GRADIENT DESCENT GAN OPTIMIZATION IS LOCALLY STABLE.

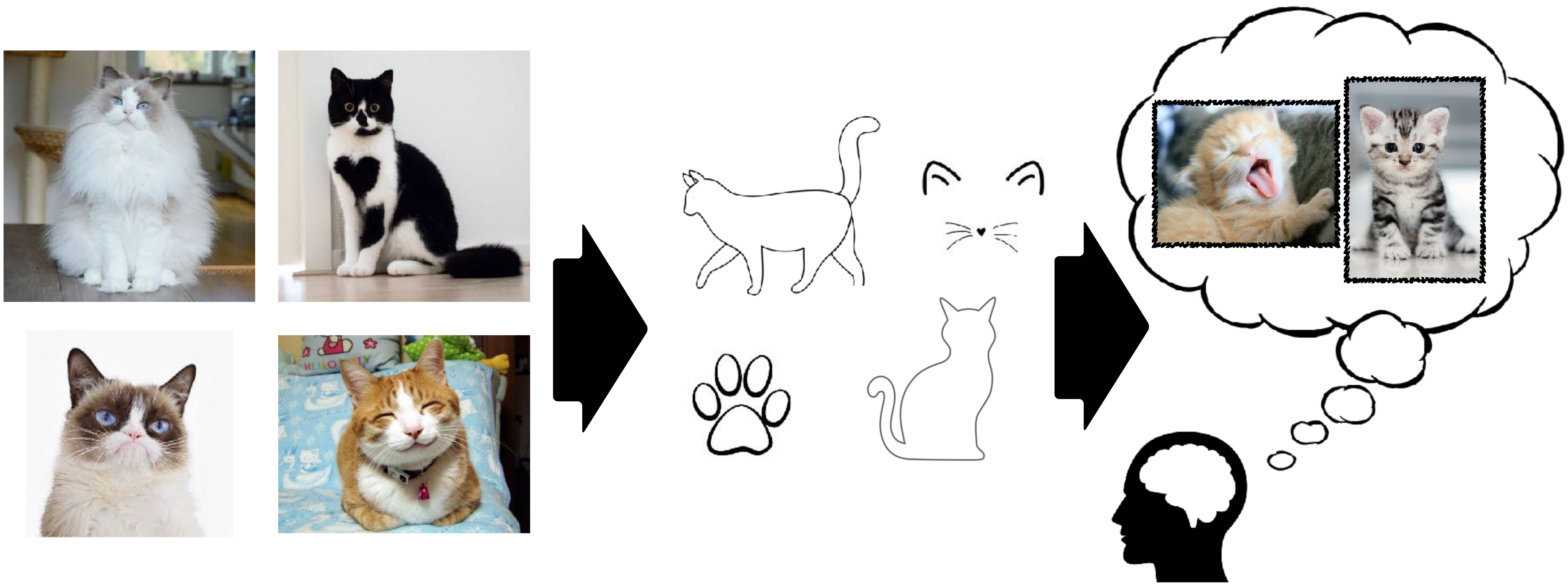
Vaishnavh Nagarajan | Zico Kolter

(Based on NIPS '17 Oral paper)

These slides are adapted from a 1hr talk I presented at CMU for a general CS audience.

Generative Adversarial Networks (GANs)

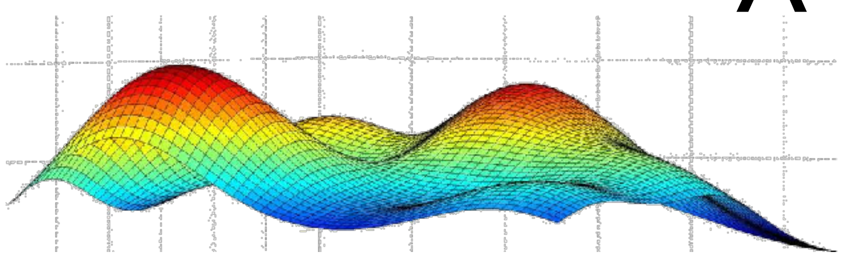
A goal of AI: “Understand” data



Build an agent that generates new data (which it does by learning an abstract representation of training data)

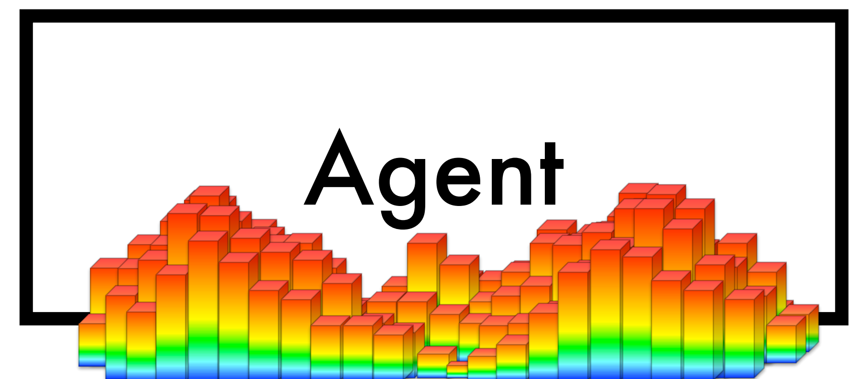
GENERATIVE ADVERSARIAL NETWORKS (GANS)

A generative model



TRUE DISTRIBUTION
(over e.g., cat images)

TRAINING DATA



(Implicitly)
LEARNED DISTRIBUTION

NEW DATA

PAST WORK

- **GANs were introduced by Goodfellow et al., '14**

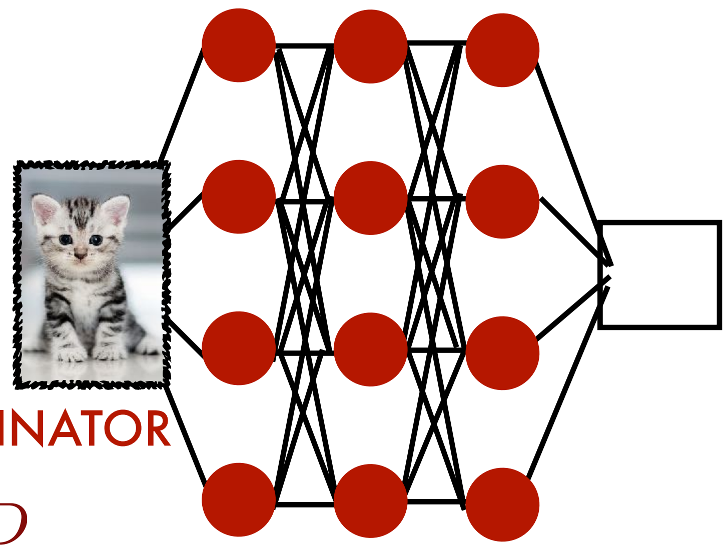
- **Many, many variants**: Improved GAN, WGAN, Improved WGAN, Unrolled GAN , InfoGAN MMD-GAN, McGAN, f-GAN, Fisher GAN, EBGAN, …

- **Wide-ranging applications:** image generation (DCGAN), text-to-image generation (StackGAN), super-resolution (SRGAN) …

PAST WORK

"One hour of imaginary celebrities" [Karras et al., '17]

GENERATIVE ADVERSARIAL NETWORKS (GANS)

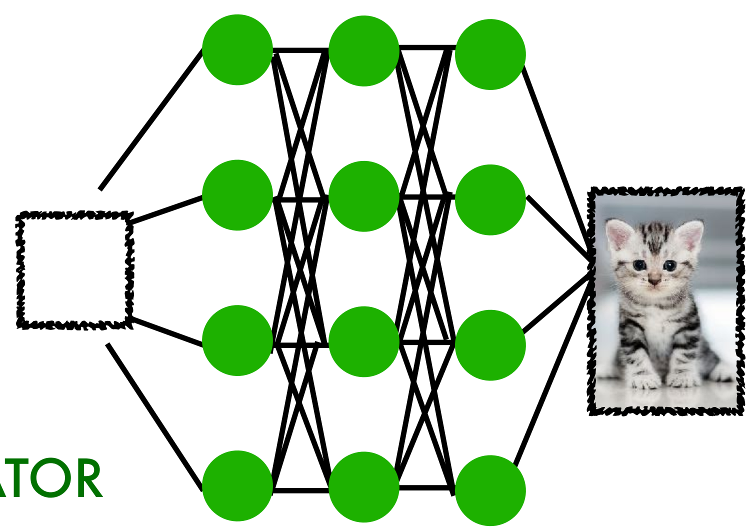


DISCRIMINATOR θ_D

tries its best to tell apart generated images from real images

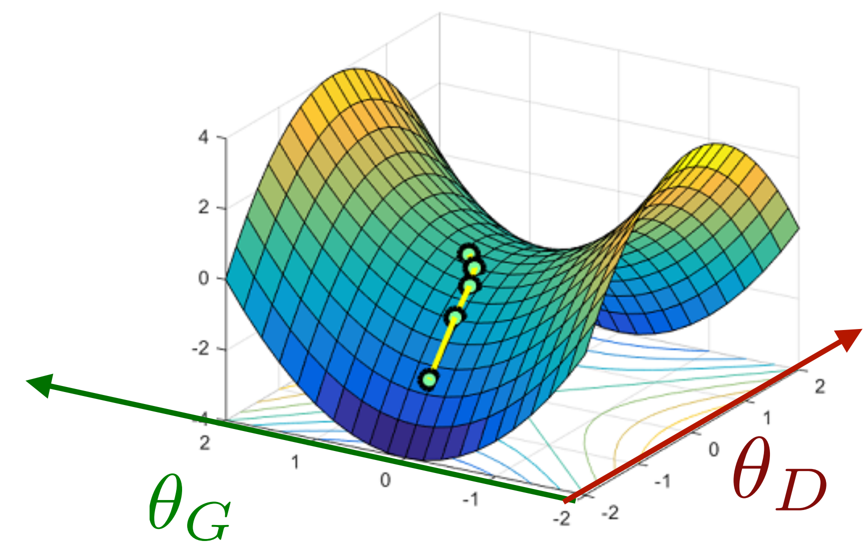
like a game

$$\min_{\theta_G} \max_{\theta_D} V(\theta_G, \theta_D)$$



GENERATOR θ_G

tries its best to generate images that discriminator finds real



GAN OPTIMIZATION: Parameters of two models are iteratively updated (in a standard way) to find "**equilibrium**" of a "min-max objective".

We study dynamics of standard GAN optimization:

Is the equilibrium “locally stable”?
When it is not, how do we make it stable?

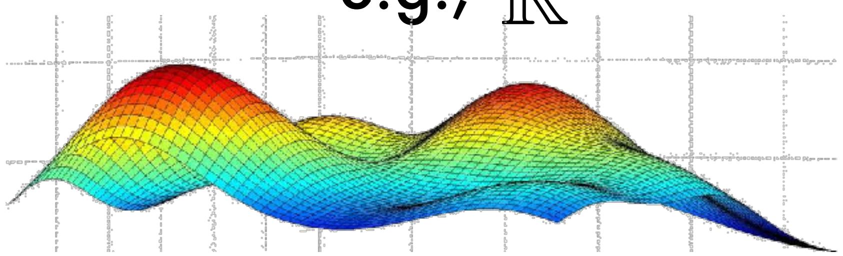
OUTLINE

- **GAN Formulation**
- Toolbox: *Non-linear systems*
- Challenge: *Why is proving stability hard?*
- Main result
- Stabilizing WGANs

GAN FORMULATION

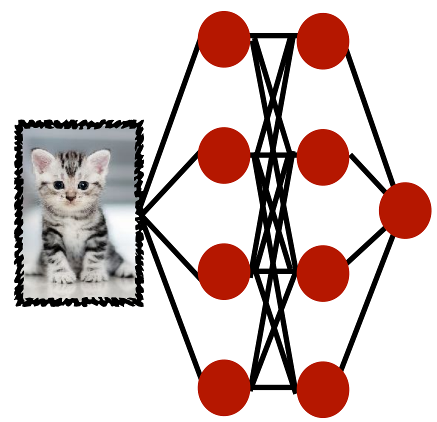
Unknown
TRUE P.D.F $p_{\text{data}}(\cdot)$
over INPUT DOMAIN $\mathcal{X}$

e.g., $\mathbb{R}^{32\times 32}$



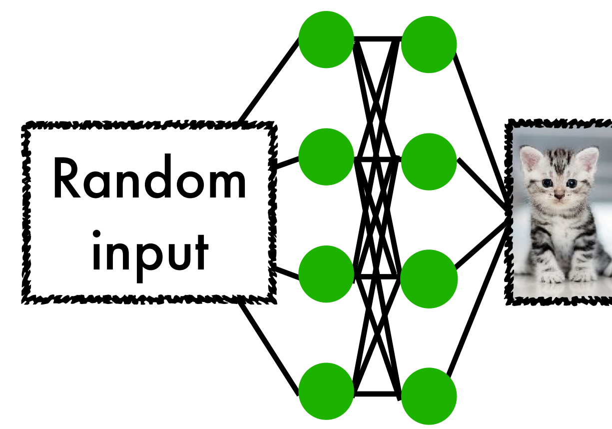
DISCRIMINATOR θ_D

$D\colon \mathcal{X} \to \mathbb{R}$

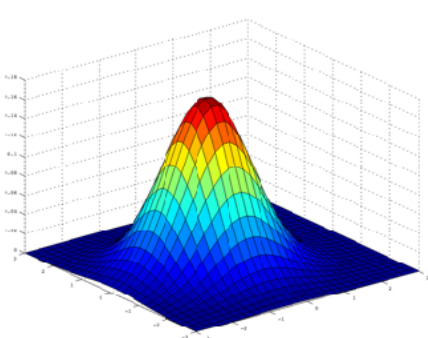


GENERATOR θ_G

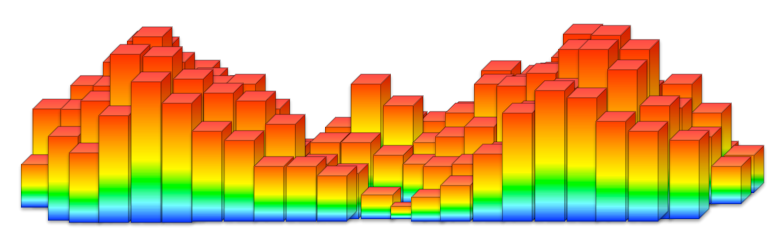
$G : \mathcal{Z} \to \mathcal{X}$



Known distribution over latent space $\mathcal{Z}$ with P.D.F $p_{\text{latent}}(\cdot)$



Generated distribution of $G(z)$ over $\mathcal{X}$ with P.D.F $p_{\theta_G}(\cdot)$



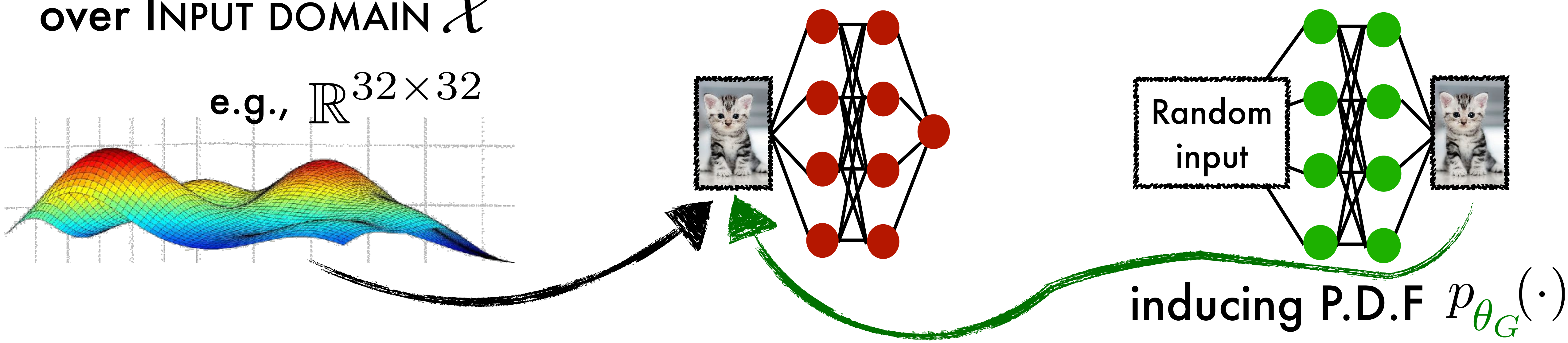
GAN Formulation

Unknown
True P.D.F $p_{\text{data}}(\cdot)$
over Input domain $\mathcal{X}$

e.g., $\mathbb{R}^{32\times 32}$

Discriminator θ_D
$D\colon \mathcal{X} \to \mathbb{R}$

Generator θ_G
$G : \mathcal{Z} \to \mathcal{X}$



Discriminator's objective: Tell real and generated data apart

D thinks x is:

$D(x) > 0$	real
$D(x) < 0$	generated
$D(x) = 0$	equally both

GAN Formulation

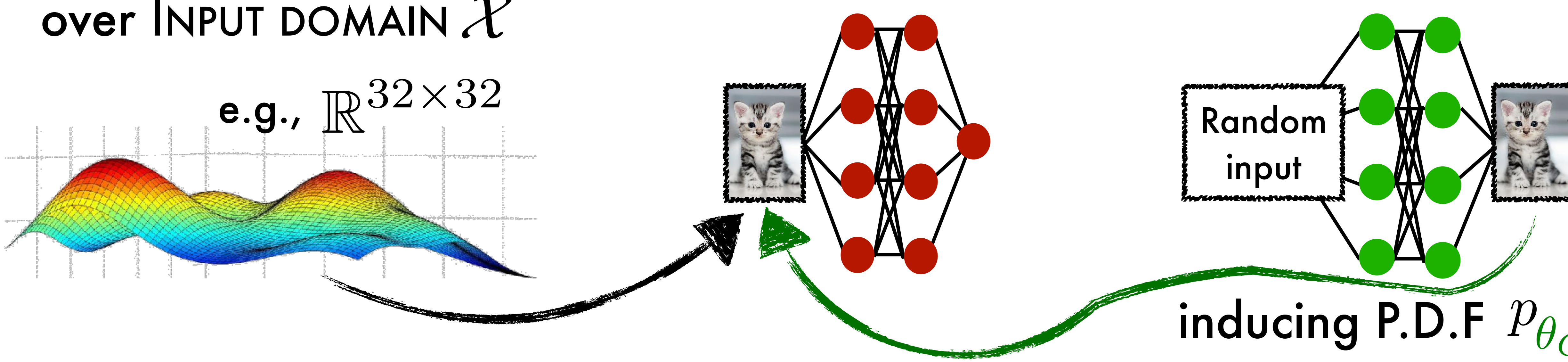
Unknown
True P.D.F $p_{\text{data}}(\cdot)$
over Input domain $\mathcal{X}$

Discriminator θ_D

$D\colon \mathcal{X} \to \mathbb{R}$

Generator θ_G

$G : \mathcal{Z} \to \mathcal{X}$



Discriminator's objective: Tell real and generated data apart

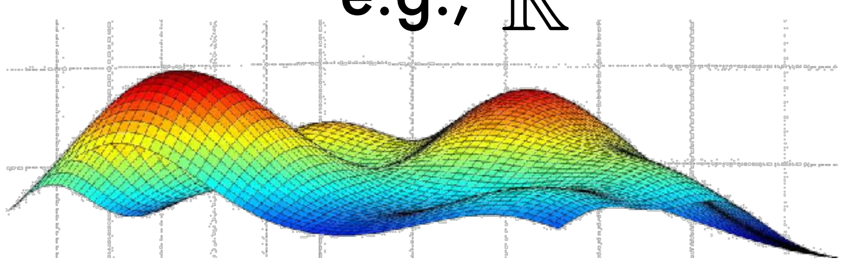
$$\max_{\theta_D} V(\theta_G, \theta_D)$$

$$= \mathbb{E}_{x\sim p_{\text{data}}}\big[\underbrace{f(D(x))}_{\text{How real x is, according to the discriminator}}\big] + \mathbb{E}_{z\sim p_{\text{latent}}}\big[\underbrace{f(-D(G(z)))}_{\text{How ``generated'' } G(z) \text{ looks according to the discriminator}}\big]$$

GAN Formulation

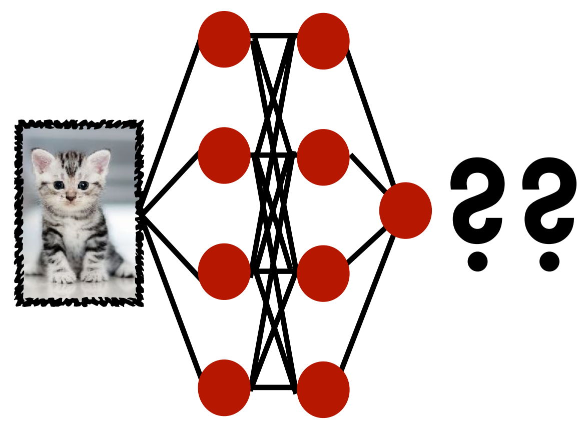
Unknown
True P.D.F $p_{\text{data}}(\cdot)$
over Input domain $\mathcal{X}$

e.g., $\mathbb{R}^{32\times 32}$



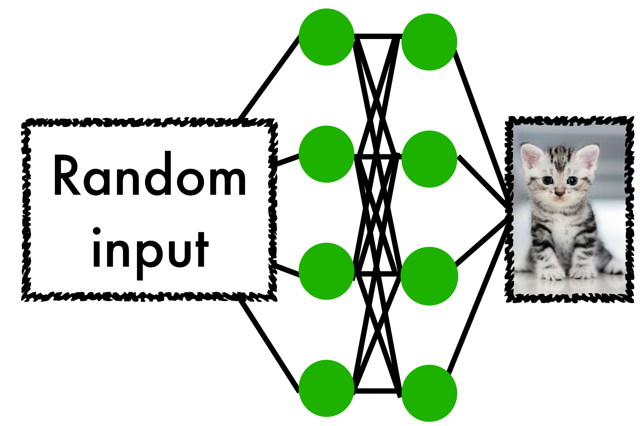
Discriminator θ_D

$D\colon \mathcal{X} \to \mathbb{R}$



Generator θ_G

$G : \mathcal{Z} \to \mathcal{X}$



inducing P.D.F $p_{\theta_G}(\cdot)$

Generator’s objective: Generate data that even the best discriminator can’t tell apart from real data

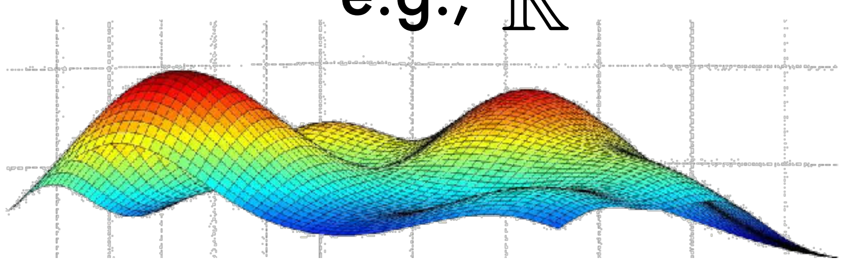
$$\min_{\theta_G}\left[\max_{\theta_D} V(\theta_G,\theta_D)\right]$$

$$= \mathbb{E}_{x\sim p_{\text{data}}}\left[\underbrace{f(D(x))}_{\substack{\text{How real x is,}\\ \text{according to the discriminator}}}\right] + \mathbb{E}_{z\sim p_{\text{latent}}}\left[\underbrace{f(-D(G(z)))}_{\substack{\text{How “generated” } G(z) \text{ looks}\\ \text{according to the discriminator}}}\right]$$

GAN FORMULATION

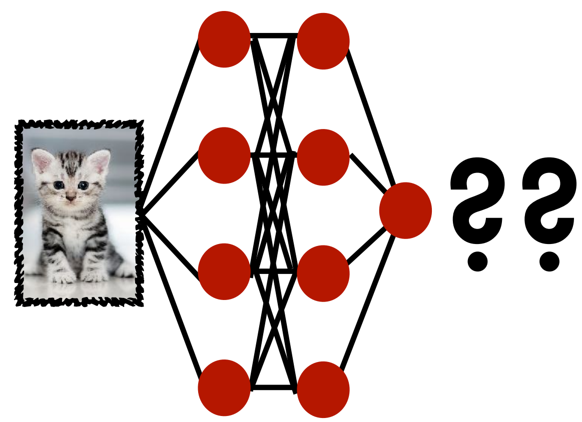
Unknown
TRUE P.D.F $p_{\text{data}}(\cdot)$
over INPUT DOMAIN $\mathcal{X}$

e.g., $\mathbb{R}^{32\times 32}$



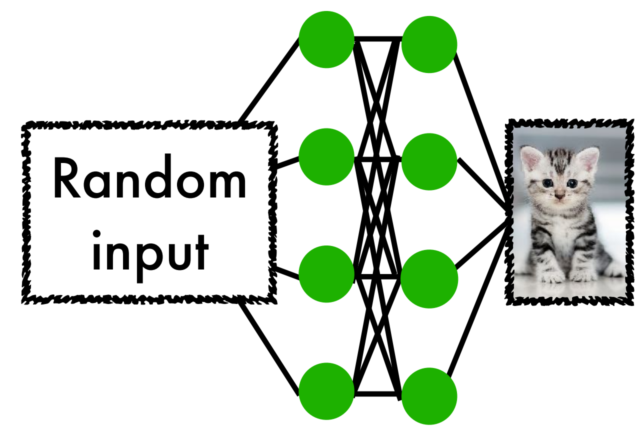
DISCRIMINATOR θ_D

$D\colon \mathcal{X} \to \mathbb{R}$



GENERATOR θ_G

$G : \mathcal{Z} \to \mathcal{X}$



inducing P.D.F $p_{\theta_G}(\cdot)$

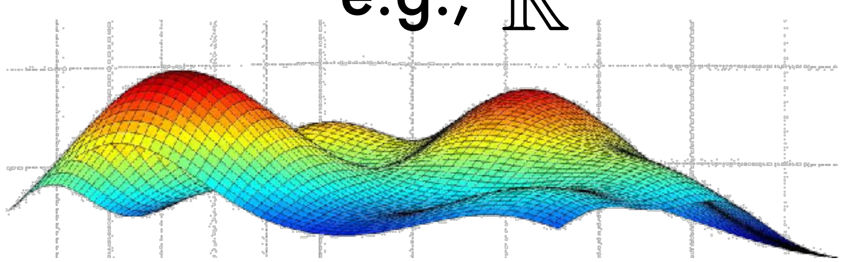
Traditional GAN	Wasserstein GAN (WGAN)
$f(t) = \log\left(\dfrac{1}{1+\exp(-t)}\right)$	$f(t) = t$

$$= \mathbb{E}_{x\sim p_{\text{data}}}\left[\underbrace{f(D(x))}_{\text{How real x is, according to the discriminator}}\right] + \mathbb{E}_{z\sim p_{\text{latent}}}\left[\underbrace{f(-D(G(z)))}_{\text{How ``generated'' } G(z) \text{ looks according to the discriminator}}\right]$$

GAN FORMULATION

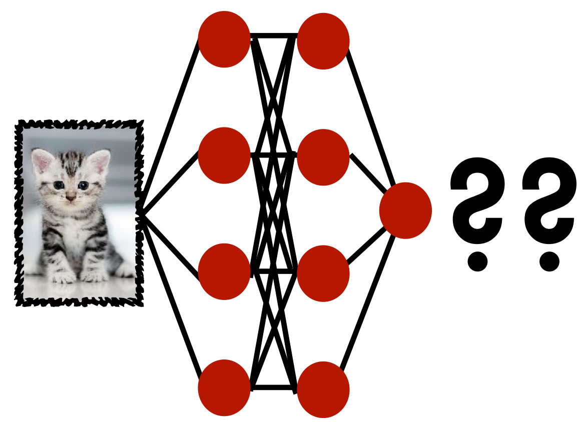
Unknown
TRUE P.D.F $p_{\text{data}}(\cdot)$
over INPUT DOMAIN $\mathcal{X}$

e.g., $\mathbb{R}^{32\times32}$



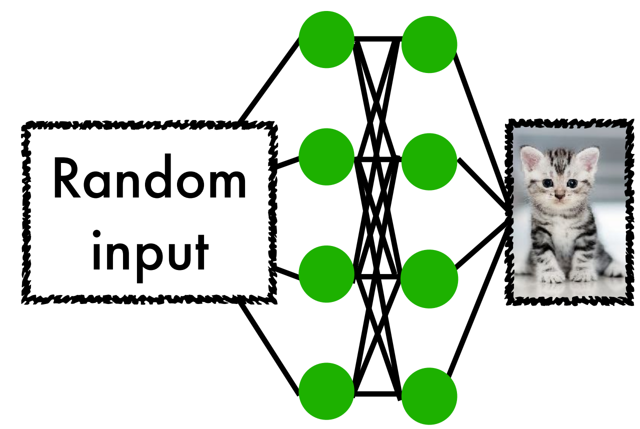
DISCRIMINATOR θ_D

$D\colon \mathcal{X} \to \mathbb{R}$



GENERATOR θ_G

$G\colon \mathcal{Z} \to \mathcal{X}$



inducing P.D.F $p_{\theta_G}(\cdot)$

SOLUTION: Generator matches true distribution and discriminator cannot tell apart data from either. How do we find this solution?

$$\min_{\theta_G}\left[\max_{\theta_D} V(\theta_G,\theta_D)\right]$$

$$= \mathbb{E}_{x\sim p_{\text{data}}}\big[\underbrace{f(D(x))}_{\text{How real x is, according to the discriminator}}\big] + \mathbb{E}_{z\sim p_{\text{latent}}}\big[\underbrace{f(-D(G(z)))}_{\text{How ``generated'' } G(z) \text{ looks according to the discriminator}}\big]$$

GAN Optimization

We consider: **infinitesimal, simultaneous** gradient ascent/descent updates

$$\min_{\theta_G} \left[\max_{\theta_D} V(\theta_G, \theta_D) \right]$$

Repeat simultaneously:

time derivative

$$\dot{\theta}_D = \nabla_{\theta_D} V(\theta_G, \theta_D)$$

$$\dot{\theta}_G = -\nabla_{\theta_G} V(\theta_G, \theta_D)$$

until **equilibrium**:

$$\dot{\theta}_D = 0$$

$$\dot{\theta}_G = 0$$

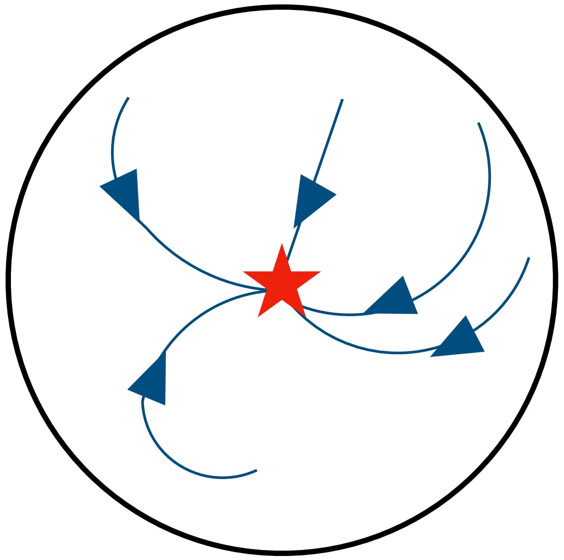
Outline

- GAN Formulation
- **Toolbox: *Non-linear systems***
- Challenge: *Why is proving stability hard?*
- Main result
- Stabilizing WGANs

Locally Exponentially Stable

Consider a dynamical system $\dot{\theta} = h(\theta)$ for which $\theta^\star$ is an equilibrium point i.e., $h(\theta^\star) = 0$

Informal Definition: The equilibrium point is **locally exponentially stable** if **any** initialization of the system sufficiently close to the equilibrium, converges to the equilibrium point "very quickly" (distance to equilibrium decays at the rate $\propto e^{-O(t)}$)



PROVING STABILITY

Consider a dynamical system $\dot{\theta} = h(\theta)$ for which $\theta^\star$ is an equilibrium point i.e., $h(\theta^\star) = 0$

LINEARIZATION THEOREM: The equilibrium of this (non-linear) system is locally exponentially stable if and only if its Jacobian at equilibrium HAS EIGENVALUES WITH **STRICTLY NEGATIVE REAL PARTS:**

$$J = \left.\frac{\partial h(\theta)}{\partial \theta}\right|_{\theta^\star} = \begin{bmatrix} \frac{\partial h_1(\theta)}{\partial \theta_1} & \frac{\partial h_1(\theta)}{\partial \theta_2} & \cdots \\ \frac{\partial h_2(\theta)}{\partial \theta_1} & \frac{\partial h_2(\theta)}{\partial \theta_2} & \cdots \\ \frac{\partial h_3(\theta)}{\partial \theta_1} & \vdots & \vdots\vdots\vdots \end{bmatrix}_{\theta=\theta^\star}$$

(asymmetric, real square matrix with possibly complex eigenvalues)

$$Jv = \lambda v \implies Re(\lambda) < 0$$

Proving Stability

Consider a dynamic
equilibrium point i.e.

Linearization Theor

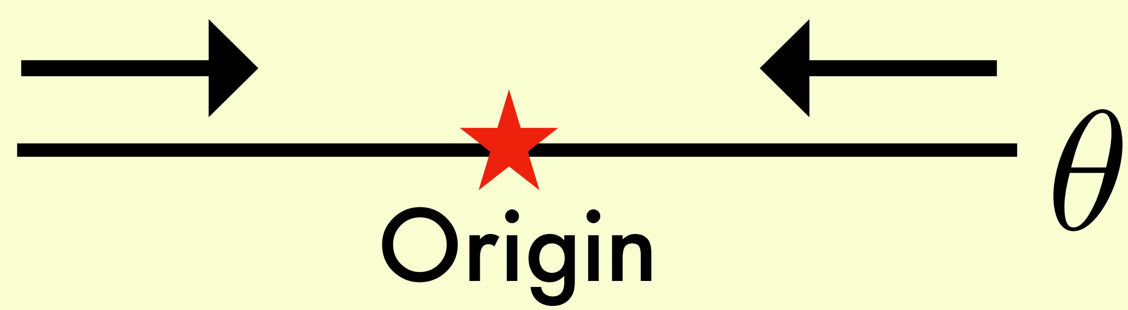
locally exponentially

has eigenvalues wit

$$J = \left.\frac{\partial h(\theta)}{\partial \theta}\right|_{\theta^\star}$$

1D Sanity Check/Intuition:

$$\dot{\theta} = -\theta$$



$$J = -1 < 0$$

(asymmetric, real square matrix with possibly complex eigenvalues)

$$Jv = \lambda v \implies Re(\lambda) < 0$$

OUTLINE

- GAN Formulation
- Toolbox: Non-linear systems
- **Challenge: *Why is proving stability hard?***
- Main result
- Stabilizing GANs

RECALL: GAN OPTIMIZATION

We consider: **infinitesimal, simultaneous** gradient descent updates

$$\min_{\theta_G} \left[\max_{\theta_D} V(\theta_G, \theta_D) \right]$$

Repeat simultaneously:

time derivative

$$\dot{\theta}_D = \nabla_{\theta_D} V(\theta_G, \theta_D)$$

$$\dot{\theta}_G = -\nabla_{\theta_G} V(\theta_G, \theta_D)$$

until equilibrium:

$$\dot{\theta}_D = 0$$

$$\dot{\theta}_G = 0$$

Why is proving GAN stability hard?

GAN involves concave-minimization—concave-maximization, even for a linear discriminator and a generator.

$$D(x) = \theta_D x \qquad\qquad G(z) = \theta_G z$$

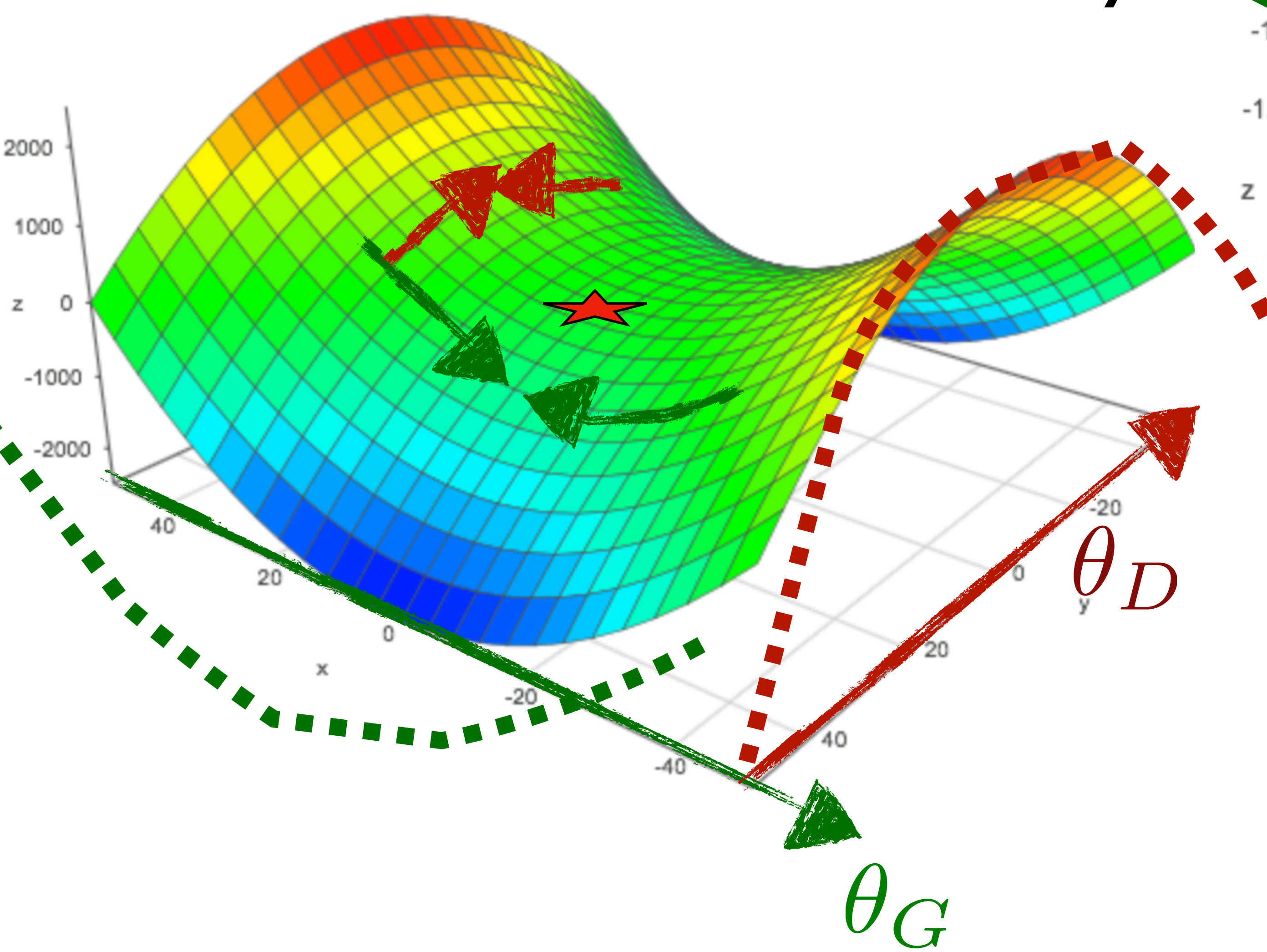
$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x \sim p_{\text{data}}} [f(\theta_D x)] + \mathbb{E}_{z \sim p_{\text{latent}}} [f(-\theta_D \theta_G z)]$$

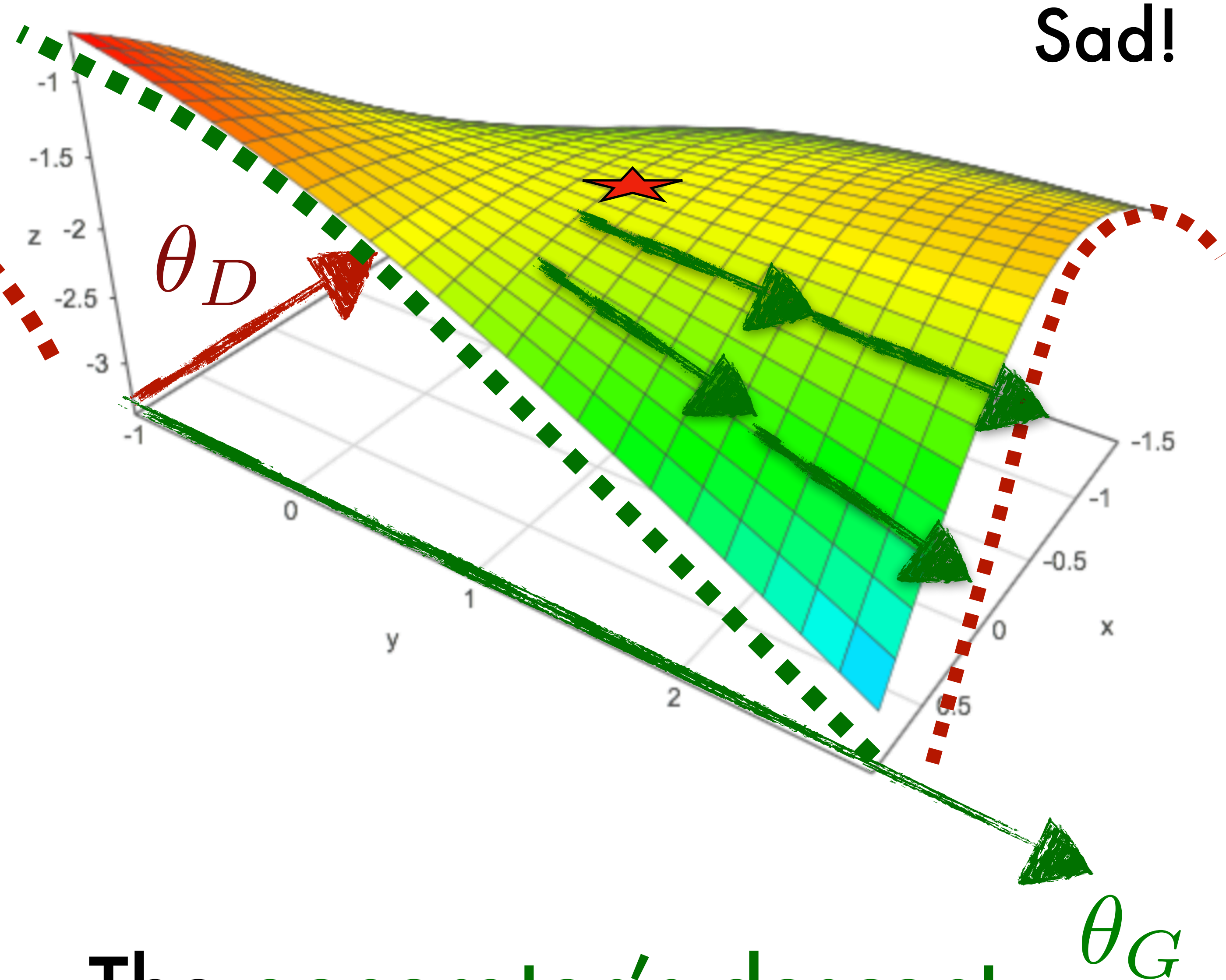
Given f is concave for GANs ,
objective is concave w.r.t θ_G.

Why is proving GAN stability hard?

If objective were convex-concave, would've been easy!

but for GANs, it is concave-concave. Sad!





The descent-ascent updates individually point in the direction of equilibrium.

The generator's descent updates take us away from equilibrium!

WHY IS PROVING GAN STABILITY HARD?

SOME CONCURRENT WORK:

Mescheder et al., '17: GANs may **not** be stable.

Heusel et al., '17, Li et al., '17: Stable provided discriminator updates "dominate" generator updates in some way. e.g.,

$$\dot{\theta}_D = \nabla_{\theta_D} V(\theta_G, \theta_D) \text{ x 100}$$

$$\dot{\theta}_G = -\nabla_{\theta_G} V(\theta_G, \theta_D)$$

But GANs in practice: updated with "equal weights"...

Despite a concave-concave objective,

simultaneous gradient descent GAN equilibrium

is

"locally exponentially stable"

under suitable conditions

on the representational powers of

the discriminator & generator.

OUTLINE

- GAN Formulation
- Toolbox: *Non-linear systems*
- Challenge: *Why is proving stability hard?*
- **Main result: *GANs are stable***
- Stabilizing WGANs

ASSUMPTION 1

Consider an equilibrium point $(\theta_D^\star, \theta_G^\star)$ such that generated distribution matches true distribution:

$$p_{\theta_G^\star}(\cdot) = p_{\text{data}}(\cdot)$$

and discriminator cannot tell real and generated data apart:

$$D_{\theta_D^\star}(x) = 0 \quad \text{for all } x$$

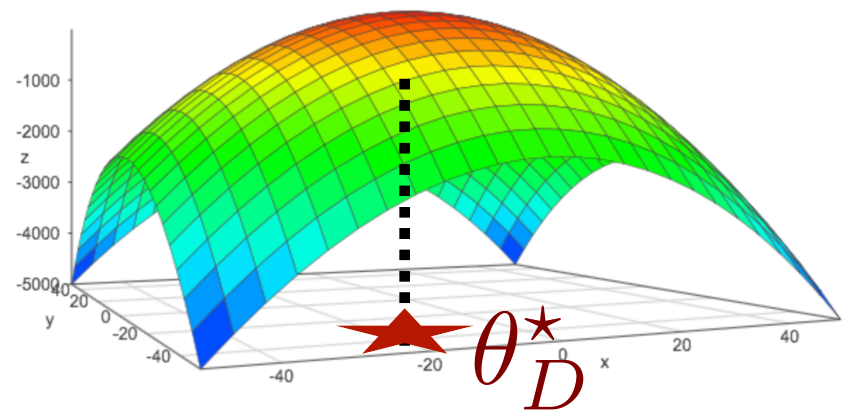
NOTE:

1. This *is* an equilibrium point (updates are 0 here).
2. Other kinds of equilibria may exist.
3. More relaxations in the paper, but at the cost of other restrictions

ASSUMPTION 2

Consider the objective at the equilibrium generator,
as a function of the discriminator.

$$V(\theta_D, \theta_G^\star)$$



AT EQUILIBRIUM DISCRIMINATOR, this is already a concave function.

We assume stronger curvature.
the corresponding Hessian $\nabla^2_{\theta_D} V(\theta_D, \theta_G^\star)$ evaluated at equilibrium discriminator is **NEGATIVE DEFINITE.**

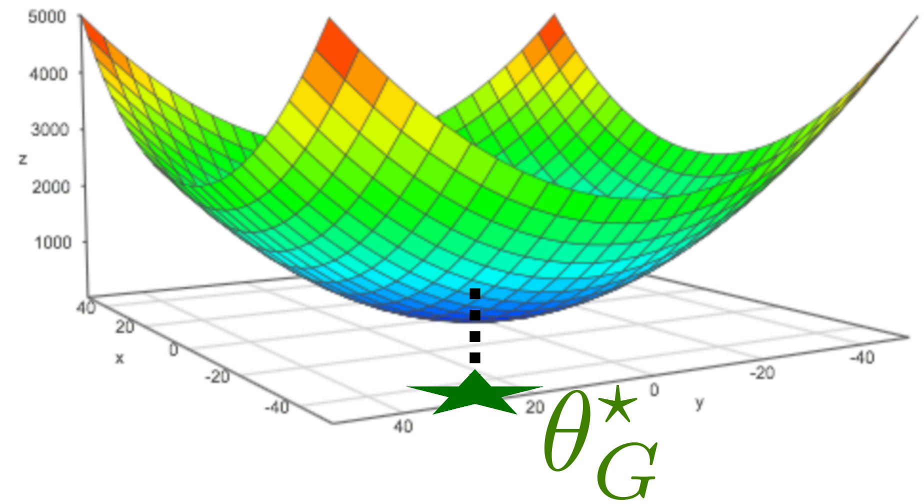
ASSUMPTION 3

Consider

"the magnitude of the objective's gradient w.r.t equilibrium discriminator",

as a function of the generator.

$$\|\nabla_{\theta_D} V(\theta_D, \theta_G)\|^2\Big|_{\theta_D = \theta_D^\star}$$



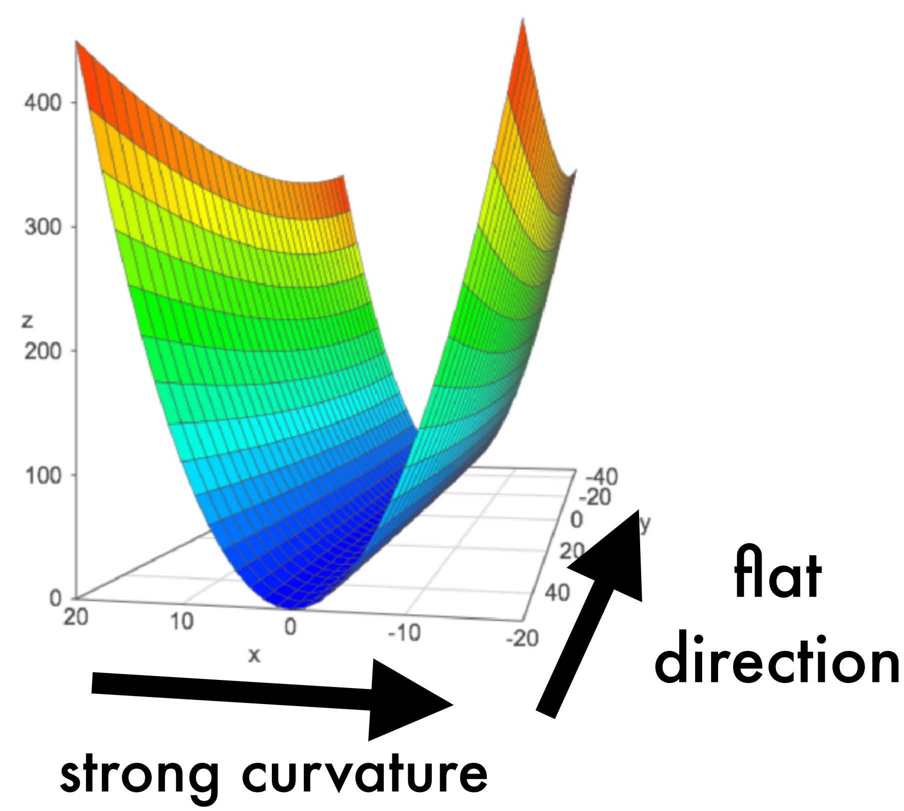
AT EQUILIBRIUM GENERATOR, this is already a convex function.

We assume stronger curvature.

the Hessian $\nabla^2_{\theta_G} \|\nabla_{\theta_D} V(\theta_D, \theta_G)\|^2\Big|_{\theta_D = \theta_D^\star}$

evaluated at equilibrium generator is **POSITIVE DEFINITE**.

These strong curvature assumptions
imply a locally unique equilibrium.
We also consider a specific relaxation allowing
a subspace of equilibria.



RECALL: GAN OPTIMIZATION

We consider: **infinitesimal, simultaneous** gradient descent updates

$$\min_{\theta_G} \left[\max_{\theta_D} V(\theta_G, \theta_D) \right]$$

Repeat simultaneously:

$$\dot{\theta}_D = \nabla_{\theta_D} V(\theta_G, \theta_D)$$

time derivative

$$\dot{\theta}_G = -\nabla_{\theta_G} V(\theta_G, \theta_D)$$

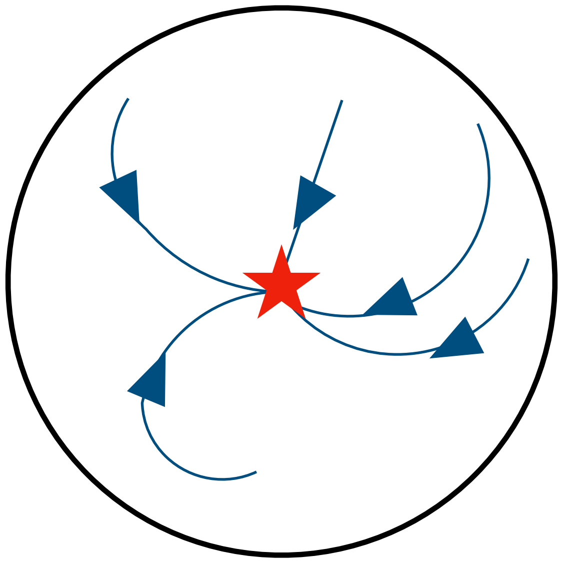
until equilibrium:

$$\dot{\theta}_D = 0$$

$$\dot{\theta}_G = 0$$

MAIN RESULT

THEOREM: Under assumptions 1-3, the equilibrium of the simultaneous gradient descent GAN system is locally exponentially stable.



MAIN RESULT

THEOREM: Under assumptions 1-3, the equilibrium of the simultaneous gradient descent GAN system is locally exponentially stable.

Specifically, the Jacobian at equilibrium has eigenvalues with strictly negative real parts.

$$J = \left.\frac{\partial h(\theta)}{\partial \theta}\right|_{\theta^\star} = \begin{bmatrix} \frac{\partial h_1(\theta)}{\partial \theta_1} & \frac{\partial h_1(\theta)}{\partial \theta_2} & \cdots \\ \frac{\partial h_2(\theta)}{\partial \theta_1} & \frac{\partial h_2(\theta)}{\partial \theta_2} & \cdots \\ \frac{\partial h_3(\theta)}{\partial \theta_1} & \vdots & \vdots\vdots\vdots \end{bmatrix}_{\theta=\theta^\star}$$

(asymmetric, real square matrix with possibly complex eigenvalues)

$$Jv = \lambda v \implies Re(\lambda) < 0$$

Proof Outline

Jacobian at equilibrium:

$$\begin{bmatrix} \partial\dot{\theta}_D / \partial\theta_D & \partial\dot{\theta}_D / \partial\theta_G \\ \partial\dot{\theta}_G / \partial\theta_D & \partial\dot{\theta}_G / \partial\theta_G \end{bmatrix}$$

PROOF OUTLINE

Jacobian at equilibrium:

$$\begin{bmatrix} \boxed{\nabla^2_{\theta_D} V(\theta_D, \theta_G)} & \partial \dot{\theta}_D / \partial \theta_G \\ \partial \dot{\theta}_G / \partial \theta_D & \partial \dot{\theta}_G / \partial \theta_G \end{bmatrix} = \begin{bmatrix} \text{negative definite} & \\ & \end{bmatrix}$$

A negative definite diagonal matrix makes it more likely that the whole matrix has eigenvalues with negative real parts.

Proof Outline

Jacobian at equilibrium:

$$\begin{bmatrix} \nabla^2_{\theta_D} V(\theta_D,\theta_G) & \boxed{\nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D,\theta_G)} \\ \boxed{-\left(\nabla_{\theta_G}\nabla_{\theta_D}V(\theta_D,\theta_G)\right)^T} & \partial\dot{\theta}_G / \partial\theta_G \end{bmatrix} = \begin{bmatrix} \text{negative definite} & \text{full column rank} \\ \text{negative transpose} & \end{bmatrix}$$

$$\left(\nabla_{\theta_G}\nabla_{\theta_D}V(\theta_D,\theta_G)\right)^T \nabla_{\theta_G}\nabla_{\theta_D}V(\theta_D,\theta_G) = \nabla^2_{\theta_G}\left\|\nabla_{\theta_D}V(\theta_D,\theta_G)\right\|^2\Big|_{\theta_D=\theta_D^\star}$$

Assumption 3: positive definite

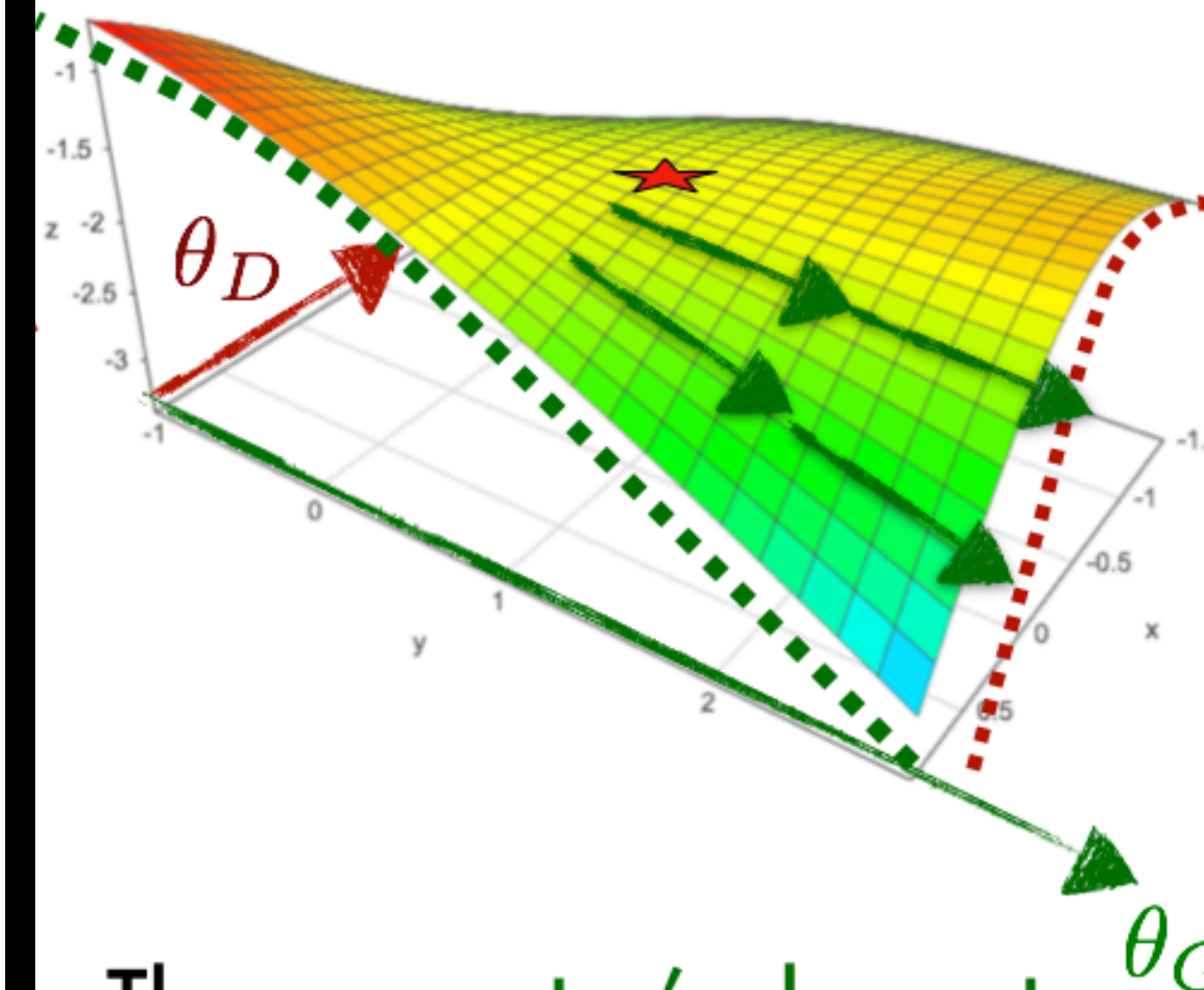
A negative definite diagonal matrix makes it more likely that the whole matrix has eigenvalues with negative real parts.

Proof Outline

Jacobian at equilibrium:

$$\begin{bmatrix} \nabla^2_{\theta_D} V(\theta_D, \theta_G) & \nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G) \\ -\left(\nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G)\right)^T & \boxed{-\nabla^2_{\theta_G} V(\theta_D, \theta_G)} \end{bmatrix}$$

but it is concave-concave!



The generator's descent updates take us away from equilibrium!

25

A negative definite diagonal m

whole matrix has eigenva

PROOF OUTLINE

Jacobian at equilibrium:

$$\begin{bmatrix} \nabla^2_{\theta_D} V(\theta_D, \theta_G) & \nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G) \\ -\left(\nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G)\right)^T & \boxed{-\nabla^2_{\theta_G} V(\theta_D, \theta_G)} \end{bmatrix} = \begin{bmatrix} \text{negative definite} & \text{full column rank} \\ \text{negative transpose} & 0 \end{bmatrix}$$

could be (– negative definite)
i.e., positive definite!

A negative definite diagonal matrix makes it more likely that the whole matrix has eigenvalues with negative real parts.

PROOF OUTLINE

Jacobian at equilibrium:

$$\begin{bmatrix} \nabla^2_{\theta_D} V(\theta_D, \theta_G) & \nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G) \\ -\left(\nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G)\right)^T & \boxed{-\nabla^2_{\theta_G} V(\theta_D, \theta_G)} \end{bmatrix} = \begin{bmatrix} \text{negative definite} & \text{full column rank} \\ \text{negative transpose} & 0 \end{bmatrix}$$

fix discriminator as all-zero equilibrium discriminator, objective is a constant:

$$\mathbb{E}_{p_{\text{data}}}[f(0)] + \mathbb{E}_{p_{\theta_G}}[f(0)] = 2f(0)$$

A negative definite diagonal matrix makes it more likely that the whole matrix has eigenvalues with negative real parts.

Proof Outline

Jacobian at equilibrium:

$$\begin{bmatrix} \nabla^2_{\theta_D} V(\theta_D, \theta_G) & \nabla_{\theta_G} \nabla_{\theta_D} V(\theta_D, \theta_G) \\ -\left(\nabla_{\theta_G} \nabla_{\theta_D} V(\theta_D, \theta_G)\right)^T & -\nabla^2_{\theta_G} V(\theta_D, \theta_G) \end{bmatrix} = \begin{bmatrix} \text{negative definite} & \text{full column rank} \\ \text{negative transpose} & 0 \end{bmatrix}$$

Main Lemma: Matrices J of this form have eigenvalues with **strictly** negative real parts:

$$Jv = \lambda v \implies Re(\lambda) < 0$$

Thus, the GAN equilibrium is locally exponentially stable.

OUTLINE

- GAN Formulation
- Toolbox: *Non-linear systems*
- Challenge: *Why is proving stability hard?*
- Main result: *GANs are stable*
- **Stabilizing WGANs**

WGAN

Jacobian at equilibrium:

$$\begin{bmatrix} \boxed{\nabla^2_{\theta_D} V(\theta_D, \theta_G)} & \nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G) \\ -\left(\nabla_{\theta_G}\nabla_{\theta_D} V(\theta_D, \theta_G)\right)^T & -\nabla^2_{\theta_G} V(\theta_D, \theta_G) \end{bmatrix} = \begin{bmatrix} 0 & \text{full column rank} \\ \text{negative transpose} & 0 \end{bmatrix}$$

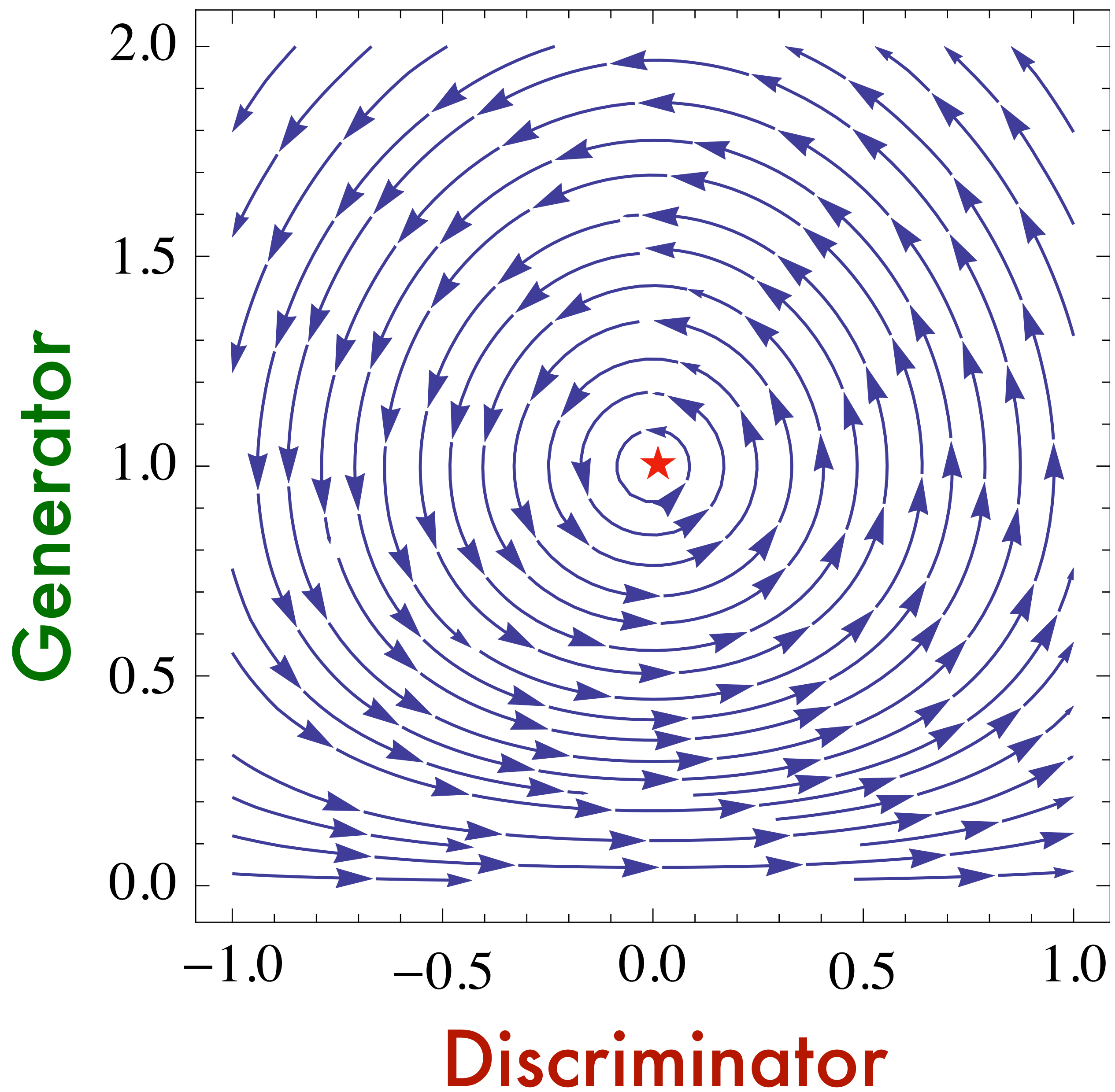
$$f(t) = t$$

fix generator to be at equilibrium:

$$\mathbb{E}_{p_{\text{data}}}[D(x)] + \mathbb{E}_{p_{\theta^\star_G}}[-D(x)] = 0$$

THEOREM: There exists an equilibrium for simultaneous gradient descent WGAN that does not converge locally.

WGAN



A system learning a uniform distribution.

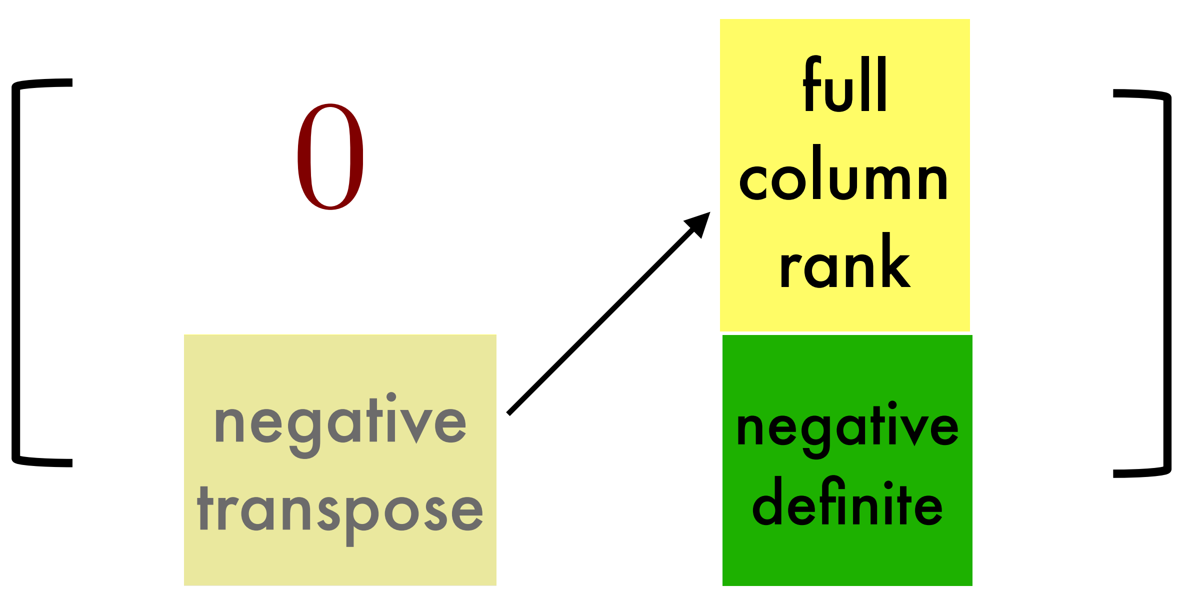
THEOREM: There exists an equilibrium for simultaneous gradient descent WGAN that does not converge locally.

Gradient-norm based Regularization

$$\dot{\theta_D} = \nabla_{\theta_D} V(\theta_D, \theta_G)$$

$$\dot{\theta_G} = -\nabla_{\theta_G} V(\theta_D, \theta_G) - \eta \nabla_{\theta_G} \|\nabla_{\theta_D} V(\theta_D, \theta_G)\|^2$$

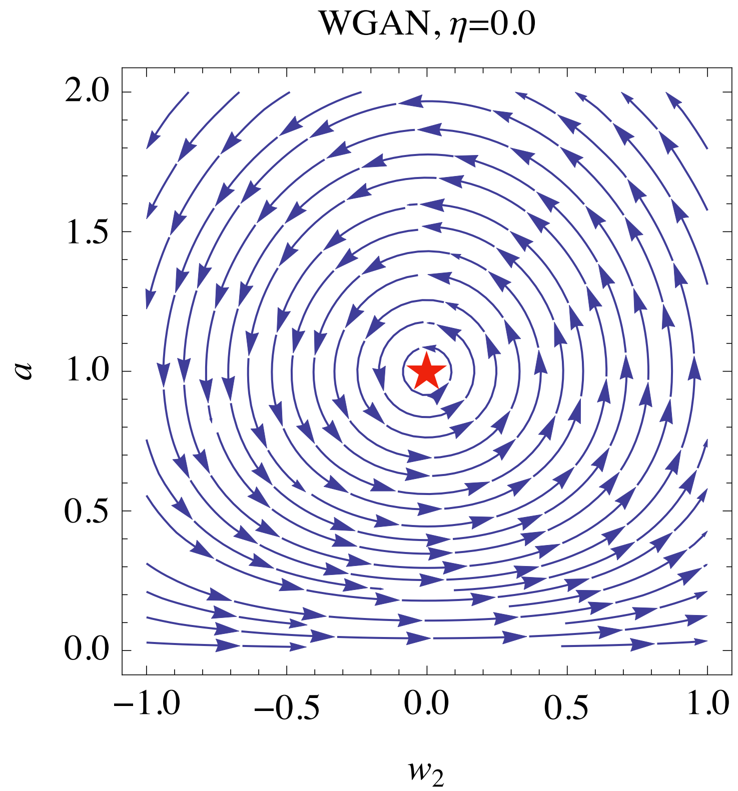
Generator minimizes (the objective + the norm of the discriminator's gradient).



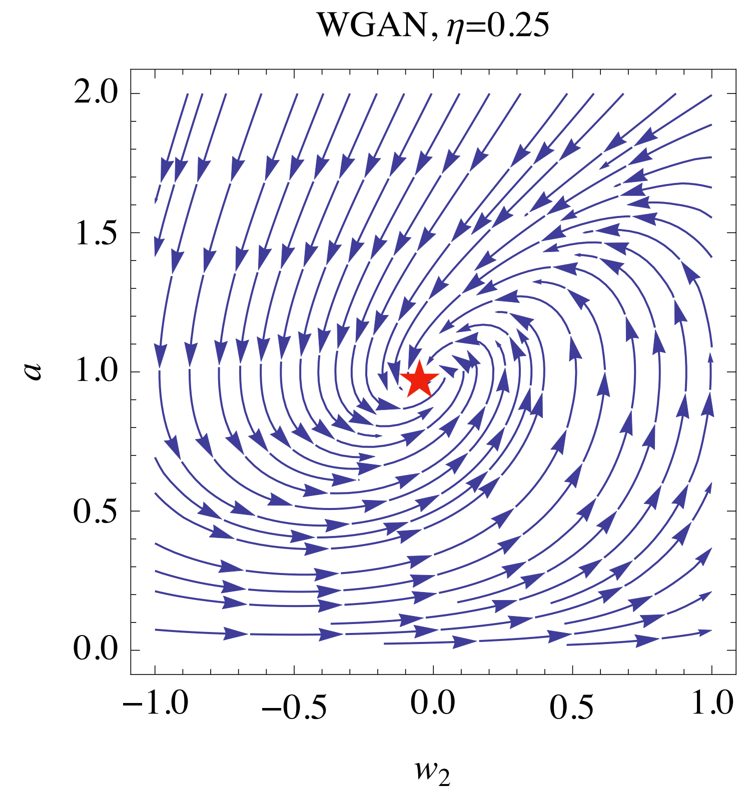
Theorem: Under similar assumptions, the equilibrium of the regularized simultaneous gradient descent (W)GAN system is locally exponentially stable when η not too large.

REGULARIZED WGAN
(learning a uniform distribution)

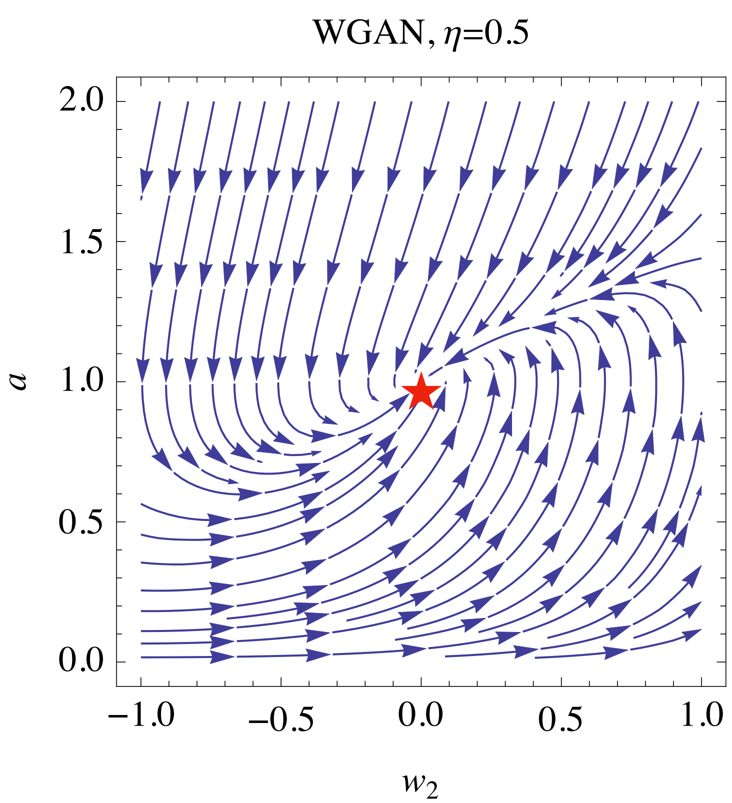
$\eta = 0$



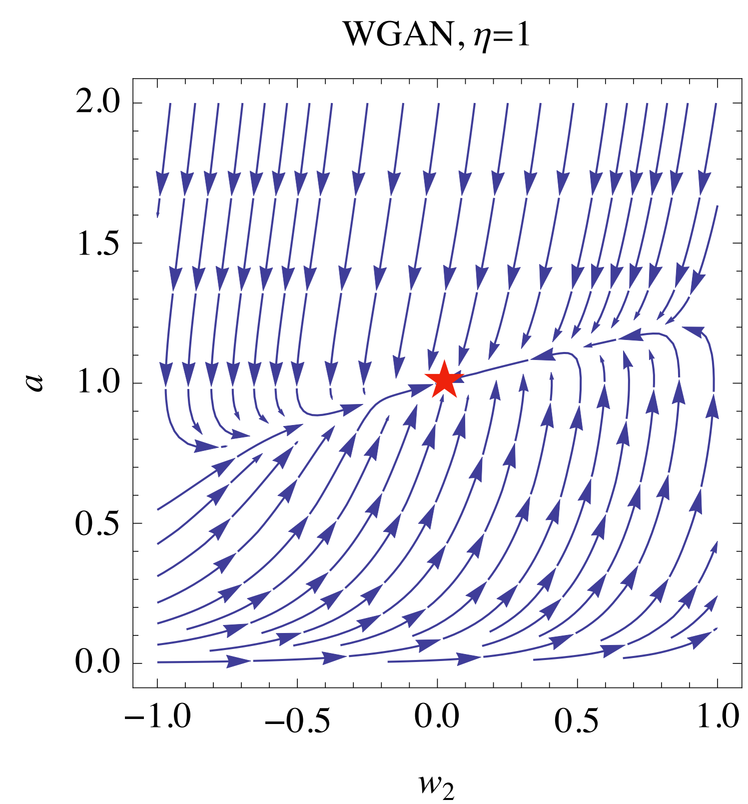
$\eta = 0.25$



$\eta = 0.5$



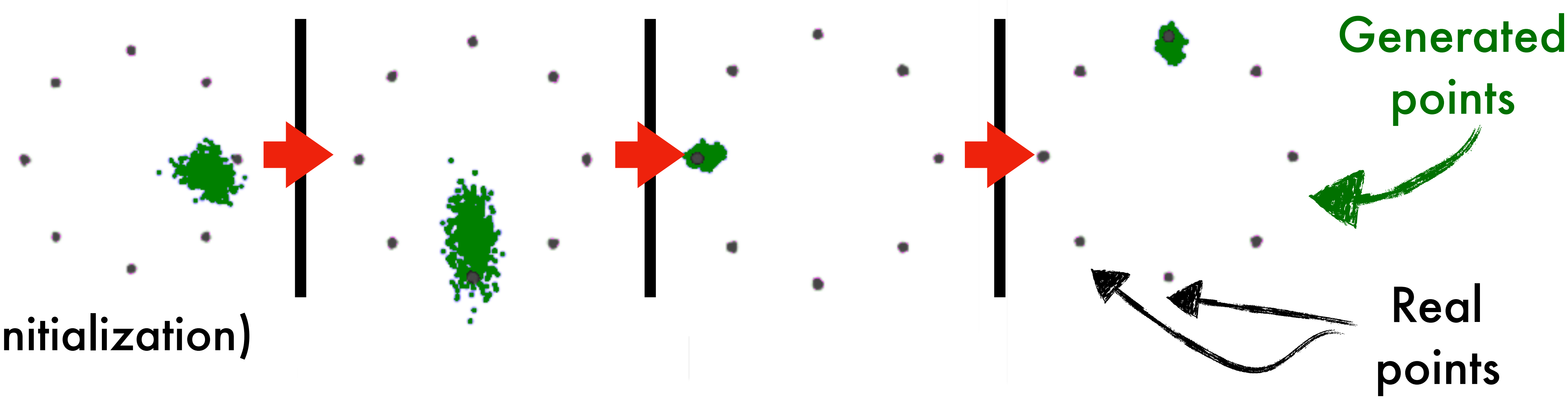
$\eta = 1.0$



FORESIGHTED GENERATOR

GAN training: a game where discriminator and generator try to outdo each other until neither can outdo the other.

Traditional GAN Generator is "greedy"

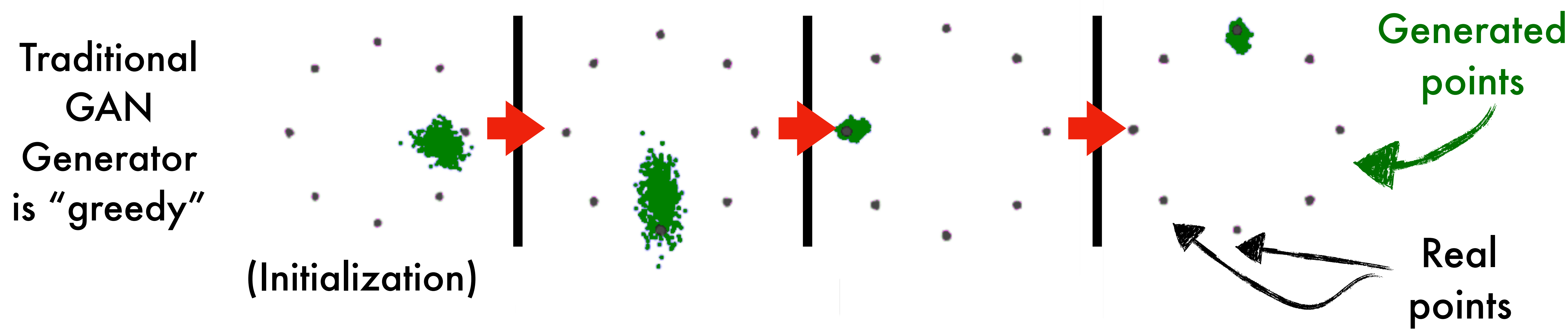


(Initialization)

Greedy generator strategy: Generate only one data point: the one to which discriminator has assigned highest value ("most real" according to discriminator).

FORESIGHTED GENERATOR

GAN training: a game where discriminator and generator try to outdo each other until neither can outdo the other.



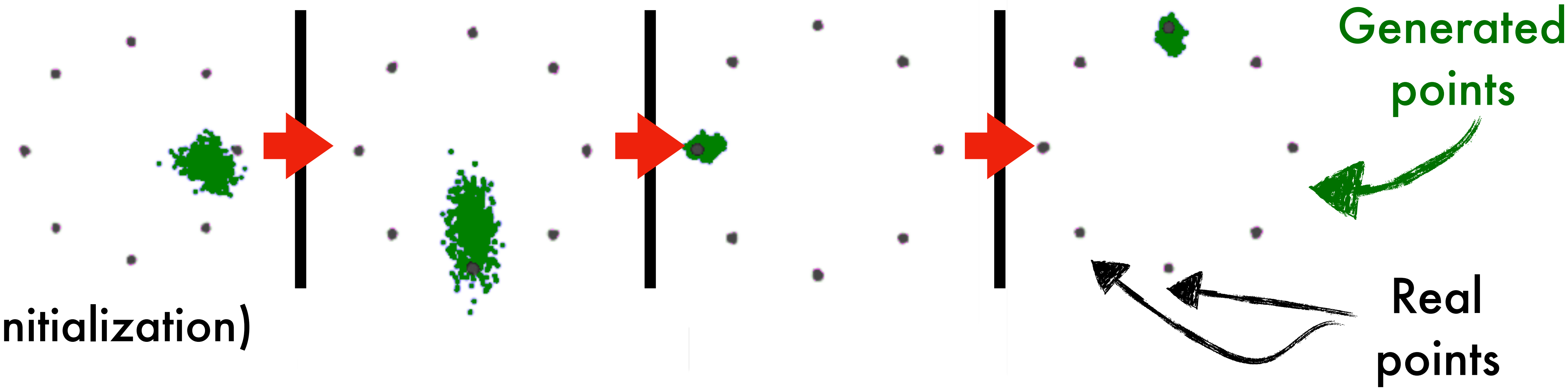
OBSERVATION: Generator keeps updating to state where objective $V(\theta_G, \theta_D)$ is small but discriminator update $\|\nabla_{\theta_D} V(\theta_D, \theta_G)\|^2$ is large.

SOLUTION: Generator explicitly seeks state where objective $V(\theta_G, \theta_D)$ is small AND discriminator update $\|\nabla_{\theta_D} V(\theta_D, \theta_G)\|^2$ is small.

FORESIGHTED GENERATOR

GAN training: a game where discriminator and generator try to outdo each other until neither can outdo the other.

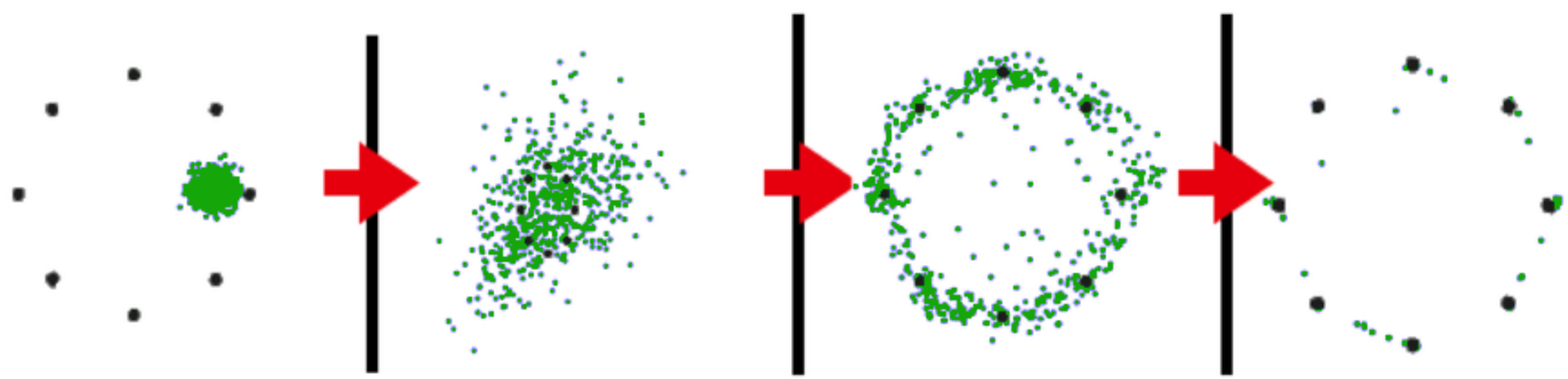
Traditional GAN Generator is "greedy"



(Initialization)

$$\dot{\theta}_G = -\nabla_{\theta_G} V(\theta_D, \theta_G) - \eta \nabla_{\theta_G} \|\nabla_{\theta_D} V(\theta_D, \theta_G)\|^2$$

Traditional GAN but with Regularized Generator



(Initialization)

CONCLUSION

- Theoretical analysis of local convergence/stability of simultaneous gradient descent GANs using non-linear systems.
- GAN objective is concave-concave, yet simultaneous gradient descent is locally stable — perhaps why GANs have worked well in practice.
- Our analysis yields a regularization term that provides more stability.

Open Questions

- Prove local stability for a more general case
- Global convergence?
- Many more theoretical questions in GANs: when do equilibria exist? Do they generalize?

Thank You.
Questions?