Which Training Methods for GANs do actually Converge?

Lars Mescheder [1] **Andreas Geiger** [1 2] **Sebastian Nowozin** [3]

Abstract

Recent work has shown local convergence of GAN training for absolutely continuous data and generator distributions. In this paper, we show that the requirement of absolute continuity is necessary: we describe a simple yet prototypical counterexample showing that in the more realistic case of distributions that are not absolutely continuous, unregularized GAN training is not always convergent. Furthermore, we discuss regularization strategies that were recently proposed to stabilize GAN training. Our analysis shows that GAN training with instance noise or zero-centered gradient penalties converges. On the other hand, we show that Wasserstein-GANs and WGAN-GP with a finite number of discriminator updates per generator update do not always converge to the equilibrium point. We discuss these results, leading us to a new explanation for the stability problems of GAN training. Based on our analysis, we extend our convergence results to more general GANs and prove local convergence for simplified gradient penalties even if the generator and data distributions lie on lower dimensional manifolds. We find these penalties to work well in practice and use them to learn high-resolution generative image models for a variety of datasets with little hyperparameter tuning.

1. Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are powerful latent variable models that can be used to learn complex real-world distributions. Especially for images, GANs have emerged as one of the dominant approaches for generating new realistically looking samples after the model has been trained on some dataset.

Method	Local convergence (a.c. case)	Local convergence (general case)
unregularized (Goodfellow et al., 2014)	✓	✗
WGAN (Arjovsky et al., 2017)	✗	✗
WGAN-GP (Gulrajani et al., 2017)	✗	✗
DRAGAN (Kodali et al., 2017)	✓	✗
Instance noise (Sønderby et al., 2016)	✓	✓
ConOpt (Mescheder et al., 2017)	✓	✓
Gradient penalties (Roth et al., 2017)	✓	✓
Gradient penalty on real data only	✓	✓
Gradient penalty on fake data only	✓	✓

Table 1. Convergence properties of different GAN training algorithms for general GAN-architectures. Here, we distinguish between the case where both the data and generator distributions are absolute continuous (a.c.) and the general case where they may lie on lower dimensional manifolds.

However, while very powerful, GANs can be hard to train and in practice it is often observed that gradient descent based GAN optimization does not lead to convergence. As a result, a lot of recent research has focused on finding better training algorithms (Arjovsky et al., 2017; Gulrajani et al., 2017; Kodali et al., 2017; Sønderby et al., 2016; Roth et al., 2017) for GANs as well as gaining better theoretically understanding of their training dynamics (Arjovsky et al., 2017; Arjovsky & Bottou, 2017; Mescheder et al., 2017; Nagarajan & Kolter, 2017; Heusel et al., 2017).

Despite practical advances, the training dynamics of GANs are still not completely understood. Recently, Mescheder et al. (2017) and Nagarajan & Kolter (2017) showed that local convergence and stability properties of GAN training can be analyzed by examining the eigenvalues of the Jacobian of the the associated gradient vector field: if the Jacobian has only eigenvalues with negative real-part at the equilibrium point, GAN training converges locally for small enough learning rates. On the other hand, if the Jacobian has eigenvalues on the imaginary axis, it is generally not locally convergent. Moreover, Mescheder et al. (2017) showed that if there are eigenvalues close but not on the imaginary axis, the training algorithm can require intractably small learning rates to achieve convergence. While Mescheder et al. (2017) observe eigenvalues close to the imaginary axis in practice, this observation does not answer the question if eigenvalues close to the imaginary axis are a general phenomenon and if yes, whether they are indeed the root cause for the training

[1] MPI Tübingen, Germany [2] ETH Zürich, Switzerland [3] Microsoft Research, Cambridge, UK. Correspondence to: Lars Mescheder <lars.mescheder@tue.mpg.de>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

instabilities that people observe in practice.

A partial answer to this question was given by Nagarajan & Kolter (2017), who showed that for absolutely continuous data and generator distributions[1] all eigenvalues of the Jacobian have negative real-part. As a result, GANs are locally convergent for small enough learning rates in this case. However, the assumption of absolute continuity is not true for common use cases of GANs, where both distributions lie on lower dimensional manifolds (Sønderby et al., 2016; Arjovsky & Bottou, 2017).

In this paper we show that this assumption is indeed necessary: by considering a simple yet prototypical example of GAN training we analytically show that (unregularized) GAN training is not always locally convergent. We also discuss how recent techniques for stabilizing GAN training affect local convergence on our example problem. Our findings show that neither Wasserstein GANs (WGANs) (Arjovsky et al., 2017) nor Wasserstein GANs with Gradient Penalty (WGAN-GP) (Gulrajani et al., 2017) nor DRAGAN (Kodali et al., 2017) converge on this simple example for a fixed number of discriminator updates per generator update. On the other hand, we show that instance noise (Sønderby et al., 2016; Arjovsky & Bottou, 2017), zero-centered gradient penalties (Roth et al., 2017) and consensus optimization (Mescheder et al., 2017) lead to local convergence.

Based on our analysis, we give a new explanation for the instabilities commonly observed when training GANs based on discriminator gradients orthogonal to the tangent space of the data manifold. We also introduce simplified gradient penalties for which we prove local convergence. We find that these gradient penalties work well in practice, allowing us, among others, to learn a generative image model of all 1000 Imagenet classes in a single GAN.

In summary, our contributions are as follows:

- We identify a simple yet prototypical counterexample showing that (unregularized) gradient descent based GAN optimization is not always locally convergent
- We discuss if and how recently introduced regularization techniques stabilize the training
- We introduce simplified gradient penalties and prove local convergence for the regularized GAN training dynamics

All proofs can be found in the supplementary material.

2. Instabilities in GAN training

2.1. Background

GANs are defined by a min-max two-player game between a discriminative network $D_\psi(x)$ and generative network $G_\theta(z)$. While the discriminator tries to distinguish between real data point and data points produced by the generator, the generator tries to fool the discriminator. It can be shown (Goodfellow et al., 2014) that if both the generator and discriminator are powerful enough to approximate any real-valued function, the unique Nash-equilibrium of this two player game is given by a generator that produces the true data distribution and a discriminator which is 0 everywhere on the data distribution.

Following the notation of Nagarajan & Kolter (2017), the training objective for the two players can be described by an objective function of the form

$$L(\theta, \psi) = \mathrm{E}_{p(z)}\left[f(D_\psi(G_\theta(z)))\right] + \mathrm{E}_{p_\mathcal{D}(x)}\left[f(-D_\psi(x))\right] \quad (1)$$

for some real-valued function f. The common choice $f(t) = -\log(1 + \exp(-t))$ leads to the loss function considered in the original GAN paper (Goodfellow et al., 2014). For technical reasons we assume that f is continuously differentiable and satisfies $f'(t) \neq 0$ for all $t \in \mathbb{R}$.

The goal of the generator is to minimize this loss whereas the discriminator tries to maximize it. Our goal when training GANs is to find a Nash-equilibrium, i.e. a parameter assignment (θ^*, ψ^*) where neither the discriminator nor the generator can improve their utilities.

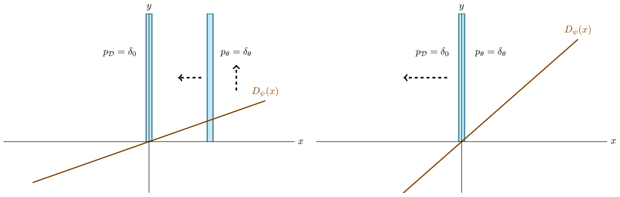
GANs are usually trained using Simultaneous or Alternating Gradient Descent (SimGD and AltGD). Both algorithms can be described as fixed point algorithms (Mescheder et al., 2017) that apply some operator $F_h(\theta, \psi)$ to the parameter values (θ, ψ) of the generator and discriminator, respectively. For example, simultaneous gradient descent corresponds to the operator $F_h(\theta, \psi) = (\theta, \psi) + h\, v(\theta, \psi)$, where $v(\theta, \psi)$ denotes the *gradient vector field*

$$v(\theta, \psi) := \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ \nabla_\psi L(\theta, \psi) \end{pmatrix}. \quad (2)$$

Similarly, alternating gradient descent can be described by an operator $F_h = F_{2,h} \circ F_{1,h}$ where $F_{1,h}$ and $F_{2,h}$ perform an update for the generator and discriminator, respectively.

Recently, it was shown (Mescheder et al., 2017) that local convergence of GAN training near an equilibrium point (θ^*, ψ^*) can be analyzed by looking at the spectrum of the Jacobian $F_h'(\theta^*, \psi^*)$ at the equilibrium: if $F_h'(\theta^*, \psi^*)$ has eigenvalues with absolute value bigger than 1 , the training algorithm will generally not converge to (θ^*, ψ^*). On the other hand, if all eigenvalues have absolute value smaller

[1] Nagarajan & Kolter (2017) also proved local convergence for a slightly more general family of probability distributions where the support of the generator is equal to the support of the true data distribution near the equilibrium point. Alternatively, they show that their results also hold when the discriminator satisfies certain (strong) smoothness conditions. However, these conditions are usually hard to satisfy in practice without prior knowledge about the support of the true data distribution.


y
$p_{\mathcal{D}} = \delta_0$
$p_\theta = \delta_\theta$
$D_\psi(x)$
x
y
$p_{\mathcal{D}} = \delta_0$
$p_\theta = \delta_\theta$
$D_\psi(x)$
x

(a) $t = t_0$ (b) $t = t_1$

Figure 1. Visualization of the counterexample showing that in the general case, gradient descent GAN optimization is not convergent: (a) In the beginning, the discriminator pushes the generator towards the true data distribution and the discriminator's slope increases. (b) When the generator reaches the target distribution, the slope of the discriminator is largest, pushing the generator away from the target distribution. This results in oscillatory training dynamics that never converge.

than 1, the training algorithm will converge to (θ^*, ψ^*) with linear rate $\mathcal{O}(|\lambda_{\max}|^k)$ where $\lambda_{\max}$ is the eigenvalue of $F'(\theta^*, \psi^*)$ with the biggest absolute value. If all eigenvalues $F'(\theta^*, \psi^*)$ are on the unit circle, the algorithm can be convergent, divergent or neither, but if it is convergent it will generally converge with a sublinear rate. A similar result (Khalil, 1996; Nagarajan & Kolter, 2017) also holds for the (idealized) continuous system

$$\begin{pmatrix} \dot\theta(t) \\ \dot\psi(t) \end{pmatrix} = \begin{pmatrix} -\nabla_\psi L(\theta, \psi) \\ \nabla_\theta L(\theta, \psi) \end{pmatrix} \tag{3}$$

which corresponds to training the GAN with infinitely small learning rate: if all eigenvalues of the Jacobian $v'(\theta^*, \psi^*)$ at a stationary point (θ^*, ψ^*) have negative real-part, the continuous system converges locally to (θ^*, ψ^*) with linear convergence rate. On the other hand, if $v'(\theta^*, \psi^*)$ has eigenvalues with positive real-part, the continuous system is not locally convergent. If all eigenvalues have zero real-part, it can be convergent, divergent or neither, but if it is convergent, it will generally converge with a sublinear rate.

For simultaneous gradient descent linear convergence can be achieved if and only if all eigenvalues of the Jacobian of the gradient vector field $v(\theta, \psi)$ have negative real part (Mescheder et al., 2017). This situation was also considered by Nagarajan & Kolter (2017) who examined the asymptotic case of step sizes h that go to 0 and proved local convergence for absolutely continuous generator and data distributions under certain regularity assumptions.

2.2. The Dirac-GAN

> Simple experiments, simple theorems are the building blocks that help us understand more complicated systems.
>
> *Ali Rahimi - Test of Time Award speech, NIPS 2017*

In this section, we describe a simple yet prototypical counterexample which shows that in the general case, unregularized GAN training is neither locally nor globally convergent.

Definition 2.1. *In the Dirac-GAN, the true (univariate) data distribution $p_{\mathcal{D}}$ is given by $p_{\mathcal{D}} = \delta_0$ and the generator is given by $p_\theta = \delta_\theta$. The discriminator is given given by a linear function: $D_\psi(x) = \psi \cdot x$.*

Note that in the Dirac-GAN, both the generator and the discriminator have exactly one parameter. This situation is visualized in Figure 1. In this setup, the GAN training objective (1) is given by

$$L(\theta, \psi) = f(\psi\theta) + f(0) \tag{4}$$

While using linear discriminators might appear restrictive, the class of linear discriminators is in fact as powerful as the class of all real-valued functions for this example: when we use $f(t) = -\log(1 + \exp(-t))$ and we take the supremum over ψ in (4), we obtain (up to scalar and additive constants) the Jensen-Shannon divergence between p_θ and $p_{\mathcal{D}}$. The same holds true for the Wasserstein-divergence, when we use $f(t) = t$ and put a Lipschitz constraint on the discriminator (see Section 3.1).

We show that the training dynamics of GANs lead to divergent behavior in this simple setup.

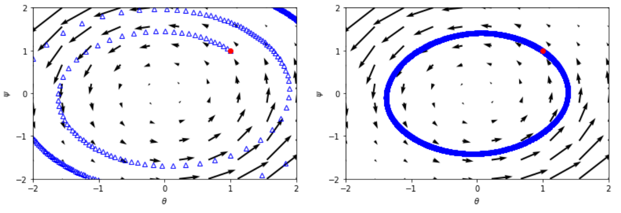
Lemma 2.2. *The unique equilibrium point of the training objective in (4) is given by $\theta = \psi = 0$. Moreover, the Jacobian of the gradient vector field at the equilibrium point has the two eigenvalues $\pm f'(0)\, i$ which are both on the imaginary axis.*

We now take a closer look at the training dynamics produced by various algorithms for training the Dirac-GAN. First, we consider the (idealized) continuous system in (3): while Lemma 2.2 shows that the continuous system is generally not linearly convergent to the equilibrium point, it could in principle converge with a sublinear convergence rate. However, this is not the case as the next lemma shows:

Lemma 2.3. *The integral curves of the gradient vector field $v(\theta, \psi)$ do not converge to the Nash-equilibrium. More specifically, every integral curve $(\theta(t), \psi(t))$ of the gradient vector field $v(\theta, \psi)$ satisfies $\theta(t)^2 + \psi(t)^2 = const$ for all $t \in [0, \infty)$.*

Note that our results do not contradict the results of Nagarajan & Kolter (2017) and Heusel et al. (2017): our example violates Assumption IV in Nagarajan & Kolter (2017) that the support of the generator distribution is equal to the support of the true data distribution near the equilibrium. It also violates the assumption in Heusel et al. (2017) that the optimal discriminator parameter vector is a continuous function of the current generator parameters[2]. In fact, unless $\theta = 0$,

[2] This assumption is usually even violated by Wasserstein-GANs, as the optimal discriminator parameter vector as a function of the current generator parameters can have discontinuities near the Nash-equilibrium. See Section 3.1 for details.



(a) SimGD (b) AltGD

Figure 2. Training behavior of the Dirac-GAN. The starting iterate is marked in red.

there is not even an optimal discriminator parameter vector for the Dirac-GAN. Indeed, we find that two-time scale updates as suggested by Heusel et al. (2017) do not help convergence towards the Nash-equilibrium (see Figure 22 in the supplementary material). However, our example seems to be a prototypical situation for (unregularized) GAN training which usually deals with distributions that are concentrated on lower dimensional manifolds (Arjovsky & Bottou, 2017).

We now take a closer look at the *discretized system*.

Lemma 2.4. *For simultaneous gradient descent, the Jacobian of the update operator $F_h(\theta, \psi)$ has eigenvalues $\lambda_{1/2} = 1 \pm hf'(0)i$ with absolute values $\sqrt{1 + h^2 f'(0)^2}$ at the Nash-equilibrium. Independently of the learning rate, simultaneous gradient descent is therefore not stable near the equilibrium. Even stronger, for every initial condition and learning rate $h > 0$, the norm of the iterates (θ_k, ψ_k) obtained by simultaneous gradient descent is monotonically increasing.*

The behavior of simultaneous gradient descent on our example problem is visualized in Figure 2a.

Similarly, for alternating gradient descent we have

Lemma 2.5. *For alternating gradient descent with n_g generator and n_d discriminator updates, the Jacobian of the update operator $F_h(\theta, \psi)$ has eigenvalues*

$$\lambda_{1/2} = 1 - \frac{\alpha^2}{2} \pm \sqrt{\left(1 - \frac{\alpha^2}{2}\right)^2 - 1}. \tag{5}$$

with $\alpha := \sqrt{n_g n_d} hf'(0)$. For $\alpha \leq 2$, all eigenvalues are hence on the unit circle. Moreover for $\alpha > 2$, there are eigenvalues outside the unit circle.

Even though Lemma 2.5 shows that alternating gradient descent does not converge linearly to the Nash-equilibrium, it could in principle converge with a sublinear convergence rate. However, this is very unlikely because – as Lemma 2.3 shows – even the continuous system does not converge. Indeed, we empirically found that alternating gradient descent oscillates in stable cycles around the equilibrium and shows no sign of convergence (Figure 2b).

2.3. Where do instabilities come from?

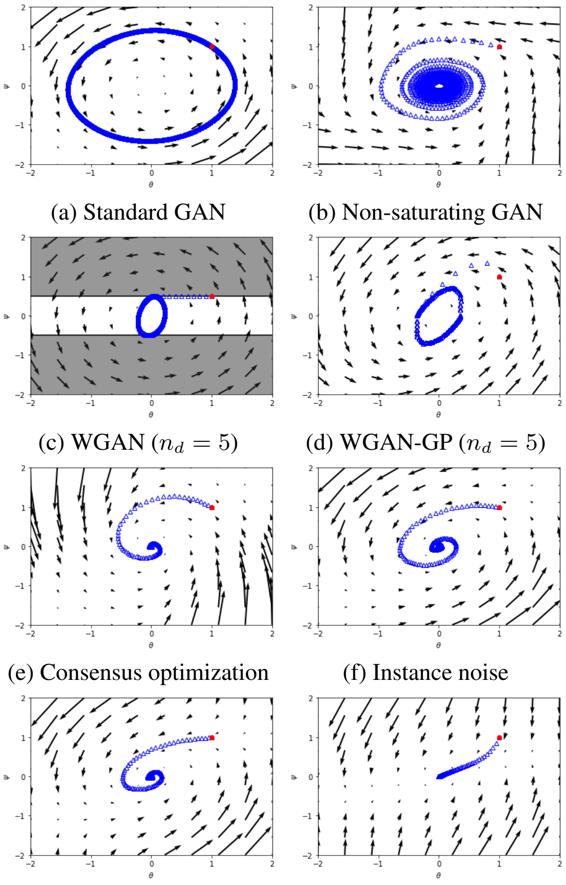
Our simple example shows that naive gradient based GAN optimization does not always converge to the equilibrium point. To get a better understanding of what can go wrong for more complicated GANs, it is instructive to analyze these instabilities in depth for this simple example problem.

To understand the instabilities, we have to take a closer look at the oscillatory behavior that GANs exhibit both for the Dirac-GAN and for more complex systems. An intuitive explanation for the oscillations is given in Figure 1: when the generator is far from the true data distribution, the discriminator pushes the generator towards the true data distribution. At the same time, the discriminator becomes more certain, which increases the discriminator's slope (Figure 1a). Now, when the generator reaches the target distribution (Figure 1b), the slope of the discriminator is largest, pushing the generator away from the target distribution. As a result, the generator moves away again from the true data distribution and the discriminator has to change its slope from positive to negative. After a while, we end up with a similar situation as in the beginning of training, only on the other side of the true data distribution. This process repeats indefinitely and does not converge.

Another way to look at this is to consider the local behavior of the training algorithm near the Nash-equilibrium. Indeed, near the Nash-equilibrium, there is nothing that pushes the discriminator towards having zero slope on the true data distribution. Even if the generator is initialized *exactly* on the target distribution, there is no incentive for the discriminator to move to the equilibrium discriminator. As a result, training is unstable near the equilibrium point.

This phenomenon of discriminator gradients orthogonal to the data distribution can also arise for more complex examples: as long as the data distribution is concentrated on a low dimensional manifold and the class of discriminators is big enough, there is no incentive for the discriminator to produce zero gradients orthogonal to the tangent space of the data manifold and hence converge to the equilibrium discriminator. Even if the generator produces *exactly* the true data distribution, there is no incentive for the discriminator to produce zero gradients orthogonal to the tangent space. When this happens, the discriminator does not provide useful gradients for the generator orthogonal to the data distribution and the generator does not converge.

Note that these instabilities can only arise if the true data distribution is concentrated on a lower dimensional manifold. Indeed, Nagarajan & Kolter (2017) showed that - under some suitable assumptions - gradient descent based GAN optimization is locally convergent for absolutely continuous distributions. Unfortunately, this assumption may not be satisfied for data distributions like natural images to



(a) Standard GAN

(b) Non-saturating GAN

(c) WGAN ($n_d = 5$)

(d) WGAN-GP ($n_d = 5$)

(e) Consensus optimization

(f) Instance noise

(g) Gradient penalty

(h) Gradient penalty (CR)

Figure 3. Convergence properties of different GAN training algorithms using alternating gradient descent with recommended number of discriminator updates per generator update ($n_d = 1$ if not noted otherwise). The shaded area in Figure 3c visualizes the set of forbidden values for the discriminator parameter ψ. The starting iterate is marked in red.

which GANs are commonly applied (Arjovsky & Bottou, 2017). Moreover, even if the data distribution is absolutely continuous but concentrated along some lower dimensional manifold, the eigenvalues of the Jacobian of the gradient vector field will be very close to the imaginary axis, resulting in a highly ill-conditioned problem. This was observed by Mescheder et al. (2017) who examined the spectrum of the Jacobian for a data distribution given by a circular mixture of Gaussians with small variance.

3. Regularization strategies

As we have seen in Section 2, unregularized GAN training does not always converge to the Nash-equilibrium. In this section, we discuss how several regularization techniques that have recently been proposed, influence convergence of the Dirac-GAN.

Interestingly, we also find that the non-saturating loss proposed in the original GAN paper (Goodfellow et al., 2014) leads to convergence of the continuous system, albeit with an extremely slow convergence rate. A more detailed discussion and an analysis of Consensus optimization (Mescheder et al., 2017) can be found in the supplementary material.

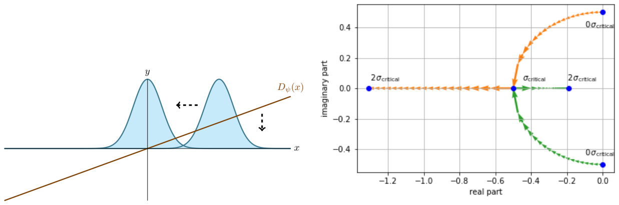
3.1. Wasserstein GAN

The two-player GAN game can be interpreted as minimizing a probabilistic divergence between the true data distribution and the distribution produced by the generator (Nowozin et al., 2016; Goodfellow et al., 2014). This divergence is obtained by considering the best-response strategy for the discriminator, resulting in an objective function that only contains the generator parameters. Many recent regularization techniques for GANs are based on the observation (Arjovsky & Bottou, 2017) that this divergence may be discontinuous with respect to the parameters of the generator or may even take on infinite values if the support of the data distribution and the generator distribution do not match.

To make the divergence continuous with respect to the parameters of the generator, Wasserstein GANs (WGANs) Arjovsky et al. (2017) replace the Jensen-Shannon divergence used in the original derivation of GANs (Goodfellow et al., 2014) with the Wasserstein-divergence. As a result, Arjovsky et al. (2017) propose to use $f(t) = t$ and restrict the class of discriminators to Lipschitz continuous functions with Lipschitz constant equal to some $g_0 > 0$. While a WGAN converges if the discriminator is always trained until convergence, in practice WGANs are usually trained by running only a fixed finite number of discriminator updates per generator update. However, near the Nash-equilibrium the optimal discriminator parameters can have a discontinuity as a function of the current generator parameters: in the Dirac-GAN, the optimal discriminator has to move from $\psi = -1$ to $\psi = 1$ when θ changes signs. As the gradients get smaller near the equilibrium point, the gradient updates do not lead to convergence for the discriminator. Overall, the training dynamics are again determined by the Jacobian of the gradient vector field near the Nash-equilibrium:

Lemma 3.1. *WGANs trained with simultaneous or alternating gradient descent with a fixed number of discriminator updates per generator update and a fixed learning rate $h > 0$ do generally not converge to the Nash equilibrium for the Dirac-GAN.*

The training behavior of the WGAN is visualized in Figure 3c. We stress that this analysis only holds if the discriminator is trained with a fixed number of discriminator updates (as it is usually done in practice). More careful training that ensures that the discriminator is kept exactly optimal or two-timescale training (Heusel et al., 2017) might be able to ensure convergence for WGANs.



(a) Example with instance noise

(b) Eigenvalues

Figure 4. Dirac-GAN with instance noise. While unregularized GAN training is inherently unstable, instance noise can stabilize it: (a) Near the Nash-equilibrium, the discriminator is pushed towards the zero discriminator. (b) As we increase the noise level σ from 0 to σ_{critical}, the real part of the eigenvalues at the equilibrium point becomes negative and the absolute value of the imaginary part becomes smaller. For noise levels bigger than σ_{critical} all eigenvalues are real-valued and GAN training hence behaves like a normal optimization problem.

The convergence properties of WGANs were also considered by Nagarajan & Kolter (2017) who showed that even for absolutely continuous densities and infinitesimal learning rates, WGANs are not always locally convergent.

We also found that WGAN-GP (Gulrajani et al., 2017) does not converge for the Dirac-GAN (Figure 3d). Please see the supplementary material for details.[3]

3.2. Instance noise

A common technique to stabilize GANs is to add *instance noise* (Sønderby et al., 2016; Arjovsky & Bottou, 2017), i.e. independent Gaussian noise, to the data points. While the original motivation was to make the probabilistic divergence between data and generator distribution well-defined for distributions that do not have common support, this does not clarify the effects of instance noise on the *training algorithm* itself and its ability to find a Nash-equilibrium. Interestingly, however, it was recently shown (Nagarajan & Kolter, 2017) that in the case of absolutely continuous distributions, gradient descent based GAN optimization is - under suitable assumptions - locally convergent.

Indeed, for the Dirac-GAN we have:

Lemma 3.2. *When using Gaussian instance noise with standard deviation σ, the eigenvalues of the Jacobian of the gradient vector field are given by*

$$\lambda_{1/2} = f''(0)\sigma^2 \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}. \tag{6}$$

In particular, all eigenvalues of the Jacobian have negative real-part at the Nash-equilibrium if $f''(0) < 0$ and $\sigma > 0$. Hence, simultaneous and alternating gradient descent are both locally convergent for small enough learning rates.

Interestingly, Lemma 3.2 shows that there is a critical noise level given by $\sigma^2_{\text{critical}} = |f'(0)|/|f''(0)|$. If the noise level is smaller than the critical noise level, the eigenvalues of the Jacobian have non-zero imaginary part which results in a rotational component in the gradient vector field near the equilibrium point. If the noise level is larger than the critical noise level, all eigenvalues of the Jacobian become real-valued and the rotational component in the gradient vector field disappears. The optimization problem is best behaved when we select $\sigma = \sigma_{\text{critical}}$: in this case we can even achieve quadratic convergence for $h = |f'(0)|^{-1}$. The effect of instance noise on the eigenvalues is visualized in Figure 4b, which shows the traces of the two eigenvalues as we increase σ from 0 to $2\sigma_{\text{critical}}$.

Figure 3f shows the training behavior of the GAN with instance noise, showing that instance noise indeed creates a strong radial component in the gradient vector field which makes the training algorithm converge.

3.3. Zero-centered gradient penalties

Motivated by the success of instance noise to make the f-divergence between two distributions well-defined, Roth et al. (2017) derived a local approximation to instance noise that results in a zero-centered[4] gradient penalty for the discriminator.

In our simple example, a penalty on the squared norm of the gradients of the discriminator (no matter where) results in the regularizer

$$R(\psi) = \frac{\gamma}{2}\psi^2. \tag{7}$$

This regularizer does not include the weighting terms considered by Roth et al. (2017). However, the same analysis can also be applied to the regularizer with the additional weighting, yielding almost exactly the same results (see Section D.2 of the supplementary material).

Lemma 3.3. *The eigenvalues of the Jacobian of the gradient vector field for the gradient-regularized GAN at the equilibrium point are given by*

$$\lambda_{1/2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - f'(0)^2}. \tag{8}$$

In particular, for $\gamma > 0$ all eigenvalues have negative real part. Hence, simultaneous and alternating gradient descent are both locally convergent for small enough learning rates.

As for instance noise, there is a critical regularization parameter $\gamma_{\text{critical}} = 2|f'(0)|$ that results in a locally rotation free vector field. A visualization of the training behavior of the Dirac-GAN with gradient penalty is shown in Figure 3g. Figure 3h illustrates the training behavior of the GAN with

[3] Despite these negative results, WGAN-GP has been successfully applied in practice (Gulrajani et al., 2017; Karras et al., 2017) and we leave a theoretical analysis of these empirical results to future research.

[4] In contrast to the gradient regularizers used in WGAN-GP and DRAGAN which are not zero-centered.

gradient penalty and critical regularization (CR). In particular, we see that near the Nash-equilibrium the vector field does not have a rotational component anymore and hence behaves like a normal optimization problem.

4. General convergence results

In Section 3 we analyzed the convergence properties of various regularization strategies for the Dirac-GAN. In this section, we consider general GAN problems. First, we introduce two simplified versions of the zero-centered gradient penalty proposed by Roth et al. (2017). We then show that these gradient penalties allow us to extend the convergence proof by Nagarajan & Kolter (2017) to the case where the generator and data distribution do not locally have the same support.[5] As a result, our convergence proof for the regularized training dynamics also holds for the more realistic case where both the generator and data distributions may lie on lower dimensional manifolds.

4.1. Simplified gradient penalties

Our analysis suggests that the main effect of the zero-centered gradient penalties proposed by Roth et al. (2017) on local stability is to penalize the discriminator for deviating from the Nash-equilibrium. The simplest way to achieve this is to penalize the gradient on real data alone: when the generator distribution produces the true data distribution and the discriminator is equal to 0 on the data manifold, the gradient penalty ensures that the discriminator cannot create a non-zero gradient orthogonal to the data manifold without suffering a loss in the GAN game.

This leads to the following regularization term:

$$R_1(\psi) := \frac{\gamma}{2} \mathrm{E}_{p_\mathcal{D}(x)} \left[\|\nabla D_\psi(x)\|^2 \right]. \tag{9}$$

Note that this regularizer is a simplified version of to the regularizer derived by Roth et al. (2017). However, our regularizer does not contain the additional weighting terms and penalizes the discriminator gradients only on the true data distribution.

We also consider a similar regularization term given by

$$R_2(\theta, \psi) := \frac{\gamma}{2} \mathrm{E}_{p_\theta(x)} \left[\|\nabla D_\psi(x)\|^2 \right] \tag{10}$$

where we penalize the discriminator gradients on the current generator distribution instead of the true data distribution.

Note that on the Dirac-GAN from Section 2, both regularizers reduce to the gradient penalty from Section 3.3 whose behavior is visualized in Figure 3g and Figure 3h.

4.2. Convergence

In this section we present convergence results for the regularized GAN-training dynamics for both regularization terms $R_1(\psi)$ and $R_2(\psi)$ under some suitable assumptions.[6]

Let (θ^*, ψ^*) denote an equilibrium point of the regularized training dynamics. In our convergence analysis, we consider the realizable case, i.e. we assume that there are generator parameters that make the generator produce the true data distribution:

Assumption I. *We have $p_{\theta^*} = p_\mathcal{D}$ and $D_{\psi^*}(x) = 0$ in some local neighborhood of* $\operatorname{supp} p_\mathcal{D}$.

Like Nagarajan & Kolter (2017), we assume that f satisfies the following property:

Assumption II. *We have $f'(0) \neq 0$ and $f''(0) < 0$.*

An extension of our convergence proof for $f(t) = t$ (as in WGANs) can be found in the supplementary material.

The convergence proof is complicated by the fact that for neural networks, there generally is not a single equilibrium point (θ^*, ψ^*), but a submanifold of equivalent equilibria corresponding to different parameterizations of the same function. We therefore define the *reparameterization manifolds $\mathcal{M}_G$* and $\mathcal{M}_D$. To this end, let

$$h(\psi) := \mathrm{E}_{p_\mathcal{D}(x)} \left[|D_\psi(x)|^2 + \|\nabla_x D_\psi(x)\|^2 \right]. \tag{11}$$

The *reparameterization manifolds* are then defined as

$$\mathcal{M}_G := \{\theta \mid p_\theta = p_\mathcal{D}\} \quad \mathcal{M}_D := \{\psi \mid h(\psi) = 0\}. \tag{12}$$

To prove local convergence, we have to assume some regularity properties for $\mathcal{M}_G$ and $\mathcal{M}_D$ near the equilibrium point. To state these assumptions, we need

$$g(\theta) := \mathrm{E}_{p_\theta(x)} \left[\nabla_\psi D_\psi(x) |_{\psi = \psi^*} \right]. \tag{13}$$

Assumption III. *There are ϵ-balls $B_\epsilon(\theta^*)$ and $B_\epsilon(\psi^*)$ around θ^* and ψ^* so that $\mathcal{M}_G \cap B_\epsilon(\theta^*)$ and $\mathcal{M}_D \cap B_\epsilon(\psi^*)$ define $\mathcal{C}^1$- manifolds. Moreover, the following holds:*

(i) if $v \in \mathbb{R}^n$ is not in the tangent space of $\mathcal{M}_D$ at ψ^*, then $\partial_v^2 h(\psi^*) \neq 0$.
(ii) if $w \in \mathbb{R}^m$ is not in the tangent space of $\mathcal{M}_G$ at θ^*, then $\partial_w g(\theta^*) \neq 0$.

While formally similar, the two conditions in Assumption III have very different meanings: the first condition is a simple regularity property that means that the geometry of $\mathcal{M}_D$ can be locally described by the second derivative of h. The second condition implies that the discriminator is strong

[5] Assumption IV in Nagarajan & Kolter (2017)

[6] Our results also hold for any convex combination of R_1 and R_2 and the regularizer with the additional weighting terms derived by Roth et al. (2017). See the supplementary material for details.

enough so that it can detect any deviation from the equilibrium generator distribution. Indeed, this is the only point where we assume that the class of representable discriminators is sufficiently expressive (and excludes, for example, the trivial case $D_\psi = 0$ for all ψ).

We are now ready to state our main convergence result. To this end, consider the regularized gradient vector field

$$\tilde{v}_i(\theta, \psi) := \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ \nabla_\psi L(\theta, \psi) - \nabla_\psi R_i(\theta, \psi) \end{pmatrix}. \tag{14}$$

Theorem 4.1. *Assume Assumption I, II and III hold for (θ^*, ψ^*). For small enough learning rates, simultaneous and alternating gradient descent for $\tilde{v}_1$ and $\tilde{v}_2$ are both convergent to $\mathcal{M}_G \times \mathcal{M}_D$ in a neighborhood of (θ^*, ψ^*). Moreover, the rate of convergence is at least linear.*

Theorem 4.1 shows that GAN training with our gradient penalties is convergent when initialized sufficiently close the equilibrium point. While this does not show that the method is globally convergent, it at least shows that near the equilibrium the method is well-behaved.

4.3. Stable equilibria for unregularized GAN training

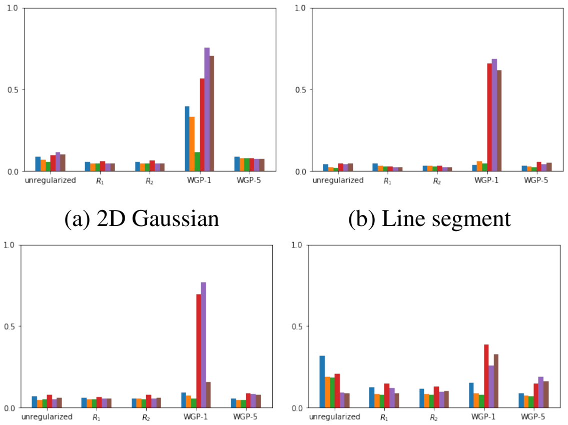
As we have seen in Section 2, unregularized GAN training does not always converge to the Nash-equilibrium. However, this does not rule out the existence of stable equilibria for every GAN architecture. In Section E of the supplementary material, we identify two forms of stable equilibria that may exist for unregularized GAN training (*Energy Solutions* and *Full-Rank Solutions*). However, it is not yet clear under what conditions such solutions exist for high dimensional data distributions.

5. Experiments

2D-Problems Measuring convergence for GANs is hard for high dimensional problems, because we lack an evaluation metric that can reliably detect non-convergent behavior. We therefore first examine the behavior of the different regularizers on simple 2D examples where we can assess convergence using an estimate of the Wasserstein-1-distance.

To this end, we run 5 different training algorithms on 4 different 2D-examples for 6 different GAN architectures. For each method, we try both stochastic gradient descent and RMS-Prop with 4 different learning rates. For the R_1-, R_2- and WGAN-GP-regularizers we try 3 different regularization parameters. We train all methods for 50k iterations and report the results for the best hyperparameter setup. Please see the supplementary material for details.

The results are shown in Figure 5. We see that the R_1- and R_2-regularizers perform similarly and they achieve slightly better results than unregularized training or training with WGAN-GP. In the supplementary material we show that the R_1- and R_2-regularizers find solutions where the discriminator is 0 in a neighborhood of the true data distribution, whereas unregularized training and WGAN-GP converge to *energy solutions* which we define in Section E.1 of the supplementary material.



(a) 2D Gaussian

(b) Line segment

(c) Circle

(d) Four line segments

Figure 5. Wasserstein-1-distance to true data distribution for 4 different 2D-data-distributions, 6 different architectures (small bars) and 5 different training methods. Here, we abbreviate WGAN-GP with 1 and 5 discriminator update(s) per generator update as WGP-1 and WGP-5.

Imagenet To test how well the gradient penalties from Section 4.1 perform on more complicated tasks, we train a convolutional GAN consisting of ResNet-architectures (He et al., 2016) for both the generator and discriminator on the ILSVRC dataset (Russakovsky et al., 2015). While we find that unregularized GAN training quickly leads to mode-collapse on this architecture, our simple R_1-regularizer enables stable training. Some samples from the model after 35 epochs of training and more details on the experimental setup can be found in the supplementary material.

6. Conclusion

In this paper, we analyzed the stability of GAN training on a simple yet prototypical example. Due to the simplicity of the example, we were able to analyze the convergence properties of the training dynamics analytically and we showed that (unregularized) gradient based GAN optimization is not always locally convergent. Our findings also show that WGANs and WGAN-GP do not always lead to local convergence whereas instance noise and zero-centered gradient penalties do. Based on our analysis, we extended our results to more general GANs and we proved local convergence for simplified zero-centered gradient penalties under suitable assumptions. In the future, we would like to extend our theory to the non-realizable case and examine the effect of finite sampling sizes on the GAN training dynamics.

Acknowledgements

We would like to thank Vaishnavh Nagarajan and Kevin Roth for insightful discussions. We also thank Vaishnavh Nagarajan for giving helpful feedback on an early draft of this manuscript. We thank NVIDIA for donating the GPUs for the experiments presented in the supplementary material. This work was supported by Microsoft Research through its PhD Scholarship Programme.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P. A., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016.*, pp. 265–283, 2016.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *CoRR*, abs/1701.04862, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

Barratt, S. and Sharma, R. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Berthelot, D., Schumm, T., and Metz, L. BEGAN: boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717, 2017.

Bertsekas, D. P. *Nonlinear programming*. Athena scientific Belmont, 1999.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5769–5779, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 6629–6640, 2017.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Khalil, H. K. Nonlinear systems. *Prentice-Hall, New Jersey*, 2(5):5–1, 1996.

Kodali, N., Abernethy, J. D., Hays, J., and Kira, Z. How to train your DRAGAN. *CoRR*, abs/1705.07215, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Mescheder, L. M., Nowozin, S., and Geiger, A. The numerics of gans. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 1823–1833, 2017.

Nagarajan, V. and Kolter, J. Z. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5591–5600, 2017.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 271–279, 2016.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 2015–2025, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2226–2234, 2016.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised MAP inference for image super-resolution. *CoRR*, abs/1610.04490, 2016.

Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, 2012.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhao, J. J., Mathieu, M., and LeCun, Y. Energy-based generative adversarial network. *CoRR*, abs/1609.03126, 2016.