathbf{B}^\top)) \end{aligned} \tag{46}$$

where we used Schur complement. For $\mathrm{Det}(\lambda\mathbf{I} - \nabla F_\eta^{\mathrm{Alt}}(\mathbf{z}^*))$ to be zero, $\lambda$ has to one of the eigenvalues of the following matrix:

$$\mathbf{M} \triangleq \mathbf{I} - \eta\mathbf{A} - \eta^2\mathbf{B}\mathbf{B}^\top - \eta^2\mathbf{B}(\mathbf{I}-\eta\mathbf{C})(\lambda\mathbf{I}-\mathbf{I}+\eta\mathbf{C})^{-1}\mathbf{B}^\top \tag{47}$$

**In the case of $\lambda$ being real**, for any $\eta \le \frac{1}{2L}$, it is easy to show that $\lambda$, as one of the eigenvalues of $\mathbf{M}$, is within $(0,1)$. Let us first assume $\lambda > 1-\eta\beta_{\min}$, we have $\mathbf{B}(\mathbf{I}-\eta\mathbf{C})(\lambda\mathbf{I}-\mathbf{I}+\eta\mathbf{C})^{-1}\mathbf{B}^\top \succeq \mathbf{0}$. Hence, we have

$$\mathbf{M} \preceq \mathbf{I} - \eta\mathbf{A} - \eta^2\mathbf{B}\mathbf{B}^\top \preceq \mathbf{I} - \eta^2\mathbf{B}\mathbf{B}^\top \preceq (1-\eta^2\mu_{\mathbf{xy}}^2)\mathbf{I} \tag{48}$$

As a consequence, we know $|\lambda| \le 1 - \eta^2\mu_{\mathbf{xy}}^2$.

**In the case of $\lambda$ being complex**, we could reuse the result of (33)

$$|\lambda| \le \sqrt{(1-\eta\alpha_{\min})(1-\eta\beta_{\min})} \le \sqrt{1-\eta\mu_{\mathbf{y}}} \tag{49}$$

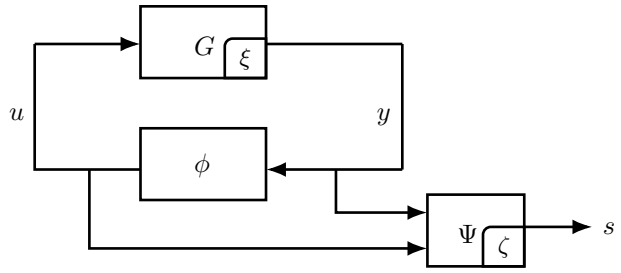



**Figure 7:** Feedback interconnection between a system $G$ (optimization algorithm) with state matrices $(A, B, C, D)$ and a nonlinearity $\phi$. An IQC is a constraint on $(y, u)$ satisfied by $\phi$.

## B Details about IQC framework

Borrowing the notations from Lessard et al. (2016), we frame various first-order algorithms as a unified linear dynamical system[11] in feedback with a nonlinearity $\phi : \mathbb{R}^d \to \mathbb{R}^d$,

$$\begin{aligned} \xi_{t+1} &= A\xi_t + Bu_t \\ y_t &= C\xi_t + Du_t \\ u_t &= \phi(y_t). \end{aligned} \tag{50}$$

At each iteration $t = 0, 1, ...$, $u_t \in \mathbb{R}^d$ is the control input, $y_t \in \mathbb{R}^d$ is the output, and $\xi_t \in \mathbb{R}^{nd}$ is the state for algorithms with $n$ step of memory. The state matrices $A, B, C, D$ differ for various algorithms. For most algorithms we consider in the paper, they have the general form:

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{cc|c} (1+\beta)\mathbf{I}_d & -\beta\mathbf{I}_d & -\eta\mathbf{I}_d \\ \mathbf{I}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline (1+\alpha)\mathbf{I}_d & -\alpha\mathbf{I}_d & \mathbf{0}_d \end{array}\right],$$

where $\mathbf{I}_d$ and $\mathbf{0}_d$ are the identity and zero matrix of size $d \times d$, respectively. Often, the nonlinear function $\phi$ is the troublesome function we wish to analyze. Although we do not know $\phi$ exactly, we assume to have some knowledge of the constraints it imposes on the input-output pair $(y, u)$. For example, we may assume $\phi$ to be $L$-Lipschitz, which implies $\|u_t - u^*\|_2 \le L\|y_t - y^*\|_2$ for all $t$ with $u^* = \phi(y^*)$ as a fixed point. In matrix form, this is

$$\begin{bmatrix} y_t - y^* \\ u_t - u^* \end{bmatrix}^\top \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix} \begin{bmatrix} y_t - y^* \\ u_t - u^* \end{bmatrix} \ge 0. \tag{51}$$

We can also characterize strong convexity of $f$ and $g$ by similar quadratic constraints. Notably, the above constraint is very special in that it only manifests itself as separate quadratic constraints on each $(y_t, u_t)$. It is possible to specify quadratic constraints that couple different $t$ values. To achieve that, we follow Lessard et al. (2016) and adopt auxiliary sequences $\zeta, s$ together with a map $\Psi$ characterized by matrices $(A_\Psi, B^y_\Psi, B^u_\Psi, C_\Psi, D^y_\Psi, D^u_\Psi)$:

$$\begin{aligned} \zeta_{t+1} &= A_\Psi\zeta_t + B^y_\Psi y_t + B^u_\Psi u_t, \\ s_t &= C_\Psi\zeta_t + D^y_\Psi y_t + D^u_\Psi u_t. \end{aligned} \tag{52}$$

The equations (52) define an affine map $s = \Psi(y, u)$, where $s_t$ could be a function of all past $y_i$ and $u_i$ with $i \le t$. We consider the quadratic form $(s_t - s^*)^\top M(s_t - s^*)$ for a given matrix $M$ with $s^*$ and $\xi^*$ fixed points of (52). We note that the quadratic form is a function of $(y_0, \dots, y_t, u_0, \dots, u_t)$ that is determined by our choice of $(\Psi, M)$. In particular, we can recover constraint (51) with

$$\Psi = \left[\begin{array}{c|cc} A_\Psi & B^y_\Psi & B^u_\Psi \\ \hline C_\Psi & D^y_\Psi & D^u_\Psi \end{array}\right] = \left[\begin{array}{c|cc} \mathbf{0}_d & \mathbf{0}_d & \mathbf{0}_d \\ \hline \mathbf{0}_d & \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{0}_d & \mathbf{I}_d \end{array}\right], \qquad M = \begin{bmatrix} L^2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{I}_d \end{bmatrix}. \tag{53}$$

[11] This linear dynamical system can represent any first-order methods.

Combining the dynamics (50) with the map $\Psi$ (by eliminating $y_t$), we obtain

$$\begin{aligned}\begin{bmatrix}\xi_{t+1}\\ \zeta_{t+1}\end{bmatrix} &= \begin{bmatrix}A & 0\\ B_\Psi^y C & A_\Psi\end{bmatrix}\begin{bmatrix}\xi_t\\ \zeta_t\end{bmatrix} + \begin{bmatrix}B\\ B_\Psi^u + B_\Psi^y D\end{bmatrix} u_t,\\ s_t &= \begin{bmatrix}D_\Psi^y C & C_\Psi\end{bmatrix}\begin{bmatrix}\xi_t\\ \zeta_t\end{bmatrix} + \begin{bmatrix}D_\Psi^u + D_\Psi^y D\end{bmatrix} u_t.\end{aligned} \tag{54}$$

More succinctly, (54) can be written as

$$\begin{aligned}x_{t+1} &= \hat{A}x_t + \hat{B}u_t\\ s_t &= \hat{C}x_t + \hat{D}u_t\end{aligned}, \qquad \text{where } x_t \triangleq \begin{bmatrix}\xi_t\\ \zeta_t\end{bmatrix}. \tag{55}$$

With these definitions in hand, we now state the main result of verifying exponential convergence. Basically, we build a Linear Matrix Inequality (LMI) to guide the search for the parameters of a quadratic Lyapunov function in order to establish a rate bound.

**Theorem 9** (Zhang et al. (2021))**.** *Consider the dynamical system (50). Suppose the vector field $F$ satisfies the IQC $(\Psi, M)$ and define $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ according to (53)–(55), we have the following linear matrix inequality (LMI):*

$$\begin{bmatrix}\hat{A}^\top P\hat{A} - \rho^2 P & \hat{A}^\top P\hat{B}\\ \hat{B}^\top P\hat{A} & \hat{B}^\top P\hat{B}\end{bmatrix} + \lambda\begin{bmatrix}\hat{C} & \hat{D}\end{bmatrix}^\top M\begin{bmatrix}\hat{C} & \hat{D}\end{bmatrix} \preceq 0. \tag{56}$$

*If this LMI is feasible for some $P \succ 0$, $\lambda \geq 0$ and $\rho > 0$, we have*

$$(x_{t+1} - x^*)^\top (P \otimes \mathbf{I}_d)(x_{t+1} - x^*) \leq \rho^2 (x_t - x^*)^\top (P \otimes \mathbf{I}_d)(x_t - x^*). \tag{57}$$

*Consequently, for any $\xi_0$ and $\zeta_0 = \zeta^*$, we obtain*

$$\|\xi_t - \xi^*\|_2^2 \leq \mathrm{cond}(P)\rho^{2t}\|\xi_0 - \xi^*\|_2^2. \tag{58}$$

**Remark 2.** *The LMI (56) can be extended to the case of multiple constraints with $(\Psi_i, M_i)$ (see (Lessard et al., 2016, Page 12) for details).*

To apply Theorem 9, we seek to solve the semidefinite program (SDP) of finding the minimal $\rho$ such that the LMI (56) is feasible. For simple algorithms, one can typically solve the SDP analytically. Nevertheless, one may only get a numerical proof when the algorithm of interest is complicated and the resulting SDP is hard to solve.

### B.1 Analyzing Alt-GDA with IQC framework

Recall that we are concerned with the bilinear saddle point problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m}\max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x}) + \mathbf{x}^\top\mathbf{B}\mathbf{y} - g(\mathbf{y}). \tag{59}$$

For Alt-GDA, it has the following update rule:

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t - \eta\mathbf{B}\mathbf{y}_t - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_t)\\ \mathbf{y}_{t+1} &= \mathbf{y}_t + \eta\mathbf{B}^\top\mathbf{x}_{t+1} - \eta\nabla_{\mathbf{y}} g(\mathbf{y}_t)\end{aligned} \tag{60}$$

We can frame it as a linear dynamical system in feedback with the state matrices:

$$\left[\begin{array}{c|c}A & B\\ \hline C & D\end{array}\right] = \left[\begin{array}{cc|cc}\mathbf{I}_m & -\eta\mathbf{B} & -\eta\mathbf{I}_m & \mathbf{0}_{m\times n}\\ \eta\mathbf{B}^\top & \mathbf{I}_n - \eta^2\mathbf{B}^\top\mathbf{B} & -\eta^2\mathbf{B}^\top & -\eta\mathbf{I}_n\\ \hline \mathbf{I}_m & \mathbf{0}_{m\times n} & \mathbf{0}_m & \mathbf{0}_{m\times n}\\ \mathbf{0}_{n\times m} & \mathbf{I}_n & \mathbf{0}_{n\times m} & \mathbf{0}_n\end{array}\right], \tag{61}$$

In our case, we have $\xi_t = y_t = [\mathbf{x}_t^\top, \mathbf{y}_t^\top]^\top$ and $u_t = [\nabla_{\mathbf{x}} f(\mathbf{x}), -\nabla_{\mathbf{y}} g(\mathbf{y})]$. Further, we use the weighted off-by-one IQC defined in Lessard et al. (2016) with the following representation (let $d = m + n$):

$$\Psi = \left[\begin{array}{c|cc}A_\Psi & B_\Psi^y & B_\Psi^u\\ \hline C_\Psi & D_\Psi^y & D_\Psi^u\end{array}\right] = \left[\begin{array}{c|cc}\mathbf{0}_d & -L\mathbf{I}_d & \mathbf{I}_d\\ \hline \rho^2\mathbf{I}_d & L\mathbf{I}_d & -\mathbf{I}_d\\ \mathbf{0}_d & -\mu\mathbf{I}_d & \mathbf{I}_d\end{array}\right], \tag{62}$$

and

$$M_1 = \begin{bmatrix} \mathbf{0}_m & \mathbf{0}_{m\times n} & \mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \mathbf{0}_n & \mathbf{0}_{n\times m} & \mathbf{0}_n \\ \mathbf{I}_m & \mathbf{0}_{m\times n} & \mathbf{0}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \mathbf{0}_n & \mathbf{0}_{n\times m} & \mathbf{0}_n \end{bmatrix}, \quad M_2 = \begin{bmatrix} \mathbf{0}_m & \mathbf{0}_{m\times n} & \mathbf{0}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \mathbf{0}_n & \mathbf{0}_{n\times m} & \mathbf{I}_n \\ \mathbf{0}_m & \mathbf{0}_{m\times n} & \mathbf{0}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \mathbf{I}_n & \mathbf{0}_{n\times m} & \mathbf{0}_n \end{bmatrix}. \tag{63}$$

With all these matrices defined, we can solve the SDP problem (56) with bisection search on $\rho$. However, we can only solve SDPs with relatively small $m$ and $n$ in practice. Towards this end, we prove that we can reduce any problem of (8) to the case with $m = n = 1$, which is numerically easy to solve. First, we assume without loss of generality that $m \leq n$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ is diagonal. In the case that $\mathbf{B}$ is not diagonal, we can do singular value decomposition to get $\mathbf{B} = \mathbf{U}\hat{\mathbf{B}}\mathbf{V}^\top$ where $\hat{\mathbf{B}} \in \mathbb{R}^{m\times n}$ is diagonal and then absorb $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ into $\mathbf{x}$ and $\mathbf{y}$ to get the following equivalent problem:

$$\min_{\hat{\mathbf{x}}\in\mathbb{R}^m} \max_{\hat{\mathbf{y}}\in\mathbb{R}^n} f(\mathbf{U}\hat{\mathbf{x}}) + \hat{\mathbf{x}}^\top \hat{\mathbf{B}}\hat{\mathbf{y}} - g(\mathbf{V}\hat{\mathbf{y}}), \tag{64}$$

where $\hat{\mathbf{x}} = \mathbf{U}^\top\mathbf{x}$ and $\hat{\mathbf{y}} = \mathbf{V}^\top\mathbf{y}$. Further, one can show $f'(\mathbf{x}) = f(\mathbf{U}\mathbf{x})$ (or $g'(\mathbf{y}) = g(\mathbf{V}\mathbf{y})$) is also $L$-smooth and $\mu$-strongly convex as $f$ (or $g$) is. This is because $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal matrices. Therefore, we can assume $\mathbf{B}$ to be diagonal without loss of generality. Next, we prove that the IQC-certified rate of Alt-GDA for (8) is no worse than the rate of the same problem with $m = n = 1$. Formally, we have the following theorem.

**Theorem 8.** *Using the IQC framework to analyze the convergence rate of Alt-GDA on problem (8), we can simply assume $m = n = 1$ if $\mathbf{B}$ is diagonal. Let $\rho_{m,n}$ be the IQC-certified rate for problem (8) with $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, then we have $\rho_{m,n} \leq \rho_{1,1}$.*

*Proof.* For IQC, we basically search for a quadratic Lyapunov function by solving a SDP problem (56). By (61)-(63), we have the linear matrix inequality (LMI) as follows:

$$\left[\hat{A}, \hat{B}\right]^\top P \left[\hat{A}, \hat{B}\right] + \left[\hat{C}, \hat{D}\right]^\top M \left[\hat{C}, \hat{D}\right] \preceq \rho^2 \left[\mathbf{I}, \mathbf{0}\right]^\top P \left[\mathbf{I}, \mathbf{0}\right], \tag{65}$$

with

$$\hat{A} = \begin{bmatrix} \begin{bmatrix} \mathbf{I}_m & -\eta\mathbf{B} \\ \eta\mathbf{B}^\top & \mathbf{I}_n - \eta^2\mathbf{B}^\top\mathbf{B} \end{bmatrix} & \mathbf{0}_d \\ -L\mathbf{I}_d & \mathbf{0}_d \end{bmatrix}, \ \hat{B} = \begin{bmatrix} \begin{bmatrix} -\eta\mathbf{I}_m & \mathbf{0}_{m\times n} \\ -\eta^2\mathbf{B}^\top & -\eta\mathbf{I}_n \end{bmatrix} \\ \mathbf{I}_d \end{bmatrix},$$

$$\hat{C} = \begin{bmatrix} L\mathbf{I}_d & \rho^2\mathbf{I}_d \\ -\mu\mathbf{I}_d & \mathbf{0}_d \end{bmatrix}, \ \hat{D} = \begin{bmatrix} -\mathbf{I}_d \\ \mathbf{I}_d \end{bmatrix}, \ M = \begin{bmatrix} \mathbf{0}_m & \mathbf{0}_{m\times n} & \lambda_1\mathbf{I}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \mathbf{0}_n & \mathbf{0}_{n\times m} & \lambda_2\mathbf{I}_n \\ \lambda_1\mathbf{I}_m & \mathbf{0}_{m\times n} & \mathbf{0}_m & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & \lambda_2\mathbf{I}_n & \mathbf{0}_{n\times m} & \mathbf{0}_n \end{bmatrix}.$$

Given that $\mathbf{B}$ is diagonal, one can show that both $\left[\hat{A}, \hat{B}\right]$ and $\left[\hat{C}, \hat{D}\right]$ have very special structure. In particular, we can permute them column-wise and row-wise to get block-diagonal matrices:

$$\left[\hat{A}, \hat{B}\right] = \mathbf{U} \underbrace{\begin{bmatrix} \begin{bmatrix} 1 & -\eta\mathbf{B}_{11} & 0 & 0 & -\eta & 0 \\ \eta\mathbf{B}_{11} & 1-\eta^2\mathbf{B}_{11}^2 & 0 & 0 & -\eta^2\mathbf{B}_{11} & -\eta \\ -L & 0 & 0 & 0 & 1 & 0 \\ 0 & -L & 0 & 0 & 0 & 1 \end{bmatrix} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \begin{bmatrix} 1 & 0 & -\eta \\ -L & 0 & 1 \end{bmatrix} \end{bmatrix}}_{\triangleq \mathbf{Q}_1} \mathbf{V} \tag{66}$$

where both $\mathbf{U}$ and $\mathbf{V}$ are permutation matrices. In more detail, the $(1,1)$ block is repeated $r = \min(m, n)$ times, where $\mathbf{B}_{11}$ is replaced by $\mathbf{B}_{ii}$, $i = 1, \ldots, r$ in each successive block, and the smaller block is repeated $d - r$ times. Also we can do the same for $\left[\hat{C}, \hat{D}\right] = \mathbf{U}\mathbf{Q}_2\mathbf{V}$ and $\left[\mathbf{I}, \mathbf{0}\right] = \mathbf{U}\mathbf{Q}_3\mathbf{V}$. Each diagonal block of $\mathbf{Q}_3$ is either $\left[\mathbf{I}_4, \mathbf{0}_{4\times 2}\right]$ or $\left[\mathbf{I}_2, \mathbf{0}_{2\times 1}\right]$. Hence, we can write (65) in the following form:

$$\mathbf{V}^\top\mathbf{Q}_1^\top\mathbf{U}^\top P\mathbf{U}\mathbf{Q}_1\mathbf{V} + \mathbf{V}^\top\mathbf{Q}_2^\top\mathbf{U}^\top M\mathbf{U}\mathbf{Q}_2\mathbf{V} \preceq \rho^2\mathbf{V}^\top\mathbf{Q}_3^\top\mathbf{U}^\top P\mathbf{U}\mathbf{Q}_3\mathbf{V}, \tag{67}$$

whose feasible set is equivalent to

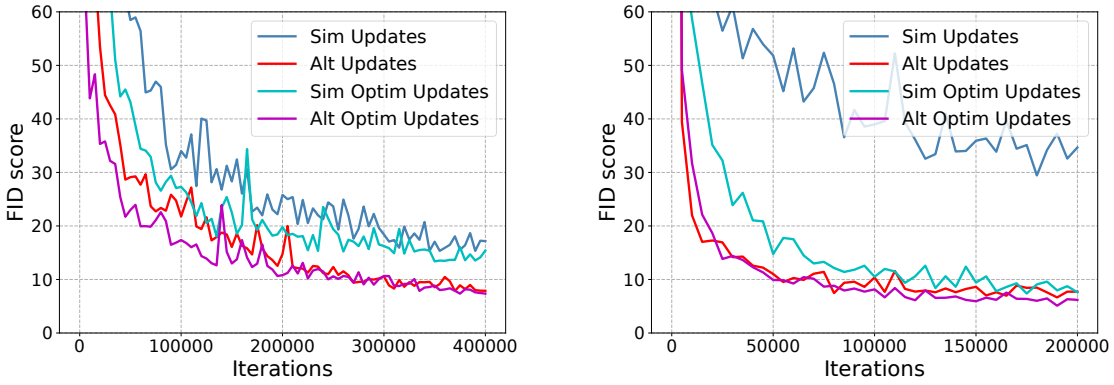
$$\mathbf{Q}_1^\top\mathbf{U}^\top P\mathbf{U}\mathbf{Q}_1 + \mathbf{Q}_2^\top\mathbf{U}^\top M\mathbf{U}\mathbf{Q}_2 \preceq \rho^2\mathbf{Q}_3^\top\mathbf{U}^\top P\mathbf{U}\mathbf{Q}_3. \tag{68}$$

It is easy to see that $\mathbf{U}^\top M \mathbf{U}$ has the same block-diagonal structure as $\mathbf{Q}_1$ and $\mathbf{Q}_2$. If we further restrict $\mathbf{U}^\top P \mathbf{U}$ to have the same block-diagonal structure as $\mathbf{Q}_1$ and $\mathbf{Q}_2$, then it suffices to pick a $\rho$ so that each diagonal block of the LMI (68) holds. Moreover, the LMI of each diagonal block is the LMI of the case $m = n = 1$, except for the last $n - m$ blocks, for which we have

$$\begin{bmatrix} 1 & 0 \\ -L & 1 \\ -\eta & 0 \end{bmatrix}^\top P \begin{bmatrix} 1 & 0 \\ -L & 1 \\ -\eta & 0 \end{bmatrix} + \begin{bmatrix} L & \rho^2 & -1 \\ -\mu & 0 & 1 \end{bmatrix}^\top \begin{bmatrix} 0 & \lambda_2 \\ \lambda_2 & 0 \end{bmatrix} \begin{bmatrix} L & \rho^2 & -1 \\ -\mu & 0 & 1 \end{bmatrix} \preceq \rho^2 \left[\mathbf{I}, \mathbf{0}\right]^\top P \left[\mathbf{I}, \mathbf{0}\right]. \tag{69}$$

This is the LMI for minimizing a $\mu$-strongly convex $L$-smooth function (see Lessard et al. (2016)), which has a better convergence rate compared to our minimax problem (i.e., any feasible $\rho$ of the LMI for 1-dimension minimax problem is also feasible for (69)) because it is a special case of (8) with $\mathbf{B} = 0$ in the 1-dimensional case. So far, we show that as long as the LMI for the case of $m = n = 1$ holds, then the general case also holds since the general case can be decomposed into many 1-dimensional problems. This completes the proof. □

## C Additional Results on SVHN




**(a)** SGD **(b)** AMSGrad

**Figure 8:** **(a)** ResNet model trained with SGD on SVHN. **(b)** ResNet model trained with AMSGrad on SVHN. Alternating algorithms dominate simultaneous ones. Again, the use of optimism makes little difference for alternating algorithms.

## D Implementation Details for Generative Adversarial Networks

For our experiments, we used the PyTorch[12] deep learning framework. For experiments, we compute the FID score using the provided implementation in Tensorflow[13] for consistency with related works.

**Optimistic update rule:** the simultaneous version of optimistic gradient descent-ascent takes the following form:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - 2\eta \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) + \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \\ \mathbf{y}_{t+1} &= \mathbf{y}_t + 2\eta \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) - \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \end{aligned} \tag{70}$$

By comparison, the alternating version iterates as follows:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - 2\eta \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) + \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \\ \mathbf{y}_{t+1} &= \mathbf{y}_t + 2\eta \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t) - \eta \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_{t-1}) \end{aligned} \tag{71}$$

**Loss functions:** For DCGAN experiments, we used WGAN-GP objective (Gulrajani et al., 2017). For ResNet experiments, we used the hinge version of the adversarial non-saturating loss, see Miyato et al. (2018). As a reference, our ResNet architectures for CIFAR-10 and SVHN (Netzer et al., 2011) have approximately 85 layers in total for the generator and discriminator, including the nonlinearity and the normalization layers. This ResNet architecture was also used in Chavdarova et al. (2021), see Appendix E 2.2 of Chavdarova et al. (2021).

[12] https://pytorch.org/
[13] https://github.com/bioinf-jku/TTUR

**Hyperparameters:** We conduct grid search over the step size (and $\beta_2$ for AMSGrad) for each setting. For SGD, the search range of step-size is $\{5e{-}4, 1e{-}3, 2e{-}3, 5e{-}3, 1e{-}2, 2e{-}2\}$. For AMSGrad, the search range of step-size is $\{5e{-}5, 1e{-}4, 2e{-}4, 5e{-}4, 1e{-}3, 2e{-}3\}$ while the search range of $\beta_2$ is $\{0.9, 0.99, 0.999\}$. We report the optimal hyperparameters used in the following tables. All these hyperparameters are tuned with random seed 1. We have also tried other seeds (including seed 2 and 3) and the optimal hyperparameters could be different with different random seeds. However, the optimal curves across different random seeds look similar.

**Table 1:** Hyperparameters for DCGAN experiments.

| Parameter | Sim-SGD | Alt-SGD | Sim-AMSGrad | Alt-AMSGrad |
|---|---|---|---|---|
| batch-size | 128 | 128 | 128 | 128 |
| step-size (G) | 0.002 | 0.005 | 0.0005 | 0.0005 |
| step-size (D) | 0.002 | 0.005 | 0.0005 | 0.0005 |
| momentum | 0.0 | 0.0 | - | - |
| $\beta_1$ | - | - | 0.0 | 0.0 |
| $\beta_2$ | - | - | 0.999 | 0.999 |

**Table 2:** Hyperparameters for ResNet experiments on CIFAR-10.

| Parameter | Simultaneous | | | | Alternating | | | |
|---|---|---|---|---|---|---|---|---|
| | SGD | OSGD | AMSGrad | OAMSGrad | SGD | OSGD | AMSGrad | OAMSGrad |
| batch-size | 128 | 128 | 128 | 128 | 128 | 128 | 128 | 128 |
| step-size (G) | 0.005 | 0.005 | 0.0002 | 0.0002 | 0.01 | 0.01 | 0.0005 | 0.001 |
| step-size (D) | 0.005 | 0.005 | 0.0002 | 0.0002 | 0.01 | 0.01 | 0.0005 | 0.001 |
| momentum | 0.0 | 0.0 | - | - | 0.0 | 0.0 | - | - |
| $\beta_1$ | - | - | 0.0 | 0.0 | - | - | 0.0 | 0.0 |
| $\beta_2$ | - | - | 0.999 | 0.99 | - | - | 0.999 | 0.999 |

**Table 3:** Hyperparameters for ResNet experiments on SVHN.

| Parameter | Simultaneous | | | | Alternating | | | |
|---|---|---|---|---|---|---|---|---|
| | SGD | OSGD | AMSGrad | OAMSGrad | SGD | OSGD | AMSGrad | OAMSGrad |
| batch-size | 128 | 128 | 128 | 128 | 128 | 128 | 128 | 128 |
| step-size (G) | 0.002 | 0.005 | 0.0001 | 0.0002 | 0.005 | 0.01 | 0.0005 | 0.0002 |
| step-size (D) | 0.002 | 0.005 | 0.0001 | 0.0002 | 0.005 | 0.01 | 0.0005 | 0.0002 |
| momentum | 0.0 | 0.0 | - | - | 0.0 | 0.0 | - | - |
| $\beta_1$ | - | - | 0.0 | 0.0 | - | - | 0.0 | 0.0 |
| $\beta_2$ | - | - | 0.999 | 0.999 | - | - | 0.999 | 0.99 |