On the Limitations of First-Order Approximation in GAN Dynamics

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

Abstract

While Generative Adversarial Networks (GANs) have demonstrated promising performance on multiple vision tasks, their learning dynamics are not yet well understood, both in theory and in practice. To address this issue, we study GAN dynamics in a simple yet rich parametric model that exhibits several of the common problematic convergence behaviors such as vanishing gradients, mode collapse, and diverging or oscillatory behavior. In spite of the non-convex nature of our model, we are able to perform a rigorous theoretical analysis of its convergence behavior. Our analysis reveals an interesting dichotomy: a GAN with an optimal discriminator provably converges, while first order approximations of the discriminator steps lead to unstable GAN dynamics and mode collapse. Our result suggests that using first order discriminator steps (the de-facto standard in most existing GAN setups) might be one of the factors that makes GAN training challenging in practice.

1. Introduction

Generative Adversarial Networks (GANs) have recently been proposed as a novel framework for learning generative models (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, reliably training GANs is a challenging problem that often hinders further research and applicability in this area. Practitioners have encountered a variety of obstacles in this context 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. As a result, the root cause of frequent failures of GAN dynamics in practice remains unclear.

In this paper, we take a first step towards a principled understanding of GAN *dynamics*. Our general methodology is to 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, 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 parametric and non-trivial GAN dynamics.

Our results show a clear dichotomy between the dynamics arising from applying simultaneous gradient descent and the one that is able to use an optimal discriminator. The GAN with optimal discriminator provably converges from (essentially) *any* starting point. On the other hand, the simultaneous gradient GAN empirically often fails to converge, even when the discriminator is allowed many more gradient steps than the generator. These findings go against the common wisdom that first order methods are sufficiently strong for all deep learning applications. By carefully inspecting our models, we are able to pinpoint some of the causes of this, and we highlight a phenomena we call *discriminator collapse* which often causes first order methods to fail in our setting.

[1]MIT. Correspondence to: Jerry Li <jerryzli@mit.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

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). This formulation can also encode other common objectives: most notably, as we will show, 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, 2018). 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 large gap between the theory and practice of GANs. 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):

Our theoretical understanding of GANs is still fairly poor. In particular, to the best of the authors' knowledge, all existing analyzes of GAN dynamics for parametric models simply consider global optima and stationary points of the dynamics. There has been no rigorous study of the actual GAN dynamics, except studying it in the immediate neighborhood of such stationary points (Nagarajan & Kolter, 2017). This raises a natural question:

Can we understand the convergence behavior of GANs?

This question is difficult to tackle for many reasons. One of them is the non-convexity of the GAN objective/loss function, and of the generator and discriminator sets. Another one is that, in practice, GANs are always optimized using first order methods. That is, instead of following the "ideal" dynamics that has both the generator and discriminator always perform the optimal update, we just approximate such updates by a sequence of gradient steps. This is motivated by the fact that computing such optimal updates is, in general, algorithmically intractable, and adds an additional layer of complexity to the problem.

In this paper, we want to change this state of affairs and initiate the study of GAN dynamics from an algorithmic perspective. Specifically, we pursue the following question:

What is the impact of using first order approximation on the convergence of GAN dynamics?

Concretely, we focus on analyzing the difference between two GAN dynamics: a "first order" dynamics, in which both the generator and discriminator use first order updates; and an "optimal discriminator" dynamics, in which only the generator uses first order updates but the discriminator always makes an optimal update. Even the latter, simpler dynamics has proven to be challenging to understand. Even the question of whether using the optimal discriminator updates is the right approach has already received considerable attention. In particular, (Arjovsky & Bottou, 2017) present theoretical evidence that using the optimal discriminator at each step may not be desirable in certain settings (although these settings are very different to the one we consider in this paper).

We approach our goal by defining a simple GAN model whose dynamics, on one hand, captures many of the difficulties of real-world GANs but, on the other hand, is still simple enough to make analysis possible. We then rigorously study our questions in the context of this model. Our intention is to make the resulting understanding be the first step towards crystallizing a more general picture.

3. A Simple Model for Studying GAN 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 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. We 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, motivated by 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 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, finding the best fit in total variation distance to some unknown $G_{\mu^*} \in \mathcal{G}$ is equivalent to finding $\widehat{\mu}$ minimizing

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

is a 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 resorts to 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 specified as

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

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

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

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.

Parametrization We note here that there could be several potential GAN dynamics to consider here. Each one resulting from slightly different parametrization of the total variation distance. For instance, a completely equivalent way to define the total variation distance is

$$d_{\mathrm{TV}}(P, Q) = \max_A |P(A) - Q(A)| \,, \tag{12}$$

which does not change the value of the variational distance, but does change the induced dynamics. We do not focus on these induced dynamics in this paper since they do not exactly fit within the traditional GAN framework, i.e. it is not of the form (1) (see Appendix B). Nevertheless, it is an interesting set of dynamics and it is a natural question whether similar phenomena occur in these dynamics. In Appendix B, we show the the optimal discriminator dynamics are unchanged, and the induced first order dynamics have qualitatively similar behavior to the ones we consider in this paper. This also suggests that the phenomena we exhibit might be more fundamental.

4. Optimal Discriminator vs. First Order Dynamics

We now describe our results in more detail. We first consider the dynamics induced by the optimal discriminator. Our main theoretical result is[1]:

Theorem 4.1. *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 $d_{\mathrm{TV}}(G_{\mu^*}, G_{\widehat{\mu}^{(T)}}) \leq \delta$.*

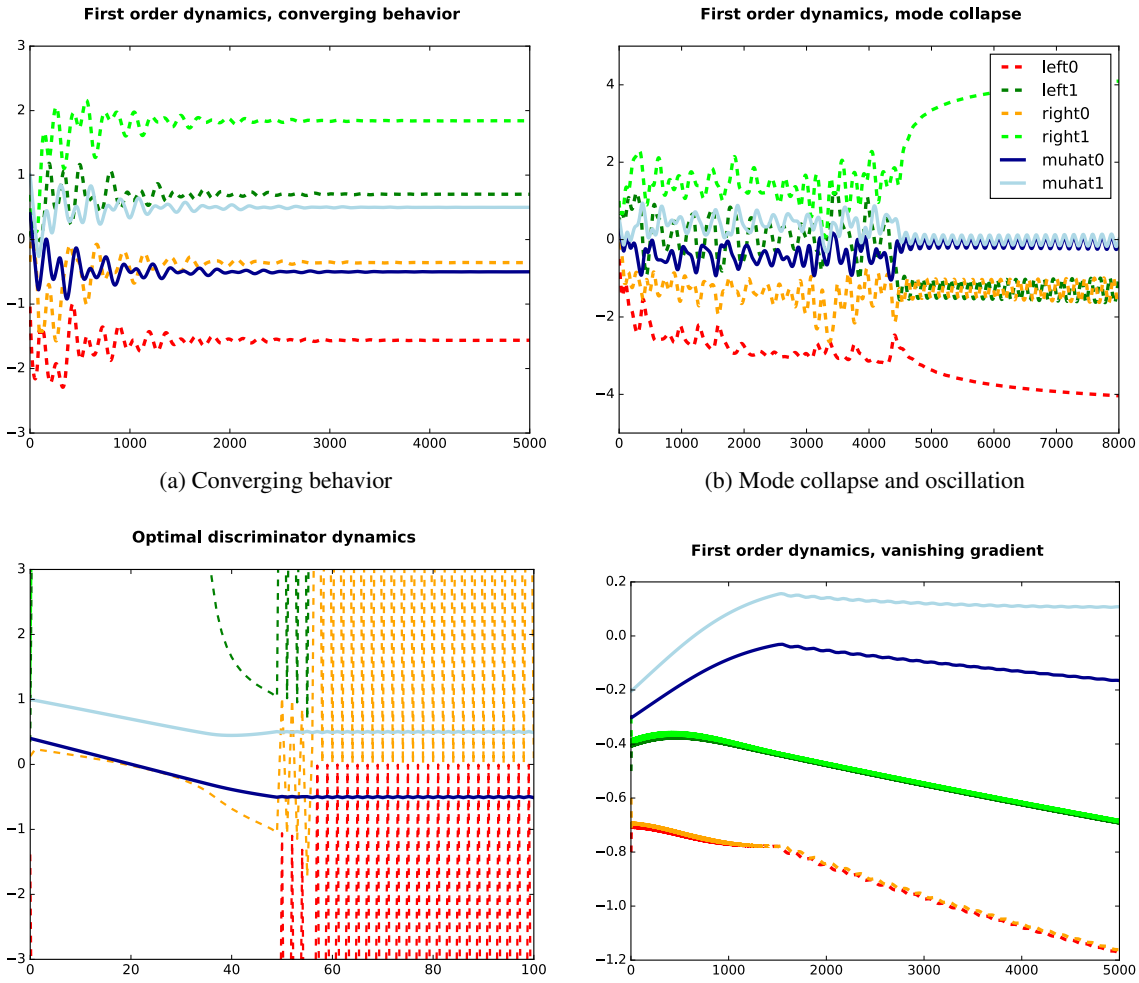
In other words, if the μ^* are bounded by a constant, and not too close together, then in time which is polynomial in the inverse of the desired accuracy δ and e^{-C^2}, where C is a bound on how far apart the μ^* and $\widehat{\mu}$ are, the optimal discriminator dynamics converge to the ground truth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

On the other hand, we provide simulation results 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 4.1. 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 for the generator. In particular, one barrier we observed for any such dynamics is something we call *discriminator collapse*, that we describe in Section 6. We do not provide a proof of convergence for the first order dynamics, but we remark that in light of our simulation results, this is simply because the first order dynamics do not converge.

4.1. Analyzing the Optimal Discriminator Dynamics

We provide now a high level overview of the proof of Theorem 4.1. 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,

[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.

even if it is non-convex and *non-smooth*, as long as the change in its derivative is smaller in magnitude than the value of the derivative, gradient descent makes progress on the function value. Note that this condition is much weaker than typical assumptions used to analyze gradient descent.

Lemma 4.2. *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

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

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 4.2. 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 4.2. We do so by establishing a strong (in fact, nearly constant) lower bound on the gradient when we are fairly away from optimality (Lemma 4.3). Then, we show a relatively weak bound on the smoothness of the function (Lemma 4.4), but which is sufficiently strong in combination with Lemma 4.3 to satisfy Lemma 4.2. Finally, we rule out the pathological cases we explicitly excluded earlier, such as mode collapse or divergent behavior (Lemmas 4.5 and 4.6). Putting all these together appropriately yields the desired statement. Our first lemma is a lower bound on the gradient value:

Lemma 4.3. *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 $\mu^* \in \mathrm{Sep}(\gamma), \widehat{\mu} \in \mathrm{Sep}(\delta)$. There is some $K = \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ 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.

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

Lemma 4.4. *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 4.3 holds with those parameters. If we have $\|\widehat{\mu}' - \widehat{\mu}\|_2 \leq \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ for appropriate choices of constants on the RHS, 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 4.2, however, there is a major caveat. Specifically, Lemma 4.4 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 4.5 (No mode collapse)**.** *Fix $\gamma > 0$, and let δ be sufficiently small. Let $\eta \leq \delta/C$ for some C 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 4.6 (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 4.1 by setting parameters appropriately. We refer the reader to the supplementary material for more details including the proofs.

5. Experiments

To illustrate more conclusively that the phenomena demonstrated in Figure 1 are not particularly 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 these 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 dynamics.

The results of the experiment are given in Figure 2. 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 in Figure 1 are not so rare, and that these first order methods are quite often unable to emulate optimal discriminator dynamics.

6. Why do first order methods get stuck?

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

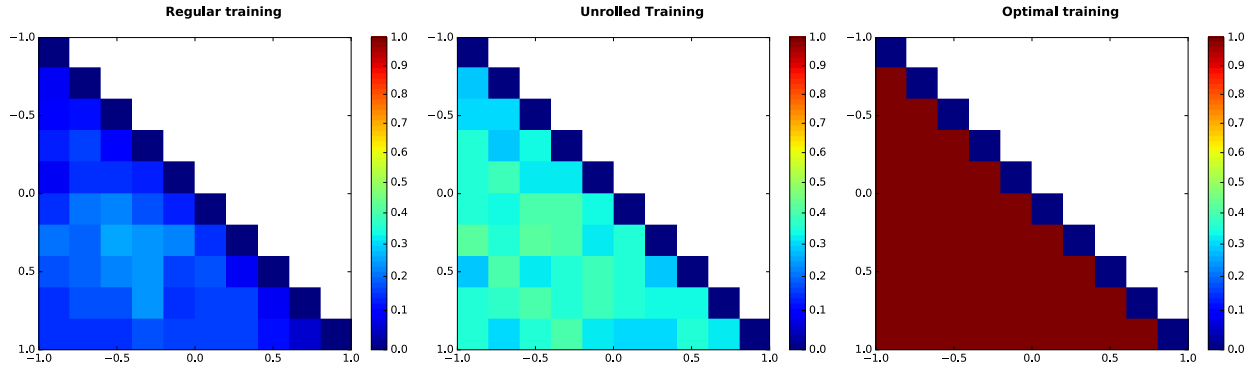


Figure 2. Heatmap of success probability for random discriminator initialization for regular GAN training, unrolled GAN training, and optimal discriminator dynamics.

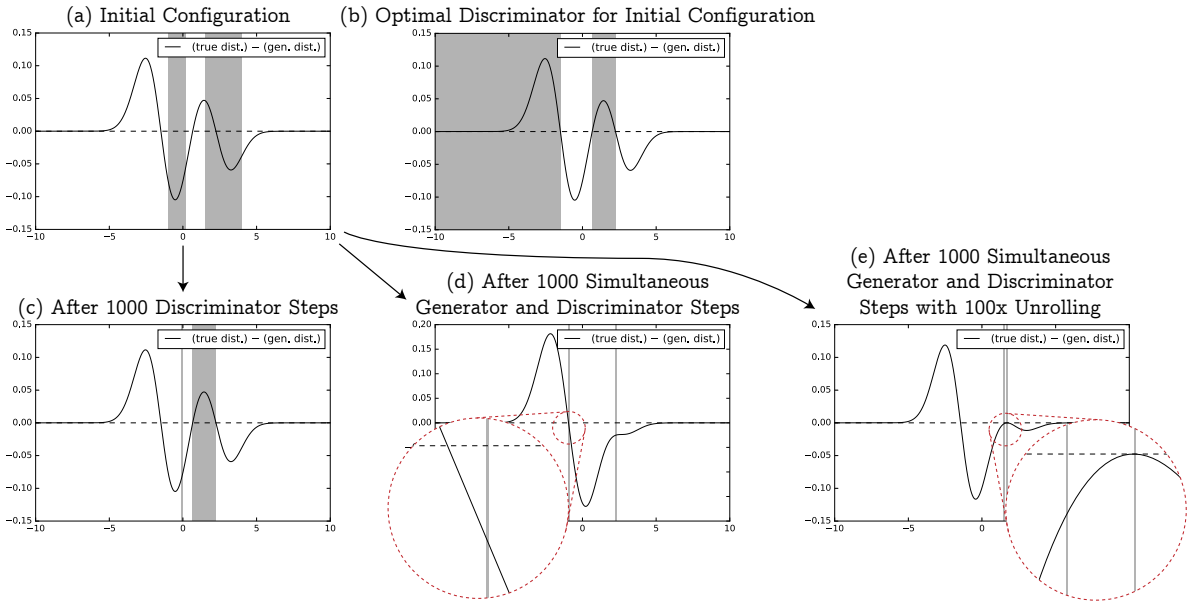


Figure 3. Example of Discriminator Collapse. The initial configuration has $\mu^* = \{-2, 2\}$, $\widehat{\mu} = \{-1, 2.5\}$, left discriminator $[-1, 0.2]$, and right discriminator $[-1, 2.5]$. The (multiplicative) step size used to generate (c), (d), and (e) was 0.3.

which does not seem to have previously been observed in the literature. We call this behavior *discriminator collapse*. At a high level, this phenomenon is when the local optimization landscape around the current discriminator encourages it to make updates which decrease its representational power. We view understanding the exact nature of discriminator collapse in more general settings and interesting research problem to explore further.

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

In Figure 3, 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 on regions 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 fixed 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 to the ground truth.

7. Related Work

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

The recent paper (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 rigorous proof of global convergence and show 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. 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).

The authors of (Arjovsky & Bottou, 2017) also take 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. Moreover, our model shows that absolute continuity and support overlap are not the only important aspects in GAN dynamics: although our distributions clearly 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. 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., 2018) views GAN training through the lens of online learning. The paper gives results for the game-theoretic minimax formulation based on results from online learning. The authors give results that go beyond the convex-concave setting, but do not address generalization questions. Moreover, their 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 our approach. We establish results for learning the unknown distribution and analyze the commonly used gradient descent approach for learning GANs.

8. 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 dynamics. To the best of our knowledge, our work is the first to establish global convergence guarantees for a parametric GAN. We find an interesting dichotomy: If we take optimal discriminator steps, the training dynamics provably converge. In contrast, we show experimentally that the dynamics often fail if we take first order discriminator steps. We believe that our results provide new insights into GAN training and point towards a rich algorithmic landscape to be explored in order to further understand GAN dynamics.

Acknowledgements

Jerry Li was supported by NSF Award CCF-1453261 (CAREER), CCF-1565235, a Google Faculty Research Award, and an NSF Graduate Research Fellowship. Aleksander Mądry was supported in part by an Alfred P. Sloan Research Fellowship, a Google Research Award, and the NSF grant CCF-1553428. John Peebles was supported by the NSF Graduate Research Fellowship under Grant No. 1122374 and by the NSF Grant No. 1065125. Ludwig Schmidt was supported by a Google PhD Fellowship.

References

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. In *ICML*, 2017.

Arora, S. and Zhang, Y. Do gans actually learn the distribution? an empirical study. In *ICLR*, 2018.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (GANs). In *ICML*, 2017.

Chan, S.-O., Diakonikolas, I., Servedio, R. A., and Sun, X. Efficient density estimation via piecewise polynomial approximation. In *STOC*, 2014.

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

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

Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, abs/1701.00160, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.

Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., and Krause, A. An online learning approach to generative adversarial networks. In *ICLR*, 2018.

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

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

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. In *ICLR*, 2017.

Nagarajan, V. and Kolter, J. Z. Gradient descent gan optimization is locally stable. In *NIPS*, 2017.