Limiting Behaviors of Nonconvex-Nonconcave Minimax Optimization via Continuous-Time Systems

Benjamin Grimmer GRIMMER@JHU.EDU
Johns Hopkins University, Baltimore MD

Haihao Lu HAIHAO.LU@CHICAGOBOOTH.EDU
University of Chicago, Chicago IL

Pratik Worah PWORAH@GOOGLE.COM
Google Research, New York NY

Vahab Mirrokni MIRROKNI@GOOGLE.COM
Google Research, New York NY

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

Unlike nonconvex optimization, where gradient descent is guaranteed to converge to a local optimizer, algorithms for nonconvex-nonconcave minimax optimization can have topologically different solution paths: sometimes converging to a solution, sometimes never converging and instead following a limit cycle, and sometimes diverging. In this paper, we study the limiting behaviors of three classic minimax algorithms: gradient descent ascent (GDA), alternating gradient descent ascent (AGDA), and the extragradient method (EGM). Numerically, we observe that all of these limiting behaviors can arise in Generative Adversarial Networks (GAN) training and are easily demonstrated even in simple GAN models. To explain these different behaviors, we study the high-order resolution continuous-time dynamics that correspond to each algorithm, which results in sufficient (and almost necessary) conditions for the local convergence by each method. Moreover, this ODE perspective allows us to characterize the phase transition between these potentially nonconvergent limiting behaviors caused by introducing regularization in the problem instance.

Keywords: Minimax Optimization, ODE Dynamics, First-Order Methods, Cycling, Divergence, Convergence

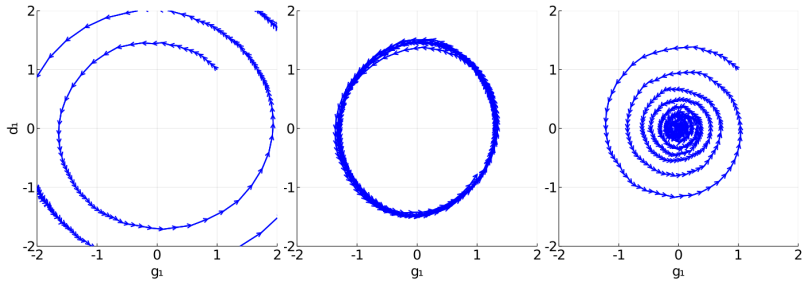
1. Introduction

In this paper, we are interested in the limiting behavior of optimizing nonconvex-nonconcave problems

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

for any differentiable objective function $L(x,y)$. Minimax optimization has found wide usage in robust optimization (Verdu and Poor, 1984; Ben-Tal et al., 2009; Bertsimas et al., 2011) and many machine learning tasks. One notable application is in GAN training (Goodfellow et al., 2014) where a generator G tries to produce new data samples from a latent distribution and a discriminator D tries to distinguish these from true data, defined by the following minimax problem

$$\min_G \max_D \mathbb{E}_{s\sim p_{data}}\left[\log D(s)\right] + \mathbb{E}_{e\sim p_{latent}}\left[\log(1-D(G(e)))\right]. \tag{2}$$



(a) GDA (b) AGDA (c) EGM

Figure 1: Sample trajectories of GDA, AGDA, and EGM with batch size 10 applied to the GAN (2) showing the first coordinates (g_1, d_1) of g and d which diverge, cycle, and converge, respectively.

We consider first-order methods for solving (1) given $F(x, y) = (\nabla_x L(x, y), -\nabla_y L(x, y))$ as a gradient oracle or an unbiased estimator of these gradients. Given such an oracle, three of the most classic first-order minimax optimization methods can be defined as follows, producing a sequence of solution pairs $z_k = (x_k, y_k)$: Gradient Descent Ascent (GDA)

$$z_{k+1} = z_k - sF(z_k), \tag{3}$$

Alternating Gradient Descent Ascent (AGDA)

$$\begin{aligned} x_{k+1} &= x_k - sF_x(x_k, y_k) \\ y_{k+1} &= y_k - sF_y(x_{k+1}, y_k), \end{aligned} \tag{4}$$

and the Extragradient Method (EGM)

$$\begin{aligned} z_{k+1/2} &= z_k - sF(z_k) \\ z_{k+1} &= z_k - sF(z_{k+1/2}). \end{aligned} \tag{5}$$

Stochastic versions of these algorithms all follow by replacing $F(x, y)$ with an unbiased estimator. In the case of GAN training, an unbiased gradient estimate is given by using a finite batch of samples from p_{data} and p_{latent} to approximate the expectations in (2).

We begin by considering the limiting behavior of these three methods on a GAN training example over CIFAR-10 data (with full experiment details given in Section 2). We take simple models for the discriminator controlling weights $d \in \mathbb{R}^n$ in the logistic function

$$D(s) = \frac{1}{1 + e^{d^T(s - \bar{s})}} \tag{6}$$

where $\bar{s} = \mathbb{E}_{s \sim p_{data}}[s]$ is the average image over the dataset and the generator controlling a translation of a normal latent distribution $g \in \mathbb{R}^n$ giving $G(e) = g + e$. Figure 1 shows that even in this simplified setting the trajectories of GDA, AGDA, and EGM vary greatly, diverging, cycling, and converging respectively.

There has been a recent surge in using continuous-time ODE models to understand the behavior of iterative optimization methods, initiated by Su et al. (2016). A typical focus is on the ODE arising

from taking the stepsize $s \to 0$ to zero. However, under this limit, all three of these methods have their solution paths converge to the *gradient flow* (GF) given by the ODE

$$\dot{z} = -F(z). \tag{7}$$

To understand the differences in limiting behavior we have observed between GDA, AGDA, and EGM, we need to consider the $O(s)$-resolution ODEs proposed by Shi et al. (2018) and Lu (2020), which capture differences between these methods when the stepsize s is nonzero.

1.1. Our Contributions

The main contribution of this work is understanding and characterizing different possible (potentially nonconvergent) limiting behaviors and limit points of minimax algorithms, which enables us to explain the differing trajectories observed in our one-layer GAN example.

1. **Divergence, cycling and limit points.** We derive necessary and sufficient conditions for stationary points to be attractive for each of GDA, AGDA, and EGM of the generic minimax problem (1). These conditions apply broadly to any sufficiently differentiable nonconvex-nonconcave minimax problems and explain the differences in convergence and divergence found in our GAN experiments. Our conditions are based on understanding the underlying ODEs but apply directly to each discrete-time algorithm.

2. **Regularization induced phase transitions.** Adding regularization to our one-layer CIFAR example eventually leads GDA and AGDA to converge. Figure 2 shows GDA transitions from divergence to having an attractive limit cycle, which then shrinks eventually collapsing into an attractive stationary point. In Theorem 1, we show this transition from a repulsive stationary point to attraction happens broadly. Finally, Section 4 presents several examples of Hopf bifurcations occurring in minimax optimization, explaining the observed transitions.

1.2. Related Works

Divergence and cycling. GDA is well known to diverge even for convex-concave problems while EGM still converges, the simplest example being $\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$. This behavior was described by Mokhtari et al. (2020) by relating it to the proximal point method. Cycling behaviors arising in nonconvex-nonconcave problems have been observed in a variety of settings (Letcher, 2020; Hsieh et al., 2020; Grimmer et al., 2020). We differ from these works in that our focus is on developing tools for characterizing attractive limit points and the transitions between different limiting behaviors for a wide range of minimax algorithms.

GAN equilibria. Arora et al. (2017) showed under moderate size requirements, GAN training will have an approximately pure equilibrium but that equilibrium point may be far from target true data distribution. These results however do not guarantee that an equilibrium point will be found, leaving the potential for cycling and divergence as seen in Figure 1.

Nonconvex-nonconcave limit points. Quantifying limit points as local Nash equilibrium with measures like stationarity $\nabla L(x, y) \approx 0$ (which is the first-order necessary condition for a Nash equilibrium) has been done for a variety of different first-order methods (Cherukuri et al., 2016; Daskalakis and Panageas, 2018; Adolphs et al., 2019; Mazumdar et al., 2020; Liang and Stokes, 2019). We follow in this vein as the limit points of GDA and AGDA necessarily have $\nabla L(x, y) = 0$.

Alternative measures of local optimality are discussed in Jin et al. (2020), where a notion of local minimax points is presented capturing the sequential nature of minimax problems and relating to GDA's limit points.

Continuous-time analysis of these limit points based on the related gradient flow has been previously considered (Ratliff et al., 2014; Nagarajan and Kolter, 2017; Mazumdar and Ratliff, 2019; Vlatakis-Gkaragkounis et al., 2019) but such an approach cannot distinguish between the behavior of GDA, PPM, AGDA, and EGM as these all share the same limiting flow.

Stability of limit points. Recently Lu (2020) presented an ODE approach to analyzing a broad class of discrete-time algorithms. Their analysis is non-trivial and requires high-order smoothness (five times differentiability) globally or on a large set. Our results build upon and improve this machinery as they require strong global conditions, whereas we just need similar conditions to hold only at the stationary point z^*. We achieve this by looking at different energy functions (i.e., $\|z - z^*\|_P^2$ vs $\|F(z)\|^2$). In doing so, we obtain the flexibility of changing norms, and indeed we find a non-standard norm is the natural one for describing AGDA's convergence (an algorithm not addressed in Lu (2020)). We present more detailed comparisons with Lu (2020) in Section 3.

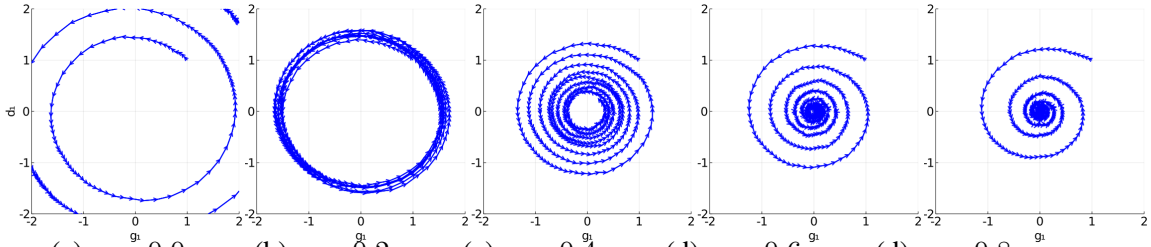
As an alternative to this ODE approach, Zhang et al. (2020) study discrete-time algorithms directly, and present conditions of stability that involve the complex eigenvalues of the non-symmetric Jacobian matrix, which cannot be easily verified even for simple problems. In contrast, our approach leads to much more transparent conditions on a p.s.d. matrix which facilitates our study on the role of the stepsize and derivatives in convergence. For example, our result shows that having a larger interaction term $\nabla_{xy}^2 L(z)$ helps EGM's convergence while hurting GDA.

Nonconvex-nonconcave convergence rates. Beyond characterizing the limit points, there has also been a recent push to establish finite-time convergence guarantees for nonconvex-nonconcave minimax problems. Many of these approaches rely on forms of convex-concave-like assumptions, such as Minty's Variational Inequality (Lin et al., 2018) and Polyak-Lojasiewicz conditions (Nouiehed et al., 2019; Yang et al., 2020). Convergence guarantees for the proximal point method on quite general nonconvex-nonconcave problems are given in (Grimmer et al., 2020) when the interaction block of the Hessian is sufficiently strong or sufficiently weak.

2. GAN Divergence, Cycling and Phase Transitions

As briefly described in the introduction, we begin by surveying the types of solution paths that arise numerically from GAN training (2). We first fix the true data distribution p_{data} as being uniformly over the set of $50,000$ training images in the CIFAR-10 dataset, each represented as vectors of length $n = 32 \times 32 \times 3$ and the latent distribution p_{latent} as a standard normal. We find that considering one-layer networks already suffices to encounter a wide range of different solution path geometries when solving (2). We consider a discriminator controlling weights $d \in \mathbb{R}^n$ in the logistic (6) and a generator controlling a translation $g \in \mathbb{R}^n$ giving $G(e) = g + e$.

For this setup, the origin $(g, d) = (0, 0)$ has $F(0, 0) \approx (0, 0)$ and so it is an approximate stationary point for all three of GDA, AGDA, and EGM. However, the solution paths of these methods vary widely when initialized near the origin at $g = (1, 0, ..., 0)$ and $d = (1, 0, ..., 0)$ with fixed $s = 0.2$. Sample solution paths using batches of 10 samples to approximate the gradients are shown in Figure 1, plotting the first coordinate of g and d which had the only nonzero initializations. Appendix A gives more sample trajectories using other batch sizes, showing the same general dynamics hold.



(a) $\alpha = 0.0$ (b) $\alpha = 0.2$ (c) $\alpha = 0.4$ (d) $\alpha = 0.6$ (d) $\alpha = 0.8$

Figure 2: GDA trajectory on a simple GAN with increasing quadratic regularization, plotting the first coordinates (g_1, d_1) at each iteration.

This experiment shows three different limiting behaviors arising: GDA diverges spiraling outward from the origin, AGDA falls into a stable cycle around the origin, and EGM converges into a stationary point around the origin. In Section 3, we develop theory based on related $O(s)$-resolution ODE models explaining these observed differences in convergence and divergence.

Since GDA and AGDA failed to converge, a typical recourse is to add regularization to the objective to deter this behavior. Using L2 penalization, we have the modified training problem objective

$$\mathbb{E}_{s\sim p_{data}}\left[\log D(s)\right] + \mathbb{E}_{e\sim p_{latent}}\left[\log(1 - D(G(e)))\right] + \frac{\alpha}{2}\|g\|^2 - \frac{\alpha}{2}\|d\|^2, \tag{8}$$

for any level of regularization $\alpha \geq 0$. Adding mild amounts of regularization $\alpha < 1$ suffices to cause the nonconvergent trajectories of GDA and AGDA to transition into convergent paths.

Figure 2 shows the trajectory of GDA (4) as the regularization parameter increases. We see that as α grows larger, the stable limit cycle shrinks until it collapses to a point for some critical value of α between $\alpha = 0.4$ and 0.6. For all α larger than this critical point, the trajectory of GDA has transitioned from nonconvergent cycling to convergence.

In the following theorem, we show this phenomenon is typical. Provided Gaussian data and latent distributions, applying GDA to (8) has a repulsive fixed point that transitions to become attractive with sufficient regularization.

Theorem 1 *For any covariance matrices $\Sigma \succeq \Sigma'$ in $p_{data} \sim N(0, \Sigma)$ and $p_{latent} \sim N(0, \Sigma')$, the origin is a stationary point of (8) which is repulsive to the GDA's dynamic for $\alpha = 0$ and transitions to be attractive whenever $\alpha \geq \lambda_{max}(\Sigma - \Sigma')/4 + O(s)$.*

Proving this theorem relies on a careful understanding the dynamics of the related $O(s)$ ODE, developed in Section 3. Then in Section 4, we explain the observed phase transitions as Hopf bifurcations occurring in the related $O(s)$-resolution ODE for each algorithm.

3. The Attractive Limit Points of the Four Dynamics

Our simple GAN experiments show the trajectories of GDA, AGDA, and EGM can be topologically different. Here we explain such differences by studying the $O(s)$-resolution ODEs for the three algorithms (which are formally defined in Section 3.1) as well as the simple gradient flow dynamics. From these, we arrive at our main theoretical result Theorem 5 in Section 3.2, giving necessary and

sufficient conditions for a stationary point to be attractive for these dynamics. Further, Section 3.3 shows that for reasonably small s, these conditions guarantee the discrete-time iterations have a linear attractor whenever their associated ODE does.

For any positive definite $P \succ 0$, let $\|z\|_P^2 = z^T P z$. Our theory applies to any stationary point z^* with second derivatives denoted by $A = \nabla_{xx}^2 L(z^*), B = \nabla_{xy}^2 L(z^*), C = -\nabla_{yy}^2 L(z^*)$.

3.1. $O(s)$-Resolution ODE Systems for Discrete-time Algorithms

The traditional way to obtain an ODE for a discrete-time algorithm is to let the stepsize s go to 0. While that may be the easiest and the most natural approach, we never use $s = 0$ in practice, and even worse, the solution trajectory of the resulting ODE and the discrete-time algorithm with any positive stepsize can be topologically different. This is exemplified by the trajectories of GDA, AGDA, and EGM seen in Figure 2 which are topologically different, despite sharing the same ODE when letting the stepsize go to 0, namely gradient flow. To overcome such limitations, high-order resolution ODEs (Shi et al., 2018; Lu, 2020) have been proposed recently to study the subtle differences between similar algorithms. We here utilize the framework proposed by Lu (2020) to study the $O(s)$-resolution ODE for these three algorithms and explain their different limiting behaviors.

More formally, we say an ODE given by generic functions f_0 and f_1

$$\dot{z} = f_0(z) + s f_1(z) \tag{9}$$

is the $O(s)$-resolution ODE of a discrete-time algorithm with iterate update $z^+ = g(z, s)$ if it satisfies that for any z and $z^+ = g(z, s)$, $\|z(s) - z^+\| = o(s^2)$. For stochastic gradients, this can be extended by considering the fluid limit as $z^+ = \mathbb{E}\, g(z, s)$ given by using large enough batches. It turns out such an ODE (9) is unique for a smooth enough discrete-time algorithm (see Theorem 1 in Lu (2020)) and we herein study the $O(s)$-resolution ODE of GDA, AGDA and EGM:

Proposition 2 *1.* ***GDA:*** *The $O(s)$-resolution ODE of GDA is*

$$\dot{z} = -F(z) - \frac{s}{2} \nabla F(z) F(z). \tag{10}$$

2. ***EGM:*** *The $O(s)$-resolution ODE of EGM is*

$$\dot{z} = -F(z) + \frac{s}{2} \nabla F(z) F(z). \tag{11}$$

3. ***AGDA:*** *The $O(s)$-resolution ODE of AGDA is*

$$\dot{z} = -F(z) - \frac{s}{2} \begin{bmatrix} A & B^T \\ B & C \end{bmatrix} F(z)\,, \tag{12}$$

Letting the stepsize go to zero, all three ODEs reduce to the previously stated gradient flow ODE (7). Note this $O(s)$-ODE machinery can also apply to algorithms employing different stepsizes or time-scales in x and y. For example, GDA with stepsize s in x and γs in y, follows the general trajectory $\dot{z} = -\begin{bmatrix} F_x(z) \\ \gamma F_y(z) \end{bmatrix} - \frac{s}{2} \begin{bmatrix} \nabla F_x(z) \\ \gamma \nabla F_y(z) \end{bmatrix} \begin{bmatrix} F_x(z) \\ \gamma F_y(z) \end{bmatrix}$. The convergence and phase transitions of this can be analyzed similar to the results presented herein for GDA, AGDA, and EGM.

3.2. Local Convergence of the $O(s)$-Resolution ODE

Armed with these more refined ODEs for GDA, AGDA, and EGM, we can formalize our necessary and sufficient conditions for a limit point to be attractive. For a generic ODE

$$\dot{z} = G(z)\ , \tag{13}$$

we need conditions for when a fixed point z^* to the ODE (13) is attractive, which will translate into our required conditions for the convergence of specific algorithms. The traditional stability theory for ODEs says that the attractivity of a fixed point relies on whether the real part of every eigenvalue of $\nabla G(z^*)$ is negative, but unfortunately, it is not transparent to directly evaluate the eigenvalue structure of $\nabla G(z^*)$ for minimax problems. Instead, to take advantage of the structure of minimax problems, we here study whether $\|z(t) - z^*\|_P$ goes to 0 for a positive definite norm matrix P. The natural norm P to use differs for different algorithms.

Definition 3 *We say a fixed point z^* is a linear attractor to an ODE system (13) in the P norm if there exists $\delta, \rho > 0$ such that for any initial solution $z(0) \in B(z^*, \delta)$, it holds that $\|z(t) - z^*\|_P \leq e^{-\rho t}\|z(0) - z^*\|_P$.*[1]

Then the linear attractors of our ODEs of interest are characterized as follows.

Theorem 4 *1. **GF:** Suppose z^* is a stationary point to GF (7). Then z^* is a linear attractor to GF (7) in Euclidean norm (that is, $P = I$) if it holds that*

$$A \succ 0\ , C \prec 0 \tag{14}$$

and not if either inequality is strictly violated.
*2. **GDA:** Suppose z^* is a stationary point to the $O(s)$-resolution ODE to GDA (10). Then z^* is a linear attractor to (10) in Euclidean norm if*

$$A + \frac{s}{2}(A^2 - BB^T) \succ 0\ , C + \frac{s}{2}(C^2 - B^TB) \succ 0 \tag{15}$$

and not if either inequality is strictly violated.
*3. **EGM:** Suppose z^* is a stationary point to the $O(s)$-resolution ODE to EGM (11). Then z^* is a linear attractor to (11) in Euclidean norm if*

$$A - \frac{s}{2}(A^2 - BB^T) \succ 0\ , C - \frac{s}{2}(C^2 - B^TB) \succ 0 \tag{16}$$

and not if either inequality is strictly violated.
*4. **AGDA:** Suppose z^* is a stationary point to the $O(s)$-resolution ODE to AGDA (12). Then z^* is a linear attractor to (12) in scaled Euclidean norm with $P = \begin{bmatrix} I & \frac{1}{2}sB^T \\ \frac{1}{2}sB & I \end{bmatrix}$ if*

$$\begin{bmatrix} A + \frac{s}{2}A^2 + \frac{s^2}{4}\left(AB^TB + B^TBA\right) & \frac{s}{2}\left(AB^T + B^TC\right) + \frac{s^2}{4}\left(A^2B^T + B^TC^2\right) \\ \frac{s}{2}\left(B^TA + CB\right) + \frac{s^2}{4}\left(BA^2 + C^2B\right) & C + \frac{s}{2}C^2 + \frac{s^2}{4}\left(CBB^T + BB^TC\right) \end{bmatrix} \succ 0 \tag{17}$$

and not if the inequality is strictly violated.

1. In ODE literature, this convergence rate sometimes is called exponential convergence. Here we choose to call it linear in order to be consistent with the optimization literature.

Similar sufficient conditions for GDA and EGM were presented in Lu (2020). Theorem 4 improves upon those by only requiring the positive definiteness conditions at z^*, whereas Lu (2020) requires these hold globally or on a potentially unbounded set like $\{z|\|\nabla F(z)\| \leq \delta\}$. This improvement comes from analyzing different energy functions ($\|z - z^*\|_P$ instead of $\|\nabla F(z)\|$). The flexibility of the norm P allows us to present the tighter convergence condition for AGDA (17).

Examining the difference between our conditions for EGM and GDA makes clear that EGM's trajectory is more attractive whenever $A^2 - BB^T$ and $C^2 - BB^T$ are negative definite. For example, whenever the interaction block of the Hessian $B = \nabla^2_{xy} L(z^*)$ is large, this will be the case. However, such a limit point may be undesirable (failing to be a local minimax solution). Weighing the tradeoff between potentially failing to converge with GDA and undesirable limit points from EGM is likely an application-specific problem.

In Section 4, we consider several example families of minimax optimization problems, where these conditions can be seen. In Figure 2, the size of A and C grow, leading GDA to develop an attractive limit point. In Figure 4, the interaction term B grows, leading EGM to develop an attractive limit point whereas Figure 5 shows GDA develops a repulsive limit point as B grows.

To prove this theorem, we first establish some basic properties of linear attractors. Consider a fixed point z^* to the generic ODE (13). Suppose the ODE is locally Lipschitz continuous around z^*, i.e., there exists $H > 0$ such that $\|\nabla G(z)\| \leq H$ for any $z \in \{z|\|z - z^*\| \leq 1\}$. The following proposition presents a sufficient condition for whether a fixed point is a linear attractor:

Proposition 5 *Consider a fixed point z^* to the dynamic (13). For any positive definite $P \succ 0$, if*

$$\frac{1}{2}\left(\nabla G(z^*)^T P + P \nabla G(z^*)\right) \prec 0\,, \tag{18}$$

then z^* is a linear attractor to the dynamic $\dot{z} = G(z)$ in the P norm.

The next proposition shows that the above sufficient condition is essentially tight. The only slack between these results is boundary case when the Jacobian-like term $\frac{1}{2}\left(\nabla G(z^*)^T P + P \nabla G(z^*)\right)$ is only negative semidefinite, in which case higher-order derivatives determine attraction.

Proposition 6 *Consider a fixed point z^* to the dynamic (13). For any positive definite $P \succ 0$, if*

$$\lambda_{max}\left(\frac{1}{2}\left(\nabla G(z^*)^T P + P \nabla G(z^*)\right)\right) > 0\,, \tag{19}$$

then z^* is not a linear attractor to the dynamic $\dot{z} = G(z)$ in the P norm.

Applying these propositions to the four continuous-time dynamics we consider is the core of our proof of Theorem 4. One key step in establishing these conditions is choosing the corresponding norms for studying different algorithms. As we can see in Figure 1, the trajectory of AGDA follows from a slightly skewed ellipsoid while that of GDA is more symmetric. This is fundamentally because GDA has simultaneous updates in primal and dual, while AGDA utilizes sequential updates. This difference results in different choices of norm P when studying the two algorithms. It turns out the natural norms for GF, GDA and EGM are the same: the Euclidean norm with $P = I$, and the natural norm for AGDA is a scaled Euclidean norm with $P = \begin{bmatrix} I & \frac{1}{2}sB^T \\ \frac{1}{2}sB & I \end{bmatrix}$. These norms are obtained by examining the dynamics of each method when solving bilinear systems $L(x, y) = y^T B x$.

Remark 7 *It holds for bilinear problem $L(x,y) = y^T Bx$ that the LHS of (17) is 0. This explains the circling over an ellipsoid behavior of AGDA when solving the bilinear problem. Indeed, the ellipsoid is given by the matrix $P = \begin{bmatrix} I & \frac{1}{2}sB^T \\ \frac{1}{2}sB & I \end{bmatrix}$. Note that not all norms are equivalent in terms of linear attractors, indeed AGDA may linearly attract in the P-norm but fail to monotonically decrease the Euclidean norm.*

3.3. Local Convergence of the Discrete-time Algorithms

Here we extend our guarantees from Theorem 4 from the underlying ODEs to the discrete-time algorithms. Similar to the results stated in Section 3.2, we first present general results and then apply them to the specific algorithms. Consider a generic discrete-time algorithm with iterate update $z^+ = g(z,s)$ and the dynamic around a fixed point z^*. As defined in Lu (2020), we say the $O(s)$-resolution ODE is gradient-based if for any $\delta > 0$, there exists a constant $c > 0$ such that it holds for any $z \in \{\|F(z)\| \leq \delta\}$ and a small enough stepsize s that

$$\|z(s) - z^+\| \leq cs^3 \|F(z)\| \,, \tag{20}$$

where $z^+ = g(z,s)$, and $z(s)$ is the solution obtained at $t = s$ following its $O(s)$-resolution ODE with initial solution $z(0) = z$. Indeed, the $O(s)$-resolution ODE of GDA, AGDA, and EGM are gradient-based when $L(x,y)$ is smooth enough. This is formalized in the follow proposition utilizing fifth-order continuously differentiablity at a fixed point z^*, which is a fairly mild condition as in practice many objective functions are analytical (i.e. infinitely differentiable), for example, any GANs with sigmoid activation functions.

Proposition 8 *Suppose $L(x,y)$ is fifth-order continuously differentiable around a fixed point z^*. Then $O(s)$-resolution ODE of GDA, AGDA, and EGM are gradient-based around z^*.*

Using this condition, we find that if the z^* is a linear attractor to the $O(s)$-resolution ODE, then it is also a linear attractor to the discrete-time algorithm.

Theorem 9 *Consider a discrete-time algorithm with iterate update $z^+ = g(z,s)$, and its $O(s)$-resolution ODE (13). Suppose z^* is a linear attractor to its $O(s)$-resolution ODE and some $b > 0$ has*

$$\lambda_{max}\left(\frac{1}{2}\left(\nabla G(z^*)^T P + P \nabla G(z^*)\right)\right) \leq -bs \,.$$

Suppose the $O(s)$-resolution ODE is gradient-based around z^*. Then there exists s^*, such that for any $s \leq s^*$, z^* is a linear attractor to the discrete-time algorithm.

As a direct consequence of Theorem 9, we know that the conditions (15), (16) and (17) are sufficient conditions for z^* being attractive for each of GDA, AGDA and EGM when the stepsize s is reasonably small (i.e., smaller than a constant). Lastly, we note that when the stepsize s is reasonably small, the stationary points of the $O(s)$-resolution ODEs we have considered are indeed also stationary points of $L(x,y)$. This completes our recovery from ODE guarantees back to the discrete-time algorithms.

Proposition 10 *When the stepsize s is sufficiently small, the stationary points to the $O(s)$-resolution ODE of GDA, EGM and AGDA, namely (10), (11) and (12) are also stationary points of (1).*

4. Phase Transitions between Limit Points and Limiting Cycles

The GAN problem considered in Section 2 demonstrated that divergence, cycling, and convergence all easily arise from a simple formulation of GAN training. Meanwhile, Figure 2 and Theorem 1 showed that adding mild amounts of regularization sufficed to cause the trajectories of the considered methods to all transition to convergence. Here we explain these phase transitions in the discrete algorithmic iterations using their related ODEs. In continuous-time dynamical systems, the transformation from an (attractive or repulsive) limit point into a limit cycle (or vice versa) can occur when the eigenvalues of the Jacobian $\nabla G(z)$ transition from being all negative real part to having some nonnegative real values. When exactly one conjugate pair of eigenvalues crosses to having positive real part, the dynamics of the system are described as a *Hopf Bifurcation*.

Such transitions are controlled by this pair of eigenvectors and hence are fundamentally two-dimensional phenomena. This notion can be generalized to multi-dimension problems using the Central Manifold Theorem (Kelley, 1967), which essentially states that one can locally decompose the multi-dimension spaces into independently evolving dynamics on a one-dimension or two-dimension manifold. Thus, one could reparameterize the dynamics $\dot{z} = G(z)$ to isolate these behaviors, however, the details of such an approach are beyond the scope of this work. We consider phase transitions of in an algorithm's solution path as a hyperparameter $\alpha \in \mathbb{R}$ changes

$$\min_x \max_y L(x, y, \alpha).$$

For example, this captures our previously considered setting of regularizing our one-layer GAN setup (8) on CIFAR-10 data and the setting of our Theorem 1.

In the rest of this section, we focus on concrete examples of Hopf bifurcations arising. For ease of presentation, we consider a two-dimension system that corresponds to a conjugate pair of eigenvalues crossing from negative to positive real part (potentially coming from reparameterizing to isolate the independently evolving two-dimension manifold). We suppose the stationary point of interest occurs at the origin $z = 0$ and the dynamics $\dot{z} = G(z)$ at the critical α^* are in the form

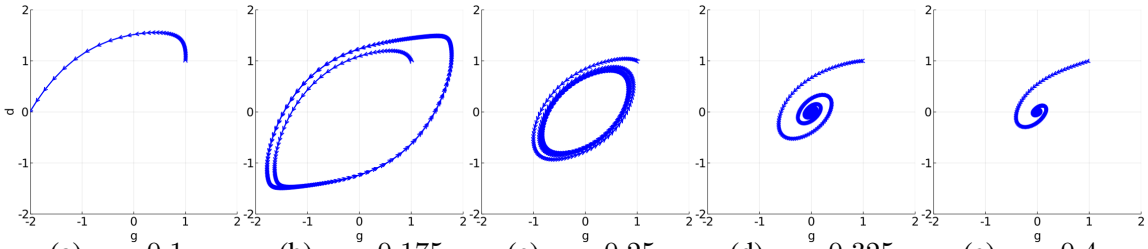
$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & -w \\ w & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} P(x, y) \\ Q(x, y) \end{bmatrix} \tag{21}$$

where $w \neq 0$ and P, Q both have $P(0, 0) = Q(0, 0) = 0$ and $\nabla P(0, 0) = \nabla Q(0, 0) = 0$. Indeed, most of the dynamic around the stationary point $z = 0$ and critical point α^* can be transformed to (21) except for some degenerated cases (Kuznetsov, 1998). For dynamics of the form (21), the shape of the Hopf bifurcation occurring for some critical α^* is determined by the *First Lyapunov Coefficient*, defined as

$$\begin{aligned} l_1(0) =& \frac{1}{8w}\left(P_{xxx} + P_{xyy} + Q_{xxy} + Q_{yyy}\right) \\ &+ \frac{1}{8w^2}\left(P_{xy}(P_{xx} + P_{yy}) - Q_{xy}(Q_{xx} + Q_{yy}) - P_{xx}Q_{xx} + P_{yy}Q_{yy}\right) \end{aligned} \tag{22}$$

where subscripts denote partial derivatives at $(0, 0)$ (e.g., $P_{xxy} = \frac{d^3}{dx^2 dy} P(0, 0)$).

Depending on the sign of this coefficient, two types of bifurcations can arise, *supercritical* and *subcritical*. A supercritical bifurcation (occurring when $l_1(0) < 0$) corresponds to an attractive limit cycle transforming into an attractive limit point and a subcritical bifurcation (occurring when



(a) $\alpha = 0.1$ (b) $\alpha = 0.175$ (c) $\alpha = 0.25$ (d) $\alpha = 0.325$ (e) $\alpha = 0.4$

Figure 3: Regularization causing GDA on the GAN (8) to undergo a supercritical Hopf bifurcation.

$l_1(0) > 0$) corresponds to a repulsive limit point transforming into an attractive limit point surrounded by a repulsive limit cycle, respectively. The following subsections demonstrate these transitions can easily arise in minimax optimization, first in GANs and second in polynomial optimization.

4.1. GAN Hopf Bifurcations

First, we illustrate a supercritical Hopf bifurcation by considering a two-dimensional case of the GAN formulation (8) fixing p_{data} and p_{latent} to both produce ± 1 with probability one half each. Consequently, if the generator plays $g = 0$, it exactly matches the true data distribution and is the desired solution. However, GDA fails to converge at all for this problem unless sufficient regularization is added. Figure 3 shows GDA with $s = 0.2$ and full gradient evaluations diverging or cycling when $\alpha \leq 0.25$ and converging when $\alpha \geq 0.325$ (much like we previously saw in Figure 2).

Numerically, we see that the transition between cycling and convergence happens around $\alpha^* \approx 0.26872$. To verify this is a supercritical Hopf bifurcation, we simply need to write the GDA dynamics (3) in the form (21) and then compute the Lyapunov coefficient. One valid reparameterization of the GDA dynamics on (g, d) comes from considering $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0.57735 & 0.57735 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} g \\ d \end{bmatrix}$ which for our example problem has $\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} \approx \begin{bmatrix} 0 & -0.43463 \\ 0.43463 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + o(\|(x, y)\|)$. Computing higher derivatives of these dynamics[2] verifies $l_1(0) \approx -0.11198 < 0$ and so indeed Figure 3 shows a supercritical Hopf bifurcation.

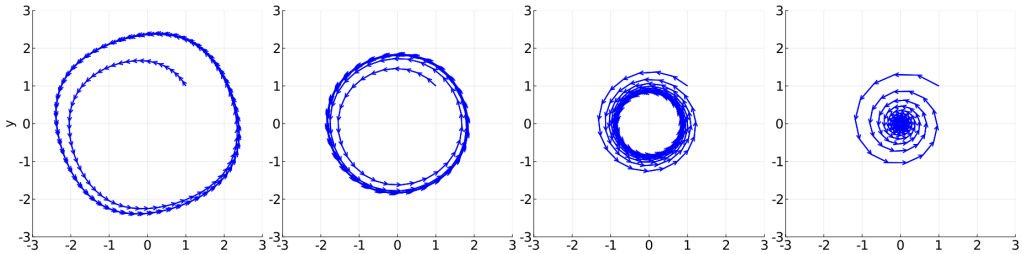
4.2. Polynomial Supercritical and Subcritical Hopf Bifurcations

Consider the polynomial minimax problem with $f(x) = (x + 3)(x + 1)(x - 1)(x - 3)$ given by

$$\min_x \max_y \; f(x) + \alpha xy - f(y) \tag{23}$$

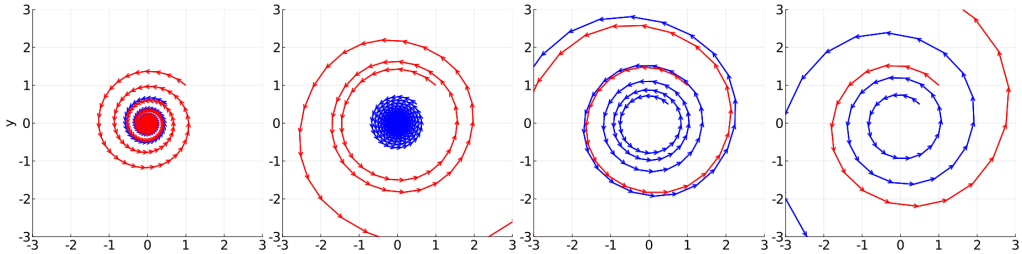
The behavior of EGM with $(x_0, y_0) = (1, 1)$ and $s = 0.002$ as α varies is shown in Figure 4. As α increases past $\alpha^* \approx 165$, the trajectories attractive limit cycle contracts towards the origin, converting the origin from a repulsive limit point to an attractive one. The related ODE dynamics (11) at $\alpha^* \approx 143$ undergo a supercritical Hopf bifurcation (with $l_1(0) \approx -0.04362$), matching the transition we observed in Figure 4.

2. At the critical $\alpha^* \approx 0.26872$, our reparameterization has $w \approx 0.434633$, $P_{xxx} \approx -0.20695$, $P_{xyy} \approx 0.01227$, $Q_{xxy} \approx -0.20695$, and $Q_{yyy} \approx 0.01227$.



(a) $\alpha = 50$ (b) $\alpha = 100$ (c) $\alpha = 150$ (d) $\alpha = 200$

Figure 4: EGM on the polynomial (23) undergoing a supercritical Hopf bifurcation as α increases.



(a) $\alpha = 100$ (b) $\alpha = 125$ (c) $\alpha = 150$ (d) $\alpha = 175$

Figure 5: GDA on (24) from two different initializations. A subcritical Hopf bifurcation occurs as a repulsive limit cycle collapses into the origin transforming from attractive to repulsive.

Similarly, the behavior of GDA applied to the polynomial minimax optimization problem

$$\min_x \max_y \; g(x) + \alpha xy - g(y), \tag{24}$$

where $g(x) = -(x+3)(x+1)(x-1)(x-3)$, undergoes a subcritical Hopf Bifurcation: As α decreases past $\alpha^* \approx 135$, the origin changes from being repulsive to attractive, surrounded by a repulsive cycle. Further decreasing α grows this repulsive cycle, creating a larger area attracted to the origin. Figure 5 shows trajectories initialized at $(1/2, 1/2)$ and $(1, 1)$ which are excluded from this attractive region by $\alpha = 125$ and $\alpha = 150$, respectively. Considering the related ODE dynamics (10) at $\alpha^* \approx 143$, the dynamics undergo a subcritical Hopf bifurcation with $l_1(0) \approx 0.04362$.

The calculation of the Lyapunov coefficient in each case is given in Appendix C.

5. Conclusion

Inspired by the topologically different limiting behaviors seen in GAN training, we derived necessary and sufficient conditions for a stationary point to be attractive for each of GDA, AGDA, and EGM that avoid sending the stepsize parameter to zero. These results are based on understanding the ODE dynamics related to each method, which further allows us to describe the observed phase transitions to convergence as Hopf bifurcations. This explains the transitions observed when adding regularization to our simplified family of GAN training problems.

References

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. volume 89 of *Proceedings of Machine Learning Research*, pages 486–495. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/adolphs19a.html.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). volume 70 of *Proceedings of Machine Learning Research*, pages 224–232, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arora17a.html.

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

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Ashish Cherukuri, Bahman Gharesifard, and Jorge Cortes. Saddle-point dynamics: conditions for asymptotic stability of saddle points, 2016. arXiv: 1510.02145.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9256–9266, Red Hook, NY, USA, 2018. Curran Associates Inc.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of the proximal point method for nonconvex-nonconcave minimax optimization, 2020. arXiv: 2006.08667.

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: convergence to spurious non-critical sets, 2020. arXiv: 2006.09065.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization?, 2020. arXiv: 1902.00618.

Al Kelley. The stable, center-stable, center, center-unstable, unstable manifolds. *Journal of Differential Equations*, 3(4):546 – 570, 1967. ISSN 0022-0396. doi: https://doi.org/10.1016/0022-0396(67)90016-2. URL http://www.sciencedirect.com/science/article/pii/0022039667900162.

Yuri A. Kuznetsov. *Elements of Applied Bifurcation Theory (2nd Ed.)*. Springer-Verlag, Berlin, Heidelberg, 1998. ISBN 0387983821.

Alistair Letcher. On the impossibility of global convergence in multi-loss optimization, 2020. arXiv: 2005.12649.

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

Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities, 2018. arXiv: 1810.10207.

Haihao Lu. An $o(s^r)$-resolution ode framework for discrete-time optimization algorithms and applications to the linear convergence of minimax problems, 2020. arXiv: 2001.08826, version 5.

Eric Mazumdar and Lillian J Ratliff. Local nash equilibria are isolated, strict local nash equilibria in 'almost all'zero-sum continuous games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6899–6904. IEEE, 2019.

Eric Mazumdar, Lillian J. Ratliff, and S. Shankar Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, Jan 2020. ISSN 2577-0187. doi: 10.1137/18m1231298. URL http://dx.doi.org/10.1137/18M1231298.

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

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. *Advances in neural information processing systems*, 30, 2017.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems 32*, pages 14934–14942. Curran Associates, Inc., 2019.

Lillian J Ratliff, Samuel A Burden, and S Shankar Sastry. Genericity and structural stability of non-degenerate differential nash equilibria. In *2014 American Control Conference*, pages 3990–3995. IEEE, 2014.

Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations, 2018. arXiv: 1810.08907.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016. URL http://jmlr.org/papers/v17/15-084.html.

Sergio Verdu and H Poor. On minimax robustness: A general approach and applications. *IEEE Transactions on Information Theory*, 30(2):328–340, 1984.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. *Advances in Neural Information Processing Systems*, 32, 2019.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems, 2020. arXiv: 2002.09621.

Guojun Zhang, Pascal Poupart, and Yaoliang Yu. Optimality and stability in non-convex smooth games, 2020.

Appendix A. More trajectories on simple CIFAR GANs

The trajectories presented in Figure 1 and 2 all use a moderate batch size of 10, reducing stochastic effects. Figure 6 presents 500 steps of a sample trajectory when this batch size is 1, 10, or 100 for each method. Largely, the dynamics of GDA, AGDA, and EGM stay the same. GDA still diverges and AGDA still converges into an attractive limit cycle but with much greater variance in how closely it follows it. EGM still converges towards the origin, eventually staying in a ball around the origin whose radius shrinks as the variance in the gradient samples shrinks.

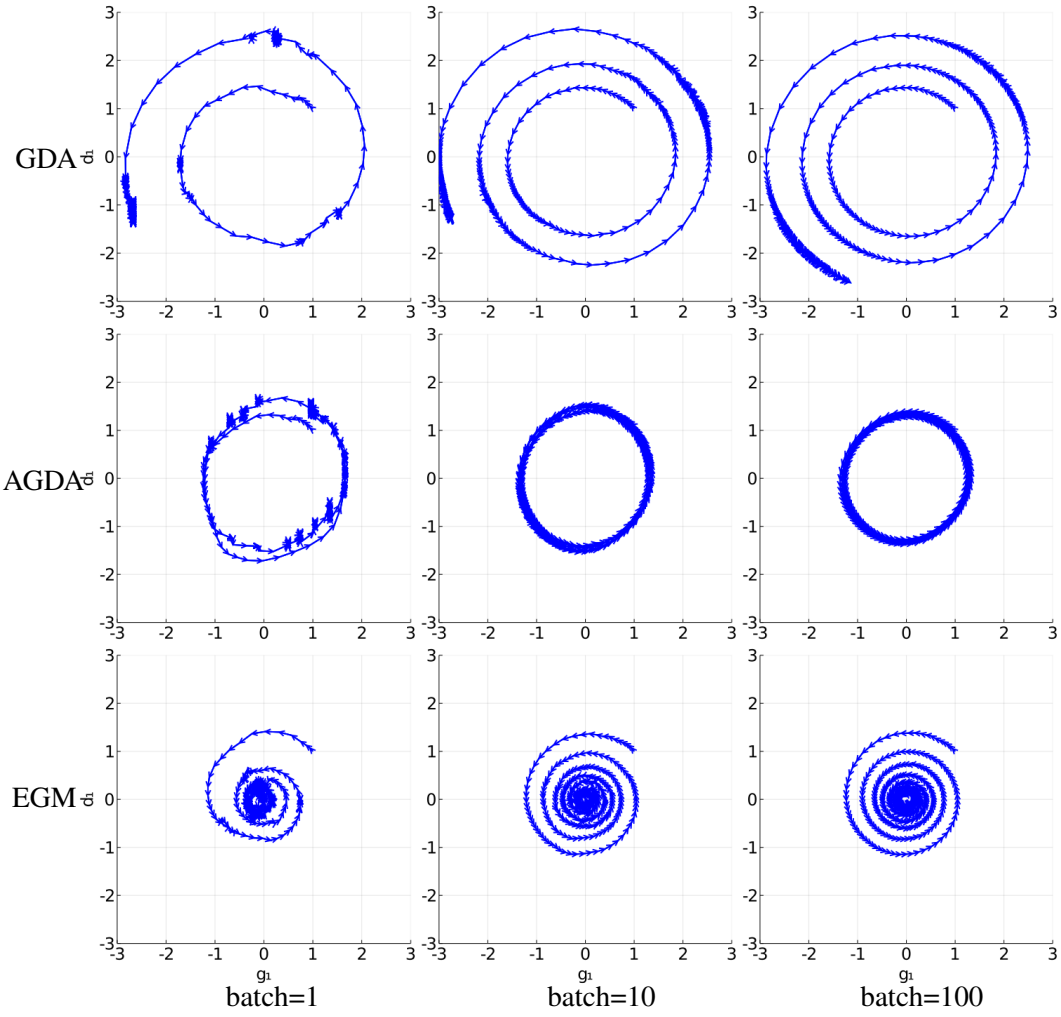


Figure 6: Sample trajectories with batch sizes 1, 10, and 100 for GDA, AGDA, and EGM on our simplified GAN over CIFAR data.

Appendix B. Deferred Proofs

B.1. Proof of Proposition 2

We utilize Theorem 1 in Lu (2020) and the notations therein to compute the $O(s)$-resolution ODE. For GDA and EGM, their $O(s)$-resolution ODEs are presented in Corollary 1 in Lu (2020). For AGDA, it follows by Taylor expansion of the update rule over stepsize s that

$$z_{k+1} = z_k + sg_1(z_k) + s^2 g_2(z_k) + s^3 g_3(z_k) + o(s^3)\,, \tag{25}$$

where $g_1(z_k) = -F(z_k)$, $g_2(z_k) = \begin{bmatrix} 0 \\ \nabla_x F_y(z_k) F_x(z_k) \end{bmatrix}$, and $g_3(z_k) = \begin{bmatrix} 0 \\ -\nabla_{xx} F_y(z_k)(F_x(z_k), F_x(z_k)) \end{bmatrix}$ with $\nabla_{xx} F_y(z_k)$ being a tensor. Now suppose the $O(s)$-resolution ODE of AGDA is $\dot{z} = f_0(z) + sf_1(z)$, then it follows by Theorem 1 in Lu (2020) that

$$f_0(z) = g_1(z) = -F(z)$$

$$f_1(z) = g_2(z) - \frac{1}{2}\nabla f_0(z) f_0(z) = -\frac{1}{2}\begin{bmatrix} A & B^T \\ B & C \end{bmatrix} F(z)\,.$$

Hence (12) is the $O(s)$-resolution ODE of AGDA.

B.2. Proof of Proposition 5

We here claim the following fact, which directly proves the proposition: Suppose there exists positive definite matrices $\alpha > 0$ such that

$$\frac{1}{2}\left(\nabla G(z^*)^T P + P\nabla G(z^*)\right) \preceq -\alpha I\,, \tag{26}$$

Let $\bar{P} = \lambda_{\max}(P)$ and $\underline{P} = \lambda_{\min}(P)$ be the maximal and minimal eigenvalue of p.s.d. matrix P. Suppose $H > 0$ satisfies that $\|\nabla^2 G(z)\| \le H$ for $z \in \{z | \|z - z^*\|_P \le 1\}$ and the initial point $z(0)$ is close enough to z^*: $\|z(0) - z^*\|_P \le \min\{\frac{\alpha\sqrt{\underline{P}}}{H\bar{P}}, 1\}$, then the trajectory $z(t)$ monotonically converges to z^* with

$$\|z(t) - z^*\|_P^2 \le e^{-\alpha t/\bar{P}} \|z(0) - z^*\|_P^2. \tag{27}$$

Consider the Lyapunov function $\frac{1}{2}\|z(t) - z^*\|_P^2$. This is monotonically decreasing as

$$\begin{aligned}
\frac{\partial}{\partial t}\frac{1}{2}\|z(t) - z^*\|_P^2 &= (z(t) - z^*)^T P G(z(t)) \\
&\le (z(t) - z^*)^T P \nabla G(z^*)(z(t) - z^*) + \frac{H\bar{P}}{2}\|z(t) - z^*\|^3 \\
&= \frac{1}{2}(z(t) - z^*)^T (\nabla G(z^*)^T P + P\nabla G(z^*))(z(t) - z^*) + \frac{H\bar{P}}{2}\|z(t) - z^*\|^3 \\
&\le -\alpha\|z(t) - z^*\|^2 + \frac{H\bar{P}}{2\sqrt{\underline{P}}}\|z(t) - z^*\|^2 \|z(t) - z^*\|_P \\
&\le -\frac{\alpha}{2}\|z(t) - z^*\|^2 \\
&\le -\frac{\alpha}{2\bar{P}}\|z(t) - z^*\|_P^2\,,
\end{aligned}$$

where the first inequality utilizes $\|\nabla G(z)\| \leq H$, the second inequality is from (26) and $\|z(t) - z^*\| \leq \frac{1}{\sqrt{\underline{P}}}\|z(t) - z^*\|_P$, and the third inequality follows from the fact that $\|z(t) - z^*\|_P \leq \frac{\alpha\sqrt{\underline{P}}}{H\bar{P}}$ by noticing $\|z(t) - z^*\|_P$ is monotonically non-increasing over t. (27) follows immediately from the above inequality.

B.3. Proof of Proposition 6

By condition (19), we know that there exists some $\alpha > 0$ such that

$$\lambda_{max}\left(\frac{1}{2}\left(\nabla G(z^*)^T P + P\nabla G(z^*)\right)\right) \geq \alpha\,.$$

Let e to be the unit eigenvector of $\left(\nabla G(z^*)^T P + P\nabla G(z^*)\right)$ that correspond to the max eigenvalue. For any $\delta > 0$, consider taking $z(0) = z^* + \min\{\frac{\alpha}{H\bar{P}}, \delta\}e$. Then it holds that

$$\begin{aligned}
\frac{\partial}{\partial t}\frac{1}{2}\|z(0) - z^*\|_P^2 &= (z(0) - z^*)^T P G(z(0)) \\
&\geq (z(0) - z^*)^T P\nabla G(z^*)(z(0) - z^*) - \frac{H\bar{P}}{2}\|z(0) - z^*\|^3 \\
&= \frac{1}{2}(z(0) - z^*)^T(\nabla G(z^*)^T P + P\nabla G(z^*))(z(0) - z^*) - \frac{H\bar{P}}{2}\|z(0) - z^*\|^2 \\
&\geq \alpha\|z(0) - z^*\|^2 - \frac{H\bar{P}}{2}\|z(0) - z^*\|^3 \\
&\geq \frac{\alpha}{2}\|z(0) - z^*\|^2 \\
&> 0\,,
\end{aligned}$$

where the first inequality utilizes $\|\nabla G(z)\| \leq H$, the second inequality is from the definition of $z(0)$, and the third inequality follows from the fact that $\|z(0) - z^*\| \leq \frac{\alpha}{H\bar{P}}$. Therefore, for a smaller enough t, it holds that $\|z(t) - z^*\|_P > \|z(0) - z^*\|_P$, which contradicts with the definition of begin a linear attractor.

B.4. Proof of Theorem 4

1. For Gradient Flow (7), we take $P = I$. Notice that

$$\frac{1}{2}\left(\nabla G(z^*)^T + \nabla G(z^*)\right) = -\frac{1}{2}\left(\nabla F(z^*) + (\nabla F(z^*))^T\right) = \begin{bmatrix} -A & \\ & -C \end{bmatrix} \prec 0\,,$$

where the inequality is due to (14). Applying Propositions 5 and 6 with $P = I$ finishes the proof.

2. For Gradient Descent Ascent (10), we take $P = I$. Notice that

$$\begin{aligned}
\frac{1}{2}\left(\nabla G(z^*)^T + \nabla G(z^*)\right) &= -\frac{1}{2}\left(\nabla F(z^*) + (\nabla F(z^*))^T\right) - \frac{s}{4}\left((\nabla F(z^*))^2 + ((\nabla F(z^*))^2)^T\right) \\
&\quad - \frac{s}{4}\left(\nabla^3 F(z^*)F(z^*) + (\nabla^3 F(z^*)F(z^*))^T\right) \\
&= -\frac{1}{2}\left(\nabla F(z^*) + (\nabla F(z^*))^T\right) - \frac{s}{4}\left((\nabla F(z^*))^2 + ((\nabla F(z^*))^2)^T\right) \\
&= \begin{bmatrix} -A - \frac{s}{2}\left(A^2 - B^T B\right) & \\ & -C - \frac{s}{2}\left(C^2 - BB^T\right) \end{bmatrix} \prec 0\,,
\end{aligned}$$

where the second equality is from $F(z^*) = 0$, the third equality uses the definition of $\nabla F(z^*)$ by noticing $\nabla F(z^*)$ and $(\nabla F(z^*))^2$ are both generalized block skew-symmetric, the first inequality utilizes $\|A\|, \|-C\| \leq \|F(z^*)\| \leq 1/s$, and the last inequality is due to (15). Applying Propositions 5 and 6 with $P = I$ finishes the proof.

3. Similar to 2., we have for EGM (11) that

$$\begin{aligned}\frac{1}{2}\left(\nabla G(z^*)^T + \nabla G(z^*)\right) &= -\frac{1}{2}\left(\nabla F(z^*) + (\nabla F(z^*))^T\right) + \frac{s}{4}\left((\nabla F(z^*))^2 + ((\nabla F(z^*))^2)^T\right) \\ &\quad + \frac{s}{4}\left(\nabla^3 F(z^*)F(z^*) + (\nabla^3 F(z^*)F(z^*))^T\right) \\ &= -\frac{1}{2}\left(\nabla F(z^*) + (\nabla F(z^*))^T\right) + \frac{s}{4}\left((\nabla F(z^*))^2 + ((\nabla F(z^*))^2)^T\right) \\ &= \begin{bmatrix} -A + \frac{s}{2}\left(A^2 - B^T B\right) & \\ & -C + \frac{s}{2}\left(C^2 - BB^T\right)\end{bmatrix} \prec 0\,,\end{aligned}$$

where the last inequality is due to (16). Applying Propositions 5 and 6 with $P = I$ finishes the proof.

4. For Alternating Gradient Descent Ascent (12), we take $P = \begin{bmatrix} I & \frac{1}{2}sB^T \\ \frac{1}{2}sB & I\end{bmatrix}$, and consider $M = \frac{1}{2}\left(P\nabla G(z^*) + \nabla G(z^*)^T P\right)$. Then M is symmetric and

$$M_{xx} = -A - \frac{s}{2}A^2 - \frac{s^2}{4}\left(AB^T B + B^T BA\right),$$

$$M_{yy} = -C - \frac{s}{2}C^2 - \frac{s^2}{4}\left(CBB^T + BB^T C\right),$$

$$M_{xy} = -\frac{s}{2}\left(BA + CB\right) - \frac{s^2}{4}\left(BA^2 + C^2 B\right).$$

Applying Propositions 5 and 6 finishes the proof.

B.5. Proof of Proposition 8

Notice that $L(x, y)$ is fifth-order continuously differentiable around z^*, thus there exists δ such that $\nabla^j F(z)$ exists and bounded for any $z \in B(z^*, \delta) := \{z | \|z - z^*\| \leq \delta\}$. Corollary 2 in Lu (2020) shows that under the condition of Proposition 8, the $O(s)$-resolution ODEs of EGM and GDA are gradient-based in $B(z^*, \delta)$. We here just show that the $O(s)$-resolution ODE of AGDA is gradient-based. Recall the definition of g_1, g_2, g_3 in (25). Notice that under the condition of the proposition, we have $\|g_j(z)\| \leq O(\|F(z)\|)$ and $\|\nabla^k g_j(z)\| \leq O(1)$ for $j = 1, 2, 3$, $k = 1, ..., 5 - j$ and $z \in B(z^*, \delta)$. Thus it follows from Theorem 3 in Lu (2020) that the $O(s)$-resolution ODE of AGDA is also gradient-based in $B(z^*, \delta)$.

B.6. Proof of Theorem 9

Throughout this proof, the norm we use is P norm. Suppose the $O(s)$-resolution ODE is gradient based in $B(z^*, \delta')$. Suppose $\|z - z^*\| \leq \delta := \min\{\frac{bs\sqrt{P}}{H\bar{P}}, 1, \delta'\}$. Let $z(s)$ be the solution of the $O(s)$-resolution ODE at $t = s$ from initial solution $z(0) = z$ and $z^+ = g(z, s)$. Then it follows by the proof of Proposition 5 that $\|z(s) - z^*\|_P \leq e^{-\frac{1}{2}bs^2/\bar{P}}\|z - z^*\|_P$. Moreover, it follows from the definition of gradient-based $O(s)$-resolution ODE that there exists $c > 0$ such that

$\|z(s) - z^+\| \le cs^3\|F(z)\|$ for $z \in B(z^*, \delta)$. Now notice $F(z)$ is Lipschtiz continuous in the close and bounded set $B(z^*, \delta)$, i.e., there exists $\gamma > 0$ such that $\|F(z)\| = \|F(z) - F(z^*)\| \le \gamma\|z - z^*\|$. Putting everything together and letting $s^* = \frac{b\sqrt{P}}{8c\gamma\bar{P}^2}$, we have for any $s \le s^*$ that

$$\begin{aligned}
\|z^+ - z^*\|_P &\le \|z(s) - z^*\|_P + \|z^+ - z(s)\|_P \\
&\le e^{-\frac{1}{2}bs^2/\bar{P}}\|z - z^*\|_P + \bar{P}\|z^+ - z(s)\| \\
&\le e^{-\frac{1}{2}bs^2/\bar{P}}\|z - z^*\|_P + \bar{P}cs^3\|F(z)\| \\
&\le \left(1 - \frac{1}{4}bs^2/\bar{P}\right)\|z - z^*\|_P + \bar{P}cs^3\gamma\|z - z^*\| \\
&\le \left(1 - \frac{1}{4}bs^2/\bar{P}\right)\|z - z^*\|_P + \frac{\bar{P}}{\sqrt{P}}cs^3\gamma\|z - z^*\|_P \\
&\le \left(1 - \frac{1}{8}\frac{bs^2}{\bar{P}}\right)\|z - z^*\|_P ,
\end{aligned}$$

where the second inequality utilizes Proposition 5 and the definition of $\bar{P}$, the third inequality utilizes (20), the fourth inequality utilizes $\|F(z)\| = \|F(z) - F(z^*)\| \le \gamma\|z - z^*\|$ and the last inequality is because $s \le s^*$. This proves the theorem by telescoping.

B.7. Proof of Proposition 10

For GDA, we have $G(z) = -\left(I + \frac{s}{2}\nabla F(z)\right)F(z)$. Notice for small enough s, it holds that $\|\frac{s}{2}\nabla F(z)\| \le \frac{1}{2}$, whereby $I + \frac{s}{2}\nabla F(z)$ is a full-rank matrix. Therefore, $G(z) = 0$ is equivalent to $F(z) = 0$, which finishes the proof for GDA. The proof for AGDA and EGM follow the same process.

B.8. Proof of Theorem 1

Recall the dynamic of GDA are given by

$$\dot{z} = -F(z) - \frac{s}{2}\nabla F(z)F(z) .$$

Further, the objective of this Guassian family of GAN training problems is given by

$$L(g, d) = \mathbb{E}_{s\sim N(0,I)}\left[\log\left(\frac{1}{1 + e^{d^T\Sigma^{1/2}s}}\right) + \log\left(1 - \frac{1}{1 + e^{d^T(\Sigma'^{1/2}s+g)}}\right)\right] + \frac{\alpha}{2}\|g\|^2 - \frac{\alpha}{2}\|d\|^2$$

which has gradient given by

$$F(g, d) = \begin{bmatrix} \nabla_g L(g, d) \\ -\nabla_d L(g, d) \end{bmatrix} = \begin{bmatrix} \mathbb{E}_{s\sim N(0,I)}\left[\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}d\right] + \alpha g \\ \mathbb{E}_{s\sim N(0,I)}\left[\frac{e^{d^T\Sigma^{1/2}s}}{1+e^{d^T\Sigma^{1/2}s}}\Sigma^{1/2}s - \frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}(\Sigma'^{1/2}s + g)\right] + \alpha d \end{bmatrix} .$$

Evaluating this at $(g, d) = (0, 0)$ verifies that the origin is stationary for any choice of $\alpha \in \mathbb{R}$ as

$$F(0, 0) = \begin{bmatrix} 0 \\ \mathbb{E}_{s\sim N(0,I)}\left[\frac{e^0}{1+e^0}\Sigma^{1/2}s - \frac{1}{1+e^0}\Sigma'^{1/2}s\right] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where the last equality relies on our consideration of zero mean data distributions. Next we evaluate the Jacobian of $F(g,d)$ to understand the related continuous dynamics giving $\nabla_g F(g,d)$ as

$$\begin{bmatrix} \mathbb{E}_{s\sim N(0,I)}\left[\frac{-e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}dd^T\right]+\alpha I \\ \mathbb{E}_{s\sim N(0,I)}\left[\frac{e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}(\Sigma'^{1/2}s+g)d^T-\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}I\right] \end{bmatrix}$$

and $\nabla_d F(g,d)$ given by

$$\begin{bmatrix} \mathbb{E}_{s\sim N(0,I)}\left[\frac{-e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}d(\Sigma'^{1/2}s+g)^T+\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}I\right] \\ \mathbb{E}_{s\sim N(0,I)}\left[\frac{e^{d^T\Sigma^{1/2}s}}{(1+e^{d^T\Sigma^{1/2}s})^2}\Sigma^{1/2}ss^T\Sigma^{1/2}-\frac{e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}(\Sigma'^{1/2}s+g)(\Sigma'^{1/2}s+g)^T\right]+\alpha I \end{bmatrix}$$

Consequently, the Gradient Flow (GF) dynamics at the origin have

$$-\nabla F(0,0)=\begin{bmatrix}-\alpha I & \frac{1}{2}I \\ \frac{-1}{2}I & \frac{1}{4}(\Sigma-\Sigma')-\alpha I\end{bmatrix}.$$

The dynamics of GDA are given by $\dot{g}$ equal to

$$\begin{aligned}
&\mathbb{E}_{s\sim N(0,I)}\left[\frac{-1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}d\right]-\alpha g \\
&+\frac{s}{2}\Bigg(\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{-e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}dd^T\right]+\alpha I\right)\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{-1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}d\right]-\alpha g\right) \\
&\qquad+\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{-e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}d(\Sigma'^{1/2}s+g)^T+\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}I\right]\right) \\
&\qquad\cdot\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{-e^{d^T\Sigma^{1/2}s}}{1+e^{d^T\Sigma^{1/2}s}}\Sigma^{1/2}s+\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}(\Sigma'^{1/2}s+g)\right]-\alpha d\right)\Bigg)
\end{aligned}$$

and $\dot{d}$ equal to

$$\begin{aligned}
&\mathbb{E}_{s\sim N(0,I)}\left[\frac{-e^{d^T\Sigma^{1/2}s}}{1+e^{d^T\Sigma^{1/2}s}}\Sigma^{1/2}s+\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}(\Sigma'^{1/2}s+g)\right]-\alpha d \\
&+\frac{s}{2}\Bigg(\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}(\Sigma'^{1/2}s+g)d^T-\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}I\right]\right) \\
&\qquad\cdot\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{-1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}d\right]-\alpha g\right) \\
&\qquad+\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{e^{d^T\Sigma^{1/2}s}}{(1+e^{d^T\Sigma^{1/2}s})^2}\Sigma^{1/2}ss^T\Sigma^{1/2}-\frac{e^{d^T(\Sigma'^{1/2}s+g)}}{(1+e^{d^T(\Sigma'^{1/2}s+g)})^2}(\Sigma'^{1/2}s+g)(\Sigma'^{1/2}s+g)^T\right]+\alpha I\right) \\
&\qquad\cdot\left(\mathbb{E}_{s\sim N(0,I)}\left[\frac{-e^{d^T\Sigma^{1/2}s}}{1+e^{d^T\Sigma^{1/2}s}}\Sigma^{1/2}s+\frac{1}{1+e^{d^T(\Sigma'^{1/2}s+g)}}(\Sigma'^{1/2}s+g)\right]-\alpha d\right)\Bigg).
\end{aligned}$$

Taking the Jacobian of these more complicated GDA dynamics at $(g, d) = (0, 0)$ provides an $O(s)$ correction to the GF dynamics above as $\nabla G(0, 0)$ is the following

$$\begin{bmatrix} -\left(\alpha + \frac{s}{2}\left(\alpha^2 - \frac{1}{4}\right)\right) I & \frac{1}{2}I + \frac{s}{16}\Sigma' \\ \frac{-1}{2}I - \frac{s}{16}\Sigma' & \frac{1}{4}(\Sigma - \Sigma') - \alpha I + \frac{s}{2}\left(\frac{1}{4}I + (\frac{1}{4}(\Sigma - \Sigma') + \alpha I)(\frac{1}{4}(\Sigma - \Sigma') - \alpha I)\right) \end{bmatrix}.$$

Next we verify that the origin is not attractive for $\alpha = 0$. This is immediate as the Jacobian above simplifies to the following very structured block-matrix

$$\nabla G(0, 0) = \begin{bmatrix} \frac{s}{8}I & \frac{1}{2}I + \frac{s}{16}\Sigma' \\ \frac{-1}{2}I - \frac{s}{16}\Sigma' & \frac{1}{4}(\Sigma - \Sigma') + \frac{s}{2}\left(\frac{1}{4}I + \frac{1}{16}(\Sigma - \Sigma')^2\right) \end{bmatrix}.$$

Since we assume $\Sigma \succeq \Sigma'$, the diagonal blocks on this matrix are both positive definite and so this matrix has at least one eigenvalue with positive real part.

Finally, we verify that the origin transitions to being attractive for all sufficiently large α. This follows from observing that the diagonal entries are all decreasing $O(-\alpha^2)$ while the off-diagonal entries have magnitude $O(\alpha)$. Then applying Gershgorin's circle theorem guarantees that once the diagonal is sufficiently negative, all of the eigenvalues of GDA's dynamics must be negative (and thus the origin becomes a linear attractor).

Appendix C. Calculation of First Lyapunov Coefficients

We give a proposition providing formulas for $l_1(0)$ for symmetric bilinear problems. Using this formula, we can compute the first Lyapunov coefficient for both polynomial examples considered in Section 4, verifying the phase transitions observed are supercritical and subcritical respectively.

Lemma 11 *Consider any symmetric and bilinear minimax problem* $\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} L(x, y, \alpha) = f(x) + \alpha xy - f(y)$ *with a stationary point at* $(x, y) = (0, 0)$. *Then the dynamics of GF* (7) *at the origin have First Lyapunov Coefficient given by*

$$w = \alpha,$$
$$l_1(0) = -\frac{f^{(4)}(0)}{4\alpha}.$$

The dynamics of GDA (10) *at the origin have First Lyapunov Coefficient given by*

$$w = \alpha(1 + sf^{(2)}(0)),$$
$$l_1(0) = -\frac{f^{(4)}(0)}{4\alpha(1 + sf^{(2)}(0))} - s\frac{4f^{(4)}(0)f^{(2)}(0) + 3f^{(3)}(0)^2}{8\alpha(1 + sf^{(2)}(0))} + \frac{2sf^{(3)}(0)^2 + 3s^2f^{(3)}(0)^2f^{(2)}(0)}{16\alpha(1 + sf^{(2)}(0))^2}.$$

The dynamics of EG (11) *at the origin have First Lyapunov Coefficient given by*

$$w = \alpha(1 - sf^{(2)}(0)),$$
$$l_1(0) = -\frac{f^{(4)}(0)}{4\alpha(1 - sf^{(2)}(0))} + s\frac{4f^{(4)}(0)f^{(2)}(0) + 3f^{(3)}(0)^2}{8\alpha(1 - sf^{(2)}(0))} - \frac{2sf^{(3)}(0)^2 - 3s^2f^{(3)}(0)^2f^{(2)}(0)}{16\alpha(1 - sf^{(2)}(0))^2}.$$

Proof First, we observe that for any problem with $L(x, y, 0) = f(x) + axy - f(y)$, each of our algorithms give dynamics of the type (21). The symmetry between x and y in the objective function and in EGM and GDA ensures that whenever the trace of $\nabla G(z)$ is zero, the two diagonal entries of $\nabla G(z)$ must be zero.

The dynamics of GF (7) for this class of symmetric problems are given by

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -f'(x) - \alpha y \\ \alpha x - f'(y) \end{bmatrix}$$

From this, the following derivatives in terms of the form (21) are given by

$$\begin{aligned} w = Q_x &= \alpha, \\ P_{xx} = Q_{yy} &= -f^{(3)}(0), \\ P_{xy} = P_{yy} = Q_{xy} = Q_{xx} &= 0, \\ P_{xxx} = Q_{yyy} &= -f^{(4)}(0), \\ P_{xyy} = Q_{xxy} &= 0. \end{aligned}$$

Then the claimed formula follows from (22).

The dynamics of GDA (10) for this class of symmetric problems are given by

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -f'(x) - \alpha y \\ \alpha x - f'(y) \end{bmatrix} - \frac{s}{2} \begin{bmatrix} f^{(2)}(x)(f'(x) + \alpha y) - \alpha(\alpha x - f'(y)) \\ \alpha(f'(x) + \alpha y) + f^{(2)}(y)(\alpha x - f'(y)) \end{bmatrix}$$

From this, the following derivatives in terms of the form (21) are given by

$$\begin{aligned} w = Q_x &= a + s\alpha f^{(2)}(0), \\ P_{xx} = Q_{yy} &= -f^{(3)}(0) - \frac{3s f^{(3)}(0) f^{(2)}(0)}{2}, \\ P_{yy} = -Q_{xx} &= \frac{-s\alpha f^{(3)}(0)}{2}, \\ P_{xy} = Q_{xy} &= 0, \\ P_{xxx} = Q_{yyy} &= -f^{(4)}(0) - \frac{s}{2}\left(4f^{(4)}(0)f^{(2)}(0) + 3f^{(3)}(0)^2\right), \\ P_{xyy} = Q_{xxy} &= 0. \end{aligned}$$

Then the claimed formula follows from (22).

The dynamics of EGM (11) for this class of symmetric problems are given by

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -f'(x) - \alpha y \\ \alpha x - f'(y) \end{bmatrix} + \frac{s}{2} \begin{bmatrix} f^{(2)}(x)(f'(x) + \alpha y) - \alpha(\alpha x - f'(y)) \\ \alpha(f'(x) + \alpha y) + f^{(2)}(y)(\alpha x - f'(y)) \end{bmatrix}$$

From this, the following derivatives in terms of the form (21) are given by

$$\begin{aligned} w = Q_x &= a - s\alpha f^{(2)}(0), \\ P_{xx} = Q_{yy} &= -f^{(3)}(0) + \frac{3sf^{(3)}(0)f^{(2)}(0)}{2}, \\ P_{yy} = -Q_{xx} &= \frac{s\alpha f^{(3)}(0)}{2}, \\ P_{xy} = Q_{xy} &= 0, \\ P_{xxx} = Q_{yyy} &= -f^{(4)}(0) + \frac{s}{2}\left(4f^{(4)}(0)f^{(2)}(0) + 3f^{(3)}(0)^2\right), \\ P_{xyy} = Q_{xxy} &= 0. \end{aligned}$$

Then the claimed formula follows from (22). ■

For the example (23), the function $f(x) = (x+3)(x+1)(x-1)(x-3)$ has

$$\begin{aligned} f^{(4)}(0) &= 24 \\ f^{(3)}(0) &= 0 \\ f^{(2)}(0) &= -20 \\ f^{(1)}(0) &= 0. \end{aligned}$$

Plugging this into Lemma 11 yields the EGM ODE dynamics at $\alpha^* = \sqrt{20400}$ having

$$\begin{aligned} w &= \alpha(1 - sf^{(2)}(0)) = \sqrt{20400}(1 - 0.002 \times -20) \approx 148.542, \\ l_1(0) &= -\frac{f^{(4)}(0)}{4\alpha(1 - sf^{(2)}(0))} + s\frac{4f^{(4)}(0)f^{(2)}(0) + 3f^{(3)}(0)^2}{8\alpha(1 - sf^{(2)}(0))} - \frac{2sf^{(3)}(0)^2 - 3s^2f^{(3)}(0)^2f^{(2)}(0)}{16\alpha(1 - sf^{(2)}(0))^2} \\ &\approx -0.04362. \end{aligned}$$

Likewise, for (24), the function $g(x) = -(x+3)(x+1)(x-1)(x-3)$ has

$$\begin{aligned} f^{(4)}(0) &= -24 \\ f^{(3)}(0) &= 0 \\ f^{(2)}(0) &= 20 \\ f^{(1)}(0) &= 0. \end{aligned}$$

Plugging this into Lemma 11 yields the GDA ODE dynamics at $\alpha^* = \sqrt{20400}$ having

$$\begin{aligned} w &= \alpha(1 + sf^{(2)}(0)) = \sqrt{20400}(1 + 0.002 \times 20) \approx 148.542, \\ l_1(0) &= -\frac{f^{(4)}(0)}{4\alpha(1 + sf^{(2)}(0))} - s\frac{4f^{(4)}(0)f^{(2)}(0) + 3f^{(3)}(0)^2}{8\alpha(1 + sf^{(2)}(0))} + \frac{2sf^{(3)}(0)^2 + 3s^2f^{(3)}(0)^2f^{(2)}(0)}{16\alpha(1 + sf^{(2)}(0))^2} \\ &\approx 0.04362. \end{aligned}$$