Which Training Methods for GANs do actually Converge?

Lars Mescheder\textsuperscript{1}, Sebastian Nowozin\textsuperscript{2}, Andreas Geiger\textsuperscript{1}

\textsuperscript{1}MPI-IS Tübingen \textsuperscript{2}University of Tübingen \textsuperscript{3}Microsoft Research Cambridge
Introduction

Generative neural networks:

Key challenge: have to learn high dimensional probability distribution
Introduction

Generative Adversarial networks (GANs):

\[ z \sim p_0 \rightarrow G_\theta \rightarrow x_{\text{fake}} \]
\[ x_{\text{real}} \sim p_D \rightarrow D_\psi \rightarrow \text{real or fake?} \]

\[
\min_\theta \max_\psi \mathbb{E}_{p_0(z)} f \left( D_\psi \left( G_\theta (z) \right) \right) + \mathbb{E}_{p_D(x)} f \left( -D_\psi (x) \right)
= : L(\theta, \psi)
\]
Generative Adversarial Networks

Alternating gradient descent

1: while not converged do
2: \[ \theta \leftarrow \theta - h \nabla_\theta L(\theta, \psi) \]
3: \[ \psi \leftarrow \psi + h \nabla_\psi L(\theta, \psi) \]
4: end while
Generative Adversarial Networks

Simultaneous gradient descent

1: \textbf{while} not converged \textbf{do}
2: \quad \nu_\theta \leftarrow -\nabla_\theta L(\theta, \psi)
3: \quad \nu_\psi \leftarrow \nabla_\psi L(\theta, \psi)
4: \quad \theta \leftarrow \theta + h \nu_\theta
5: \quad \psi \leftarrow \psi + h \nu_\psi
6: \textbf{end while}
Generative Adversarial Networks

- Does a (pure) Nash-equilibrium exist?
  - Yes, if there is $\theta$ with $p_\theta = p_D$ (Goodfellow et al., 2014)

- Does it solve the min-max problem?
  - Yes, if $p_{\theta^*} = p_D$ (Goodfellow et al., 2014)

- Do simultaneous and / or alternating gradient descent converge to the Nash-equilibrium?
Ist GAN training locally asymptotically stable?

Mescheder et al. (2017): No, if Jacobian of gradient vector field has purely imaginary eigenvalues

Nagarajan and Kolter (2017): Yes, if generator and data distributions locally have the same support

Heusel et al. (2017): Yes, if optimal discriminator parameters are continuous function of generator parameters and two-timescale annealing scheme is adopted

Is GAN training locally asymptotically stable in the general case?
Our Contributions

- **Dirac-GAN:**
  - Unregularized **GAN-training** is *not always stable* when distributions do not have the same support

- **Analysis of common regularizers:**
  - WGAN and WGAN-GP not always stable
  - Instance noise & zero-centered gradient penalties are stable

- **Simplified gradient penalties**
  - *Convergence proof* for realizable case

- **Empirical results:**
  - **High resolution** (1024x1024) **generative models** without progressively growing architectures
Convergence theory

“Simple experiments, simple theorems are the building blocks that help us understand more complicated systems.”

Ali Rahimi – Test of Time Award speech, NIPS 2017
Convergence theory

The Dirac-GAN:

\[ p_D = \delta_0 \quad p_\theta = \delta_\theta \quad D_\psi(x) = \psi \cdot x \]

\[ L(\theta, \psi) = f(\theta \psi) + f(0) \]
Convergence theory

Unregularized GAN training:
Convergence theory

Understanding the \textbf{gradient vector field}:

\[
v(\theta, \psi) = \left(\begin{array}{c}
-\nabla_\theta L(\theta, \psi) \\
\nabla_\psi L(\theta, \psi)
\end{array}\right)
\]

\textbf{Local convergence} of simultaneous and alternating gradient descent determined by \textbf{eigenvalues} of \textbf{Jacobian}

\[
v'(\theta^*, \psi^*) = \left(\begin{array}{cc}
-\nabla^2_\theta L(\theta, \psi) & -\nabla_{\theta, \psi} L(\theta, \psi) \\
\nabla_{\theta, \psi} L(\theta, \psi) & \nabla^2_\psi L(\theta, \psi)
\end{array}\right)
\]
Convergence theory

Continuous system:

![Diagram showing stability regions in the complex plane with arrows indicating the direction of change. The diagram is divided into stable and unstable regions.](image)

- Stable region
- Unstable region
Convergence theory

Continuous system:
Convergence theory

Discretized system:

\[ h = 1.0 \]
Convergence theory

Discretized system:

$h = 0.5$
Which training methods converge?

Unregularized GAN training:

Eigenvalues: \( \{ -f'(0)i, +f'(0)i \} \)
Which training methods converge?

Wasserstein-GAN\(^1\) training:

![Diagram showing eigenvalues: \{-i, +i\}](image)

\(^1\)Arjovsky et al. - Wasserstein GAN (2017)
Which training methods converge?

Wasserstein-GAN\(^1\) training
(5 discriminator updates / generator update):

Eigenvalues: \(\{-i, +i\}\)

\(^1\)Arjovsky et al. - Wasserstein GAN (2017)
Which training methods converge?

Zero-centered Gradient Penalties

\[ R(\psi) := \frac{\gamma}{2} \mathbb{E}_{P_D(x)} \left[ \| \nabla_x D_\psi(x) \|^2 \right] \]

Eigenvalues: \[ \left\{ -\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} - f'(0)^2}, -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - f'(0)^2} \right\} \]
Which training methods converge?

Zero-centered Gradient Penalties (critical)

\[ R(\psi) := \frac{\gamma_{\text{critical}}}{2} \mathbb{E}_{p_D(x)} \left[ \| \nabla_x D_\psi(x) \|^2 \right] \]

Eigenvalues: \( \{-|f'(0)|, -|f'(0)|\} \)
Which training methods converge?

Zero-centered Gradient Penalties

\[ R(\psi) := \frac{\gamma}{2} \mathbb{E}_{p_D(x)} \left[ \| \nabla_x D_\psi(x) \|^2 \right] \]

The diagram illustrates the convergence behavior for different values of \( \gamma \). Fast convergence occurs near the origin, while slow convergence and no convergence are observed at larger values of \( \gamma \).
General convergence results

- Regularizers for discriminator

\[ R_1(\psi) := \frac{\gamma}{2} \mathbb{E}_{p_d}(x) \left[ \| \nabla_x D_\psi(x) \|^2 \right] \]

\[ R_2(\theta, \psi) := \frac{\gamma}{2} \mathbb{E}_{p_\theta(x)} \left[ \| \nabla_x D_\psi(x) \|^2 \right] \]

- Regularized gradient vector field

\[ \tilde{v}_i(\theta, \psi) := \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ \nabla_\psi L(\theta, \psi) - \nabla_\psi R_i(\theta, \psi) \end{pmatrix} \]
General convergence results

**Assumption I:** the generator can represent the true data distribution

**Assumption II:** \( f'(0) \neq 0 \) and \( f''(0) \leq 0 \)

**Assumption III:** the discriminator can detect when the generator deviates from equilibrium

**Assumption IV:** the generator and data distribution have locally the same support (Nagarajan & Kolter)
General convergence results

For the Dirac-GAN:

**Assumption I:** the generator can represent the true data distribution

**Assumption II:** \( f'(0) \neq 0 \) and \( f''(0) \leq 0 \)

**Assumption III:** the discriminator can detect when the generator deviates from equilibrium

**Assumption IV:** the generator and data distribution have locally the same support (Nagarajan & Kolter)

General convergence results

For GANs in the wild:

**Assumption I:** the generator can represent the true data distribution

**Assumption II:** $f'(0) \neq 0$ and $f''(0) \leq 0$

**Assumption III:** the discriminator can detect when the generator deviates from equilibrium

**Assumption IV:** the generator and data distribution have locally the same support (Nagarajan & Kolter)
General convergence results

**Theorem:** under Assumption I, II, III and some mild technical assumptions the GAN training dynamics for the regularized training objective are locally asymptotically stable near the equilibrium point.
General convergence results

Proof (idea): (Extends prior work by Nagarajan & Kolter\(^1\))

\[ \tilde{v}_i(\theta, \psi) := \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ \nabla_\psi L(\theta, \psi) - \nabla_\psi R_i(\theta, \psi) \end{pmatrix} \]

\[ \tilde{v}'(\theta^*, \psi^*) = \begin{pmatrix} 0 & -K^T_{DG} \\ K_{DG} & K_{DD} - L_{DD} \end{pmatrix} \]

- Full column rank
- Negative definite
- (orthogonal to \( \mathcal{M}_G \times \mathcal{M}_D \))
- All eigenvalues of \( \tilde{v}'(\theta^*, \psi^*) \) have negative real part

\(^1\)Nagarajan & Kolter - Gradient descent GAN optimization is locally stable (2017)
Experiments

Imagenet (128 x 128, 1k classes)
Experiments

LSUN bedrooms (256 x 256)
Experiments

LSUN churches (256 x 256)
Experiments

LSUN towers (256 x 256)
Experiments

celebA-HQ (1024 x 1024)
Practical recommendations

- use **alternating** instead of simultaneous **gradient descent**
- **don’t use** **momentum**
- use **regularization** to stabilize the training
- simple **zero-centered gradient penalties** for the discriminator yield excellent results
- **progressively growing architectures** might be **not** all that **important** when using a good regularizer
Poster #77