Which Training Methods for GANs do actually Converge?

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

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

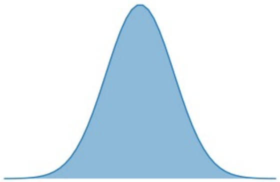
Max Planck Institute for Intelligent Systems
Autonomous Vision Group

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

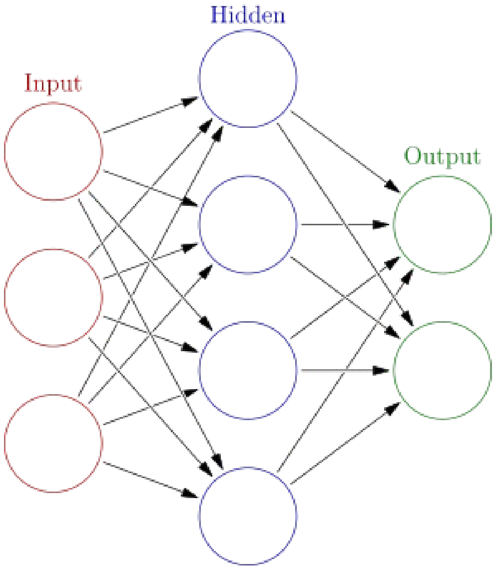
Microsoft

Introduction

Generative neural networks:



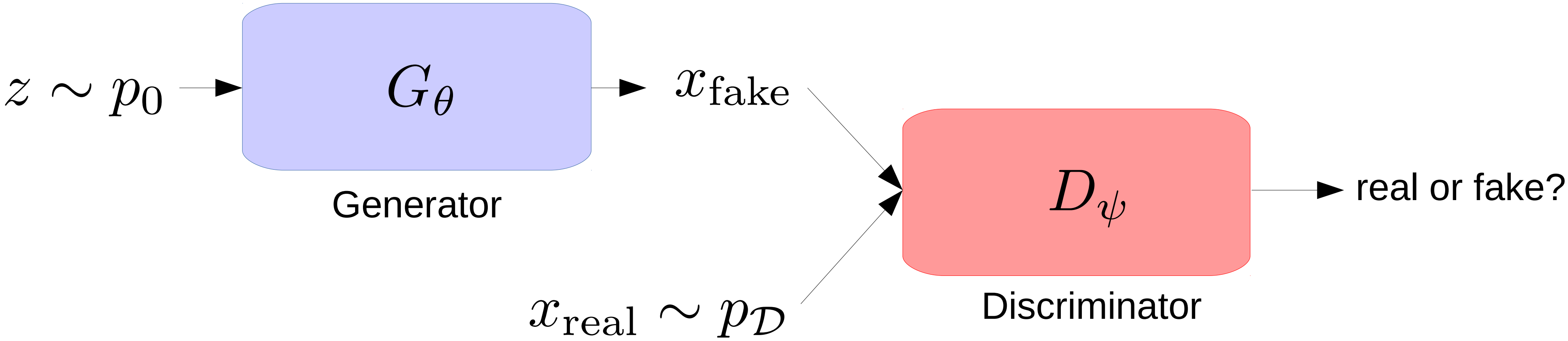
noise z



Key challenge:
have to learn high dimensional probability distribution

Introduction

Generative Adversarial networks (GANs):



$$\min_\theta \max_\psi \underbrace{\mathrm{E}_{p_0(z)} f\left(D_\psi(G_\theta(z))\right) + \mathrm{E}_{p_{\mathcal{D}}(x)} f\left(-D_\psi(x)\right)}_{=:L(\theta,\psi)}$$

Generative Adversarial Networks

Alternating gradient descent

while not converged **do**
$$\theta \leftarrow \theta - h \nabla_\theta L(\theta, \psi)$$
$$\psi \leftarrow \psi + h \nabla_\psi L(\theta, \psi)$$
end while

Generative Adversarial Networks

Simultaneous gradient descent

while not converged **do**
$$v_\theta \leftarrow -\nabla_\theta L(\theta, \psi)$$
$$v_\psi \leftarrow \nabla_\psi L(\theta, \psi)$$
$$\theta \leftarrow \theta + h v_\theta$$
$$\psi \leftarrow \psi + h v_\psi$$
end while

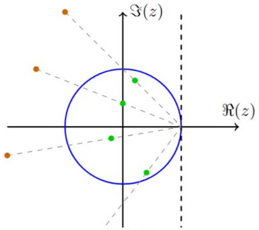
Generative Adversarial Networks

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

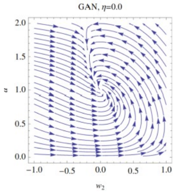
- Does it solve the min-max problem?
  - Yes, if $p_{\theta*} = p_\mathcal{D}$ (Goodfellow et al., 2014)

- $\boxed{\text{Do simultaneous and / or alternating gradient descent converge to the Nash-equilibrium?}}$

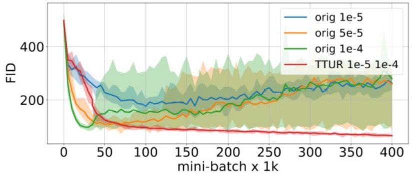
Ist GAN training locally asymptotically stable?



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

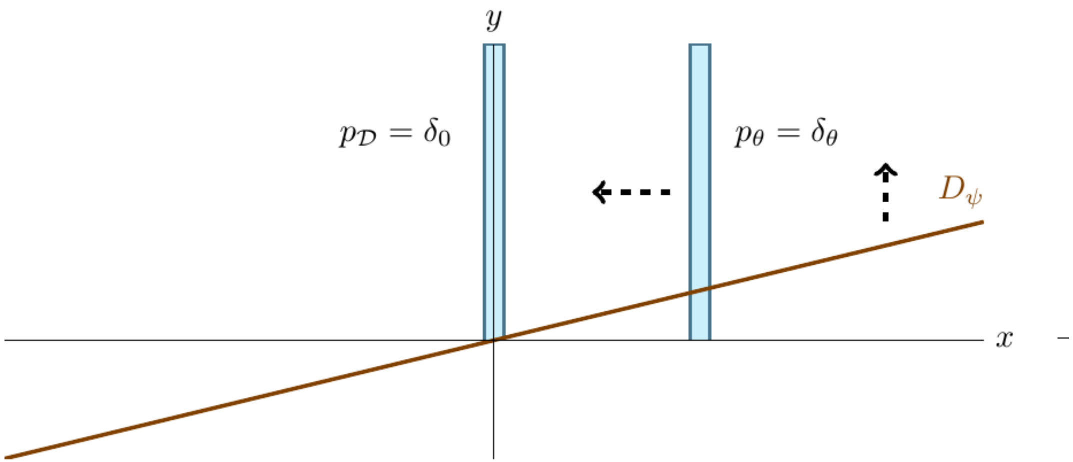
Convergence theory

“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:

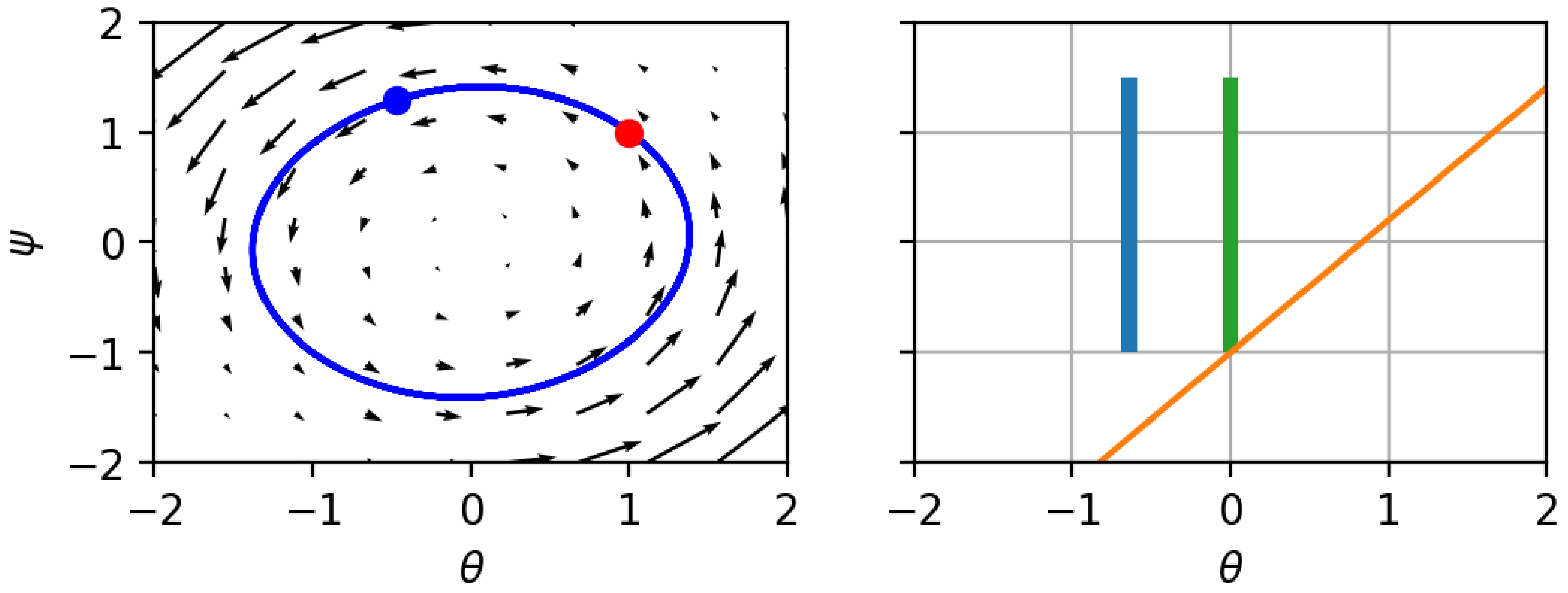


$$p_{\mathcal{D}} = \delta_0 \qquad p_\theta = \delta_\theta \qquad D_\psi(x) = \psi \cdot x$$

$$L(\theta, \psi) = f(\theta\psi) + f(0)$$

Convergence theory

Unregularized GAN training:



Convergence theory

Understanding the **gradient vector field**:

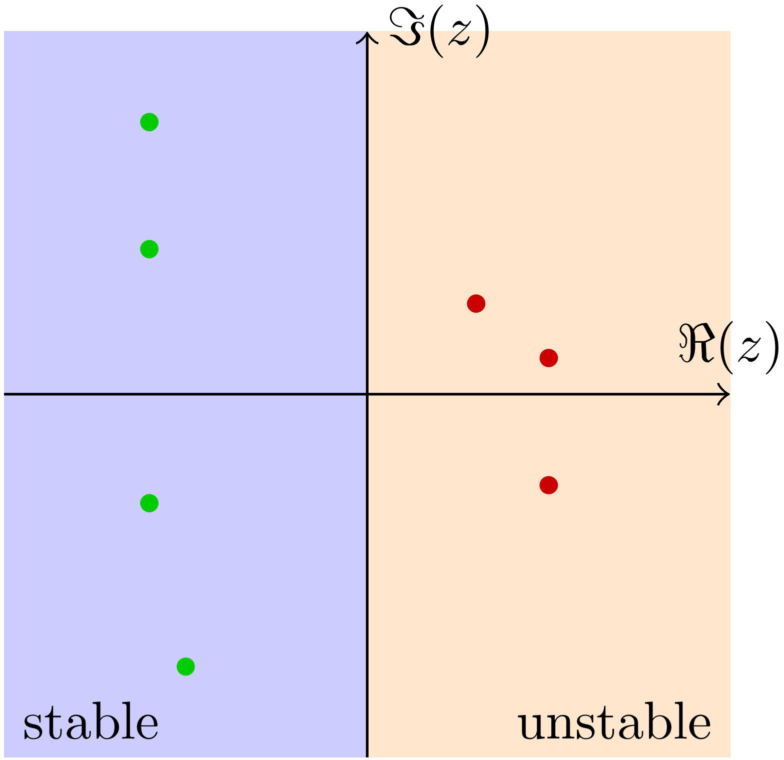
$$v(\theta, \psi) = \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ \nabla_\psi L(\theta, \psi) \end{pmatrix}$$

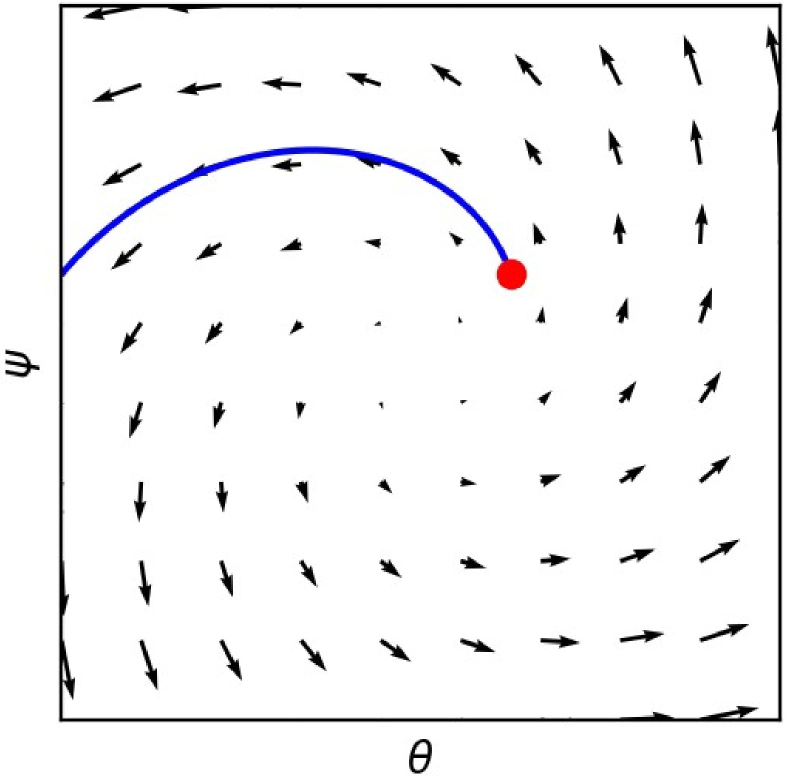
Local convergence of simultaneous and alternating gradient descent determined by **eigenvalues** of **Jacobian**

$$v'(\theta^*, \psi^*) = \begin{pmatrix} -\nabla_\theta^2 L(\theta, \psi) & -\nabla_{\theta,\psi} L(\theta, \psi) \\ \nabla_{\theta,\psi} L(\theta, \psi) & \nabla_\psi^2 L(\theta, \psi) \end{pmatrix}$$

Convergence theory

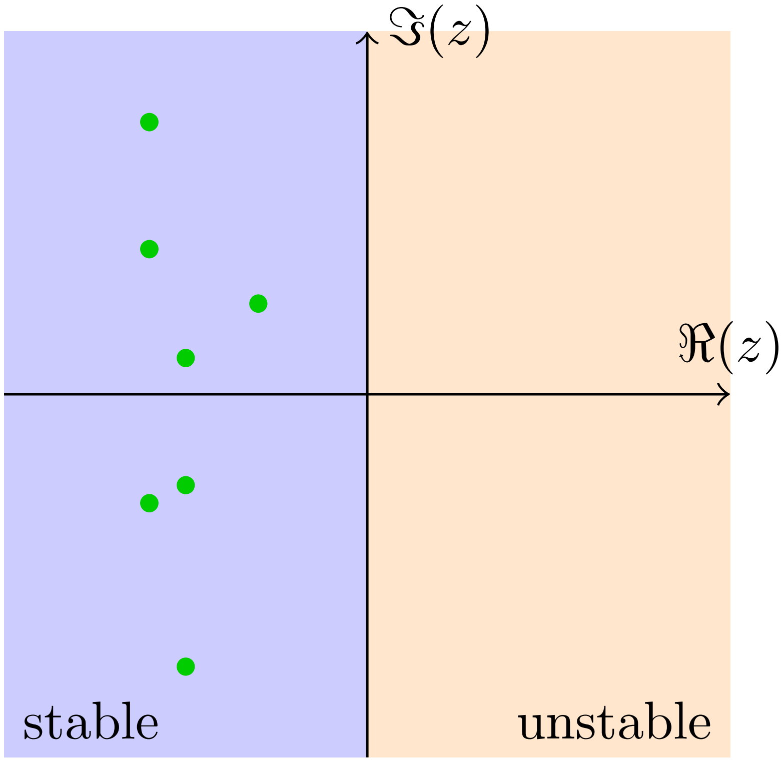
Continuous system:

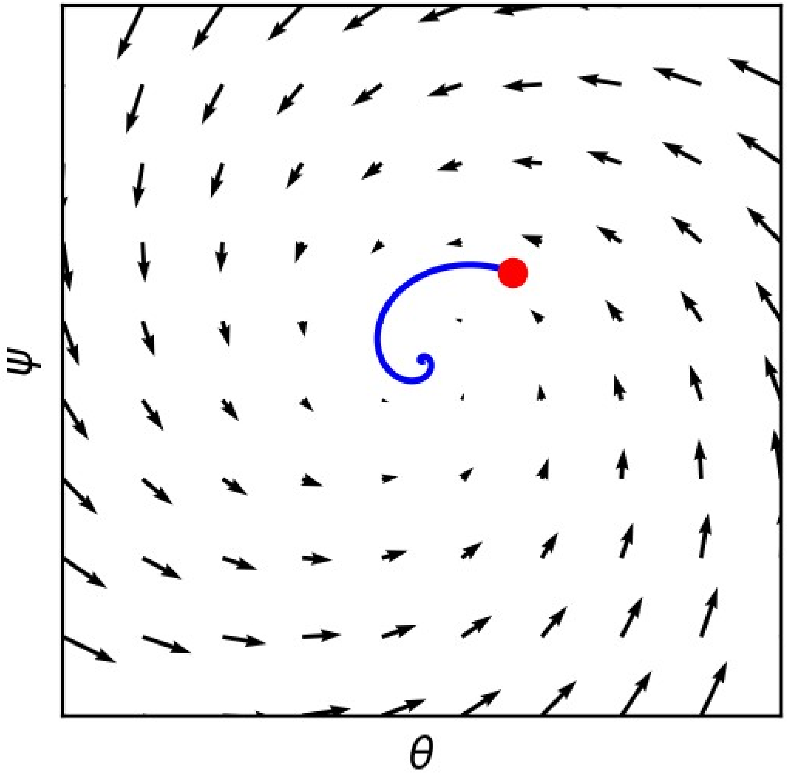




Convergence theory

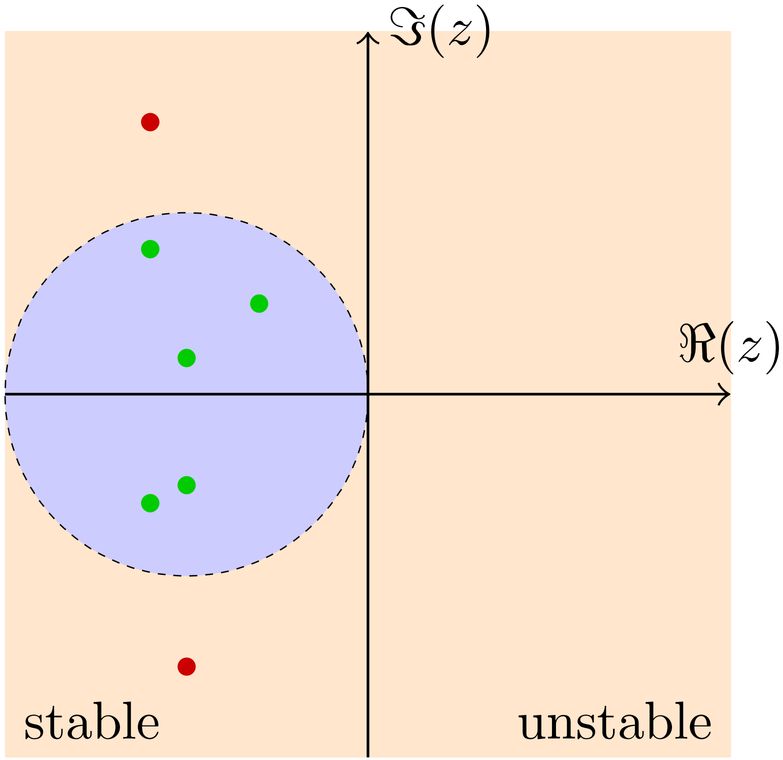
Continuous system:



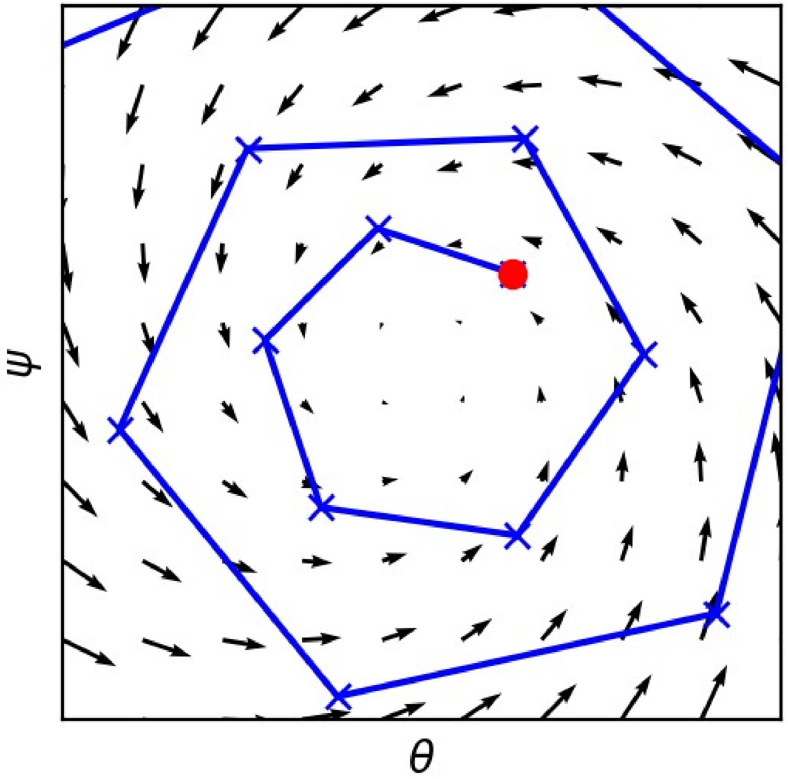


Convergence theory

Discretized system:

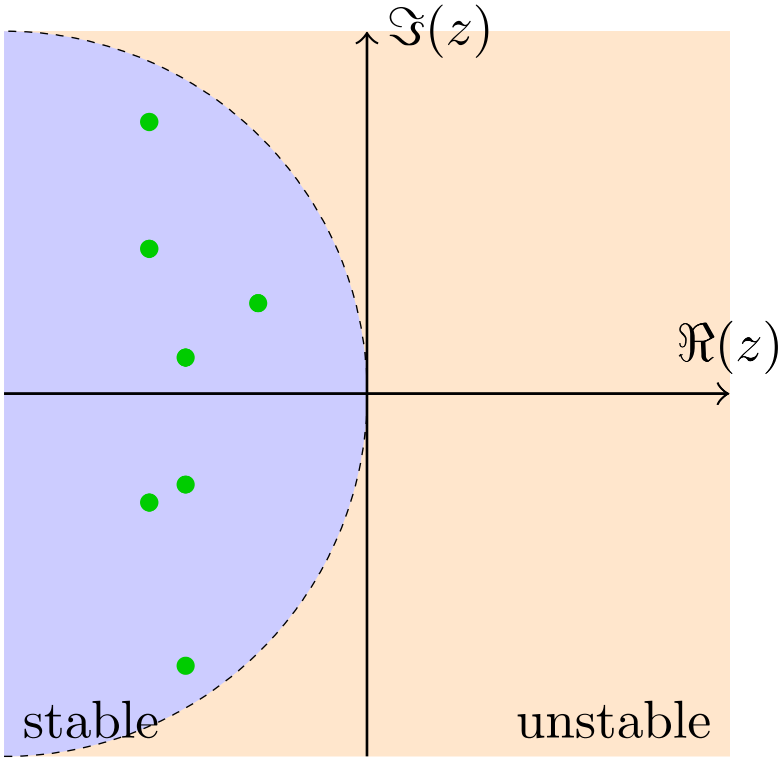


$h = 1.0$

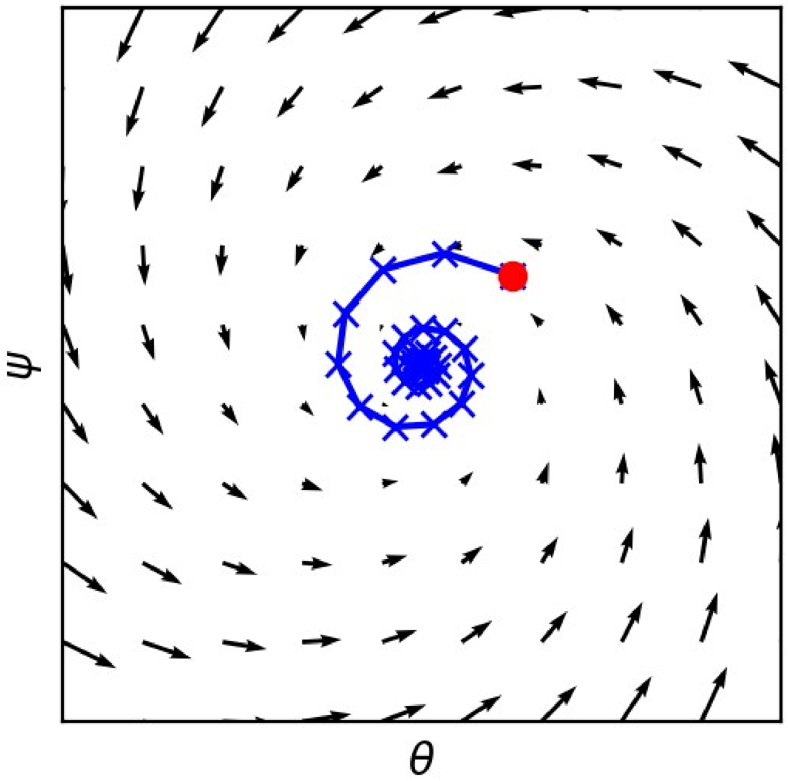


Convergence theory

Discretized system:

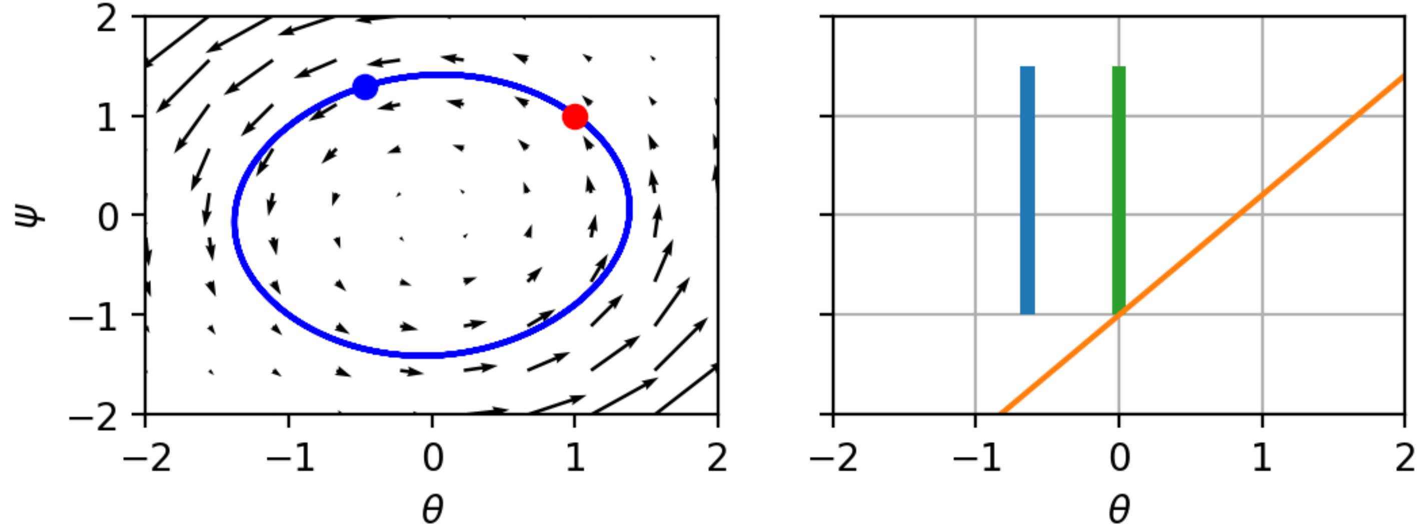


$h = 0.5$



Which training methods converge?

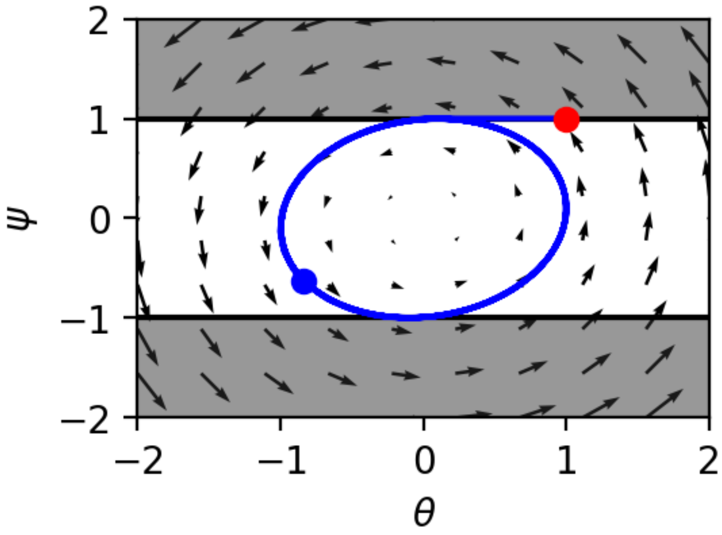
Unregularized GAN training:

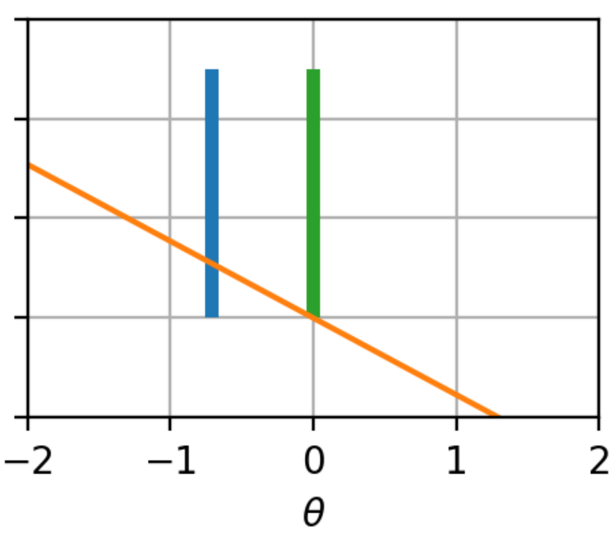


Eigenvalues: $\{-f'(0)i, +f'(0)i\}$

Which training methods converge?

Wasserstein-GAN[1] training:



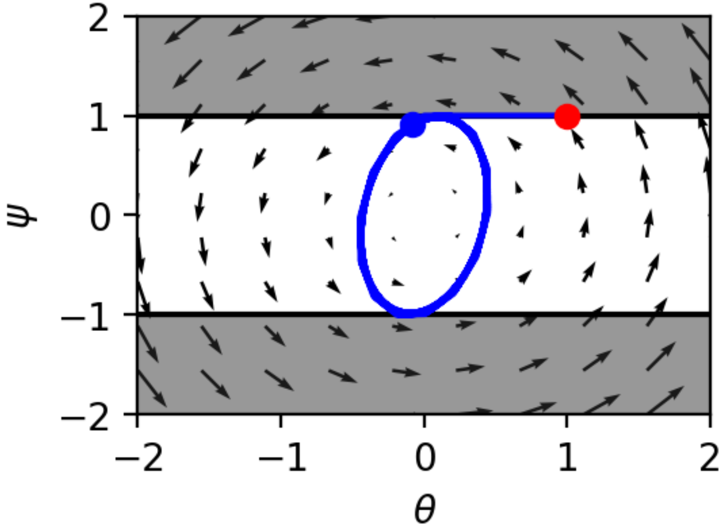


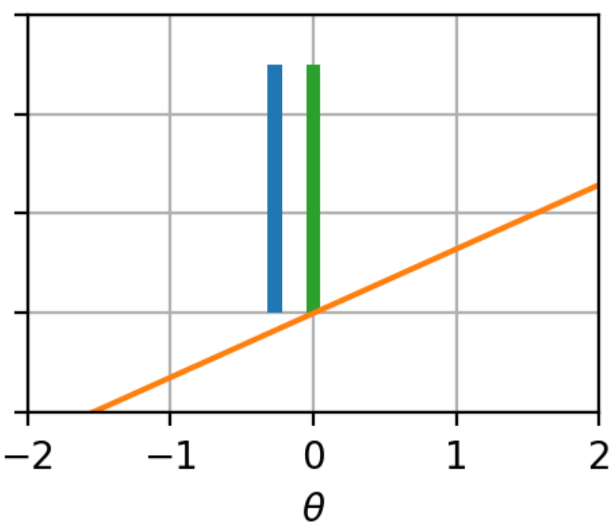
Eigenvalues: $\{-i, +i\}$

[1]Arjovsky et al. - Wasserstein GAN (2017)

Which training methods converge?

Wasserstein-GAN[1] training
(5 discriminator updates / generator update):





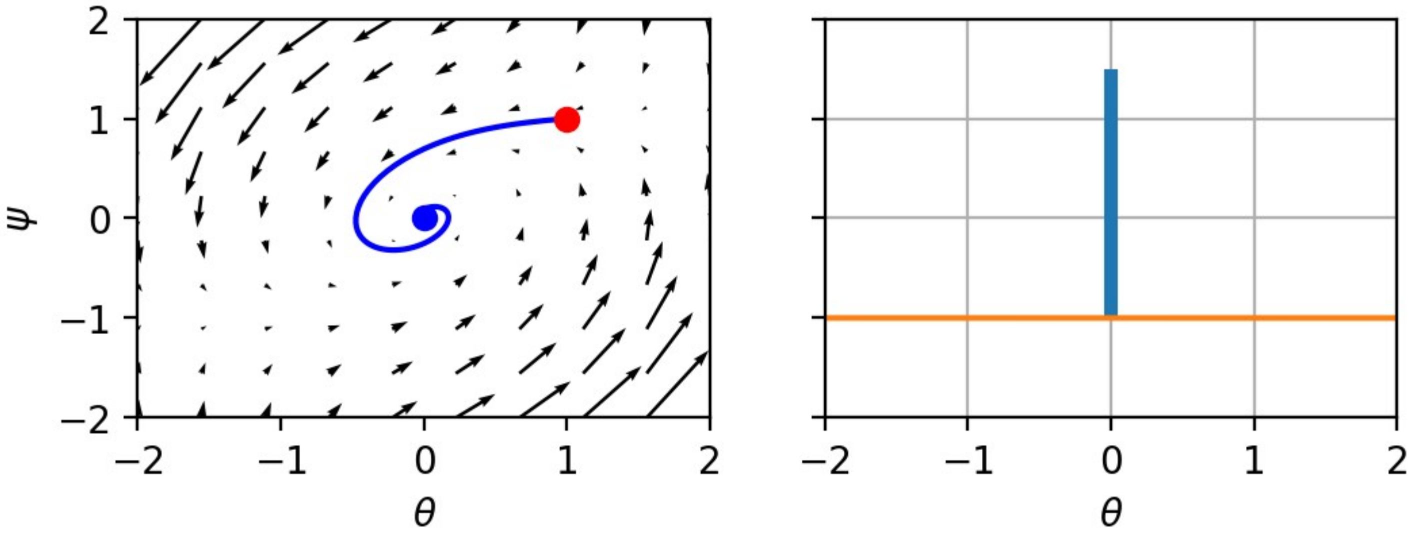
Eigenvalues: $\{-i, +i\}$

[1]Arjovsky et al. - Wasserstein GAN (2017)

Which training methods converge?

Zero-centered Gradient Penalties

$$R(\psi) := \frac{\gamma}{2} \mathrm{E}_{p_{\mathcal{D}}(x)} \left[\|\nabla_x D_\psi(x)\|^2 \right]$$

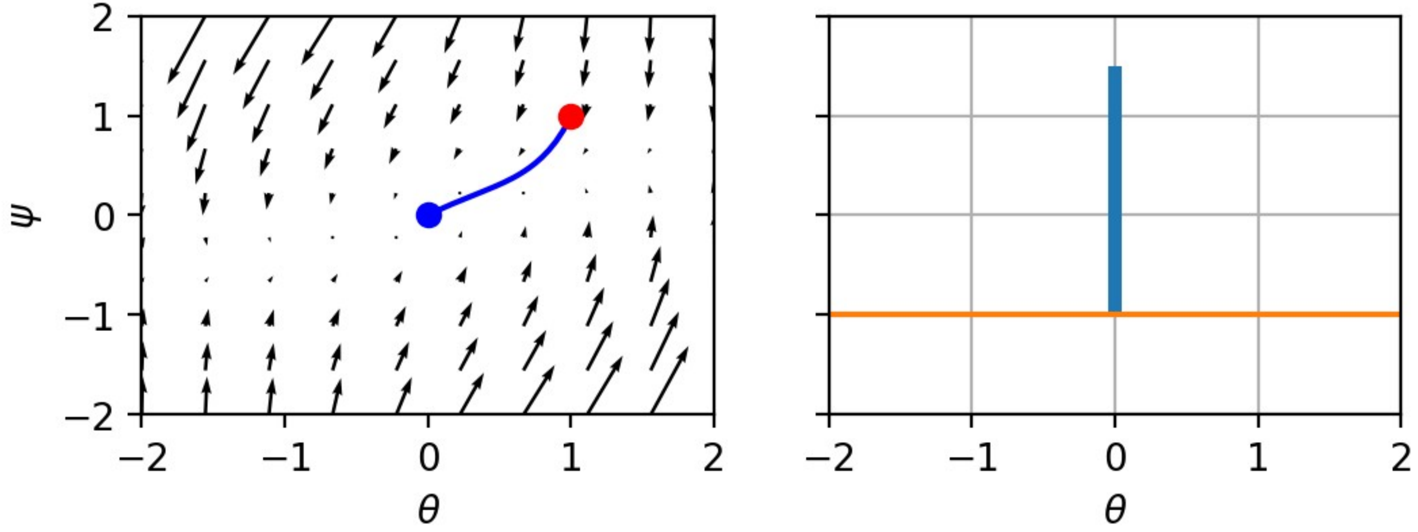


Eigenvalues: $\left\{ -\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} - f'(0)^2}, -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - f'(0)^2} \right\}$

Which training methods converge?

Zero-centered Gradient Penalties (critical)

$$R(\psi) := \frac{\gamma_{\text{critical}}}{2} \mathrm{E}_{p_{\mathcal{D}}(x)} \left[\|\nabla_x D_\psi(x)\|^2 \right]$$

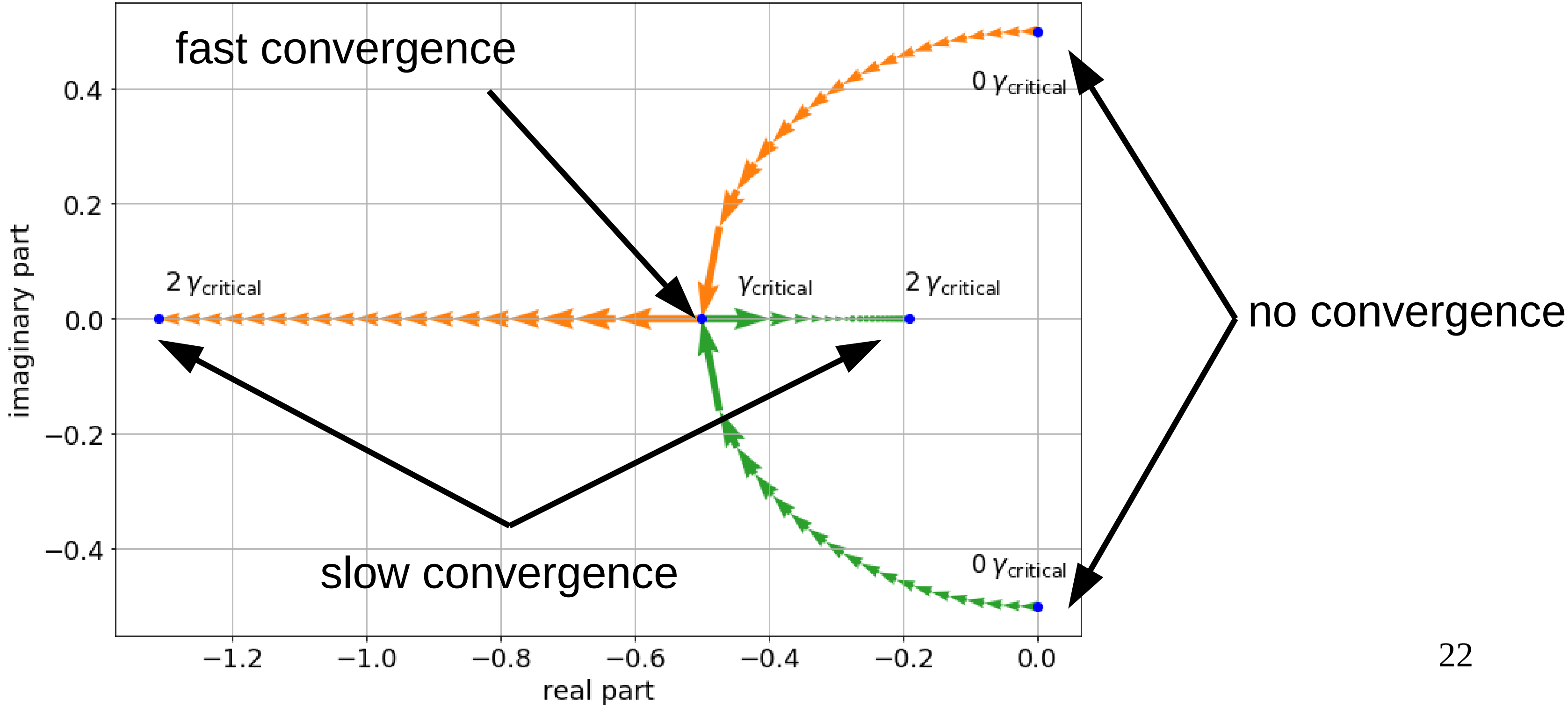


Eigenvalues: $\{-|f'(0)|, -|f'(0)|\}$

Which training methods converge?

Zero-centered Gradient Penalties

$$R(\psi) := \frac{\gamma}{2} \mathrm{E}_{p_{\mathcal{D}}(x)} \left[\|\nabla_x D_\psi(x)\|^2 \right]$$



General convergence results

- Regularizers for discriminator

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

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

- 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])

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

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

Full column rank

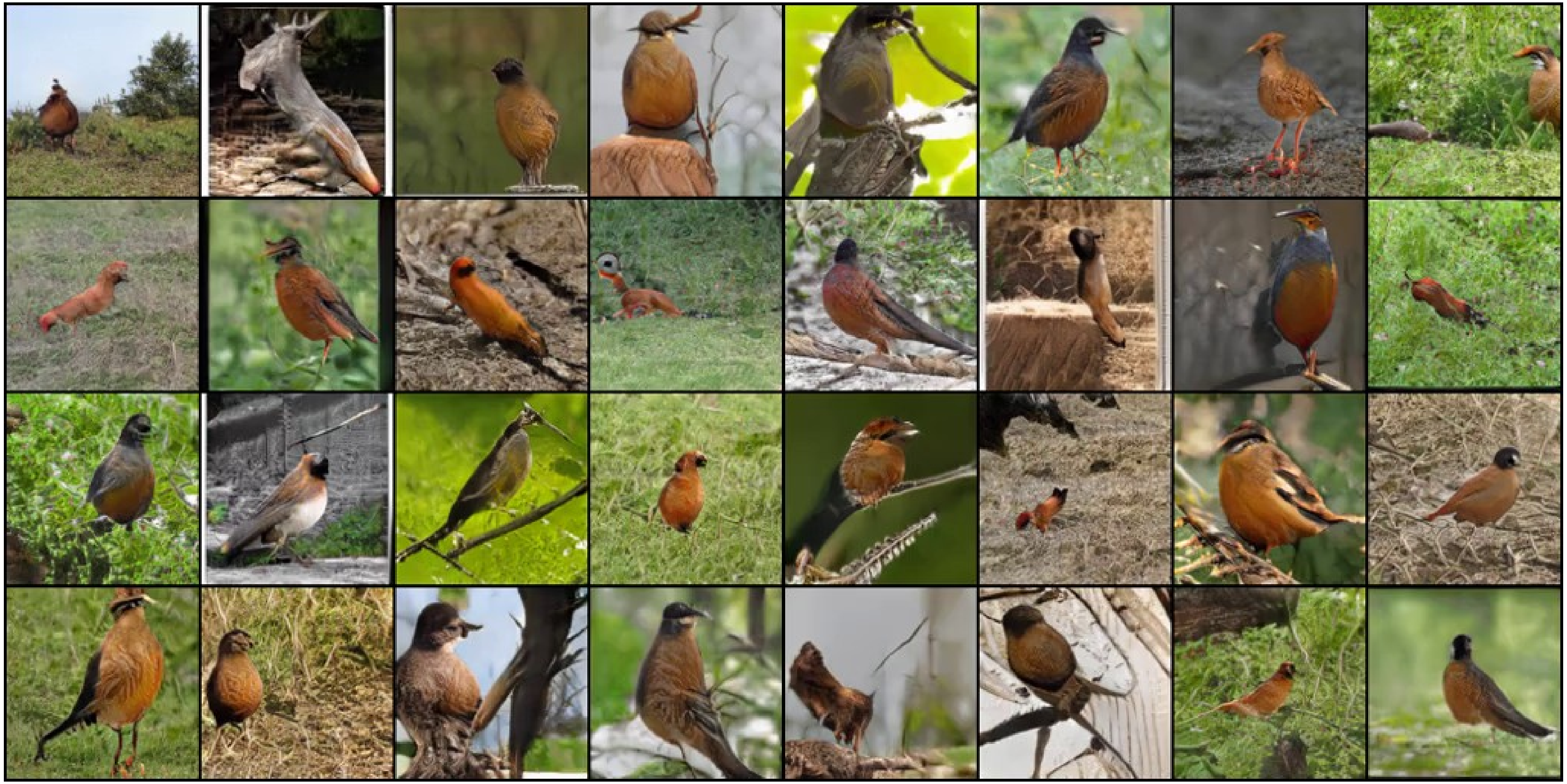
Negative definite

(orthogonal to $\mathcal{M}_G \times \mathcal{M}_D$)

$\Rightarrow$ 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)

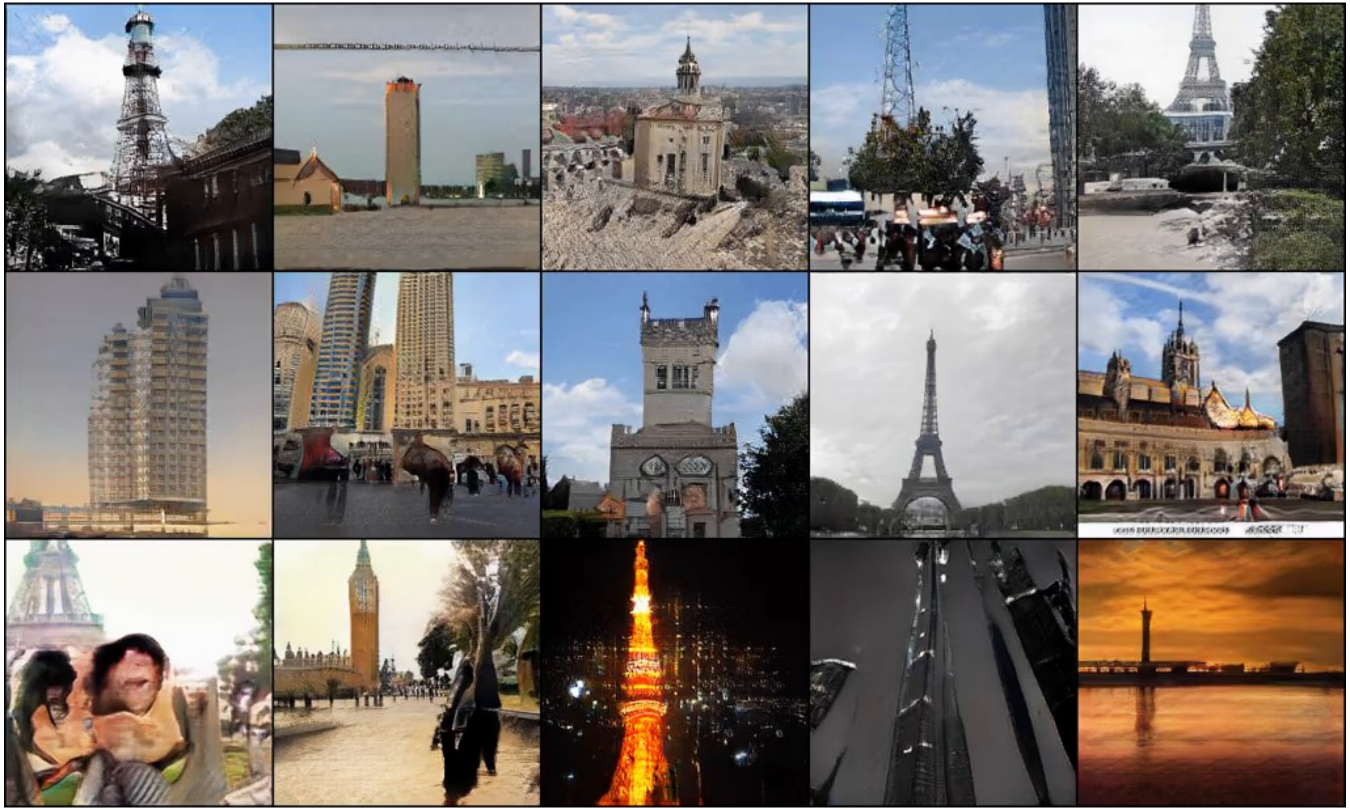
Experiments

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