Towards Understanding the Dynamics of Generative Adversarial Networks

Jerry Li [1] **Aleksander Madry** [1] **John Peebles** [1] **Ludwig Schmidt** [1]

Abstract

Generative Adversarial Networks (GANs) have recently been proposed as a promising avenue towards learning generative models with deep neural networks. While GANs have demonstrated state-of-the-art performance on multiple vision tasks, their learning dynamics are not yet well understood, both in theory and in practice. To address this issue, we take a first step towards a rigorous study of GAN dynamics. We propose a simple model that exhibits several of the problematic convergence behaviors (e.g., vanishing gradient, mode collapse, diverging or oscillatory behavior) and allows us to still establish the first convergence bounds for parametric GAN dynamics. We find an interesting dichotomy: a GAN with an optimal discriminator provably converges, while a first order approximation of the discriminator steps leads to unstable GAN dynamics and mode collapse. Our model and analysis point to a specific challenge in practical GAN training that we call *discriminator collapse*.

1. Introduction

Generative modeling is a fundamental learning task of growing importance. As we apply machine learning to increasingly sophisticated problems, we often aim to learn functions with an output domain that is far more complex than simple class labels. Common examples include image "translation" (Isola et al., 2017), speech synthesis (van den Oord et al., 2016), and robot trajectory prediction (Finn et al., 2016). Due to progress in deep learning, we now have access to powerful architectures that can represent generative models over such complex domains. However, training these generative models is a key challenge. Simpler learning problems such as classification have a clear notion of "right" and "wrong," and the approaches based on minimizing the corresponding loss functions have been tremendously successful. In contrast, training a generative model is far more nuanced because it is often unclear how "good" a sample from the model is.

Generative Adversarial Networks (GANs) have recently been proposed to address this issue (Goodfellow et al., 2014). In a nutshell, the key idea of GANs is to learn *both* the generative model and the loss function at the same time. The resulting training dynamics are usually described as a game between a generator (the generative model) and a discriminator (the loss function). The goal of the generator is to produce realistic samples that fool the discriminator, while the discriminator is trained to distinguish between the true training data and samples from the generator. GANs have shown promising results on a variety of tasks, and there is now a large body of work that explores the power of this framework (Goodfellow, 2017).

Unfortunately, training GANs turns out to be a challenging problem that often hinders further research in this area. Practitioners have encountered a variety of obstacles such as vanishing gradients, mode collapse, and diverging or oscillatory behavior (Goodfellow, 2017). At the same time, the theoretical underpinnings of GAN dynamics are not yet well understood. To date, there were no convergence proofs for GAN models even in very simple settings. A key challenge is that the overall loss function driving the GAN dynamics is usually non-convex. As a result, the root cause of failing GAN dynamics is often unclear.

In this paper, we take a first step towards a principled understanding of GAN *dynamics*. We propose and examine a problem setup that exhibits all common failure cases of GAN dynamics while remaining sufficiently simple to allow for a rigorous analysis. Concretely, we introduce and study the GMM-GAN: a variant of GAN dynamics that captures learning a mixture of two univariate Gaussians. We first show experimentally that standard gradient dynamics of the GMM-GAN often fail to converge due to mode collapse or oscillatory behavior. Interestingly, this also holds for techniques that were recently proposed to improve GAN training such as unrolled GANs (Metz et al., 2017). In contrast to these findings, we then show that GAN dynamics with an *optimal* discriminator *do* converge, both experimentally and *provably*. To the best of our knowledge, our theoretical analysis of the GMM-GAN is the first global convergence proof for concrete and non-trivial GAN dynamics.

Our results show a clear dichotomy between the GAN dynamics arising from applying simultaneous gradient descent and the one that is able to use an optimal discriminator. The GAN with optimal discriminator converges from (essentially) *any* starting point. On the other hand, the simul-

taneous gradient GAN often fails to converge, even when the discriminator is allowed many more gradient update steps than the generator. Importantly, our model allows us to understand the root cause of this dichotomy, which we call *discriminator collapse*. In regions where the generator is fooling the discriminator well, the gradient updates for the discriminator are stuck in a local minimum. As a result, the discriminator loses much of its capacity and is thus unable to properly guide the GAN dynamics, which then fails to converge to a good solution as a result. In particular, these findings go against the common wisdom that first order methods are sufficiently strong for all deep learning applications. We conjecture that this wisdom might not be correct in the context of saddle point problems underlying GANs, and that discriminator collapse is also an important issue in more complex GANs. We also believe that studying simple models like our GMM-GAN in more detail will also lead to better methods for training more general GANs.

Paper outline. In Section 2, we introduce our framework for studying GAN dynamics and describe some of its properties. Section 3 states our main theoretical results and describes the discriminator collapse phenomenon in more detail. Section 4 then gives an overview of our convergence proof. In Section 5, we provide experiments. We defer the proof details to the supplementary material.

2. Generative Adversarial Dynamics

Generative adversarial networks are commonly described as a two player game (Goodfellow et al., 2014). Given a true distribution P, a set of generators $\mathcal{G} = \{G_u, u \in \mathcal{U}\}$, a set of discriminators $\mathcal{D} = \{D_v, v \in \mathcal{V}\}$, and a monotone measuring function $m : \mathbb{R} \to \mathbb{R}$, the objective of GAN training is to find a generator u in

$$\arg\min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} \mathbb{E}_{x\sim P}[m(D_v(x))]+\mathbb{E}_{x\sim G_u}[m(1-D_v(x))] \ . \tag{1}$$

In other words, the game is between two players called the generator and discriminator, respectively. The goal of the discriminator is to distinguish between samples from the generator and the true distribution. The goal of the generator is to fool the discriminator by generating samples that are similar to the data distribution.

By varying the choice of the measuring function and the set of discriminators, one can capture a wide variety of loss functions. Typical choices that have been previously studied include the KL divergence and the Wasserstein distance (Goodfellow et al., 2014; Arjovsky et al., 2017). Moreover, this formulation can also encode other common objectives: most notably, as we will show later, the total variation distance.

To optimize the objective (1), the most common approaches are variants of simultaneous gradient descent on the generator u and the discriminator v. But despite its attractive theoretical grounding, GAN training is plagued by a variety of issues in practice. Two major problems are *mode collapse* and *vanishing gradients*. Mode collapse corresponds to situations in which the generator only learns a subset (a few modes) of the true distribution P (Goodfellow, 2017; Arora & Zhang, 2017). For instance, a GAN trained on an image modeling task would only produce variations of a small number of images. Vanishing gradients (Arjovsky et al., 2017; Arjovsky & Bottou, 2017; Arora et al., 2017) are, on the other hand, a failure case where the generator updates become vanishingly small, thus making the GAN dynamics not converge to a satisfying solution. Despite many proposed explanations and approaches to solve the vanishing gradient problem, it is still often observed in practice (Goodfellow, 2017).

2.1. Towards a principled understanding of GAN dynamics

GANs provide a powerful framework for generative modeling. However, there is a fundamental gap between these theoretical explanations and practice. Specifically, to the best of the authors' knowledge, all theoretical studies of GAN dynamics for parametric models simply consider global optima and stationary points of the dynamics, and there has been no rigorous study of the actual GAN dynamics. In practice, GANs are always optimized using first order methods, and the current theory of GANs cannot tell us whether or not these methods converge to a meaningful solution. This raises a natural question, also posed as an open problem in (Goodfellow, 2017):

Can we understand the convergence behavior of GANs?

This problem has proved difficult to understand, because of the non-convexity of the objective function, and of the generator and discriminator sets. Even simpler dynamics, for instance, when we assume we have access to optimal discriminators, but still take gradient steps for the generator, have proven challenging to understand. In fact, there is no consensus whether or not optimal discriminator steps improve the convergence behavior.

In this work, we want to change this state of affairs and initiate the study of GAN dynamics from an algorithmic perspective. Specifically, as a first step towards crystallizing a more general picture, we define a principled model that captures many of the difficulties of real-world GANs but is still simple enough to make a full analysis tractable. We present it below.

2.2. A Simple Model of Generative Adversarial Dynamics

Perhaps a tempting starting place for coming up with a simple but meaningful set of GAN dynamics is to consider the generators being univariate Gaussians with fixed variance. Indeed, in the supplementary material we give a short proof that simple GAN dynamics always converge for this class of generators. However, it seems that this class of distributions is insufficiently expressive to exhibit many of the interesting phenomena such as mode collapse mentioned above. In particular, the distributions in this class are all unimodal, and it is unclear what mode collapse would even mean in this context.

Generators. The above considerations motivate us to make our model slightly more complicated. That is, to assume that the true distribution and the generator distributions are all mixtures of two univariate Gaussians with unit variance, and uniform mixing weights. Formally, our generator set is $\mathcal{G}$, where

$$\mathcal{G} = \left\{ \frac{1}{2}\mathcal{N}(\mu_1, 1) + \frac{1}{2}\mathcal{N}(\mu_2, 1) \mid \mu_1, \mu_2 \in \mathbb{R} \right\} . \quad (2)$$

For any $\mu \in \mathbb{R}^2$, we let $G_\mu(x)$ denote the distribution in $\mathcal{G}$ with means at μ_1 and μ_2. While this is a simple change compared to a single Gaussian case, it makes a large difference in the behavior of the dynamics. In particular, many of the pathologies present in real-world GAN training begin to appear.

Loss function. While GANs are usually viewed as a generative framework, they can also be viewed as a general method for density estimation. We want to set up learning an unknown generator $G_{\mu^*} \in \mathcal{G}$ as a generative adversarial dynamics. To this end, we must first define the loss function for the density estimation problem. A well-studied goal in this setting is to recover $G_{\mu^*}(x)$ in total variation (also known as L^1 or statistical) distance, where the total variation distance between two distributions P, Q is defined as

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2}\int_\Omega |P(x) - Q(x)| dx = \max_A P(A) - Q(A) , \quad (3)$$

where the maximum is taken over all measurable events A.

Such finding the best-fit distribution in total variation distance can indeed be naturally phrased as generative adversarial dynamics. Unfortunately, for arbitrary distributions, this is algorithmically problematic, simply because the set of discriminators one would need is intractable to optimize over.

However, for distributions that are structurally simple, like mixtures of Gaussians, it turns out we can consider a much simpler set of discriminators. In Appendix A.1 in the supplementary material, by using well-known connections to VC theory, we show that for two generators $G_{\mu_1}, G_{\mu_2} \in \mathcal{G}$, we have

$$d_{\mathrm{TV}}(G_{\mu_1}, G_{\mu_2}) = \max_{E = I_1 \cup I_2} G_{\mu_1}(E) - G_{\mu_2}(E) , \quad (4)$$

where the maxima is taken over two disjoint intervals $I_1, I_2 \subseteq \mathbb{R}$. In other words, instead of considering the difference of measure between the two generators G_{μ_1}, G_{μ_2} on arbitrary events, we may restrict our attention to unions of two disjoint intervals in $\mathbb{R}$. This is a special case of a well-studied distance measure for structured distributions, known as the $\mathcal{A}_k$-distance, for $k = 2$ (Devroye & Lugosi, 2012; Chan et al., 2014). Moreover, this class of subsets has a simple parametric description.

Discriminators. Now, the above discussion motivates our definition of discriminators to be

$$\mathcal{D} = \{ \mathbb{I}_{[\ell_1, r_1]} + \mathbb{I}_{[\ell_2, r_2]} \mid \ell, r \in \mathbb{R}^2 \text{ s.t. } \ell_1 \le r_1 \le \ell_2 \le r_2 \} . \quad (5)$$

In other words, the set of discriminators is taken to be the set of indicator functions of sets which can be expressed as a union of at most two disjoint intervals. With this definition, we have that the finding the best fit in total variation distance to some unknown $G_{\mu^*} \in \mathcal{G}$ is equivalent to finding $\widehat{\mu}$ minimizing

$$\widehat{\mu} = \arg\min_\mu \max_{\ell, r} L(\mu, \ell, r) \text{ , where}$$
$$L(\mu, \ell, r) = \mathbb{E}_{x \sim G_{\mu^*}}[D(x)] + \mathbb{E}_{x \sim G_\mu}[1 - D(x)] \quad (6)$$

can be shown to be a simple, smooth function of all three parameters (see the supplementary material for details).

Dynamics. The objective in (6) is easily amenable to optimization at parameter level. A natural approach for optimizing this function would be to define $G(\widehat{\mu}) = \max_{\ell, r} L(\widehat{\mu}, \ell, r)$, and to perform (stochastic) gradient descent on this function. This corresponds to, at each step, finding the the optimal discriminator, and updating the current $\widehat{\mu}$ in that direction. We call these dynamics the *optimal discriminator dynamics*. Formally, given $\widehat{\mu}^{(0)}$ and a stepsize η_g, and a true distribution $G_{\mu^*} \in \mathcal{G}$, the optimal discriminator dynamics for $G_{\mu^*}, \mathcal{G}, \mathcal{D}$ starting at $\widehat{\mu}^{(0)}$ are given iteratively as

$$\ell^{(t)}, r^{(t)} = \arg\max_{\ell, r} L(\widehat{\mu}^{(t)}, \ell, r) , \quad (7)$$
$$\widehat{\mu}^{(t+1)} = \widehat{\mu}^{(t)} - \eta_g \nabla_\mu L(\widehat{\mu}^{(t)}, \ell^{(t)}, r^{(t)}) , \quad (8)$$

where the maximum is taken over ℓ, r which induce two disjoint intervals.

For more complicated generators and discriminators such as neural networks, these dynamics are computationally difficult to perform. Therefore, instead of the updates as in (8), one uses simultaneous gradient iterations on the generator and discriminator. These dynamics are called the *first order dynamics*. Formally, given $\widehat{\mu}^{(0)}, \ell^{(0)}, r^{(0)}$ and a stepsize η_g, η_d, and a true distribution $G_{\mu^*} \in \mathcal{G}$, the first order dynamics for $G_{\mu^*}, \mathcal{G}, \mathcal{D}$ starting at $\widehat{\mu}^{(0)}$ are given by

$$\widehat{\mu}^{(t+1)} = \widehat{\mu}^{(t)} - \eta_g \nabla_\mu L(\widehat{\mu}^{(t)}, \ell^{(t)}, r^{(t)}) \tag{9}$$

$$r^{(t+1)} = r^{(t)} + \eta_d \nabla_r L(\widehat{\mu}^{(t)}, \ell^{(t)}, r^{(t)}) \,, \tag{10}$$

$$\ell^{(t+1)} = \ell^{(t)} + \eta_d \nabla_\ell L(\widehat{\mu}^{(t)}, \ell^{(t)}, r^{(t)}) \,. \tag{11}$$

Even for our relatively simple setting, the first order dynamics can exhibit a variety of behaviors, depending on the starting conditions of the generators and discriminators. In particular, in Figure 1, we see that depending on the initialization, the dynamics can either converge to optimality, exhibit a primitive form of mode collapse, where the two generators collapse into a single generator, or converge to the wrong value, because the gradients vanish. This provides empirical justification for our model, and shows that these dynamics are complicated enough to model the complex behaviors which real-world GANs exhibit. Moreover, as we show in Section 5 below, these behaviors are not just due to very specific pathological initial conditions: indeed, when given random initial conditions, the first order dynamics still more often than not fail to converge.

3. Optimal Discriminator vs. First Order

The difference between the optimal discriminator dynamics and the first order dynamics is an important question in GAN training. While to the best of our knowledge, nobody has rigorously analyzed either set of dynamics, the question of whether or not training the discriminator to optimality is the right approach has received considerable attention. The optimal discriminator dynamics are in general algorithmically infeasible and it is not clear to what extent these first order methods can hope to match the performance of the optimal discriminator. Somewhat orthogonal to these computational issues, there is also theoretical evidence (Arjovsky & Bottou, 2017) that using the optimal discriminator at each step may not even be desirable in certain settings, though their setting is different from than we consider. All these considerations lead us to the question:

Can we understand the impact of using optimal discriminators on convergence, and how do these dynamics differ from the first order dynamics?

Our main theoretical result is[1]:

Theorem 3.1. *Fix $\delta > 0$ sufficiently small and $C > 0$. Let $\mu^* \in \mathbb{R}^2$ so that $|\mu_i^*| \leq C$, and $|\mu_1^* - \mu_2^*| \geq \delta$. Then, for all initial points $\widehat{\mu}^{(0)}$ so that $|\widehat{\mu}_i^{(0)}| \leq C$ for all i and so that $|\widehat{\mu}_1^{(0)} - \widehat{\mu}_2^{(0)}| \geq \delta$, if we let $\eta = \mathrm{poly}(1/\delta, e^{-C^2})$ and $T = \mathrm{poly}(1/\delta, e^{-C^2})$, then if $\widehat{\mu}^{(T)}$ is specified by the optimal discriminator dynamics, we have $d_{\mathrm{TV}}(G_{\mu^*}, G_{\widehat{\mu}^{(T)}}) \leq \delta$.*

In other words, if the μ^* are bounded by a constant, and not too close together, then in time which is polynomial in the inverse of the desired accuracy δ and e^{-C^2}, where C is a bound on how far apart the μ^* and $\widehat{\mu}$ are, the optimal discriminator dynamics converge to the ground truth in total variation distance. The dependence on e^{-C^2} is necessary, as if the $\widehat{\mu}$ and μ^* are initially very far apart, then the initial gradients for $\widehat{\mu}$ will necessarily be of this scale as well.

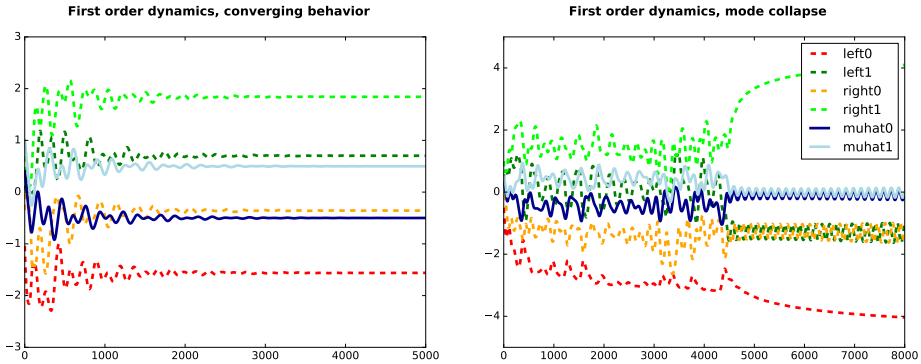
On the other hand, we provide simulations that demonstrate that first order updates, or more complicated heuristics such as unrolling, all fail to consistently converge to the true distribution, even under the same sorts of conditions as in Theorem 3.1. In Figure 1, we gave some specific examples where the first order dynamics fail to converge. In Section 5 we show that this sort of divergence is common, even with random initializations for the discriminators. In particular, the probability of convergence is generally much lower than 1, for both the regular GAN dynamics, and unrolling. In general, we believe that this phenomena should occur for *any* natural first order dynamics. In particular, one barrier we observed for any such dynamics is something we call *discriminator collapse*, that we describe below.

3.1. Discriminator Collapse: A Barrier for First Order Methods

As discussed above, our simple GAN dynamics are able to capture the same undesired behaviors that more sophisticated GANs exhibit. In addition to these behaviors, our dynamics enables us to discern another degenerate behavior which does not seem to have previously been observed in the literature. We call this behavior *discriminator collapse*.

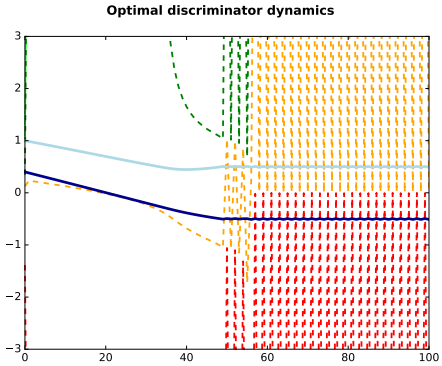
We first explain this phenomenon using language specific to our GMM-GAN dynamics. In our dynamics, discriminator collapse occurs when a discriminator interval which originally had finite width is forced by the dynamics to have its width converge to 0. This happens whenever this interval lies entirely in a region where the generator PDF is much

[1] We actually analyze a minor variation on the optimal discriminator dynamics. In particular, we do not rule out the existence of a measure zero set on which the dynamics are ill-behaved. Thus, we will analyze the optimal discriminator dynamics after adding an arbitrarily small amount of Gaussian noise. It is clear that by taking this noise to be sufficiently small (say exponentially small) then we avoid this pathological set with probability 1, and moreover the noise does not otherwise affect the convergence analysis at all. For simplicity, we will ignore this issue for the rest of the paper.

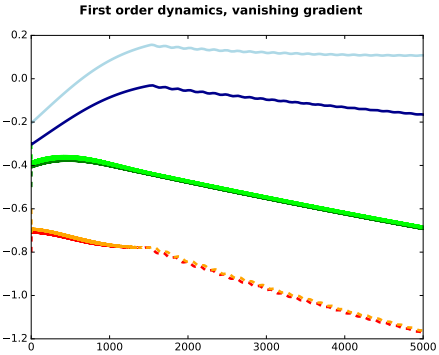


(a) Converging behavior

(b) Mode collapse and oscillation



(c) Optimal discriminator



(d) Vanishing gradient

Figure 1: A selection of different GAN behaviors. In all plots the true distribution was G_{μ^*} with $\mu^* = (-0.5, 0.5)$, and step size was taken to be 0.1. The solid lines represent the two coordinates of $\widehat{\mu}$, and the dotted lines represent the discriminator intervals. In order: (a) first order dynamics with initial conditions that converge to the true distribution. (b) First order dynamics with initial conditions that exhibit wild oscillation before mode collapse. (c) Optimal discriminator dynamics. (d) First order dynamics that exhibit vanishing gradients and converge to the wrong distribution. Observe that the optimal discriminator dynamics converge, and then the discriminator varies wildly, because the objective function is not differentiable at optimality. Despite this it remains roughly at optimality from step to step.

larger than the discriminator PDF. We will shortly argue why this is undesirable.

In Figure 2, we show an example of discriminator collapse in our dynamics. Each plot in the figure shows the true PDF minus the PDF of the generators, where the regions covered by the discriminator are shaded. Plot (a) shows the initial configuration of our example. Notice that the leftmost discriminator interval lies entirely in a region for which the true PDF minus the generators' PDF is negative. Since the discriminator is incentivized to only have mass where the difference is positive, the first order dynamics will cause the discriminator interval to collapse to have length zero if it is in a negative region. We see in Plot (c) that this discriminator collapses if we run many discriminator steps for this generator. In particular, these steps do not converge to the globally optimal discriminator shown in Plot (b).

This collapse also occurs when we run the dynamics. In Plots (d) and (e), we see that after running the first order dynamics – or even unrolled dynamics – for many iterations, eventually *both* discriminators collapse. When a discriminator interval has length zero, it can never uncollapse, and moreover, its contribution to the gradient of the generator is zero. Thus these dynamics will never converge.

For general GANs, we view discriminator collapse as a situation when the local optimization landscape around the current discriminator encourages it to make updates which decrease its representational power. For instance, this could happen because the first order updates are unable to wholly follow the evolution of the optimal discriminator due to attraction of local maxima, and thus only capture part of the optimal discriminator's structure. We view understanding the exact nature of discriminator collapse in more general settings and interesting research problem to explore further.

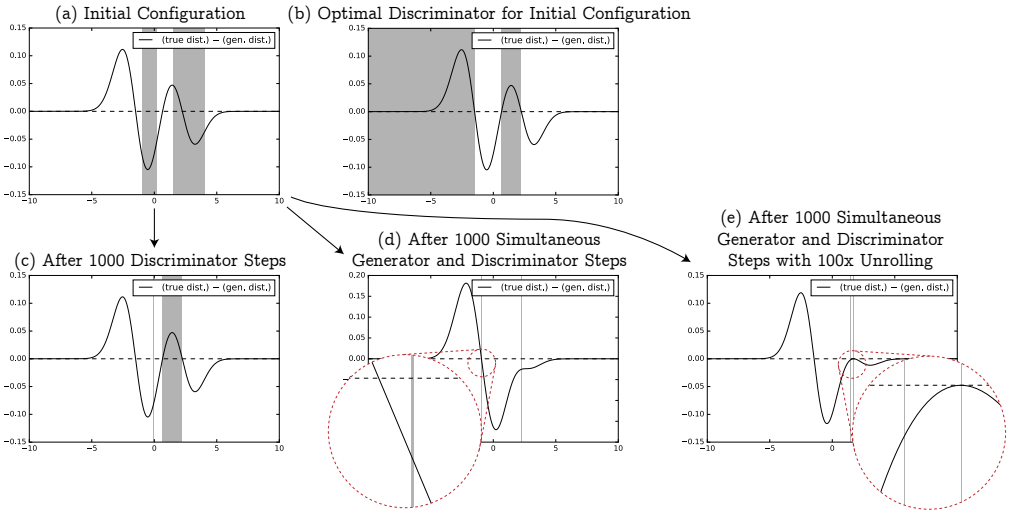


Figure 2: Example of Discriminator Collapse. The initial configuration has $\mu^* = \{-2, 2\}$, $\widehat{\mu} = \{-1, 2.5\}$, left discriminator $[-1, 0.2]$, and right discriminator $[-1, 2.5]$. The (multiplicative) step size used to generate (c), (d), and (e) was 0.3. These plots are discussed in Subsection 3.1.

4. Analyzing the Optimal Discriminator

We provide now a high level overview of the proof of Theorem 3.1.

The key element we will need in our proof is the ability to quantify the progress our updates make on converging towards the optimal solution. This is particularly challenging as our objective function is neither convex nor smooth. The following lemma is our main tool for achieving that. Roughly stated, it says that for any Lipschitz function, even if it is non-convex and *non-smooth*, as long as the change in its derivative is smaller in magnitude than the actual value of the derivative, gradient descent makes actual progress on the function value. Note that this change of derivative condition is much weaker than typical assumptions one makes to analyze gradient descent.

Lemma 4.1. *Let $g : \mathbb{R}^k \to \mathbb{R}$ be a Lipschitz function that is differentiable at some fixed $x \in \mathbb{R}^k$. For some $\eta > 0$, let $x' = x - \eta \nabla f(x)$. Suppose there exists $c < 1$ so that almost all $v \in L(x, x')$, where $L(x, y)$ denotes the line between x and y, g is differentiable, and moreover, we have $\|\nabla g(x) - \nabla g(v)\|_2 \leq c\|\nabla g(x)\|_2$. Then $g(x') - g(x) \leq -\eta(1-c)\|\nabla g(x)\|_2^2$.*

Here, we will use the convention that $\mu_1^* \leq \mu_2^*$, and during the analysis, we will always assume for simplicity of notation that $\widehat{\mu}_1 \leq \widehat{\mu}_2$. Also, in what follows, let $f(\widehat{\mu}) = f_{\mu^*}(\widehat{\mu}) = d_{\mathrm{TV}}(G_{\widehat{\mu}}, G_{\mu^*})$ and $F(\widehat{\mu}, x) = G_{\mu^*}(x) - G_{\widehat{\mu}}(x)$ be the objective function and the difference of the PDFs between the true distribution and the generator, respectively.

For any $\delta > 0$, define the sets

$$\mathrm{Rect}(\delta) = \mathrm{Rect}(\mu^*, \delta) = \{\widehat{\mu} : |\widehat{\mu}_i - \mu_j^*| < \delta \text{ for some } i, j\}$$
$$\mathrm{Opt}(\delta) = \mathrm{Opt}(\mu^*, \delta) = \{\widehat{\mu} : |\widehat{\mu}_i - \mu_i^*| < \delta \text{ for all } i\} .$$

to be the set of parameter values which have at least one parameter which is not too far from optimality, and the set of parameter values so that all parameter values are close. We also let $B(C)$ denote the box of sidelength C around the origin, and we let $\mathrm{Sep}(\gamma) = \{v \in \mathbb{R}^2 : |v_1 - v_2| > \gamma\}$ be the set of parameter vectors which are not too close together.

Our main work lies within a set of lemmas which allow us to instantiate the bounds in Lemma 4.1. We first show a pair of lemmas which show that, explicitly excluding bad cases such as mode collapse, our dynamics satisfy the conditions of Lemma 4.1. We do so by establishing a strong lower bound on the gradient when we are fairly away from optimality (Lemma 4.2). Then, we show a relatively weak bound on the smoothness of the function (Lemma 4.3), but which is sufficiently strong in combination with Lemma 4.2 to satisfy Lemma 4.1. Finally, we rule out the pathological cases we explicitly excluded earlier, such as mode collapse or divergent behavior (Lemmas 4.4 and 4.5). Putting all these together appropriately yields the desired claim.

Our first lemma is a lower bound on the gradient value:

Lemma 4.2. *Fix $C \geq 1 \geq \gamma \geq \delta > 0$. Suppose $\widehat{\mu} \notin \mathrm{Rect}(0)$, and suppose $\mu^*, \widehat{\mu} \in B(C)$ and $\mu^* \in \mathrm{Sep}(\gamma), \widehat{\mu} \in \mathrm{Sep}(\delta)$. Then, there is some $K = K(\delta, C) = \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ so that $\|\nabla f_{\mu^*}(\widehat{\mu})\|_2 \geq K$.*

The above lemma statement is slightly surprising at first glance. It says that the gradient is never 0, which would

suggest there are no local optima at all. To reconcile this, one should note that the gradient is not continuous (defined) everywhere. In fact, one could show that points near the local optima of our problem actually have some of the largest derivatives compared to other points, while the local optima themselves all have an undefined gradient.

The second states a bound on the smoothness of the function:

Lemma 4.3. *Fix $C \geq 1$ and $\gamma \geq \delta > 0$ so that δ is sufficiently small. Let $\mu^*, \widehat{\mu}, \widehat{\mu}'$ be such that $L(\widehat{\mu}, \widehat{\mu}') \cap \mathrm{Opt}(\delta) = \varnothing$, $\mu^* \in \mathrm{Sep}(\gamma)$, $\widehat{\mu}', \widehat{\mu} \in \mathrm{Sep}(\delta)$, and $\mu^*, \widehat{\mu}, \widehat{\mu}' \in B(C)$. Let $K = \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ be the K for which Lemma 4.2 holds with those parameters. If we have $\|\widehat{\mu}' - \widehat{\mu}\|_2 \leq \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$, we get*

$$\|\nabla f_{\mu^*}(\widehat{\mu}') - \nabla f_{\mu^*}(\widehat{\mu})\|_2 \leq K/2 \leq \|\nabla f_{\mu^*}(\widehat{\mu})\|_2/2.$$

These two lemmas almost suffice to prove progress as in Lemma 4.1, however, there is a major caveat. Specifically, Lemma 4.3 needs to assume that $\widehat{\mu}$ and $\widehat{\mu}'$ are sufficiently well-separated, and that they are bounded. While the $\widehat{\mu}_i$ start out separated and bounded, it is not clear that it does not mode collapse or diverge off to infinity. However, we are able to rule these sorts of behaviors out. Formally:

Lemma 4.4 (No mode collapse)**.** *Fix $\gamma > 0$, and let $\delta \leq \gamma/100$ be sufficiently small. Let $\eta \leq \delta/C$ for some C sufficiently large. Suppose $\mu^* \in \mathrm{Sep}(\gamma)$. Then, if $\widehat{\mu} \in \mathrm{Sep}(\delta)$, and $\widehat{\mu}' = \widehat{\mu} - \eta \nabla f_{\mu^*}(\widehat{\mu})$, we have $\widehat{\mu}' \in \mathrm{Sep}(\delta)$.*

Lemma 4.5 (No diverging to infinity)**.** *Let $C > 0$ be sufficiently large, and let $\eta > 0$ be sufficiently small. Suppose $\mu^* \in B(C)$, and $\widehat{\mu} \in B(2C)$. Then, if we let $\widehat{\mu}' = \widehat{\mu} - \eta \nabla f_{\mu^*}(\widehat{\mu})$, then $\widehat{\mu}' \in B(2C)$.*

Together, these four lemmas together suffice to prove Theorem 3.1 by setting parameters appropriately. We now pause to make some remarks on the proof techniques for these Lemmas. At a high level, Lemmas 4.2, 4.4, 4.5 all follow from involved case analyses. Specifically, we are able to deduce structure about the possible discriminator intervals by reasoning about the structure of the current mean estimate $\widehat{\mu}$ and the true means. From there we are able to derive bounds on how these discriminator intervals affect the derivatives and hence the update functions.

To prove Lemma 4.3, we carefully study the evolution of the optimal discriminator as we make small changes to the generator. The key idea is to show that when the generator means are far from the true means, then the zero crossings of $F(\widehat{\mu}, x)$ cannot evolve too unpredictably as we change $\widehat{\mu}$. We do so by showing that locally, in this setting F can be approximated by a low degree polynomial with large coefficients, via bounding the condition number of a certain Hermite Vandermonde matrix. This gives us sufficient control over the local behavior of zeros to deduce the desired claim. By being sufficiently careful with the bounds, we are then able to go from this to the full generality of the lemma. We defer further details to Appendix B.

5. Experiments

To illustrate more conclusively that the phenomena demonstrated in Figure 1 are not rare, and that first order dynamics do often fail to converge, we also conducted the following heatmap experiments. We set $\mu^* = (-0.5, 0.5)$ as in Figure 1. We then set a grid for the $\widehat{\mu}$, so that each coordinate is allowed to vary from -1 to 1. For each of these grid points, we randomly chose a set of initial discriminator intervals, and ran the first order dynamics for 3000 iterations, with constant stepsize 0.3. We then repeated this 120 times for each grid point, and plotted the probability that the generator converged to the truth, where we say the generator converged to the truth if the TV distance between the generator and optimality is < 0.1. The choice of these parameters was somewhat arbitrary, however, we did not observe any qualitative difference in the results by varying these numbers, and so we only report results for this specific set of parameters. We also did the same thing for the optimal discriminator dynamics, and for unrolled discriminator dynamics with 5 unrolling steps, as described in (Metz et al., 2017), which attempt to match the optimal discriminator.

The results of the experiment are given in Figure 3. Qualitatively, we see that all three methods fail when we initialize the two generator means to be the same. This makes sense, since in that regime, the generator starts out mode collapsed and it is impossible for it to un-"mode collapse", so it cannot fit the true distribution well.

Ignoring this pathology, we see that the optimal discriminator otherwise always converges to the ground truth, as our theory predicts. On the other hand, both regular first order dynamics and unrolled dynamics often times fail, although unrolled dynamics do succeed more often than regular first order dynamics. This suggests that the pathologies encountered in Figure 1 are not so rare, and that indeed, these first order methods are quite often unable to emulate optimal discriminator dynamics.

6. Related Work

GANs have received a tremendous amount of attention in the deep learning community over the past two years (Goodfellow, 2017). Hence we only compare our results to the most closely related papers here.

(Arora et al., 2017) studies generalization aspects of GANs and the existence of equilibria in the two-player game. In contrast, our paper focuses on the *dynamics* of GAN training. We provide the first proof of global convergence and show

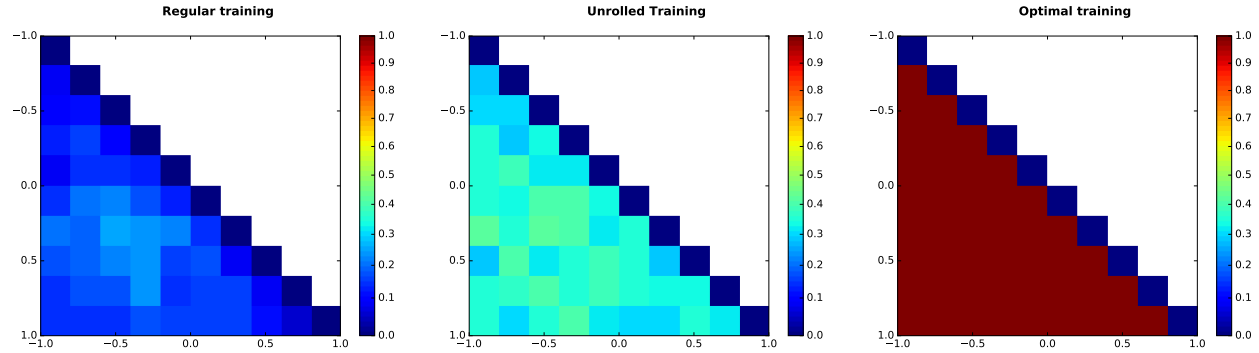


Figure 3: Heatmap of success probability for random discriminator initialization for regular GAN training, unrolled GAN training, and optimal discriminator dynamics.

that a GAN with an optimal discriminator always converges to an approximate equilibrium.

One recently proposed method for improving the convergence of GAN dynamics is the unrolled GAN (Metz et al., 2017). The paper proposes to "unroll" multiple discriminator gradient steps in the generator loss function. The authors argue that this improves the GAN dynamics by bringing the discriminator closer to an optimal discriminator response. However, our experiments show that this is not a perfect approximation: the unrolled GAN still fails to converge in multiple initial configurations (however, it does converge more often than a "vanilla" one-step discriminator).

(Arjovsky & Bottou, 2017) also takes a theoretical view on GANs. They identify two important properties of GAN dynamics: (i) Absolute continuity of the population distribution, and (ii) overlapping support between the population and generator distribution. If these conditions do not hold, they show that the GAN dynamics fail to converge in some settings. However, they do not prove that the GAN dynamics *do* converge under such assumptions. We take a complementary view: we give a convergence proof for a concrete GAN dynamics. Our simple model shows that absolute continuity and support overlap are not the only important aspects in GAN dynamics: although our 2-GMMs satisfy both of their conditions, the first-order dynamics still fail to converge.

The paper (Nagarajan & Kolter, 2017) studies the stability of equilibria in GAN training. The authors invoke stability arguments for dynamical systems based on differential equations. In contrast to our work, the results focus on *local* stability while we establish *global* convergence results. Moreover, their theorems rely on fairly strong assumptions. While the authors give a concrete model for which these assumptions are satisfied (the linear quadratic Gaussian GAN), the corresponding target and generator distributions are *unimodal*. Hence this model cannot exhibit mode collapse. We propose the GMM-GAN specifically because it is rich enough to exhibit mode collapse.

The recent work (Grnarova et al., 2017) views GAN training through the lens of online learning. The paper establishes results for the game-theoretic minimax formulation based on results from online learning (the well-known Follow-the-Regularized-Leader approach). The authors give results that go beyond the convex-concave setting, but do not address generalization questions. Moreover, their theoretical algorithm is not based on gradient descent (in contrast to essentially all practical GAN training) and relies on an oracle for minimizing the highly non-convex generator loss. This viewpoint is complementary to ours. We give results for learning the unknown distribution and analyze the commonly used gradient descent approach for learning GANs.

There is also a growing literature on non-convex optimization, e.g., (Ge et al., 2015; Lee et al., 2016). However, prior work with general results on non-convex optimization does not apply to our setting. For instance, our loss function is not smooth and the gradient does not vanish near minima (it is not even defined there).

7. Conclusions

We haven taken a step towards a principled understanding of GAN dynamics. We define a simple yet rich model of GAN training and prove convergence of the corresponding optimization dynamics. To the best of our knowledge, our work is the first to establish global convergence guarantees for a parametric GAN model. We find an interesting dichotomy: If we always take optimal discriminator steps, the training dynamics provably converge from any starting point. In contrast, we show experimentally that the training dynamics often fail if we take first order discriminator steps instead. We also identify *discriminator collapse* as a candidate barrier in GAN training and show that it often prevents our first order dynamics from converging. We believe that our results provide new insights into GAN training and point towards a rich algorithmic landscape that is to be explored in order to further understand GAN dynamics.

References

Arjovsky, Martin and Bottou, Leon. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017. URL https://arxiv.org/abs/1701.04862.

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

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

Arora, Sanjeev, Ge, Rong, Liang, Yingyu, Ma, Tengyu, and Zhang, Yi. Generalization and equilibrium in generative adversarial nets (GANs). *arXiv preprint arXiv:1703.00573*, 2017.

Chan, Siu-On, Diakonikolas, Ilias, Servedio, Rocco A, and Sun, Xiaorui. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 604–613. ACM, 2014.

Devroye, Luc and Lugosi, Gábor. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Finn, Chelsea, Christiano, Paul, Abbeel, Pieter, and Levine, Sergey. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *CoRR*, abs/1611.03852, 2016. URL http://arxiv.org/abs/1611.03852.

Gautschi, Walter. How (un) stable are Vandermonde systems? *Lecture Notes in Pure and Applied Mathematics*, 124:193–210, 1990.

Ge, Rong, Huang, Furong, Jin, Chi, and Yuan, Yang. Escaping from saddle points - online stochastic gradient for tensor decomposition. In *COLT*, 2015.

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

Goodfellow, Ian J. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, abs/1701.00160, 2017. URL http://arxiv.org/abs/1701.00160.

Grnarova, Paulina, Levy, Kfir Y., Lucchi, Aurelien, Hofmann, Thomas, and Krause, Andreas. An online learning approach to generative adversarial networks. *arXiv preprint arXiv:1706.03269*, 2017.

Hummel, R.A. and Gidas, B.C. *Zero Crossings and the Heat Equation*. New York University., 1984.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

Lee, Jason D., Simchowitz, Max, Jordan, Michael I., and Recht, Benjamin. Gradient descent only converges to minimizers. In *COLT*, 2016.

Markov, VA. On functions deviating least from zero in a given interval. *Izdat. Imp. Akad. Nauk, St. Petersburg*, pp. 218–258, 1892.

Metz, Luke, Poole, Ben, Pfau, David, and Sohl-Dickstein, Jascha. Unrolled generative adversarial networks. In *ICLR*, 2017. URL http://arxiv.org/abs/1611.02163.

Nagarajan, Vaishnavh and Kolter, J. Zico. Gradient descent gan optimization is locally stable. *arXiv preprint arXiv:1706.04156*, 2017.

van den Oord, Aäron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew W., and Kavukcuoglu, Koray. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499.