# Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks

**Tengyuan Liang**
University of Chicago, Booth School of Business

**James Stokes**
University of Pennsylvania

## Abstract

Motivated by the pursuit of a systematic computational and algorithmic understanding of Generative Adversarial Networks (GANs), we present a simple yet unified non-asymptotic local convergence theory for smooth two-player games, which subsumes several discrete-time gradient-based saddle point dynamics. The analysis reveals the surprising nature of the off-diagonal interaction term as both a blessing and a curse. On the one hand, this interaction term explains the origin of the slow-down effect in the convergence of Simultaneous Gradient Ascent (SGA) to stable Nash equilibria. On the other hand, for the unstable equilibria, exponential convergence can be proved thanks to the interaction term, for four modified dynamics proposed to stabilize GAN training: Optimistic Mirror Descent (OMD), Consensus Optimization (CO), Implicit Updates (IU) and Predictive Method (PM). The analysis uncovers the intimate connections among these stabilizing techniques, and provides detailed characterization on the choice of learning rate. As a by-product, we present a new analysis for OMD proposed in Daskalakis, Ilyas, Syrgkanis, and Zeng [2017] with improved rates.

## 1 Introduction

In this paper we consider the non-asymptotic local convergence and stability of discrete-time gradient-based optimization algorithms for solving smooth two-player zero-sum games of the form,

$$\min_{\theta\in\mathbb{R}^p}\max_{\omega\in\mathbb{R}^q} U(\theta,\omega) \ . \tag{1}$$

The motivation behind our non-asymptotic analysis follows from the observation that Generative Adversarial Networks (GANs) lack principled understanding at both the computational and algorithmic level. GAN optimization is a special case of (1), which has been developed for learning a complex and multi-modal probability distribution based on samples from $\mathcal{P}_{\text{real}}$ (over $\mathcal{X}$), through learning a generator function $g_\theta(\cdot)$ that transforms the input distribution $\mathcal{P}_{\text{input}}$ (over $\mathcal{Z}$) to match the target $\mathcal{P}_{\text{real}}$. Ignoring the parameter regularization, the value function corresponding to a GAN is of the form,

$$U(\theta,\omega) = \mathbb{E}\, h_1(f_\omega(X)) - \mathbb{E}\, h_2(f_\omega(g_\theta(Z))) \ , \tag{2}$$

where $X\sim\mathcal{P}_{\text{real}}$, $Z\sim\mathcal{P}_{\text{input}}$ and $(\theta,\omega)\in\mathbb{R}^p\times\mathbb{R}^q$ parametrizes the generator function $g_\theta:\mathcal{Z}\to\mathcal{X}$ and discriminator function $f_\omega:\mathcal{X}\to\mathbb{R}$, respectively. The original GAN [Goodfellow et al., 2014], for example, corresponds to choosing $h_1(t)=\log\sigma(t)$, $h_2(t)=-\log(1-\sigma(t))$ where $\sigma$ is the sigmoid function; Wasserstein GAN [Arjovsky et al., 2017] considers $h_1(t)=h_2(t)=t$; $f$-GAN [Nowozin et al., 2016] proposes to use $h_1(t)=t, h_2(t)=f^*(t)$, where $f^*$ denotes the Fenchel dual of $f$. Recently, several attempts have been made to understand whether GANs learn the target distribution in the statistical sense [Liu et al., 2017, Arora and Zhang, 2017, Liang, 2017, 2018, Arora et al., 2017, Liu and Chaudhuri, 2018].

Optimization of GANs (and value functions of the form (1) at large) is hard, both in theory and in practice [Singh et al., 2000, Pfau and Vinyals, 2016, Salimans et al., 2016]. Global optimization of a general value function with multiple saddle points is impractical and unstable, so we instead resort to the more modest problem of searching for a *local saddle point* $(\theta_*,\omega_*)$ such that no player has the incentive to deviate locally

$$U(\theta_*,\omega_*)\leqslant U(\theta,\omega_*) \ , \quad \text{for } \theta \text{ in an open nbhd of } \theta_* \ ,$$
$$U(\theta_*,\omega_*)\geqslant U(\theta_*,\omega) \ , \quad \text{for } \omega \text{ in an open nbhd of } \omega_* \ .$$

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

For smooth value functions, the above conditions are equivalent to the following solution concept:

**Definition 1.1** (Local Nash Equilibrium)**.** $(\theta_*, \omega_*)$ *is called a local Nash equilibrium if*

1. $\nabla_\theta U(\theta_*, \omega_*) = \mathbf{0}$, $\nabla_\omega U(\theta_*, \omega_*) = \mathbf{0}$;
2. $\nabla_{\theta\theta} U(\theta_*, \omega_*) \succeq \mathbf{0}$, $-\nabla_{\omega\omega} U(\theta_*, \omega_*) \succeq \mathbf{0}$.

Here we use $\nabla_{\theta\omega} U(\theta, \omega) \in \mathbb{R}^{p\times q}$ to denote the off-diagonal term $\partial^2 U/\partial\theta\partial\omega$, and name it the *interaction term* throughout the paper. $\nabla_\theta U(\theta, \omega) \in \mathbb{R}^p$ denotes the gradient $\partial U/\partial\theta$, and $\nabla_{\theta\theta} U(\theta, \omega) \in \mathbb{R}^{p\times p}$ for the Hessian $\partial^2 U/\partial\theta\partial\theta$.

In practice, discrete-time dynamical systems are employed to numerically approach the saddle points of $U(\theta, \omega)$, as is the case in GANs [Goodfellow et al., 2014], and in primal-dual methods for non-linear optimization [Singh et al., 2000]. The simplest possibility is *Simultaneous Gradient Ascent* (SGA), which corresponds to the following discrete-time dynamical system,

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \nabla_\theta U(\theta_t, \omega_t) \ , \\ \omega_{t+1} &= \omega_t + \eta \nabla_\omega U(\theta_t, \omega_t) \ , \end{aligned} \tag{3}$$

where $\eta$ is the step size or learning rate. In the limit of vanishing step size, SGA approximates a continuous-time autonomous dynamical system, the asymptotic convergence of which has been established in Singh et al. [2000], Cherukuri et al. [2017], Nagarajan and Kolter [2017]. In practice, however, it has been widely reported that the discrete-time SGA dynamics for GAN optimization suffers from instabilities due to the possibility of complex eigenvalues in the operator of the dynamical system [Salimans et al., 2016, Metz et al., 2016, Nagarajan and Kolter, 2017, Mescheder et al., 2017, Heusel et al., 2017]. We believe room for improvement still exists in the current theory, which we hope will render it to be more informative in practice:

- **Non-asymptotic convergence speed.** In practice, one is concerned with finite step size $\eta > 0$ which is typically subject to extensive hyperparameter tuning. Detailed characterizations on the convergence speed, and theoretical insights on the choice of learning rate can be helpful.
- **Unified simple analysis for modified saddle point dynamics.** Several attempts to fix GAN optimization have been put forth by independent researchers, which modify the dynamics [Mescheder et al., 2017, Daskalakis et al., 2017, Yadav et al., 2017] using very different insights. A unified analysis that reviews the deeper connections amongst these proposals helps to better understand the saddle point dynamics at large.

In this paper, we address the above points by studying the theory of non-asymptotic convergence of SGA and related discrete-time saddle point dynamics, namely, *Optimistic Mirror Descent* (OMD), *Consensus Optimization* (CO), Implicit Updates (IU) and *Predictive Method* (PM). More concretely, we provide the following theoretical contributions about the crucial effect of the off-diagonal interaction term $\nabla_{\theta\omega} U(\theta, \omega)$ in two-player games:

- **Stable case: curse of the interaction term.** Locally, SGA converges *exponentially* fast to a stable Nash equilibrium with a carefully chosen learning rate. This can be viewed as a generalization (rather than a special case) of the local convergence guarantee for single-player gradient descent for strongly-convex functions. In addition, we quantitatively isolate the slow-down in the convergence rate of two-player SGA compared to single-player gradient descent, due to the presence of the off-diagonal interaction term $\nabla_{\theta\omega} U(\theta, \omega)$ for the two-player game.
- **Unstable case: blessing of the interaction term.** For unstable Nash equilibria, SGA *diverges* away for any non-zero learning rate. We discover a unified non-asymptotic analysis that encompasses four proposed modified dynamics — OMD, CO, IU and PM. The analysis shows that all these algorithms, at a high level, share the same idea of utilizing the curvature introduced by the interaction term $\nabla_{\theta\omega} U(\theta, \omega)$. Unlike the slow sub-linear rate of convergence experienced by single-player gradient descent for non-strongly convex functions[1], four modified dynamics effectively exploit the interaction term to achieve *exponential* convergence to unstable Nash equilibria. The analysis also provides specific advice on the choice of learning rate for each procedure, albeit restricted to the simple case of bi-linear games.

The organization of the paper is as follows. In Section 2 we consider the situation when locally, the value function satisfies strict convexity/concavity. We show non-asymptotic exponential convergence to Nash equilibria for SGA, and identify an optimal learning rate.

[1] In fact, Nesterov [2013] constructed a convex function that is non-strongly convex, such that all first order methods suffer slow sub-linear rate of convergence (in optimization literature, linear rate refers to exponential convergence speed).

To reveal and understand the new features of the modified dynamics, we study in Section 3 the minimal unstable bilinear game, showing that proposed stabilizing techniques all achieve exponential convergence to unstable Nash equilibria. Finally, in Section 4 we take a step closer to the real world by numerically evaluating each of the proposed dynamical systems using value functions of GAN form (2), under objective evaluation metrics. Proofs are deferred to the Appendix A.

## 2 Stable Case: Non-asymptotic Local Convergence

In this section we will establish the non-asymptotic convergence of SGA dynamics to saddle points that are *stable local Nash equilibria*. With a properly chosen learning rate, the local convergence can be intuitively pictured as cycling inwards to these saddle points, where the distance to the saddle point of interest is exponentially contracting. First, let's introduce the notion of stable equilibrium.

**Definition 2.1** (Stable Local Nash Equilibrium)**.** *$(\theta_*, \omega_*)$ is called a stable local Nash equilibrium if*

1. $\nabla_\theta U(\theta_*, \omega_*) = \mathbf{0}$, $\nabla_\omega U(\theta_*, \omega_*) = \mathbf{0}$;
2. $\nabla_{\theta\theta} U(\theta_*, \omega_*) \succ \mathbf{0}$, $-\nabla_{\omega\omega} U(\theta_*, \omega_*) \succ \mathbf{0}$.

The above notion of stability is stronger than the Definition 1.1, in the sense that $\nabla_{\theta\theta} U(\theta_*, \omega_*)$ and $-\nabla_{\omega\omega} U(\theta_*, \omega_*)$ have smallest eigenvalues bounded away from 0.

**Assumption 2.1** (Local Strong Convexity-Concavity)**.** *Consider $U(\theta, \omega) : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ that is smooth and twice differentiable, and let $(\theta_*, \omega_*)$ be a stable local Nash equilibrium as in Definition 2.1. Assume that for some $r > 0$, there exists an open neighborhood near $(\theta_*, \omega_*)$ such that for all $(\theta, \omega) \in B_2((\theta_*, \omega_*), r)$, the following strong convexity-concavity condition holds,*

$$\nabla_{\theta\theta} U(\theta, \omega) \succ \mathbf{0}, \quad -\nabla_{\omega\omega} U(\theta, \omega) \succ \mathbf{0} \ .$$

It will prove convenient to introduce some notation before introducing the main theorem. Let us define the following block-wise abbreviation for the matrix of second derivatives,

$$\begin{bmatrix} \nabla_{\theta\theta} U(\theta, \omega) & \nabla_{\theta\omega} U(\theta, \omega) \\ -\nabla_{\omega\theta} U(\theta, \omega) & -\nabla_{\omega\omega} U(\theta, \omega) \end{bmatrix} := \begin{bmatrix} A_{\theta,\omega} & C_{\theta,\omega} \\ -C_{\theta,\omega}^T & B_{\theta,\omega} \end{bmatrix} , \tag{4}$$

and define $\alpha, \beta$ as

$$\begin{aligned} \alpha &:= \min_{(\theta,\omega) \in B_2((\theta_*,\omega_*),r)} \lambda_{\min}\left(\mathrm{diag}(A_{\theta,\omega}^2, B_{\theta,\omega}^2)\right) \ , \\ \beta &:= \max_{(\theta,\omega) \in B_2((\theta_*,\omega_*),r)} \lambda_{\max}\left(F_{\theta,\omega}\right) \ , \end{aligned} \tag{5}$$

$$F_{\theta,\omega} := \begin{bmatrix} A_{\theta,\omega}^2 + C_{\theta,\omega} C_{\theta,\omega}^T & -A_{\theta,\omega} C_{\theta,\omega} + C_{\theta,\omega} B_{\theta,\omega} \\ -C_{\theta,\omega}^T A_{\theta,\omega} + B_{\theta,\omega} C_{\theta,\omega}^T & B_{\theta,\omega}^2 + C_{\theta,\omega}^T C_{\theta,\omega} \end{bmatrix}$$

where $\lambda_{\max}(M), \lambda_{\min}(M)$ denote the largest and smallest eigenvalue of matrix $M$.

**Theorem 1** (Exponential Convergence: SGA)**.** *Consider $U(\theta, \omega) : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ that satisfies Assumption 2.1 for some radius $r > 0$ near a stable local Nash equilibrium $(\theta_*, \omega_*)$ as in Definition 2.1. Suppose the initialization satisfies $(\theta_0, \omega_0) \in B_2((\theta_*, \omega_*), r)$. Then the SGA dynamics* (3) *with fixed learning rate*

$$\eta = \sqrt{\alpha}/\beta \ ,$$

*($\alpha, \beta$ defined in Eqn.* (5)*) obtains an $\epsilon$-minimizer such that $(\theta_T, \omega_T) \in B_2((\theta_*, \omega_*), \epsilon)$, as long as*

$$T \geqslant T_{\mathrm{SGA}} := \left\lceil 2\frac{\beta}{\alpha} \log \frac{r}{\epsilon} \right\rceil \ .$$

**Remark 1.** It is interesting to compare the convergence speed of the saddle point dynamics to conventional gradient descent in one variable, for a strongly-convex function. We remind the reader that to obtain an $\epsilon$-minimizer for a strongly-convex function, one needs the following number of iterations of gradient descent,

$$T_{\mathrm{GD}} := \max\left\{ \frac{\lambda_{\max}\left(A_{\theta,\omega}\right)}{\lambda_{\min}\left(A_{\theta,\omega}\right)} \log \frac{r}{\epsilon}, \quad \frac{\lambda_{\max}\left(B_{\theta,\omega}\right)}{\lambda_{\min}\left(B_{\theta,\omega}\right)} \log \frac{r}{\epsilon} \right\} \ ,$$

depending on whether we are optimizing with respect to $\theta$ or $\omega$, respectively. It is now evident that due to the presence of $C_{\theta,\omega} C_{\theta,\omega}^T$, the convergence of two-player SGA to a saddle-point can be significantly slower than convergence of single-player gradient descent. In particular, applying the eigenvalue interlacing theorem to the principal submatrix $A_{\theta,\omega}^2 + C_{\theta,\omega} C_{\theta,\omega}^T$ of $F_{\theta,\omega}$ we obtain

$$\begin{aligned} \lambda_{\max}\left(F_{\theta,\omega}\right) &\geqslant \lambda_{\max}\left(A_{\theta,\omega}^2 + C_{\theta,\omega} C_{\theta,\omega}^T\right) \ , \\ &\geqslant \lambda_{\max}\left(A_{\theta,\omega}^2\right) \ . \end{aligned}$$

Therefore the convergence of SGA is slower than that in the conventional GD[2]

$$T_{\mathrm{SGA}} \geqslant T_{\mathrm{GD}} \ .$$

We would like to emphasize that for the saddle point convergence, the slow-down effect of the interaction term $C_{\theta,\omega}$ is explicit in our non-asymptotic analysis.

[2]Recall that $\frac{\lambda_{\max}(A_{\theta,\omega}^2)}{\lambda_{\min}(A_{\theta,\omega}^2)} \geqslant \frac{\lambda_{\max}(A_{\theta,\omega})}{\lambda_{\min}(A_{\theta,\omega})}$.

The intuition that the discrete-time SGA dynamics cycles inward to a stable Nash equilibrium exponentially fast can be seen in the following way. The presence of the off-diagonal anti-symmetric component in Eqn. (4) means that the associated linear operator of the discrete-time dynamics has complex eigenvalues, which results in periodic cycling behavior. However, due to the explicit choice of $\eta$, the distance to stable Nash equilibrium is shrinking exponentially fast. The local exponential stability in the infinitesimal/asymptotic case when $\eta \to 0$ has already been studied in a paper Nagarajan and Kolter [2017] (Theorem 3.1 therein) by showing the Jacobian matrix of a particular form of GAN objective is Hurwitz (has all strictly negative eigenvalues). There are two distinct differences in our result: (1) we provide non-asymptotic convergence, with specific guidance on the choice of learning rate $\eta$; (2) our analysis goes through analyzing the singular values (which is rather different from the modulus of eigenvalue for a general matrix), instead of involving the complex eigenvalues, and this simple technique generalizes to four other modified saddle point dynamics which we discuss in the next section.

In fact, one can show that the slow-down effect of the interaction term $C_{\theta,\omega}$ for SGA in the above theorem is indeed necessary.

**Corollary 1** (Simple Lower Bound for SGA)**.** *Consider $U(\theta,\omega) = \frac{1}{2}\theta^T\theta - \frac{1}{2}\omega^T\omega + \theta^T C\omega$ with $p = q$ and $C \in \mathbb{R}^{p\times q}$ full rank. Then the SGA dynamics* (3) *with any fixed learning rate $\eta$ satisfies*

$$\|\theta_{t+1}\|^2 + \|\omega_{t+1}\|^2 \geqslant \frac{\lambda_{\min}(C^TC)}{1+\lambda_{\min}(C^TC)}\left(\|(\theta_t\|^2 + \|\omega_t\|^2\right).$$

The above corollary shows that for any stepsize $\eta$, to obtain $\epsilon$-solution, the number of SGA iteration is at least $\Omega((1+\lambda_{\min}(C^TC))\log(1/\epsilon))$. Namely, the interaction term is indeed a curse to the convergence rate.

## 3 Unstable Case: Local Bi-Linear Problem

Oscillation and instability for SGA occurs when the problem is non-strongly convex-concave, as in the bi-linear game (or more precisely, at least linear in one player). This observation was first pointed out using a very simple linear game $U(x,y) = xy$ in Salimans et al. [2016]. More generally, as a result of Theorem 1, this phenomenon occurs when the local Nash equilibrium is non-stable,

$$\begin{aligned}&\operatorname{diag}(A_{\theta_*,\omega_*}, B_{\theta_*,\omega_*}) \approx \mathbf{0} \iff \\ &\begin{bmatrix} A_{\theta_*,\omega_*} & C_{\theta_*,\omega_*} \\ -C^T_{\theta_*,\omega_*} & B_{\theta_*,\omega_*}\end{bmatrix} \approx \begin{bmatrix} \mathbf{0} & C_{\theta_*,\omega_*} \\ -C^T_{\theta_*,\omega_*} & \mathbf{0}\end{bmatrix}.\end{aligned}$$

Let's consider an extreme case when $A_{\theta_*,\omega_*} = \mathbf{0}$ and $B_{\theta_*,\omega_*} = \mathbf{0}$. In this case, we will use a novel unified non-asymptotic analysis to show that the following proposed dynamics can fix the oscillation problem and provide exponential convergence to unstable Nash equilibria:

(1) *Optimistic Mirror Descent* (OMD) in Daskalakis et al. [2017]

(2) A modified version of *Predictive Methods* (PM) motivated from Yadav et al. [2017]

(3) *Implicit Updates*

(4) *Consensus Optimization* (CO) introduced in Mescheder et al. [2017]

Our analysis shows that these stabilizing techniques, at a high level, all manipulate the dynamics to utilize the curvature generated by the interaction term $C_{\theta_*,\omega_*}C^T_{\theta_*,\omega_*}$ — which we refer to as the "blessing" of the interaction term, to contrast with the "slow-down effect" of the interaction term in the strongly convex-concave case (Theorem 1). Once again, as alluded to in the introduction, this fast linear-rate convergence result in the non-strongly convex-concave two-player game should be contrasted with the significantly slower sub-linear convergence rate for all first-order-methods in convex but non-strongly convex single-player optimization. The latter was proved by a lower bound argument in [Nesterov, 2013, Theorem 2.1.7]. The main result proved in this section is informally stated

**Theorem 2** (Informal: Unstable Case)**.** *All these four modified dynamics, in the bi-linear game, enjoy the last iterate exponential convergence guarantee.*

The bilinear game can be motivated by considering the Taylor expansion of a general smooth two-player game around a non-stable Nash equilibrium ($A, B \approx \mathbf{0}$), assuming that $(\theta_*,\omega_*) = (\mathbf{0},\mathbf{0})$. Now consider the simple bi-linear game $U(\theta,\omega) = \theta^T C\omega$. With the SGA dynamics defined in (3), one can easily verify that

$$\begin{aligned}\|\theta_{t+1}\|^2 &\geqslant \left(1+\eta^2\lambda_{\min}(CC^T)\right)\|\theta_t\|^2\ , \\ \|\omega_{t+1}\|^2 &\geqslant \left(1+\eta^2\lambda_{\min}(C^TC)\right)\|\omega_t\|^2\ .\end{aligned}$$

Therefore, the continuous limit $\eta \to 0$ is cycling around a sphere, while with any practical learning rate $\eta \neq 0$, the distance to the Nash equilibrium can be increasing exponentially instead of converging. Per Theorem 1 and the discussion above, instability for SGA only occurs when the local game is approximately bilinear. From now on, therefore, we will focus on the simplest unstable form of the game, the bi-linear game, to isolate the main idea behind fixing the instability problem. The proof technique can be extended to more general settings, with a sacrifice of simplicity.

### 3.1 (Improved) Optimistic Mirror Descent

Daskalakis, Ilyas, Syrgkanis, and Zeng [2017] employed Optimistic Mirror Descent (OMD) [Rakhlin and Sridharan, 2013] motivated by online learning to solve the instability problem in GANs. Here we provide a stronger result, showing that the last iterate of OMD enjoys *exponential convergence* for bi-linear games. We note that although the last-iterate convergence of this OMD procedure was already rigorously proved in Daskalakis et al. [2017], the exponential convergence is not known to the best of our knowledge.

**Theorem 3** (Exponential Convergence: OMD)**.** *Consider a bi-linear game $U(\theta,\omega)=\theta^T C\omega$. Assume $p=q$ and $C$ is full rank. Then the OMD dynamics,*

$$\begin{aligned}\theta_{t+1} &= \theta_t - 2\eta\nabla_\theta U(\theta_t,\omega_t) + \eta\nabla_\theta U(\theta_{t-1},\omega_{t-1}) \ , \\ \omega_{t+1} &= \omega_t + 2\eta\nabla_\omega U(\theta_t,\omega_t) - \eta\nabla_\omega U(\theta_{t-1},\omega_{t-1}) \ ,\end{aligned} \tag{6}$$

*with the learning rate*

$$\eta = \frac{1}{2\sqrt{2\lambda_{\max}(CC^T)}} \ ,$$

*obtains an $\epsilon$-minimizer such that $(\theta_T,\omega_T)\in B_2(\epsilon)$, provided*

$$T \geqslant T_{\mathrm{OMD}} := \left\lceil 16\frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)}\log\frac{4\sqrt{2}r}{\epsilon}\right\rceil \ ,$$

*under the assumption that $\|(\theta_0,\omega_0)\|, \|(\theta_1,\omega_1)\| \leqslant r$.*

Let us compare our result with the last-iterate convergence result in Daskalakis et al. [2017]. Roughly speaking, [Daskalakis et al., 2017, Theorem 1] asserts that to obtain an $\epsilon$-minimizer, one requires a learning rate scaling as $\eta(\epsilon)\asymp\epsilon^2$ and a number of iterations bounded by

$$T \gtrsim \epsilon^{-4}\log\frac{1}{\epsilon}\cdot \mathrm{Poly}\left(\frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)}\right) \ .$$

In contrast, we show that with step size $\eta$ chosen independently of $\epsilon$, the last iterate of OMD falls within $\epsilon$ of the saddle point after a number of iterations given by

$$T \gtrsim \log\frac{1}{\epsilon}\cdot\frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)} \ .$$

In other words, we improved the dependence on $\epsilon$ from polynomial to logarithmic. This improved analysis also coincides with the exponential convergence found in simulations.

### 3.2 (Modified) Predictive Methods

From a very different motivation in ordinary differential equations, Yadav et al. [2017] proposed Predictive Methods (PM) to fix the instability problem. The intuition is to evaluate the gradient at a predictive future location and then perform the update. In this section, we propose and analyze a modified version of the predictive method (for simultaneous gradient updates), inspired by Yadav et al. [2017].

Consider the following modified PM dynamics,

$$\begin{aligned}\text{predictive step:}\quad \theta_{t+1/2} &= \theta_t - \gamma\nabla_\theta U(\theta_t,\omega_t) \ , \\ \omega_{t+1/2} &= \omega_t + \gamma\nabla_\omega U(\theta_t,\omega_t) \ ; \\ \text{gradient step:}\quad \theta_{t+1} &= \theta_t - \eta\nabla_\theta U(\theta_{t+1/2},\omega_{t+1/2}) \ , \\ \omega_{t+1} &= \omega_t + \eta\nabla_\omega U(\theta_{t+1/2},\omega_{t+1/2}) \ .\end{aligned} \tag{7}$$

**Theorem 4** (Exponential Convergence: PM)**.** *Consider a bi-linear game $U(\theta,\omega)=\theta^T C\omega$. Assume $p=q$ and $C$ is full rank. Fix some $\gamma>0$. Then the PM dynamics in Eqn. (7) with learning rate*

$$\eta = \frac{\gamma\lambda_{\min}(CC^T)}{\lambda_{\max}(CC^T)+\gamma^2\lambda^2_{\max}(CC^T)} \ ,$$

*obtains an $\epsilon$-minimizer such that $(\theta_T,\omega_T)\in B_2(\epsilon)$, provided*

$$T \geqslant T_{\mathrm{PM}} := \left\lceil 2\frac{\gamma^2\lambda^2_{\max}(CC^T)+\lambda_{\max}(CC^T)}{\gamma^2\lambda^2_{\min}(CC^T)}\log\frac{r}{\epsilon}\right\rceil \ ,$$

*under the assumption that $\|(\theta_0,\omega_0)\|\leqslant r$.*

### 3.3 Implicit Updates

Implicit Update (IU) rules have been shown to be more robust compared to explicit updates, and typically match the performance or even outperform the latter empirically in online learning [Kulis and Bartlett, 2010]. We will show that a simple adaptation of implicit updates for simultaneous gradient ascent/descent resolves the instability problem in the bi-linear case.

**Theorem 5** (Exponential Convergence: IU)**.** *Consider a bi-linear game $U(\theta,\omega)=\theta^T C\omega$. Assume $p=q$ and $C$ is full rank. Then the implicit updates*

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta\nabla_\theta U(\theta_{t+1},\omega_{t+1}) \ , \\ \omega_{t+1} &= \omega_t + \eta\nabla_\omega U(\theta_{t+1},\omega_{t+1}) \ ,\end{aligned}$$

*with the learning rate*

$$\eta = \frac{1}{\sqrt{\lambda_{\max}(CC^T)}}$$

*obtains an $\epsilon$-minimizer such that $(\theta_T,\omega_T)\in B_2(\epsilon)$, provided*

$$T \geqslant T_{\mathrm{IU}} := \left\lceil (2+\sqrt{2})\frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)}\log\frac{r}{\epsilon}\right\rceil$$

*under the assumption that $\|(\theta_0,\omega_0)\|\leqslant r$.*

### 3.4 Consensus Optimization

Consensus Optimization (CO) is another elegant attempt to fix the aforementioned problem, proposed in Mescheder et al. [2017]. The idea is to add a potential component to the pure-curl vector field associated with SGA in the bi-linear game, in order to attract the dynamics to the critical points. [Mescheder et al., 2017, Nagarajan and Kolter, 2017] analyzed the infinitesimal flow version of the consensus optimization, and intuitively showed that it pushes the real part of the eigenvalue away from 1, to ensure asymptotic convergence. In this section, we provide a simple convergence analysis of the discretized dynamics, of the same flavor as the previous section. An upshot of the analysis is that it sheds light on possible choices of learning rate.

Recall that the regularization term defining consensus optimization is given by,

$$R(\theta,\omega) = \frac{1}{2}\left(\|\nabla_\theta U(\theta,\omega)\|^2 + \|\nabla_\omega U(\theta,\omega)\|^2\right) \quad . \tag{8}$$

Surprisingly, we find that the consensus optimization coincides with the modified predictive method for the bi-linear game, as described by the following

**Theorem 6** (Exponential Convergence: CO)**.** *Consider a bi-linear game $U(\theta,\omega) = \theta^T C\omega$. Assume $p = q$ and $C$ is full rank. Recall $R(\theta,\omega)$ defined in Eqn. (8), and fix some $\gamma > 0$. Then the CO dynamics with the same learning rate $\eta$ as in Thm. 4,*

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta\left[\nabla_\theta U(\theta_t,\omega_t) + \gamma\nabla_\theta R(\theta_t,\omega_t)\right] \quad ,\\ \omega_{t+1} &= \omega_t + \eta\left[\nabla_\omega U(\theta_t,\omega_t) - \gamma\nabla_\omega R(\theta_t,\omega_t)\right] \quad ,\end{aligned} \tag{9}$$

*converges exponentially fast in the same way as the PM dynamics in Thm. 4.*

## 4 Experiments

In the simplistic setting of bilinear games we have seen that exponential convergence can be achieved for appropriate choice of learning rate and this is indeed confirmed by numerical experiments as shown in Fig. 1. In reality, however, the assumption of bilinearity is not applicable to value functions of GAN form and indeed recent large-scale studies of GAN optimization [Lucic et al., 2017] suggest that improvements from algorithmic changes mostly disappeared after taking into account hyper-parameter tuning and randomness of initialization. They conclude that "future GAN research should be based on more systematic and objective evaluation procedures." Inspired by this conclusion, we conduct a systematic evaluation of the proposed optimization algorithms on two basic density learning problems, and introduce corresponding objective evaluation metrics. The goal of this analysis

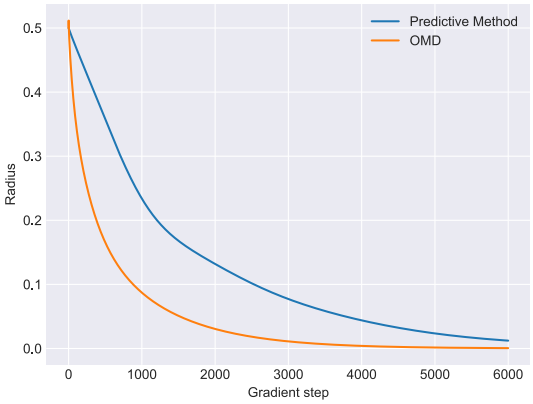


Figure 1: Distance to Nash equilibrium as a function of gradient iteration for the bilinear game assuming $p = q = 5$, $\gamma = 1$ and $r = 0.5$. The components of the interaction matrix $C$ were chosen i.i.d. uniform on $[0, 1]$.

is not to achieve state-of-art performance, but rather to compare and contrast the existing proposals in a carefully controlled learning environment. We focus on the Wasserstein GAN formulation so that the value function is given by

$$U(\theta,\omega) = \mathop{\mathbb{E}}_{X\sim\mathcal{P}_{\text{real}}} f_\omega(X) - \mathop{\mathbb{E}}_{Z\sim N(0,I_k)} f_\omega(g_\theta(Z)) \quad ,$$

where $f_\omega : \mathbb{R}^d \to \mathbb{R}$ is a multi-layer neural network with $L$ hidden layers and rectifier non-linearities and the input distribution $\mathcal{P}_{\text{input}}$ was chosen to be $k$-dimensional standard Gaussian noise. Following Gulrajani et al. [2017], we impose the Lipschitz-1 constraint on the discriminator network using the two-sided gradient penalty term $\Lambda(\omega)$ introduced in [Gulrajani et al., 2017, Eqn. (3)]. The consensus optimization loss is defined with respect to the value function as in (8) without including gradient penalty. The combined loss function of the discriminator and generator are respectively,

$$\begin{aligned}L_{\text{dis}}(\theta,\omega) &= -U(\theta,\omega) + \gamma\, R(\theta,\omega) + \lambda\,\Lambda(\omega) \quad ,\\ L_{\text{gen}}(\theta,\omega) &= U(\theta,\omega) + \gamma\, R(\theta,\omega) \quad .\end{aligned}$$

The coefficients of the gradient penalty and consensus optimization terms were determined by a coarse parameter search and then locked to $\lambda = \gamma = 1$ throughout. In order to make close contact with our theoretical formalism, we optimize the above loss functions using simultaneous gradient updates with fixed learning rate of $\eta = 10^{-3}$.

### 4.1 Learning Covariance of Multivariate Gaussian

Consider the problem of learning the covariance matrix $\Sigma \in S^d_{++}$ of a $d$-dimensional multivariate Gaussian distribution $\mathcal{P}_{\text{real}} = N(0, \Sigma)$ with non-degenerate covariance $\Sigma \succ 0$. Note that the learning problem is well-specified if we choose the generator function $g_\theta : \mathbb{R}^k \to \mathbb{R}^d$ to be a simple linear transformation of the $k$-dimensional latent space ($k \geqslant d$). Although the GAN approach is clearly overkill for this simple density estimation problem, we find this example illuminating because it affords some analytical tractability for the otherwise intractable general GAN value function (2). Specifically, if we choose the discriminator function $f_\omega : \mathbb{R}^d \to \mathbb{R}$ to be a neural network with $L = 1$ hidden layer consisting of $H$ hidden units with rectifier nonlinearities, and set biases to zero, then the explicit functional forms of discriminator and generator are respectively,

$$f_\omega(x) = \sum_{i=1}^{H} v_i \langle w_i, x \rangle \mathbf{1}_{\{\langle w_i, x \rangle \geqslant 0\}} \ , \qquad g_\theta(z) = Vz \ ,$$

where $\omega \in \{w_i \in \mathbb{R}^d, v_i \in \mathbb{R} : \forall i \in [H]\}$ and $\theta \in \{V \in \mathbb{R}^{d \times k}\}$ are the discriminator and generator parameters, respectively. If, moreover, we express the covariance matrix as $\Sigma = AA^T$, then the value function can be expressed in closed form as,

$$U(\theta, \omega) = \text{const} \times \sum_{i=1}^{H} v_i \big[ \|A^T w_i\| - \|V^T w_i\| \big] \ . \quad (10)$$

The above analytical form of the value function sheds some light on the nature of the local Nash equilibrium solution concept. In particular, if one solves for the condition of being a Nash equilibrium, one does not conclude that $V_* V_*^T = \Sigma$. The result depends on the rank of the matrix $[w_1, \ldots, w_H]$.

The evaluation of different optimization algorithms involved comparing the target density $\mathcal{P}_{\text{real}} = N(0, \Sigma)$ and the analytical generator density $\mathcal{P}_{\text{fake}} = N(0, VV^T)$ after $t = 10^5$ training iterations (Fig. 2). For simplicity, we chose the evaluation metric to be the Frobenius norm of the difference between the covariance matrices $\|\Sigma - VV^T\|_{\text{F}}$. The covariance learning experiments were conducted in the well-specified and over-parametrized regime ($k = 16$, $d = 2$) using $H = 128$ hidden units for the discriminator network.

### 4.2 Mixture of Gaussians

In practical applications, GANs are typically trained using the empirical distribution of the samples, where the samples are drawn from an idealized multi-modal probability distribution. To capture the notion of a multi-modal data distribution, we focus on a mixture of 8 Gaussians with means located at the vertices of a regular octagon inscribed in the unit circle, where each component has a fixed diagonal covariance of width $\sigma = 0.03$. In contrast to previous visual-based evaluations, we estimate the Wasserstein-1 distance $W_1(\mathcal{P}_{\text{real}}, \mathcal{P}_{\text{fake}})$ between the target density $\mathcal{P}_{\text{real}}$ and the distribution $\mathcal{P}_{\text{fake}}$ of the random variable $g_\theta(Z)$ implied by the trained generator network. The estimate is obtained by solving the linear program which computes the earth mover's distance between the sample estimates $\widehat{\mathcal{P}}_{\text{real}} = \frac{1}{m} \sum_{i=1}^m \delta_{X_i}$ and $\widehat{\mathcal{P}}_{\text{fake}} = \frac{1}{m} \sum_{i=1}^m \delta_{g_\theta(Z_i)}$, respectively, and approaches the population version $W_1(\mathcal{P}_{\text{real}}, \mathcal{P}_{\text{fake}})$ as the number of samples $m \to \infty$.

The experiments with the mixture of Gaussians used 2 dimensional Gaussian as input ($k = 2$). Both the generator and discriminator networks consisted of $L = 4$ hidden layers with $H = 128$ units per hidden layer. The estimate of the Wasserstein-1 distance was calculated using a sample size of $m = 512$ after training for $t = 5 \cdot 10^4$ iterations. It is clear from Fig. 3 that the Wasserstein-1 distance correlates closely with the visual fit to the target distribution. The empirical evaluation (Fig. 2) shows that the separation between consensus optimization and competing algorithms disappears on the mixture distribution, suggesting that the qualitative ranking is not robust to the choice of loss landscape. These findings demand deeper understanding of the global structure of the landscape, including the formulation of regularization to tame the notoriously difficult GAN optimization [Arbel et al., 2018], which is not captured by our local stability analysis.

## 5 Conclusions and Future Work

In this paper we made a first step towards understanding the local convergence rate of the discrete-time gradient-based saddle point dynamics for solving smooth two-player zero-sum games, including GANs as the leading motivation. The focus of the paper is on illustrating how local geometry affects the convergence speed and choice of learning-rate for both stable and unstable local Nash Equilibria. A curious fact we proved is that modified first-order dynamics such as OMD converge with linear-rate to unstable local Nash Equilibria, as a consequence of the interaction term between the two players.

We acknowledge that there are still critical steps left open by our analysis in order to understand the effectiveness of heuristic methods for training GANs for distribution learning. Solving this problem requires an understanding of the ability of various stable/unstable local Nash Equilibria to represent distributions in the

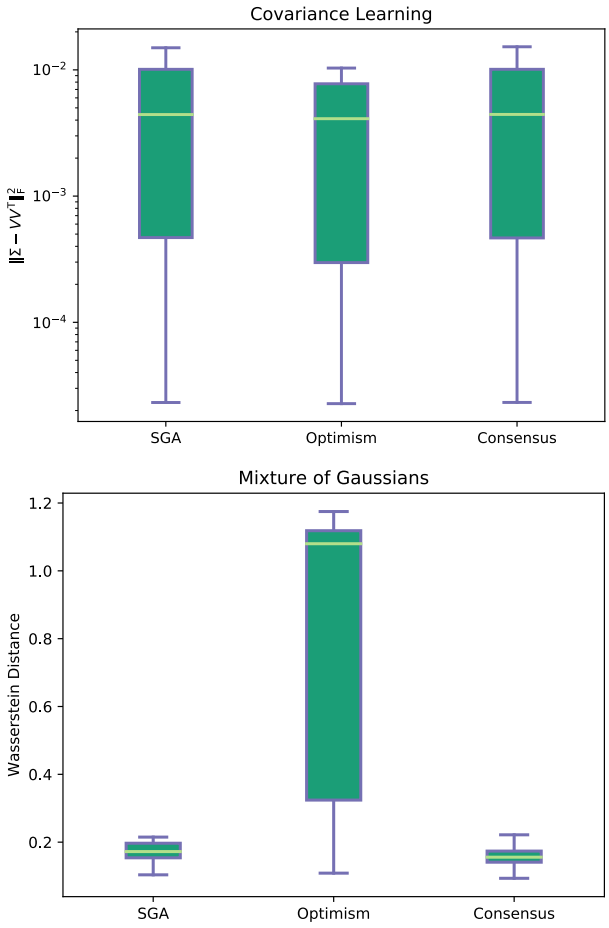


Figure 2: Evaluation metrics for covariance learning (top) and mixture of Gaussians learning (bottom) using different dynamical systems after $t = 10^5$ and $t = 5 \cdot 10^4$ training iterations, respectively and 16 random seeds. Note that for covariance learning, we use the log-scale on $y$-axis.

statistical sense. To the best of our knowledge, it remains unclear whether convergence to the stable local solution concept (Definition 2.1) is better than converging/oscillating/escaping an unstable local solution, in terms of distribution learning. Recent progress by Daskalakis and Panageas [2018b,a] employs the stable manifold Theorem [Galor, 2007] to show that certain dynamics avoid unstable local solutions (barring initialization in a set of Lebesgue measure zero). Overall, a satisfactory theory — in both the computational and statistical sense — for answering how heuristic gradient-based saddle point dynamics for GANs are able to learn distributions is still wide open for future investigation.

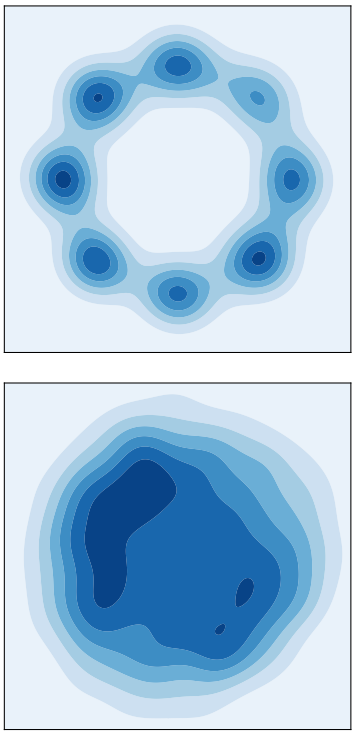

Figure 3: Density plots of best and worst generator distribution measured by empirical Wasserstein-1 distance from the target distribution, across all baselines amongst 16 random seeds (excluding non-convergent runs) after training for $5 \times 10^4$ iterations. Top: Consensus ($W_1 = 0.093$). Bottom: OMD ($W_1 = 0.367$).

## References

Michael Arbel, Dougal J Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for mmd gans. *arXiv preprint arXiv:1805.11565*, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Theoretical limitations of encoder-decoder gan architectures. *arXiv preprint arXiv:1711.02651*, 2017.

Ashish Cherukuri, Bahman Gharesifard, and Jorge Cortes. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1):486–511, 2017.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and con-

strained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018a.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *arXiv preprint arXiv:1807.03907*, 2018b.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Oded Galor. *Discrete dynamical systems*. Springer Science & Business Media, 2007.

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

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6629–6640, 2017.

Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 575–582, 2010.

Tengyuan Liang. How well can generative adversarial networks learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018.

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. *arXiv preprint arXiv:1705.08991*, 2017.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. *arXiv preprint arXiv:1705.10461*, 2017.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In *Advances in Neural Information Processing Systems*, pages 5591–5600, 2017.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

Satinder Singh, Michael Kearns, and Yishay Mansour. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 541–548. Morgan Kaufmann Publishers Inc., 2000.

Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. *Accepted at ICLR 2018*, 2017.