

Towards a Better Global Loss Landscape of GANs

Sun et al., 2020 NeurIPS

CS598 GAN 4

Outline

Introduction

Motivation and Method

Results

Outline

Introduction

Motivation and Method

Results

Introduction

- ▶ Theoretical efforts to understanding of GANs focus on statistics or optimization.
- ▶ Statistics:
 - ▶ In the **GAN** paper, Goodfellow et al. (2014) linked the min-max formulation and the JS divergence.
 - ▶ **WGAN** (Arjovsky et al., 2017) proposed a loss function based on the Wasserstein distance.
 - ▶ Wasserstein distance and JS distance are not generalizable but real metric used in practice is generalizable (Arora et al., 2017).
- ▶ Optimization:
 - ▶ Cyclic behavior, non-convergence, for min-max optimization: cycle around a stable point to slowly converge or diverge.
 - ▶ Sub-optimal local minima issues for **GANs**: current works only perform local analysis or global analysis with simple setting.
- ▶ Therefore, main goal of the paper is to perform global analysis on **GANs** for general data distribution.

Introduction

Table 1: Comparison of theoretical works.







	Supervised Learning		GANs	
	paper	brief description	paper	brief description
Generalization analysis	[9]	generalization bound for neural-nets	[5]	generalization bound for GANs
Convergence analysis	[77]	convex problem, divergence of Adam convergence of AMSGrad	[23]	bi-linear game, non-convergence of GDA convergence of optimistic GDA
Global landscape	[73] [50]	Any distinct input data Wide neural-nets have no sub-optimal basins	This work	Any distinct input data SepGAN has bad basins; RpGAN does not

* This table does NOT show a complete list of works. The goal is to list various types of works. Only one or two works are listed as examples of that class.

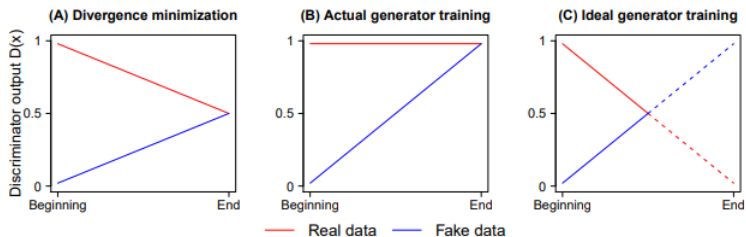
Relativistic GANs

- ▶ Discriminator should utilize both real and fake data to measure the reality of samples, therefore, the measurement of absolute reality is substituted with relative reality.
- ▶ Having the priori knowledge that input samples consist of half real and half fake samples, discriminator estimates real samples with lower score as the fake samples are more realistic than the real samples.

Relativistic GANs

Scenario	Absolute probability (Standard GAN)	Relative probability (Relativistic average Standard GAN)
Real image looks real and fake images look fake		
	$C(x_r) = 8$ $P(x_r \text{ is bread}) = 1$	$\overline{C(x_f)} = -5$ $P(x_r \text{ is bread} \overline{C(x_f)}) = 1$
Real image looks real but fake images look similarly real on average		
	$C(x_r) = 8$ $P(x_r \text{ is bread}) = 1$	$\overline{C(x_f)} = 7$ $P(x_r \text{ is bread} \overline{C(x_f)}) = .73$
Real image looks fake but fake images look more fake on average		
	$C(x_r) = -3$ $P(x_r \text{ is bread}) = .05$	$\overline{C(x_f)} = -5$ $P(x_r \text{ is bread} \overline{C(x_f)}) = .88$

Relativistic GANs



- ▶ The training is problematic by only pushing $D(fake)$ to 1 and ignoring $D(real)$.

Relativistic GANs

- ▶ To make discriminator relativistic, we can sample real/fake data pairs $\hat{x} = (x_r, x_f)$ with $D(\hat{x}) = \textit{sigmoid}(C(x_r) - C(x_f))$.
- ▶ The discriminator estimates the probability that the given real data is more realistic than a randomly sampled fake data.

Relativistic GANs

- ▶ Therefore, we have loss in general form:

$$\begin{aligned}L_D^{RGAN} &= \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [f_1 (C(x_r) - C(x_f))] \\ &\quad + \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [f_2 (C(x_f) - C(x_r))] \\ L_G^{RGAN} &= \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [g_1 (C(x_r) - C(x_f))] \\ &\quad + \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [g_2 (C(x_f) - C(x_r))]\end{aligned}$$

- ▶ The point is to use real/fake paired data for training.

Outline

Introduction

Motivation and Method

Results

Intuition

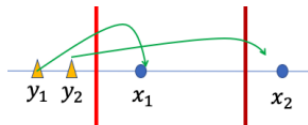
- ▶ Suppose we have two real samples x_1, x_2 and two generated samples y_1, y_2 in one dimension, then we can intuitively illustrate the training process of standard GANs as:



- ▶ Easily lead to mode collapse.

Intuition

- ▶ Having the same setting, the training process of relativistic paired GANs is:



- ▶ Seem to have no mode collapse.

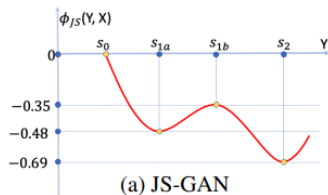
Bad Local Minima

- ▶ To further investigate how pairing influence local minima, the paper sets a two-point case:
 - ▶ Two real samples x_1, x_2 and two generated samples y_1, y_2 .
 - ▶ Four states s_0, s_{1a}, s_{1b}, s_2 represent $|\{x_1, x_2\} \cap \{y_1, y_2\}| = 0$,
 $y_1 = y_2 \in \{x_1, x_2\}$, $|\{x_1, x_2\} \cap \{y_1, y_2\}| = 1$,
 $\{x_1, x_2\} = \{y_1, y_2\}$.
- ▶ s_0 : no overlap
- ▶ s_2 : perfect match
- ▶ s_{1a} : mode collapse
- ▶ s_{1b} : mode dropping

Bad Local Minima

$$\phi_{JS}(Y, X) = \begin{cases} -\log 2 \approx -0.6931 & \text{if } s_2, \\ -\log 2/2 \approx -0.3467 & \text{if } s_{1b}, \\ \frac{1}{4}(2 \log 2 - 3 \log 3) \approx -0.4774 & \text{if } s_{1a}, \\ 0 & \text{if } s_0 \end{cases}$$

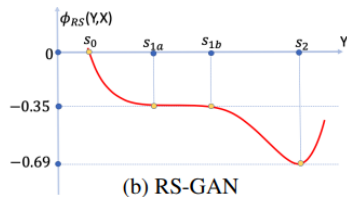
where ϕ_{JS} is the divergence measurement in JS-GAN.



Sub-optimal local minima at s_{1a} .

Bad Local Minima

$$\phi_{RS}(Y, X) = \begin{cases} -\log 2 \approx -0.6931 & \text{if } s_2, \\ -\frac{1}{2} \log 2 \approx -0.3466 & \text{if } s_{1a}, s_{1b}, \\ 0 & \text{if } s_0 \end{cases}$$



Landscape Results in Function Space

- ▶ Extend from $n=2$ to general n .

Theorem 1.

Suppose $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ are distinct. Suppose h_1, h_2 satisfy Assumptions 4.1, 4.2 and 4.3. Then for separable-GAN loss $g_{\text{SP}}(Y)$ defined in Eq. (5), we have:

(i) The global minimal value is $-\frac{1}{2} \sup_{t \in \mathbb{R}} (h_1(t) + h_2(-t))$, which is achieved iff $\{y_1, \dots, y_n\} = \{x_1, \dots, x_n\}$.

(ii) If $y_i \in \{x_1, \dots, x_n\}$, $i \in \{1, 2, \dots, n\}$ and $y_i = y_j$ for some $i \neq j$, then Y is a sub-optimal strict local minimum. Therefore, $g_{\text{SP}}(Y)$ has $(n^n - n!)$ sub-optimal strict local minima.

- ▶ The theorem is saying that (i) For standard GANs, global minimal achieves when two sets of points perfectly match; (ii) Local minima always exists in some scenarios.

Landscape Results in Function Space

Definition (global-min-reachable)

We say a point w is global-min-reachable for a function $F(w)$ if there exists a continuous path from w to one global minimum of F along which the value of $F(w)$ is non-increasing.

Theorem 2.

Suppose $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ are distinct. Suppose h satisfies Assumptions 4.4 and 4.5. Then for RpGAN loss $g_{\mathbb{R}}$ defined in Eq. (6): (i) The global minimal value is $h(0)$, which is achieved iff $\{y_1, \dots, y_n\} = \{x_1, \dots, x_n\}$. (ii) Any Y is global-min-reachable for the function $g_{\mathbb{R}}(Y)$.

- ▶ Theorem 2 means (i) For RpGAN, global minimal achieves when two sets of points perfectly match; (ii) Only global minima.

Landscape Results in Parameter Space

- ▶ Different from before, where we optimize over y_i and f (function space), we now optimize over w and θ (parameter space).
- ▶ Assume generator and discriminator are extremely expressive, the analysis of optimization on parameter space gives the same results.

Outline

Introduction

Motivation and Method

Results

Results of Case Study

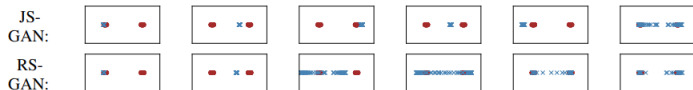


Figure 5: Training process of JS-GAN and RS-GAN for two-cluster data. True data are red, fake data are blue. RS-GAN escapes from mode collapse faster than JS-GAN.

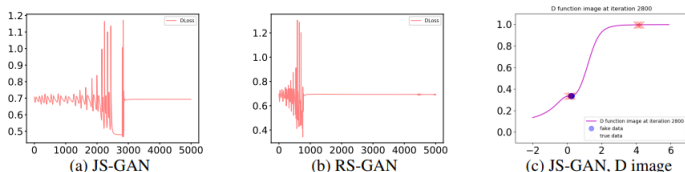


Figure 6: (a) and (b): Evolution of D loss over iterations. RS-GAN is $3-4\times$ faster than JS-GAN. (c) For JS-GAN training in (a), we plot (Y, D) together at iteration 2800. Y are represented in blue points, and they are near $c_1 = 0$. D is near the optimal $D^*(s_{1a})$ since $D(0) \approx 1/3$ and $D(4) \approx 1$. Interestingly, this bad attractor (Y, D) is similar to the one discussed in Fig. 1, so the intuition of “local-min” is verified in (c).

Results of Real Data Experiments

	CIFAR-10				STL-10		
	Inception Score \uparrow	FID \downarrow	FID Gap	Model size	Inception Score \uparrow	FID \downarrow	FID Gap
Real Dataset	11.24 \pm 0.19	5.18			24.45 \pm 0.41	5.34	
Standard CNN							
WGAN-GP	6.68 \pm 0.06	39.66			8.11 \pm 0.09	55.64	
JS-GAN	6.27 \pm 0.10	49.13	15.34	100%	8.01 \pm 0.07	50.38	2.16
RS-GAN	7.02 \pm 0.07	33.79			7.62 \pm 0.08	52.54	
JS-GAN+ SN	7.42 \pm 0.08	28.07	0.91	100%	8.32 \pm 0.10	44.06	0.18
RS-GAN+ SN	7.32 \pm 0.08	27.16			8.29 \pm 0.13	43.88	
JS-GAN+SN; GD channel/2	6.85 \pm 0.08	33.90	1.16	29.0%	7.69 \pm 0.05	57.16	4.69
RS-GAN+SN; GD channel/2	6.74 \pm 0.04	32.74			7.95 \pm 0.10	52.47	
JS-GAN + SN; GD channel/4	5.83 \pm 0.07	52.63	7.26	9.2%	6.90 \pm 0.06	72.96	9.35
RS-GAN + SN; GD channel/4	5.94 \pm 0.09	45.37			7.27 \pm 0.11	63.61	
ResNet							
JS-GAN+ SN	8.12 \pm 0.14	20.13	0.82	100%	8.87 \pm 0.07	36.33	1.56
RS-GAN + SN	7.92 \pm 0.13	19.31			8.96 \pm 0.10	34.77	
JS-GAN + SN; GD channel/2	7.67 \pm 0.04	23.29	1.51	27.5%	8.45 \pm 0.05	44.39	2.21
RS-GAN + SN; GD channel/2	7.63 \pm 0.07	21.78			8.47 \pm 0.09	42.18	
JS-GAN + SN; GD channel/4	6.65 \pm 0.06	45.20	13.94	10.4%	8.21 \pm 0.12	53.57	1.48
RS-GAN+ SN; GD channel/4	7.08 \pm 0.05	31.26			8.46 \pm 0.11	52.09	
JS-GAN + SN; BottleNeck	7.60 \pm 0.07	26.98	1.54	16.8%	8.29 \pm 0.05	50.38	3.80
RS-GAN+ SN; BottleNeck	7.57 \pm 0.09	25.44			8.52 \pm 0.11	46.58	

Table 2: Inception score (IS) (higher is better) and Fréchet Inception distance (FID) (lower is better) for JS-GAN, WGAN-GP and RS-GAN on CIFAR-10 and STL-10. We also show FID gap between JS-GAN and RS-GAN and show the relative model size of narrow nets vs. regular nets (“regular”: CNN and ResNet of [67]).

Conclusion

- ▶ Analyze the landscape of JS-GAN and generalize the results to the standard GANs.
- ▶ The results show that the minimization problem has many sub-optimal local minima and each leads to a model collapse.
- ▶ Prove minimization problem of the relativistic pairing GANs has no local minima.
- ▶ Conduct experiments that support the landscape theories.

References

- ▶ I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NeurIPS, 2014.
- ▶ M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. In ICML, 2017.
- ▶ S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In ICML, 2017.
- ▶ A. Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In ICLR, 2018.
- ▶ R. Sun, T. Fang, A. Schwing. Towards a Better Global Loss Landscape of GANs. In NeurIPS, 2020.