

# Which Training Methods for GANs do actually Converge? (ICML 2018)

Lars Mescheder, Andreas Geiger, Sebastian Nowozin

Presenter:

Minhao Jiang (minhaoj2)

- The recent works show the local convergence of GAN training for absolutely continuous data and generator distributions.
- In this paper, the author discussed a counterexample showing that in the more realistic case of distributions that are not **absolutely continuous**, unregularized GAN training is not always convergent.
- The paper also showed that how recent techniques for stabilizing GAN training affect local convergence on the example problem, as WGAN, WGAN-GP, and DRAGAN do not converge on this example. And based on this observation, the paper introduced **simplified gradient penalties** and prove local convergence for the regularized GAN training dynamics.

## Problem Definition:

- We consider the traditional GAN training objective function as

$$L(\theta, \psi) = \mathbb{E}_{p(z)}[f(D_\psi(G_\theta(z)))] + \mathbb{E}_{p_D(x)}[f(-D_\psi(x))]$$

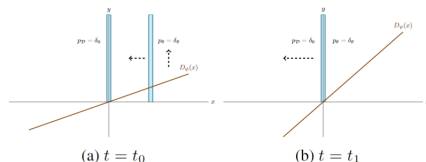
where the common choice  $f(t) = -\log(1 + \exp(-t))$

- Recently, it was shown that local convergence of GAN training near an equilibrium point  $(\theta^*, \psi^*)$  can be analyzed by looking at the spectrum of the Jacobian  $F'_h(\theta^*, \psi^*)$  at the equilibrium:
  - If  $F'_h(\theta^*, \psi^*)$  has eigenvalues with absolute value bigger than 1, the training algorithm will generally not converge to  $(\theta^*, \psi^*)$ .
  - On the other hand, If  $F'_h(\theta^*, \psi^*)$  has eigenvalues with absolute value smaller than 1, the algorithm will converge in sublinear time.
- Gradient vector field:

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

## Definition 1

The Dirac-GAN consists of a (univariate) generator distribution  $p_\theta = \gamma_\theta$  and a linear discriminator  $D_\psi(x) = \psi \cdot x$ . The true data distribution  $p_D$  is given by a Dirac-distribution concentrated at 0.



In this setup, the GAN training objective is given by  $L(\theta, \psi) = f(\psi\theta) + f(0)$

## Lemma 2

The unique equilibrium point of the training objective is given by  $\theta = \psi = 0$ . And the Jacobian of the gradient vector field has two eigenvalues  $\pm f'(0)i$ .

Considering the idealized continuous systems in GAN training dynamics, in the previous works, it was assumed that the optimal discriminator parameter vector is a continuous function of the current generator parameters.

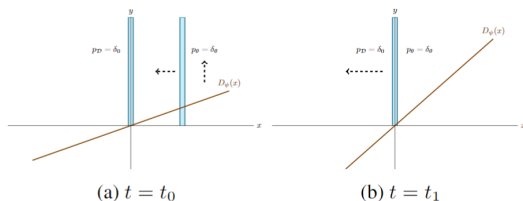
## Lemma 2.3

The integral curves of the gradient vector field  $v(\theta, \psi)$  do not converge to the Nash-equilibrium. Every integral curve  $(\theta(t), \psi(t))$  of the gradient vector field  $v(\theta, \psi)$  satisfies  $\theta(t)^2 + \psi(t)^2 = \text{const}$  for all  $t \in [0, \infty)$

In this case, unless  $\theta = 0$ , there is not even an optimal discriminator parameter for the Dirac-GAN.

And the following theorems showed that in two normal training dynamics of GAN: SimGD and AltGD, both encounter such instabilities.

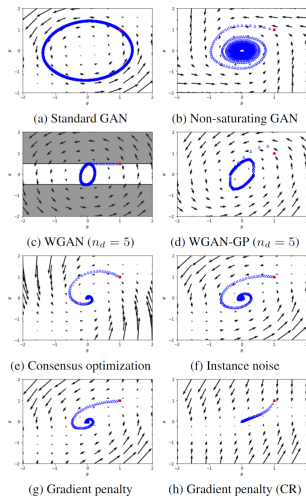
**Where do these instabilities come from?**



**Figure:** (a): In the beginning, the discriminator pushes the generator towards the true data distribution and the slope increases. (b): When the generator reaches the target distribution, the slope of the discriminator is the largest, pushing it away from the target distribution. This results in the oscillating behavior that will never converge.

Another way to look at it is to consider the local behavior of the training algorithm near the Nash-equilibrium, where there is no incentive for the discriminator to move to the equilibrium discriminator.

Note that WGAN and WGAN-GP both do not converge on this example.



**Figure:** Converging properties of different GAN training algorithms using alternating gradient descent.

A common technique to stabilize GANs is to add instance noise, i.e., independent Gaussian noise, to the data points.

For the Dirac-GAN:

## Lemma 3.2

When using Gaussian instance noise with standard deviation  $\sigma$ , the eigenvalues of the Jacobian of the gradient vector field are given by

$$\lambda_{1/2} = f''(0)\sigma^2 \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}$$

This theorem also implies that in the case of absolutely continuous distributions, gradient descent based GAN optimization is, under suitable assumptions, locally convergent.



## Zero-centered gradient penalties.

A penalty on the squared norm of the gradients of the discriminator results in the regularizer

$$R(\psi) = \frac{\gamma}{2}\psi^2$$

### Lemma 3.3

The eigenvalues of the Jacobian of the gradient vector field for the gradient-regularized Dirac-GAN at the equilibrium point are given by

$$\lambda_{1/2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - f'(0)^2}$$

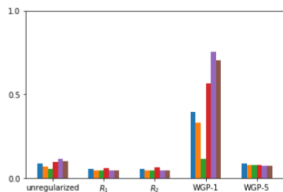
Like instance noise, there is a critical regularization parameter  $\gamma_{\text{critical}} = 2|f'(0)|$  that results in a locally rotation free vector field. And in this case, simultaneous and alternating gradient descent are both locally convergent.

The analysis suggests that the main effect of the zero-centered gradient penalties on local stability is to penalize the discriminator for deviating from the Nash-equilibrium. Then we can derive the following gradient penalties.

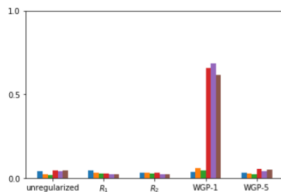
$$R_1(\psi) = \frac{\gamma}{2} \mathbb{E}_{p_D(x)} [\|\nabla D_\psi(x)\|^2] \quad (1)$$

$$R_2(\psi) = \frac{\gamma}{2} \mathbb{E}_{p_\theta(x)} [\|\nabla D_\psi(x)\|^2] \quad (2)$$

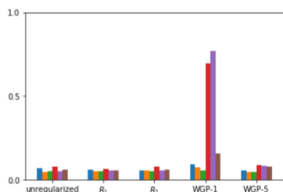
# Experiment Results



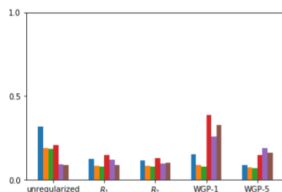
(a) 2D Gaussian



(b) Line segment



(c) Circle



(d) Four line segments

Figure: Experiments on 2D-Problems

- Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 2015–2025, 2017
- Nagarajan, V. and Kolter, J. Z. Gradient descent GAN optimization is locally stable. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5591–5600, 2017.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5769–5779, 2017.

# The relativistic discriminator: a key element missing from standard GAN (ICLR 2018)

Alexia Jolicoeur-Martineau

Presenter:

Minhao Jiang (minhaoj2)

In standard generative adversarial network (SGAN), the discriminator  $D$  estimates the probability that the input data is real. The generator  $G$  is trained to increase the probability that fake data is real. In this paper, the authors argue that it should also simultaneously decrease the probability that real data is real because

- 1 This would account for a priori knowledge that half of the data in the mini-batch is fake.
- 2 This would be observed with divergence minimization.
- 3 In optimal settings, SGAN would be equivalent to integral probability metric (IPM) GANs.

**Problem Definition:** GANs can be defined generally in terms of the discriminator in the following way

$$L_D = \mathbb{E}_{x_r \sim \mathbb{P}}[\tilde{f}_1(D(x_r))] + \mathbb{E}_{z \sim \mathbb{P}_z}[\tilde{f}_2(D(G(z)))] \quad (1)$$

$$L_G = \mathbb{E}_{x_r \sim \mathbb{P}}[\tilde{g}_1(D(x_r))] + \mathbb{E}_{z \sim \mathbb{P}_z}[\tilde{g}_2(D(G(z)))] \quad (2)$$

where  $\tilde{f}_1, \tilde{f}_2, \tilde{g}_1, \tilde{g}_2$  are scalar-to-scalar functions.  $\mathbb{P}$  is the distribution of the real data.

**Integral Probability Metrics (IPM):** IPMs are statistical divergences represented mathematically as

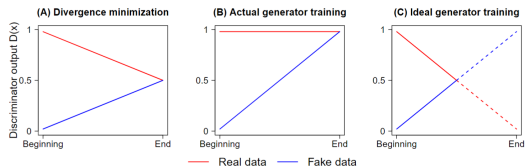
$$IPM_F(\mathbb{P}||\mathbb{Q}) = \sup_{C \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}}[C(x)] - \mathbb{E}_{x \sim \mathbb{Q}}[C(x)] \quad (3)$$

IPM-based GANs can be defined using equation 1 and 2 assuming

$\tilde{f}_1(D(x)) = \tilde{g}_2(D(x)) = -D(x)$  and  $\tilde{f}_2(D(x)) = \tilde{g}_1(D(x)) = D(x)$  and  $D(x) = C(x)$

# Missing Property of SGAN

In this paper, the authors argued that the key missing property of SGAN is that the probability of real data being real ( $D(x_r)$ ) should decrease as the probability of fake data being real ( $D(x_f)$ ) increase.



**Figure:** Expected discriminator output of the real and fake data for the direct minimization of the JSD, actual training of the generator to minimize its loss function, and ideal training of the generator to minimize its loss function.

SGAN completely ignores the a priori knowledge that half of the mini-batch samples are fake. And IPM-based GANs implicitly account for the fact that some of the samples must be fake because they compare how realistic real data is compared to fake data.



In SGAN, the discriminator loss function is equal to the Jensen-Shannon divergence. Thus, it can be represented as solving the following maximum problem

$$JSD(\mathbb{P}||\mathbb{Q}) = \frac{1}{2}(\log(4) + \max_{D:x \rightarrow [0,1]}) \mathbb{E}_{x_r \sim \mathbb{P}}[\log(D(x_r))] + \mathbb{E}_{x_f \sim \mathbb{Q}}[\log(1 - D(x_f))] \quad (4)$$

In terms of the gradient steps of SGAN and IPM-based GANs,

$$\nabla_w L_D^{GAN} = -\mathbb{E}_{x_r \sim \mathbb{P}}[(1 - D(x_r))\nabla_w C(x_r)] + \mathbb{E}_{x_f \sim \mathbb{Q}_\theta}[D(x_f)\nabla_w C(x_f)] \quad (5)$$

$$\nabla_\theta L_G^{GAN} = -\mathbb{E}_{z \sim \mathbb{P}_z}[(1 - D(G(z)))\nabla_x C(G_z)J_\theta G(z)] \quad (6)$$

$$\nabla_w L_D^{IPM} = -\mathbb{E}_{x_r \sim \mathbb{P}}[\nabla_w C(x_r)] + \mathbb{E}_{x_f \sim \mathbb{Q}_\theta}[\nabla_w C(x_f)] \quad (7)$$

$$\nabla_\theta L_G^{IPM} = -\mathbb{E}_{z \sim \mathbb{P}_z}[\nabla_x C(G_z)J_\theta G(z)] \quad (8)$$

In IPMs, oth real and fake data equally contribute to the gradient of the discriminator's loss function. However, in SGAN, if the discriminator reach optimality, the gradient completely ignores real data, which means if  $D(x_r)$  does not indirectly change when training the discriminator to reduce  $D(x_f)$ , the discriminator will stop learning what it means for data to be "real" and training will focus entirely on fake data.

The discriminator estimates the probability that the given real data is more realistic than a randomly sampled fake data. When the discriminator is defined only on  $C(x)$ . Then we have the discriminator and generator loss functions of the Relativistic Standard GAN

$$L_D^{RSGAN} = -\mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [\log(\text{sigmoid}(C(x_r) - C(x_f)))] \quad (9)$$

$$L_G^{RSGAN} = -\mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [\log(\text{sigmoid}(C(x_f) - C(x_r)))] \quad (10)$$

And for discriminator defined as  $a(C(x_r) - C(x_f))$

$$L_D^{RGAN} = \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [f_1(C(x_r) - C(x_f))] + \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [f_2(C(x_f) - C(x_r))] \quad (11)$$

$$L_G^{RGAN} = \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [g_1(C(x_r) - C(x_f))] + \mathbb{E}_{(x_r, x_f) \sim (\mathbb{P}, \mathbb{Q})} [g_2(C(x_f) - C(x_r))] \quad (12)$$

In RGANs,  $g_1$  is influenced by fake data, thus by the generator. This means that in most RGANs, the generator is trained to minimize the full loss function envisioned rather than only half of it.

Although the relative discriminator provide the missing property that we want in GANs (i.e.  $G$  influencing  $D(x_r)$ ), its interpretation is different from the standard discriminator. Rather than measuring “the probability that the input data is real”, it is now measuring “the probability that the input data is more realistic than a randomly sampled data of the opposing type.

So we define that

$$P(x_r \text{ is real}) = \mathbb{E}_{x_r \sim \mathbb{Q}}[D(x_r, x_f)] \quad (13)$$







$$P(x_f \text{ is real}) = \mathbb{E}_{x_f \sim \mathbb{P}}[D(x_f, x_r)] \quad (14)$$

where  $D(x_r, x_f) = \text{sigmoid}(C(x_r) - C(x_f))$

$$L_D^{RaGAN} = \mathbb{E}_{x_r \sim \mathbb{P}}[f_1(C(x_r) - \mathbb{E}_{x_f \sim \mathbb{Q}} C(x_f))] + \mathbb{E}_{x_f \sim \mathbb{Q}}[f_2(C(x_f) - \mathbb{E}_{x_r \sim \mathbb{P}} C(x_r))] \quad (15)$$

$$L_G^{RaGAN} = \mathbb{E}_{x_r \sim \mathbb{P}}[g_1(C(x_r) - \mathbb{E}_{x_f \sim \mathbb{Q}} C(x_f))] + \mathbb{E}_{x_f \sim \mathbb{Q}}[g_2(C(x_f) - \mathbb{E}_{x_r \sim \mathbb{P}} C(x_r))] \quad (16)$$

# Experiment Results

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

## Evaluation Metrics:

Loss	$lr = .0002$ $\beta = (.50, .999)$ $n_D = 1$	$lr = .0001$ $\beta = (.50, .9)$ $n_D = 5$	Loss	$lr = .001$	$\beta = (.9, .9)$	No BN	Tanh
SGAN	40.64	41.32	SGAN	154.20	35.29	35.54	59.17
RSGAN	36.61	55.29	RSGAN	50.95	45.12	37.11	77.21
RaSGAN	31.98	37.92	RaSGAN	55.55	43.46	41.96	54.42
LSGAN	29.53	187.01	LSGAN	52.27	225.94	38.54	147.87
RaLSGAN	30.92	219.39	RaLSGAN	<b>33.33</b>	48.92	<b>34.66</b>	53.07
HingeGAN	49.53	80.85	HingeGAN	43.28	<b>33.47</b>	34.21	58.51
RaHingeGAN	39.12	37.72	RaHingeGAN	51.05	42.78	43.75	<b>50.69</b>
WGAN-GP	83.89	<b>27.81</b>	WGAN-GP	61.97	104.95	85.27	59.94
RSGAN-GP	<b>25.60</b>	28.13					
RaSGAN-GP	331.86						

**Figure:** Experimental results of different GAN loss functions on CIFAR-10 datasets. Measured with FID scores.

- In the first paper, we analyzed the stability of GAN training on a simple yet prototypical example and we showed that (unregularized) gradient based GAN optimization is not always locally convergent. And the authors extended the local convergence with simplified zero-centered gradient penalties under suitable assumptions.
- In the second paper, the authors proposed the relativistic discriminator as a way to fix and improve on standard GAN. We further generalized this approach to any GAN loss and introduced a generally more stable variant called RaD. Our results suggest that relativism significantly improve data quality and stability of GANs at no computational cost