Global Landscape of GANs: Analysis and Improvement

—how **2 lines of code change** makes difference

Ruoyu Sun
Assistant Professor, ISE department; CSL and ECE (affiliated)
University of Illinois at Urbana-Champaign

BAAI 2020

Joint with Tiantian Fang, Alex Schwing of UIUC

GAN: Generative Models

- **What I cannot create, I do not understand. —R. Feynman**
- 

source: Goodfellow, ICLR'19 tutorial. https://www.iangoodfellow.com/slides/2019-05-07.pdf

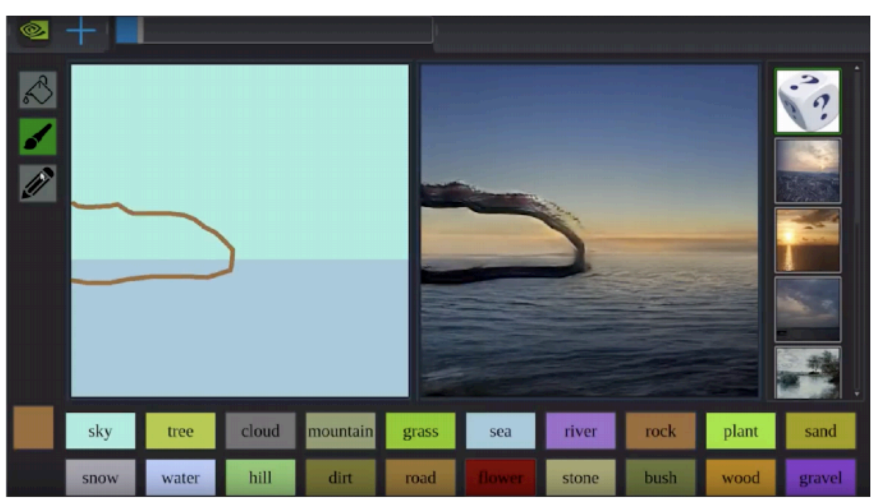
- **GAN (generative adversarial network)** has achieved great success: image generation, image-to-image translation, super-resolution, etc.

GAN Applications

Image Painting. Liu et al.'18

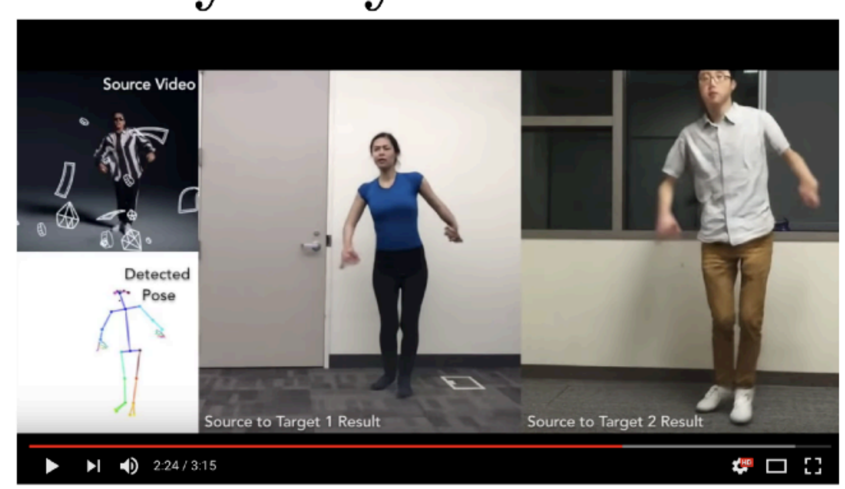
DiscoGAN. Kim et al.'17

GauGAN



(Park et al 2019)

Everybody Dance Now



(Chan et al 2018)

Motivation: Theory

- **Hard** to tune
- **Huge**: BigGAN requires **8 V100**, **15 days**

Nvidia Tesla v100 16GB
$7,720.00
& **FREE Shipping**
Arrives: **June 30 - July 6**
Deliver to Ruoyu - Champaign 61822
Only 2 left in stock - order soon.

Theory democratizes deep learning/AI techniques.
(besides improve understanding and design)

Example: 20 years ago, neural-net training is magic

Now: neural-net tricks are *partially* understood; easy to use

(**R. Sun,** Optimization for deep learning: an overview. JORSC 2020)

What's in This Talk?

1) For **GAN researchers:**

—More understanding of global dynamics of GANs

—Advocate R-GAN class

2) For **general audience**:

—Simple **intuition.** Toy **demo** of how GAN works.

3) For **mathematicians:**

—The power of equilibrium analysis (generic math trick)

Our Contributions

We analyze global landscape of the **empirical loss of GANs (with neural-nets).**

Theory:

1) JS-GAN has exponentially many bad basin, each of them is mode-collapse

2) Relativistic GANs (R-GAN) have no bad basin

Experiments:

0) **R-GAN** used by practitioners already; two lines of code change

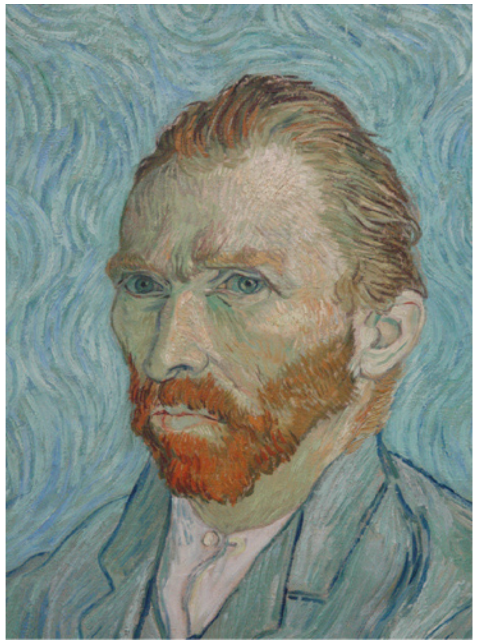
1) Verify "better landscape": narrower nets; more robust to initial point.

2) We **explain the training process** by our theory (for simple cases)

Part I Review of GAN and Literature

Generating Data

- Want to find a **new distribution** that is close the **true distribution**
- **Analogy**: you want to generate “paintings” (generated data), that match masterpieces (true data
- Who measures the progress? A critic, who tells the gap between your paintings and masterpieces



Documentary: China’s Van Goghs

Original JS-GAN

- The problem is $\min_{p_g} \phi(p_g, p_{\text{data}}),$ (1)

where $\phi(p_g, p_{\text{data}}) = \max_D E_{x \sim p_{\text{data}}, y \sim p_g} \log(D(x)) + \log(1 - D(y))\,.$

- Equivalent to min max L(p_{g} , D), for certain L.

- **Sanity check:** Loss $\phi(p_g, p_{\text{data}})$ is minimized iff $p_g = p_{\text{data}}$.

- **Math subject:** min-max optimization, game theory, probability

Theoretical Research

- **Statistical analysis**:

    —Relation to JS-distance [Goodfellow et al'14] Wasserstein GAN [Arjovsky & Bottou, 2017], f-GAN [Nowozin et al.'16]

    –Generalization bounds [Arora, Ge, Liang, Ma, and Zhang, 2017]

    –Mode collapse: PacGAN [Lin, Khetan, Fanti, and Oh'2018]

- **Optimization analysis:**

    —Convergence to local-min or stationary points:

Daskalakis et al., 2018; Daskalakis & Panageas, 2018; Azizian et al., 2019; Gidel et al., 2019; Mazumdar et al.; Yazıcı et al., 2019; Jin et al., 2019; Sanjabi et al., 2018

Bridge from simple to complex theoretical models

S1) pdf space

S2) generator function space

S3) parameter space

Source: Adapted from Goodfellow 17'tutorial, bridging theory and practice

Optimization Theory Steps

O3) converge to it? **O4) How quickly?**

O1) Is global-min desired? **O2) Is there bad local-min?**

Source: Adapted from Goodfellow 17'tutorial, bridging theory and practice

Optimization Analysis of GAN

	(S1) pdf space	(S2) G function space	(S3) parameter space
(O1) Sanity check	[Goodfellow et al. 14]	This work	This work
(O2) Local-min are good?	[Goodfellow et al. 14]	This work	This work
(O3,4) Convergence to local-min	Nagarajan & Kolter, 2017;		Mescheder et al. '18 (linear D), Sanjabi et al.'18, Jin et al.'19, Chu et al. '20, Daskalakis et al.'18,Yazıcı et al.'19, Gidel et al.'19

Part II Empirical Loss v.s Population Loss

Classical Analysis of GAN

- Problem: minimize $\min_{p_g} \max_{D} E_{x \sim p_{\text{data}}, y \sim p_g} \log(D(x)) + \log(1 - D(y))$.
- **Claim** [Goodfellow et al. 14] Function $\phi_{JS}(p_g, p_{\text{data}})$ is convex in p_g .

Probability space formulations are very popular in GANs, e.g.

—**Theory** papers: [Chu, Blanchet and Glynn'19], [Johnson and Zhang'19]

—**Empirical** papers: [Gong et al'19, TAC-GAN]

Pros: "**Convexify**" the problem by viewing the problem as in pdf-space.

Classical Analysis of GAN

Essence of the proof: any linear functional of the probability density is convex**.**

Claim: For any function f, $E_{y \sim p_g} f(y)$ **is convex in** p_g **.**

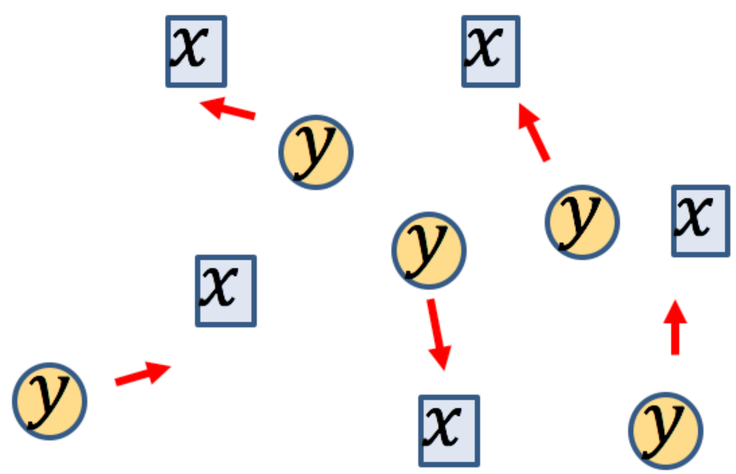
For instance, the problem $E_{y \sim p_g}[sin(y^2 + 1) + cos(y) + y^5]$ is convex in p_g

Observation: pdf space view does not utilize the structure of GANs.

Empirical Loss

"A good strategy to simplify a model for theoretical purposes is to work in **function space**."

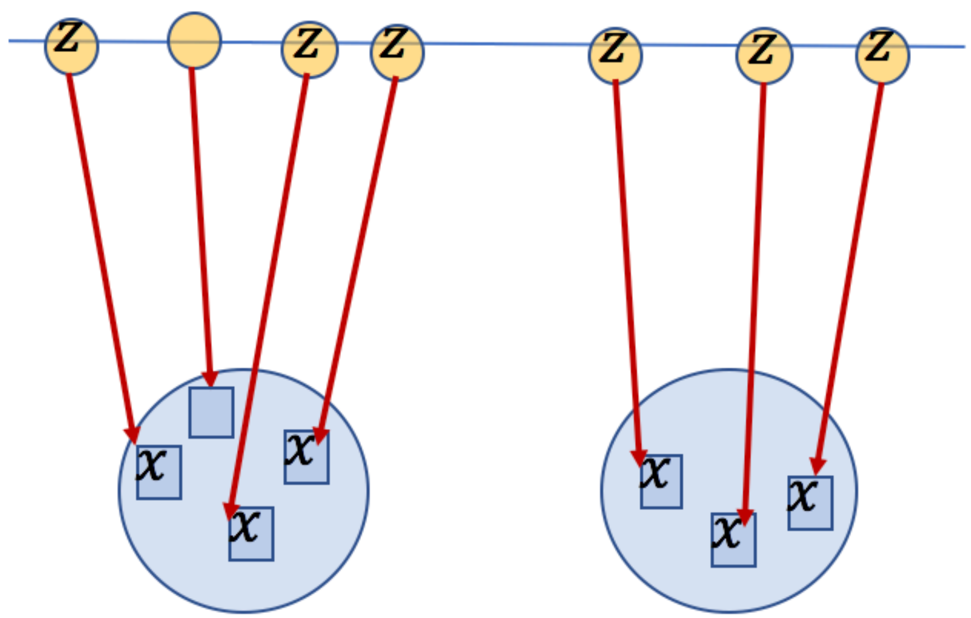
- **Empirical loss in function space:**
  - —**Data distribution**: fixed set of data points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$.
  - —**Generated distribution**: function space of samples $Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^{n \times d}$.



We will talk about neural-net **param space** results as well.

Generalization

Will this cause overfitting (memorizing)?Not necessarily memorizing



Generalization is possible; [Arora et al'18] gives concrete bounds on generalization.

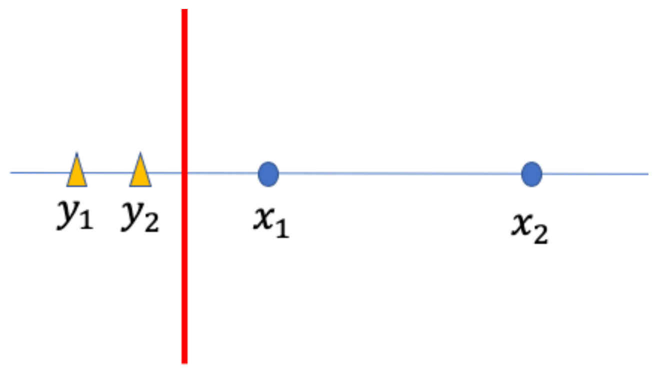
NOT the focus of this talk.

Part II Analysis of JS-GAN and RSGAN

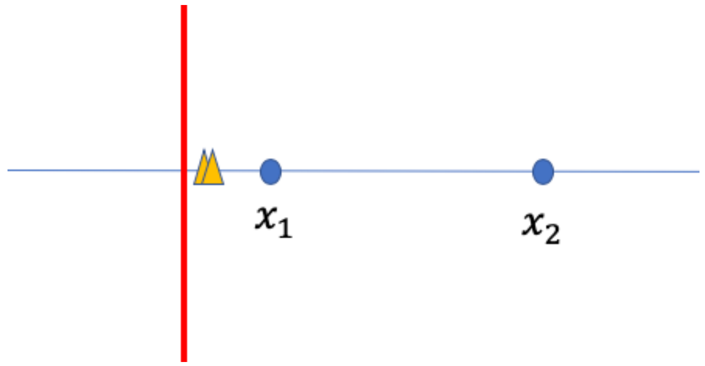
Intuition: Why GAN May Fail

Consider generating two points $Y = \{y_1, y_2\}$

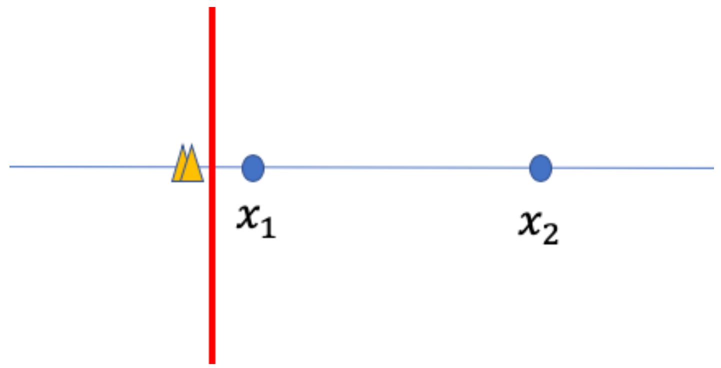
First, D successfully classifies Y and X



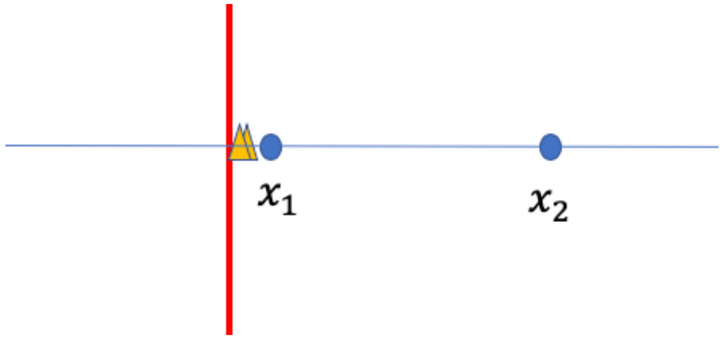
Second, Y moves right, to cross D.



Third, D moves right, to classify Y and X



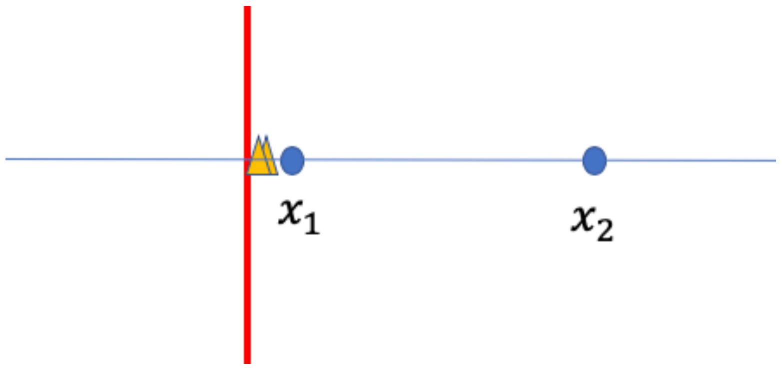
Fourth, Y moves right, to cross D



JS-GAN: Stuck at One Mode

In JS-GAN, the generated points are around one p**oint (mode).**

This is mode collapse**.**



Optimization-wise, seems to be a local-min?

Will formalize later.

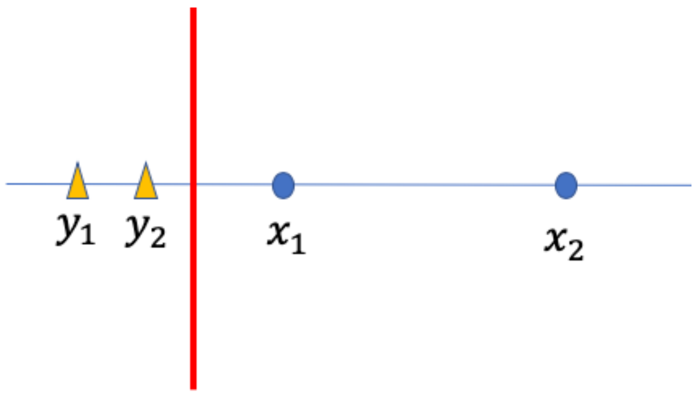
Recently, we learned that Li, Malik'2017 proposed similar intuition, when analyzing why mode collapse happens. But no formal proof of local-min.

Solution: “personalized criteria”

The issue is: a single criterion for every generated point.

Consider teaching a class, with many students.

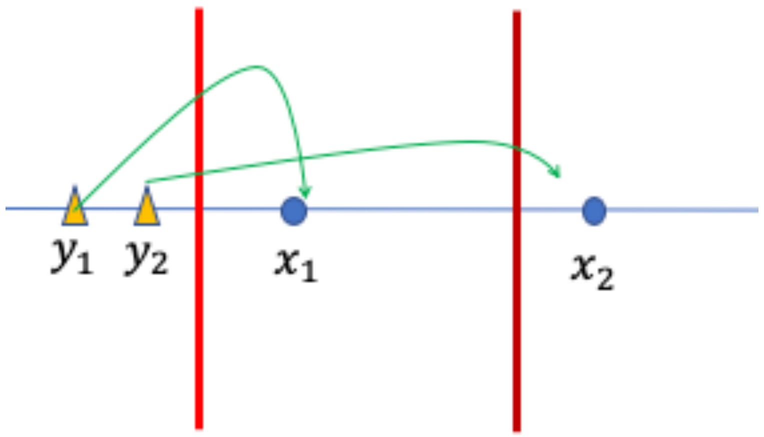
Universal criterion**:** If 60 points is enough, **t**hen most people will rest, after getting 60 points.



Personalized criterion:

—telling top 20%, criterion is 90 points, for grad school.

—telling other 80%, criterion is 60 points, for passing.



Key: **break locality.**

h-GAN and R-h-GAN

h-GAN: $\min_X \phi_h(Y, X)$, where $\phi_h(Y, X) = \max_f \frac{1}{2n} \sum_{i=1}^{n} h(f(x_i)) + \sum_{i=1}^{n} h(-f(y_i))$.

Example: in JS-GAN, $h(u) = \log(\frac{1}{1 + e^{-f(u)}})$

Relativistic GAN: $\min_Y \phi_{h,\mathrm{R}}(Y, X)$ where $\phi_{h,\mathrm{R}}(Y, X) = \max_f \frac{1}{2n} \sum_{i=1}^{n} h(f(x_i) - f(y_i))$.

Example: in relativistic standard GAN (**RS-GAN**), $h(u) = \log(\frac{1}{1 + e^{-f(u)}})$

Relativistic GAN

We proposed it in early version of the work (and called it coupled-GAN).

- Later, we found Jolicoeur-Martineau'2019 [JM’19] also proposed the same formulation, and call it “relativistic GAN”.

- It has different motivation (statistical): our motivation is to “break locality”

- [JM’19] showed convincing empirical results of relativisitic GANs.

Motivation from W-GAN: "Coupling" is Crucial

Wasserstein GAN:

$$\phi_{\mathrm{W}}(Y, X) = \max_{|f|_L \leq 1} \frac{1}{n} \sum_i [f(x_i) - f(y_i)]$$

W-GAN is different from JS-GAN in two aspects:

1) Change logistic regression loss to linear;

2) (Automatically) **Couple** X and Y. It is a special case of R-h-GAN.

We suspect that that "**coupling**" improves landscape, and is critical.

The first difference of changing "log(1+exp(…))" to linear does not help much.

Conjecture: if keeping log(1+exp(…)), but coupled, it should work better than WGAN.
—This is exactly RS-GAN.

Recent models BigGAN, SN-GAN, etc. use hinge loss. W-GAN is known to be slow.

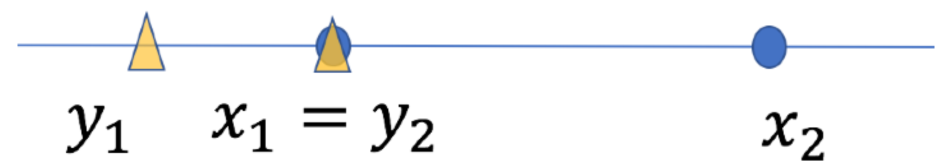
Part III Landscape Analysis: Formal Results

2-Point Example

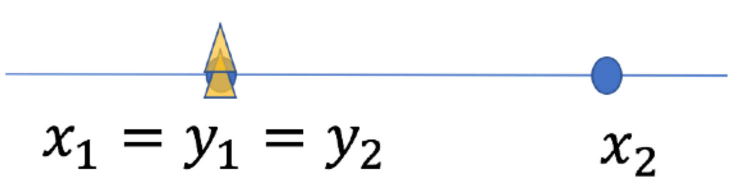
We compute the values of the objective for all Y.

Mainly four patterns.

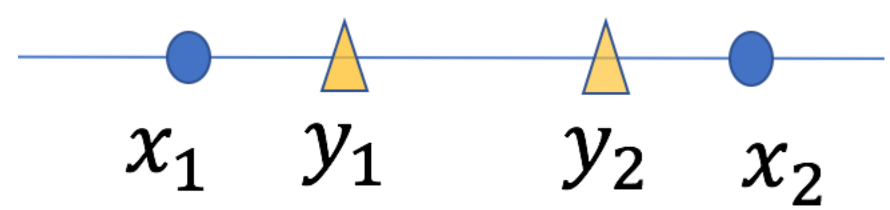
State 0: Perfect generation.



State 1b: mode dropping.



State 1a: mode collapse



State 2: Both points fake.

2-Point: Compute Values

Claim 1:

Suppose $n = 2$ *and* $x_1 \neq x_2 \in \mathbb{R}^d$. *Then*

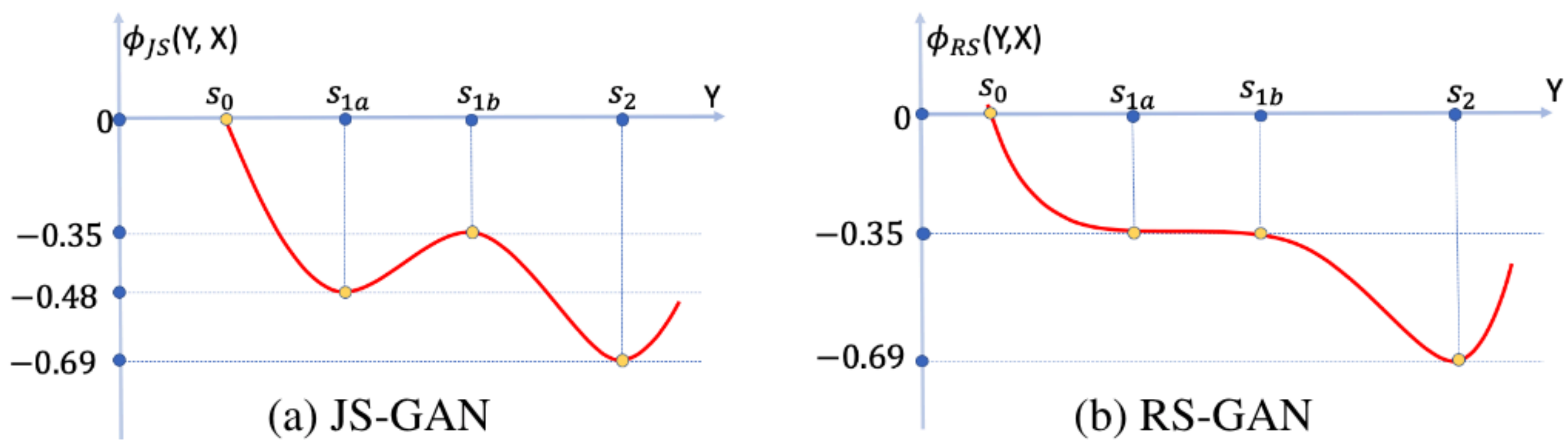
$$\phi_{\mathrm{JS}}(Y, X) = \begin{cases} -2\log 2 \approx -1.3862, & \textit{if } \{x_1, x_2\} = \{y_1, y_2\} \\ -\log 2 \approx -0.6931, & \textit{if } |\{x_1, x_2\} \cap \{y_1, y_2\}| = 1, \\ \log 2 - 1.5\log 3 \approx -0.9548, & \textit{if } y_1 = y_2 \in \{x_1, x_2\}, \\ 0 & \textit{if } |\{x_1, x_2\} \cap \{y_1, y_2\}| = \emptyset. \end{cases}$$

$$\phi_{\mathrm{RS}}(Y, X) = \begin{cases} -\log 2 \approx -0.6931, & \textit{if } \{x_1, x_2\} = \{y_1, y_2\} \\ -\frac{1}{2}\log 2 \approx -0.3466, & \textit{if } |\{i : x_i = y_i\}| = 1 \\ 0 & \textit{otherwise.} \end{cases}$$

Corollary 1: $(y_1, y_2) = (x_1, x_1)$ is a strict local-min for JS-GAN; but RS-GAN has no strict local-min.

2-point Example

Smoothed version of the loss landscape:



(a) JS-GAN (b) RS-GAN

Observation: mode-collapse s_{1a} causes a basin in JS-GAN, but NOT in RS-GAN.

Intuition: JS-GAN views mode collapse as **worse than** mode dropping (one fake data is good, another is noise), causing bad basin.

RS-GAN views mode-collapse, mode dropping **as equally bad**, thus mode collapse does not create a basin.

Disclaimer: the loss function are actually discontinuous, but we connect the points to make it smooth. In practical training, we inexactly optimize D, which smoothes the landscape.

Non-basin v.s. basin

Non-strict local-min
Weak attractor

Basin
Strong attractor

h-GAN has basin: general n

Assumption 1: $\sup_t h(t) = 0$; $h(0) < 0$; h is concave.

Recall: $\phi_h(Y, X) = \max_f \frac{1}{2n} \sum_{i=1}^{n} h(f(x_i)) + \sum_{i=1}^{n} h(-f(y_i))$.

Theorem 1 If all $y_i \in \{x_1, x_2, \ldots, x_n\}$ but some x_i is not in the generated data set, then Y is a sub-optimal strict local-min of $\phi_h(Y, X)$.

- **In words:** "mode-collapse" = "bad basin"
- $(n^n - n!)$ basins in h-GAN (e.g. JS-GAN) landscape.

R-GAN is nice: general n

$$\phi_{h,\mathrm{R}}(Y,X) = \max_f \frac{1}{2n} \sum_{i=1}^{n} h(f(x_i) - f(y_i)) \, .$$

Global-min-reachable (GMR): If from any point u, there is a continuous path from u to a global minimum of F such that F is non-increasing along the path, we say F satisfies GMR.

- **Theorem 2:** Y is a global-min of $g(Y) = \phi_{h,\mathrm{R}}(Y,X)$ iff $\{x_1, x_2, \ldots, x_n\} = \{y_1, y_2, \ldots, y_n\}$. In addition, **g is GMR.**

- This implies: R-GAN (including RS-GAN) does not have bad basins.

Results in Parameter Space

Assume the generator neural-net is $G_w(z)$, and the discriminator neural-net is $f_\theta(u)$.

Assumption 1 (informal): Both $G_w(z)$ and $f_\theta(u)$ have enough representation power.

$$\min_w \varphi_h(w) \quad \text{where} \quad \varphi_h(w) = \max_\theta \frac{1}{2n} \sum_{i=1}^{n} h(f_\theta(x_i) - f_\theta(G_w(z_i))) \,.$$

Proposition 1 (informal) The loss function $\varphi_h(w)$ is NOT global-min-reachable.

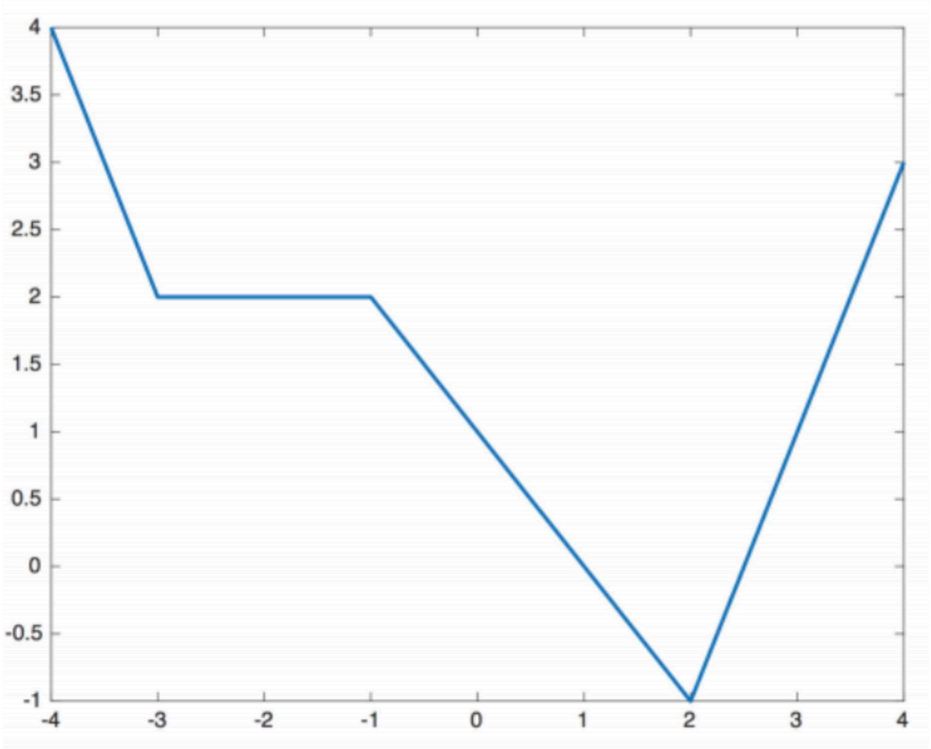
$$\min_w \varphi_{h,\mathrm{R}}(w) \quad \text{where} \quad \varphi_{h,\mathrm{R}}(Y, X) = \max_\theta \frac{1}{2n} \sum_{i=1}^{n} h(f_\theta(x_i) - f_\theta(G_w(z_i))) \,.$$

Proposition 2 (informal) The loss function $\varphi_{h,\mathrm{R}}(w)$ is global-min-reachable.

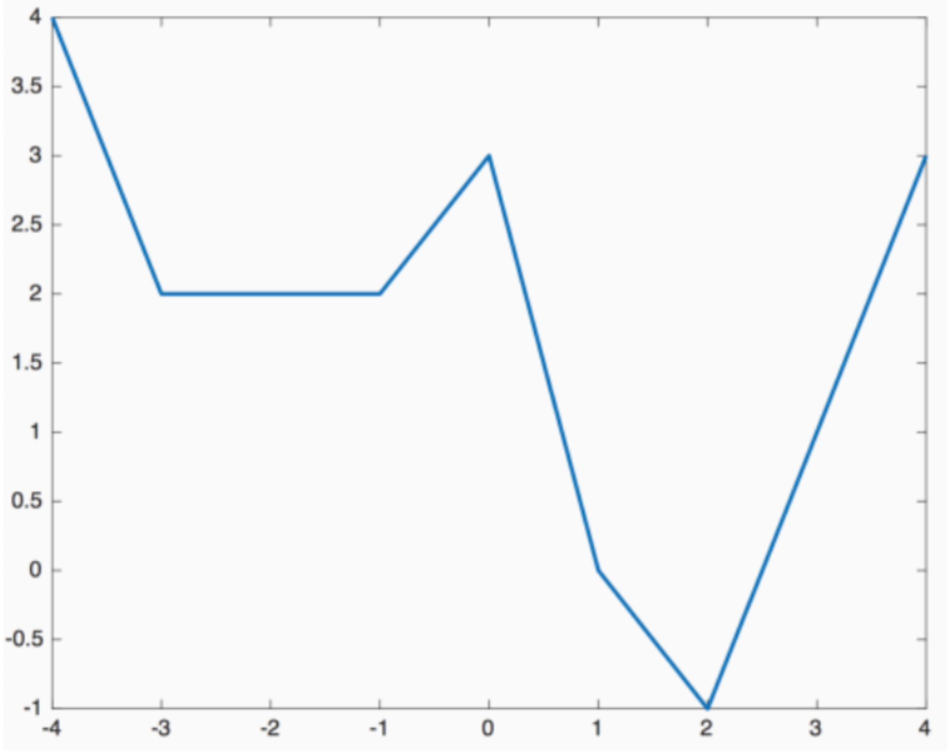
Neural-net landscape

Basin (informal): a region with no non-increasing path to globa-min. See [Li, Ding, Sun'2019] for "no bad basin" in neural-nets.

Simple examples of without and with sub-optimal basin.



No Bad Basin (with flat bad local-min)



One Bad Basin

Width eliminates bad basin

A useful concept in understanding neural-net landscape.

There is a phase transition from under to over-parameterized networks: [Li, Ding, Sun’2019]

—with <= n-1 neurons, a 1-hidden-layer neural-net can have bad basins (for certain settings)

—with >= n neurons in the last layer, a deep neural-net can have no bad basin, almost all settings..

Proof for R-GAN: Graph theory

Proof Sketch of Theorem 2:

1) Build a directed graph, with points representing x_i and y_i's, and directed edges from x_i and y_i.

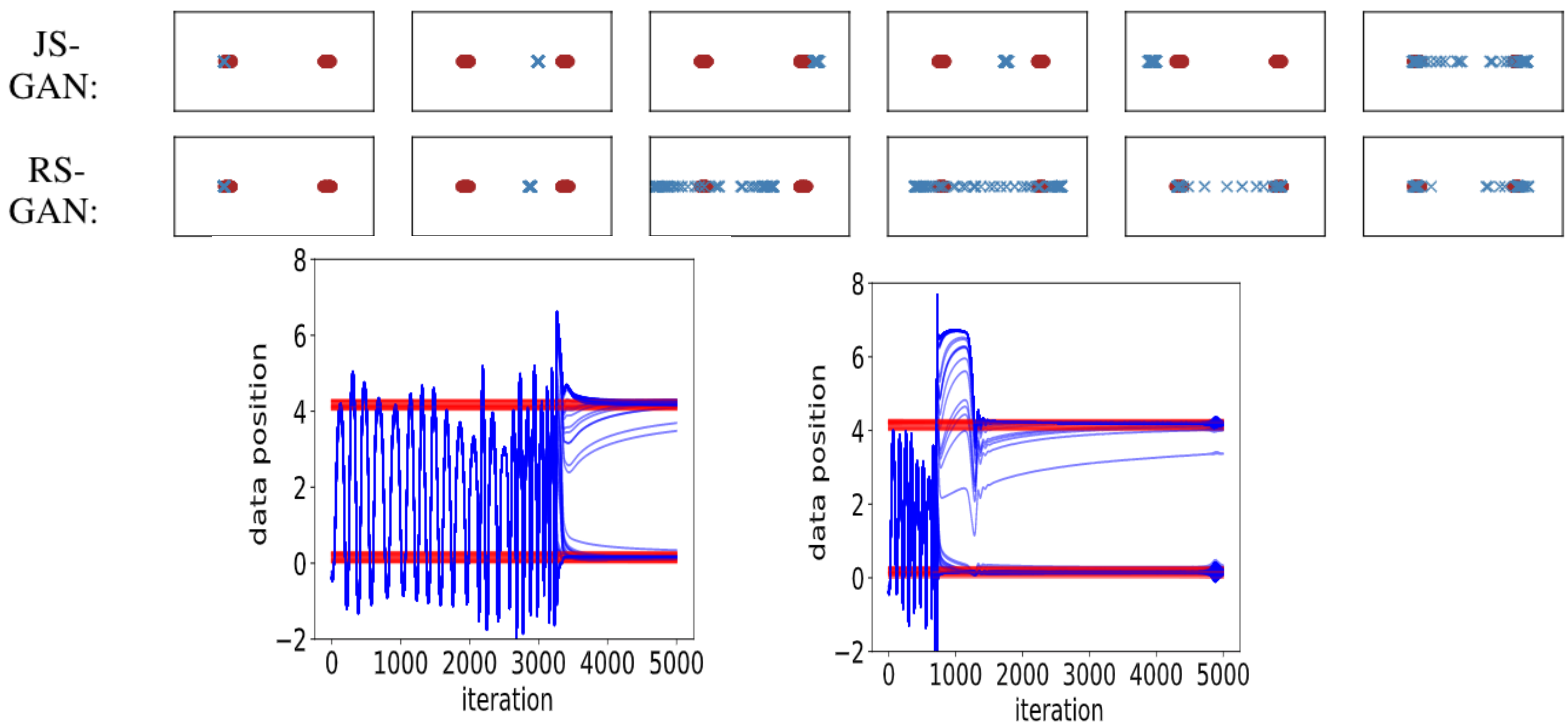
2) A directed graph with out-degree <= 1 can be decomposed into cycles and trees.

3) Each length-K cycle contributes $-(K/n)$ log 2 to the function value. Each tree contributes 0.

Part IV Explainig Two-Cluster Experiments

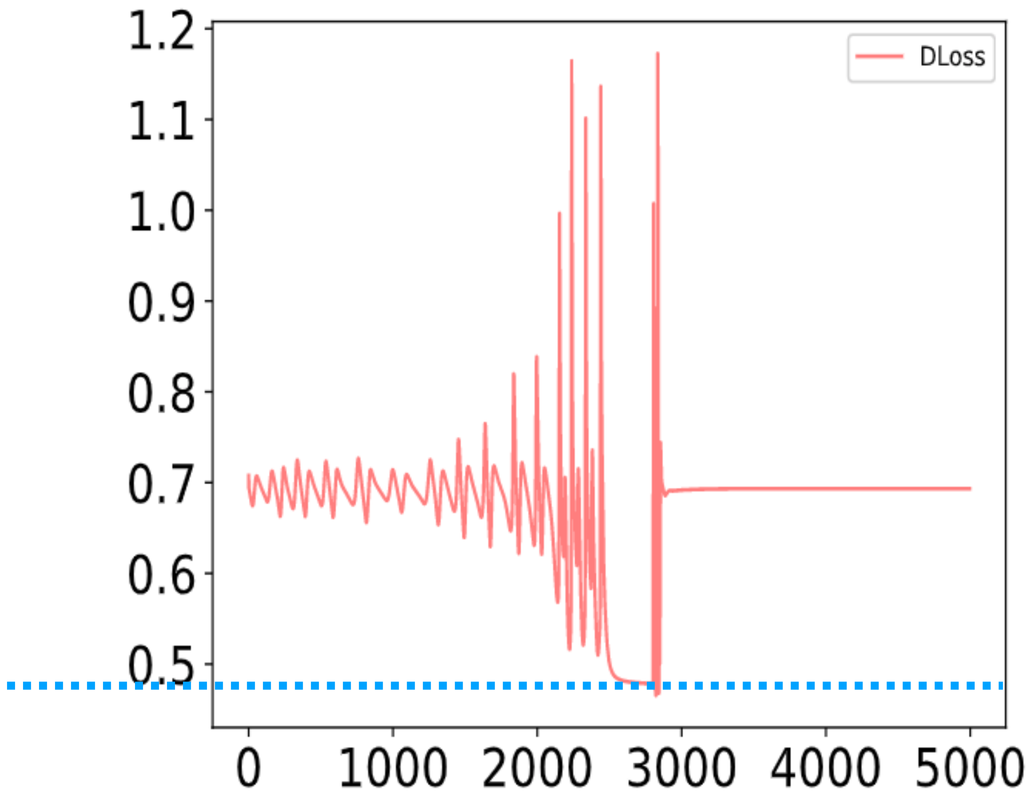
Understanding Training

True data**:** two clusters (red).
Fake data: blue points.
4-layer neural-net; standard training (alternating gradient descent ascent)

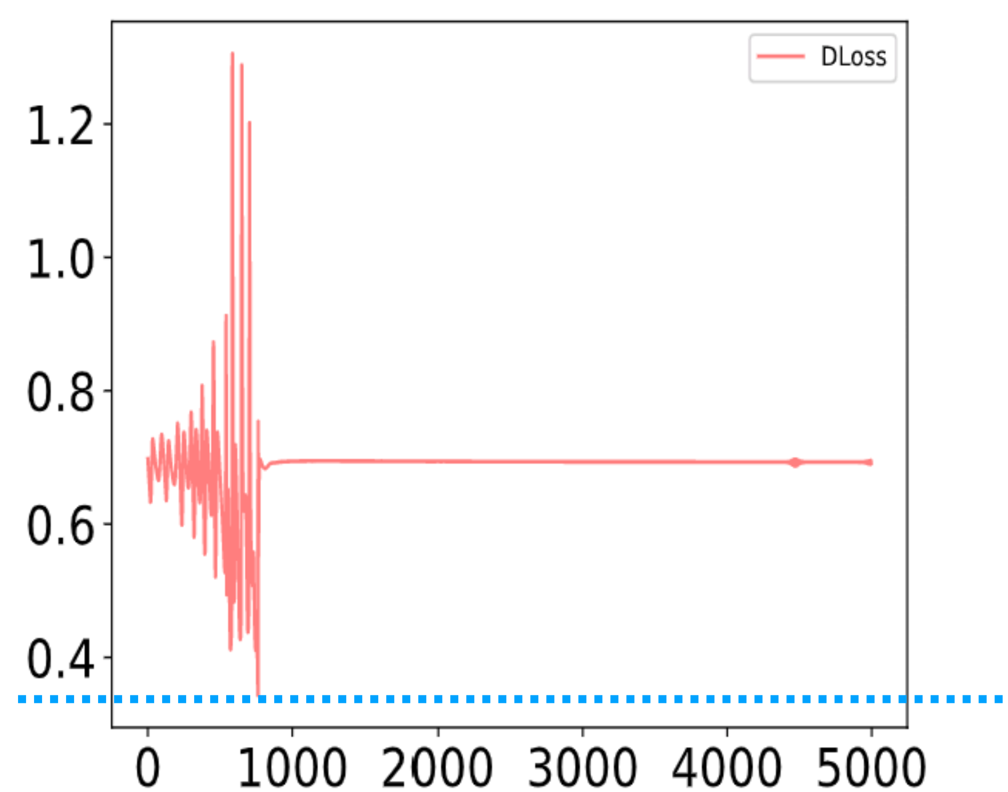


RS-GAN is faster than JS-GAN.

loss over iteration: mysterious?



(a) JS-GAN



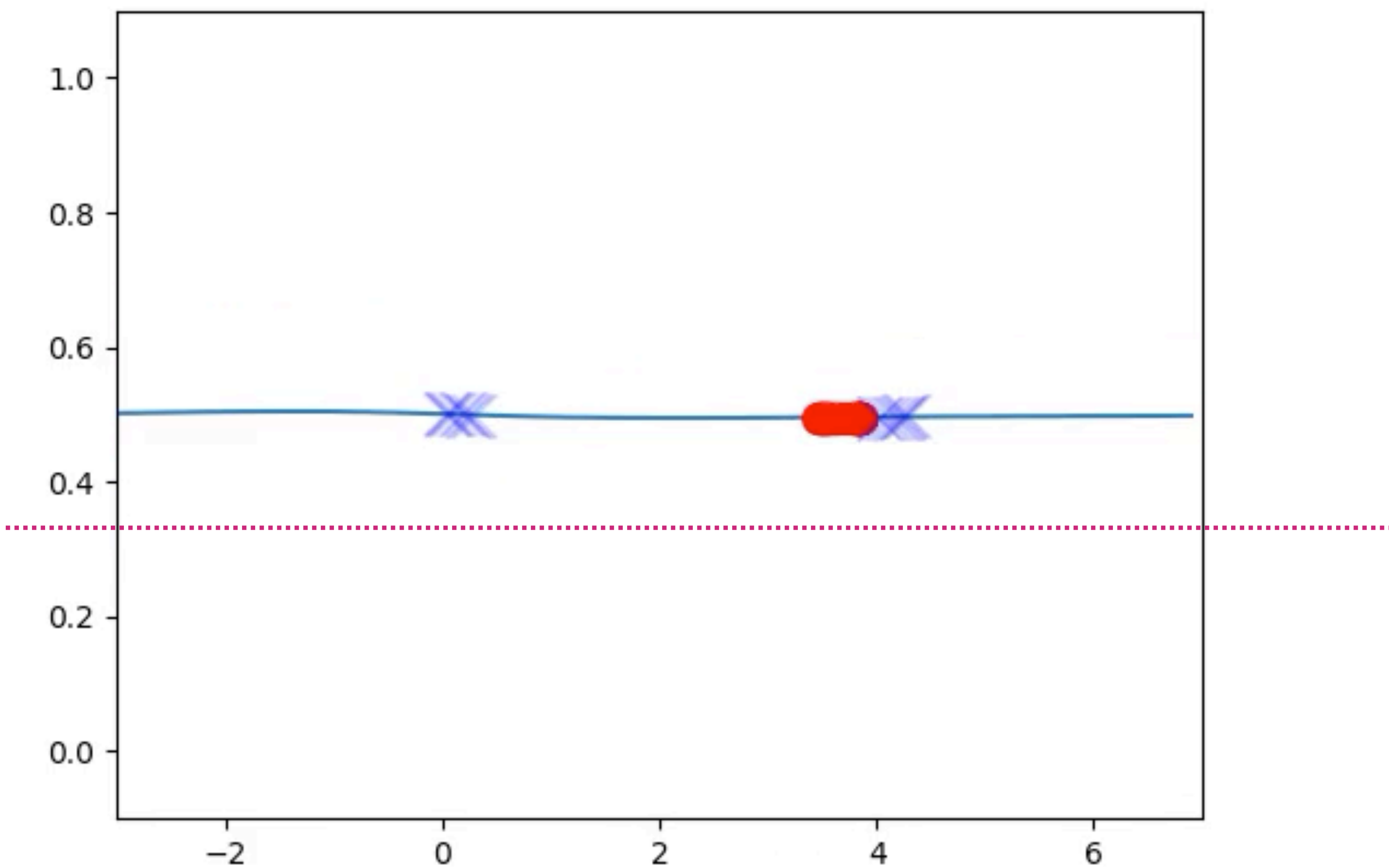
(b) RS-GAN

We draw the loss over iteration.

Unlike pure minimization problem, the plot is hard to interpret.

Suggestion 1: Check **minimal loss value**. Left: **0.48**; Right: 0.35.

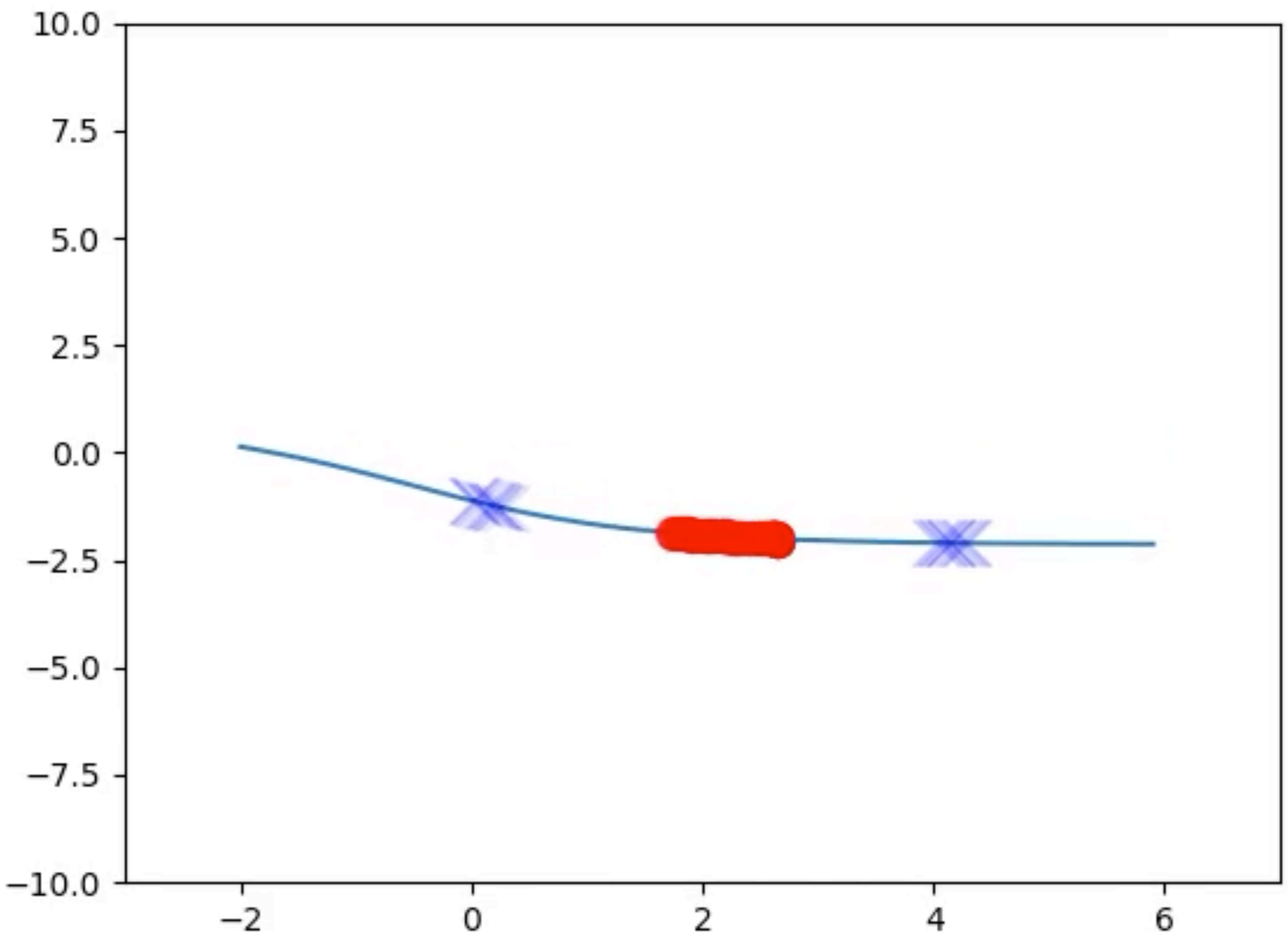
JS-GAN training process



Y: red points, want to climb up D: function; want to push Y down

Basin (equilibrium) (D, Y): D(0) = 1/3, D(1) = 1. Y is mode collapse

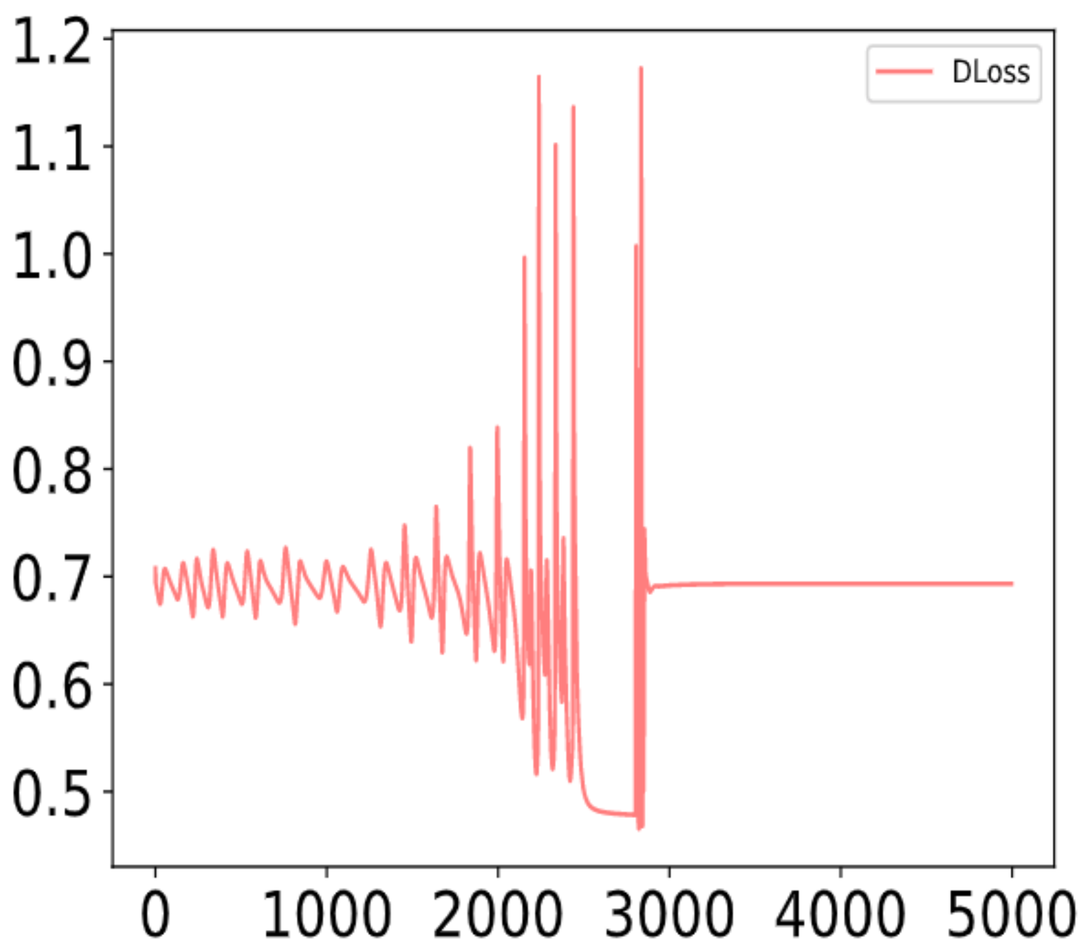
RS-GAN Training Process

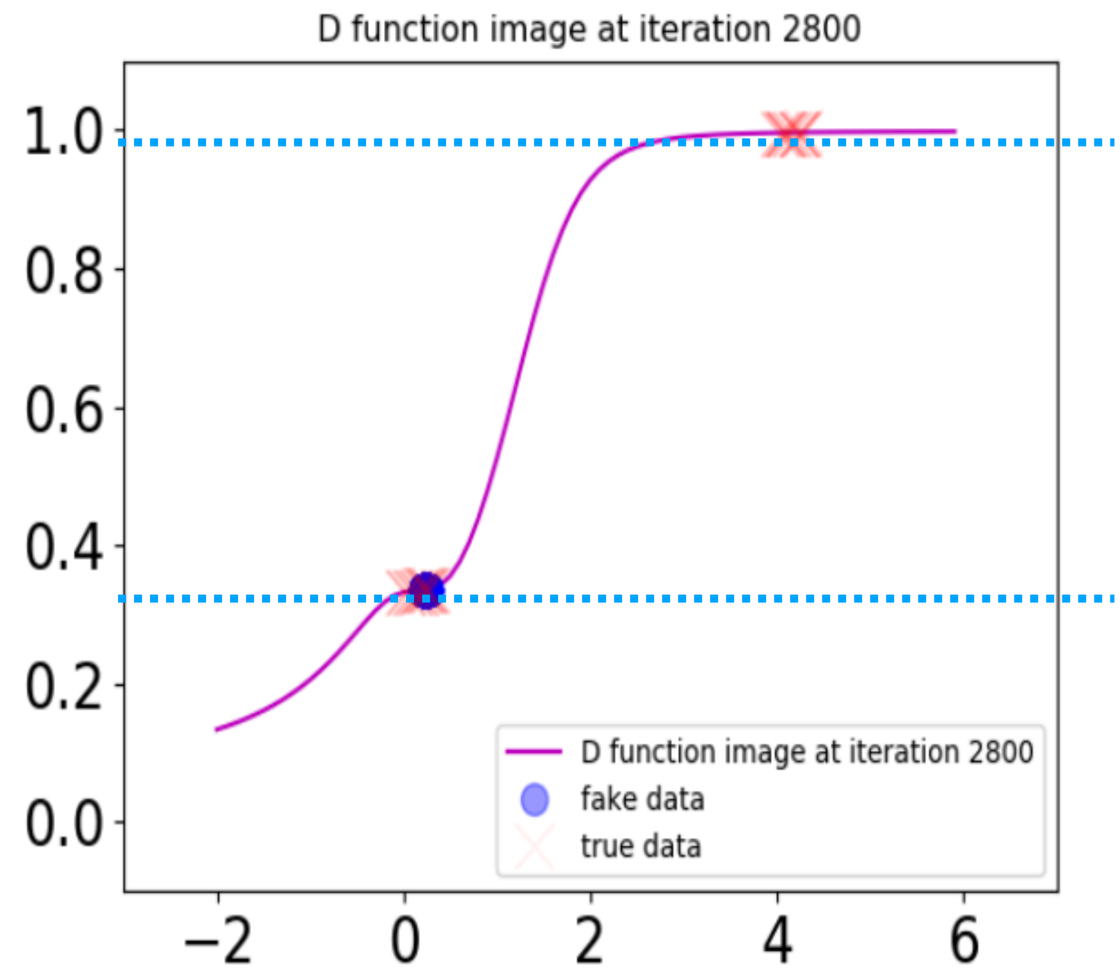


Y: red points, want to climb up D: function; want to push Y down

No basin. Mode collapse will not attract iterates strongly.

Understanding Training





u* = (mode-collapse Y, optimal D for Y) is attractor.

By theory: **D*(0) = 1/3; D*(1) = 1. Match right plot.**

Right plot: visualization attractor in space of (samples Y; **function D**)

Math Essence: Equilibrium Points

Non-linear dynamics is very complicated.

(Poincare, Smale, …: I said so!)

This work: Let's identify **equilibrium points**, ignore details of dynamics for now.

Real-data Experiments

Two Lines of Code Change

Plug-and-Play Change: two lines of change in code

Original GAN (D and G loss):

```
return (self.BLL(logitX, torch.ones_like(logitX)) + self.BLL(logitG, torch.zeros_like(logitG)))/2

return self.BLL(logitG, torch.ones_like(logitG))
```

RS-GAN (D and G loss; can swap the two)

```
return self.BLL(logitG - logitX, torch.ones_like(logitX))

return self.BLL(logitX - logitG, torch.ones_like(logitX))
```

Predictions

Predictions:

P0) JS-GAN is better than RS-GAN; sometimes huge gap

P1) For narrow net, the gap is larger.

(reason: **wide nets have better landscape**, thus help JS-GAN to escape basins).

P2) Exists bad initial point that JS-GAN training fails.

P0) Previous Achievement

Achievement 1: ESRGAN (Wang et al., 2018) applied a variant of RSGAN, as a major improvement over SRGAN, and which won the PIRM2018- SR competition (region 3).

Achievement 2: CAT data set, R-GANs can work; standard GANs fail. **2k images.**

Ian Goodfellow @goodfellow_ian · Jul 3, 2018

This new family of GAN loss functions looks promising! I'm especially excited about Fig 4-6, where we see that the new loss results in much faster learning during the first several iterations of training. I implemented

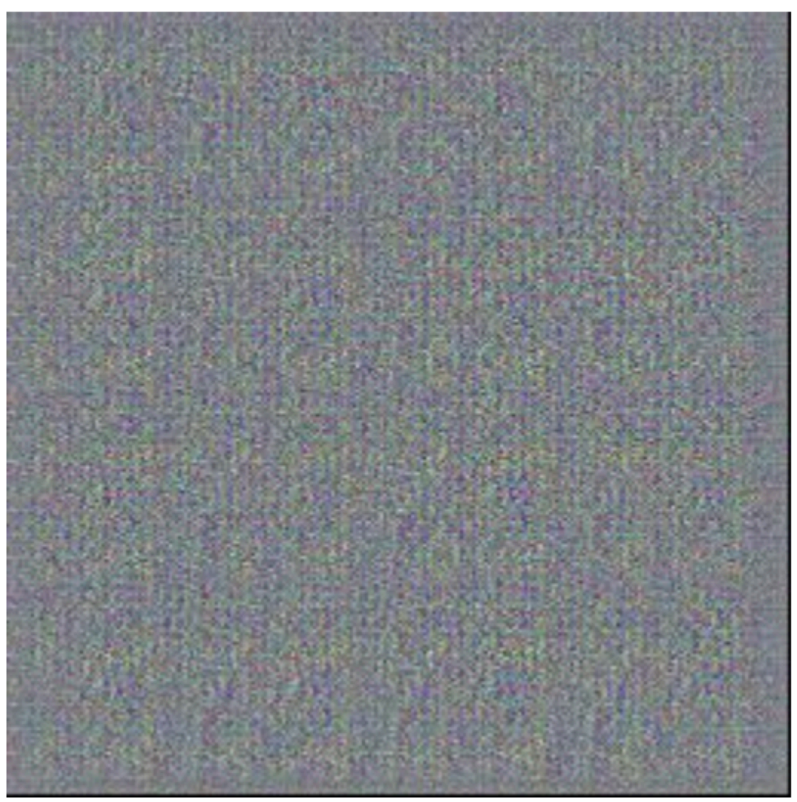


Figure 4: 256x256 cats with GAN (5k iterations)

JS-GAN; Source: [JM'19]

Figure 6: 256x256 cats with RaSGAN (FID = 32.11)

RS-GAN variant; Source: [JM'19]

P0) JS-GAN v.s. RS-GAN: Regular gap

Scores on CIFAR-10. After extensive tuning to achieve best results for each case. SN (spectral normalization) shrinks the gap.

FID score: lower better. **IS**: higher better.

	CIFAR-10		
	Inception Score ↑	FID ↓	Model size
Real Dataset	11.24±0.19	5.18	
Standard CNN			
JS-GAN	6.27±0.10	49.13	100%
WGAN-GP	6.68±0.06	39.66	100%
RS-GAN	7.02±0.07	33.79	100%
JS-GAN+ SN	7.42±0.08	28.07	100%
RS-GAN+ SN	7.32±0.08	27.16	100%

Gap: 15.3

P1) Narrower ==> Bigger gap

SN paper, BigGAN paper use **hinge loss**.

We compare hingeGAN, and R-hingeGAN. **5-10 FID score gap**.

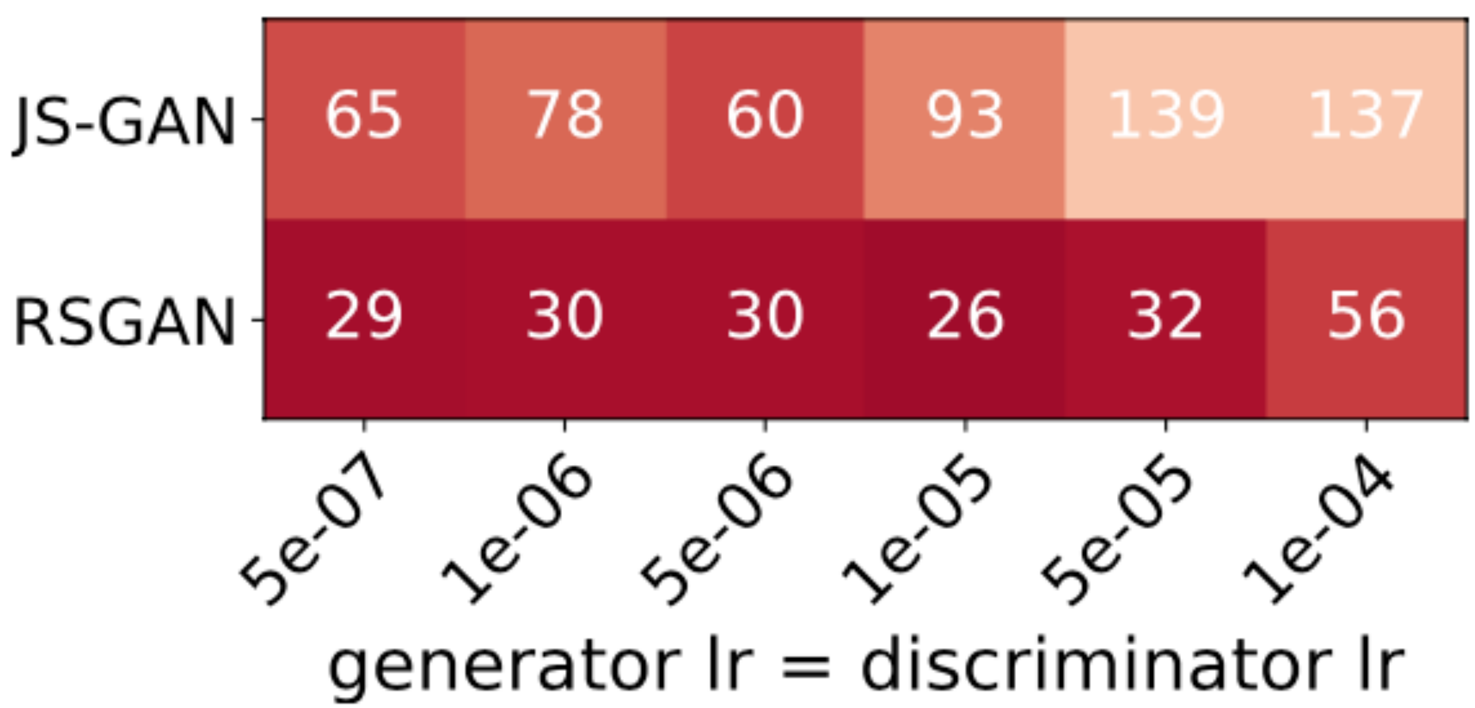
	CIFAR-10	
	IS ↑	FID ↓
ResNet + Hinge Loss		
JS^{hinge}	7.92±0.08	21.30
JS^{hinge}+GD channel/2	7.63±0.05	27.21
JS^{hinge}+GD channel/4	6.79±0.09	37.51
JS^{hinge}+BottleNeck	7.16±0.10	33.24
R^{hinge_HL}	8.03±0.09	19.07
R^{hinge_HL}+GD channel/2	7.69±0.10	22.79
R^{hinge_HL}+GD channel/4	7.11±0.06	32.35
R^{hinge_HL}+BottleNeck	7.52±0.05	24.07

Gap: 1.2

Gap: 9.2
with 16% size

P2) Bad initial point exists

Find one initial point to distinguish them. MNIST.



FID score: Lower is Better.

Concluding Remarks

Summary

- We theoretically analyze **empirical version** of GANs, in function space and parameter space (for neural-nets).
- JS-GAN has **bad basin**; they are **mode collapse**
- **RS-GAN does not have bad basin**
- **Simulation**: 0) RS-GAN outperforms JS-GAN

  1) Narrower nets: RS-GAN even better.

  2) Evidence for "better landscape of RSGAN": distinguishing initial point

Summary: Big Picture

- We hope to provide a "linear regression model of GANs": a simplest model that is analyzable globally
- A non convex-concave model that is possibly tractable
- Mathematically speaking, identifying "equilibrium points" in a complex game is a common approach

Future Directions

Theory:

- Better understanding of GAN behavior
- Optimization theory on special classes of games

Practice:

- Efficient GAN training (BigGAN is too big…)

Reference: On the global landscape of generative adversarial networks. **Ruoyu Sun**, Tiantian Fang, Alex Schwing. (under review)

—happy to share upon request.

Thank you for listening!