On Convergence of Lookahead in Smooth Games

Junsoo Ha
Seoul National University

Gunhee Kim
Seoul National University

Abstract

A key challenge in smooth games is that there is no general guarantee for gradient methods to converge to an equilibrium. Recently, Chavdarova et al. (2021) reported a promising empirical observation that Lookahead (Zhang et al., 2019) significantly improves GAN training. While promising, few theoretical guarantees has been studied for Lookahead in smooth games. In this work, we establish the first convergence guarantees of Lookahead for smooth games. We present a spectral analysis and provide a geometric explanation of how and when it actually improves the convergence around a stationary point. Based on the analysis, we derive sufficient conditions for Lookahead to stabilize or accelerate the local convergence in smooth games. Our study reveals that Lookahead provides a general mechanism for stabilization and acceleration in smooth games.

1 INTRODUCTION

In the last few years, a plethora of learning problems have been formulated as a game between multiple players (Goodfellow et al., 2014; Brock et al., 2019; Karras et al., 2019; Goodfellow et al., 2015; Silver et al., 2018; Vinyals et al., 2019). However, optimization of interdependent objectives is a non-trivial problem both in terms of complexity (Daskalakis et al., 2006, 2020) and convergence (Mertikopoulos et al., 2018). In particular, gradient-based methods often fail to converge and oscillate around an equilibrium even in a simple setting (Mescheder et al., 2018; Mertikopoulos et al., 2018). To tackle such non-convergence, a great effort has been devoted to developing efficient methods with provable convergence guarantees (Heusel et al., 2017; Mescheder et al., 2017, 2018; Balduzzi et al., 2018; Yadav et al., 2018; Mertikopoulos et al., 2019; Daskalakis et al., 2018; Letcher et al., 2019; Gidel et al., 2019b,a; Chavdarova et al., 2019; Adolphs et al., 2019; Mazumdar et al., 2019; Schäfer and Anandkumar, 2019; Peng et al., 2020; Lin et al., 2020; Mishchenko et al., 2020; Jelassi et al., 2020; Antonakopoulos et al., 2021).

Recently, Chavdarova et al. (2021) reported a promising empirical observation that Lookahead (Zhang et al., 2019) greatly improves the dynamics of bilinear games and GANs. In particular, they empirically report that Lookahead converges in bilinear games where gradient descent fails to, and GANs trained by Lookahead can outperform BigGANs (Brock et al., 2019) even with 1/30 parameters and small computation overhead.

Despite its great promise, the study of Chavdarova et al. (2021) was limited to empirical observations; hence, Lookahead optimizer still lacks theoretical investigation in smooth games. Specifically, crucial questions, such as convergence guarantees and effects of its hyperparameters on convergence, remain unexplained. In this work, we answer such questions for the first time. Our contributions are summarized as follows:

- We present a spectral analysis of Lookahead, and provide a geometric explanation of how and when it actually improves the local convergence. We interpret Lookahead as a geometric transformation of the eigenvalues, which improves the convergence around a stationary point.

- Based on the analysis, we derive sufficient conditions for Lookahead to stabilize or accelerate the local convergence in general smooth games, and global convergence in bilinear games. Our study reveals that Lookahead provides a general mechanism for stabilization and acceleration in smooth games. We summarize our findings in Table 1.

Our notation follows Goodfellow et al. (2016) and is summarized in Table A.1 of Appendix A. We defer the proofs of all theorems in this work to Appendix C and the results of additional experiments to Appendix E.

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

Table 1: **Summarized new results of this work**. The convergence rate ρ is defined by the spectral radius each method. We define *stabilization* as reduction of a radius from $\rho \geq 1$ to $\rho < 1$, and *acceleration* as further reduction from $\rho < 1$. Yellow checkmarks require extra assumptions on eigenvalues. See Section 2.1 for each method.

Method	Game	Converges	Stabilization / Acceleration	Conditions
LA-$\mathcal{A}$, $\mathcal{A}$ unstable	General	✓	Local stabilization $\rho_{\text{LA-}\mathcal{A}} < 1$	Theorem 4
LA-$\mathcal{A}$, $\mathcal{A}$ stable	General	✓	Local acceleration $\rho_{\text{LA-}\mathcal{A}} < \rho_{\mathcal{A}}$	Theorem 5
LA-GD_{Alt}	Bilinear	✓	Stabilization $\rho_{\text{LA-GD}_{\text{Alt}}} < 1$	Corollary 6
LA-GD_{Sim}	Bilinear	✓	Stabilization $\rho_{\text{LA-GD}_{\text{Sim}}} < 1$	Corollary 7
LA-EG_{Sim}	Bilinear	✓	Acceleration $\rho_{\text{LA-EG}_{\text{Sim}}} < \rho_{\text{EG}_{\text{Sim}}}$	Corollary 8
LA-PP_{Sim}	Bilinear	✓	Acceleration $\rho_{\text{LA-PP}_{\text{Sim}}} < \rho_{\text{PP}_{\text{Sim}}}$	Corollary C

2 PRELIMINARIES

2.1 Smooth Game Optimization

Game (Neumann and Morgenstern, 1944) is a model of interactions between multiple players; following Balduzzi et al. (2018), we define smooth games as follows.

Definition 1 (Smooth game)**.** *A set of smooth scalar functions* $\{f_i\}_{i=1}^n$ *with* $f_i : \mathbb{R}^d \to \mathbb{R}$ *such that* $d = \sum_{i=1}^n d_i$ *is called a smooth game between players* $i = 1, \ldots, n$ *with strategy spaces* $\{\mathbb{R}^{d_i}\}_{i=1}^n$.

Intuitively, each f_i represents the cost of player i's strategy $\mathbf{x}_i \in \mathbb{R}^{d_i}$ with respect to the other players' strategies $\mathbf{x}_{-i}$. The holy grail of game optimization is finding a Nash equilibrium (Nash, 1951), which is a strategy profile where no player has unilateral incentive to change its own strategy.

Definition 2 (Nash equilibrium)**.** *For a smooth game* $\{f_i\}_{i=1}^n$ *with strategy spaces* $\{\mathbb{R}^{d_i}\}_{i=1}^n$ *such that* $d = \sum_{i=1}^n d_i$, $\mathbf{x}^* \in \mathbb{R}^d$ *is a Nash equilibrium if* $f_i(\mathbf{x}^*) \leq f_i(\mathbf{x}_i, \mathbf{x}_{-i}^*), \forall \mathbf{x}_i \in \mathbb{R}^{d_i}$ *for each* i.

A strategy profile that merely exhibits zero gradient with respect to each player is called a stationary point of the game. A straightforward computational approach to find an equilibrium is to design a strategy update rule for each player. Such update rules define iterative *plays* between the players, and is often referred to as a *dynamics* of the game. However, it is known that gradient-based dynamics often fail to converge and oscillate around an equilibrium (Mertikopoulos et al., 2018; Gidel et al., 2019b). Such non-convergence is mainly due to (non-cooperative) interactions between multiple players, and is considered as a key challenge in smooth game optimization (Mescheder et al., 2017, 2018; Balduzzi et al., 2018; Schäfer and Anandkumar, 2019; Gidel et al., 2019b,a; Berard et al., 2020).

Below, we introduce a few first-order game dynamics. For notational simplicity, we use the derivative $\nabla_{\mathbf{x}} f(\cdot)$ to denote the concatenated partial derivatives $(\nabla_{\mathbf{x}_1} f_1(\cdot), \ldots, \nabla_{\mathbf{x}_n} f_n(\cdot))$ of a smooth game $\{f_i\}_{i=1}^n$, where each $\nabla_{\mathbf{x}_i} f_i(\cdot)$ denotes a derivative of a player i's cost function with respect to its own strategy.

Gradient Descent (GD) minimizes the cost function of each player with gradient descent. Its simultaneous dynamics $F_{\text{GD}_{\text{Sim}}}$ with a learning rate $\eta > 0$ is

$$\mathbf{x}^{(t+1)} = F_{\text{GD}_{\text{Sim}}}(\mathbf{x}^{(t)}) \overset{\text{def}}{=} \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}). \tag{1}$$

Meanwhile, its alternating variant $F_{\text{GD}_{\text{Alt}}}$ follows

$$F_{\text{GD}_{\text{Alt}}}(\mathbf{x}^{(t)}) \overset{\text{def}}{=} F_1 \circ \ldots \circ F_n(\mathbf{x}^{(t)}), \text{ where} \tag{2}$$

$$F_i(\mathbf{x}) \overset{\text{def}}{=} (\ldots, \mathbf{x}_{i-1}, \mathbf{x}_i - \eta \nabla_{\mathbf{x}_i} f_i(\mathbf{x}), \mathbf{x}_{i+1}, \ldots). \tag{3}$$

Proximal Point (PP) (Martinet, 1970) method computes an update by solving a proximal subproblem at each iteration. Its simultaneous dynamics $F_{\text{PP}_{\text{Sim}}}$ with a learning rate $\eta > 0$ is

$$\mathbf{x}^{(t+1)} = F_{\text{PP}_{\text{Sim}}}(\mathbf{x}^{(t)}) \overset{\text{def}}{=} \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^{(t+1)}). \tag{4}$$

Extragradient (EG) (Korpelevich, 1976) computes an update with an *extrapolated* gradient. Its simultaneous dynamics $F_{\text{EG}_{\text{Sim}}}$ with a learning rate $\eta > 0$ is

$$\mathbf{x}^{(t+1)} = F_{\text{EG}_{\text{Sim}}}(\mathbf{x}^{(t)}) \overset{\text{def}}{=} \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}^{(t)}), \text{ where} \tag{5}$$

$$\hat{\mathbf{x}}^{(t)} \overset{\text{def}}{=} \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}). \tag{6}$$

Lookahead (LA) (Zhang et al., 2019) is an optimizer that wraps around a base optimizer and takes a *backward* step for each k *forward* steps. Given a base dynamics $F_{\mathcal{A}}$ induced by an optimization method $\mathcal{A}$, its Lookahead dynamics $G_{\text{LA-}\mathcal{A}}$ with a period $k \in \mathbb{N}$ and a rate $\alpha \in (0, 1)$ is

$$\mathbf{x}^{(t+1)} = G_{\text{LA-}\mathcal{A}}(\mathbf{x}^{(t)}) \overset{\text{def}}{=} (1 - \alpha)\mathbf{x}^{(t)} + \alpha F_{\mathcal{A}}^k(\mathbf{x}^{(t)}). \tag{7}$$

2.2 Related Work

The convergence analysis of smooth games dates several decades back and has been established in the saddle-point (Korpelevich, 1976; Benzi et al., 2005) and variational inequality problems (Rockafellar, 1976; Tseng, 1995), where each can be reformulated as n-player games under certain assumptions (Scutari et al., 2010). Specifically, Rockafellar (1976) proved the linear convergence of PP in bilinear and strongly monotone games, and Tseng (1995); Facchinei and Pang (2003) did the same for EG. Similarly, Nemirovski (2004) proved the linear convergence of EG in monotone games, and Juditsky et al. (2011) did the same in stochastic settings.

As a variety of learning problems are formulated as a game between multiple players (Goodfellow et al., 2014; Madry et al., 2018; Vinyals et al., 2019), game optimization has regained considerable attentions. For instance, (Daskalakis et al., 2018) rediscovered optimistic gradient (OG) Popov (1980) for GAN training, and Gidel et al. (2019a) proved its linear convergence for strongly monotone games. Chavdarova et al. (2019); Jelassi et al. (2020); Mishchenko et al. (2020); Antonakopoulos et al. (2021) proposed variants of EG, and Mokhtari et al. (2020) established an unifying theory for PP, EG and OG in strongly-convex strongly-concave games.

Meanwhile, recent studies have shown that a careful manipulation of a game dynamics can improve its convergence. For example, Gidel et al. (2019b) proved that adding a negative momentum can make non-convergent $\mathrm{GD}_{\mathrm{Alt}}$ to converge in bilinear games, and Azizian et al. (2020) showed that a momentum can accelerate EG in smooth games. Yoon and Ryu (2021) proposed an *anchoring* method for EG in convex-concave games, and established an acceleration in gradient norms. Regularizers that induce better convergence guarantees have been extensively studied as well (Mescheder et al., 2017; Balduzzi et al., 2018; Schäfer and Anandkumar, 2019; Letcher et al., 2019; Adolphs et al., 2019; Mazumdar et al., 2019; Wang et al., 2020; Hemmat et al., 2020).

Lastly, the recent study of Chavdarova et al. (2021) has shown that augmenting a game dynamics with Lookahead (Zhang et al., 2019), i.e., taking a *backward* step for each k *forward* steps, significantly improves GAN training. In this work, we establish the first local convergence guarantees of Lookahead in smooth games, and show that it provides a general mechanism for local stabilization and acceleration in smooth games.

3 THE SPECTRAL CONTRACTION

In this section, we interpret Lookahead as a geometric transformation of the eigenvalues which improves the convergence of smooth games by reducing the spectral radius of a game dynamics. We demonstrate such *spectral contraction effect* by analyzing a simple exemplar bilinear game that has a unique Nash equilibrium $(0,0)$:

$$\min_{x_1\in\mathbb{R}}\max_{x_2\in\mathbb{R}}\ x_1\cdot x_2. \tag{8}$$

This game has been studied as a representative toy example in game optimization due to its oscillating dynamics (Gidel et al., 2019b,a). Notably, simultaneous gradient descent $\mathrm{GD}_{\mathrm{Sim}}$ diverges from the Nash equilibrium of Equation 8, and even an advanced method such as negative momentum (Gidel et al., 2019b) fails to stabilize such an instability (Zhang and Yu, 2020). The following result shows Lookahead can stabilize $\mathrm{GD}_{\mathrm{Sim}}$.

Example 1 (Stabilization)**.** *Lookahead dynamics $G_{LA\text{-}GD_{Sim}}$ with $\eta > 0, k \in \mathbb{N}, \alpha \in (0,1)$ converges to the Nash equilibrium of Equation 8 if k satisfies $\Re((1+i\eta)^k) < 1$ and α is small enough.*

A precise threshold for α and a similar result for $\mathrm{GD}_{\mathrm{Alt}}$ can be found in Appendix D. For a small $\eta > 0$, there exists $k \in \mathbb{N}$ that makes the real part of the complex number $(1+i\eta)^k$ negative, i.e., $\Re((1+i\eta)^k) < 0$. Hence, Example 3 implies that Lookahead can stabilize $\mathrm{GD}_{\mathrm{Sim}}$.

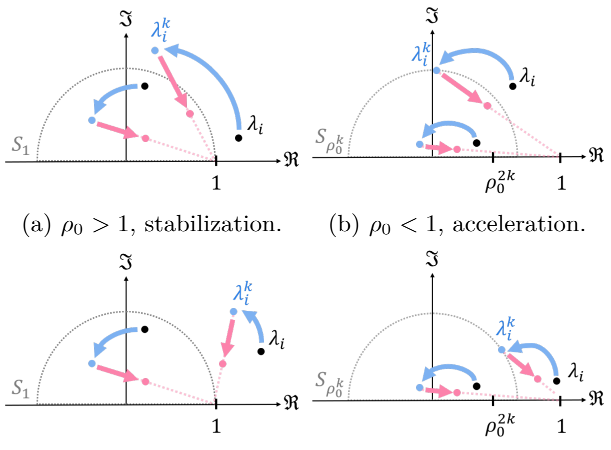
However, such stabilization effect raises a natural question: would there be an advantage for using Lookahead when its base dynamics is already stable? The next example studies Lookahead dynamics of $\mathrm{PP}_{\mathrm{Sim}}$, which is known to be convergent in Equation 8 (Gidel et al., 2019a), and provides an affirmative answer.

Example 2 (Acceleration)**.** *Lookahead dynamics $G_{LA\text{-}PP_{Sim}}$ with $\eta \in (0,1), k \in \mathbb{N}, \alpha \in (0,1)$ converges to the Nash equilibrium of Equation 8. The rate of convergence improves upon its base dynamics $F_{PP_{Sim}}$ if k satisfies $\Re((1+i\eta)^k) < 1$ and α is large enough.*

The threshold for α and a similar result for $\mathrm{EG}_{\mathrm{Sim}}$ can be found in Appendix D. Figure 1 illustrates a geometric interpretation of Lookahead. In short, Lookahead improves the convergence by rotating and pulling the eigenvalues of its base dynamics. Specifically, k *forward* steps of each Lookahead iteration rotates the eigenvalues, and a *backward* step pulls them into a circle with a radius smaller than their maximal modulus. This results in a reduction of the spectral radius, which determines the local convergence rate around a stationary point (Azizian et al., 2020). The following proposition captures such *spectral contraction effect* of Lookahead; we denote the spectral radius by $\rho(\cdot)$, and the sets of modulus-filtered eigenvalues by $\lambda_{\geq 1}(\cdot)$ and $\lambda_{\max}(\cdot)$.

Proposition 1 (Spectral contraction)**.** *Let $\mathbf{X} \in \mathbb{R}^{m\times m}$ be the Jacobian of a dynamics at a stationary point. Denote its spectral radius by $\rho_0 \overset{\mathsf{def}}{=} \rho(\mathbf{X})$ and the radius of its Lookahead dynamics with $k\in\mathbb{N}, \alpha\in(0,1)$ by $\rho_k(\alpha) \overset{\mathsf{def}}{=} \rho((1-\alpha)\boldsymbol{I} + \alpha\mathbf{X}^k)$. Then, we get either stabilization $(\rho_k(\alpha) < 1)$ or acceleration $(\rho_k(\alpha) < \rho_0^k)$ depending on the spectral radius ρ_0 of its base dynamics as follows:*

- *For $\rho_0 > 1$, $\rho_k(\alpha) < 1 \iff \tau_{k|\geq 1} < 1, \alpha < c_1$,*
- *For $\rho_0 = 1$, $\rho_k(\alpha) < 1 \iff \tau_{k|max} < 1$,*
- *For $\rho_0 < 1$, $\rho_k(\alpha) < \rho_0^k \iff \tau_{k|max} < \rho_0^{2k}, \alpha > c_2$,*



(a) $\rho_0 > 1$, stabilization.

(b) $\rho_0 < 1$, acceleration.

(c) $\rho_0 > 1$, k not satisfied.

(d) $\rho_0 < 1$, k not satisfied.

Figure 1: **Illustration of the spectral contraction effect in Proposition 1.** The dots represent the eigenvalues of a base dynamics (black), their rotated values (blue), and the eigenvalues of a Lookahead dynamics (red). We denote the spectral radius of each base dynamics by ρ_0. **Top:** k *forward* steps of each Lookahead iteration rotate the eigenvalues λ to λ^k, and a *backward* step pulls them into a circle with a smaller radius. This results in a reduction of the spectral radius, which improves the stability and convergence to a stationary point. **Bottom:** However, when there exists an eigenvalue that is not rotated left enough, i.e., $\tau_{k|\geq 1} \geq 1$ or $\tau_{k|\max} \geq \rho_0^{2k}$, no α can reduce the radius.

where $\tau_{k|\geq 1} \overset{\text{def}}{=} \max_{\lambda_i \in \lambda_{\geq 1}(\mathbf{X})} \Re(\lambda_i^k)$ *and* $\tau_{k|max} \overset{\text{def}}{=} \max_{\lambda_i \in \lambda_{max}(\mathbf{X})} \Re(\lambda_i^k)$ *are the right-most real parts of the rotated eigenvalues, and* $c_1, c_2 \in \mathbb{R}$ *are some constants that depend on* k*.*

The Jacobian of a Lookahead dynamics at a stationary point can be written as $(1-\alpha)\mathbf{I} + \alpha\mathbf{X}^k$. Hence, each statement provides an exact condition for Lookahead to reduce the spectral radius ρ_0 of its base dynamics. For an unstable base dynamics, i.e., $\rho_0 > 1$, the first case implies that a small α can reduce the spectral radius $\rho_k(\alpha)$ to smaller than 1 when k rotates the eigenvalues $\lambda_{\geq 1}(\mathbf{X})$ *left enough*, i.e., $\tau_{k|\geq 1} < 1$. For a stable base dynamics, i.e., $\rho_0 < 1$, the last case implies that a large α can further reduce the radius ρ_0 to $\rho_k(\alpha)$ if k rotates the eigenvalues $\lambda_{\max}(\mathbf{X})$ *left enough*, i.e., $\tau_{k|\max} < \rho_0^{2k}$.

While the theorem shows that Lookahead can reduce the spectral radius, it does not predict the amount of reduction that could be made for a given k and α. To fill this gap, we derive precise bounds on the optimal contraction for a fixed k in terms of spectral quantities.

Proposition 2 (Contraction bounds)**.** *Let* $\mathbf{X} \in \mathbb{R}^{m\times m}$ *be the Jacobian of a dynamics at an equilibrium, and denote its spectral radius by* $\rho_0 \overset{\text{def}}{=} \rho(\mathbf{X})$ *and the optimal radius of its Lookahead dynamics with* $k \in \mathbb{N}$ *by* $\rho_k^* \overset{\text{def}}{=} \inf_{\alpha\in(0,1)} \rho_k(\alpha)$*. Then, for* $\tau_k \overset{\text{def}}{=} \max_{\lambda_i \in \lambda(\mathbf{X})} \Re(\lambda_i^k)$*, the following statements hold:*

- *For* $\rho_0 \geq 1$, ${\rho_k^*}^2 \leq 1 - \frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} < 1$ *if* $\tau_k < 1$*,*
- *For* $\rho_0 < 1$, ${\rho_k^*}^2 \leq 1 - \frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} < \rho_0^{2k}$ *if* $\tau_k < \rho_0^{2k}$*,*
- *A lower bound* $\rho_k^* \geq \max_{\lambda_i \in D} |\lambda_i|^k$ *holds for the eigenvalues inside the disk* $D \overset{\text{def}}{=} \{\lambda_i \in \lambda(\mathbf{X}) : |\lambda_i^k - \frac{1}{2}| < \frac{1}{2}\}$*.*

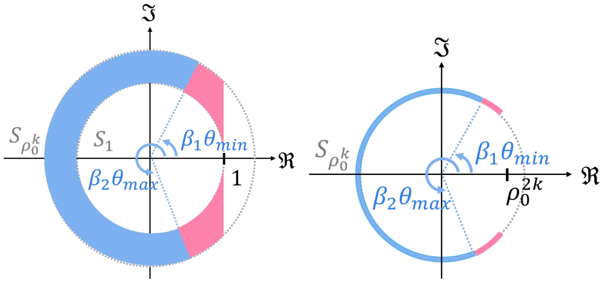
The upper bounds are monotonically increasing with respect to the right-most real part of the rotated eigenvalues, i.e., τ_k. Hence, the upper bounds show that τ_k is the key quantity that determines the amount of contraction. For instance, if we could choose k such that $\tau_k < 0$, i.e., rotates all the eigenvalues to the left half-plane, we may expect a spectral contraction $\frac{\rho_k^*}{\rho_0^k} < \frac{1}{\sqrt{1+\rho_0^{2k}}}$. On the other hand, the lower bound highlights the possible failure case of the contraction. For instance, when there exists a large eigenvalue that resists to be rotated and remains inside the disk $\left|z - \frac{1}{2}\right| < \frac{1}{2}$, e.g., a real eigenvalue $\lambda_i < 1$ such that $\lambda_i \approx 1$, the contraction ends up with a restrictive lower bound $\rho_k^* \geq |\lambda_i|^k \approx 1$.

So far, we have seen that a proper choice of k is crucial for spectral contraction. For example, Theorem 1 shows the contraction takes place if and only if $\tau_{k|\geq 1} < 1$ and $\tau_{k|\max} < \rho_0^{2k}$. At this point, a natural question arises: how do we choose such k? We answer this question with Lemma 3, sufficient conditions for k to rotate the eigenvalues *left enough*, i.e., $\tau_{k|\geq 1} < 1$ and $\tau_{k|\max} < \rho_0^{2k}$.

Lemma 3 (Sufficient conditions for left-rotating k)**.** *Let* $\mathbf{X} \in \mathbb{R}^{m\times m}$ *be a Jacobian that can be written as* $\mathbf{X} = \mathbf{I} - \eta\mathbf{J}$ *for some* $\mathbf{J} \in \mathbb{R}^{m\times m}$ *and* $\eta > 0$*. Assume that a subset of the eigenvalues* $S \subseteq \lambda(\mathbf{X})$ *contains non-reals only, and every element of* S *has its conjugate pair in* S*. Then, for* $\rho_0 \overset{\text{def}}{=} \rho(\mathbf{X})$, $\tau_k \overset{\text{def}}{=} \max_{\lambda_i \in S} \Re(\lambda_i^k)$, $\theta_{min} \overset{\text{def}}{=} \min_{\lambda_i \in S} |\operatorname{Arg}(\lambda_i)|$, $\theta_{max} \overset{\text{def}}{=} \max_{\lambda_i \in S} |\operatorname{Arg}(\lambda_i)|$*, the following statements hold:*

- *When* $\rho_0 > 1$*, the eigenvalues* S *are left-rotated so that* $\tau_k < 1$ *if* $k \in (\beta_1, \beta_2)$*, where* $\beta_1, \beta_2 > 0$ *are such that* $\beta_1\theta_{min} = \arccos \rho_0^{-\beta_1}$ *and* $\beta_2\theta_{max} = 2\pi - \arccos \rho_0^{-\beta_2}$*.*
- *When* $\rho_0 < 1$*, the eigenvalues* S *are left-rotated so that* $\tau_k < \rho_0^{2k}$ *if* $k \in (\beta_1, \beta_2)$*, where* $\beta_1, \beta_2 > 0$ *are such that* $\beta_1\theta_{min} = \arccos \rho_0^{\beta_1}$ *and* $\beta_2\theta_{max} = 2\pi - \arccos \rho_0^{\beta_2}$*.*

The existence of a feasible $k \in (\beta_1, \beta_2)$ *is guaranteed for a small enough* $\eta > 0$ *when the imaginary conditioning* $\max_{\lambda_i, \lambda_j \in S} |\Im(\lambda_i)/\Im(\lambda_j)|$ *of* S *is smaller than* 3*.*



(a) Left-rotated $\lambda_{\geq 1}(\mathbf{X})$. (b) Left-rotated $\lambda_{\max}(\mathbf{X})$.

Figure 2: **Illustration of the eigenvalues rotated by Lemma 3**. Each blue region contains a subset $S \subseteq \lambda(\mathbf{X})$ of the eigenvalues rotated by $k \in (\beta_1, \beta_2)$, i.e., $\{\lambda_i^k | \lambda_i \in S\}$, where $\beta_1, \beta_2 > 0$ are defined as in Lemma 3. Each red region contributes to the non-necessity. **(a)** The first case where the base dynamics is unstable, i.e., $\rho_0 > 1$ and $S = \lambda_{\geq 1}(\mathbf{X})$. The lower bound $k > \beta_1$ rotates the eigenvalues S *left enough*, i.e., $\tau_{k|\geq 1} < 1$, and the upper bound $k < \beta_2$ prevents S from being over-rotated. **(b)** The second case where the base dynamics is stable, i.e., $\rho_0 < 1$ and $S = \lambda_{\max}(\mathbf{X})$. Similarly, the lower bound $k > \beta_1$ rotates S and the upper bound $k < \beta_2$ prevents S from being over-rotated.

Note that τ_k is defined over a *subset* of the eigenvalues, and η can be seen as a learning rate of the dynamics. Hence, Lemma 3 can be interpreted as sufficient conditions for k to left-rotate the eigenvalues S, and suggests that such k exists for a small learning rate when the imaginary conditioning of S is smaller than 3. For an unstable dynamics, i.e., $\rho_0 > 1$ and $S = \lambda_{\geq 1}(\mathbf{X})$, the first case gives a condition for k to rotate the eigenvalues so that $\tau_{k|\geq 1} < 1$. On the other hand, for a stable dynamics, i.e., $\rho_0 < 1$ and $S = \lambda_{\max}(\mathbf{X})$, the second case gives a condition for k to rotate the eigenvalues so that $\tau_{k|\max} < \rho_0^{2k}$. For such k's, Theorem 1 promises a spectral contraction. This combination of Proposition 1 and Lemma 3 establishes stabilization and acceleration guarantees of Lookahead in smooth games. Below, we denote the Jacobian of a dynamics F by $\nabla_{\mathbf{x}} F(\cdot)$, and the largest and smallest absolute principal values of a set of complex numbers by $\theta_{\max}(\cdot)$ and $\theta_{\min}(\cdot)$.

Theorem 4 (Local stabilization)**.** *Let* $\mathbf{x}^* \in \mathbb{R}^n$ *be a stationary point of a dynamics* F *with spectral radius* $\rho_0 \geq 1$*. Assume each element of* $S = \lambda_{\geq 1}(\nabla_{\mathbf{x}} F(\mathbf{x}^*))$ *is non-real. Then, its Lookahead dynamics with* $k \in \mathbb{N}, \alpha \in (0, 1)$ *locally converges to* $\mathbf{x}^*$ *if* $k \in (\beta_1, \beta_2)$ *and* α *is small enough, where* $\beta_1, \beta_2 > 0$ *satisfy* $\beta_1 \theta_{min}(S) = \arccos \rho_0^{-\beta_1}, \beta_2 \theta_{max}(S) = 2\pi - \arccos \rho_0^{-\beta_2}$.

Theorem 4 implies that, under certain assumptions on the eigenvalues, carefully chosen Lookahead hyperparameters can stabilize unstable equilibria. Specifically, by Lemma 3, the existence of a feasible $k \in (\beta_1, \beta_2)$ is guaranteed when the eigenvalues $S = \lambda_{\geq 1}(\mathbf{X})$ has imaginary conditioning less than 3. Therefore, any unstable points with such eigenvalues can be stabilized by Lookahead. In Appendix E, we verify this can be realistic even for a practical non-linear game like GANs.

The next theorem shows that Lookahead can further accelerate the local convergence of its base dynamics.

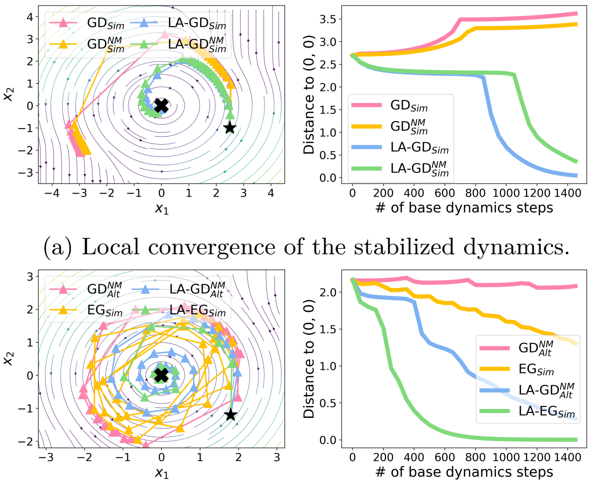
Theorem 5 (Local acceleration)**.** *Let* $\mathbf{x}^* \in \mathbb{R}^n$ *be a stationary point of a dynamics* F *with spectral radius* $\rho_0 < 1$*. Assume each element of* $S = \lambda_{max}(\nabla_{\mathbf{x}} F(\mathbf{x}^*))$ *is non-real. Then, the local convergence rate to* $\mathbf{x}^*$ *in its Lookahead dynamics with* $k \in \mathbb{N}, \alpha \in (0, 1)$ *improves upon* F *if* $k \in (\beta_1, \beta_2)$ *and* α *is large enough, where* $\beta_1, \beta_2 > 0$ *satisfy* $\beta_1 \theta_{min}(S) = \arccos \rho_0^{\beta_1}, \beta_2 \theta_{max}(S) = 2\pi - \arccos \rho_0^{\beta_2}$.

The precise threshold for α can be found in Appendix D. In contrast to Theorem 4, the acceleration requires a large α. Such a difference suggests that there exists a trade-off between the stabilization and acceleration that can be adjusted by α. For instance, one could trade-off the acceleration for stability by choosing a relatively small α; however, a prohibitively small α will introduce undesirable stable points. We discuss such a *spurious stabilization effect* in Section 5.

An example. We emphasize that our main results, i.e., Theorem 4–5, applies to an arbitrary base dynamics; hence, Lookahead provides a general mechanism for stabilization and acceleration in smooth games. To demonstrate our theoretic results, we exemplify a non-linear game with a local Nash equilibrium at $(0, 0)$:

$$\min_{x_1 \in \mathbb{R}} \max_{x_2 \in \mathbb{R}} \; -\log(2 + \exp(-x_1 \cdot x_2)) + \epsilon \cdot \phi(x_2). \quad (9)$$

where $\phi(x) \overset{\text{def}}{=} -x^2/2 + x^4/4$. The first term is a variant of Dirac-GAN example proposed by Mescheder et al. (2018), and introduces a strong rotational force around the equilibrium. The second term induces divergent trajectories along the x_2-axis. For $\epsilon = 0.001$ and a base learning rate $\eta = 0.1$, the equilibrium becomes unstable for simultaneous gradient descent GD_{Sim}, and asymptotically stable for extragradient EG_{Sim}. To verify the stabilization and acceleration guarantees, we compute the constants β_1, β_2 and the thresholds on α of Theorem 4–5, and choose k and α that stabilize GD_{Sim} and accelerate EG_{Sim}. Figure 3 (a) shows that GD_{Sim} diverges from the equilibrium, and even negative momentum $\text{GD}_{\text{Sim}}^{\text{NM}}$ fails to stabilize GD_{Sim}. However, k and α predicted by Theorem 4 successfully stabilize both GD_{Sim} and $\text{GD}_{\text{Sim}}^{\text{NM}}$. This suggests a stronger stabilization effect of Lookahead upon negative momentum. Meanwhile, Figure 3 (b) demonstrates the acceleration of EG_{Sim}, and shows that Lookahead can accelerate negative momentum $\text{GD}_{\text{Alt}}^{\text{NM}}$ as well. These results substantiate Theorem 4–5 and suggest that Lookahead



(a) Local convergence of the stabilized dynamics.

(b) Local convergence of the accelerated dynamics.

Figure 3: **Illustration of the stabilization and acceleration effect** in Equation 9. **(Top)** Simultaneous gradient descent $\mathrm{GD}_{\mathrm{Sim}}$ diverges from the local Nash equilibrium $(0,0)$, and negative momentum $\mathrm{GD}^{\mathrm{NM}}_{\mathrm{Sim}}$ fails to stabilize the equilibrium. However, the Lookahead hyperparameters predicted by Theorem 4, denoted by LA-$\mathrm{GD}_{\mathrm{Sim}}$ and LA-$\mathrm{GD}^{\mathrm{NM}}_{\mathrm{Sim}}$, successfully converges to the equilibrium. **(Bottom)** Negative momentum $\mathrm{GD}^{\mathrm{NM}}_{\mathrm{Alt}}$ and extragradient $\mathrm{EG}_{\mathrm{Sim}}$ slowly converges to the equilibrium due to the strong oscillation around the equilibrium; however, the Lookahead hyperaprameters predicted by Theorem 5 accelerate the convergence.

provides a general mechanism for stabilization and acceleration in smooth games. We refer the readers to Appendix D for further experimental details.

4 GENERAL BILINEAR GAMES

Even though Theorem 4–5 provide *local* guarantees for general smooth games, we can derive global stabilization and acceleration guarantees for bilinear games:

$$\min_{\mathbf{x}_1\in\mathbb{R}^m}\max_{\mathbf{x}_2\in\mathbb{R}^n} \mathbf{x}_1^T\mathbf{A}\mathbf{x}_2 - \mathbf{x}_1^T\mathbf{b}_1 - \mathbf{x}_2^T\mathbf{b}_2, \tag{10}$$

where $\mathbf{A}\in\mathbb{R}^{m\times n}, \mathbf{b}_1\in\mathbb{R}^m, \mathbf{b}_2\in\mathbb{R}^n$ admits $\mathbf{x}_1^*\in\mathbb{R}^m, \mathbf{x}_2^*\in\mathbb{R}^n$ such that $\mathbf{A}^T\mathbf{x}_1^* = \mathbf{b}_2, \mathbf{A}\mathbf{x}_2^* = \mathbf{b}_1$. This game has been extensively studied as an archetype of game optimization in the recent few years (Daskalakis et al., 2018; Gidel et al., 2019b,a; Zhang and Yu, 2020)

The first corollary shows that Lookahead can stabilize alternating $\mathrm{GD}_{\mathrm{Alt}}$, which is non-convergent and oscillatory around the Nash equilibria (Gidel et al., 2019b,a). We denote the singular values of the matrix $\mathbf{A}$ by σ_i, and their largest and smallest values by $\sigma_{\max}$ and $\sigma_{\min}$.

Corollary 6 (Stabilization of $\mathrm{GD}_{\mathrm{Alt}}$)**.** *Lookahead dynamics $G_{LA\text{-}GD_{Alt}}$ with $\eta\in(0,2\sigma_{max}^{-1}), k\in\mathbb{N}, \alpha\in(0,1)$ converges to a Nash equilibrium of Equation 14 if $k\arccos(1-\eta^2\sigma_i^2/2) \bmod 2\pi \neq 0, \forall\sigma_i$.*

The modulo condition breaks when there exists a singular value σ_i such that $k\arccos(1-\eta^2\sigma_i^2/2)$ exactly matches a multiple of 2π. Hence, Corollary 6 implies Lookahead LA-$\mathrm{GD}_{\mathrm{Alt}}$ converges to a Nash equilibrium for almost any $k\in\mathbb{N}, \alpha\in(0,1)$. This is in contrast to negative momentum (Gidel et al., 2019b) which works only for carefully chosen coefficients. The next result shows Lookahead can even stabilize divergent $\mathrm{GD}_{\mathrm{Sim}}$.

Corollary 7 (Stabilization of $\mathrm{GD}_{\mathrm{Sim}}$)**.** *Lookahead dynamics $G_{LA\text{-}GD_{Sim}}$ with $\eta>0, k\in\mathbb{N}, \alpha\in(0,1)$ converges to a Nash equilibrium of Equation 14 if $k\in(\beta_1,\beta_2), \alpha<c$, where $\beta_1,\beta_2>0$ satisfies*

$$\beta_1\arctan\eta\sigma_{min} = \arccos\rho_0^{-\beta_1},$$
$$\beta_2\arctan\eta\sigma_{max} = 2\pi - \arccos\rho_0^{-\beta_2}$$

with $\rho_0 \overset{\mathrm{def}}{=} \sqrt{1+\eta^2\sigma_{max}^2}$, and $c\in\mathbb{R}$ is a constant dependent on k such that $c>0, \forall k\in(\beta_1,\beta_2)$.

The precise threshold c can be found in Appendix C. Even though β_1,β_2 are defined implicitly, they are easily computable as each term in the first two equations is monotone with respect to β_1,β_2. Since Lemma 3 guarantees the existence of a feasible $k\in(\beta_1,\beta_2)$ for a small conditioning $\frac{\sigma_{\max}}{\sigma_{\min}}<3$, the corollary implies that Lookahead can stabilize $\mathrm{GD}_{\mathrm{Sim}}$ for well-conditioned bilinear games. This result qualitatively seperates Lookahead from negative momentum (Gidel et al., 2019b), which fails to stabilize $\mathrm{GD}_{\mathrm{Sim}}$ for all hyperparameters (Zhang and Yu, 2020). This qualitative difference suggests that Lookahead has a stronger stabilization capability.

Corollary 8 (Acceleration of $\mathrm{EG}_{\mathrm{Sim}}$)**.** *Lookahead dynamics $G_{LA\text{-}EG_{Sim}}$ with $\eta\in(0,\sigma_{max}^{-1}), k\in\mathbb{N}, \alpha\in(0,1)$ converges to a Nash equilibrium of Equation 14. The rate of convergence is improved upon its base dynamics $F_{EG_{Sim}}$ if $\eta\in(0,\sigma_{max}^{-1}/2), k\in(\beta_1,\beta_2), \alpha>c$, where $\beta_1,\beta_2>0$ are such that*

$$\beta_1\arctan\eta\sigma_{min}(1-\eta\sigma_{min})^{-1} = \arccos\rho_0^{\beta_1},$$
$$\beta_2\arctan\eta\sigma_{min}(1-\eta\sigma_{min})^{-1} = 2\pi - \arccos\rho_0^{\beta_2}$$

with $\rho_0 \overset{\mathrm{def}}{=} \sqrt{1-2\eta\sigma_{min}+2\eta^2\sigma_{min}^2}$, $c\in\mathbb{R}$ is a constant dependent on k such that $c<1, \forall k\in(\beta_1,\beta_2)$.

β_1,β_2 are computable and the threshold c can be found in Appendix C. In contrast to Theorem 7, there always exists a feasible $k\in(\beta_1,\beta_2)$ for any game conditioning $\frac{\sigma_{\max}}{\sigma_{\min}}$; the inequality $\arccos(\cdot)<2\pi-\arccos(\cdot)$ yields $\beta_1<\beta_2$, implying the existence of a feasible k for a small $\eta>0$. Hence, Theorem 8 implies Lookahead can always accelerate $\mathrm{EG}_{\mathrm{Sim}}$, whose existing rates (Gidel et al., 2019a; Zhang and Yu, 2020) are known to be

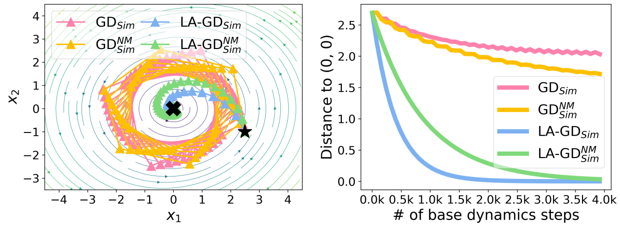


Figure 4: **Spurious stabilization effect** in Equation 11 with $\epsilon = 0.5$. Gradient descent GD$_{\text{Sim}}$ and negative momentum GD$^{\text{NM}}_{\text{Sim}}$ avoids the undesirable local maximin $(0, 0)$ due to its inherent instability. However, Lookahead dynamics LA-GD$_{\text{Sim}}$ and LA-GD$^{\text{NM}}_{\text{Sim}}$ blindly stabilize the local maximin, and introduces a *spurious convergence* to the undesirable stationary point.

first-order suboptimal in bilinear games (Azizian et al., 2020). This acceleration can be derived for *any* convergent dynamics, and we derive a similar result for PP$_{\text{Sim}}$ in Appendix C. However, as the corollary provides no explicit rate, it remains open whether it can achieve the first-order optimal rate (Azizian et al., 2020).

5 THE STABILIZATION EFFECT

Spurious stabilization. Recent studies (Mazumdar et al., 2019; Hsieh et al., 2020) point out that most game dynamics suffer from *spurious convergence*, i.e., convergence to an undesirable stationary point. In this section, we show that the stabilization of Lookahead can induce an additional spurious convergence, and the benefits of the stabilization heavily depend on the game structure. Then, we show that GANs do not suffer from the spurious stabilization both in theory and practice. First, we consider the following nonlinear game proposed by Hsieh et al. (2020):

$$\min_{x_1 \in \mathbb{R}} \max_{x_2 \in \mathbb{R}} \; x_1 \cdot x_2 + \epsilon \cdot (x_2^2/2 - x_2^4/4), \tag{11}$$

This game can be considered as a bilinear game with a small nonlinear perturbation. In contrast to bilinear games, the origin $(0, 0)$ becomes an undesirable local *maximin* for $\epsilon > 0$, and a stable (local) Nash equilibrium for $\epsilon \leq 0$. As Theorem 4–5 both hold for an arbitrary stationary point, Lookahead can either stabilize or accelerate the local convergence around the origin $(0, 0)$ in both cases. Hence, while Lookahead can improve the convergence towards the (local) Nash equilibrium for $\epsilon \leq 0$, it can also create an undesirable stability for $\epsilon > 0$. Figure 4 illustrates the latter case, i.e., the *spurious stabilization* phenomenon. In Figure 4, gradient descent and negative momentum successfully avoids the undesirable local maximin $(0, 0)$; however, Lookahead blindly stabilizes $(0, 0)$ and creates a spurious convergence. This clearly shows that the stabilization of Lookahead is a double-edged sword, and its benefits heavily depend on the game structure.

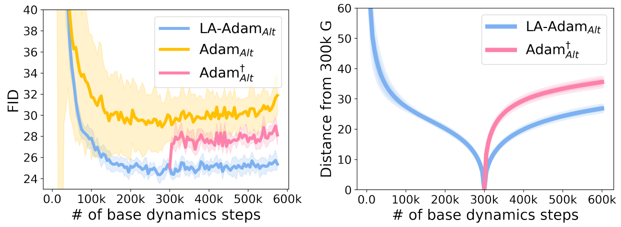


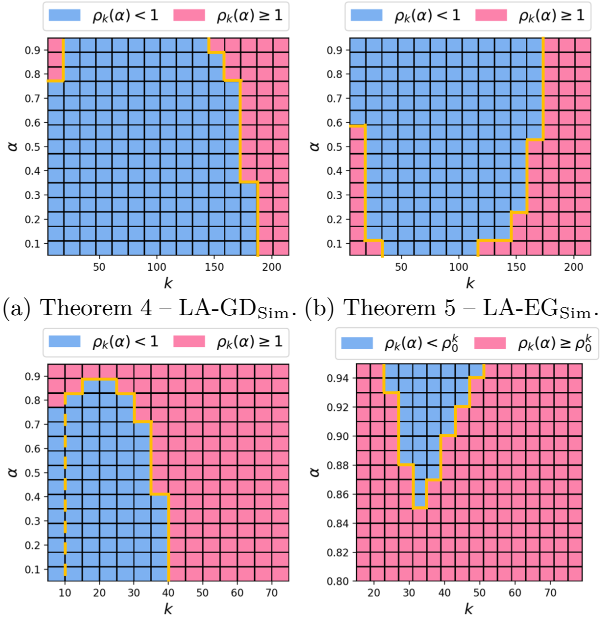
Figure 5: **Stabilization of Lookahead in GANs trained with CIFAR-10**. The solid lines and shades represent the mean and standard deviation of each metric over 8 runs, respectively. Lookahead LA-Adam$_{\text{Alt}}$ with $k = 5000, \alpha = 0.5$ achieves much lower FID scores than Adam$_{\text{Alt}}$. However, as Lookahead iteration stops after 300k steps, Adam$^{\dagger}_{\text{Alt}}$ quickly diverges from the region that contains Lookahead trajectories, and suffers from a severe performance degradation. This suggests that Lookahead stabilizes a small region that contains highly-performant, yet unstable generators of GANs.

GANs. Nevertheless, Chavdarova et al. (2021) report that Lookahead significantly improves GAN training. Such an empirical success suggests that GANs might exhibit a special structure that could be exploited by Lookahead. Notably, No et al. (2021) gives a positive answer to this hypothesis for 2-layer random-feature WGANs. Under certain assumptions on discriminator and random features, they prove that 2-layer random-feature WGANs with wide generator have no spurious stationary points, i.e., each stationary point is a Nash equilibrium (No et al. (2021), Theorem 8). As a crucial corollary, we can guarantee that Lookahead does not induce a spurious stabilization in such settings.

Corollary 9 (No spurious stabilization for wide 2-layer WGANs; informal)**.** *Under certain assumptions on 2-layer random-feature fully-connected WGANs with sigmoidal activation and $\ell 2$-regularized discriminator, for any $\zeta > 0$, there exists a large enough generator hidden layer width $N_g \in \mathbb{N}$ such that the following statement holds with probability at least $1 - \zeta$: a set of stationary points that are stabilized by Lookahead contains no other points than the Nash equilibria of the game.*

The key assumptions of No et al. (2021) include (i) a small discriminator, and (ii) sample space spanning random features, where the first assumption can be dropped for an infinitely wide generator. We defer the precise statements of the assumptions and the corollary to Appendix C. This result gives a promising guarantee that Lookahead does not create a spurious stabilization in simple settings and only stabilizes Nash equilibria.

However, there still exists a gap between the corollary and practice: modern GANs adopt deep convolutional architectures, and neither use a small discriminator



(a) Theorem 4 – LA-GD$_{\text{Sim}}$. (b) Theorem 5 – LA-EG$_{\text{Sim}}$.

(c) Corollary 7 – LA-GD$_{\text{Sim}}$. (d) Corollary 8 – LA-EG$_{\text{Sim}}$.

Figure 6: **Stabilization and acceleration over a range** of (k, α) in the (a–b) nonlinear and (c–d) bilinear game experiments. Each plot illustrates the relative size of the spectral radius of Lookahead $\rho_k(\alpha)$ over a range of (k, α). Each blue cell represents (k, α) that achieves either stabilization $\rho_k(\alpha) < 1$ or acceleration $\rho_k(\alpha) < \rho_0^k$ with respect to the spectral radius ρ_0 of its base dynamics. The blue cells contained in the yellow lines represent the improvements predicted by each theorem, i.e., (k, α) that satisfies $k \in (\beta_1, \beta_2)$ and $\alpha < c$ or $\alpha > c$, where c is the threshold for α predicted by each theorem in Appendix C. Each red cell indicates (k, α) that fails to improve the convergence.

nor an infinitely wide generator. Hence, we conjecture that the success of Lookahead in GANs originates from the stabilization of a small region that contains unstable, yet highly-performant generators, which may not necessarily be the Nash equilibria.

To verify our hypothesis, we train a SNDCGAN (Miyato et al., 2018) on CIFAR-10 using LA-Adam$_{\text{Alt}}$ with $k = 5000, \alpha = 0.5$. Specifically, we train GANs with Lookahead until the convergence of FID scores (Heusel et al., 2017), and turn off the Lookahead iteration to see if the base dynamics Adam$_{\text{Alt}}$ quickly diverges from the highly-performant generators found by Lookahead. We illustrate our results over 8 runs in Figure 5.

Figure 5 shows that Lookahead LA-Adam$_{\text{Alt}}$ achieves much lower FID scores than its base dynamics Adam$_{\text{Alt}}$. However, as we stop the Lookahead iteration after 300k steps, Adam$_{\text{Alt}}$ quickly diverges from the region that contains highly performant generators found by Lookahead, and suffers from a severe performance drop. This suggests that Lookahead stabilizes a small region that contains highly-performant, yet unstable generators, and justifies the usage of Lookahead in practical GANs.

6 NUMERICAL EXPERIMENTS

Nonlinear game. We numerically verify the predictions of Theorem 4–5 on Equation 9 with $\epsilon = 0.001$. We fix the learning rate $\eta = 0.1$, under which the local Nash equilibrium $(0, 0)$ becomes unstable for GD$_{\text{Sim}}$ and stable for EG$_{\text{Sim}}$. We derive the constants β_1, β_2 and the threshold c from each theorem to stabilize GD$_{\text{Sim}}$, and accelerate EG$_{\text{Sim}}$ at $(0, 0)$. Then, we inspect the spectral radius of Lookahead at $(0, 0)$ over a range of k and α to verify whether the local stabilization $\rho_k(\alpha) < 1$ or acceleration $\rho_k(\alpha) < \rho_0^k$ promised by the theorems actually hold. We illustrate the results in Figure 6 (a)–(b). Figure 6 (a) shows that the local stabilization promised by Theorem 4 actually holds for GD$_{\text{Sim}}$. Similarly, Figure 6 (b) verifies the acceleration of EG$_{\text{Sim}}$ promised by Theorem 5. We do not observe any non-necessity in our sufficient conditions.

Bilinear game. We verify Corollary 7–8 on a bilinear game with $\mathbf{A} = \mathbf{I}_n + \epsilon \cdot \mathbf{E}_n$ and $\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{0}$, where each element of $\mathbf{E}_n \in \mathbb{R}^{n \times n}$ is sampled from $\mathcal{N}(0, 1)$. For the learning rate $\eta = 0.1$, we derive the constants β_1, β_2 and the precise threshold c from each theorem. We compute the spectral radius of Lookahead $\rho_k(\alpha)$ over a range of (k, α) that contains the sufficient conditions of Corollary 7–8. Then, we inspect the spectral radius $\rho_k(\alpha)$ to verify whether the global stabilization $\rho_k(\alpha) < 1$ or acceleration $\rho_k(\alpha) < \rho_0^k$ promised by the corollaries actually hold. As Corollary 7 is provably non-vacuous for the games with conditioning less than 3, we report our results using $n = 1000$ and $\epsilon = 0.01$, which gives a sample of $\mathbf{A}$ with conditioning $\frac{\sigma_{\max}}{\sigma_{\min}} = 2.5 < 3$; we report the results on a larger conditioning in Appendix E. Figure 6 (c)–(d) illustrate the results. Figure 6 (c) verifies the stabilization guarantee of Corollary 7, and Figure 6 (d) confirms the acceleration of Corollary 8. Most blue cells are tightly contained in the yellow lines, suggesting the sharpness of our conditions.

7 CONCLUSION

In this work, we established the first convergence guarantees of Lookahead in smooth games. Our results reveal that Lookahead provides a general mechanism for stabilization and acceleration, and points to several future research directions. The first step would be analyzing Lookahead in stochastic settings; as Chavdarova et al. (2021) report Lookahead tends to be especially effective for stochastic games, we expect a variance reduction effect in Lookahead. Another important direction would be studying whether our local results could be transferred to global guarantees of GANs.

Acknowledgements

This work was supported by Samsung Advanced Institute of Technology, Basic Science Research Program through the National Research Foundation of Korea (NRF) (2020R1A2B5B03095585) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01082, SW StarLab).

References

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. (2019). Local saddle point optimization: A curvature exploitation approach. In *AISTATS*.

Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. (2021). Adaptive extra-gradient methods for min-max optimization and games. In *ICLR*.

Azizian, W., Scieur, D., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. (2020). Accelerating smooth games by manipulating spectral shapes. In *AISTATS*.

Balduzzi, D., Racanière, S., Martens, J., Foerster, J. N., Tuyls, K., and Graepel, T. (2018). The Mechanics of n-Player Differentiable Games. In *ICML*.

Benzi, M., Golub, G., and Liesen, J. (2005). Numerical solution of saddle point problems. *Acta Numerica*, 14:1 – 137.

Berard, H., Gidel, G., Almahairi, A., Vincent, P., and Lacoste-Julien, S. (2020). A closer look at the optimization landscapes of generative adversarial networks. In *ICLR*.

Bertsekas, D. (1999). *Nonlinear Programming*. Athena Scientific.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. (2019). Reducing noise in gan training with variance reduced extragradient. In *NeurIPS*.

Chavdarova, T., Pagliardini, M., Stich, S. U., Fleuret, F., and Jaggi, M. (2021). Taming gans with lookahead-minmax. In *ICLR*.

Chen, C. (1995). *Linear System Theory and Design*. Oxford University Press.

Daskalakis, C., Goldberg, P., and Papadimitriou, C. (2006). The Complexity of Computing a Nash Equilibrium. *Electron. Colloquium Comput. Complex.*

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2018). Training GANs with Optimism. In *ICLR*.

Daskalakis, C., Skoulakis, S., and Zampetakis, M. (2020). The complexity of constrained min-max optimization. *ArXiv*, abs/2009.09623.

Facchinei, F. and Pang, J. (2003). *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer.

Gidel, G., Berard, H., Vincent, P., and Lacoste-Julien, S. (2019a). A Variational Inequality Perspective on Generative Adversarial Nets. In *ICLR*.

Gidel, G., Hemmat, R. A., Pezeshki, M., Huang, G., Priol, R. L., Lacoste-Julien, S., and Mitliagkas, I. (2019b). Negative Momentum for Improved Game Dynamics. In *AISTAT*.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative Adversarial Nets. In *NIPS*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *CoRR*, abs/1412.6572.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *ArXiv*, abs/1704.00028.

Hemmat, R. A., Mitra, A., Lajoie, G., and Mitliagkas, I. (2020). Lead: Least-action dynamics for min-max optimization. *arXiv preprint arXiv:2010.13846*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NIPS*.

Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. (2020). The Limits of Min-Max Optimization Algorithms: Convergence to Spurious Non-critical Sets. *ArXiv*, abs/2006.09065.

Jelassi, S., Domingo-Enrich, C., Scieur, D., Mensch, A., and Bruna, J. (2020). Extragradient with player sampling for faster nash equilibrium finding. In *ICML*.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58.

Karras, T., Laine, S., and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *ICLR*.

Korpelevich, G. M. (1976). The Extragradient Method for Finding Saddle Points and Other Problems. *Ekonomika i Matematicheskie Metody*, 12:747–756.

Kurach, K., Lucic, M., Zhai, X., Michalski, M., and Gelly, S. (2019). A large-scale study on regularization and normalization in gans. In *ICML*.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Letcher, A., Foerster, J. N., Balduzzi, D., Rocktäschel, T., and Whiteson, S. (2019). Stable opponent shaping in differentiable games. In *ICLR*.

Lin, T., Jin, C., and Jordan, M. I. (2020). Near-optimal algorithms for minimax optimization. In *COLT*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

Martinet, B. (1970). Brève Communication. Régularisation d'inéquations Variationnelles par Approximations Successives. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 4(R3):154–158.

Mazumdar, E. V., Jordan, M. I., and Sastry, S. (2019). On Finding Local Nash Equilibria (and Only Local Nash Equilibria) in Zero-Sum Games. *CoRR*, abs/1901.00838.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. (2019). Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*.

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. (2018). Cycles in Adversarial Regularized Learning. In *SODA*.

Mescheder, L. M., Geiger, A., and Nowozin, S. (2018). Which Training Methods for GANs do actually Converge? In *ICML*.

Mescheder, L. M., Nowozin, S., and Geiger, A. (2017). The Numerics of GANs. In *NIPS*.

Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. (2020). Revisiting stochastic extragradient. In *AISTATS*.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *ArXiv*, abs/1802.05957.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020). A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *AISTATS*.

Mortici, C. and Srivastava, H. M. (2014). Estimates for the Arctangent Function Related to Shafer's Inequality. *Colloq. Math*, 136(2):263–270.

Nash, J. (1951). Non-Cooperative Games. *Annals of Mathematics*, 54:286.

Nemirovski, A. (2004). 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 J. Optim.*, 15:229–251.

Neumann, J. and Morgenstern, O. (1944). Theory of games and economic behavior. *Journal of the American Statistical Association*, 40:263.

No, A., Yoon, T., Kwon, S.-H., and Ryu, E. K. (2021). WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points. In *ICML*.

Peng, W., Dai, Y., Zhang, H. B., and Cheng, L. (2020). Training gans with centripetal acceleration. *Optimization Methods and Software*, 35:955 – 973.

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

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.

Rockafellar, R. T. (1976). Monotone Operators and the Proximal Point Algorithm. *Siam Journal on Control and Optimization*, 14:877–898.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *ArXiv*, abs/1606.03498.

Schäfer, F. and Anandkumar, A. (2019). Competitive Gradient Descent. In *NeurIPS*.

Scutari, G., Palomar, D., Facchinei, F., and Pang, J. (2010). Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 27:35–49.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-play. *Science*, 362:1140 – 1144.

Tseng, P. (1995). On Linear Convergence of Iterative Methods for the Variational Inequality Problem. *Journal of Computational and Applied Mathematics*, 60:237–252.

Vinyals, O., Babuschkin, I., Czarnecki, W., Mathieu, M., Dudzik, A., Chung, J., Choi, D., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster Level in StarCraft II using Multi-agent Reinforcement Learning. *Nature*, pages 1–5.

Wang, Y., Zhang, G., and Ba, J. (2020). On solving minimax optimization locally: A follow-the-ridge approach. In *ICLR*.

Yadav, A., Shah, S., Xu, Z., Jacobs, D., and Goldstein, T. (2018). Stabilizing adversarial nets with prediction methods. In *ICLR*.

Yoon, T. and Ryu, E. K. (2021). Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *ICML*.

Zhang, G. and Yu, Y. (2020). Convergence of Gradient Methods on Bilinear Zero-Sum Games. In *ICLR*.

Zhang, M., Lucas, J., Hinton, G. E., and Ba, J. (2019). Lookahead Optimizer: k Steps Forward, 1 Step Back. In *NeurIPS*.

Supplementary Material:
On Convergence of Lookahead in Smooth Games

A NOTATION

Table A.1: The notation used throughout the paper.

Symbol	Definition
a	A scalar.
$\mathbf{A}$	A matrix.
$\mathbf{A}^T$	Transpose of matrix A.
$\mathbf{I}$	Identity matrix with its shape implied by context.
$\mathbf{I}_n$	Identity matrix with n rows and n columns.
$\mathbb{R}$	The set of real numbers
$[a, b]$	The real interval including a and b.
$(a, b]$	The real interval excluding a but including b.
$\lVert\cdot\rVert$	L^2 norm.
x_i	i-th element of a vector $\mathbf{x} = (x_1, \ldots, x_n)$.
$\mathbf{x}_i$	i-th vector of the concatenated $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.
$\mathbf{x}_{-i}$	$(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)$.
$\mathbf{1}_{\text{condition}}$	1 if the condition is true, 0 otherwise.
$\nabla_x f(x')$	The derivative of a function f evaluated at x'.
S_r	The zero-centered circle of radius $r > 0$ in $\mathbb{C}$.
$\Re(z)$	The real part of $z \in \mathbb{C}$.
$\Im(z)$	The imaginary part of $z \in \mathbb{C}$.
$\text{Arg}(z)$	The angle between $z \in \mathbb{C}$ and the real axis.
$\sigma(\mathbf{A})$	The set of singular values of $\mathbf{A} \in \mathbb{R}^{m\times n}$.
$\rho(\mathbf{A})$	The spectral radius of $\mathbf{A} \in \mathbb{R}^{m\times m}$.
$\lambda(\mathbf{A})$	The set of eigenvalues of $\mathbf{A} \in \mathbb{R}^{m\times m}$.
$\lambda_{\geq a}(\mathbf{A})$	$\{\lambda_i \in \lambda(\mathbf{A}) : \lvert\lambda_i\rvert \geq a\}$
$\lambda_{\max}(\mathbf{A})$	$\{\lambda_i \in \lambda(\mathbf{A}) : \lvert\lambda_i\rvert = \max_{\lambda_i \in \lambda(\mathbf{A})} \lvert\lambda_i\rvert\}$.

B USEFUL FACTS

B.1 Standard Results on Convergence

Lemma B.1 (Bertsekas (1999)). *Let $F : \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable, and let $\boldsymbol{x}^* \in \mathbb{R}^m$ be a fixed point of F such that $\rho(\nabla_{\boldsymbol{x}} F(\boldsymbol{x}^*)) < 1$. Then, there exists an open neighborhood $U_{\boldsymbol{x}^*}$ of $\boldsymbol{x}^*$ such that for any $\boldsymbol{x} \in U_{\boldsymbol{x}^*}$, $\lVert F^t(\boldsymbol{x}) - \boldsymbol{x}^*\rVert_2 \in \mathcal{O}(\rho(\nabla_{\boldsymbol{x}} F(\boldsymbol{x}^*))^t)$ for $t \to \infty$.*

Lemma B.2 (Gidel et al. (2019b)). *Let $\boldsymbol{M} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{u}^{(t)}$ be a sequence of iterates such that, $\boldsymbol{u}^{(t+1)} = \boldsymbol{M}\boldsymbol{u}^{(t)}$. Then, we have three cases of interest for the spectral radius $\rho(\boldsymbol{M})$:*

- *If $\rho(\boldsymbol{M}) < 1$ and $\boldsymbol{M}$ is diagonalizable,[1] then $\lVert\boldsymbol{u}^{(t)}\rVert_2 \in \mathcal{O}(\rho(\boldsymbol{M})^t \lVert\boldsymbol{u}^{(0)}\rVert_2)$.*
- *If $\rho(\boldsymbol{M}) > 1$, then there exists $\boldsymbol{u}^{(0)}$ such that $\lVert\boldsymbol{u}^{(t)}\rVert_2 \in \Omega(\rho(\boldsymbol{M})^t \lVert\boldsymbol{u}^{(0)}\rVert_2)$.*
- *If $\lvert\lambda_i\rvert = 1, \forall \lambda_i \in \lambda(\boldsymbol{M})$, and $\boldsymbol{M}$ is diagonalizable, then $\lVert\boldsymbol{u}^{(t)}\rVert_2 \in \Theta(\lVert\boldsymbol{u}^{(0)}\rVert_2)$.*

[1] In fact, $\mathbf{M}$ does not has to be diagonalizable; see Theorem 5.4 and Theorem 5.D4 in Chen (1995).

B.2 Characteristic Equations for Bilinear Games

Recent work Zhang and Yu (2020) provides an exact and optimal conditions for popular first-order methods to converge in bilinear games. They also derive the characteristic equation for each first-order dynamics. As our proofs rely on spectral arguments, we restate a simplified version of the equations for completeness. We denote the singular values of a game matrix by σ_i, and the eigenvalues of each dynamics' by λ_i.

$$\begin{aligned} \mathrm{GD}_{\mathrm{Alt}}:&\quad (\lambda_i-1)^2+\eta^2\sigma_i^2\lambda_i=0,\\ \mathrm{GD}_{\mathrm{Sim}}:&\quad (\lambda_i-1)^2+\eta^2\sigma_i^2=0,\\ \mathrm{PP}_{\mathrm{Sim}}:&\quad (\lambda_i^{-1}-1)^2+\eta^2\sigma_i^2=0,\\ \mathrm{EG}_{\mathrm{Sim}}:&\quad (\lambda_i-1)^2+2\eta\sigma_i^2(\lambda_i-1)+\eta^2\sigma_i^2+\eta^2\sigma_i^4=0. \end{aligned}$$

B.3 Bilinear Game Reduction

In this paper, we consider the bilinear games of general form

$$\min_{\mathbf{x}_1\in\mathbb{R}^m}\max_{\mathbf{x}_2\in\mathbb{R}^n}\ \mathbf{x}_1^T\mathbf{A}\mathbf{x}_2-\mathbf{x}_1^T\mathbf{b}_1-\mathbf{x}_2^T\mathbf{b}_2, \tag{12}$$

where $\mathbf{A}\in\mathbb{R}^{m\times n},\mathbf{b}_1\in\mathbb{R}^m,\mathbf{b}_2\in\mathbb{R}^n$ admits $\mathbf{x}_1^*\in\mathbb{R}^m,\mathbf{x}_2^*\in\mathbb{R}^n$ such that $\mathbf{A}^T\mathbf{x}_1^*=\mathbf{b}_2,\mathbf{A}\mathbf{x}_2^*=\mathbf{b}_1$. The existence of $\mathbf{x}_1^*,\mathbf{x}_2^*$ allows us to rewrite the game as

$$\min_{\mathbf{x}_1\in\mathbb{R}^m}\max_{\mathbf{x}_2\in\mathbb{R}^n}\ (\mathbf{x}_1-\mathbf{x}_1^*)^T\mathbf{U}\begin{bmatrix}\Sigma_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{V}^T(\mathbf{x}_2-\mathbf{x}_2^*), \tag{13}$$

where $\mathbf{U}\in\mathbb{R}^{m\times m},\Sigma_r\in\mathbb{R}^{r\times r},\mathbf{V}\in\mathbb{R}^{n\times n}$ is the SVD of $\mathbf{A}$ with $r\overset{\text{def}}{=}\operatorname{rank}(\mathbf{A})$. Hence, we can analyze the convergence of a dynamics in Equation 12 by inspecting a rather simpler problem

$$\min_{\mathbf{x}_1\in\mathbb{R}^r}\max_{\mathbf{x}_2\in\mathbb{R}^r}\ \mathbf{x}_1^T\Sigma_r\mathbf{x}_2, \tag{14}$$

as they are equivalent up to some rotations and translations. Therefore, we can establish the convergence guarantees for Lookahead in Equation 14 without loss of generality. Since Σ_r is non-singular, the reduced bilinear game of Equation 14 has the unique Nash equilibrium at the origin. This bilinear game reduction is well-known, and has been widely used for simplifying the analysis Gidel et al. (2019b,a); Mokhtari et al. (2020); Zhang and Yu (2020).

B.4 Local Nash Equilibrium

Following Balduzzi et al. (2018), the concept of Nash equilibrium can be generalized to its *local* variant as follows.

Definition B.1 (Local Nash equilibrium)**.** *For a smooth game $\{f_i\}_{i=1}^n$ with strategy spaces $\{\mathbb{R}^{d_i}\}_{i=1}^n$ such that $d=\sum_{i=1}^n d_i$, $\mathbf{x}^*\in\mathbb{R}^d$ is a local Nash equilibrium of the game if, for each $i=1,\ldots,n$, there exists an open neighborhood $U_i\subseteq\mathbb{R}^{d_i}$ of $\mathbf{x}_i^*$ such that satisfies $f_i(\mathbf{x}^*)\le f_i(\mathbf{x}_i,\mathbf{x}_{-i}^*),\forall\mathbf{x}_i\in U_i$.*

However, it is not straightforward from the definition how to test whether an equilibriuma is a local Nash or not. The following result from Mescheder et al. (2017) provides a tool for verifying a local Nash equilibrium.

Proposition B.3 (Mescheder et al. (2017))**.** *For any two-player zero-sum game $\{f,-f\}$ with strategy spaces $\{\mathbb{R}^{d_1},\mathbb{R}^{d_2}\}$, $\mathbf{x}^*=(\mathbf{x}_1^*,\mathbf{x}_2^*)\in\mathbb{R}^{d_1+d_2}$ is a local Nash equilibrium if $\nabla_{\mathbf{x}_1}f(\mathbf{x}^*)=\nabla_{\mathbf{x}_2}f(\mathbf{x}^*)=\mathbf{0}$, and the matrix*

$$\mathbf{J}(\mathbf{x}^*)=\begin{bmatrix}-\nabla^2_{\mathbf{x}_1}f(\mathbf{x}^*) & -\nabla_{\mathbf{x}_1,\mathbf{x}_2}f(\mathbf{x}^*)\\ \nabla_{\mathbf{x}_1,\mathbf{x}_2}f(\mathbf{x}^*) & \nabla^2_{\mathbf{x}_2}f(\mathbf{x}^*)\end{bmatrix}$$

is negative definite, i.e., $\mathbf{x}^T\mathbf{J}(\mathbf{x}^*)\mathbf{x}<0$ for any non-zero $\mathbf{x}\in\mathbb{R}^{d_1+d_2}$.

C OMITTED PROOFS

C.1 Examples

Example 3 (Stabilization)**.** *Lookahead dynamics $G_{LA\text{-}GD_{Sim}}$ with $\eta>0,k\in\mathbb{N},\alpha\in(0,1)$ converges to the Nash equilibrium of Equation 8 if k satisfies $\Re((1+i\eta)^k)<1$ and α is small enough.*

Proof. The eigenvalues of $\nabla_{\mathbf{x}} F_{\mathrm{GD_{Sim}}}$ are $\lambda_\pm \overset{\mathsf{def}}{=} 1 \pm i\eta$. Hence, plugging $\lambda_\pm$ into the first case of Proposition 1 gives $\rho(\nabla_{\mathbf{x}} G_{\text{LA-GD}_{\text{Sim}}}) < 1$ if $\Re((1 \pm i\eta)^k) < 1$ and $\alpha < c_1$, where $c_1 > 0$ if $\Re((1 \pm i\eta)^k) < 1$. Therefore, combining the results with the first case of Lemma B.2 proves the statement. □

Example 4 (Acceleration)**.** *Lookahead dynamics $G_{\textit{LA-PP}_{\textit{Sim}}}$ with $\eta \in (0,1), k \in \mathbb{N}, \alpha \in (0,1)$ converges to the Nash equilibrium of Equation 8. The rate of convergence improves upon its base dynamics $F_{\textit{PP}_{\textit{Sim}}}$ if k satisfies $\Re((1+i\eta)^k) < 1$ and α is large enough.*

Proof. The eigenvalues of $\nabla_{\mathbf{x}} F_{\mathrm{PP_{Sim}}}$ are $\lambda_\pm = \frac{1 \pm i\eta}{1+\eta^2}$, and each has modulus $|\lambda_{\pm i}| = \frac{1}{\sqrt{1+\eta^2}} < 1$. It follows $\rho(\nabla_{\mathbf{x}} G_{\text{LA-PP}_{\text{Sim}}}) = |1 - \alpha + \alpha\lambda_\pm^k| < 1$ as $1 - \alpha + \alpha\lambda_\pm^k$ is an interpolation between two distinct points on the unit disk. Therefore, the first case of Lemma B.2 implies that Lookahead dynamics $\mathrm{G}_{\text{LA-PP}_{\text{Sim}}}$ with $k \in \mathbb{N}$ and $\alpha \in (0,1)$ converges to the Nash equilibrium. Furthermore, plugging $\lambda_\pm$ into the last case of Proposition 1 yields $\rho(\nabla_{\mathbf{x}} G_{\text{LA-PP}_{\text{Sim}}}) < \rho(\nabla_{\mathbf{x}} F_{\mathrm{PP_{Sim}}})^k$ for $\Re((1+i\eta)^k) < 1$ and $\alpha > c_2$, where $c_2 < 1$ if $\Re((1+i\eta)^k) < 1$. Therefore, combining the results with the first case of Lemma B.2 proves the last statement on the improved convergence rate. □

Example 5 (Stabilization)**.** *Lookahead dynamics $G_{\textit{LA-GD}_{\textit{Alt}}}$ with $\eta \in (0,2), k \in \mathbb{N}, \alpha \in (0,1)$ converges to the Nash equilibrium of Equation 8 if and only if $k \arccos(1 - \eta^2/2) \bmod 2\pi \neq 0$.*

Proof. The eigenvalues of $\nabla_{\mathbf{x}} F_{\mathrm{GD_{Sim}}}$ are $\lambda_\pm \overset{\mathsf{def}}{=} 1 - \eta^2/2 \pm i\eta\sqrt{1 - \eta^2/4}$, and each has the modulus $|\lambda_\pm| = 1$ for any $\eta \in (0,2)$. Hence, plugging $\lambda_\pm$ into the second case of Proposition 1 gives $\rho(\nabla_{\mathbf{x}} G_{\text{LA-GD}_{\text{Alt}}}) < 1$ if and only if $\Re(\lambda_\pm^k) < 1$. However, $\Re(\lambda_\pm^k) < 1$ holds if and only if $k \arccos(1 - \eta^2/2) \bmod 2\pi \neq 0$, and combining the results with the first case of Lemma B.2 proves the statement. □

Example 6 (Acceleration)**.** *Lookahead dynamics $G_{\textit{LA-EG}_{\textit{Sim}}}$ with $\eta \in (0,1), k \in \mathbb{N}, \alpha \in (0,1)$ converges to the Nash equilibrium of Equation 8. The rate of convergence improves upon its base dynamics $F_{\textit{EG}_{\textit{Sim}}}$ if k satisfies $\Re((1 - \eta + i\eta)^k) < (1 - 2\eta + 2\eta^2)^k$ and α is large enough.*

Proof. The eigenvalues of $\nabla_{\mathbf{x}} F_{\mathrm{EG_{Sim}}}$ are $\lambda_\pm = 1 - \eta \pm i\eta$, and each has the modulus $|\lambda_{\pm i}| = \sqrt{1 - 2\eta + 2\eta^2} < 1$ for any $\eta \in (0,1)$. It follows $\rho(\nabla_{\mathbf{x}} G_{\text{LA-EG}_{\text{Sim}}}) = |1 - \alpha + \alpha\lambda_\pm^k| < 1$ since $1 - \alpha + \alpha\lambda_\pm^k$ is an interpolation between two distinct points on the unit disk. Therefore, the first case of Lemma B.2 implies Lookahead dynamics $\mathrm{G}_{\text{LA-EG}_{\text{Sim}}}$ with $k \in \mathbb{N}$ and $\alpha \in (0,1)$ converges to the Nash equilibrium. Furthermore, plugging $\lambda_\pm$ into the last case of Proposition 1 yields $\rho(\nabla_{\mathbf{x}} G_{\text{LA-EG}_{\text{Sim}}}) < \rho(\nabla_{\mathbf{x}} F_{\mathrm{EG_{Sim}}})^k$ if $\Re((1 - \eta + i\eta)^k) < (1 - 2\eta + 2\eta^2)^k$ and $\alpha > c_2$, where $c_2 < 1$ if $\Re((1 - \eta + i\eta)^k) < (1 - 2\eta + 2\eta^2)^k$. Therefore, combining the results with the first case of Lemma B.2 proves the last statement on the improved convergence rate. □

C.2 The Spectral Contraction

Proposition 1 (Spectral contraction)**.** *Let $\mathbf{X} \in \mathbb{R}^{m \times m}$ be a Jacobian of a dynamics at an equilibrium. Denote its spectral radius by $\rho_0 \overset{\mathsf{def}}{=} \rho(\mathbf{X})$ and the radius of its Lookahead dynamics with $k \in \mathbb{N}, \alpha \in (0,1)$ by $\rho_k(\alpha) \overset{\mathsf{def}}{=} \rho((1-\alpha)\boldsymbol{I} + \alpha\mathbf{X}^k)$. Then, either stabilization ($\rho_k(\alpha) < 1$) or acceleration ($\rho_k(\alpha) < \rho_0^k$) is achieved according to whether the base dynamics is stable ($\rho_0 < 1$) or not ($\rho_0 > 1$) as follows.*

- *For $\rho_0 > 1$, $\rho_k(\alpha) < 1$ if and only if $\tau_{k|\geq 1} < 1, \alpha < c_1$,*
- *For $\rho_0 = 1$, $\rho_k(\alpha) < 1$ if and only if $\tau_{k|max} < 1$,*
- *For $\rho_0 < 1$, $\rho_k(\alpha) < \rho_0^k$ if and only if $\tau_{k|max} < \rho_0^{2k}, \alpha > c_2$,*

where $\tau_{k|\geq 1} \overset{\mathsf{def}}{=} \max\limits_{\lambda_i \in \lambda_{\geq 1}(\mathbf{X})} \Re(\lambda_i^k)$, $\tau_{k|max} \overset{\mathsf{def}}{=} \max\limits_{\lambda_i \in \lambda_{max}(\mathbf{X})} \Re(\lambda_i^k)$, and $c_1, c_2 \in \mathbb{R}$ are such that

$$c_1 \overset{\mathsf{def}}{=} \min_{\lambda_i \in \lambda_{\geq 1}(\boldsymbol{X})} \frac{2\cos\phi_i}{|1 - \lambda_i^k|} > 0 \iff \tau_{k|\geq 1} < 1,$$

$$c_2 \overset{\mathsf{def}}{=} \max_{\lambda_i \in \lambda(\boldsymbol{X})} \frac{\cos\phi_i - \Delta_i}{|1 - \lambda_i^k|} < 1 \overset{\rho_0 < 1}{\iff} \tau_{k|max} < \rho_0^{2k},$$

with $\phi_i \overset{\mathsf{def}}{=} \mathrm{Arg}(1 - \lambda_i^k)$, $\Delta_i \overset{\mathsf{def}}{=} \sqrt{\rho_0^{2k} - \sin^2\phi_i}$.

Proof. We prove each of the statements in their order.

The case for $\rho_0 > 1$: Assume $\tau_{k|\geq 1} < 1$. Then, for each $\lambda_i \in \lambda_{\geq 1}(\mathbf{X})$, λ_i^k can be visualized as B in Figure C.7 (a), where the existence of D is guaranteed by $\tau_{k|\geq 1} < 1$. Then, we can see that, for A, C, D in Figure C.7 (a),

$$|\overline{AC}| = \alpha|1-\lambda^k| < 2\cos\phi_i = |\overline{AD}| \tag{15}$$

is sufficient to place $1-\alpha+\alpha\lambda^k$ inside S_1. Furthermore, for any $\lambda_j \in \lambda(\mathbf{X})$ such that $|\lambda_j| < 1$, $1-\alpha+\alpha\lambda_j^k$ lies inside S_1 since $1-\alpha+\alpha\lambda_j^k$ is an interpolation between two distinct points on/inside S_1. Hence, we conclude $\rho_k(\alpha) < 1$.

Conversely, assume $\rho_k(\alpha) < 1$ and suppose that there exists $\lambda_i \in \lambda_{\geq 1}(\mathbf{X})$ such that $\Re(\lambda_i^k) \geq 1$, i.e., $\tau_{k|\geq 1} \geq 1$. Then, we have a contradiction $\rho_k(\alpha) \geq 1$ since $|1-\alpha+\alpha\lambda_i^k| \geq 1$ for such λ_i. Additionally, suppose that there is $\lambda_i \in \lambda_{\geq 1}(\mathbf{X})$ such that $\Re(\lambda_i^k) < 1$ but $\alpha \geq \frac{2\cos\phi_i}{|\lambda_i^k - 1|}$. For such λ_i, we have

$$|\overline{AC}| = \alpha|1-\lambda_i^k| \geq 2\cos\phi_i = |\overline{AD}| \tag{16}$$

for A, C, D in Figure C.7 (a). This implies $|1-\alpha+\alpha\lambda^k| \geq 1$, a contradiction to the assumption $\rho_k(\alpha) < 1$. Therefore, we conclude $\tau_{k|\geq 1} < 1$ and $\alpha < c_1$.

The case for $\rho_0 = 1$: Assume $\tau_{k|\max} < 1$. Then, $\rho_k(\alpha) < 1$ is immediate since for any $\lambda_i \in \lambda(\mathbf{X})$, $1-\alpha+\alpha\lambda_i^k$ is an interpolation between two distinct points 1 and λ_i^k on the unit disk. Conversely assume $\rho_k(\alpha) < 1$. Then, by the definition of spectral radius, we have $\tau_{k|\max} < 1$.

The case for $\rho_0 < 1$: Assume $\tau_{k|\max} < \rho_0^{2k}$. Then, for any $\lambda_i \in \lambda(\mathbf{X})$, λ_i^k can be visualized as B in Figure C.7 (b), where the existence of D is guaranteed by $\Re(\lambda^k) < \rho_0^{2k}$. Then, we can see that, for A, C, D in Figure C.7 (b),

$$|\overline{AC}| = \alpha|1-\lambda_i^k| > \cos\phi_i - \Delta_i = |\overline{AD}| \tag{17}$$

is sufficient to place $1-\alpha+\alpha\lambda_i^k$ inside $S_{\rho_0^k}$.

Conversely, assume $\rho_k(\alpha) < \rho_0^k$ and suppose that there exists $\lambda_i \in \lambda_{\max}(\mathbf{X})$ such that $\Re(\lambda_i^k) \geq \rho_0^{2k}$, i.e., $\tau_{k|\max} \geq \rho_0^{2k}$. Then, we have $|1-\alpha+\alpha\lambda_i^k| \geq \rho_0^k$ for such λ_i since the line between 1 and λ_i^k cannot be secant to $S_{\rho_0^k}$. This contradicts the assumption $\rho_k(\alpha) < \rho_0^k$. Now suppose that there exists $\lambda_i \in \lambda(\mathbf{X})$ such that $\alpha \leq \frac{\cos\phi_i - \Delta_i}{|\lambda_i^k - 1|}$. For such λ_i, we have

$$|\overline{AC}| = \alpha|1-\lambda_i^k| \leq \cos\phi_i - \Delta_i = |\overline{AD}| \tag{18}$$

for A, C, D in Figure C.7 (b), implying $|1-\alpha+\alpha\lambda_i^k| \geq \rho_0^k$, which contradicts the assumption $\rho_k(\alpha) < \rho_0^k$. Therefore, we conclude $\tau_{k|\max} < \rho_0^{2k}$ and $\alpha > c_2$.

The inequality on c_1: Assume $c_1 > 0$. Then, we have $\cos\phi_i > 0$ for all $\lambda_i \in \lambda_{\geq 1}(\mathbf{X})$, and by definition of ϕ_i, it follows that every λ_i^k lies on the left side of the vertical line $\Re(z) = 1$, i.e. $\tau_{k|\geq 1} < 1$. Conversely, if $\tau_{k|\geq 1} < 1$, every λ_i^k lies on the left side of the vertical line $\Re(z) = 1$. Then, by definition of ϕ_i, we have $\cos\phi_i > 0$ for all $\lambda_i \in \lambda_{\geq 1}(\mathbf{X})$.

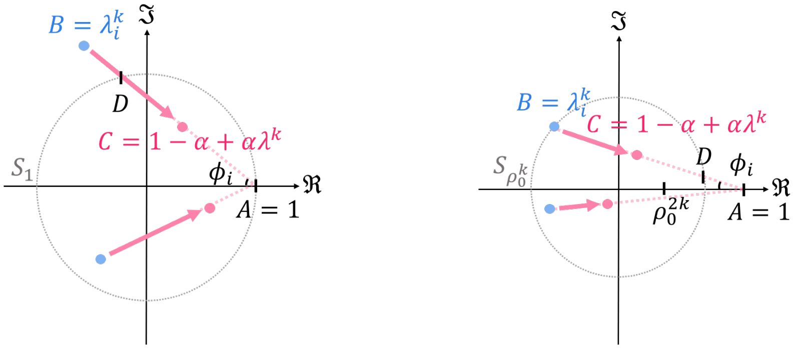
The inequality on c_2: Assume $\rho_0 < 1$, $c_2 < 1$, and suppose $\tau_{k|\max} \geq \rho_0^{2k}$. Then, there exists $\lambda_i \in \lambda_{\max}(\mathbf{X})$ such that $\Re(\lambda_i^k) \geq \rho_0^{2k}$. For such λ_i, it follows that

$$|\overline{AB}| = |1-\lambda_i^k| = \cos\phi_i - \Delta_i = |\overline{AD}| \tag{19}$$

for A, B, D in Figure C.7 (b). Therefore, we have $c_2 \geq 1$, a contradiction to the assumption $c_2 < 1$. Hence, we conclude $\tau_{k|\max} < \rho_0^{2k}$. Conversely, assume $\rho_0 < 1$, $\tau_{k|\max} < \rho_0^{2k}$ and suppose $c_2 \geq 1$. Then, there exists $\lambda_i \in \lambda(\mathbf{X})$ such that

$$|\overline{AB}| = |1-\lambda_i^k| \leq \cos\phi_i - \Delta_i = |\overline{AD}| \tag{20}$$

for A, B, D in Figure C.7 (b), implying $|\overline{AB}| = |\overline{AD}|$ since $|\overline{AB}| \geq |\overline{AD}|$. This implies λ_i lying on the circle $S_{\rho_0^k}$, and therefore, we have $\lambda_i \in \lambda_{\max}(\mathbf{X})$. Then, the assumption $\tau_{k|\max} < \rho_0^{2k}$ gives a contradiction $|\overline{AB}| > |\overline{AD}|$ for such λ_i, since it guarantees the line between A and B to be secant to the circle $S_{\rho_0^k}$. Therefore, we conclude $c_2 < 1$. □



(a) The case for $\rho_0 > 1$.

(b) The case for $\rho_0 < 1$.

Figure C.7: Eigenvalues of $(1-\alpha)\mathbf{I} + \alpha\mathbf{X}^k$ in the proof of Proposition 1.

Proposition 2 (Contraction bounds)**.** *Let $\mathbf{X} \in \mathbb{R}^{m\times m}$ be the Jacobian of a dynamics at an equilibrium, and denote its spectral radius by $\rho_0 \overset{\text{def}}{=} \rho(\mathbf{X})$ and the optimal radius of its Lookahead dynamics with $k\in\mathbb{N}$ by $\rho_k^* \overset{\text{def}}{=} \inf_{\alpha\in(0,1)} \rho_k(\alpha)$. Then, for $\tau_k \overset{\text{def}}{=} \max_{\lambda_i\in\lambda(\mathbf{X})} \Re(\lambda_i^k)$, the following statements hold:*

- *For $\rho_0 \geq 1$, the upper bound ${\rho_k^*}^2 \leq 1 - \frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} < 1$ holds if $\tau_k < 1$,*
- *For $\rho_0 < 1$, the upper bound ${\rho_k^*}^2 \leq 1 - \frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} < \rho_0^{2k}$ holds if $\tau_k < \rho_0^{2k}$,*
- *The lower bound $\rho_k^* \geq \max_{\lambda_i\in D} |\lambda_i|^k$ holds for the eigenvalues $D \overset{\text{def}}{=} \{\lambda_i\in\lambda(\mathbf{X}) : |\lambda_i^k - \frac{1}{2}| < \frac{1}{2}\}$.*

Proof. We start by observing the following inequalities:

$$
\begin{aligned}
{\rho_k^*}^2 &= \min_{\alpha\in(0,1)} \max_{\lambda_i\in\lambda(\mathbf{X})} |1-\alpha+\alpha\lambda_i^k|^2 && (21)\\
&\leq \min_{\alpha\in(0,1)} (\alpha-1)^2 + 2\alpha(1-\alpha)\tau_k + \alpha^2\rho_0^{2k} && (22)\\
&\leq (\alpha-1)^2 + 2\alpha(1-\alpha)\tau_k + \alpha^2\rho_0^{2k} && (23)\\
&= (1+\rho_0^{2k}-2\tau_k)\left(\alpha - \frac{1-\tau_k}{1+\rho_0^{2k}-2\tau_k}\right)^2 + 1 - \frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k}, && (24)
\end{aligned}
$$

where the first inequality follows from the maximum over each terms, and the second inequality holds for any $\alpha \in (0,1)$. We can see from the last equation that the upper bound $1 - \frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k}$ can be achieved when $\frac{1-\tau_k}{1+\rho_0^{2k}-2\tau_k} \in (0,1)$.

The upper bound of ${\rho_k^*}^2$ for $\rho_0 \geq 1$**:** Assume $\tau_k < 1$. Then, we have $\frac{1-\tau_k}{1+\rho_0^{2k}-2\tau_k} \in (0,1)$, since:

$$
\begin{aligned}
&1-\tau_k > 0, && (25)\\
&1+\rho_0^{2k}-2\tau_k \geq (1-\rho_0^k)^2 > 0, && (26)\\
&(1-\tau_k) - (1+\rho_0^{2k}-2\tau_k) = \tau_k - \rho^{2k} < 0. && (27)
\end{aligned}
$$

Therefore, we conclude

$$
{\rho_k^*}^2 \leq 1 - \frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} < 1. \qquad (28)
$$

The upper bound of ρ_k^{*2} for $\rho_0 < 1$: Assume $\tau_k < \rho_0^{2k}$. Then, we have $\frac{1-\tau_k}{1+\rho_0^{2k}-2\tau_k} \in (0,1)$, since:

$$1-\tau_k \geq 1-\rho_0^k > 0, \tag{29}$$
$$1+\rho_0^{2k}-2\tau_k \geq (1-\rho_0^k)^2 > 0, \tag{30}$$
$$(1-\tau_k)-(1+\rho_0^{2k}-2\tau_k) = \tau_k - \rho_0^{2k} < 0. \tag{31}$$

Furthermore, we also have

$$1-\frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} - \rho_0^{2k} = -\frac{(\rho_0^{2k}-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} < 0. \tag{32}$$

Therefore, we conclude

$$\rho_k^{*2} \leq 1-\frac{(1-\tau_k)^2}{1+\rho_0^{2k}-2\tau_k} < \rho_0^{2k}. \tag{33}$$

The lower bound of ρ_k^{*2}: The lower bound is immediate from the following inequalities:

$$\begin{aligned}
\rho_k^{*2} &= \min_{\alpha\in(0,1)} \max_{\lambda_i\in\lambda(\mathbf{X})} |1-\alpha+\alpha\lambda_i^k|^2 & (34)\\
&\geq \max_{\lambda_i\in\lambda(\mathbf{X})} \min_{\alpha\in(0,1)} |1-\alpha+\alpha\lambda_i^k|^2 & (35)\\
&= \max_{\lambda_i\in\lambda(\mathbf{X})} \min_{\alpha\in(0,1)} |1-\lambda_i^k|^2 \left(\alpha - \frac{\Re(1-\lambda_i^k)}{|1-\lambda_i^k|^2}\right)^2 + 1 - \left(\frac{\Re(1-\lambda_i^k)}{|1-\lambda_i^k|}\right)^2 & (36)\\
&\geq \max_{\lambda_i\in\lambda(\mathbf{X})} (1-a_i)\sin^2\phi(\lambda_i) + a_i|\lambda_i|^{2k} & (37)\\
&\geq \max_{\lambda_i\in\lambda(\mathbf{X})} a_i|\lambda_i|^{2k} = \max_{\lambda_i\in D} |\lambda_i|^{2k}, & (38)
\end{aligned}$$

where we define $a_i \overset{\text{def}}{=} \mathbf{1}_{|\lambda_i^k-\frac{1}{2}|<\frac{1}{2}}$. The first inequality follows from the min-max inequality, and the second inequality follows from the minimum of the quadratic with constraints $\alpha \in (0,1)$. □

Lemma 3 (Sufficient conditions for left-rotating k)**.** *Let $\mathbf{X}\in\mathbb{R}^{m\times m}$ be a Jacobian that can be written as $\mathbf{X} = \mathbf{I}-\eta\mathbf{J}$ for some $\mathbf{J}\in\mathbb{R}^{m\times m}$ and $\eta > 0$. Assume that a subset of the eigenvalues $S \subseteq \lambda(\mathbf{X})$ contains non-reals only, and every element of S has its conjugate pair in S. Then, for $\rho_0 \overset{\text{def}}{=} \rho(\mathbf{X})$, $\tau_k \overset{\text{def}}{=} \max_{\lambda_i\in S} \Re(\lambda_i^k)$, $\theta_{min} \overset{\text{def}}{=} \min_{\lambda_i\in S} |\operatorname{Arg}(\lambda_i)|$, $\theta_{max} \overset{\text{def}}{=} \max_{\lambda_i\in S} |\operatorname{Arg}(\lambda_i)|$, the following statements hold:*

- *When $\rho_0 > 1$, the eigenvalues in S are left-rotated so that $\tau_k < 1$ if $k \in (\beta_1,\beta_2)$, where $\beta_1,\beta_2 > 0$ are such that $\beta_1\theta_{min} = \arccos\rho_0^{-\beta_1}$ and $\beta_2\theta_{max} = 2\pi - \arccos\rho_0^{-\beta_2}$.*
- *When $\rho_0 < 1$, the eigenvalues in S are left-rotated so that $\tau_k < \rho_0^{2k}$ if $k \in (\beta_1,\beta_2)$, where $\beta_1,\beta_2 > 0$ are such that $\beta_1\theta_{min} = \arccos\rho_0^{\beta_1}$ and $\beta_2\theta_{max} = 2\pi - \arccos\rho_0^{\beta_2}$.*

The existence of a feasible $k \in (\beta_1,\beta_2)$ is guaranteed for a small enough $\eta > 0$ if the imaginary conditioning $\max_{\lambda_i,\lambda_j\in S} |\Im(\lambda_i)/\Im(\lambda_j)|$ of the subset of the eigenvalues S is less than 3.

Proof. We prove each of the statements in their order.

The case for $\rho_0 > 1$: Let $k \in (\beta_1,\beta_2)$ and define $\theta_i \overset{\text{def}}{=} \operatorname{Arg}(\lambda_i)$ for each $\lambda_i \in \lambda(S)$. Then, we have

$$k\theta_{\min} > \arccos 1/\rho_0^k, \tag{39}$$
$$k\theta_{\max} < 2\pi - \arccos 1/\rho_0^k, \tag{40}$$

since $\arccos 1/\rho_0^x$ is monotone and bounded for any $x > 0$. Hence, for each θ_i, we have

$$\arccos 1/\rho_0^k < k\theta_i < 2\pi - \arccos 1/\rho_0^k. \tag{41}$$

Notice that this implies $\cos k\theta_i < 1/\rho_0^k$ for each θ_i, since $\arccos 1/\rho_0^k \in (0, \pi/2)$ for any $k > 0$. Hence, we conclude

$$\tau_k = \max_{\lambda_i \in S} |\lambda_i|^k \cos k\theta_i \leq \rho_0^k \max_{\lambda_i \in S} \cos k\theta_i < 1. \tag{42}$$

The case for $\rho_0 < 1$: Let $k \in (\beta_1, \beta_2)$ and let $\theta_i \overset{\text{def}}{=} |\operatorname{Arg}(\lambda_i)|$ for each $\lambda_i \in \lambda(S)$. Then, we have

$$k\theta_{\min} > \arccos \rho_0^k, \tag{43}$$

$$k\theta_{\max} < 2\pi - \arccos \rho_0^k, \tag{44}$$

since $\arccos \rho_0^x$ is monotone and bounded for any $x > 0$. Hence, for each θ_i, we have

$$\arccos \rho_0^k < k\theta_i < 2\pi - \arccos \rho_0^k. \tag{45}$$

Notice that this implies $\cos k\theta_i < \rho_0^k$ for each θ_i, since $\arccos 1/\rho_0^k \in (0, \pi/2)$ for any $k > 0$. Hence, we conclude

$$\tau_k = \max_{\lambda_i \in S} |\lambda_i|^k \cos k\theta_i \leq \rho_0^k \max_{\lambda_i \in S} \cos k\theta_i < \rho_0^{2k}. \tag{46}$$

The existence of a feasible k: Note that we have the inequalities

$$\beta_1 < \frac{\pi}{2\theta_{\min}}, \quad \frac{3\pi}{2\theta_{\max}} < \beta_2, \tag{47}$$

since $\arccos(\cdot) < \frac{\pi}{2}$ and $2\pi - \arccos(\cdot) > \frac{3}{2}\pi$ for any positive numbers. Therefore, we have $\left(\frac{\pi}{2\theta_{\min}}, \frac{3\pi}{2\theta_{\max}}\right) \subset (\beta_1, \beta_2)$. For a scalar function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \frac{3\pi x}{\pi + 2x}$, one can easily verify that $\theta_{\max} < f(\theta_{\min})$ is equivalent to $\frac{3\pi}{2\theta_{\max}} - \frac{\pi}{2\theta_{\min}} > 1$, which implies nonempty $\mathbb{N} \cap (\beta_1, \beta_2)$. Hence, for $\Im_{\min} \overset{\text{def}}{=} \min_{\lambda_i \in S} |\Im(\lambda_i)|$ and $\Im_{\max} \overset{\text{def}}{=} \max_{\lambda_i \in S} |\Im(\lambda_i)|$, it suffices to show that the imaginary conditioning $\frac{\Im_{\max}}{\Im_{\min}} < 3$ implies the existence of $\delta > 0$ such that $\theta_{\max} < f(\theta_{\min})$ for any $\eta \in (0, \delta)$.

Let us define $\Re_{\min} \overset{\text{def}}{=} \min_{\lambda_i \in S} \Re(\lambda_i)$, $\Re_{\max} \overset{\text{def}}{=} \max_{\lambda_i \in S} \Re(\lambda_i)$, and a scalar function $H : \mathbb{R} \to \mathbb{R}$ as

$$H(\eta) \overset{\text{def}}{=} \left(1 + \frac{2{\theta_{\max}}^+}{\pi}\right) \left(\frac{1 + \eta\Re_{\max}}{1 + \eta\Re_{\min}}\right) \left(\frac{1 + 2\sec {\theta_{\min}}^-}{1 + 2\sec {\theta_{\max}}^+} + b\right),$$

where ${\theta_{\min}}^- \overset{\text{def}}{=} \arctan \frac{\eta\Im_{\min}}{1 + \eta\Re_{\max}}$, ${\theta_{\max}}^+ \overset{\text{def}}{=} \arctan \frac{\eta\Im_{\max}}{1 + \eta\Re_{\min}}$, and $b \overset{\text{def}}{=} \frac{(1 + 2\sec {\theta_{\min}}^-) \tan^4 {\theta_{\max}}^+}{540}$.

We show that the inequality

$$\frac{\Im_{\max}}{\Im_{\min}} < \frac{3}{H(\eta)} \tag{48}$$

implies $\theta_{\max} < f(\theta_{\min})$, and that there exists $\delta > 0$ such that Equation 48 holds for any $\eta \in (0, \delta)$ when $\frac{\Im_{\max}}{\Im_{\min}} < 3$.

The inequalities ${\theta_{\min}}^- \leq \theta_{\min}$ and $\theta_{\max} \leq {\theta_{\max}}^+$ directly follow from the definitions of ${\theta_{\min}}^-$ and ${\theta_{\max}}^+$. Furthermore, using the Shafer-type double inequalities (Mortici and Srivastava, 2014) for $\arctan(\cdot)$, we can obtain

$$\begin{aligned}
{\theta_{\min}}^- &\geq \frac{3 \tan {\theta_{\min}}^-}{1 + 2\sqrt{1 + \tan^2 {\theta_{\min}}^-}} = \frac{3\eta\Im_{\min}}{(1 + \eta\Re_{\max})(1 + 2\sec {\theta_{\min}}^-)}, \\
{\theta_{\max}}^+ &\leq \frac{3 \tan {\theta_{\max}}^+}{1 + 2\sqrt{1 + \tan^2 {\theta_{\max}}^+}} + \frac{1}{180} \tan^5 {\theta_{\max}}^+ \\
&= \frac{3\eta\Im_{\max}}{(1 + \eta\Re_{\min})(1 + 2\sec {\theta_{\max}}^+)} + \frac{\eta\Im_{\max} \tan^4 {\theta_{\max}}^+}{180(1 + \eta\Re_{\min})},
\end{aligned}$$

from which follows that

$$\frac{\theta_{\max}}{\theta_{\min}} \leq \frac{{\theta_{\max}}^+}{{\theta_{\min}}^-} = \frac{\Im_{\max}}{\Im_{\min}} \left(\frac{1 + \eta\Re_{\max}}{1 + \eta\Re_{\min}}\right) \left(\frac{1 + 2\sec {\theta_{\min}}^-}{1 + 2\sec {\theta_{\max}}^+} + b\right). \tag{49}$$

On the other hand, assuming Equation 48, we can derive

$$\frac{\Im_{\max}}{\Im_{\min}}\left(\frac{1+\eta\Re_{\max}}{1+\eta\Re_{\min}}\right)\left(\frac{1+2\sec{\theta_{\min}}^{-}}{1+2\sec{\theta_{\max}}^{+}}+b\right)<\frac{f({\theta_{\max}}^{+})}{{\theta_{\max}}^{+}}. \tag{50}$$

Furthermore, f is concave and monotonically increasing since $f'(x)=\frac{3\pi^2}{(\pi+2x)^2}$. Hence, we have

$$\frac{f({\theta_{\max}}^{+})}{{\theta_{\max}}^{+}}<\frac{f(\theta_{\min})}{\theta_{\min}}. \tag{51}$$

Consequently, combining the inequalities of Equation 49-51 gives us $\theta_{\max}<f(\theta_{\min})$.

Assume $\frac{\Im_{\max}}{\Im_{\min}}<3$ and define $\epsilon\overset{\mathsf{def}}{=}3-\frac{\Im_{\max}}{\Im_{\min}}>0$. By the continuity of $\frac{3}{H(\cdot)}$ at $\eta=0$ and $H(0)=1$, there exists $\delta>0$ such that $\left|3-\frac{3}{H(\eta)}\right|<\epsilon$ for any $\eta\in(0,\delta)$. Hence, we have $\frac{\Im_{\max}}{\Im_{\min}}=3-\epsilon<\frac{3}{H(\eta)}$ for any $\eta\in(0,\delta)$. Then, we obtain the chain of implications $\frac{\Im_{\max}}{\Im_{\min}}<\frac{3}{H(\eta)}\Longrightarrow\theta_{\max}<f(\theta_{\min})\Longrightarrow\frac{3\pi}{2\theta_{\max}}-\frac{\pi}{2\theta_{\min}}>1$, which implies nonempty $\mathbb{N}\cap(\beta_1,\beta_2)$, i.e., the existence of a feasible k. □

Theorem 4 (Local stabilization)**.** *Let $\mathbf{x}^*\in\mathbb{R}^n$ be an unstable equilibrium of dynamics F with the spectral radius ρ_0. Assume each element of $S=\lambda_{\geq 1}(\nabla_{\mathbf{x}}F(\mathbf{x}^*))$ is non-real. Then, $\mathbf{x}^*$ becomes locally asymptotically stable in its Lookahead dynamics with $k\in\mathbb{N},\alpha\in(0,1)$ if $k\in(\beta_1,\beta_2)$ and $\alpha<c$, where $\beta_1,\beta_2>0,c\in\mathbb{R}$ are such that $\beta_1\theta_{min}(S)=\arccos\rho_0^{-\beta_1},\beta_2\theta_{max}(S)=2\pi-\arccos\rho_0^{-\beta_2}$, and $c=\min\limits_{\lambda_i\in S}\frac{2\cos\phi_i}{|1-\lambda_i^k|}>0,\forall k\in(\beta_1,\beta_2)$ with $\phi_i\overset{\mathsf{def}}{=}\operatorname{Arg}(1-\lambda_i^k)$.*

Proof. By the instability, we have $\rho_0>1$. Furthermore, by the assumption, each element of S is non-real and has its conjugate pair in S. Hence, for any $k\in(\beta_1,\beta_2)$, the first case of Lemma 3 gives $\tau_{k|\geq 1}\overset{\mathsf{def}}{=}\max\limits_{\lambda_i\in S}\Re(\lambda_i^K)<1$, and Proposition 1 guarantees $c>0$. As a result, the first case of Proposition 1 gives $\rho(\nabla_{\mathbf{x}}G(\mathbf{x}^*))<1$ for any $k\in(\beta_1,\beta_2)$ and $\alpha<c$. Then, it follows from Lemma B.1 that $\mathbf{x}^*$ is locally asymptotically stable in G for any $k\in(\beta_1,\beta_2)$ and $\alpha<c$. □

Theorem 5 (Local acceleration)**.** *Let $\mathbf{x}^*\in\mathbb{R}^n$ be an equilibrium of dynamics F with the spectral radius $\rho_0<1$. Assume each element of $S=\lambda_{max}(\nabla_{\mathbf{x}}F(\mathbf{x}^*))$ is non-real. Then, the local convergence rate to $\mathbf{x}^*$ in its Lookahead dynamics with $k\in\mathbb{N},\alpha\in(0,1)$ improves upon F if $k\in(\beta_1,\beta_2)$ and $\alpha>c$, where $\beta_1,\beta_2>0,c\in\mathbb{R}$ are such that $\beta_1\theta_{min}(S)=\arccos\rho_0^{\beta_1},\beta_2\theta_{max}(S)=2\pi-\arccos\rho_0^{\beta_2}$, and $c=\max\limits_{\lambda_i\in S}\frac{\cos\phi_i-\sqrt{\rho_0^{2k}-\sin^2\phi_i}}{|1-\lambda_i^k|}<1,\forall k\in(\beta_1,\beta_2)$ with $\phi_i\overset{\mathsf{def}}{=}\operatorname{Arg}(1-\lambda_i^k)$.*

Proof. By the assumption, we have $\rho_0<1$. Furthermore, each element of S is non-real and has its conjugate pair in S. Then, for any $k\in(\beta_1,\beta_2)$, the second case of Lemma 3 gives $\tau_{k|\max}\overset{\mathsf{def}}{=}\max\limits_{\lambda_i\in S}\Re(\lambda_i^k)<\rho_0^{2k}$, and Proposition 1 guarantees $c<1$. As a result, the last case of Proposition 1 yields $\rho(\nabla_{\mathbf{x}}G(\mathbf{x}^*))<\rho_0^k$ for any $k\in(\beta_1,\beta_2)$ and $\alpha>c$. Then, it follows from Lemma B.1 that the local convergence rate of a Lookahead dynamics G is improves upon its base dynamics F if $k\in(\beta_1,\beta_2)$ and $\alpha>c$, assuming the amortized computation over k forward steps. □

C.3 General Bilinear Games

Corollary 6 (Stabilization of $\mathrm{GD}_{\mathrm{Alt}}$)**.** *Lookahead dynamics $G_{LA\text{-}GD_{Alt}}$ with $\eta\in(0,2\sigma_{max}^{-1}),k\in\mathbb{N},\alpha\in(0,1)$ converges to a Nash equilibrium of Equation 10 if $k\arccos(1-\eta^2\sigma_i^2/2)\bmod 2\pi\neq 0,\forall\sigma_i$.*

Proof. For any learning rate $\eta\in(0,2\sigma_{\max}^{-1})$, the eigenvalues of the Jacobian $\nabla_{\mathbf{x}}G_{\text{LA-GD}_{\text{Alt}}}$ can be written as $1-\alpha+\alpha\lambda_{\pm i}^k$ for each $\lambda_{\pm i}\overset{\mathsf{def}}{=}1-\eta^2\sigma_i^2/2\pm i\eta\sigma_i\sqrt{1-\eta^2\sigma_i^2/4}\in\lambda(\nabla_{\mathbf{x}}F_{\text{GD}_{\text{Alt}}})$ with unit modulus $|\lambda_{\pm i}|=1$. Note that $\Re(\lambda_{\pm i}^k)<1$ holds for each σ_i if $k\arccos(1-\eta^2\sigma_i^2/2)\bmod 2\pi\neq 0$ for each $\lambda_{\pm i}$, and the second case of Proposition 1 gives $\rho(\nabla_{\mathbf{x}}G_{\text{LA-GD}_{\text{Alt}}})<1$ if $\Re(\lambda_{\pm i}^k)<1$ for each $\lambda_{\pm i}$. Therefore, the first case of Lemma B.2 implies that a Lookahead dynamics LA-GD_{Alt} converges to a Nash equilibrium of Equation 10 if $k\arccos(1-\eta^2\sigma_i^2/2)\bmod 2\pi\neq 0,\forall\sigma_i\in\sigma_i(\Sigma_r)$. □

Corollary 7 (Stabilization of $\mathrm{GD}_{\mathrm{Sim}}$)**.** *Lookahead dynamics $G_{LA\text{-}GD_{Sim}}$ with $\eta > 0, k \in \mathbb{N}, \alpha \in (0,1)$ converges to a Nash equilibrium of Equation 10 if $k \in (\beta_1, \beta_2), \alpha < c$, where $\beta_1, \beta_2 > 0, c \in \mathbb{R}$ are such that*

$$\beta_1 \arctan \eta\sigma_{min} = \arccos \rho_0^{-\beta_1},$$
$$\beta_2 \arctan \eta\sigma_{max} = 2\pi - \arccos \rho_0^{-\beta_2}$$

$$c = \min_{\sigma_i \in \sigma(\Sigma_r)} \frac{2\cos\phi_i}{|1 - (1 + i\eta\sigma_i)^k|} > 0, \forall k \in (\beta_1, \beta_2),$$

with $\rho_0 \overset{\text{def}}{=} \sqrt{1 + \eta^2\sigma_{max}^2}$, $\phi_i \overset{\text{def}}{=} \mathrm{Arg}(1 - (1 + i\eta\sigma_i)^k)$.

Proof. For any base learning rate $\eta > 0$, the eigenvalues of the Jacobian $\nabla_{\mathbf{x}} G_{\text{LA-GD}_{\text{Sim}}}$ can be written as $1-\alpha+\alpha\lambda_{\pm i}^k$ for each $\lambda_{\pm i} \overset{\text{def}}{=} 1 \pm i\eta\sigma_i \in \lambda(\nabla_{\mathbf{x}} F_{\text{GD}_{\text{Sim}}})$ with modulus $|\lambda_{\pm i}| > 1$. Define the smallest and largest angles of the eigenvalues $\lambda_{\pm i}$ as $\theta_{\min} \overset{\text{def}}{=} \min_{\sigma_i} \arctan \eta\sigma_i = \arctan \eta\sigma_{\min}$ and $\theta_{\max} \overset{\text{def}}{=} \max_{\sigma_i} \arctan \eta\sigma_i = \arctan \eta\sigma_{\max}$, respectively. Then, for any $k \in (\beta_1, \beta_2)$, the first case of Lemma 3 gives $\tau_{k|\geq 1} \overset{\text{def}}{=} \max_{\lambda_{\pm i}} \Re(\lambda_{\pm i}^k) < 1$, and Proposition 1 guarantees $c > 0$. Furthermore, for any $k \in (\beta_1, \beta_2)$ and $\alpha < c$, the first case of Proposition 1 implies $\rho(\lambda(\nabla_{\mathbf{x}} F_{\text{GD}_{\text{Sim}}})) < 1$. Therefore, the first case of Lemma B.2 implies that a Lookahead dynamics $\text{LA-GD}_{\text{Sim}}$ converges to a Nash equilibrium of Equation 10 if $k \in (\beta_1, \beta_2)$ and $\alpha < c$. □

Corollary 8 (Acceleration of $\text{LA-EG}_{\text{Sim}}$)**.** *Lookahead dynamics $G_{LA\text{-}EG_{Sim}}$ with $\eta \in (0, \sigma_{max}^{-1}), k \in \mathbb{N}, \alpha \in (0,1)$ converges to a Nash equilibrium of Equation 10. The rate of convergence is improved upon its base dynamics $F_{EG_{Sim}}$ if $\eta \in (0, \sigma_{max}^{-1}/2), k \in (\beta_1, \beta_2), \alpha > c$, where $\beta_1, \beta_2 > 0$ and $c \in \mathbb{R}$ are such that*

$$\beta_1 \arctan \eta\sigma_{min}(1 - \eta\sigma_{min})^{-1} = \arccos \rho_0^{\beta_1},$$
$$\beta_2 \arctan \eta\sigma_{min}(1 - \eta\sigma_{min})^{-1} = 2\pi - \arccos \rho_0^{\beta_2},$$
$$c = \max_{\sigma_i \in \sigma(\Sigma_r)} \frac{\cos\phi_i - \sqrt{\rho_0^{2k} - \sin^2\phi_i}}{|1 - (1 - \eta\sigma_i + i\eta\sigma_i)^k|} < 1, \forall k \in (\beta_1, \beta_2),$$

with $\rho_0 \overset{\text{def}}{=} \sqrt{1 - 2\eta\sigma_{min} + 2\eta^2\sigma_{min}^2}$, $\phi_i \overset{\text{def}}{=} \mathrm{Arg}(1 - (1 - \eta\sigma_i + i\eta\sigma_i)^k)$.

Proof. For any learning rate $\eta > 0$, the eigenvalues of the Jacobian $\nabla_{\mathbf{x}} G_{\text{LA-EG}_{\text{Sim}}}$ can be written as $1 - \alpha + \alpha\lambda_{\pm i}^k$ for each $\lambda_{\pm i} \overset{\text{def}}{=} 1 - \eta\sigma_i + i\eta\sigma_i \in \lambda(\nabla_{\mathbf{x}} F_{\text{EG}_{\text{Sim}}})$. Additionally, we have the modulus $|\lambda_{\pm i}| = \sqrt{1 - 2\eta\sigma_i + 2\eta^2\sigma_i^2} < 1$ for each $\lambda_{\pm i}$ when $\eta < \sigma_{\max}^{-1}$. Hence, $|1 - \alpha + \alpha\lambda_{\pm i}^k| < 1$ holds for each $\lambda_{\pm i}$ for any $\eta < \sigma_{\max}^{-1}$, since $1 - \alpha + \alpha\lambda_{\pm i}^k$ is an interpolation between two distinct points on the unit disk. Therefore, it follows from the first case of Lemma B.2 that Lookahead dynamics $\text{LA-EG}_{\text{Sim}}$ converges to a Nash equilibrium of Equation 10 for any $\eta \in (0, \sigma_{\max}^{-1}), k \in \mathbb{N}, \alpha \in (0,1)$.

Now assume $\eta \in (0, \sigma_{\max}^{-1}/2)$ and notice that the quadratic $|\lambda_{\pm i}|^2 = 2\eta^2(\sigma_i - \frac{1}{2\eta})^2 + \frac{1}{2}$ holds for each $\lambda_{\pm i}$. As a result, we have the set of eigenvalues with the maximal modulus $\lambda_{\max}(\nabla_{\mathbf{x}} F_{\text{EG}_{\text{Sim}}}) = \{1 - \eta\sigma_{\min} \pm i\eta\sigma_{\min}\}$ and their angles $\theta_{\min} = \theta_{\max} = \arctan \eta\sigma_{\min}(1 - \eta\sigma_{\min})^{-1}$. Hence, for any $k \in (\beta_1, \beta_2)$, the second case of Lemma 3 gives us $\tau_{k|\max} \overset{\text{def}}{=} \max_{\lambda_{\pm i}} \Re(\lambda_{\pm i}^k) < \rho_0^{2k}$, and Proposition 1 guarantees $c < 1$ for any $k \in (\beta_1, \beta_2)$. Furthermore, for any $k \in (\beta_1, \beta_2)$ and $\alpha > c$, the last case of Proposition 1 implies $\rho(\nabla_{\mathbf{x}} G_{\text{LA-EG}_{\text{Sim}}}) < \rho_0^k$. Then, it follows from the first case of Lemma B.2 that the convergence rate of Lookahead dynamics $\text{LA-EG}_{\text{Sim}}$ improves upon its base dynamics EG_{Sim} if $\eta \in (0, \sigma_{\max}^{-1}/2), k \in (\beta_1, \beta_2), \alpha > c$, assuming the amortized computation over k forward steps. □

Corollary C (Acceleration of $\text{LA-PP}_{\text{Sim}}$)**.** *Lookahead dynamics $G_{LA\text{-}PP_{Sim}}$ with $\eta > 0, k \in \mathbb{N}, \alpha \in (0,1)$ converges to a Nash equilibrium of Equation 10. The rate of convergence is improved upon its base dynamics $F_{PP_{Sim}}$ if $k \in (\beta_1, \beta_2)$ and $\alpha > c$, where $\beta_1, \beta_2 > 0$ and $c \in \mathbb{R}$ are such that*

$$\beta_1 \arctan \eta\sigma_{min} = \arccos \rho_0^{\beta_1},$$
$$\beta_2 \arctan \eta\sigma_{min} = 2\pi - \arccos \rho_0^{\beta_2},$$

$$c = \max_{\sigma_i \in \sigma(\Sigma_r)} \frac{\cos\phi_i - \sqrt{\rho_0^{2k} - \sin^2\phi_i}}{\left|1 - \left(\frac{1+i\eta\sigma_i}{1+\eta^2\sigma_i^2}\right)^k\right|} < 1, \forall k \in (\beta_1, \beta_2),$$

with $\rho_0 \overset{\mathsf{def}}{=} \frac{1}{\sqrt{1+\eta^2\sigma_{min}^2}}$, $\phi_i \overset{\mathsf{def}}{=} \mathrm{Arg}\left(1 - \left(\frac{1+i\eta\sigma_i}{1+\eta^2\sigma_i^2}\right)^k\right)$.

Proof. For any learning rate $\eta > 0$, the eigenvalues of the Jacobian $\nabla_{\mathbf{x}} G_{\text{LA-PP}_{\text{Sim}}}$ can be written as $1 - \alpha + \alpha\lambda_{\pm i}^k$ for each $\lambda_{\pm i} \overset{\mathsf{def}}{=} \frac{1 \pm i\eta\sigma_i}{1+\eta^2\sigma_i^2} \in \lambda(\nabla_{\mathbf{x}} F_{\text{PP}_{\text{Sim}}})$ with modulus $|\lambda_{\pm i}| = \frac{1}{\sqrt{1+\eta^2\sigma_i^2}} < 1$. As a result, $|1 - \alpha + \alpha\lambda_{\pm i}^k| < 1$ holds for each $\lambda_{\pm i}$ since $1 - \alpha + \alpha\lambda_{\pm i}^k$ is an interpolation between two distinct points on the unit disk. Therefore, it follows from the first case of Lemma B.2 that Lookahead dynamics LA-PP$_{\text{Sim}}$ converges to a Nash equilibrium of Equation 10 for any $\eta > 0, k \in \mathbb{N}, \alpha \in (0, 1)$.

Notice we have the set of eigenvalues with the maximal modulus $\lambda_{\max}(\nabla_{\mathbf{x}} F_{\text{PP}_{\text{Sim}}}) = \left\{\frac{1 \pm i\eta\sigma_{\min}}{1+\eta^2\sigma_{\min}^2}\right\}$ and their angles $\theta_{\min} = \theta_{\max} = \arctan \eta\sigma_{\min}$. Hence, for any $k \in (\beta_1, \beta_2)$, the second case of Lemma 3 gives $\tau_{k|\max} \overset{\mathsf{def}}{=} \max_{\lambda_{\pm i}} \Re(\lambda_{\pm i}^k) < \rho_0^{2k}$, and Proposition 1 guarantees $c < 1$. Furthermore, for any $k \in (\beta_1, \beta_2)$ and $\alpha > c$, the last case of Proposition 1 implies $\rho(\nabla_{\mathbf{x}} G_{\text{LA-PP}_{\text{Sim}}}) < \rho_0^k$. Then, it follows from the first case of Lemma B.2 that the convergence rate of Lookahead dynamics LA-PP$_{\text{Sim}}$ improves upon its base dynamics PP$_{\text{Sim}}$ if $k \in (\beta_1, \beta_2)$ and $\alpha > c$, assuming the amortized computation over k forward steps. □

C.4 Stabilization Effect in GANs

Recently, No et al. (2021) have shown that 2-layer random-feature fully-connected WGANs with wide generator exhibit no spurious stationary points, i.e., each stationary point is a Nash equilibrium of the game. As a simple, yet crucial corollary, we can guarantee that Lookahead does not introduce a spurious stabilization in such settings. For completeness, we restate the underlying assumptions and one of the main result of No et al. (2021) below.

Assumption 1 (AL). *A continuous random vector $Z \in \mathbb{R}^k$ has a Lipschitz continuous probability density function $q_Z(z)$ satisfying $q_Z(z) > 0$ for all $z \in \mathbb{R}^k$.*

Assumption 2 (AG). *Let $\mathcal{G} = \{\phi(\cdot;\kappa)\kappa \in \mathbb{R}^p|\}$, where $\phi(\cdot;\kappa) : \mathbb{R}^k \to \mathbb{R}^n$, be a collection of generator feature functions such that $\phi \in \mathcal{G}$ are of form $\phi(z;\kappa) = \sigma_g(\kappa_w z + \kappa_b)$, where $\kappa = (\kappa_w, \kappa_b) \in \mathbb{R}^{n\times k} \times \mathbb{R}^n$, and $\sigma_g : \mathbb{R} \to \mathbb{R}$ is a bounded continuous activation function satisfying $\lim_{r\to-\infty} \sigma_g(r) < \lim_{r\to\infty} \sigma_g(r)$.*

Assumption 3 (AD). *Let $\mathcal{D} = \{\psi_1, \ldots, \psi_{N_d}\}$ be a class of discriminator feature functions $\psi_j : \mathbb{R}^n \to \mathbb{R}$ for each $1 \le j \le N_d$ such that each $\psi_j \in \mathcal{D}$ has a form of $\psi_j(x) = \sigma(a_j^T x + b_j)$ for some $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$. The twice differentiable activation function σ satisfies $\sigma'(x) > 0$ for all $x \in \mathbb{R}$ and $\sup_{x\in\mathbb{R}} |\sigma(x)| + |\sigma'(x)| + |\sigma''(x)| \le \infty$. The weights $a_1, \ldots, a_{N_d}$ and biases $b_1, \ldots, b_{N_d}$ are sampled (IID) from a distribution with a probability density function.*

Assumption 4. *The first n parameters $\{\kappa_i\}_{i=1}^n$ of generator random feature functions are chosen so that $\{\phi_i\}_{i=1}^n$ are constant functions spanning the sample space R^n, and the remaining parameters $\{\kappa_i\}_{i=n+1}^{N_g}$ are sampled (IID) from a probability distribution that has a continuous and strictly positive density function.*

Generator. For the generator feature functions $\phi_1, \ldots, \phi_{N_g} \in \mathcal{G}$ with $1 \le N_g < \infty$ and $\theta \in \mathbb{R}^{N_g}$, a 2-layer random-feature generator is given by

$$g_\theta(z) = \sum_{i=1}^{N_g} \theta_i \phi_i(z), \tag{52}$$

and the class of generators constructed from the feature functions $\{\phi_i\}_{i=1}^{N_g}$ is written as

$$\mathrm{span}(\{\phi_i\}_{i=1}^{N_g}) = \{g_\theta : \theta \in \mathbb{R}^{N_g}\}. \tag{53}$$

Discriminator. For $\eta \in \mathbb{R}^{N_d}$ and $\Psi(x) = (\psi_1(x), \dots, \psi_{N_d}(x)) \in \mathbb{R}^{N_d}$, a 2-layer random-feature discriminator is given by

$$f_\eta(x) = \sum_{j=1}^{N_d} \eta_j \psi_j(x) = \eta^T \Psi(x), \tag{54}$$

and the class of discriminators constructed from the feature functions in $\mathcal{D}$ is written as

$$\operatorname{span}(\mathcal{D}) = \{f_\eta : \eta \in \mathbb{R}^{N_d}\}. \tag{55}$$

The game. For the Wasserstein GAN loss function

$$L(\theta, \eta) = \mathbb{E}_X[f_\eta(X)] - \mathbb{E}_Z[f_\eta(g_\theta(Z))] - \frac{1}{2}\left\|\eta\right\|_2^2, \tag{56}$$

where the Lipschitz constraint on the discriminator has been replaced with an explicit $\ell 2$ regularizer, the game between the generator g_θ and the discriminator f_η is given by

$$\inf_\theta \sup_\eta L(\theta, \eta) = \inf_\theta J(\theta), \text{ where } J(\theta) \overset{\text{def}}{=} \sup_\eta L(\theta, \eta). \tag{57}$$

Under Assumption 1–4, No et al. (2021) shows that 2-layer random-feature WGANs with a small discriminator, and a sufficiently wide generator does not exhibit a spurious stationary point.

Theorem (Theorem 9, No et al. (2021)). *Let the discriminator hidden layer width be $N_d \leq n$ for the sample space dimension $n \in \mathbb{N}$. Assume Assumption 1–4. Then, for any $C > 0$ and $\zeta > 0$, there exists a large enough generator hidden layer width $N_g \in \mathbb{N}$ such that the following statement holds with probability at least $1 - \zeta$: any stationary point $\theta_s \in \mathbb{R}^{N_g}$ satisfying $\left\|\theta_s\right\|_1 \leq C$ is a global minimum of $J(\cdot)$.*

A simple, yet crucial corollary of the above theorem is that the stabilization effect of Lookahead introduces benefits for the game optimization and has no spurious effect, i.e., only stabilizes the Nash equilibria of the game.

Corollary 9 (No spurious stabilization for wide 2-layer WGANs). *For the sample space dimension $n \in \mathbb{N}$, under Assumption 1–4 on 2-layer random-feature WGANs with discriminator hidden layer width $N_d \leq n$ and $\ell 2$-regularized discriminator, for any $C > 0$ and $\zeta > 0$, there exists a large enough generator hidden layer width $N_g \in \mathbb{N}$ such that the following statement holds with probability at least $1 - \zeta$: a set of stationary points $\{(\theta_s, \eta_s) \in \mathbb{R}^{N_g + N_d} : \nabla_{(\theta_s, \eta_s)} L(\theta_s, \eta_s) = 0, \left\|\theta_s\right\|_1 \leq C\}$ that are stabilized by Lookahead contains no other points than the Nash equilibria of Equation 57.*

Proof. By the equivalance between gradient descent dynamics of Equation 57 and gradient descent minimization of $J(\cdot)$ (Section 2.3, No et al. (2021)), each stationary point of $J(\cdot)$ is also a stationary point of the game. As a result, each stationary of the game is a Nash equilibrium (Theorem 9, No et al. (2021)), and therefore, the stationary points that are stabilized by Lookahead contains no other points than the Nash equilibra of Equation 57. □

D EXPERIMENTAL DETAILS

D.1 Nonlinear Game

We use $\epsilon = 0.001$ for the nonlinear game experiment in Figure 3 and Section 6. Using the automatic differentiation package provided by PyTorch, we verify with Proposition B.3 that $(0, 0)$ is a local Nash equilibrium of the game. For a fixed learning rate $\eta = 0.1$, we use the automatic differentiation to inspect the spectral radius of each dynamics, and actually verify that the equilibrium $(0, 0)$ is unstable for simultaneous gradient descent GD_{Sim}, and asymptotically stable for extragradient EG_{Sim}. Then, we use Theorem 4 to stabilize GD_{Sim} and Theorem 5 to accelerate EG_{Sim}. For each theorem, we compute the constants β_1, β_2 by solving the implicit equations in each theorem for β_1, β_2 with the numerical solver provided by WolframAlpha. Then, we evaluate the precise threshold c for α over a range of k that covers (β_1, β_2). Lastly, we compute the spectral radius of each Lookahead dynamics LA-GD_{Sim} and LA-EG_{Sim} for a range of (k, α) that covers the sufficient conditions of Theorem 4–5, and illustrate their relative sizes in Figure 6 (a)–(b). We report the actual values of (β_1, β_2) and c in Table D.2.

Table D.2: The constants β_1, β_2, c in the nonlinear game experiments of Section 6.

Name	β_1, β_2	(k, c)
Theorem 4	1.50, 26.58	(2, 0.24), (4, 0.60), (6, 0.71), (8, 0.76), (10, 0.79), (12, 0.79), (14, 0.79), (16, 0.77), (18, 0.73), (20, 0.67), (22, 0.57), (24, 0.40), (26, 0.11), (28, -0.36)
Theorem 5	0.04, 27.16	(2, 0.20), (4, 0.10), (6, 0.07), (8, 0.06), (10, 0.05), (12, 0.05), (14, 0.06), (16, 0.06), (18, 0.08), (20, 0.11), (22, 0.16), (24, 0.27), (26, 0.56), (28, 1.00)
Corollary 7	6.09, 37.51	(5, 0.79), (10, 0.88), (15, 0.89), (20, 0.89), (25, 0.86), (30, 0.76), (35, 0.46), (40, -1.17), (45, -3.63), (50, -4.13), (55, -4.63), (60, -5.13), (65, -5.64), (70, -6.14)
Corollary 8	20.31, 76.16	(15, 1.00), (19, 1.00), (23, 0.93), (27, 0.88), (31, 0.85), (35, 0.87), (39, 0.90), (43, 0.92), (47, 0.94), (51, 0.95), (55, 0.96), (59, 0.97), (63, 0.97), (67, 0.98), (71, 0.98), (75, 0.99)

For the experiments in Section 5, we choose k and α that are predicted by Theorem 4–5 to stabilize GD_{Sim} and accelerate EG_{Sim}. Specifically, we choose $(k, \alpha) = (50, 0.3)$ for $\text{LA-GD}_{\text{Sim}}$ and $\text{LA-GD}^{\text{NM}}_{\text{Sim}}$, and $(k, \alpha) = (50, 0.7)$ for $\text{LA-EG}_{\text{Sim}}$ and $\text{LA-GD}^{\text{NM}}_{\text{Alt}}$. We use negative momentum coefficients $m = -0.2$ for both $\text{GD}^{\text{NM}}_{\text{Sim}}$ and $\text{LA-GD}^{\text{NM}}_{\text{Sim}}$, and $m = -0.9$ for $\text{GD}^{\text{NM}}_{\text{Alt}}$ and $\text{LA-GD}^{\text{NM}}_{\text{Alt}}$. The trajectories and progress in Figure 3 are evaluated for every k base-dynamics steps.

D.2 Bilinear Game

For the bilinear game experiments in Section 6, we use $\epsilon = 0.01$ and $n = 1000$. This gives a sample of the game matrix $\mathbf{A}_n$ with $\sigma_{\max} = 1.462$ and $\sigma_{\min} = 0.583$, which yields the conditioning $\frac{\sigma_{\max}}{\sigma_{\min}} = 2.505 < 3$. We fix the base learning rate $\eta = 0.1$ for all the experiments. For Corollary 7–8, we compute the constants β_1, β_2 by solving the implicit equations of β_1, β_2 in each theorem using the numerical solver provided by WolframAlpha. Then, for each theorem, we evaluate the precise threshold c for α over a range of k that covers (β_1, β_2). Lastly, we compute the spectral radius of each Lookahead dynamics $\text{LA-GD}_{\text{Alt}}$, $\text{LA-GD}_{\text{Sim}}$, $\text{LA-PP}_{\text{Sim}}$ and $\text{LA-EG}_{\text{Sim}}$ for a range of (k, α) that covers the sufficient conditions of Corollary 6–8, and illustrate their relative sizes in Figure 6 (c)–(d). We report the actual values of (β_1, β_2) and c in Table D.2.

D.3 GANs

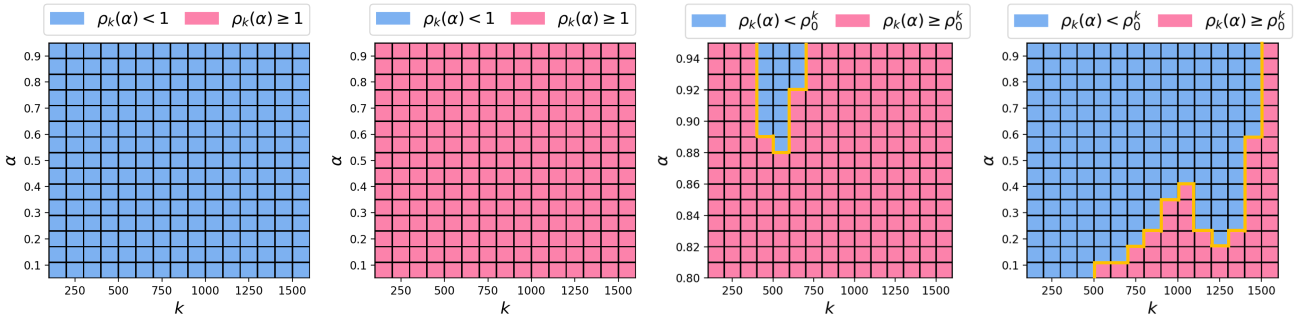
For the GAN experiment in Section 5, we train a SNDCGAN Miyato et al. (2018) with the non-saturating loss function Goodfellow et al. (2014) on CIFAR-10 using $\text{LA-Adam}_{\text{Alt}}$ with $k = 5000, \alpha = 0.5$. We use a base learning rate 0.0003 for the discriminator, and 0.0001 for the generator. We use Adam hyperparameters $\beta_{\text{Adam}} = (0.5, 0.999)$ and weight decay 0.0003 for the both networks. Following Kurach et al. (2019), we report the FID Heusel et al. (2017) scores between 10k generated samples and 10k test samples of CIFAR 10 dataset. We report our results over 8 different random initializations, and present the mean and standard deviation of each metric with the solid lines and the shaded area of Figure 5, respectively.

E ADDITIONAL EXPERIMENTS

E.1 Bilinear Game

In this section, we verify our results on a bilinear game with larger conditioning. Specifically, we use $n = 2000$ and $\epsilon = 0.02$ to obtain a sample of game matrix $\mathbf{A}_n$ with larger conditioning $\frac{\sigma_{\max}}{\sigma_{\min}} = 56.84 > 3$. We fix the learning rate $\eta = 0.1$ for all the dynamics throughout the experiments, and follow the same protocol as in the bilinear experiment of Section 6. We illustrate the results in Figure E.8 and report the actual values of the constants β_1, β_2, c in Table E.3.

Figure E.8 (a) shows that the convergence guarantee of Corollary 6 still holds for any k and α even in a game of larger conditioning. On the other hand, Figure E.8 (b) shows that Lookahead fails to stabilize GD_{Sim} for any k and α. This result suggests that the non-necessity of Lemma 3 might be small in practice. Figure E.8

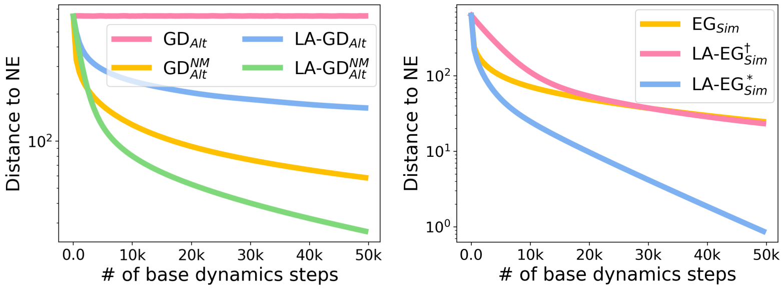


(a) Corollary 6 – LA-GD$_{\text{Alt}}$. (b) Corollary 7 – LA-GD$_{\text{Sim}}$. (c) Corollary 8 – LA-EG$_{\text{Sim}}$. (d) Corollary C – LA-PP$_{\text{Sim}}$.

Figure E.8: **Stabilization and acceleration over a range** of (k, α) in the bilinear game experiment in Section E.1. Each plot illustrates the relative size of spectral radius of Lookahead $\rho_k(\alpha)$ over a range of (k, α). Each blue cell represents (k, α) that achieves either stabilization $\rho_k(\alpha) < 1$ or acceleration $\rho_k(\alpha) < \rho_0^k$ with respect to the spectral radius ρ_0 of its base dynamics. The blue cells contained in the yellow lines represent the improvements predicted by each theorem, i.e., (k, α) that satisfies $k \in (\beta_1, \beta_2)$ and $\alpha < c$ or $\alpha > c$. Each red cell indicates (k, α) that fails to improve the convergence.

Table E.3: The constants β_1, β_2, c in the bilinear game experiment in Section E.1.

Name	β_1, β_2	(k, c)
Corollary C	1, 1442.94	(100, 0.01), (200, 0.01), (300, 0.04), (400, 0.07), (500, 0.11), (600, 0.16), (700, 0.20), (800, 0.28), (900, 0.36), (1000, 0.45), (1100, 0.28), (1200, 0.17), (1300, 0.26), (1400, 0.59), (1500, 1.00)
Corollary 8	302.85, 1108.05	(100, 1.00), (200, 1.00), (300, 1.00), (400, 1.00), (500, 0.89), (600, 0.88), (700, 0.92), (800, 0.95), (900, 0.96), (1000, 0.98), (1100, 0.99), (1200, 1.00), (1300, 1.00), (1400, 1.00), (1500, 1.00)

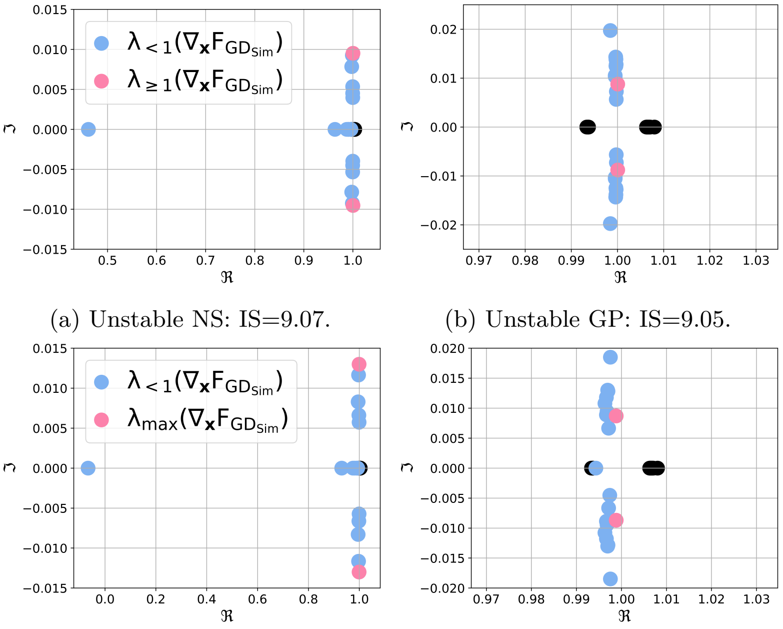


(a) Convergence of LA-GD$_{\text{Alt}}$. (b) Convergence of LA-EG$_{\text{Sim}}$.

Figure E.9: **Convergence of each Lookahead dynamics in the bilinear game experiment** in Section E.1. **(a)**. The base dynamics GD$_{\text{Alt}}$ fails to converge towards the Nash equilibrium of the game. However, as predicted by Corollary 6, its Lookahead dynamics LA-GD$_{\text{Alt}}$ with $k = 500$ and $\alpha = 0.25$ successfully converges to the Nash equilibrium. **(b)**. As predicted by Corollary 8, LA-EG$_{\text{Sim}}$ with $k = 500$ and $\alpha = 0.9$, denoted by LA-EG$^*_{\text{Sim}}$, accelerates its base dynamics EG$_{\text{Sim}}$. However, LA-EG$_{\text{Sim}}$ with $k = 500$ and $\alpha = 0.1$, denoted by LA-EG$^\dagger_{\text{Sim}}$, fails to accelerate its base dynamics and slows down the convergence.

(c)–(d) verify that the acceleration guarantees in Corollary 8 and Corollary C still hold for the game with larger conditioning. Again, we do not observe any non-necessity of our sufficient conditions in Figure E.8 (c)–(d).

To observe the actual improvements in convergence progress of each Lookahead dynamics, we choose k and α from Figure E.8 (a) and Figure E.8 (d), and measure the distance to the origin, which is the unique Nash equilibrium of the non-singular bilinear game. For LA-GD$_{\text{Alt}}$, we choose $k = 500$ and $\alpha = 0.25$, which are guaranteed by Corollary 6 to converge towards the Nash equilibrium. We test the same configuration for the negative momentum



(a) Unstable NS: IS=9.07.

(b) Unstable GP: IS=9.05.

(c) Stable NS: IS=8.925.

(d) Stable GP: IS=9.16.

Figure E.10: **Top 20 largest eigenvalues of GANs** trained with non-saturating (NS) Goodfellow et al. (2014) and WGAN-GP (GP) Gulrajani et al. (2017) loss functions on MNIST dataset. The black dots in each plot represent the eigenvalues at initialization. **Top**: Unstable points with competitive inception score (IS) 9.07 and 9.05. In (a)–(b), only a single pair of conjugate eigenvalues consists of the eigenvalues $\lambda_{\geq 1}(\nabla_{\mathbf{x}} F_{\mathrm{GD\,Sim}})$; hence, they have the imaginary conditioning 1, and therefore, both of them can be locally stabilized by Theorem 4. **Bottom:** Stable points with competitive IS=8.925 and IS=9.16. The largest eigenvalues $\lambda_{\max}(\nabla_{\mathbf{x}} F_{\mathrm{GD\,Sim}})$ are non-real; hence, they have the imaginary conditioning 1, and therefore, the local convergence towards both of them can be accelerated by Theorem 5.

method $\mathrm{GD}_{Alt}^{\mathrm{NM}}$ with momentum coefficient $m = -0.5$. Similarly, for $\text{LA-EG}_{\mathrm{Sim}}$, we choose $k = 500$ and $\alpha = 0.9$, which are predicted by Corollary 8 to accelerate the convergence. We also test $\text{LA-EG}_{\mathrm{Sim}}$ with $k = 500$ and $\alpha = 0.1$, which are predicted by Figure E.8 to slow down the convergence of its base dynamics $\mathrm{EG}_{\mathrm{Sim}}$.

We illustrate the results in Figure E.9. Figure E.9 (a) verifies the convergence guarantee of Corollary 6 and shows that $\text{LA-GD}_{\mathrm{Alt}}$ indeed converges towards the Nash equilibrium even for a game with larger conditioning. Furthermore, $\text{LA-GD}_{\mathrm{Alt}}^{\mathrm{NM}}$ successfully accelerates the negative momentum $\mathrm{GD}_{\mathrm{Alt}}^{\mathrm{NM}}$. The result in Figure E.9 (b) verifies the acceleration guarantee in Corollary 8. However, the result in Figure E.9 (b) suggests that badly configured Lookahead can slow down the convergence.

E.2 GANs

The local stabilization and acceleration guarantees given by Theorem 4–5 assume the eigenvalues $\lambda_{\geq 1}(\nabla_{\mathbf{x}} F(\mathbf{x}^*))$ and $\lambda_{\max}(\nabla_{\mathbf{x}} F(\mathbf{x}^*))$ to be non-reals with imaginary conditioning less than 3; otherwise, they lose the provable guarantee for the existence of a feasible $k \in (\beta_1, \beta_2)$ from Lemma 3. We verify whether such assumptions can be realistic in a practical nonlinear game like GANs. Specifically, we train GANs on MNIST dataset LeCun and Cortes (2010) with two different loss functions, namely non-saturating (NS) Goodfellow et al. (2014) and WGAN-GP (GP) Gulrajani et al. (2017) with Adam Kingma and Ba (2015). Then, we visualize the top 20 largest eigenvalues of $\nabla_{\mathbf{x}} F_{\mathrm{GD}_{\mathrm{Sim}}}$ at well-performing checkpoints, i.e., the weights of GANs where the generators achieve high inception scores (IS) Salimans et al. (2016). We use a small variant of DCGAN Radford et al. (2016) architecture with spectral normalization Miyato et al. (2018), and use the alternating updates for both NS and WGAN-GP loss functions. We perform 5 discriminator updates for each generator update in WGAN-GP experiments, and perform a single discriminator update for each iteration in NS experiments. We use a batch size of 100 and Adam hyperparameters $\beta_1 = 0.5$, $\beta_2 = 0.9$ with a fixed learning rate 0.0001. For each loss function, we perform 8 runs of training, and report top 20 eigenvalues of 2 representative points in Figure E.10.

The results in Figure E.10 (a)–(b) suggest that there are unstable, yet highly-performant points in GANs, and show that the assumptions in Theorem 4 can be realistic even for a practical nonlinear game like GANs. Specifically, the unstable points illustrated in Figure E.10 (a)–(b) have the eigenvalues with modulus greater than or equal to 1 $\lambda_{\geq 1}(\nabla_{\mathbf{x}} F_{\text{GDSim}})$ of imaginary conditioning 1; hence, Theorem 4 can stabilize such unstable points. This verifies that the eigenvalue assumptions in Theorem 4 is can be realistic even for a practical nonlinear game like GANs. On the other hand, the results in Figure E.10 (c)–(d) show that the well-performing stable points of GANs can exhibit non-real maximum eigenvalues. For such points, Theorem 5 can accelerate the local convergence. This verifies that the eigenvalue assumptions in Theorem 5 can be realistic in a practical nonlinear game as well.