Towards a Better Understanding and Regularization of GAN Training Dynamics

Weili Nie and Ankit B. Patel
Dept. of Electrical and Computer Engineering, Rice University
Dept. of Neuroscience, Baylor College of Medicine

Background

- GANs can be formulated as a *minimax game*

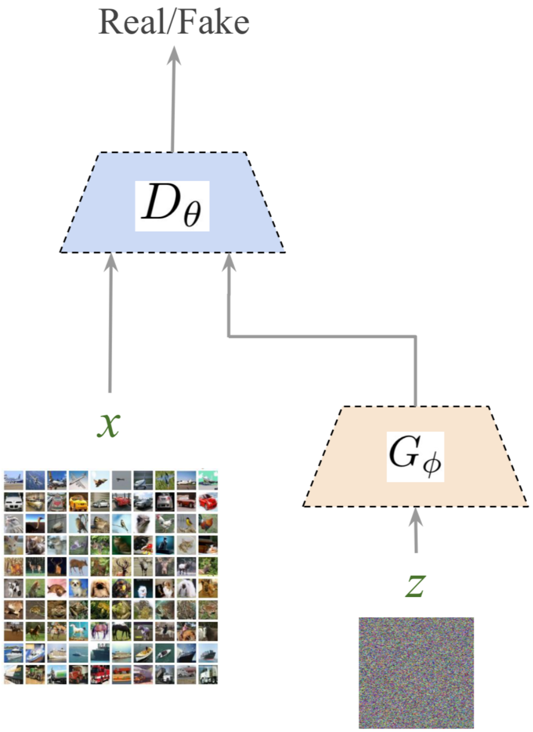
$$\min_{\phi} \max_{\theta} \; f(\phi, \theta)$$

$$f(\phi, \theta) \triangleq \mathbb{E}_{x \sim P_r}[g_1(D_\theta(x))] + \mathbb{E}_{z \sim P_0}[g_2(D_\theta(G_\phi(z)))]$$

- Updates via simultaneous gradient descent (*SimGD*)

$$\boxed{\begin{aligned} \phi^{(k+1)} &= \phi^{(k)} - \eta \nabla_\phi f(\phi^{(k)}, \theta^{(k)}) \\ \theta^{(k+1)} &= \theta^{(k)} + \eta \nabla_\theta f(\phi^{(k)}, \theta^{(k)}) \end{aligned}}$$

- Still an open question to understand *training dynamics* of GANs
  - Global convergence analysis in general impossible without *convex-concave* assumption [Nowozin et al., 2016; Yadav et al., 2018; Gidel et al., 2019]
  - Necessary to analyze *local convergence* near equilibrium [Nagarajan and Kolter, 2017; Mescheder et al., 2018; Liang and Stokes, 2019]



GAN framework

A Simple GAN Example

- *A linear GAN*: Transform latent Gaussian $z \sim \mathcal{N}(0, \sigma^2 I)$ to real Gaussian $x \sim \mathcal{N}(v, \sigma^2 I)$

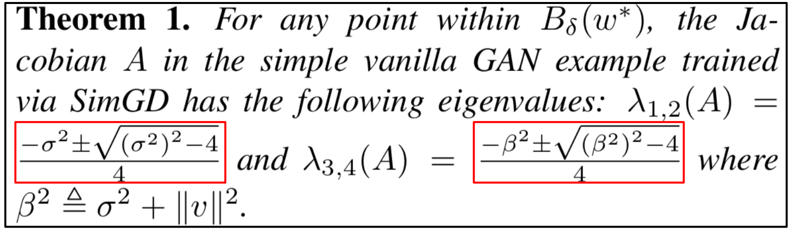
$$f(\phi, \theta) = \mathbb{E}_{x \sim \mathcal{N}(v, \sigma^2 I)}[g_1(\theta^T x)] + \mathbb{E}_{z \sim \mathcal{N}(0, \sigma^2 I)}[g_2(\theta^T(\phi + z))]$$

- With *parametrization* $w \triangleq (\phi - v, \theta)$, the equilibrium point is $w^* = 0$
- By definition, the Jacobian of *SimGD* in the example:

$$A(w^{(k)}) = \begin{bmatrix} -\nabla^2_{\phi\phi} f(w^{(k)}) & -\nabla^2_{\phi\theta} f(w^{(k)}) \\ \nabla^2_{\theta\phi} f(w^{(k)}) & \nabla^2_{\theta\theta} f(w^{(k)}) \end{bmatrix}$$

Not symmetric!

- Eigenvalues of the Jacobian:



Theorem 1. *For any point within $B_\delta(w^*)$, the Jacobian A in the simple vanilla GAN example trained via SimGD has the following eigenvalues:* $\lambda_{1,2}(A) = \boxed{\frac{-\sigma^2 \pm \sqrt{(\sigma^2)^2 - 4}}{4}}$ *and* $\lambda_{3,4}(A) = \boxed{\frac{-\beta^2 \pm \sqrt{(\beta^2)^2 - 4}}{4}}$ *where* $\beta^2 \triangleq \sigma^2 + \|v\|^2$.

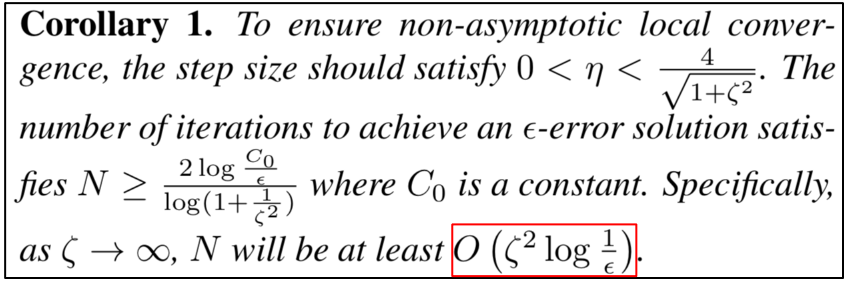
Properties of the eigenvalues depend on data distribution parameters: (v, σ)

A Simple GAN Example (cont.)

- The local convergence behavior

Our analysis is divided into *two cases*: Complex or real eigenvalues

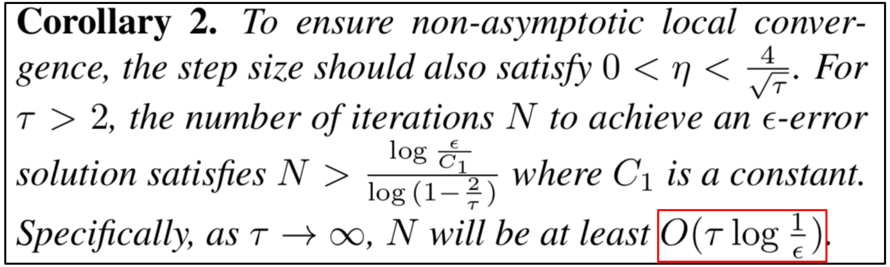
Complex eigenvalues



Corollary 1. *To ensure non-asymptotic local convergence, the step size should satisfy* $0 < \eta < \frac{4}{\sqrt{1+\zeta^2}}$. *The number of iterations to achieve an* ϵ*-error solution satisfies* $N \geq \frac{2 \log \frac{C_0}{\epsilon}}{\log(1+\frac{1}{\zeta^2})}$ *where* C_0 *is a constant. Specifically, as* $\zeta \to \infty$, N *will be at least* $\boxed{O\left(\zeta^2 \log \frac{1}{\epsilon}\right)}$.

Training steps for convergence increases quadratically with the imaginary-to-real ratio (ζ)

Real eigenvalues



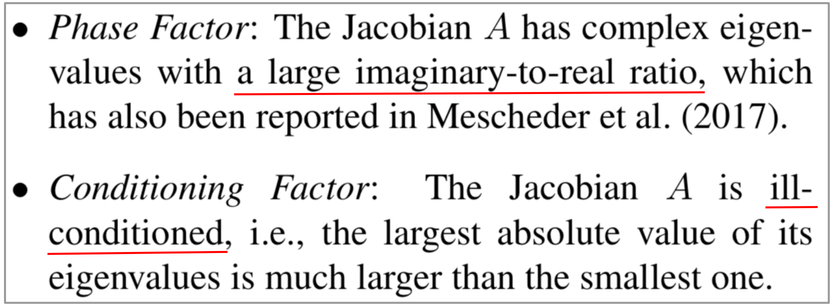
Corollary 2. *To ensure non-asymptotic local convergence, the step size should also satisfy* $0 < \eta < \frac{4}{\sqrt{\tau}}$. *For* $\tau > 2$, *the number of iterations* N *to achieve an* ϵ*-error solution satisfies* $N > \frac{\log \frac{\epsilon}{C_1}}{\log\left(1-\frac{2}{\tau}\right)}$ *where* C_1 *is a constant. Specifically, as* $\tau \to \infty$, N *will be at least* $\boxed{O(\tau \log \frac{1}{\epsilon})}$.

Training steps for convergence increases linearly with the condition number (τ)

A Simple GAN Example (cont.)

- Main results

There may exist the following *two factors* of the Jacobian in GANs *simultaneously* that result in the GAN training issues:



- *Phase Factor*: The Jacobian A has complex eigenvalues with a large imaginary-to-real ratio, which has also been reported in Mescheder et al. (2017).
- *Conditioning Factor*: The Jacobian A is ill-conditioned, i.e., the largest absolute value of its eigenvalues is much larger than the smallest one.

In the simple GAN example, we can show

the data distribution parameters (v, σ) controls the impact of two factors

and both should *not be too small or too large*, a relatively strict requirement for local convergence.

The Proposed Method - JARE

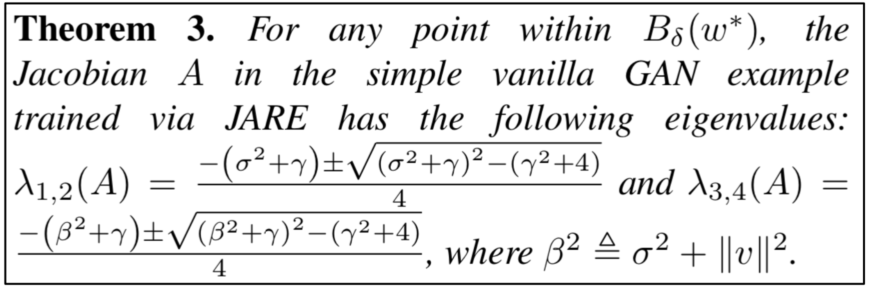
- The G and D updates regularized via JARE are

$$\phi^{(k+1)} = \phi^{(k)} - \eta\nabla_\phi f(w^{(k)}) \boxed{- \frac{1}{2}\eta\gamma\nabla_\phi \left\|\nabla_\theta f(w^{(k)})\right\|^2}$$

$$\theta^{(k+1)} = \theta^{(k)} + \eta\nabla_\theta f(w^{(k)}) \boxed{- \frac{1}{2}\eta\gamma\nabla_\theta \left\|\nabla_\phi f(w^{(k)})\right\|^2}$$

The hyperparameter γ controls the regularization terms

- Eigenvalues of the regularized Jacobian:



Theorem 3. *For any point within $B_\delta(w^*)$, the Jacobian A in the simple vanilla GAN example trained via JARE has the following eigenvalues:* $\lambda_{1,2}(A) = \frac{-(\sigma^2+\gamma)\pm\sqrt{(\sigma^2+\gamma)^2-(\gamma^2+4)}}{4}$ *and* $\lambda_{3,4}(A) = \frac{-(\beta^2+\gamma)\pm\sqrt{(\beta^2+\gamma)^2-(\gamma^2+4)}}{4}$, *where* $\beta^2 \triangleq \sigma^2 + \|v\|^2$.

Properties of the eigenvalues now depend on the hyperparameter γ

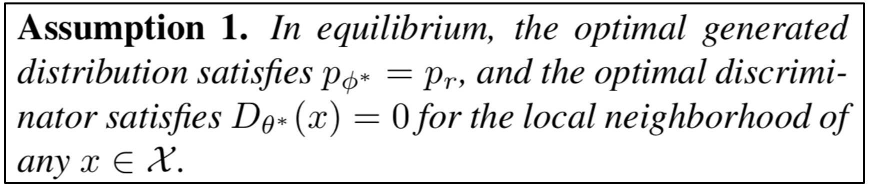
Asymptotically as $\gamma \to \infty$, we get $\boxed{\zeta \to 0}$ and $\boxed{\tau \to 1}$.

No complex eigenvalues

well-conditioned

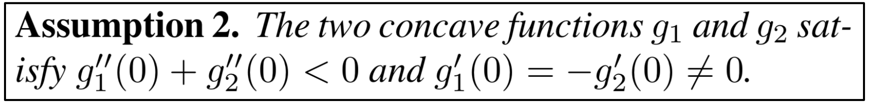
Extensions to General GANs

- Two assumptions [Nagarajan and Kolter, 2017; Mescheder et al., 2018]



Assumption 1. *In equilibrium, the optimal generated distribution satisfies $p_{\phi^*} = p_r$, and the optimal discriminator satisfies $D_{\theta^*}(x) = 0$ for the local neighborhood of any $x \in \mathcal{X}$.*

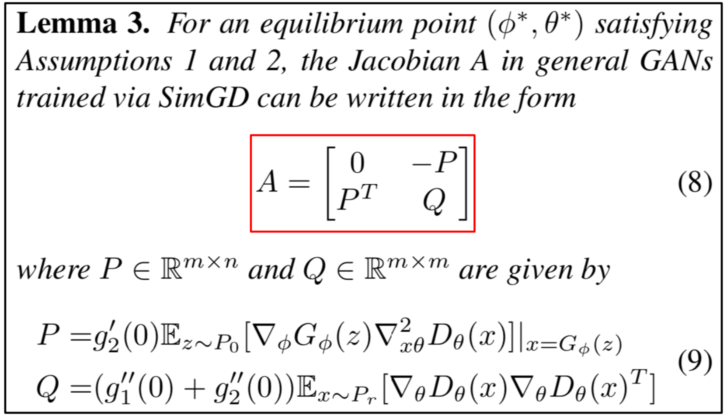
Make sure it is the optimal equilibrium point



Assumption 2. *The two concave functions g_1 and g_2 satisfy $g_1''(0) + g_2''(0) < 0$ and $g_1'(0) = -g_2'(0) \neq 0$.*

Avoid trivial solutions

- Jacobian of general GANs via SimGD



Lemma 3. *For an equilibrium point (ϕ^*, θ^*) satisfying Assumptions 1 and 2, the Jacobian A in general GANs trained via SimGD can be written in the form*

$$\boxed{A = \begin{bmatrix} 0 & -P \\ P^T & Q \end{bmatrix}} \tag{8}$$

where $P \in \mathbb{R}^{m \times n}$ and $Q \in \mathbb{R}^{m \times m}$ are given by

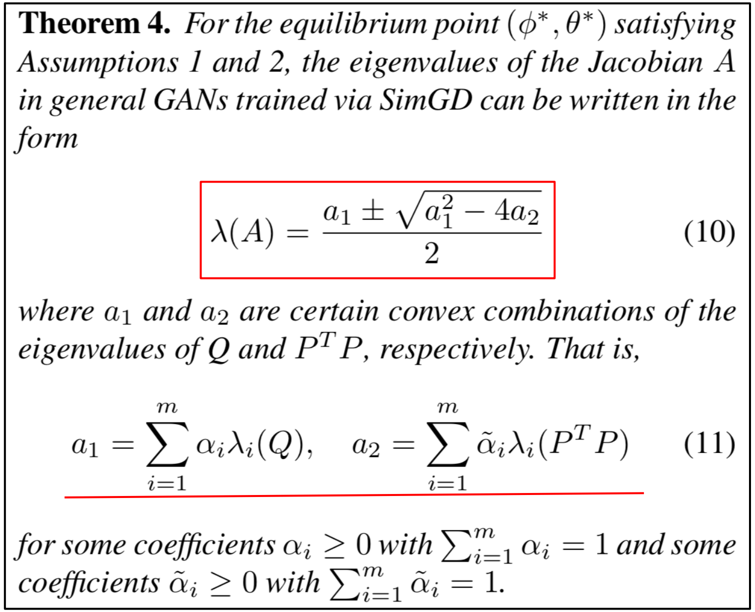
$$\begin{aligned} P &= g_2'(0)\mathbb{E}_{z \sim P_0}[\nabla_\phi G_\phi(z) \nabla^2_{x\theta} D_\theta(x)]|_{x=G_\phi(z)} \\ Q &= (g_1''(0) + g_2''(0))\mathbb{E}_{x \sim P_r}[\nabla_\theta D_\theta(x) \nabla_\theta D_\theta(x)^T] \end{aligned} \tag{9}$$

P represents how sensitive D is to local updates of G

Q represents the local geometry of D (just like Fisher Information)

Extensions to General GANs (cont.)

- Eigenvalues of Jacobian via SimGD



Theorem 4. *For the equilibrium point (ϕ^*, θ^*) satisfying Assumptions 1 and 2, the eigenvalues of the Jacobian A in general GANs trained via SimGD can be written in the form*

$$\boxed{\lambda(A) = \frac{a_1 \pm \sqrt{a_1^2 - 4a_2}}{2}} \qquad (10)$$

where a_1 and a_2 are certain convex combinations of the eigenvalues of Q and $P^T P$, respectively. That is,

$$a_1 = \sum_{i=1}^{m} \alpha_i \lambda_i(Q), \quad a_2 = \sum_{i=1}^{m} \tilde{\alpha}_i \lambda_i(P^T P) \qquad (11)$$

for some coefficients $\alpha_i \geq 0$ with $\sum_{i=1}^{m} \alpha_i = 1$ and some coefficients $\tilde{\alpha}_i \geq 0$ with $\sum_{i=1}^{m} \tilde{\alpha}_i = 1$.

By analysis, we require

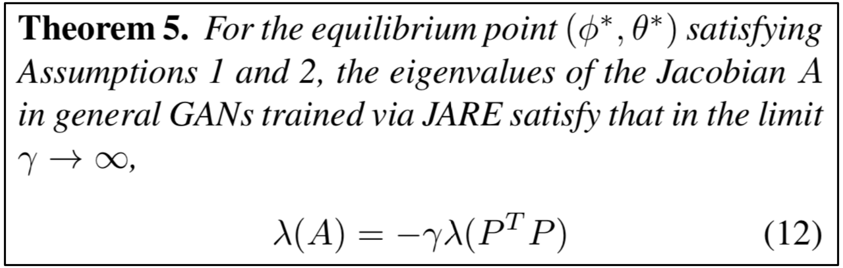
$$Q \text{ and } P^T P$$

- *<u>both are well-conditioned</u>* (which requires good G and D architectures)
- *<u>have similar eigenvalues</u>* (which requires D to well match G)

to avoid the above two factors: ***Phase*** and ***Conditioning*** Factor

Extensions to General GANs (cont.)

- Eigenvalues of Regularized Jacobian via JARE



Theorem 5. *For the equilibrium point (ϕ^*, θ^*) satisfying Assumptions 1 and 2, the eigenvalues of the Jacobian A in general GANs trained via JARE satisfy that in the limit $\gamma \to \infty$,*

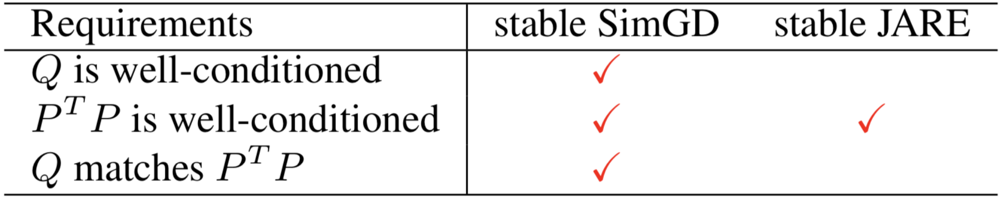
$$\lambda(A) = -\gamma\lambda(P^T P) \qquad (12)$$

The *imbalance* between eigenvalues of

Q and $P^T P$

will *not be an issue* in general GANs via JARE

- In summary, we compare SimGD and JARE in terms of ensuring good (local) training dynamics

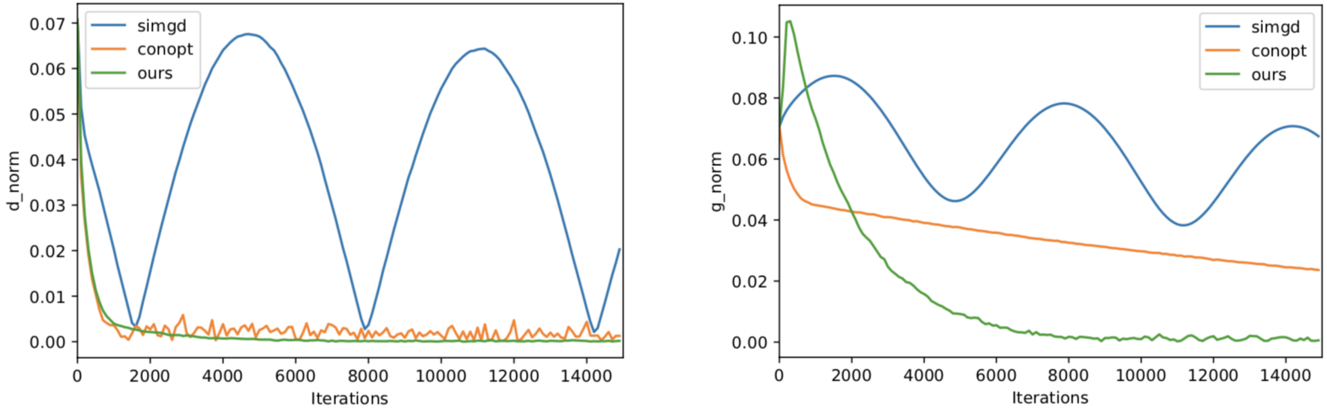


Requirements	stable SimGD	stable JARE
Q is well-conditioned	✓	
$P^T P$ is well-conditioned	✓	✓
Q matches $P^T P$	✓	

JARE could be easier to train, with better robustness than SimGD

Experiments

- Synthetic Data - Isotropic Gaussians



(a) Discriminator training curve

(b) Generator training curve

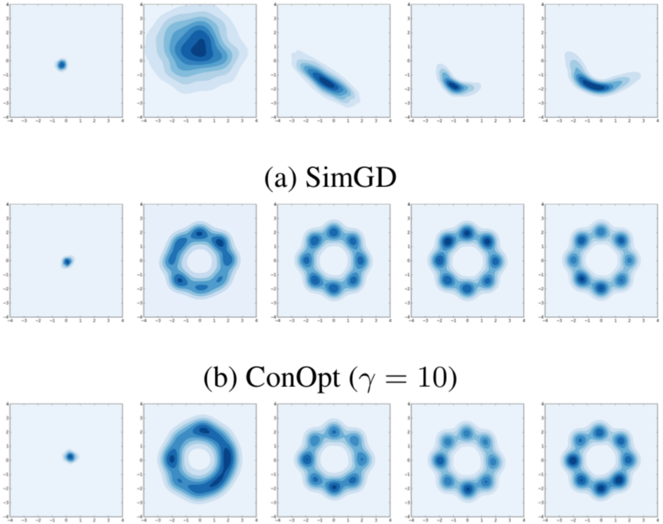
Training dynamics of ***SimGD***, ***ConOpt*** (Mescheder et al., 2017) and ***JARE*** (Ours) in the simple vanilla GAN example

SimGD suffers from both Phase and Conditioning Factor

ConOpt suffers from Conditioning Factor

Experiments (cont.)

- Synthetic Data - Mixture of Gaussians



(a) SimGD

(b) ConOpt ($\gamma = 10$)

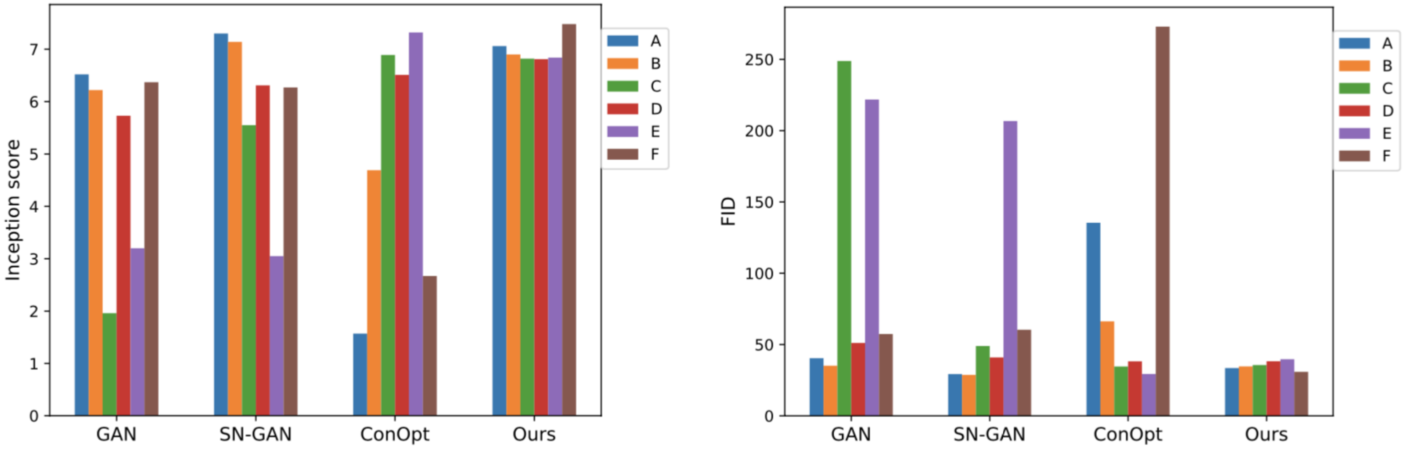
(d) Ours ($\gamma = 10$)

Comparison of *SimGD*, *ConOpt* (Mescheder et al., 2017) and *JARE* (Ours) on the mixture of Gaussians over iterations

Only SimGD oscillates among different modes and fails to converge

Experiments (cont.)

- CIFAR-10 (Inception Score: higher is better, FID: lower is better)

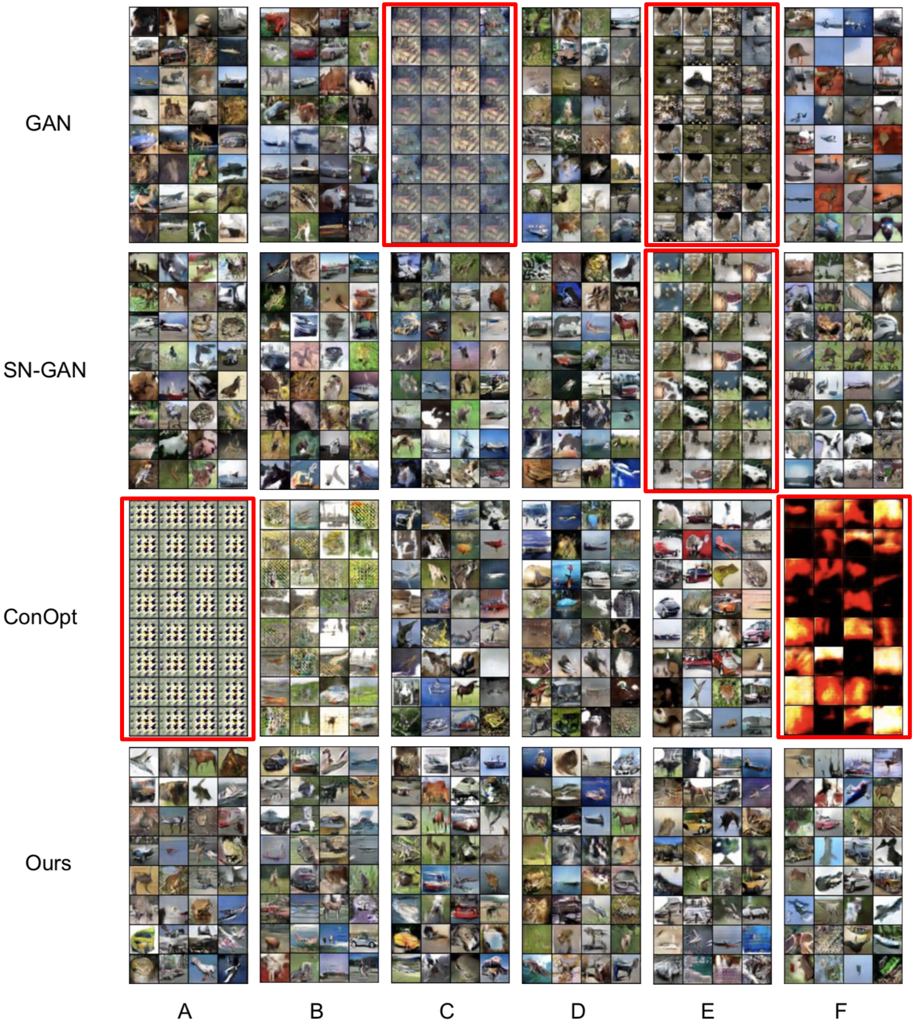


(a) Inception score

(b) FID

Quantitative evaluation of *GAN* (Goodfellow et al., 2014), *ConOpt* (Mescheder et al., 2017), *SN-GAN* (Miyato et al., 2018) and *JARE* (Ours) on CIFAR-10 with different network architectures A-F

JARE is more robust than previous methods across these different settings



Visually, JARE is able to generate good samples across these different settings

Any questions?