Motivation

- **GANs** are powerful but **hard to train**
- **Training dynamics** are not completely understood
- Recently, a variety of **techniques** have been proposed to **stabilize GAN training**
- For which training methods can we actually prove local convergence?

The Dirac GAN

GAN-objective: 
\[ L(\theta, \psi) = f(\theta \psi) + f(0) \]

Gradient vector field: 
\[ v(\theta, \psi) = \begin{pmatrix} -\psi f'(\theta \psi) \\ \theta f'(\theta \psi) \end{pmatrix} \]

Zero-centered Gradient Penalties

\[ R_1(\psi) := \frac{\gamma}{2} \mathbb{E}_{p_D(x)} \| \nabla D_\psi(x) \|^2 \quad R_2(\theta, \psi) := \frac{\gamma}{2} \mathbb{E}_{p_W(x)} \| \nabla D_\psi(x) \|^2 \]

**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 the equilibrium

**Assumption IV:** the generator and data distributions have the same support near the equilibrium point (Nagarajan & Kolter, 2017)

**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

Proof (idea): 
(builds on prior work by Nagarajan & Kolter, 2017)

\[ \tilde{v}(\theta^*, \psi^*) = \begin{pmatrix} 0 \\ -K_{CG} \end{pmatrix} \begin{pmatrix} K_{DD} - K_{LD} \end{pmatrix} \]

- full column rank
- negative definite
- all eigenvalues have negative real part

Summary

- **Question:** under what conditions can we guarantee local convergence of GAN training?
- **Negative finding:** unregularized training of GANs and WGANs is not always locally convergent near the equilibrium point
- **Positive finding:** GAN training with instance noise or zero-centered gradient penalties is provably locally convergent in the realizable case
- **Experiments:** simple zero-centered gradient penalties yield excellent results for high-dimensional image distributions

Motivation

**Which Training Methods for GANs do actually Converge?**

Lars Mescheder\textsuperscript{1,2}, Sebastian Nowozin\textsuperscript{3}, Andreas Geiger\textsuperscript{1,2}

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

lars.mescheder@tue.mpg.de

2D-Experiments

Gaussian
Line
Circle
Four lines

GT
unregularized
WGAN-GP (n\textsuperscript{d}=1)

WGAN-GP (n\textsuperscript{d}=5)

R
\textsuperscript{1}
R
\textsuperscript{2}

Qualitative Results

Imagenet (128 x 128, 1k classes)

LSUN-bedroom (256 x 256)

CelebA-HQ (1024 x 1024)