GAN Stabilization Under Practical Training Assumptions
Supplementary Material

Joshua DeOliveira[1], Walter Gerych[2], and Elke Rundensteiner[1]

[1]Worcester Polytechnic Institute, Worcester, MA
[2]Massachusetts Institute of Technology, Cambridge, MA

***Abstract*—This document contains supplementary materials that were omitted from the paper "GAN Stabilization Under Practical Training Assumptions"**

I. Extension of Preliminaries

In this section, we will describe previous theoretical contributions by Mescheder et al. (2018) to prove under what assumption can GANs exhibit convergence or at the least stability under a discretized dynamical system.

A. Convergence Theory from Prior Work

As described by Nagarajan & Kolter (2017), the vector field of GAN training at any point (θ, ψ) can be described by the gradient vector field $v(\theta, \psi)$:

$$v(\theta, \psi) = \begin{pmatrix} -\nabla_\theta \mathcal{L}(\theta, \psi) \\ \nabla_\psi \mathcal{L}(\theta, \psi) \end{pmatrix} \tag{1}$$

Given that an update operator F follows the form

$$F(\theta, \psi) = I + \eta V(\theta, \psi)$$

for some arbitrary field V, Mescheder et al. (2017) showed that if the Jacobian of an update operator F has eigenvalues with absolute values greater than 1 at the saddle-point, then GAN training will generally not converge. Additionally, if the Jacobian of an update operator F has eigenvalues with absolute values less than 1 at the saddle-point, then GAN training will converge linearly with a rate of $\mathcal{O}(|\lambda_{\text{MAX}}|^k)$, where λ_{MAX} is the eigenvalue with the greatest magnitude from $F^{'}$. Similarly, when all eigenvalues lie on the unit circle, convergence is at best sub-linear.

Mescheder et al. (2017) notably also showed that it is a necessity, but not necessarily sufficient, that the Jacobian of $v(\theta^*, \psi^*)$ has eigenvalues all with a negative real part for there to be linear convergence when training with either Alt-GDA or Sim-GDA.

Mescheder et al. (2018) then showed for GANs trained with a non-infinitesimal learning rate η via Sim-GDA, that training will converge at best sub-linearly if and only if the Jacobian of the update operator

$$F_3(\theta, \psi) = \begin{pmatrix} \theta - \eta \nabla_\theta \mathcal{L}(\theta, \psi) \\ \psi + \eta \nabla_\psi \mathcal{L}(\theta, \psi) \end{pmatrix}$$

has eigenvalues that all have a negative real part at the saddle-point (θ^*, ψ^*), and η must be within the bound seen in Eq. 2 for all eigenvalues λ.

$$\eta_{\text{SIM}} < \frac{1}{|\text{Re}(\lambda)|} \frac{2}{1 + \left(\frac{\text{Im}(\lambda)}{\text{Re}(\lambda)}\right)^2} \tag{2}$$

Similarly, Mescheder et al. (2018) also showed for GANs trained according to Alt-GDA, that $v(\theta^*, \psi^*)$ must have eigenvalues all with a negative real part, and η must be infinitesimally small to ensure the eigenvalues of the Jacobian of the Alt-GDA update operator lie on the unit circle. In terms of update operators F_1 and F_2, the update operator of Alt-GDA is $F_2 \circ F_1$, where

$$F_1(\theta, \psi) = \begin{pmatrix} \theta - \eta \nabla_\theta \mathcal{L}(\theta, \psi) \\ \psi \end{pmatrix}$$

$$F_2(\theta, \psi) = \begin{pmatrix} \theta \\ \psi + \eta \nabla_\psi \mathcal{L}(\theta, \psi) \end{pmatrix}$$

II. Lyapunov Stability

Lyapunov stability of a dynamic system is a form of stability that ensures a notion of *hovering* around equilibria. While not as strong as asymptotic stability, where every initial condition approaches an equilibrium point as $t \to \infty$, systems with Lyapunov stability prevent initial conditions from diverging infinitely far away from equilibria.

For continuous-time systems, where f is a dynamic system such that $f(\mathbf{x}(t)) = \dot{\mathbf{x}}$, and f has an equilibrium $\mathbf{x}^*$ such that $f(\mathbf{x}^*) = 0$, f is said to be Lyapunov stable, if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that if $||\mathbf{x}(0) - \epsilon|| < \delta$, then for every $t > 0$, $||\mathbf{x}(t) - x^*|| < \delta$. To prove a continuous-time system is Lyapunov stable, a Lyapunov potential function P is used. P can be any potential function such that P is symmetric and positive definite. For a linear system that is defined via the matrix A where $\dot{x} = Ax$, then A is Lyapunov stable if $A^T P + PA$ is negative semi-definite [1] for all $x \neq 0$. For nonlinear systems, one must show that there exists a P such that $\nabla P(x) \cdot f(x)$ is negative semi-definite. Continuous-time systems are effective for modeling when training under an infinitesimally small learning rate. For non-infinitesimal learning rates, we use discrete-time systems where a similar

definition holds, where a linear system that is defined via the matrix A where $x(k+1) = Ax(k)$, then A is Lyapunov stable if $A^T PA - P$ is negative semi-definite for all $x \neq 0$. For nonlinear systems, one must show that there exists a P such that $P(f(x(k+1))) - P(x(k)$ is negative or non-increasing for all k.

III. Proofs and Remarks

ζ^{Dirac} is an instantiation of a switching dynamic system [2], where ζ^{Dirac} is equipped with a set of systems $\mathbf{V} = \{V_1, V_2, V_3\}$, and has a conditional update operator, and consequently optimizes along potentially different gradient vector fields each iteration. When we can define the dynamic system formed by ζ^{Dirac} under an infinitesimal and non-infinitesimal learning rate, each found in Equations 3 and 4 respectively.

$$\zeta^{\text{Dirac}}(\mathbf{x}) = \begin{cases} V_1 \text{ if } \psi\theta < 0 \\ V_2 \text{ if } \psi\theta > 0 \end{cases} \tag{3}$$

$$\zeta^{\text{Dirac}}(\mathbf{x};\eta) = \begin{cases} \mathbf{x} - \eta V_1 \text{ if } \psi\theta < 0 \\ \mathbf{x} + \eta V_2 \text{ if } \psi\theta > 0 \end{cases} \tag{4}$$

For solving for saddle-points, we impose that we want to solve

$$\nabla^2_\theta \mathcal{L}(\theta,\psi) < 0 < \nabla^2_\psi \mathcal{L}(\theta,\psi),$$

$$\nabla_\theta \mathcal{L}(\theta,\psi) = \nabla_\psi \mathcal{L}(\theta,\psi) = 0.$$

So, the equilibrium point lies at $v(0) = 0$ using the gradient vector field in Equation 1.

For a variety of our proofs, we will use the Potential function H in Equation 5,

$$H(\mathbf{x}) = \begin{pmatrix} \frac{1}{2}\theta^2 & 0 \\ 0 & \frac{1}{2}\psi^2 \end{pmatrix} \tag{5}$$

Lemma 1. *Dirac-GANs trained via ζ^{Dirac} are both Lyapunov stable for infinitesimal learning rates when*
$0 \geq -\theta\nabla_\theta \mathcal{L}(\theta,\psi)$ *so long as* $\psi\theta < 0$,
or when $0 \geq \psi\nabla_\psi \mathcal{L}(\theta,\psi)$ *so long as* $\psi\theta > 0$.
Proof.
We utilize a Lyapunov potential function H in Equation 5. Then

$$\nabla H \cdot \zeta^{Dirac} = \begin{cases} \begin{pmatrix} -\theta\mathcal{L}(\theta,\psi) & 0 \\ 0 & 0 \end{pmatrix} : \psi\theta < 0 \\ \begin{pmatrix} 0 & 0 \\ 0 & \psi\mathcal{L}(\theta,\psi) \end{pmatrix} : \psi\theta > 0 \end{cases}$$

Thus, we can see that for a given $\mathcal{L}$, $\nabla H(\theta,\psi) \cdot \zeta^{Dirac}(\theta,\psi)$ is negative semi definite so long as
$0 \geq -\theta\mathcal{L}(\theta,\psi)$ *when* $\psi\theta < 0$
and
$0 \geq \psi\mathcal{L}(\theta,\psi)$ *when* $\psi\theta > 0$
□

Lemma 2. *Dirac-GANs trained via ζ^{Dirac} are both Lyapunov stable for non-infinitesimal learning rates*

$$\eta \leq \frac{\theta^2 - 2\theta}{2\nabla_\theta \mathcal{L}(\theta,\psi)} \text{ when } 0 \geq \frac{-\psi^2}{2} \text{ and } \psi\theta < 0$$

$$\eta \leq \frac{\psi^2 - 2\psi}{2\nabla_\psi \mathcal{L}(\theta,\psi)} \text{ and } 0 \geq \frac{-\theta^2}{2} \text{ when } \psi\theta > 0$$

Proof.
We utilize a Lyapunov potential function H in Equation 5. Then to follow the form $P(f(x(k+1))) - P(x(k) =$, then $H(\zeta^{Dirac}(\theta_{k+1},\psi_{k+1})) - H(\theta,\psi) =$

$$\begin{cases} \begin{pmatrix} \theta - \eta\nabla_\theta \mathcal{L}(\theta,\psi) - \frac{\theta^2}{2} & 0 \\ 0 & \frac{-\psi^2}{2} \end{pmatrix} : \psi\theta < 0 \\ \begin{pmatrix} \frac{-\theta^2}{2} & 0 \\ 0 & \psi + \eta\nabla_\psi \mathcal{L}(\theta,\psi) - \frac{\psi^2}{2} \end{pmatrix} : \psi\theta > 0 \end{cases}$$

Then we see
$0 \geq \theta - \eta\nabla_\theta \mathcal{L}(\theta,\psi) - \frac{\theta^2}{2}$ *and* $0 \geq \frac{-\psi^2}{2}$ *when* $\psi\theta < 0$
$0 \geq \psi + \eta\nabla_\psi \mathcal{L}(\theta,\psi) - \frac{\psi^2}{2}$ *and* $0 \geq \frac{-\theta^2}{2}$ *when* $\psi\theta > 0$
We can rewrite this in terms of η:

$$\psi\theta < 0 \implies \eta \leq \frac{\theta^2 - 2\theta}{2\nabla_\theta \mathcal{L}(\theta,\psi)} \text{ and } 0 \geq \frac{-\psi^2}{2}$$

$$\psi\theta > 0 \implies \eta \leq \frac{\psi^2 - 2\psi}{2\nabla_\psi \mathcal{L}(\theta,\psi)} \text{ and } 0 \geq \frac{-\theta^2}{2}$$

□

Lemma 3. *Dirac-GANs trained with Wasserstein loss via ζ^{Dirac} are Lyapunov stable for an infinitesimal learning rate.*
Proof.
Wasserstein loss is defined by $\mathcal{L}(\theta,\psi) = -\psi\theta$ and it follows that
$\nabla_\theta \mathcal{L}(\theta,\psi) = -\psi$
$\nabla_\psi \mathcal{L}(\theta,\psi) = -\theta$
Thus, when considering that for this $\mathcal{L}$, $\zeta^{Dirac}_{\mathbf{V}}$ requires
$\psi\theta < 0 \implies 0 \geq -\theta\nabla_\theta \mathcal{L}(\theta,\psi)$
$\psi\theta > 0 \implies 0 \geq \psi\nabla_\psi \mathcal{L}(\theta,\psi)$,
To maintain Lyapunov stability. We see that
$\psi\theta < 0 \implies 0 \geq \theta\psi$
$\psi\theta > 0 \implies 0 \geq -\psi\theta$,
Holds trivially. Since Dirac-GANs with ζ^{Dirac} are Lyapunov stable according to Lemma 1, we have demonstrated ζ^{Dirac} will coerce Dirac-GANs trained with Wasserstein loss to be Lyapunov stable. □

Lemma 4. *Dirac-GANs trained with BCE loss via ζ^{Dirac} are Lyapunov stable for an infinitesimal learning rate.*
Proof.
BCE loss is defined for Dirac-GANs

$$\mathcal{L}(\theta, \psi) = \log(\sigma(\psi \cdot 0)) + \log(1 - \sigma(\psi\theta))$$

where σ is the sigmoid activation function with property $\sigma^{'}(x) = \sigma(x)(1 - \sigma(x))x^{'}$

and it follows that

$$\nabla_\theta \mathcal{L}(\theta, \psi) = \frac{\sigma(\psi\theta)(1 - \sigma(\psi\theta))\psi}{1 - \sigma(\psi\theta)}$$

$$\nabla_\psi \mathcal{L}(\theta, \psi) = \frac{\sigma(\psi\theta)(1 - \sigma(\psi\theta))\theta}{1 - \sigma(\psi\theta)}$$

Thus, when considering that for this $\mathcal{L}$, $\zeta_{\mathbf{V}}^{Dirac}$ requires
$\psi\theta < 0 \implies 0 \geq -\theta\nabla_\theta \mathcal{L}(\theta, \psi)$
$\psi\theta > 0 \implies 0 \geq \psi\nabla_\psi \mathcal{L}(\theta, \psi)$,
To maintain Lyapunov stability per Lemma 1. We see that
$\psi\theta < 0 \implies 0 \geq -\theta\psi\sigma(\psi\theta)$
$\psi\theta > 0 \implies 0 \geq \psi\theta\sigma(\psi\theta)$,
Holds trivially since $\sigma(\psi\theta)$ is striclty bounded between the exclusive interval $(0, 1)$. Since Dirac-GANs with ζ^{Dirac} are Lyapunov stable according to Lemma 1, we have demonstrated ζ^{Dirac} will coerce Dirac-GANs trained with BCE loss to be Lyapunov stable. □

Lemma 5. *Dirac-GANs trained with Wasserstein loss via ζ^{Dirac} are Lyapunov stable for a non-infinitesimal learning rate η where*

$$\eta \leq 1 - \frac{\theta}{2} \text{ when } \zeta^{Dirac} \to V_1$$

$$\eta \leq 1 - \frac{\psi}{2} \text{ when } \zeta^{Dirac} \to V_2$$

Proof.

Wasserstein loss is defined by $\mathcal{L}(\theta, \psi) = -\psi\theta$ and it follows that

$$\nabla_\theta \mathcal{L}(\theta, \psi) = -\psi$$

$$\nabla_\psi \mathcal{L}(\theta, \psi) = -\theta$$

Thus, when considering that for this $\mathcal{L}$, $\zeta_{\mathbf{V}}^{Dirac}$ requires η to satisfy the inequalities:

$$\psi\theta < 0 \implies \eta \leq \frac{\theta^2 - 2\theta}{2\nabla_\theta \mathcal{L}(\theta, \psi)} \text{ and } 0 \geq \frac{-\psi^2}{2}$$

$$\psi\theta > 0 \implies \eta \leq \frac{\psi^2 - 2\psi}{2\nabla_\psi \mathcal{L}(\theta, \psi)} \text{ and } 0 \geq \frac{-\theta^2}{2}$$

to maintain Lyapunov stability per Lemma 2. We see that

$$\psi\theta < 0 \implies \eta \leq 1 - \frac{\theta}{2}, 0 \geq \frac{-\psi^2}{2}$$

$$\psi\theta > 0 \implies \eta \leq 1 - \frac{\psi}{2}, 0 \geq \frac{-\theta^2}{2}$$

Holds for relatively large η even when $||(\theta, \psi) - (0, 0)||$ is small. Since Dirac-GANs with ζ^{Dirac} are Lyapunov stable according to Lemma 2, we have demonstrated ζ^{Dirac} will coerce Dirac-GANs trained with Wasserstein loss to be Lyapunov stable for sizeable, non-infinitesimal learning rates. □

Lemma 6. *Dirac-GANs trained with BCE loss via ζ^{Dirac} are Lyapunov stable for a non-infinitesimal learning rate η where*

Proof.

BCE loss is defined for Dirac-GANs

$$\mathcal{L}(\theta, \psi) = \log(\sigma(\psi \cdot 0)) + \log(1 - \sigma(\psi\theta))$$

where σ is the sigmoid activation function with property $\sigma^{'}(x) = \sigma(x)(1 - \sigma(x))x^{'}$

and it follows that

$$\nabla_\theta \mathcal{L}(\theta, \psi) = \frac{\sigma(\psi\theta)(1 - \sigma(\psi\theta))\psi}{1 - \sigma(\psi\theta)}$$

$$\nabla_\psi \mathcal{L}(\theta, \psi) = \frac{\sigma(\psi\theta)(1 - \sigma(\psi\theta))\theta}{1 - \sigma(\psi\theta)}$$

Thus, when considering that for this $\mathcal{L}$, $\zeta_{\mathbf{V}}^{Dirac}$ requires η to satisfy the inequalities:

$$\psi\theta < 0 \implies \eta \leq \frac{\theta^2 - 2\theta}{2\nabla_\theta \mathcal{L}(\theta, \psi)} \text{ and } 0 \geq \frac{-\psi^2}{2}$$

$$\psi\theta > 0 \implies \eta \leq \frac{\psi^2 - 2\psi}{2\nabla_\psi \mathcal{L}(\theta, \psi)} \text{ and } 0 \geq \frac{-\theta^2}{2}$$

to maintain Lyapunov stability per Lemma 2. We see that

$$\psi\theta < 0 \implies \eta \leq \frac{\theta^2 - 2\theta}{2\sigma(\psi\theta)\psi} \text{ and } 0 \geq \frac{-\psi^2}{2}$$

$$\psi\theta > 0 \implies \eta \leq \frac{\psi^2 - 2\psi}{2\sigma(\psi\theta)\theta} \text{ and } 0 \geq \frac{-\theta^2}{2}$$

Holds for a relatively large η even when $||(\theta, \psi) - (0, 0)||$ is large, as the bound on $\sigma(\psi\theta)$ will saturate the denominator when $\psi\theta < 0$. Since Dirac-GANs with ζ^{Dirac} are Lyapunov stable according to Lemma 2, we have demonstrated ζ^{Dirac} will coerce Dirac-GANs trained with BCE loss to be Lyapunov stable. □

Theorem ?? Dirac-GANs trained with Wasserstein or BCE loss via ζ^{Dirac} are both Lyapunov stable for both infinitesimal and non-infinitesimal learning rates, and have a tighter bound on stability around saddle-points than Alt-GDA or Sim-GDA for non-infinitesimal learning rates.

Proof.

We prove ζ^{Dirac} is Lyapunov stable for Wasserstein GANs with infinitesimal learning rates in Lemma 3, and non-infinitesimal learning rates in Lemma 5.

Similarly, we prove ζ^{Dirac} is Lyapunov stable for BCE GANs with infinitesimal learning rates in Lemma 4, and non-infinitesimal learning rates in Lemma 6.

[3] proved that Sim-GDA is not stable near equilibria for Dirac-GANs even when an infinitesimal learning rate is used.

Worse, Sim-GDA diverge will all non-infinitesimal learning rates.

[3] also shows that for Dirac-GANs equipped with Alt-GDA can only achieve non-divergent behavior when

$$\eta \leq \frac{2}{\sqrt{n_g n_d}\nabla\mathcal{L}(\theta,\psi)} \tag{6}$$

where n_g is the number of generator updates after performing n_d discriminator updates. The inequality of η in Equation 6 places tighter bounds on sufficient $\eta's$ than the bound place by ζ^{Dirac}:

$$\eta \leq \frac{\theta^2 - 2\theta}{2\nabla_\theta\mathcal{L}(\theta,\psi)}$$

when updating θ, and

$$\eta \leq \frac{\psi^2 - 2\psi}{2\nabla_\psi\mathcal{L}(\theta,\psi)}$$

when updating ψ. □

Theorem ??. *For a discriminator d, and real/synthetic data $\hat{x} \sim \mathbb{P}_\cdot, x \sim \mathbb{P}_\mathbf{r}$, if $\forall\hat{x}, x : d(\hat{x};\psi) > d(x;\psi)$ then the divergence between $\mathbb{P}_\cdot$ and $\mathbb{P}_\mathbf{r}$ will generally not decrease if θ updates according to gradient descent.*

Proof.

If a given d is a Wasserstein critic, then, the locally optimal critic will asymptotically send $d(\hat{x};\psi)$ towards $-\infty$ and send $d(x;\psi)$ towards ∞ per the definition of Wasserstein loss. Thus, in the asymptotics of optimizing ψ, in the hopes of ultimately approaching a locally optimal discriminator, we expect

$$\mathbb{E}_{\hat{x}\sim\mathbb{P}_\cdot} d(\hat{x};\psi) - \mathbb{E}_{x\sim\mathbb{P}_\mathbf{r}} d(x;\psi)$$

to increase. Thus, if we were then to optimize θ, we would expect to observe θ iteratively maximize the projection of $\mathbb{P}_\cdot$. However, if $\mathbb{E}_{\hat{x}\sim\mathbb{P}_\cdot} d(\hat{x};\psi) > \mathbb{E}_{\hat{x}\sim\mathbb{P}_\cdot} d(\hat{x};\psi)$ then there are erroneous regions in the feature-space d projects to a greater degree outside the support of $\mathbb{P}_\mathbf{r}$. Therefore, if in extreme cases, $\inf\{d(\hat{x};\psi)\} > \sup\{d(x;\psi)\}$, then θ will be optimized such that it travels towards erroneously project regions. This has no guarantee on improving the divergence between $\mathbb{P}_\cdot$ and $\mathbb{P}_\mathbf{r}$, and may even cause detrimental, divergent behavior of $\mathbb{P}_\cdot$.

Furthermore, for bounded discriminators such as the discriminators employed for use with BCE loss, it is well known that a locally optimal discriminator's projection converges towards the real-to-synthetic mass ratio at each point in the feature space. Thus, when $\mathbb{E}_{\hat{x}\sim\mathbb{P}_\cdot} d(\hat{x};\psi) > \mathbb{E}_{x\sim\mathbb{P}_\mathbf{r}} d(x;\psi)$ or in the extreme case, $\inf\{d(\hat{x};\psi)\} > \sup\{d(x;\psi)\}$, then the discriminator's projection of the real-to-synthetic mass ration cannot be faithful to the feature-space, and may cause $\mathbb{P}_\cdot$ to converge to towards erroneous regions with zero density of $\mathbb{P}_\mathbf{r}$. □

Lemma ?? *If the real and synthetic distributions are not disjoint in the feature-space, then the projections of these distributions in the discriminator space will also be not disjoint: $\mathbb{P}_\cdot \cap \mathbb{P}_\mathbf{r} \neq \emptyset \implies d(\mathbb{P}_\cdot;\psi) \cap d(\mathbb{P}_\mathbf{r};\psi) \neq \emptyset$.*

Proof.

We prove this by contraction.

Let's first assume the contradictory statement, where $\mathbb{P}_\cdot$ and $\mathbb{P}_\mathbf{r}$ are non-disjoint in the feature-space, but $d(\mathbb{P}_\cdot;\psi)$ and $d(\mathbb{P}_\mathbf{r};\psi)$ are disjoint. This implies there is a point p that lies in the support of $\mathbb{P}_\mathbf{r}$ and the support of $\mathbb{P}_\cdot$ simultaneously.

Since $d(\mathbb{P}_\cdot;\psi)$ and $d(\mathbb{P}_\mathbf{r};\psi)$ are disjoint, there does not exist a point q, where q lies in the support of $d(\mathbb{P}_\cdot;\psi)$ and $d(\mathbb{P}_\mathbf{r};\psi)$ simultaneously.

However, since d is a continuous operator on the feature-space, infinitesimal perturbations in the feature-space will be infinitesimally small in the image of d.

This is a contradiction since the lack of existence of q violates the continuity assumption of d. Therefore, $d(\mathbb{P}_\cdot;\psi)$ and $d(\mathbb{P}_\mathbf{r};\psi)$ must not be disjoint. □

Lemma ?? *If a discriminator is optimal concerning a fixed generator, and the real and synthetic distributions remain not disjoint in the discriminator's projection-space, then these distributions are not disjoint in the feature-space.*

Proof.

We will prove this by contradiction.

Let's first assume the contradictory statement, where $\mathbb{P}_\cdot$ and $\mathbb{P}_\mathbf{r}$ are non-disjoint in the feature-space.

When we optimize ψ, the extremum that ψ reaches the projection of d such that $d(\mathbb{P}_\cdot;\psi)$ and $d(\mathbb{P}_\mathbf{r};\psi)$ is maximally separated. If $d(\mathbb{P}_\cdot;\psi)$ and $d(\mathbb{P}_\mathbf{r};\psi)$ remain not disjoint at the locally extremum for a fixed θ then there must exist a point p such that p lies in the support of $d(\mathbb{P}_\mathbf{r};\psi)$ and the support of $d(\mathbb{P}_\cdot;\psi)$ simultaneously.

Since d is a continuous operator on the feature-space that has aimed to maximize the distance between $d(\mathbb{P}_\mathbf{r};\psi)$ and $d(\mathbb{P}_\cdot;\psi)$, any infinitesimal perturbations in the feature-space will be infinitesimally small in the image of d.

This is a contradiction since there must exist a point q that exists in the supports of $\mathbb{P}_\cdot$ and $\mathbb{P}_\mathbf{r}$ simultaneously. By assuming $\mathbb{P}_\cdot$ and $\mathbb{P}_\mathbf{r}$ were disjoint, we have violated the continuity assumption of d. Therefore, $\mathbb{P}_\cdot$ and $\mathbb{P}_\mathbf{r}$ must not be disjoint. □

Theorem ??. *If a discriminator is optimal concerning a fixed generator, and that $\mathbb{P}_\mathbf{f}$ and $\mathbb{P}_\mathbf{r}$ are not disjoint in the discriminator's projection-space, and when θ updates according to gradient descent the divergence between $\mathbb{P}_\mathbf{f}$ and $\mathbb{P}_\mathbf{r}$ decreases in the discriminator's projection-space, then the divergence between $\mathbb{P}_\mathbf{f}$ and $\mathbb{P}_\mathbf{r}$ will decrease in the feature-space.*

Proof.

We will prove this by contradiction.

*First consider the local optimality of the discriminator for a fixed generator. Once locally optimal, the discriminator has achieved maximum separation between $d(\mathbb{P}_\mathbf{r};\psi^*)$ and $d(\mathbb{P}_\cdot;\psi^*)$ in the discriminator's projection. If $d(\mathbb{P}_\mathbf{r};\psi^*)$ and $d(\mathbb{P}_\cdot;\psi^*)$ are then disjoint, then we can point to Lemma **??** to show that the distributions must then be disjoint. Here, we instead consider when $d(\mathbb{P}_\mathbf{r};\psi^*)$ and $d(\mathbb{P}_\cdot;\psi^*)$ are still not disjoint. Then there exists a smooth path due to the continuity of d for θ to update.*

If we observe that for a single update iteration f θ according to an infinitesimally smaller learning rate decreased the

divergence between $d(\mathbb{P}_{\mathbf{r}};\psi^*)$ and $d(\mathbb{P}_{\diamond};\psi^*)$, but increase the divergence between $\mathbb{P}_{\mathbf{f}}$ and $\mathbb{P}_{\mathbf{r}}$ in the feature space, then we should observe either less synthetic mass in the support of $\mathbb{P}_{\mathbf{r}}$, or redistribution of mass in the support of $\mathbb{P}_{\mathbf{r}}$.

In the first case, this is a contradiction since θ is optimized to minimize the distance between $d(\mathbb{P}_{\mathbf{r}};\psi^*)$ and $d(\mathbb{P}_{\diamond};\psi^*)$. If the divergence between $\mathbb{P}_{\mathbf{f}}$ and $\mathbb{P}_{\mathbf{r}}$ increased, then we wouldn't observe a decrease in the divergence of the projection. The second case is also a contradiction since a sole mass redistribution of in the feature-space that increases divergence violates the assumption that a single infinitesimal optimization of θ that decreases the separation of $d(\mathbb{P}_{\mathbf{r}};\psi^*)$ and $d(\mathbb{P}_{\diamond};\psi^*)$ □

Theorem ?? *GANs trained via Dynamic-GDA are Lyapunov stable. Also, in conditions where Sim-GDA and Alt-GDA are locally convergent, Dyn-GDA is too.*

Proof. (Sketch)

To prove GANs with Dynamic-GDA are Lyapunov stable, we need to show that for an arbitrary learning rate η, there exist finite upper-bounds on the distance (θ,ψ) will ever be from equilibrium.

To show this, we consider when Dynamic-GDA $c_1 = 1^+$ and $c_2 = 0^+$, then Dynamic-GDA will optimize ψ until a locally optimal discriminator ψ^* is achieved, or $\sup\{d(\mathbb{P}_{\diamond};\psi)\} < \inf\{d(\mathbb{P}_{\mathbf{r}};\psi)\}$ for a bounded discriminator. If the discriminator is unbounded, such as in a WGAN, we halt the optimization of ψ once $\sup\{d(\mathbb{P}_{\diamond};\psi)\} < \inf\{d(\mathbb{P}_{\mathbf{r}};\psi)\}$.

Importantly this logic is to help achieve GAN agnosticism, where Dyn-GDA can begin to optimize ψ until a "faithful" enough projection of the discriminator space is found in order to safely update θ for at least 1 iteration without promoting diverging dynamics, nor needing to know if d is bounded or not.

Furthermore, as [3] points out, when using neural works in the place of g and d, there exists a set of equilibria due to the set of reparameterizations that produce identical functions. Thus, using assumptions $I^{'}, II, III^{'}$ in [3] to consider training according to arbitrarily parameterized forms of neural networks g and d, so long as there exists (θ^*,ψ^*) in the space $\boldsymbol{\theta}\times\boldsymbol{\psi}$, we can show that Dyn-GDA is at least linearly convergent with infinitesimal learning rates locally around equilibria due to using the same set of update fields as Alt-GDA. □

IV. Experimental Setup Details

We conduct all experiments on an RTX 4090 and A100 GPU, and use the PyTorch [4] framework. All experiments used the Adam optimizer for GANs with bounded discriminators, and RMSProp for GANs with unbounded discriminators. For GANs with regulation strategies baked into the architecture of the discriminator, Dynamic-GDA "turns off" these regularizations to decide the optimization to be conducted, and the gradient calculation for the updating uses the discriminator's regularization.

For experiments done pertaining to **celebA**, **CIFAR-10**, and **MNIST**, we train for 30,000 iterations, utilize a learning rate of $2\cdot10^{-4}$, a batch size of 128, and employ a 100-dimensional latent space sampled according to $\mathcal{N}(0,I)$. We normalize all images according to the distribution $\mathcal{N}(0.5,0.5)$ for each channel. For Dynamic-GDA, we use $c_1 = 1.2, c_2 = 0.8$ on all image datasets.

For all $2D$-Datasets, we train for 30,000 iterations, utilize a learning rate of $2\cdot10^{-4}$, a batch size of 128, and employ a 10-dimensional latent space sampled according to $\mathcal{N}(0,I)$ as our $c_1 = 1.1, c_2 = 0.75$ for tuning Dynamic-GDA.

V. Additional Visualizations

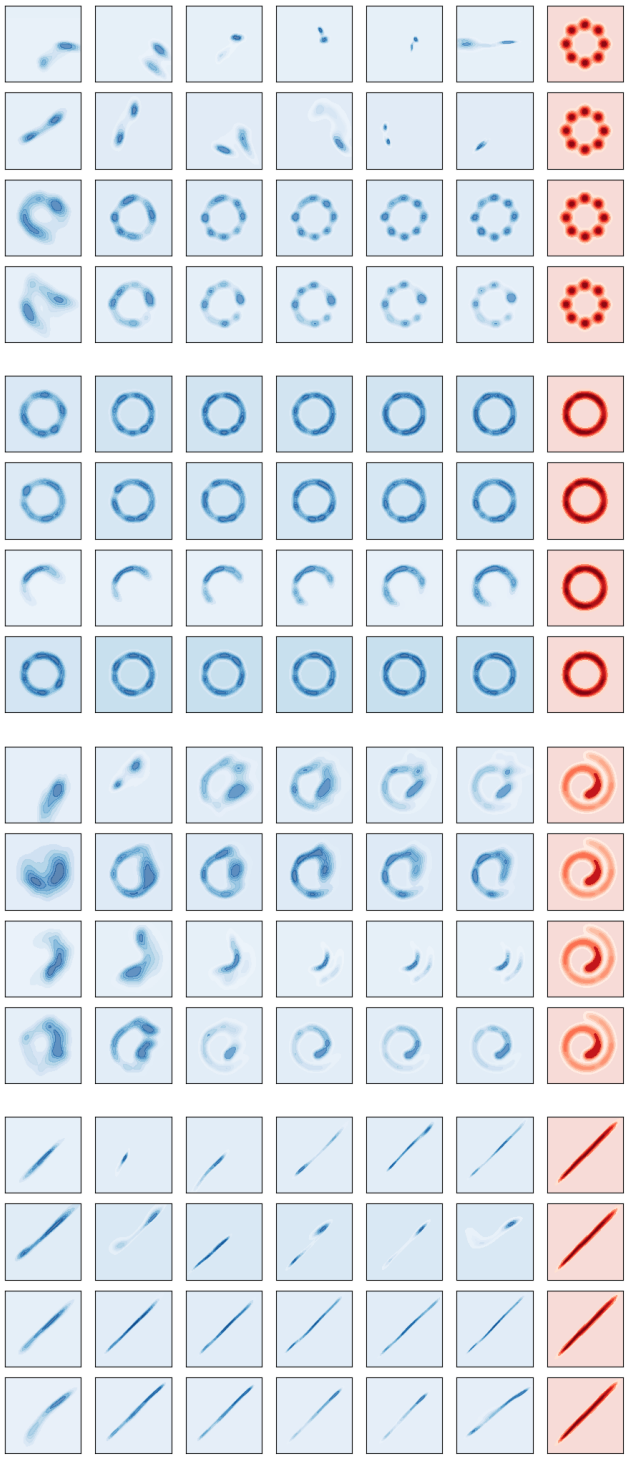

Figure 1. Distributions generated by BCE GANs at (from left to right) 5k, 10k, 15k, 20k, 25k, and 30k iterations when tasked with learning the target distributions (marked red) when equipped with either (from top to bottom) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA. Grouped top-to-bottom are the Gaussian Ring, Circle, Spiral, and Line Segment datasets.

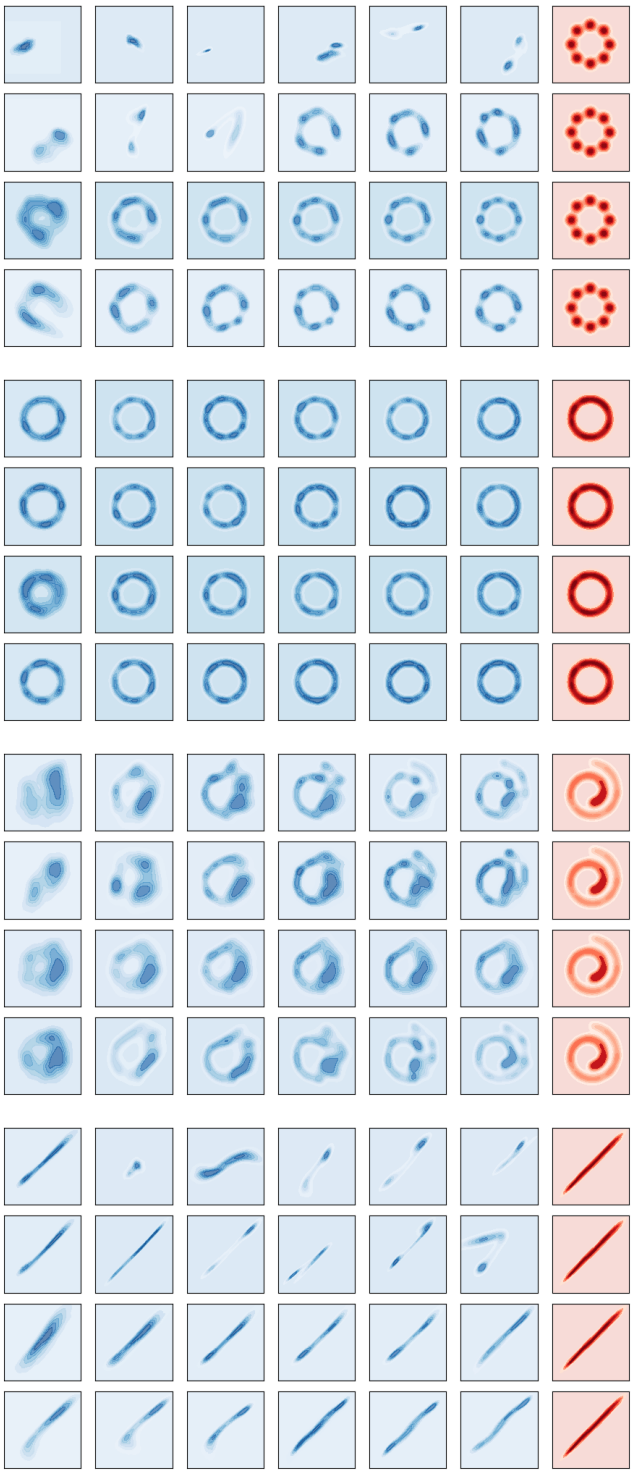

Figure 2. Distributions generated by GANs with JS-Regularization at (from left to right) 5k, 10k, 15k, 20k, 25k, and 30k iterations when tasked with learning the target distributions (marked red) when equipped with either (from top to bottom) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA. Grouped top-to-bottom are the Gaussian Ring, Circle, Spiral, and Line Segment datasets.

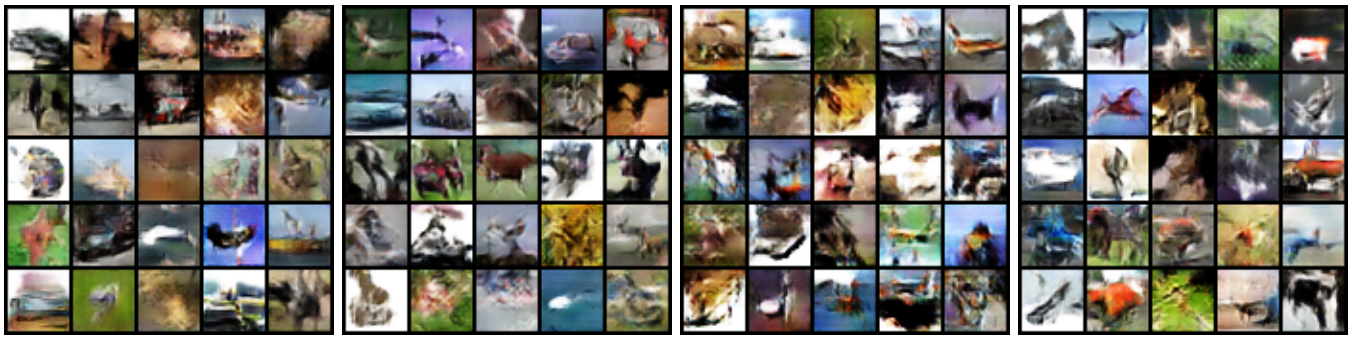

Figure 3. Synthetic samples generated after 30k iterations from Instance Noise GANs equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

Figure 4. Synthetic samples generated after 30k iterations from WGANs equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

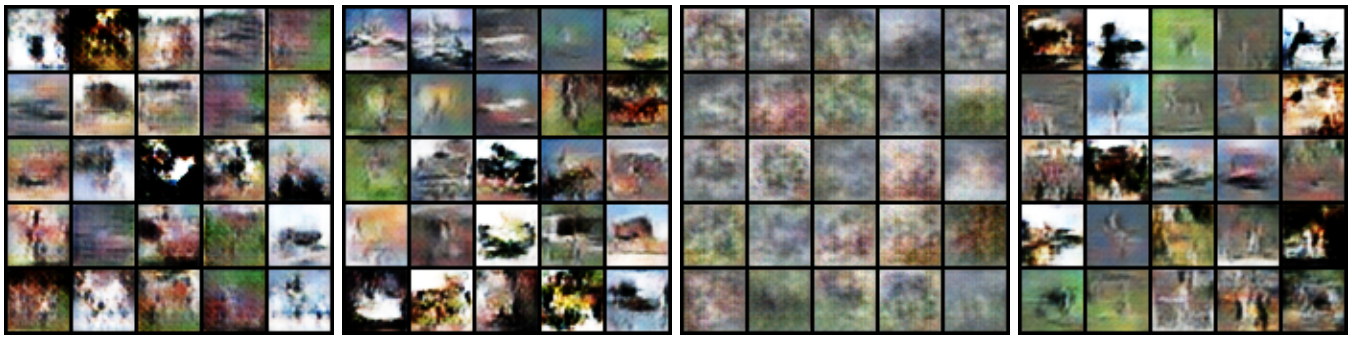

Figure 5. Synthetic samples generated after 30k iterations from GANs with Spectral Norm regularization equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

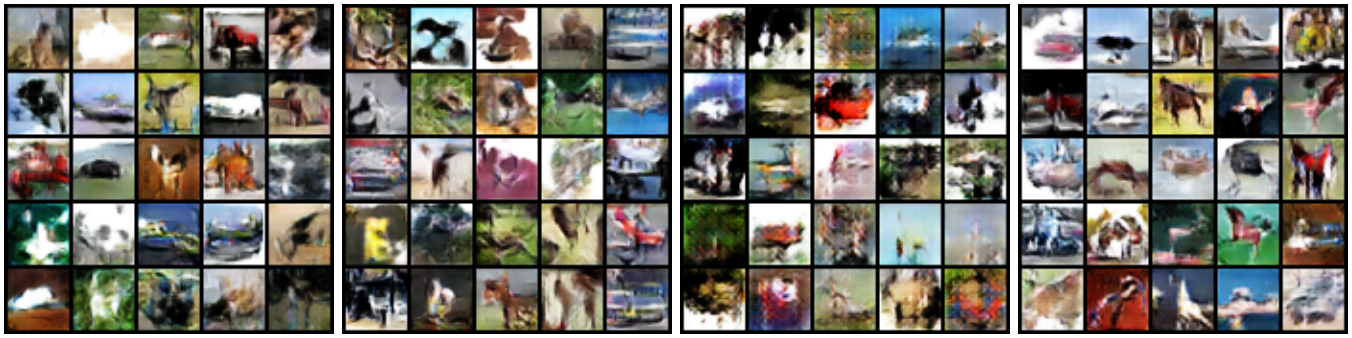

Figure 6. Synthetic samples generated after 30k iterations from GANs with JS-Regularization equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

Figure 7. Synthetic samples generated after 30k iterations from Instance Noise GANs equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

Figure 8. Synthetic samples generated after 30k iterations from WGANs equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

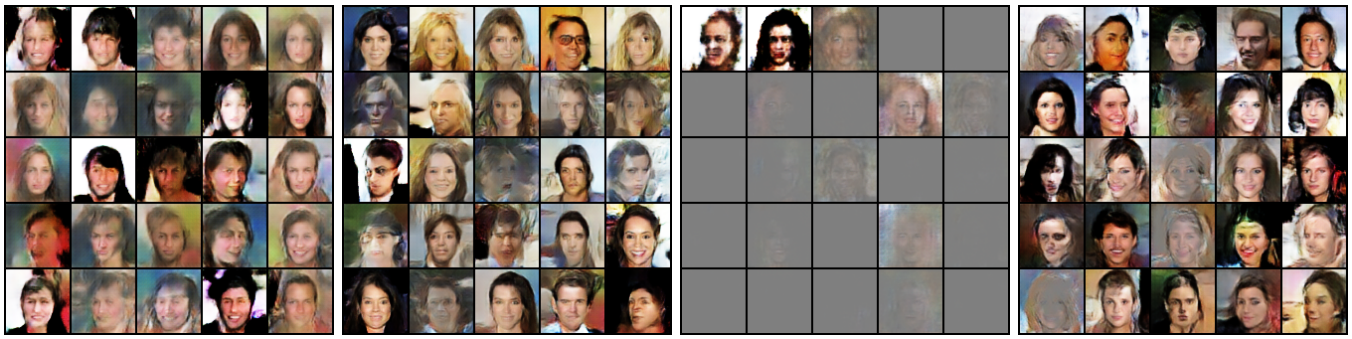

Figure 9. Synthetic samples generated after 30k iterations from GANs with Spectral Norm regularization equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

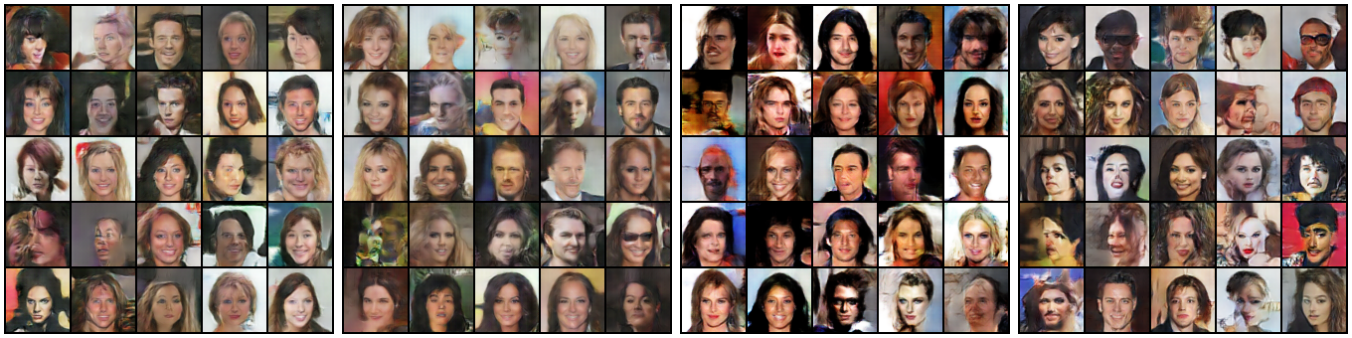

Figure 10. Synthetic samples generated after 30k iterations from GANs with JS-Regularization equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

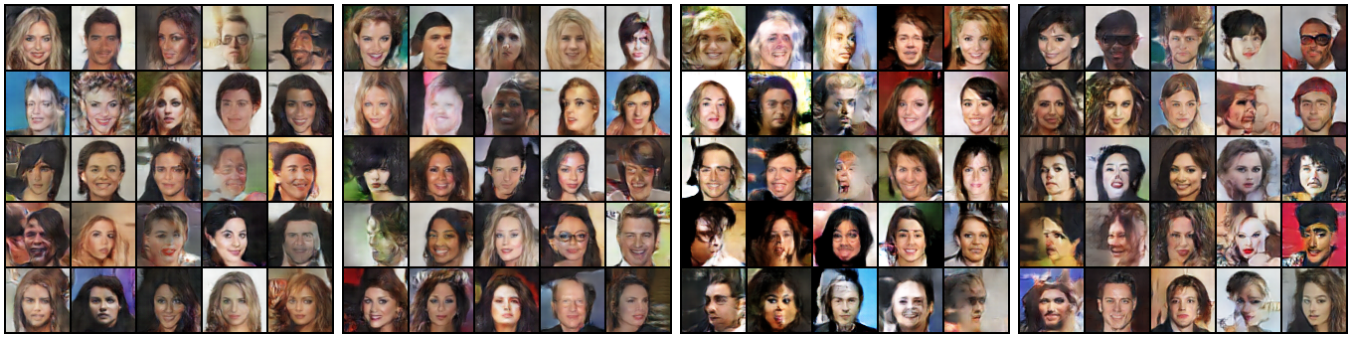

Figure 11. Synthetic samples generated after 30k iterations from GANs with R2-GP equipped with (from left to right) Sim-GDA, Alt-GDA$_{n_d=1}$, Alt-GDA$_{n_d=5}$, or Dyn-GDA during training.

REFERENCES

[1] W. Hahn *et al.*, *Stability of motion*, vol. 138. Springer, 1967.

[2] G. Ackerson and K. Fu, "On state estimation in switching environments," *IEEE transactions on automatic control*, vol. 15, no. 1, pp. 10–17, 1970.

[3] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?," in *International conference on machine learning*, pp. 3481–3490, 2018.

[4] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8024–8035, 2019.