Implicit competitive regularization

Implicit competitive regularization in GANs

Schäfer, Florian, Zheng, Hongkai and Anandkumar, Anima

arXiv preprint https://arxiv.org/abs/1910.05852

Blogpost: https://f-t-s.github.io/projects/icr/

Caltech

Why do GANs work?

- GANs generate images using adversarial training
- Vastly superior image quality compared to VAE
- Why do they work?

Produced by StyleGan2 on https://thispersondoesnotexist.com/

The minimax interpretation:

- Original formulation:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{x\sim P_{\mathrm{data}}}\left[\log\left(\mathcal{D}(x)\right)\right] + \mathbb{E}_{z\sim\mathcal{N}}\left[\log\left(1 - \mathcal{D}\left(\mathcal{G}(z)\right)\right)\right]$$

- Max over $\mathcal{D}$ yields Jensen-Shannon divergence

$$\min_{\mathcal{G}} \mathrm{JSD}(\mathcal{G}, P_{\mathrm{data}})$$

- Common interpretation: GAN ≈ minimizes JSD

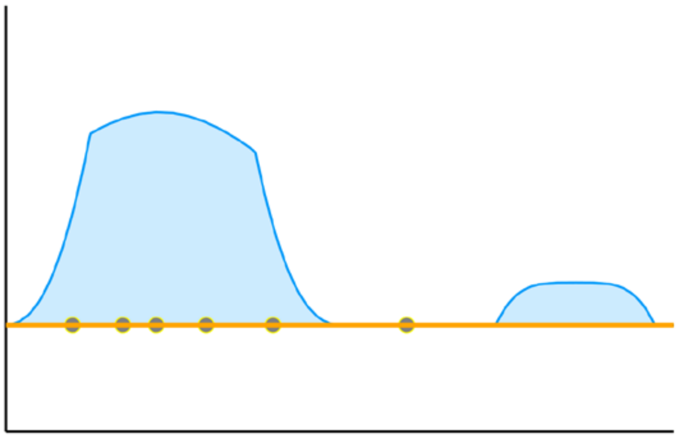
- Other GAN variants minimize other metrics/divs.

The GAN-dilemma:

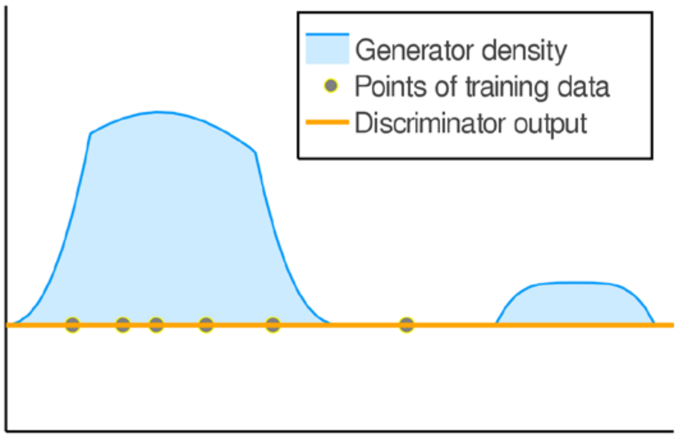
- Minimax interpretation has fundamental problem
  - Without regularity constraints, discriminator can (almost) always be perfect
  - Imposing constraints needs measure of similarity of images, which is hard to obtain.

Unconstrained discriminators:

- Unconstrained discriminators can (almost) always be perfect:

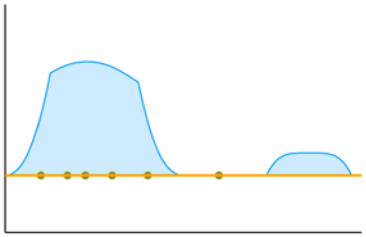


What the discriminator should do:

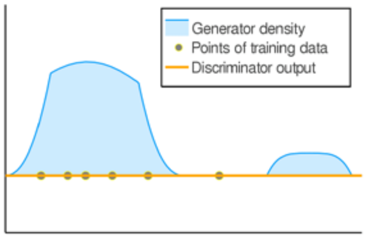


What it could do instead:

Caltech



Adding constraints:



- Idea: Impose regularity of discriminator.

- Example : WGAN

$$\min_{\mathcal{G}} \max_{\mathcal{D}:\, ||\nabla\mathcal{D}|| \leq 1} \mathbb{E}_{x\sim P_{\text{data}}}[\mathcal{D}(x)] - \mathbb{E}_{z\sim\mathcal{N}}\left[\mathcal{D}\big(\mathcal{G}(z)\big)\right].$$

- Constraint on $\nabla\mathcal{D}$ prevents *pointy* discriminators. But how to choose $||\cdot||$?

How to choose $|| \cdot ||$?:

- Usually $|| \cdot ||$ is Euclidean norm
- But $\mathcal{D}$ maps *images* to real numbers.
- We measure similarity of images by Euclidean distance of pixel-vectors!

A riddle:

Order the three columns of images by their pairwise Euclidean distance in pixel space:

A riddle:

The columns are ordered in *increasing* order of distance, from left to right

Caltech

A riddle :

The Euclidean distance is a *terrible* measure of similarity of images!

What now?

- Idea: Choose a different $|| \cdot ||$.
- But finding a measure of visual similarity is a *hard* and *unsolved* problem!
- Yet GANs work in practice, why?

A way out?

- DL outperforms classical methods
- → NN capture visual similarity
- But NN-uncertainty is poorly calibrated
- → Optimal discriminator not informative

Implicit competitive regularization

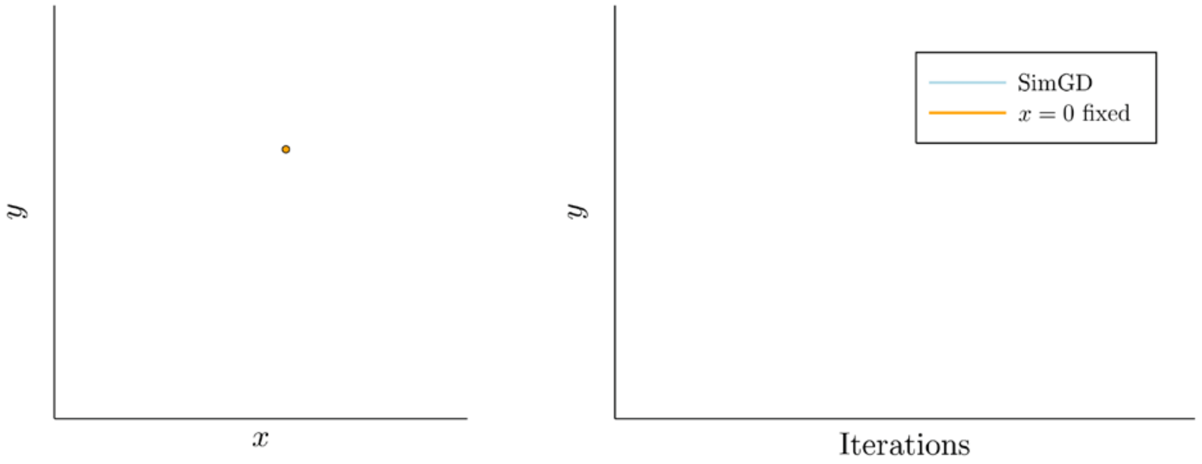
- SimGD has stable points, unstable for GD

$$\min_x \max_y x^2 + 10xy + y^2$$

- Optimal $y \to \infty$ but in SimGD $(x, y) \to (0,0)$

$\eta_x = 0.09$

$\eta_y = 0.01$



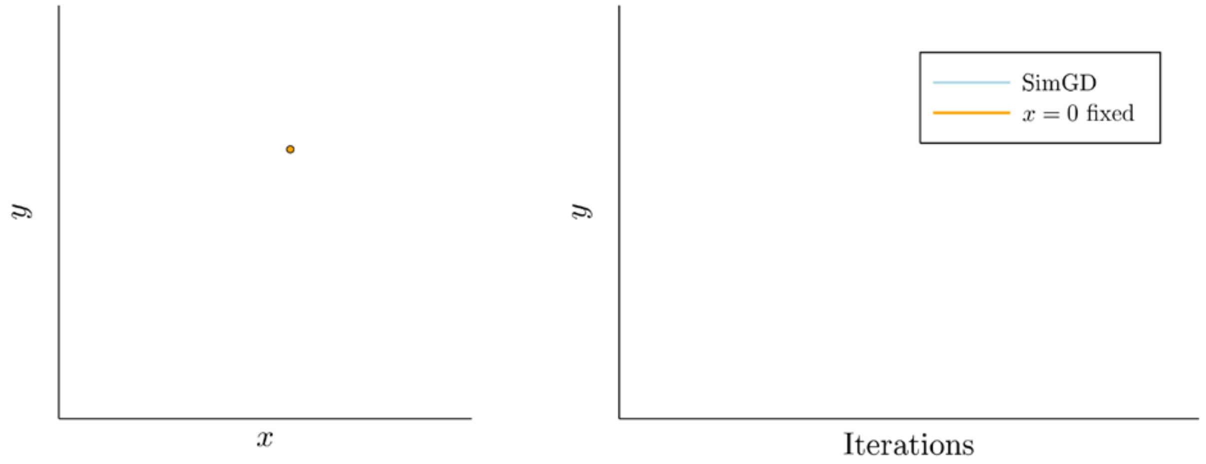
Caltech

Implicit competitive regularization

- Commonly seen as flaw of SimGD
- Just like y, $\mathcal{D}$ can always improve
  $\rightarrow$ICR crucial for convergence in GANs

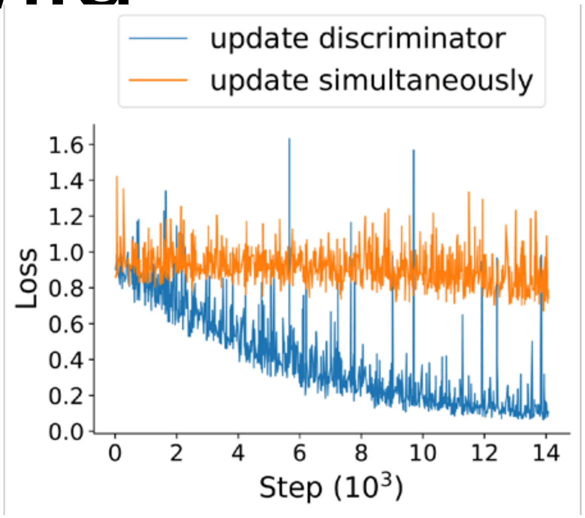
$\eta_x = 0.09$

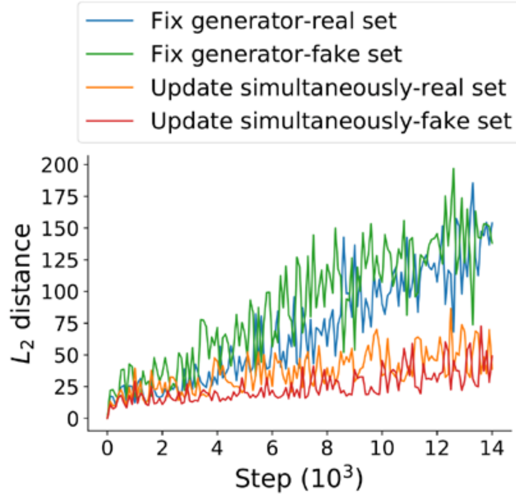
$\eta_y = 0.01$



ICR in the wild

- Train GAN on MNIST to *good* checkpoint
- Continue training with either SimGD or GD
- Measure changes in loss and output on reference images



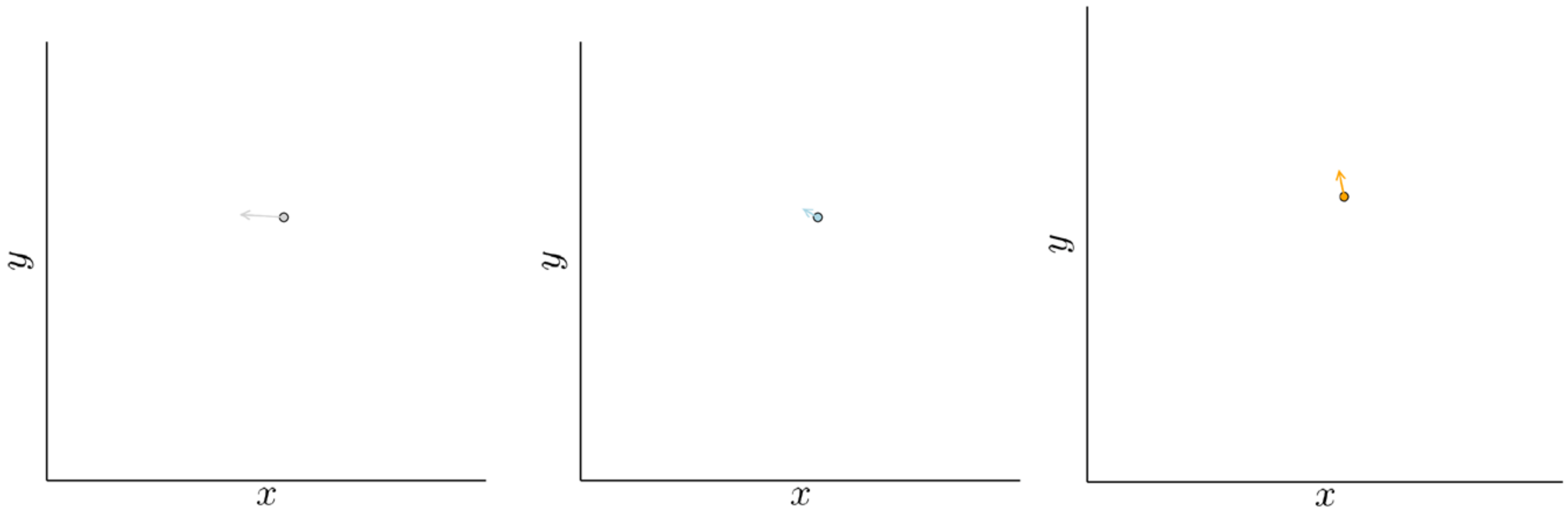


Caltech

ICR and speed of learning

- Consider different learning rates

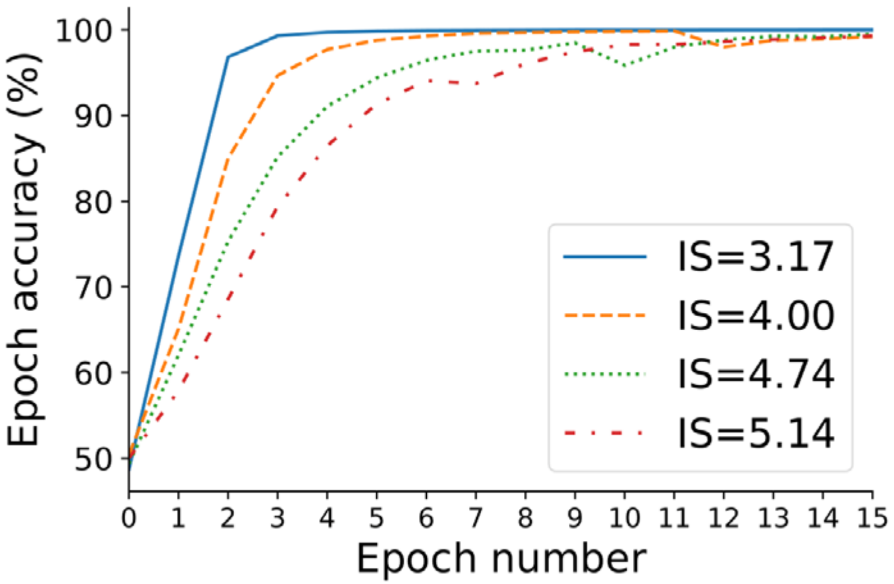
$\eta_x = 0.09, \eta_y = 0.01$ $\eta_x = 0.03, \eta_y = 0.03$ $\eta_x = 0.01, \eta_y = 0.09$



Hypothesis on NN behavior

- Hypothesis: Visually obvious mistakes are identified *more quickly* by the discriminator

- First evidence on CIFAR 10:



Caltech

Thesis: ICR enables GANs

- ICR selects critical points where the discriminator learns only *slowly*

- These correspond to realistic images

- Thus, ICR allows GANs to use *inductive biases* of NN for generative modeling

Caltech

Strengthening ICR

- GANs suffer from training instabilities
- Could stronger ICR stabilize training?
- A game-theoretic outlook will help

A game-theoretic view on ICR

$$\min_x \max_y x^2 + 10xy + y^2$$

- Start in $(x, y) = (0,1)$
- $\nabla_y = 2$, y wants to increase to $1 + \delta > 1$
- $\nabla_x = 10$, x wants to decrease to $-\epsilon < 0$
- y incurs additional loss of $10\epsilon\delta$!
- y vulnerable to *counterattack*!

Caltech

A game-theoretic view on ICR

$$\min_x \max_y x^2 + 10xy + y^2$$

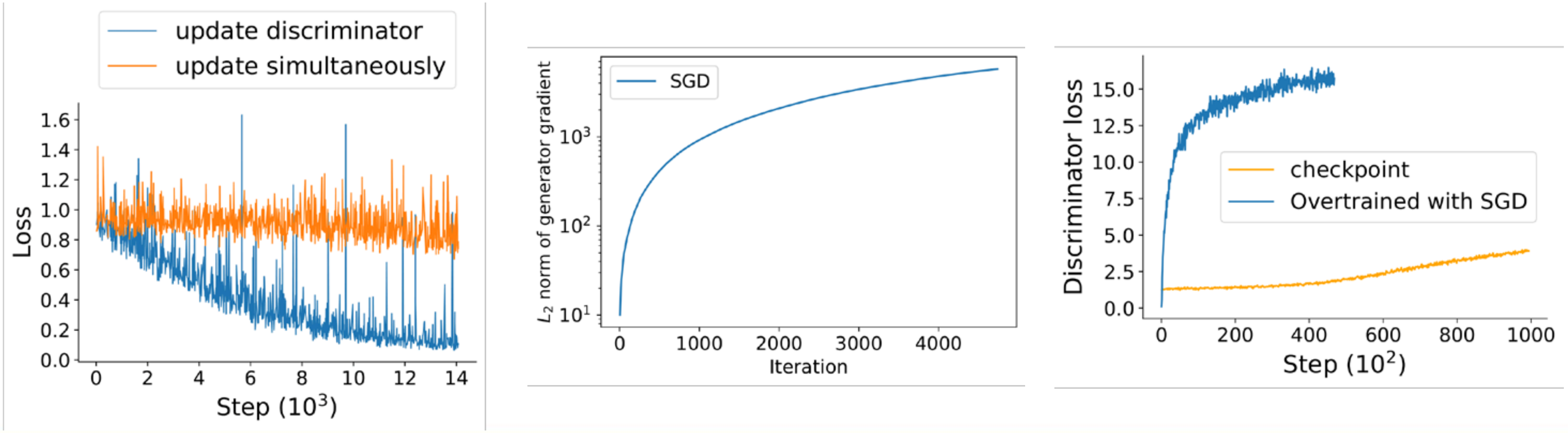
- y feels effect of x's action only *afterwards*

- This can lead to divergence in SimGD

Caltech

A game-theoretic view on ICR

- Effect present in real GANs:
- (Near) perfect discriminator vulnerable to counterattack of the generator



Caltech

Competitive gradient descent

- *Competitive gradient descent (CGD)*, solves bilinear approximation at each step

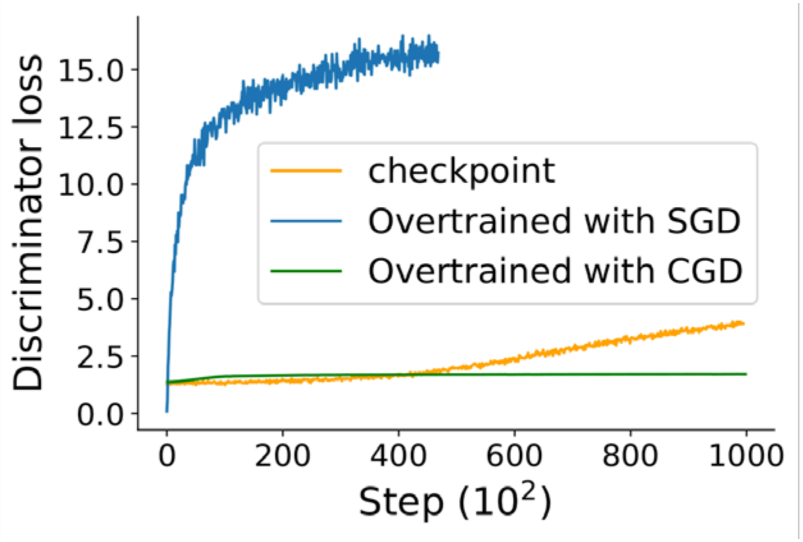
$$x_{k+1} = x_k + \mathrm{argmin}_x \, [D_x f]x + y^T\left[D_{yx}^2 f\right]x + \left[D_y f\right]y + \frac{||x||^2}{2\eta_x}$$

$$y_{k+1} = y_k + \mathrm{argmin}_y \, \left[\mathrm{D}_y \mathrm{g}\right]\mathrm{y} + x^T\left[D_{xy}^2 g\right]y + [D_x f]x + \frac{||y||^2}{2\eta_y}$$

- Leads to improved convergence behavior and stronger ICR.

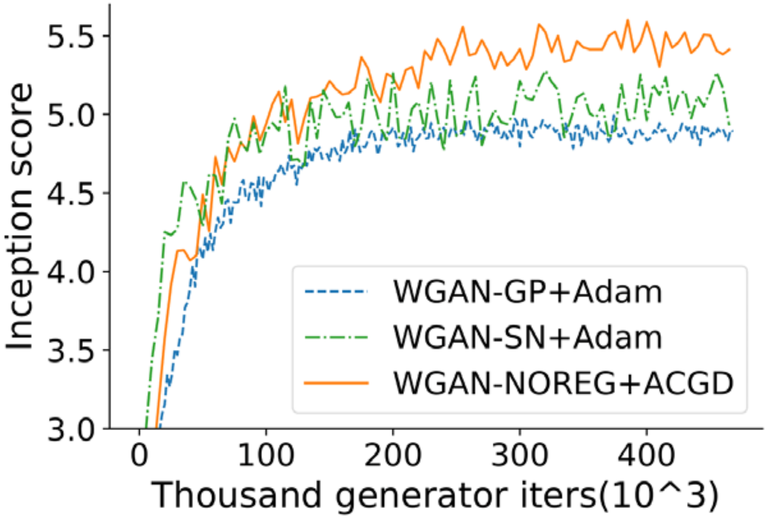
Competitive gradient descent

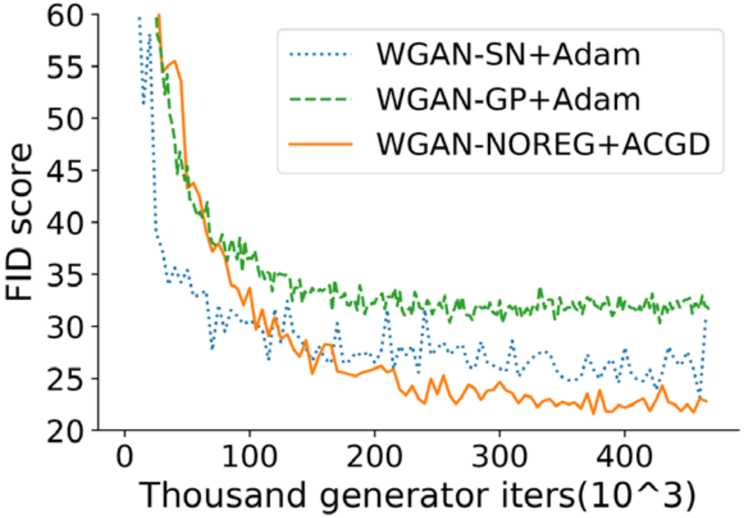
- Avoids fragility of overtrained discriminator



Competitive gradient descent

- How about more complicated images?
- Use WGAN-GP model, remove gradient penalty, train with (adaptive) CGD.





Caltech

Conclusion

- GAN-dilemma foils past explanations of GANs
- ICR explains GAN performance based on *dynamics* of *simultaneous* training
- Motivates to use CGD for stronger ICR
- Improved IS and FID using CGD on CIFAR 10

Caltech