Which Training Methods for GANs do actually Converge?

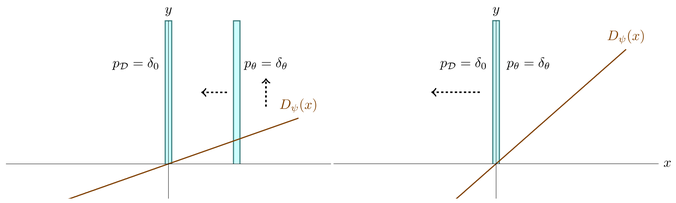
Lars Mescheder[1,2], Sebastian Nowozin[3], Andreas Geiger[1,2]

[1]MPI-IS Tübingen, [2]University of Tübingen, [3]Microsoft Research

Motivation

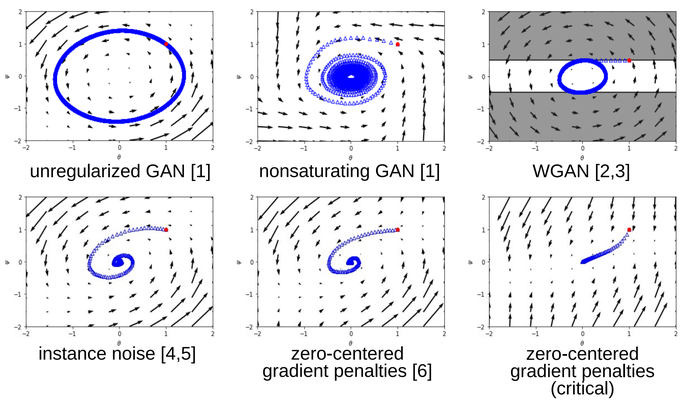
- **GANs** are powerful but **hard to train**
- **Training dynamics** are **not** completely **understood**
- Recently, a variety of **techniques** have been proposed to **stabilize GAN training**
- For which training methods can we actually **prove local convergence**?

The Dirac GAN



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

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



[1] Goodfellow et al. "Generative adversarial nets." (2014)
[2] Arjovsky, Chintala & Bottou. "Wasserstein gan." (2017)
[3] Gulrajani et al. "Improved training of wasserstein gans." (2017)
[4] Arjovsky & Bottou. "Towards principled methods for training generative adversarial networks." (2017)
[5] Sønderby, Casper Kaae, et al. "Amortised map inference for image super-resolution." (2016)
[6] Roth et al. "Stabilizing training of generative adversarial networks through regularization." (2017)

Zero-centered Gradient Penalties

$$R_1(\psi) := \frac{\gamma}{2}\mathrm{E}_{p_{\mathcal{D}}(x)}\left[\|\nabla D_\psi(x)\|^2\right] \quad R_2(\theta, \psi) := \frac{\gamma}{2}\mathrm{E}_{p_\theta(x)}\left[\|\nabla D_\psi(x)\|^2\right]$$

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

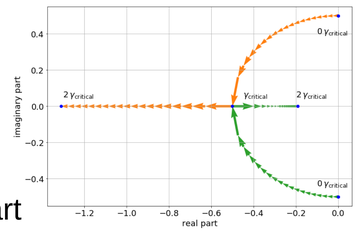
~~**Assumption IV:** the generator and data distributions have the same support near the equilibrium point (Nagarajan & Kolter, 2017)~~

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)

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

K_{DG}: full column rank; $K_{DD} - L_{DD}$: negative definite

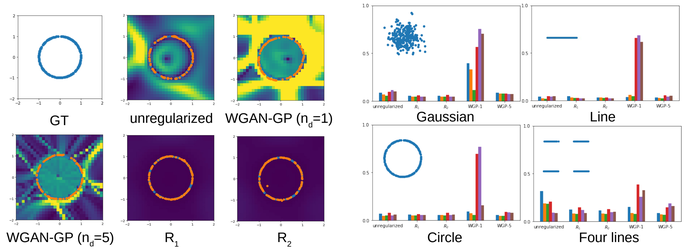


all eigenvalues have negative real part

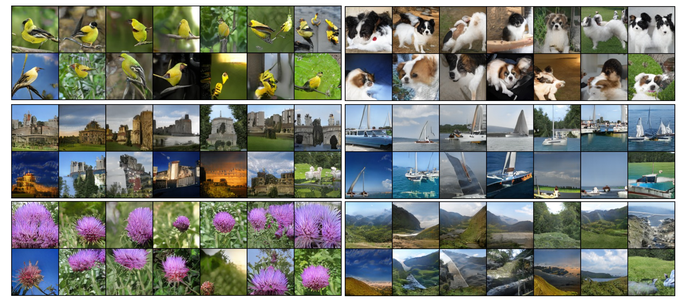
Summary

- **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

2D-Experiments



Qualitative Results



Imagenet (128 x 128, 1k classes)

LSUN-bedroom (256 x 256)

CelebA-HQ (1024 x 1024)

Max Planck Institute for Intelligent Systems – Autonomous Vision Group

lars.mescheder@tue.mpg.de