Which Training Methods for GANs do actually Converge? Supplementary material

A. Preliminaries

In this section we first summarize some results from the theory of discrete dynamical systems. We also prove a discrete version of a basic convergence theorem for continuous dynamical systems from Nagarajan & Kolter (2017) which allows us to make statements about training algorithms for GANs for finite learning rates. Afterwards, we summarize some results from Mescheder et al. (2017) about the convergence properties of simultaneous and alternating gradient descent. Moreover, we state some eigenvalue bounds that were derived by Nagarajan & Kolter (2017) which we need to prove Theorem 4.1 on the convergence of the regularized GAN training dynamics.

A.1. Discrete dynamical systems

In this section, we recall some basic definitions from the theory of discrete nonlinear dynamical systems. For a similar description of the theory of continuous nonlinear dynamical systems see for example Khalil (1996) and Nagarajan & Kolter (2017).

In this paper, we consider continuously differentiable operators $F : \Omega \to \Omega$ acting on an open set $\Omega \subset \mathbb{R}^n$. A fixed point of F is a point $\bar{x} \in \Omega$ such that $F(\bar{x}) = \bar{x}$. We are interested in stability and convergence of the fixed point iteration $F^{(k)}(x)$ near the fixed point. To this end, we first have to define what we mean by stability and local convergence:

Definition A.1. *Let $\bar{x} \in \Omega$ be a fixed point of a continuously differentiable operator $F : \Omega \to \Omega$. We call $\bar{x}$*

- stable *if for every $\epsilon > 0$ there is $\delta > 0$ such that $\|x - \bar{x}\| < \delta$ implies $\|F^{(k)}(x) - \bar{x}\| < \epsilon$ for all $k \in \mathbb{N}$.*
- asymptotically stable *if it is stable and there is $\delta > 0$ such that $\|x - \bar{x}\| < \delta$ implies that $F^{(k)}(x)$ converges to $\bar{x}$*
- exponentially stable *if there is $\lambda \in [0, 1)$, $\delta > 0$ and $C > 0$ such that $\|x - \bar{x}\| < \delta$ implies*

$$\|F^{(k)}(x) - \bar{x}\| < C\|x - \bar{x}\|\lambda^k \tag{15}$$

for all $k \in \mathbb{N}$.

If $\bar{x}$ is asymptotically stable fixed point of F, we call the algorithm obtained by iteratively applying F *locally convergent* to $\bar{x}$. If $\bar{x}$ is exponentially stable, we call the corresponding algorithm linearly convergent. Moreover, if $\bar{x}$ is exponentially stable, we call the infimum of all λ so that (15) holds for some $C > 0$ the *convergence rate* of the fixed point iteration.

As it turns out, local convergence of fixed point iterations can be analyzed by examining the spectrum of the Jacobian of the fixed point operator. We have the following central Theorem:

Theorem A.2. *Let $F : \Omega \to \Omega$ be a $\mathcal{C}^1$-mapping on an open subset Ω of $\mathbb{R}^n$ and $\bar{x} \in \Omega$ be a fixed point of F. Assume that the absolute values of the eigenvalues of the Jacobian $F'(\bar{x})$ are all smaller than 1. Then the fixed point iteration $F^{(k)}(x)$ is locally convergent to $\bar{x}$. Moreover, the rate of convergence is at least linear with convergence rate $|\lambda_{max}|$ where λ_{max} denotes the eigenvalue of $F'(\bar{x})$ with the largest absolute value.*

Proof. See Bertsekas (1999), Proposition 4.4.1. □

For the proof of Theorem 4.1 in Section D, we need a generalization of Theorem A.2 that takes into account submanifolds of fixed points. The next theorem is a discrete version of Theorem A.4 from Nagarajan & Kolter (2017) and we prove it in a similar way:

Theorem A.3. *Let $F(\alpha, \gamma)$ define a $\mathcal{C}^1$-mapping that maps some domain Ω to itself. Assume that there is a local neighborhood U of 0 such that $F(0, \gamma) = (0, \gamma)$ for $\gamma \in U$. Moreover, assume that all eigenvalues of $J := \nabla_\alpha F(\alpha, 0) \mid_{\alpha=0}$ have absolute value smaller than 1. Then the fixed point iteration defined by F is locally convergent to $\mathcal{M} := \{(0, \gamma) \mid \gamma \in U\}$ with linear convergence rate in a neighborhood of $(0, 0)$. Moreover, the convergence rate is $|\lambda_{\max}|$ with $\lambda_{\max}$ the eigenvalue of J with largest absolute value.*

Proof. In the following, we write $F(\alpha, \gamma) = (F_1(\alpha, \gamma), F_2(\alpha, \gamma))$, so that the fixed point iteration can be written as

$$\alpha_{k+1} = F_1(\alpha_k, \gamma_k) \quad \gamma_{k+1} = F_2(\alpha_k, \gamma_k). \tag{16}$$

We first examine the behavior of F_1 near $(0, 0)$. To this end, we develop F_1 into a Taylor-Series

$$F_1(\alpha, \gamma) = J\alpha + g_1(\alpha, \gamma) \tag{17}$$

We first show that for any $c > 0$ we have $\|g_1(\alpha,\gamma)\| \leq c\|\alpha\|$ sufficiently close to $(0,0)$: because $F_1(0,\gamma) = 0$ for all γ close to 0, $g_1(\alpha,\gamma)$ must be of the form $g_1(\alpha,\gamma) = h_1(\alpha,\gamma)\alpha$ with $h_1(0,0) = 0$. This shows that for any $c > 0$ there is indeed an open neighborhood V of $(0,0)$ so that $|g_1(\alpha,\gamma)| \leq c\|\alpha\|$ for all $(\alpha,\gamma) \in V$.

According to Bertsekas (1999), Proposition A 15, we can select for every $\epsilon > 0$ a norm $\|\cdot\|_Q$ on $\mathbb{R}^n$ such that

$$\|J\alpha\|_Q < (|\lambda_{\max}| + \epsilon)\|\alpha\|_Q \tag{18}$$

for $\alpha \in \mathbb{R}^n$ where $|\lambda_{\max}|$ denotes the eigenvalue of J with the largest absolute value.

Hence, for $(\alpha,\gamma) \in V$,

$$\begin{aligned}\|F_1(\alpha,\gamma)\|_Q &\leq \|J\alpha\|_Q + \|g_1(\alpha,\gamma)\|_Q \\ &< (|\lambda_{\max}| + \epsilon + c)\|\alpha\|_Q \end{aligned} \tag{19}$$

Because we can make $c + \epsilon$ as small as we want, this shows that $\|\alpha_k\| \leq C\lambda^k\|\alpha_0\|$ for some $C > 0$ and $\lambda \in [0,1)$, if α_0 and all γ_l for $l = 0,\dots,k-1$ are sufficiently close to 0 . We therefore have to show that the iterates γ_k stay in a given local neighborhood of 0, i.e. $\|\gamma_k\| \leq d$ for some $d > 0$, when α_0 and γ_0 are initialized sufficiently close to 0.

To show this, we develop F_2 into a Taylor-series around 0:

$$F_2(\alpha,\gamma) = \gamma + g_2(\alpha,\gamma). \tag{20}$$

Again, we see that g_2 must be of the form $g_2(\alpha,\gamma) = h_2(\alpha,\gamma)\alpha$, showing that $\|g_2(\alpha,\gamma)\| \leq c'\|\alpha\|_Q$ for some fixed constant $c' > 0$ (note that in general $h_2(0,0) \neq 0$). We therefore have

$$\begin{aligned}\|\gamma_k - \gamma_0\| &\leq \sum_{l=0}^{k-1} \|g_2(\alpha_l,\gamma_l)\| \leq \sum_{l=0}^{k-1} c'\|\alpha_l\|_Q \\ &\leq \sum_{l=0}^{k-1} Cc'\lambda^l\|\alpha_0\|_Q \leq \frac{Cc'}{1-\lambda}\|\alpha_0\|_Q \end{aligned} \tag{21}$$

Hence, if we initialize α_0 within $\|\alpha_0\|_Q \leq \frac{1-\lambda}{2CC'}d$ and γ_0 within $\|\gamma_0\| \leq \frac{d}{2}$, we have $\|\gamma_k\| \leq d$ for all $k \in \mathbb{N}$, concluding the proof. □

A.2. Simultaneous and Alternating Gradient Descent

In this section, we recall some results by Mescheder et al. (2017) about the convergence properties of simultaneous and alternating gradient descent as algorithms for training generative adversarial networks.

Recall that simultaneous gradient descent can be described by an update operator of the form

$$F_h(\theta,\psi) = \begin{pmatrix} \theta - h\nabla_\theta L(\theta,\psi) \\ \psi + h\nabla_\psi L(\theta,\psi) \end{pmatrix} \tag{22}$$

where $L(\theta,\psi)$ is the GAN training objective defined in (1).

Similarly, alternating gradient descent can be described by an update operator of the form $F_h = F_{2,h} \circ F_{1,h}$ where $F_{1,h}$ and $F_{2,h}$ are given by

$$F_{1,h}(\theta,\psi) = \begin{pmatrix} \theta - h\nabla_\theta L(\theta,\psi) \\ \psi \end{pmatrix} \tag{23}$$

$$F_{2,h}(\theta,\psi) = \begin{pmatrix} \theta \\ \psi + h\nabla_\psi L(\theta,\psi) \end{pmatrix}. \tag{24}$$

Moreover, we defined the gradient vector field

$$v(\theta,\psi) = \begin{pmatrix} -\nabla_\theta L(\theta,\psi) \\ \nabla_\psi L(\theta,\psi) \end{pmatrix}. \tag{25}$$

To understand convergence of simultaneous and alternating gradient descent, we have to understand when the Jacobian of the corresponding update operator has only eigenvalues with absolute value smaller than 1.

Lemma A.4. *The eigenvalues of the Jacobian of the update operator for simultaneous gradient descent are given by $\lambda = 1 + h\mu$ with μ the eigenvalues of $v'(\theta^*,\psi^*)$. Assume that $v'(\theta^*,\psi^*)$ has only eigenvalues with negative real part. The eigenvalues of the Jacobian of the update operator F_h for simultaneous gradient descent are then all in the unit circle if and only if*

$$h < \frac{1}{|\operatorname{Re}(\lambda)|} \frac{2}{1 + \left(\frac{\operatorname{Im}(\lambda)}{\operatorname{Re}(\lambda)}\right)^2} \tag{26}$$

for all eigenvalues λ of $v'(\theta^*,\psi^*)$.

Proof. For simultaneous gradient descent we have

$$F_h(\theta,\psi) = (\theta,\psi) + hv(\theta,\psi) \tag{27}$$

and hence $F_h'(\theta^*,\psi^*) = I + hv'(\theta^*,\psi^*)$. Therefore the eigenvalues are given by $\lambda = 1 + h\mu$ with μ the eigenvalues of $v'(\theta^*,\psi^*)$.

To see when $|\lambda| < 1$, we write $\mu = -a + ib$ with $a,b \in \mathbb{R}$ and $a > 0$. Then

$$|\lambda|^2 = (1 - ha)^2 + h^2b^2 \tag{28}$$

which is smaller than 1 if and only if

$$h < \frac{2a}{a^2 + b^2}. \tag{29}$$

Dividing both the numerator and denominator by a^2 shows the assertion. □

Lemma A.5. *Assume that $v'(\theta^*,\psi^*)$ has only eigenvalues with negative real part. For $h > 0$ small enough, the eigenvalues of the Jacobian of the update operator F_h for alternating gradient descent are then all in the unit circle.*

Proof. The Jacobian of the update operator $F_h = F_{h,2} \circ F_{h,1}$ at an equilibrium is

$$F_h'(\theta^*, \psi^*) = F_{h,2}'(\theta^*, \psi^*) \cdot F_{h,1}'(\theta^*, \psi^*). \tag{30}$$

However, we have

$$F_{h,i}'(\theta^*, \psi^*) = I + h v_i'(\theta^*, \psi^*) \tag{31}$$

for $i \in \{1, 2\}$ where

$$v_1(\theta, \psi) = \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ 0 \end{pmatrix} \tag{32}$$

$$v_2(\theta, \psi) = \begin{pmatrix} 0 \\ \nabla_\psi L(\theta, \psi) \end{pmatrix} \tag{33}$$

denote the components of the gradient vector field. Hence

$$\begin{aligned} F_h'(\theta^*, \psi^*) &= I + h(v_1'(\theta^*, \psi^*) + v_2'(\theta^*, \psi^*)) \\ &\quad + h^2 v_2'(\theta^*, \psi^*) v_1'(\theta^*, \psi^*) \\ &= I + h(v'(\theta^*, \psi^*) + h\, R(\theta^*, \psi^*)). \end{aligned} \tag{34}$$

with $R(\theta^*, \psi^*) := v_2'(\theta^*, \psi^*) v_1'(\theta^*, \psi^*)$. For $h > 0$ small enough, all eigenvalues of $v'(\theta^*, \psi^*) + h\, R(\theta^*, \psi^*)$ will be arbitrarily close to the eigenvalues of $v'(\theta^*, \psi^*)$. Because all eigenvalues of $v'(\theta^*, \psi^*)$ have negative real-part, all eigenvalues of $F_h'(\theta^*, \psi^*)$ will hence lie inside the unit circle for $h > 0$ small enough. □

In the proof of Theorem 4.1 we will use local coordinates, i.e. a diffeomorphism ϕ that maps a local neighborhood of (θ^*, ψ^*) to an open subset of $\mathbb{R}^{n+m}$. The vector field v and the update operator F then have the following representation in the local coordinates:

$$F_h^\phi(\alpha) := \phi \circ F_h \circ \phi^{-1}(\alpha) \tag{35}$$

$$v^\phi(\alpha) = \phi'(\theta, \psi) \cdot (v \circ \phi^{-1}(\alpha)) \tag{36}$$

While in local coordinates, the simple relationships between $F_h^\phi(\alpha)$ and $v^\phi(\alpha)$ needed to prove Lemma A.4 and Lemma A.5 do not hold anymore, the spectrum can be described in the same way:

Remark A.6. *Assume (θ^*, ψ^*) is a fixed point of F_h and a stationary point of v. Let $\alpha^* = \phi(\theta^*, \psi^*)$. Then*

$$(F_h^\phi)'(\alpha^*) = \phi'(\theta^*, \psi^*) F_h'(\theta^*, \psi^*) \phi'(\theta^*, \psi^*)^{-1} \tag{37}$$

$$(v^\phi)'(\alpha^*) = \phi'(\theta^*, \psi^*) v'(\theta^*, \psi^*) \phi'(\theta^*, \psi^*)^{-1} \tag{38}$$

Hence, $(F_h^\phi)'(\alpha^*)$ and $F_h'(\theta^*, \psi^*)$ have the same spectrum. The same also holds for $(v^\phi)'(\alpha^*)$ and $v'(\theta^*, \psi^*)$.

Proof. This follows from the chain and product rules by using the fact that $F_h(\theta^*, \psi^*) = (\theta^*, \psi^*)$ and $v(\theta^*, \psi^*) = 0$. □

As we will see in the proof of Theorem 4.1, Remark A.6 allows us to apply Theorem A.3 to situations where the stationary points lie on a lower dimensional manifold instead of a space of the form $\{0\}^k \times \mathbb{R}^{n+m-k}$.

A.3. Eigenvalue bounds

When analyzing the convergence properties of GANs, we have to analyze the spectrum of real-valued matrices of the from

$$\begin{pmatrix} 0 & -B^\mathsf{T} \\ B & -Q \end{pmatrix} \tag{39}$$

with Q symmetric positive definite. To this end, we need the following important theorem from Nagarajan & Kolter (2017) which gives explicit bounds on the real part of the eigenvalues:

Theorem A.7. *Assume $J \in \mathbb{R}^{(n+m)\times(n+m)}$ is of the following form:*

$$J = \begin{pmatrix} 0 & -B^\mathsf{T} \\ B & -Q \end{pmatrix} \tag{40}$$

where $Q \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix and $B \in \mathbb{R}^{m\times n}$ has full column rank. Then all eigenvalues λ of J satisfy $\mathrm{Re}(\lambda) < 0$. More precisely

- *if $\mathrm{Im}(\lambda) = 0$*

$$\mathrm{Re}(\lambda) \leq -\frac{\lambda_{\min}(Q)\lambda_{\min}(B^\mathsf{T}B)}{\lambda_{\max}(Q)\lambda_{\min}(Q) + \lambda_{\min}(B^\mathsf{T}B)} \tag{41}$$

- *if $\mathrm{Im}(\lambda) \neq 0$*

$$\mathrm{Re}(\lambda) \leq -\frac{\lambda_{\min}(Q)}{2} \tag{42}$$

Proof. See Nagarajan & Kolter (2017), Lemma G.2. □

In Section E.1, we need a generalization of Theorem A.7. Using almost exactly the same proof as for Theorem A.7, we obtain

Theorem A.8. *Assume $J \in \mathbb{R}^{(n+m)\times(n+m)}$ is of the following form:*

$$J = \begin{pmatrix} -P & -B^\mathsf{T} \\ B & -Q \end{pmatrix} \tag{43}$$

where $P \in \mathbb{R}^{n\times n}$ is a symmetric positive semi-definite matrix, $Q \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix and $B \in \mathbb{R}^{m\times n}$ has full column rank. Then all eigenvalues λ of J satisfy $\mathrm{Re}(\lambda) < 0$.

Proof. Let $v^\mathsf{T} = (a^\mathsf{T}, b^\mathsf{T})$ denote some eigenvector of J with corresponding eigenvalues $\lambda = \lambda_r + i\lambda_i$, where $\lambda_r, \lambda_i \in \mathbb{R}$. Then

$$\lambda_r = \frac{1}{2}\bar{v}^\mathsf{T}(J + J^\mathsf{T})v = -\bar{a}^\mathsf{T}Pa - \bar{b}^\mathsf{T}Qb. \tag{44}$$

Because both P and Q are positive semi-definite, we have $\lambda_r \leq 0$. Because Q is positive definite, it suffices to show that $b \neq 0$ to prove $\lambda_r < 0$.

Assume that $b = 0$. Because v is an eigenvector of J, we have $Ba - Qb = \lambda b$ and therefore $Ba = 0$. Because B has full-column rank, this shows $a = 0$ and hence $v = 0$. However, this contradicts the fact that v is an eigenvector of J. All in all, this show that $b \neq 0$ and thus $\lambda_r \leq -\bar{b}^\mathsf{T} Q b < 0$ as required. □

For applying Theorems A.2, we have to show that the Jacobian of the update operator F_h only has eigenvalues with absolute value smaller than 1. For simultaneous and alternating gradient descent this can be achieved (Lemma A.4 and A.5), if the Jacobian of the gradient vector field v only has eigenvalues with negative real-part. While this condition suffices to prove convergence for small learning rates, Mescheder et al. (2017) showed that simultaneous and alternating gradient descent might still require intractably small learning rates if the imaginary part of the eigenvalues is large. However, in our case we have the following simple bound on the imaginary part of the eigenvalues:

Lemma A.9. *Let*

$$J = \begin{pmatrix} -P & -B^\mathsf{T} \\ B & -Q \end{pmatrix} \tag{45}$$

where $P \in \mathbb{R}^{n \times n}$ and $Q \in \mathbb{R}^{m \times m}$ are symmetric. All eigenvalues λ of J then satisfy

$$|\operatorname{Im}(\lambda)| \leq \sqrt{\lambda_{\max}(B^T B)}. \tag{46}$$

Note that this bound is independent from P and Q.

Proof. Assume v, $\|v\| = 1$, is an eigenvector of J with eigenvalue λ. Then

$$\operatorname{Im}(\lambda) = \bar{v}^\mathsf{T} J_a v. \tag{47}$$

with $J_a := \frac{1}{2i}(J - J^\mathsf{T})$. Hence, by the Cauchy-Schwarz inequality

$$|\operatorname{Im}(\lambda)| \leq \|v\| \|J_a v\| = \|J_a v\|. \tag{48}$$

But, if $v^\mathsf{T} = (a^\mathsf{T}, b^\mathsf{T})$,

$$\|J_a v\|^2 = bBB^\mathsf{T} b + aB^\mathsf{T} Ba \leq \lambda_{\max}(B^\mathsf{T} B). \tag{49}$$

This shows

$$|\operatorname{Im}(\lambda)| \leq \sqrt{\lambda_{\max}(B^T B)}. \tag{50}$$

□

B. Proofs for the Dirac-GAN

This section contains the proofs for our results from Section 2 and Section 3 on the properties of the Dirac-GAN.

Lemma 2.2. *The unique equilibrium point of the training objective in (4) is given by $\theta = \psi = 0$. Moreover, the Jacobian of the gradient vector field at the equilibrium point has the two eigenvalues $\pm f'(0)\, i$ which are both on the imaginary axis.*

Proof. The loss in (4) can be rewritten as

$$L(\theta, \psi) = f(\theta\psi) + const \tag{51}$$

It is easy to check that the gradient vector field is given by

$$v(\theta, \psi) = \begin{pmatrix} -f'(\theta\psi)\psi \\ f'(\theta\psi)\theta \end{pmatrix}. \tag{52}$$

Because $L(\theta, 0) = L(0, \psi) = const$ for all $\theta, \psi \in \mathbb{R}$, $(\theta, \psi) = (0, 0)$ is indeed a Nash-equilibrium for the game defined by (51). Because we assume $f'(t) \neq 0$ for all $t \in \mathbb{R}$, we have $v(\theta, \psi) = 0$ if and only if $(\theta, \psi) = (0, 0)$, showing that $(0, 0)$ is indeed the unique Nash-equilibrium.

Moreover, the Jacobian $v'(\theta, \psi)$ of v is given by

$$\begin{pmatrix} -f''(\theta\psi)\psi^2 & -f'(\theta\psi) - f''(\theta\psi)\theta\psi \\ f'(\theta\psi) + f''(\theta\psi)\theta\psi & f''(\theta\psi)\theta^2 \end{pmatrix}. \tag{53}$$

Evaluating it at the Nash equilibrium $\theta = \psi = 0$, we obtain

$$v'(0, 0) = \begin{pmatrix} 0 & -f'(0) \\ f'(0) & 0 \end{pmatrix} \tag{54}$$

which has the eigenvalues $\pm f'(0) i$. □

Lemma 2.3. *The integral curves of the gradient vector field $v(\theta, \psi)$ do not converge to the Nash-equilibrium. More specifically, every integral curve $(\theta(t), \psi(t))$ of the gradient vector field $v(\theta, \psi)$ satisfies $\theta(t)^2 + \psi(t)^2 = const$ for all $t \in [0, \infty)$.*

Proof. Let $R(\theta, \psi) := \frac{1}{2}(\theta^2 + \psi^2)$. Then

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} R(\theta(t), \psi(t)) \\ &\quad = \theta(t) v_1(\theta(t), \psi(t)) + \psi(t) v_2(\theta(t), \psi(t)) = 0, \end{aligned} \tag{55}$$

showing that $R(\theta, \psi)$ is indeed constant for all $t \in [0, \infty)$.

□

Lemma 2.4. *For simultaneous gradient descent, the Jacobian of the update operator $F_h(\theta, \psi)$ has eigenvalues $\lambda_{1/2} = 1 \pm h f'(0) i$ with absolute values $\sqrt{1 + h^2 f'(0)^2}$ at the Nash-equilibrium. Independently of the learning rate, simultaneous gradient descent is therefore not stable near*

the equilibrium. Even stronger, for every initial condition and learning rate $h > 0$, the norm of the iterates (θ_k, ψ_k) obtained by simultaneous gradient descent is monotonically increasing.

Proof. The first part is a direct consequence of Lemma A.4 and Lemma 2.2.

To see the the norms of the iterates (θ_k, ψ_k) is monotonically increasing, we calculate

$$\begin{aligned}\theta_{k+1}^2 + \psi_{k+1}^2 \\ &= (\theta_k - hf'(\theta_k\psi_k)\psi_k)^2 + (\psi_k + hf'(\theta_k\psi_k)\theta_k)^2 \\ &= \theta_k^2 + \psi_k^2 + h^2 f'(\theta_k\psi_k)^2(\theta_k^2 + \psi_k^2) \\ &\geq \theta_k^2 + \psi_k^2. \end{aligned} \tag{56}$$

□

Lemma 2.5. *For alternating gradient descent with n_g generator and n_d discriminator updates, the Jacobian of the update operator $F_h(\theta, \psi)$ has eigenvalues*

$$\lambda_{1/2} = 1 - \frac{\alpha^2}{2} \pm \sqrt{\left(1 - \frac{\alpha^2}{2}\right)^2 - 1}. \tag{5}$$

with $\alpha := \sqrt{n_g n_d} h f'(0)$. For $\alpha \leq 2$, all eigenvalues are hence on the unit circle. Moreover for $\alpha > 2$, there are eigenvalues outside the unit circle.

Proof. The update operators for alternating gradient descent are given by

$$F_1(\theta, \psi) = \begin{pmatrix} \theta - hf'(\theta\psi)\psi \\ \psi \end{pmatrix} \tag{57}$$

$$F_2(\theta, \psi) = \begin{pmatrix} \theta \\ \psi + hf'(\theta\psi)\theta \end{pmatrix}. \tag{58}$$

Hence, the Jacobians of these operators at 0 are given by

$$F_1'(0,0) = \begin{pmatrix} 1 & -hf'(0) \\ 0 & 1 \end{pmatrix} \tag{59}$$

$$F_2'(0,0) = \begin{pmatrix} 1 & 0 \\ hf'(0) & 1 \end{pmatrix}. \tag{60}$$

As a result, the Jacobian of the combined update operator is

$$\begin{aligned}(F_2^{n_d} \circ F_1^{n_g})'(0,0) &= F_2'(0,0)^{n_d} \cdot F_1'(0,0)^{n_g} \\ &= \begin{pmatrix} 1 & -n_g hf'(0) \\ n_d hf'(0) & -n_g n_d h^2 f'(0)^2 + 1 \end{pmatrix}. \end{aligned} \tag{61}$$

An easy calculation shows that the eigenvalues of this matrix are

$$\lambda_{1/2} = 1 - \frac{\alpha^2}{2} \pm \sqrt{\left(1 - \frac{\alpha^2}{2}\right)^2 - 1} \tag{62}$$

with $\alpha = \sqrt{n_g n_d} h f'(0)$ which are on the unit circle if and only if $\alpha \leq 2$.

□

Lemma 3.1. *WGANs trained with simultaneous or alternating gradient descent with a fixed number of discriminator updates per generator update and a fixed learning rate $h > 0$ do generally not converge to the Nash equilibrium for the Dirac-GAN.*

Proof. First, consider simultaneous gradient descent. Assume that the iterates (θ_k, ψ_k) converge towards the equilibrium point $(0, 0)$. Note that $(\theta_{k+1}, \psi_{k+1}) \neq 0$ if $(\theta_k, \psi_k) \neq 0$. We can therefore assume without loss of generality that $(\theta_k, \psi_k) \neq 0$ for all $k \in \mathbb{N}$.

Because $\lim_{k\to\infty} \psi_k = 0$, there exists k_0 such that for all $k \geq k_0$ we have $|\psi_k| < 1$. For $k \geq k_0$ we therefore have

$$\begin{pmatrix} \theta_{k+1} \\ \psi_{k+1} \end{pmatrix} = \begin{pmatrix} 1 & -h \\ h & 1 \end{pmatrix} \begin{pmatrix} \theta_k \\ \psi_k \end{pmatrix}. \tag{63}$$

For $k \geq k_0$, the iterates are therefore given by

$$\begin{pmatrix} \theta_k \\ \psi_k \end{pmatrix} = A^{k-k_0} \begin{pmatrix} \theta_{k_0} \\ \psi_{k_0} \end{pmatrix} \quad \text{with} \quad A = \begin{pmatrix} 1 & -h \\ h & 1 \end{pmatrix}. \tag{64}$$

However, the eigenvalues of A are given by $\lambda_{1/2} = 1 \pm hi$ which both have absolute value $\sqrt{1 + h^2} > 1$. This contradicts the assumption that (θ_k, ψ_k) converges to $(0, 0)$.

A similar argument also hold for alternating gradient descent. In this case, A is given by

$$\begin{pmatrix} 1 & 0 \\ h & 1 \end{pmatrix}^{n_d} \begin{pmatrix} 1 & -h \\ 0 & 1 \end{pmatrix}^{n_g} = \begin{pmatrix} 1 & -hn_g \\ hn_d & 1 - h^2 n_g n_d \end{pmatrix}. \tag{65}$$

The eigenvalues of A as in (65) are given by

$$1 - \frac{h^2 n_g n_h}{2} \pm \sqrt{\left(1 - \frac{h^2 n_g n_h}{2}\right)^2 - 1}. \tag{66}$$

At least one of these eigenvalues has absolute value greater or equal to 1. Note that for almost all initial conditions (θ_0, ψ_0), the the inner product between the eigenvector corresponding to the eigenvalue with modulus bigger than 1 will be nonzero for all $k \in \mathbb{N}$. Since the recursion in (63) is linear, this contradicts the fact that $(\theta_k, \psi_k) \to (0, 0)$, showing that alternating gradient descent generally does not converge to the Nash-equilibrium either. □

Lemma 3.2. *When using Gaussian instance noise with standard deviation σ, the eigenvalues of the Jacobian of the gradient vector field are given by*

$$\lambda_{1/2} = f''(0)\sigma^2 \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}. \tag{6}$$

In particular, all eigenvalues of the Jacobian have negative real-part at the Nash-equilibrium if $f''(0) < 0$ and $\sigma > 0$. Hence, simultaneous and alternating gradient descent are both locally convergent for small enough learning rates.

Proof. When using instance noise, the GAN training objective (1) is given by

$$\mathrm{E}_{\tilde{\theta}\sim\mathcal{N}(\theta,\sigma^2)}\left[f(\tilde{\theta}\psi)\right] + \mathrm{E}_{x\sim\mathcal{N}(0,\sigma^2)}\left[f(-x\psi)\right]. \tag{67}$$

The corresponding gradient vector field is hence given by

$$\tilde{v}(\theta,\psi) = \mathrm{E}_{\tilde{\theta},x}\begin{pmatrix} -\psi f'(\tilde{\theta}\psi) \\ \tilde{\theta} f'(\tilde{\theta}\psi) - x f'(-x\psi)\end{pmatrix}. \tag{68}$$

The Jacobian $\tilde{v}'(\theta,\psi)$ is therefore

$$\mathrm{E}_{\tilde{\theta},x}\begin{pmatrix} -f''(\tilde{\theta}\psi)\psi^2 & -f'(\tilde{\theta}\psi) - f''(\tilde{\theta}\psi)\tilde{\theta}\psi \\ f'(\tilde{\theta}\psi) + f''(\tilde{\theta}\psi)\tilde{\theta}\psi & f''(\tilde{\theta}\psi)\tilde{\theta}^2 + x^2 f(-x\psi)\end{pmatrix} \tag{69}$$

Evaluating it at $\theta = \psi = 0$ yields

$$\tilde{v}'(0,0) = \begin{pmatrix} 0 & -f'(0) \\ f'(0) & 2f''(0)\sigma^2 \end{pmatrix} \tag{70}$$

whose eigenvalues are given by

$$\lambda_{1/2} = f''(0)\sigma^2 \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}. \tag{71}$$

□

Lemma 3.3. *The eigenvalues of the Jacobian of the gradient vector field for the gradient-regularized GAN at the equilibrium point are given by*

$$\lambda_{1/2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - f'(0)^2}. \tag{8}$$

In particular, for $\gamma > 0$ all eigenvalues have negative real part. Hence, simultaneous and alternating gradient descent are both locally convergent for small enough learning rates.

Proof. The regularized gradient vector field becomes

$$\tilde{v}(\theta,\psi) = \begin{pmatrix} -f'(\theta\psi)\psi \\ f'(\theta\psi)\theta - \gamma\psi \end{pmatrix}. \tag{72}$$

The Jacobian $\tilde{v}'(\theta,\psi)$ is therefore given by

$$\begin{pmatrix} -f''(\theta\psi)\psi^2 & -f'(\theta\psi) - f''(\theta\psi)\theta\psi \\ f'(\theta\psi) + f''(\theta\psi)\theta\psi & f''(\theta\psi)\theta^2 - \gamma \end{pmatrix}. \tag{73}$$

Evaluating it at $\theta = \psi = 0$ yields

$$\tilde{v}'(0,0) = \begin{pmatrix} 0 & -f'(0) \\ f'(0) & -\gamma \end{pmatrix} \tag{74}$$

whose eigenvalues are given by

$$\lambda_{1/2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - f'(0)^2}. \tag{75}$$

□

C. Other regularization strategies

In this section we discuss further regularization techniques for GANs on our example problem that were omitted in the main text due to space constraints.

C.1. Nonsaturating GAN

Especially in the beginning of training, the discriminator can reject samples produced by the generator with high confidence (Goodfellow et al., 2014). When this happens, the loss for the generator may saturate so that the generator receives almost no gradient information anymore.

To circumvent this problem Goodfellow et al. (2014) introduced a nonsaturating objective for the generator. In nonsaturating GANs, the generator objective is replaced with[7]

$$\max_{\theta} \mathrm{E}_{p_\theta(x)} f(-D_\psi(x)). \tag{76}$$

In our example, this is $\max_\theta f(-\psi\theta)$.

While the nonsaturating generator objective was originally motivated by global stability considerations, we investigate its effect on local convergence. A linear analysis similar to normal GANs yields

Lemma C.1. *The unique Nash-equilibrium for the nonsaturating GAN on the example problem is given by $\theta = \psi = 0$. The eigenvalues of the Jacobian of the gradient vector field at the equilibrium are $\pm f'(0)i$ which are both on the imaginary axis.*

Proof. The gradient vector field for the nonsaturating GAN is given by

$$v(\theta,\psi) = \begin{pmatrix} -f'(-\theta\psi)\psi \\ f'(\theta\psi)\theta \end{pmatrix}. \tag{77}$$

As in the proof of Lemma 2.2, we see that $(\psi,\theta) = (0,0)$ defines the unique Nash-equilibrium for the nonsaturating GAN.

Moreover, the Jacobian $v'(\theta,\psi)$ is

$$\begin{pmatrix} f''(-\theta\psi)\psi^2 & -f'(-\theta\psi) + f''(-\theta\psi)\theta\psi \\ f'(\theta\psi) + f''(\theta\psi)\theta\psi & f''(\theta\psi)\theta^2 \end{pmatrix}. \tag{78}$$

At $\theta = \psi = 0$ we therefore have

$$v'(0,0) = \begin{pmatrix} 0 & -f'(0) \\ f'(0) & 0 \end{pmatrix}. \tag{79}$$

with eigenvalues $\lambda_{1/2} = \pm f'(0)i$. □

Lemma C.1 implies that simultaneous gradient descent is not locally convergent for a nonsaturating GAN and any

[7] Goodfellow et al. (2014) used $f(t) = -\log(1+\exp(-t))$.

learning rate $h > 0$, because the eigenvalues of the Jacobian of the corresponding update operator F_h all have absolute value larger than 1 (Lemma A.4). While Lemma C.1 also rules out linear convergence towards the Nash-equilibrium in the continuous case (i.e. for $h \to 0$), the continuous training dynamics could in principle still converge with a sublinear convergence rate. Indeed, we find this to be the case for the Dirac-GAN. We have

Lemma C.2. *For every integral curve of the gradient vector field of the nonsaturating Dirac-GAN we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}(\theta(t)^2 + \psi(t)^2) = 2\left[f'(\theta\psi) - f'(-\theta\psi)\right]\theta\psi. \tag{80}$$

For concave f this is nonpositive. Moreover, for $f''(0) < 0$, the continuous training dynamics of the nonsaturating Dirac-GAN converge with logarithmic convergence rate.

Proof. The gradient vector field for the nonsaturating Dirac-GAN is given by

$$v(\theta, \psi) = \begin{pmatrix} -f'(-\theta\psi)\psi \\ f'(\theta\psi)\theta \end{pmatrix}. \tag{81}$$

Hence, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\theta(t)^2 + \psi(t)^2) &= v_1(\theta,\psi)\theta + v_2(\theta,\psi)\psi \\ &= 2\theta\psi\left[f'(\theta\psi) - f'(-\theta\psi)\right]. \end{aligned}\tag{82}$$

For concave f, we have

$$\frac{f'(\theta\psi) - f'(-\theta\psi)}{2\theta\psi} \leq 0 \tag{83}$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}(\theta(t)^2 + \psi(t)^2) \leq 0. \tag{84}$$

Now assume that $f'(0) \neq 0$ and $f''(0) < 0$.

To intuitively understand why the continuous system converges with logarithmic convergence rate, note that near the equilibrium point we asymptotically have in polar coordinates $(\theta, \psi) = (\sqrt{w}\cos(\phi), \sqrt{w}\sin(\phi))$:

$$\dot{\phi} = f'(0) + \mathcal{O}(|w|^{1/2}) \tag{85}$$

$$\begin{aligned}\dot{w} &= 4f''(0)\theta^2\psi^2 + \mathcal{O}(|\theta\psi|^4) &(86)\\ &= f''(0)w^2\sin^2(2\phi) + \mathcal{O}(|w|^4). &(87)\end{aligned}$$

When we ignore higher order terms, we can solve this system explicitly[8] for ϕ and w:

$$\phi(t) = f'(0)(t - t_0) \tag{89}$$

$$w(t) = \frac{2}{-f''(0)t + \frac{f''(0)}{4f'(0)}\sin(4f'(0)(t - t_0)) + c} \tag{90}$$

The training dynamics are hence convergent with logarithmic convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$.

For a more formal proof, first note that w is nonincreasing by the first part of the proof. Moreover, for every $\epsilon > 0$ there is $\delta > 0$ such that for $w < \delta$:

$$f'(0) - \epsilon \leq \dot{\phi} \leq f'(0) + \epsilon \tag{91}$$

$$\dot{w} \leq (f''(0)\sin^2(2\phi) + \epsilon)w^2. \tag{92}$$

This implies that for every time interval $[0, T]$, $\phi(t)$ is in

$$\bigcup_{k\in\mathbb{Z}}\left[\frac{\pi}{8} + k\frac{\pi}{2}, \frac{3\pi}{8} + k\frac{\pi}{2}\right] \tag{93}$$

for t in a union of intervals $Q_T \subseteq [0, T]$ with total length at least $\beta\lfloor\alpha T\rfloor$ with some constants $\alpha, \beta > 0$ which are independent of T.

For these $t \in Q_T$ we have $\sin^2(2\phi(t)) \geq \frac{1}{2}$. Because $f''(0) < 0$, this shows

$$\dot{w}(t) \leq \left(\frac{1}{2}f''(0) + \epsilon\right)w(t)^2 \tag{94}$$

for $t \in Q_T$ and ϵ small enough. Solving the right hand formally yields

$$w(t) \leq \frac{1}{-(\frac{1}{2}f''(0) + \epsilon)t + c}. \tag{95}$$

As $w(t)$ is nonincreasing for $t \notin Q_T$ and the total length of Q_T is at least $\beta\lfloor\alpha T\rfloor$ this shows that

$$w(T) \leq \frac{1}{-(\frac{1}{2}f''(0) + \epsilon)\beta\lfloor\alpha T\rfloor + c}. \tag{96}$$

The training dynamics hence converge with logarithmic convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$. □

Note that the standard choice $f(t) = -\log(1 + \exp(-t))$ is concave and satisfies $f''(0) = -\frac{1}{4} < 0$. Lemma C.1 is hence applicable and shows that the GAN training dynamics for the standard choice of f converge with logarithmic convergence rate in the continuous case. The training behavior of the nonsaturating GAN on our example problem is visualized in Figure 3b.

[8] For solving the ODE we use the *separation of variables*-technique and the identity

$$\int 2\sin^2(ax)\mathrm{d}x = x - \frac{\sin(2ax)}{2a}. \tag{88}$$

C.2. Wasserstein GAN-GP

In practice, it can be hard to enforce the Lipschitz-constraint for WGANs. A practical solution to this problem was given by Gulrajani et al. (2017), who derived a simple gradient penalty with a similar effect as the Lipschitz-constraint. The resulting training objective is commonly referred to as WGAN-GP.

Similarly to WGANs, we find that WGAN-GP does not converge for the Dirac-GAN. A similar analysis also applies to the DRAGAN-regularizer proposed in (Kodali et al., 2017).

The regularizer proposed by Gulrajani et al. (2017) is given by

$$R(\psi) = \frac{\gamma}{2}\mathrm{E}_{\hat{x}}\left(\|\nabla_x D_\psi(\hat{x})\| - g_0\right)^2 \tag{97}$$

where $\hat{x}$ is sampled uniformly on the line segment between two random points $x_1 \sim p_\theta(x_1)$, $x_2 \sim p_\mathcal{D}(x_2)$.

For the Dirac-GAN, it simplifies to

$$R(\psi) = \frac{\gamma}{2}(|\psi| - g_0)^2 \tag{98}$$

The corresponding gradient vector field is given by

$$\tilde{v}(\theta, \psi) = \begin{pmatrix} -\psi \\ \theta - \mathrm{sign}(\psi)\gamma(|\psi| - g_0) \end{pmatrix}. \tag{99}$$

Note that the gradient vector field has a discontinuity at the equilibrium point, as the gradient vector field takes on values with norm bigger than some fixed constant in every neighborhood of the equilibrium point. As a result, we have

Lemma C.3. *WGAN-GP trained with simultaneous or alternating gradient descent with a fixed number of generator and discriminator updates and a fixed learning rate $h > 0$ does not converge locally to the Nash equilibrium for the Dirac-GAN.*

Proof. First, consider simultaneous gradient descent. Assume that the iterates (θ_k, ψ_k) converge towards the equilibrium point $(0, 0)$. For almost all initial conditions[9] we have $(\theta_k, \psi_k) \neq (0, 0)$ for all $k \in \mathbb{N}$. This implies

$$|\psi_{k+1} - \psi_k| = h|\theta_k - \gamma\psi_k - \mathrm{sign}(\psi_k)g_0| \tag{100}$$

and hence $\lim_{k\to\infty} |\psi_{k+1} - \psi_k| = h|g_0| \neq 0$, showing that (θ_k, ψ_k) is not a Cauchy sequence. This contradicts the assumption that (θ_k, ψ_k) converges to the equilibrium point $(0, 0)$.

A similar argument also holds for alternating gradient descent. □

The training behavior of WGAN-GP on our example problem is visualized in Figure 3d.

[9] Depending on γ, h and g_0 modulo a set of measure 0.

As for WGANs, we stress that this analysis only holds if the discriminator is trained with a fixed number of discriminator updates per generator update. Again, more careful training that ensures that the discriminator is kept exactly optimal or two-timescale training (Heusel et al., 2017) might be able to ensure convergence for WGAN-GP.

C.3. Consensus optimization

Consensus optimization (Mescheder et al., 2017) is an algorithm that attempts to solve the problem of eigenvalues with zero real-part by introducing a regularization term that explicitly moves the eigenvalues to the left. The regularization term in consensus optimization is given by

$$\begin{aligned} R(\theta, \psi) &= \frac{\gamma}{2}\|v(\theta, \psi)\|^2 \\ &= \frac{\gamma}{2}(\|\nabla_\theta L(\theta, \psi)\|^2 + \|\nabla_\psi L(\theta, \psi)\|^2). \end{aligned} \tag{101}$$

As was proved by Mescheder et al. (2017), consensus optimization converges locally for small learning rates $h > 0$ provided that the Jacobian $v'(\theta^*, \psi^*)$ is invertible.[10]

Indeed, for the Dirac-GAN we have

Lemma C.4. *The eigenvalues of the Jacobian of the gradient vector field for consensus optimization at the equilibrium point are given by*

$$\lambda_{1/2} = -\gamma f'(0)^2 \pm i f'(0) \tag{102}$$

In particular, all eigenvalues have a negative real part $-\gamma f'(0)^2$. Hence, simultaneous and alternating gradient descent are both locally convergent using consensus optimization for small enough learning rates.

Proof. As was shown by Mescheder et al. (2017), the Jacobian of the modified vector field $\tilde{v}$ at the equilibrium point is

$$\tilde{v}'(0, 0) = v'(0, 0) - \gamma v'(0, 0)^\intercal v'(0, 0). \tag{103}$$

In our case, this is

$$\begin{pmatrix} -\gamma f'(0)^2 & -f'(0) \\ f'(0) & -\gamma f'(0)^2. \end{pmatrix} \tag{104}$$

A simple calculation shows that the eigenvalues of $\tilde{v}'(0, 0)$ are given by

$$\lambda_{1/2} = -\gamma f'(0)^2 \pm i f'(0). \tag{105}$$

This concludes the proof. □

[10] Mescheder et al. (2017) considered only the case of isolated equilibrium points. However, by applying Theorem A.3, it is straightforward to generalize their proof to the case where we are confronted with a submanifold of equivalent equilibrium points.

A visualization of consensus optimization for the Dirac-GAN is given in Figure 3e.

Unfortunately, consensus optimization has the drawback that it can introduce new spurious points of attraction to the GAN training dynamics. While this is usually not a problem for simple examples, it can be a problem for more complex ones like deep neural networks.

A similar regularization term as in consensus optimization was also independently proposed by Nagarajan & Kolter (2017). However, Nagarajan & Kolter (2017) proposed to only regularize the component $\nabla_\psi L(\theta, \psi)$ of the gradient vector field corresponding to the discriminator parameters. Moreover, the regularization term is only added to the generator objective to give the generator more foresight. It can be shown (Nagarajan & Kolter, 2017) that this simplified regularization term can in certain situations also make the training dynamics locally convergent, but might be better behaved at stationary points of the GAN training dynamics that do not correspond to a local Nash-equilibrium. Indeed, a more detailed analysis shows that this simplified regularization term behaves similarly to instance noise and gradient penalties (which we discussed in Section 3.2 and Section 3.3) for the Dirac-GAN.

D. General convergence results

In this section, we prove Theorem 4.1. To this end, we extend the convergence proof by Nagarajan & Kolter (2017) to our setting. We show that by introducing the gradient penalty terms $R_i(\theta, \psi)$, we can get rid of the assumption that the generator and data distributions locally have the same support. As we have seen, this makes the theory applicable to more realistic cases, where both the generator and data distributions typically lie on lower dimensional manifolds.

D.1. Convergence proof

To prove Theorem 4.1, we first need to understand the local structure of the gradient vector field $v(\theta, \psi)$. Recall that the gradient vector field $v(\theta, \psi)$ is defined as

$$v(\theta, \psi) := \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ \nabla_\psi L(\theta, \psi) \end{pmatrix} \tag{106}$$

with

$$L(\theta, \psi) = \mathrm{E}_{p(z)} \left[f(D_\psi(G_\theta(z))) \right] + \mathrm{E}_{p_\mathcal{D}(x)} \left[f(-D_\psi(x)) \right]. \tag{107}$$

Lemma D.1. *The gradient of $L(\theta, \psi)$ with respect to θ is given by*

$$\nabla_\theta L(\theta, \psi) = \mathrm{E}_{p(z)} \left[f'(D_\psi(G_\theta(z)) \left[\nabla_\theta G_\theta(z) \right]^\mathsf{T} \cdot \nabla_x D_\psi(G_\theta(z)) \right]. \tag{108}$$

Similarly, the gradient of $L(\theta, \psi)$ with respect to ψ is given by

$$\nabla_\psi L(\theta, \psi) = \mathrm{E}_{p_\theta(x)} \left[f'(D_\psi(x)) \nabla_\psi D_\psi(x) \right] - \mathrm{E}_{p_\mathcal{D}(x)} \left[f'(-D_\psi(x)) \nabla_\psi D_\psi(x) \right]. \tag{109}$$

Proof. This is just the chain rule. □

Lemma D.2. *Assume that (θ^*, ψ^*) satisfies Assumption I. The Jacobian of the gradient vector field $v(\theta, \psi)$ at (θ^*, ψ^*) is then*

$$v'(\theta^*, \psi^*) = \begin{pmatrix} 0 & -K_{DG}^\mathsf{T} \\ K_{DG} & K_{DD} \end{pmatrix}. \tag{110}$$

The terms K_{DD} and K_{DG} are given by

$$K_{DD} = 2 f''(0) \, \mathrm{E}_{p_\mathcal{D}(x)} \left[\nabla_\psi D_{\psi^*}(x) \nabla_\psi D_{\psi^*}(x)^\mathsf{T} \right] \tag{111}$$

$$K_{DG} = f'(0) \nabla_\theta \, \mathrm{E}_{p_\theta(x)} \left[\nabla_\psi D_{\psi^*}(x) \right] |_{\theta=\theta^*} \tag{112}$$

Proof. First note that by the definition of $v(\theta, \psi)$ in (106), the Jacobian $v'(\theta^*, \psi^*)$ of $v(\theta, \psi)$ is given by

$$\begin{pmatrix} -\nabla_\theta^2 L(\theta^*, \psi^*) & -\nabla_{\theta,\psi}^2 L(\theta^*, \psi^*) \\ \nabla_{\theta,\psi}^2 L(\theta^*, \psi^*) & \nabla_\psi^2 L(\theta^*, \psi^*) \end{pmatrix}. \tag{113}$$

By Assumption I, $D_{\psi^*}(x) = 0$ in some neighborhood of $\operatorname{supp} p_\mathcal{D}$. Hence, we also have $\nabla_x D_{\psi^*}(x) = 0$ and $\nabla_x^2 D_{\psi^*}(x) = 0$ for $x \in \operatorname{supp} p_\mathcal{D}$. By taking the derivative of (108) with respect to θ and using $\nabla_x D_{\psi^*}(x) = 0$ and $\nabla_x^2 D_{\psi^*}(x) = 0$ for $x \in \operatorname{supp} p_\mathcal{D}$ we see that $\nabla_\theta^2 L(\theta^*, \psi^*) = 0$.

To show (111) and (112), simply take the derivative of (109) with respect to θ and ψ and evaluate at it at $(\theta, \psi) = (\theta^*, \psi^*)$. □

We now take a closer look at the regularized vector field. Recall that we consider the two regularization terms

$$R_1(\theta, \psi) := \frac{\gamma}{2} \mathrm{E}_{p_\mathcal{D}(x)} \left[\|\nabla_x D_\psi(x)\|^2 \right] \tag{114}$$

$$R_2(\theta, \psi) := \frac{\gamma}{2} \mathrm{E}_{p_\theta(x)} \left[\|\nabla_x D_\psi(x)\|^2 \right]. \tag{115}$$

As discussed in Section 4.1, the regularization is only applied to the discriminator. The regularized vector field is hence given by

$$\tilde{v}(\theta, \psi) := \begin{pmatrix} -\nabla_\theta L(\theta, \psi) \\ \nabla_\psi L(\theta, \psi) - \nabla_\psi R_i(\theta, \psi) \end{pmatrix}. \tag{116}$$

Lemma D.3. *The gradient $\nabla_\psi R_i(\theta, \psi)$ of the regularization terms R_i, $i \in \{1, 2\}$, with respect to ψ are*

$$\nabla_\psi R_1(\theta, \psi) = \gamma \, \mathrm{E}_{p_\mathcal{D}(x)} \left[\nabla_{\psi,x} D_\psi(x) \nabla_x D_\psi(x) \right] \tag{117}$$

$$\nabla_\psi R_2(\theta, \psi) = \gamma \, \mathrm{E}_{p_\theta(x)} \left[\nabla_{\psi,x} D_\psi(x) \nabla_x D_\psi(x) \right]. \tag{118}$$

Proof. These equations can be derived by taking the derivative of (114) and (115) with respect to ψ. □

Lemma D.4. *The second derivatives $\nabla^2_\psi R_i(\theta^*, \psi^*)$ of the regularization terms R_i, $i \in \{1, 2\}$, with respect to ψ at (θ^*, ψ^*) are both given by*

$$L_{DD} := \gamma \, \mathrm{E}_{p_\mathcal{D}(x)} \left[\nabla_{\psi,x} D_{\psi^*}(x) \nabla_{\psi,x} D_{\psi^*}(x)^\mathsf{T} \right]. \quad (119)$$

Moreover, both regularization terms satisfy $\nabla_{\theta,\psi} R_i(\theta^*, \psi^*) = 0$.

Proof. $\nabla^2_\psi R_i(\theta^*, \psi^*)$, $i \in \{1, 2\}$, can be computed by taking the derivative of (117) and (118) with respect to ψ and using the fact that $\nabla_x D_{\psi^*}(x) = 0$ in a neighborhood of $\operatorname{supp} p_\mathcal{D}$.

Moreover, we clearly have $\nabla_{\theta,\psi} R_1(\theta^*, \psi^*) = 0$, because R_1 does not depend on θ. To see that $\nabla_{\theta,\psi} R_2(\theta^*, \psi^*) = 0$, take the derivative of (118) with respect to θ and use the fact that $\nabla_x D_{\psi^*}(x) = 0$ and $\nabla^2_x D_{\psi^*}(x) = 0$ for $x \in \operatorname{supp} p_\mathcal{D}$. □

As a result, the Jacobian $\tilde{v}'(\theta^*, \psi^*)$ of the regularized gradient vector field at the equilibrium point is given by

$$\tilde{v}'(\theta^*, \psi^*) = \begin{pmatrix} 0 & -K_{DG}^\mathsf{T} \\ K_{DG} & K_{DD} - L_{DD} \end{pmatrix}. \quad (120)$$

For brevity, we define $M_{DD} := K_{DD} - L_{DD}$.

To prove Theorem 4.1, we have to show that $\tilde{v}'(\theta^*, \psi^*)$ is well behaved when restricting it to the space orthogonal to the tangent space of $\mathcal{M}_G \times \mathcal{M}_D$ at (θ^*, ψ^*):

Lemma D.5. *Assume that Assumptions II and III hold. If $v \neq 0$ is not in the tangent space of $\mathcal{M}_D$ at ψ^*, then $\bar{v}^\mathsf{T} M_{DD} v < 0$.*

Proof. By Lemma D.2, we have

$$v^\mathsf{T} K_{DD} v = 2 f''(0) \, \mathrm{E}_{p_\mathcal{D}(x)} \left[\left(\nabla_\psi D_{\psi^*}(x)^\mathsf{T} v \right)^2 \right] \quad (121)$$

and by Lemma D.4

$$v^\mathsf{T} L_{DD} v = \gamma \, \mathrm{E}_{p_\mathcal{D}(x)} \left[\left\| \nabla_{x,\psi} D_{\psi^*}(x) v \right\|^2 \right]. \quad (122)$$

By Assumption II, we have $f''(0) < 0$. Hence, $v^\mathsf{T} M_{DD} v \leq 0$ and $v^\mathsf{T} M_{DD} v = 0$ implies

$$\nabla_\psi D_{\psi^*}(x)^\mathsf{T} v = 0 \quad \text{and} \quad \nabla_{x,\psi} D_{\psi^*}(x) v = 0 \quad (123)$$

for all $x \in \operatorname{supp} p_\mathcal{D}$.

Let

$$h(\psi) := \mathrm{E}_{p_\mathcal{D}(x)} \left[|D_\psi(x)|^2 + \|\nabla_x D_\psi(x)\|^2 \right]. \quad (124)$$

Using the fact that $D_\psi(x) = 0$ and $\nabla_x D_\psi(x) = 0$ for $x \in \operatorname{supp} p_\mathcal{D}$, we see that the Hessian of $h(\psi)$ at ψ^* is

$$\nabla^2_\psi h(\psi^*) = 2 \, \mathrm{E}_{p_\mathcal{D}(x)} \big[\nabla_\psi D_\psi(x) \nabla_\psi D_\psi(x)^\mathsf{T} + \nabla_{\psi,x} D_\psi(x) \nabla_{\psi,x} D_\psi(x)^\mathsf{T} \big] \quad (125)$$

The second directional derivate $\partial^2_v h(\psi)$ is therefore

$$\partial^2_v h(\psi) = 2 \, \mathrm{E}_{p_\mathcal{D}(x)} \big[|\nabla_\psi D_\psi(x)^\mathsf{T} v|^2 + \|\nabla_{x,\psi} D_\psi(x) v\|^2 \big] = 0. \quad (126)$$

By Assumption III, this can only hold if v is in the tangent space of $\mathcal{M}_D$ at ψ^*. □

Lemma D.6. *Assume that Assumption III holds. If $w \neq 0$ is not in the tangent space of $\mathcal{M}_G$ at θ^*, then $K_{DG} w \neq 0$.*

Proof. By Lemma D.2, we have

$$K_{DG} w = f'(0) \left[\nabla_\theta \, \mathrm{E}_{p_\theta(x)} \left[\nabla_\psi D_{\psi^*}(x) \right] |_{\theta=\theta^*} \right] w = f'(0) \partial_w g(\theta). \quad (127)$$

for

$$g(\theta) := \mathrm{E}_{p_\theta(x)} \left[\nabla_\psi D_{\psi^*}(x) \right]. \quad (128)$$

By Assumption III, this implies $K_{DG} w \neq 0$ if w is not in the tangent space of $\mathcal{M}_G$ at θ^*. □

We are now ready to prove Theorem 4.1:

Theorem 4.1. *Assume Assumption I, II and III hold for (θ^*, ψ^*). For small enough learning rates, simultaneous and alternating gradient descent for $\tilde{v}_1$ and $\tilde{v}_2$ are both convergent to $\mathcal{M}_G \times \mathcal{M}_D$ in a neighborhood of (θ^*, ψ^*). Moreover, the rate of convergence is at least linear.*

Proof. First note that by Lemma D.1 and Lemma D.3 $v(\theta, \psi) = 0$ for all points $(\theta, \psi) \in \mathcal{M}_G \times \mathcal{M}_D$, because $D_\psi(x) = 0$ and $\nabla_x D_\psi(x) = 0$ for all $x \in \operatorname{supp} p_\mathcal{D}$ and $\psi \in \mathcal{M}_D$. Hence, $\mathcal{M}_G \times \mathcal{M}_D$ consists only of equilibrium points of the regularized gradient vector fields.

Let $\mathcal{T}_{\theta^*} \mathcal{M}_G$ and $\mathcal{T}_{\psi^*} \mathcal{M}_D$ denote the tangent spaces of $\mathcal{M}_G$ and $\mathcal{M}_D$ at θ^* and ψ^*.

We now want to show that both simultaneous and alternating gradient descent are locally convergent to $\mathcal{M}_G \times \mathcal{M}_D$ for the regularized gradient vector field $\tilde{v}(\theta, \psi)$. To this end, we want to apply Theorem A.3. By choosing local coordinates $\theta(\alpha, \gamma_G)$ and $\psi(\beta, \gamma_D)$ for $\mathcal{M}_G$ and $\mathcal{M}_D$ and using Remark A.6, we can assume without loss of generality that $\theta^* = 0$, $\psi^* = 0$ as well as

$$\mathcal{M}_G = \mathcal{T}_{\theta^*} \mathcal{M}_G = \{0\}^k \times \mathbb{R}^{n-k} \quad (129)$$

$$\mathcal{M}_D = \mathcal{T}_{\psi^*} \mathcal{M}_D = \{0\}^l \times \mathbb{R}^{m-l}. \quad (130)$$

This allows us to write[11] $\tilde{v}(\theta, \psi) = \tilde{v}(\alpha, \gamma_G, \beta, \gamma_D)$ In order to apply Theorem A.3, we have to show that $\nabla_{(\alpha,\beta)}\tilde{v}(\theta^*, \psi^*)$ only has eigenvalues with negative real-part.

By Lemma D.2, $\nabla_{(\alpha,\beta)}\tilde{v}(\theta^*, \psi^*)$ is of the form

$$\begin{pmatrix} 0 & -\tilde{K}_{DG}^\mathsf{T} \\ \tilde{K}_{DG} & \tilde{K}_{DD} - \tilde{L}_{DD} \end{pmatrix} \tag{131}$$

where $\tilde{K}_{DD}$, $\tilde{K}_{DG}$ and $\tilde{L}_{DD}$ denote the submatrices of K_{DD}, K_{DG} and L_{DD} corresponding to the (α, β) coordinates.

We now show that $\tilde{M}_{DD} := \tilde{K}_{DD} - \tilde{L}_{DD}$ is negative definite and $\tilde{K}_{DG}$ has full column rank.

To this end, first note that

$$\tilde{v}^\mathsf{T}\tilde{M}_{DD}\tilde{v} = v^\mathsf{T}M_{DD}v \tag{132}$$

with $v^\mathsf{T} := (\tilde{v}^\mathsf{T}, 0)$. Note that $v \notin \mathcal{T}_{\psi^*}\mathcal{M}_D$ for $\tilde{v} \neq 0$. Hence, by Lemma D.5 we have that $\tilde{v}^\mathsf{T}\tilde{M}_{DD}\tilde{v} < 0$ if $\tilde{v} \neq 0$. As a result, we see that $\tilde{M}_{DD}$ is symmetric negative definite.

Similarly, for $w^\mathsf{T} := (\tilde{w}^\mathsf{T}, 0)$, the components of $K_{DG}w$ corresponding to the β-coordinates are given by $\tilde{K}_{DG}\tilde{w}$. Again, we have $w \notin \mathcal{T}_{\theta^*}\mathcal{M}_G$ for $\tilde{w} \neq 0$. Hence, by Lemma D.6 we have that $K_{DG}w \neq 0$ if $\tilde{w} \neq 0$. Because the components of $K_{DG}w$ corresponding to the γ_D coordinates are 0, this shows that $\tilde{K}_{DG}\tilde{w} \neq 0$. $\tilde{K}_{DG}$ therefore has full column rank.

Theorem A.7 now implies that all eigenvalues of $\nabla_{(\alpha,\beta)}\tilde{v}(\theta^*, \psi^*)$ have negative real part. By Lemma A.4, Lemma A.5 and Theorem A.3, simultaneous and alternating gradient descent are therefore both convergent to $\mathcal{M}_G \times \mathcal{M}_D$ near (θ^*, ψ^*) for small enough learning rates. Moreover, the rate of convergence is at least linear. □

D.2. Extensions

In the proof of Theorem 4.1 we have assumed that $f''(0) < 0$. This excludes the function $f(t) = t$ which is used in Wasserstein-GANs. We now show that our convergence proof extends to the case where $f(t) = t$ when we modify Assumption III as little bit:

Remark D.7. *When we replace $h(\psi)$ with*

$$\tilde{h}(\psi) := \mathrm{E}_{p_\mathcal{D}(x)}\left[\|\nabla_x D_\psi(x)\|^2\right] \tag{133}$$

and $\mathcal{M}_D$ with $\tilde{\mathcal{M}}_D := \{\psi \mid \tilde{h}(\psi) = 0\}$ the results of Theorem 4.1 still hold for $f(t) = t$.

Proof. Almost everything in the proof of Theorem 4.1 still holds for these modified assumptions. The only thing that we have to show is that $\mathcal{M}_G \times \mathcal{M}_D$ still consists only of equilibrium points and that Lemma D.5 still holds in this setting.

To see the former, note that by Lemma D.1 we still have $\nabla_\theta L(\theta, \psi) = 0$ for $(\theta, \psi) \in \mathcal{M}_G \times \mathcal{M}_D$, because we have $\nabla_x D_\psi(x) = 0$ for $\psi \in \mathcal{M}_D$ and $x \in \operatorname{supp} p_\mathcal{D}$. On the other hand, for $f(t) = t$ we also have $\nabla_\psi L(\theta, \psi) = 0$ if $\theta \in \mathcal{M}_G$, because for $\theta \in \mathcal{M}_G$ the definition of $\mathcal{M}_G$ implies that $p_\theta = p_\mathcal{D}$ and hence, by Lemma D.1,

$$\begin{aligned}\nabla_\psi L(\theta, \psi) = \mathrm{E}_{x\sim p_\mathcal{D}}\left[\nabla_\psi D_\psi(x)\right] \\ - \mathrm{E}_{x\sim p_\mathcal{D}}\left[\nabla_\psi D_\psi(x)\right] = 0.\end{aligned} \tag{134}$$

To see why Lemma D.5 still holds, first note that for $f(t) = t$, we have $f''(0) = 0$, so that by Lemma D.2 $K_{DD} = 0$. Hence,

$$v^\mathsf{T}M_{DD}v = -v^\mathsf{T}L_{DD}v. \tag{135}$$

We therefore have to show that $v^\mathsf{T}L_{DD}v \neq 0$ if v is not in the tangent space of $\mathcal{M}_D$.

However, we have seen in the proof of Lemma D.5 that

$$v^\mathsf{T}L_{DD}v = \gamma\,\mathrm{E}_{p_\mathcal{D}(x)}\left[\left\|\nabla_{x,\psi}D_{\psi^*}(x)v\right\|^2\right]. \tag{136}$$

Hence $v^\mathsf{T}L_{DD}v = 0$ implies $\nabla_{x,\psi}D_{\psi^*}(x)v = 0$ for $x \in \operatorname{supp} p_\mathcal{D}$ and thus

$$\partial_v^2 h(\psi) = 2\,\mathrm{E}_{p_\mathcal{D}(x)}\left[\|\nabla_{x,\psi}D_\psi(x)v\|^2\right] = 0. \tag{137}$$

By Assumption III, this can only be the case if v is in the tangent space of $\mathcal{M}_D$. This concludes the proof. □

In Section D.1, we showed that both regularizers R_1 and R_2 from Section 4.1 make the GAN training dynamics locally convergent. A similar, but slightly more complex regularizer was also proposed by Roth et al. (2017) who tried to find a computationally efficient approximation to instance noise. The regularizer proposed by Roth et al. (2017) is given by a linear combination of R_1 and R_2 where the weighting is adaptively chosen depending on the logits of $D_\psi(x)$ of the current discriminator at a data point x:

$$\begin{aligned}R_{\mathrm{Roth}}(\theta, \psi) = \mathrm{E}_{p_\theta(x)}\left[(1 - \sigma(D_\psi(x)))^2\|\nabla_x D_\psi(x)\|^2\right] \\ + \mathrm{E}_{p_\mathcal{D}(x)}\left[\sigma(D_\psi(x))^2\|\nabla_x D_\psi(x)\|^2\right]\end{aligned} \tag{138}$$

Indeed, we can show that our convergence proof extends to this regularizer (and a slightly more general class of regularizers):

Remark D.8. *When we replace the regularization terms R_1 and R_2 with*

$$\begin{aligned}R_3(\theta, \psi) = \mathrm{E}_{p_\theta(x)}\left[w_1(D_\psi(x))\|\nabla_x D_\psi(x)\|^2\right] \\ + \mathrm{E}_{p_\mathcal{D}(x)}\left[w_2(D_\psi(x))\|\nabla_x D_\psi(x)\|^2\right]\end{aligned} \tag{139}$$

[11] By abuse of notation, we simply write $\theta = (\alpha, \gamma_G)$ and $\psi = (\beta, \gamma_D)$.

so that $w_1(0) > 0$ and $w_2(0) > 0$, the results of Theorem 4.1 still hold.

Proof. Again, we have to show that $\mathcal{M}_G \times \mathcal{M}_D$ still consists only of equilibrium points and that Lemma D.5 still holds in this setting.

However, by using $\nabla_x D_\psi(x) = 0$ for $x \in \operatorname{supp} p_\mathcal{D}$ and $\psi \in \mathcal{M}_D$, it is easy to see that $\nabla_\psi R_3(\theta, \psi) = 0$ for all $(\theta, \psi) \in \mathcal{M}_G \times \mathcal{M}_D$, which implies that $\mathcal{M}_G \times \mathcal{M}_D$ still consists only of equilibrium points.

To see why Lemma D.5 still holds in this setting, note that (after a little bit of algebra) we still have $\nabla_{\theta,\psi} R_3(\theta^*, \psi^*) = 0$ and

$$\nabla^2_\psi R_3(\theta^*, \psi^*) = \frac{1}{\gamma}(w_1(0) + w_2(0)) L_{DD}. \tag{140}$$

The proof of Lemma D.5 therefore still applies in this setting. □

E. Stable equilibria for unregularized GAN training

In Section 2, we have seen that unregularized GAN training is not always locally convergent to the equilibrium point. Moreover, in Section 4, we have shown that zero-centered gradient penalties make general GANs locally convergent under some suitable assumptions.

While our results demonstrate that we cannot expect unregularized GAN training to lead to local convergence for general GAN architectures, there can be situations where unregularized GAN training has stable equilibria. Such equilibria usually require additional assumptions on the class of representable discriminators.

In this section, we identify two types of stable equilibria. For the first class of stable equilibria, which we call *Energy Solutions*, the equilibrium discriminator forms an energy function for the true data distributions and might be a partial explanation for the success of autoencoder-based discriminators (Zhao et al., 2016; Berthelot et al., 2017). For the second class, which we call *Full-rank solutions*, the discriminator learns a representation of the data distribution with certain properties and might be a partial explanation for the success of batch-normalization for training GANs (Radford et al., 2015).

E.1. Energy Solutions

For technical reasons, we assume that $\operatorname{supp} p_\mathcal{D}$ defines a $\mathcal{C}^1$-manifold in this section.

Energy solutions are solutions where the discriminator forms a potential function for the true data distribution. Such solutions (θ^*, ψ^*) satisfy the following property:

Assumption I′. *We have $p_{\theta^*} = p_\mathcal{D}$, $D_{\psi^*}(x) = 0$, $\nabla_x D_{\psi^*}(x) = 0$ and $v^\mathsf{T} \nabla^2_x D_{\psi^*}(x) v > 0$ for all $x \in \operatorname{supp} p_\mathcal{D}$ and v not in the tangent space of $\operatorname{supp} p_\mathcal{D}$ at x.*

We also need a modified version of Assumption III which ensures certain regularity properties of the reparameterization manifolds $\mathcal{M}_G$ and $\mathcal{M}_D$ near the equilibrium (θ^*, ψ^*). To formulate Assumption III′, we need

$$\tilde{g}(\psi) := \nabla_\theta \operatorname{E}_{p_\theta(x)} [D_\psi(x)] \big|_{\theta=\theta^*}. \tag{141}$$

Assumption III′. *There are ϵ-balls $B_\epsilon(\theta^*)$ and $B_\epsilon(\psi^*)$ around θ^* and ψ^* so that $\mathcal{M}_G \cap B_\epsilon(\theta^*)$ and $\mathcal{M}_D \cap B_\epsilon(\psi^*)$ define $\mathcal{C}^1$- manifolds. Moreover, the following holds:*

- *(i) if v is not in the tangent space of $\mathcal{M}_D$ at ψ^*, then $\partial_v \tilde{g}(\psi^*) \neq 0$.*
- *(ii) if w is not in the tangent space of $\mathcal{M}_G$ at θ^*, then there is a latent code $z \in \mathbb{R}^k$ so that $\nabla_\theta G_{\theta^*}(z) w$ is not in the tangent space of $\operatorname{supp} p_\mathcal{D}$ at $G_{\theta^*}(z) \in \operatorname{supp} p_\mathcal{D}$.*

The first part of Assumption III′ implies that the generator gradients become nonzero whenever the discriminator moves away from an equilibrium discriminator. The second part of Assumption III′ means that every time the generator leaves the equilibrium, it pushes some data point aways from $\operatorname{supp} p_\mathcal{D}$, i.e. the generator is not simply redistributing mass on $\operatorname{supp} p_\mathcal{D}$.

In Theorem E.2 we show that energy solutions lead to local convergence of the unregularized GAN training dynamics. For the proof, we first need a generalization of Lemma D.2:

Lemma E.1. *Assume that (θ^*, ψ^*) satisfies Assumption I′. The Jacobian of the gradient vector field $v(\theta, \psi)$ at (θ^*, ψ^*) is then given by*

$$v'(\theta^*, \psi^*) = \begin{pmatrix} K_{GG} & -K_{DG}^\mathsf{T} \\ K_{DG} & K_{DD} \end{pmatrix}. \tag{142}$$

The terms K_{DD} and K_{DG} are given by

$$K_{GG} = -f'(0) \operatorname{E}_{p(z)} \big[[\nabla_\theta G_{\theta^*}(z)]^\mathsf{T} \nabla^2_x D_{\psi^*}(G_{\theta^*}(z)) \nabla_\theta G_{\theta^*}(z) \big] \tag{143}$$

$$K_{DD} = 2 f''(0) \operatorname{E}_{p_\mathcal{D}(x)} [\nabla_\psi D_{\psi^*}(x) \nabla_\psi D_{\psi^*}(x)^\mathsf{T}] \tag{144}$$

$$K_{DG} = f'(0) \left[\nabla_\theta \operatorname{E}_{p_\theta(x)} [\nabla_\psi D_{\psi^*}(x)] \, |_{\theta=\theta^*} \right]^\mathsf{T} \tag{145}$$

Proof. Almost all parts of the proof of Lemma D.2 are still valid. The only thing that remains to show is that $\nabla^2_\theta L(\theta^*, \psi^*) = -K_{GG}$. To see this, just take the derivative of (108) with respect to θ and use the fact that $\nabla_x D_\psi(x) = 0$ for $x \in \operatorname{supp} p_\mathcal{D}$. □

We are now ready to formulate our convergence result for energy solutions:

Theorem E.2. *Assume Assumption I′, II and III′ hold for (θ^*, ψ^*). Moreover, assume that $f'(0) > 0$. For small enough learning rates, simultaneous and alternating gradient descent for the (unregularized) gradient vector field v are both convergent to $\mathcal{M}_G \times \mathcal{M}_D$ in a neighborhood of (θ^*, ψ^*). Moreover, the rate of convergence is at least linear.*

Proof (Sketch). The proof is similar to the proof of Theorem 4.1.

First, note that $\mathcal{M}_G \times \mathcal{M}_D$ still only consists of equilibrium points. Next, we introduce local coordinates and show that for v not in the tangent space of $\mathcal{M}_G$ at θ^*, we have $v^\mathsf{T} K_{GG} v < 0$. This can be shown using Lemma E.1, Assumption I′ and the second part of Assumption III′.

Moreover, we need to show that for w not in the tangent space of $\mathcal{M}_D$ at ψ^*, we have $K_{DG}^\mathsf{T} w \neq 0$. This can be shown by applying the first part of Assumption III′.

The rest of the proof is the same as the proof of Theorem 4.1, except that we have to apply Theorem A.8 instead of Theorem A.7. □

Note that energy solutions are only possible, if the discriminator is able to satisfy Assumption III′. This is not the case for the Dirac-GAN from Section 2. However, if we use a quadratic discriminator instead, there are also energy solutions to the unregularized GAN training dynamics for the Dirac-GAN. To see this, we can parameterize $D_\psi(x)$ as

$$D_\psi(x) := \psi_1 x^2 + \psi_2 x. \tag{146}$$

It is easy to check that the Dirac-GAN with a discriminator as in (146) indeed has energy solutions: every (θ, ψ) with $\theta = 0$ and $\psi_2 = 0$ defines an equilibrium point of the Dirac-GAN and the GAN-training dynamics are locally convergent near this point if $\psi_1 > 0$. Note however, that even though all equilbria with $\psi_1 > 0$ are points of attraction for the *continuous* GAN training dynamics, they may not be attractors for the *discretized system* when ψ_1 is large and the learning rate h is fixed. In general, the conditioning of energy solutions depends on the condition numbers of the Hessians $\nabla_x^2 D_{\psi^*}(x)$ at all $x \in \operatorname{supp} p_\mathcal{D}$. Indeed, the presence of ill-conditioned energy solutions might be one possible explanation why WGAN-GP often works well in practice although it is not even locally convergent for the Dirac-GAN.

E.2. Full-Rank Solutions

In practice, $D_\psi(x)$ is usually implemented by a deep neural network. Such discriminators can be described by functions of the form

$$D_\psi(x) = \psi_1^\mathsf{T} \eta_{\psi_2}(x) \tag{147}$$

with a vector-valued $\mathcal{C}^1$-functions η_{ψ_2} and $\psi = (\psi_1, \psi_2)$. η_{ψ_2} can be regarded as a feature-representation of the data point x.

We now state several assumptions that lead to local convergence in this situation.

The first assumption can be seen as a variant of Assumption I adapted to this specific situation:

Assumption I″. *We have $p_{\theta^*} = p_\mathcal{D}$ and $\psi_1^* = 0$.*

We again consider *reparameterization manifolds*, which we define as follows in this section:

$$\mathcal{M}_G := \{\theta \mid p_\theta = p_\mathcal{D}\} \quad \mathcal{M}'_D := \{\psi \mid \psi_1 = 0\}. \tag{148}$$

Moreover, let

$$g(\theta) = \mathrm{E}_{p_\theta(x)} \left[\eta_{\psi_2^*}(x)\right]. \tag{149}$$

Assumption III now becomes:

Assumption III″. *There is an ϵ-ball $B_\epsilon(\theta^*)$ around θ^* so that $\mathcal{M}_G$ defines a $\mathcal{C}^1$- manifold[12]. Moreover, the following holds:*

(i) The matrix $\mathrm{E}_{p_\mathcal{D}(x)} \left[\eta_{\psi_2^*}(x) \eta_{\psi_2^*}(x)^\mathsf{T}\right]$ has full rank.
(ii) if w is not in the tangent space of $\mathcal{M}_G$ at θ^*, then $\partial_w g(\theta^*) \neq 0$.

We call a function $\eta_{\psi_2^*}$ that satisfies the first part of Assumption III″ a *full-rank representation* of $p_\mathcal{D}$. Moreover, if $\eta_{\psi_2^*}$ satisfies the second part of Assumption III″, we call $\eta_{\psi_2^*}$ a *complete representations*, because the second part of Assumption III″ implies that every deviation from the Nash-equilibrium $p_{\theta^*} = p_\mathcal{D}$ is detectable using $\eta_{\psi_2^*}$.

In practice, complete full-rank representations might only exist if the class of discriminators is very powerful or the class of generators is limited. Especially the second part of Assumption III″ might be hard to satisfy in practice. Moreover, finding such representations might be much harder than finding equilibria for the regularized GAN-training dynamics from Section 4.

Nonetheless, we have the following convergence result for GANs that allow for complete full-rank representations:

Theorem E.3. *Assume Assumption I′, Assumption II and III′ hold for (θ^*, ψ^*). For small enough learning rates, simultaneous and alternating gradient descent for the (unregularized) gradient vector field v are both convergent to $\mathcal{M}_G \times \mathcal{M}'_D$ in a neighborhood of (θ^*, ψ^*). Moreover, the rate of convergence is at least linear.*

Proof (Sketch). The proof is again similar to the proof of Theorem 4.1. We again introduce local coordinates and

[12] Note that $\mathcal{M}'_D$ is a $\mathcal{C}^1$-manifold by definition in this setup.

show that for w not in the tangent space of $\mathcal{M}'_D$ at ψ^*, we have $w^\intercal K_{DD} w < 0$. To see this, note that w must have a nonzero ψ_1 component if it is not in the tangent space of $\mathcal{M}'_D$ at ψ^*. However, using (111), we see that the submatrix of K_{DD} corresponding to the ψ_1 coordinates is given by

$$\tilde{K}_{DD} = 2f''(0)\,\mathrm{E}_{p_\mathcal{D}(x)}\left[\eta_{\psi_2^*}(x)\eta_{\psi_2^*}(x)^\intercal\right]. \tag{150}$$

This matrix is negative definite by Assumption II and the first part of Assumption III''.

Moreover, by applying (112), we see that the component of $K_{DG}w$, $w \in \mathbb{R}^n$, corresponding to the ψ_1 coordinates is given by

$$\partial_w g(\theta^*) = f'(0)\nabla_\theta\,\mathrm{E}_{p_\theta(x)}\left[\eta_{\psi_2^*}(x)\right]\big|_{\theta=\theta^*} w. \tag{151}$$

Using the second part of Assumption III'', we therefore see that for w not in the tangent space of $\mathcal{M}_G$ at θ^*, we have $K_{DG}w \neq 0$.

The rest of the proof is the same as the proof of Theorem 4.1. □

For the Dirac-GAN from Section 2, we can obtain a complete full-rank representation, when we parameterize the discriminator D_ψ as $D_\psi(x) = \psi\exp(x)$, i.e. if we set $\psi_1 := \psi$ and $\eta_{\psi_2}(x) := \exp(x)$. It is easy to check that η_{ψ_2} indeed defines a complete full-rank representation and that the Dirac-GAN is locally convergent to $(\theta^*, \psi^*) = (0, 0)$ for this parameterization of $D_\psi(x)$.

F. Experiments

In this section, we describe additional experiments and give more details on our experimental setup. If not noted otherwise, we always use the nonsaturating GAN-objective introduced by Goodfellow et al. (2014) for training the generator. For WGAN-GP we use the generator and discriminator objectives introduced by Gulrajani et al. (2017).

2D-Problems For the 2D-problems, we run unregularized GAN training, R_1-regularized and R_2-regularized GAN training as well WGAN-GP with 1 and 5 discriminator update per generator update. We run each method on 4 different 2D-examples for 6 different GAN architectures. The 4 data-distributions are visualized in Figure 8. All 6 GAN architectures consist of 4-layer fully connected neural networks for both the generator and discriminator, where we select the number of hidden units from $\{8, 16, 32\}$ and use select either leaky RELUs (i.e. $\varphi(t) = \max(t, 0.2t)$) or Tanh-activation functions.

For each method, we try both Stochastic Gradient Descent (SGD) and RMS-Prop with 4 different learning rates: for SGD, we select the learning rate from $\{5 \cdot 10^{-3}, 10^{-2}, 2 \cdot 10^{-2}, 5 \cdot 10^{-2}\}$. For RMSProp, we select it from $\{5 \cdot 10^{-5}, 10^{-4}, 2 \cdot 10^{-4}, 5 \cdot 10^{-4}\}$. For the R_1-, R_2- and WGAN-GP-regularizers we try the regularization parameters $\gamma = 1$, $\gamma = 3$ and $\gamma = 10$. For each method and architecture, we pick the hyperparameter setting which achieves the lowest Wasserstein-1-distance to the true data distribution. We train all methods for 50k iterations and we report the Wasserstein-1-distance averaged over the last 10k iterations. We estimate the Wasserstein-1-distance using the Python Optimal Transport package[13] by drawing 2048 samples from both the generator and the true data distributions.

The best solution found by each method on the "Circle"-distribution is shown in Figure 9. We see that the R_1- and R_2-regularizers converge to solutions for which the discriminator is 0 in a neighborhood of the true data distribution. On the other hand, unregularized training and WGAN-GP converge to *energy solutions* where the discriminator forms a potential for the true data distribution. Please see Section E.1 for details.

CIFAR-10 To test our theory on real-world tasks, we train a DC-GAN architecture (Radford et al., 2015) with 3 convolutional layers and no batch-normalization on the CIFAR-10 dataset (Krizhevsky & Hinton, 2009). We apply different regularization strategies to stabilize the training. To compare the different regularization strategies, we measure the inception score (Salimans et al., 2016) over Wall-clock-time. We implemented the network in the Tensorflow framework (Abadi et al., 2016). For all regularization techniques, we use the RMSProp optimizer (Tieleman & Hinton, 2012) with $\alpha = 0.9$ and a learning rate of 10^{-4}.

For the R_1 and R_2 regularizers from Section 4.1 we use a regularization parameter of $\gamma = 10$. For the WGAN-GP regularizer we also use a regularization parameter of $\gamma = 10$ as suggested by Gulrajani et al. (2017). We train all methods using 1 discriminator update per generator update except for WGAN-GP, for which we try both 1 and 5 discriminator updates

The inception score (Salimans et al., 2016) over time for the different regularization strategies is shown in Figure 6. As predicted by our theory, we see that the R_1 and R_2 regularizers from Section 4.1 lead to stable training whereas unregularized GAN training is not stable. We also see that WGAN-GP with 1 or 5 discriminator updates per generator update lead to similar final inception scores on this architecture. The good behavior of WGAN-GP is surprising considering the fact that it does not even converge locally for the Dirac-GAN. One possible explanation is that WGAN-GP oscillates in narrow circles around the equilibrium which might be enough to produce images of sufficiently high

[13] http://pot.readthedocs.io

quality. Another possible explanation is that WGAN-GP converges to an energy or a full-rank solution (Section E) for this example.

Imagenet For the Imagenet experiment, we use ResNet-architectures for the generator and discriminator, both having 55 layers in total. Both the generator and discriminator are conditioned on the labels of the input data. The architectures for the generator and discriminator are shown in Table 3. We use preactivation ResNet-blocks and Leaky RELU-nonlinearities everywhere. We also multiply the output of the ResNet blocks with 0.1. For the generator, we sample a latent variable z from a 256-dimensional uniform distribution on $[-1, 1]^{256}$ and concatenate it with a 256 dimensional embedding of the labels. The resulting 512-dimensional vector is then fed into the first fully connected layer of the generator. The discriminator takes as input an image and outputs a 1000 dimensional vector. Depending on the label of the input, we select the corresponding index in this vector and use it as the logits for the GAN-objective.

We implemented the network in the Pytorch framework (Paszke et al., 2017) and use the RMSProp optimizer with $\alpha = 0.99$, $\epsilon = 10^{-5}$ and an initial learning rate of 10^{-4}. We use a batch size of 128 and we train the networks on 4 GeForce GTX 1080 Ti GPUs for 35 epochs. Every 10 epochs, we anneal the learning rate by a factor of 2.

We find that while training this GAN without any regularization quickly leads to mode collapse, using the R_1-regularizers from Section 4.1 leads to stable training.

Some random (unconditional) samples can be seen in Figure 10. Moreover, Figure 11 and Figure 12 show conditional samples for some selected Imagenet classes. While not completely photorealistic, we find that our model can produce convincing samples from all 1000 Imagenet classes.

We also compare the R_1-regularizer with WGAN-GP (with 1 discriminator update per generator update) on a slightly smaller architecture[14] and no learning rate annealing. The resulting inception score[15] over the number of iterations is visualized in Figure 7. We find that for this dataset and architecture we can achieve higher inception scores when using the R_1-regularizer in place of the WGAN-GP regularizer.

celebA and LSUN To see if the R_1-regularizers helps to train GANs for high-resolution image distributions, we apply our method to the celebA dataset (Liu et al., 2015) and to 4 subsets of the LSUN dataset (Yu et al., 2015) with resolution 256×256. We use a similar training setup as for the Imagenet experiment, but we use a slightly different architecture (Table 4). As in the Imagenet-experiment, we use preactivation ResNet-blocks and Leaky RELU-nonlinearities everywhere and we multiply the output of the ResNet-blocks with 0.1. We implemented the network in the Pytorch framework and use the RMSProp optimizer with $\alpha = 0.99$ and a learning rate of 10^{-4}. As a regularization term, we use the R_1-regularizer with $\gamma = 10$. For the latent code z, we use a 256 dimensional Gaussian distribution. The batch size is 64.

We find that the R_1 successfully stabilizes training of this architecture. Some random samples can be seen in Figures 13, 14, 15, 16 and 17.

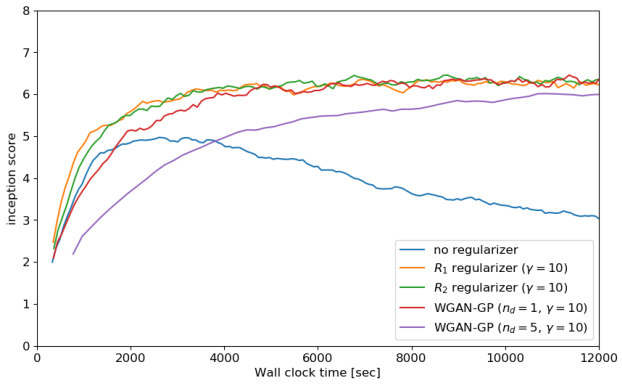


Figure 6. Inception score over time for various regularization strategies when training on CIFAR-10. While the inception score can be problematic for evaluating probabilistic models (Barratt & Sharma, 2018), it still gives a rough idea about the convergence and stability properties of different training methods.

celebA-HQ In addition to the generative model for celebA with resolution 256×256, we train a GAN on celebA-HQ (Karras et al., 2017) with resolution 1024×1024. We use almost the same architecture as for celebA (Table 4), but add two more levels to increase the resolution from 256×256 to 1024×1024 and decrease the number of features from 64 to 16. Because of memory constraints, we also decrease the batch size to 24. In contrast to Karras et al. (2017), we train our model end-to-end during the whole course of training, i.e. we do not use progressive growing of the GAN-architectures (nor any of the other techniques used by Karras et al. (2017) to stabilize the training). We find that the simple R_1-regularizer stabilizes the training, allowing our model to converge to a good (albeit not perfect) solution without using a progressively growing GAN. Some random samples are shown in Figure 18.

[14] For computational reasons we only use 2 instead of 4 RESNET-blocks in each level for this experiment.

[15] For measuring the inception score, we use the public implementation from http://github.com/sbarratt/inception-score-pytorch.

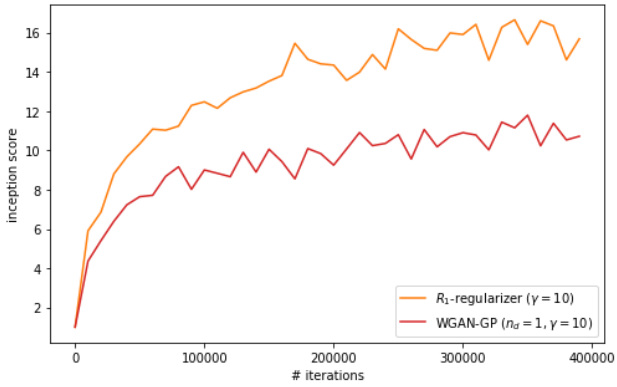


Figure 7. Inception score over the number of iterations for GAN training with R_1- and WGAN-GP-regularization when training on Imagenet. We find that R_1-regularization leads to higher inception scores for this dataset and GAN-architecture.

Layer	output size	filter
Fully Connected	$256 \cdot 4 \cdot 4$	$256 \rightarrow 256 \cdot 4 \cdot 4$
Reshape	$256 \times 4 \times 4$	-
TransposedConv2D	$128 \times 8 \times 8$	$256 \rightarrow 128$
TransposedConv2D	$64 \times 16 \times 16$	$128 \rightarrow 64$
TransposedConv2D	$3 \times 32 \times 32$	$64 \rightarrow 3$

(a) Generator architecture

Layer	output size	filter
Conv2D	$64 \times 16 \times 16$	$3 \rightarrow 64$
Conv2D	$128 \times 8 \times 8$	$64 \rightarrow 128$
Conv2D	$256 \times 4 \times 4$	$128 \rightarrow 256$
Reshape	$256 \cdot 4 \cdot 4$	-
Fully Connected	$256 \cdot 4 \cdot 4$	$256 \cdot 4 \cdot 4 \rightarrow 1$

(b) Discriminator architecture

Table 2. Architectures for CIFAR-10-experiment.

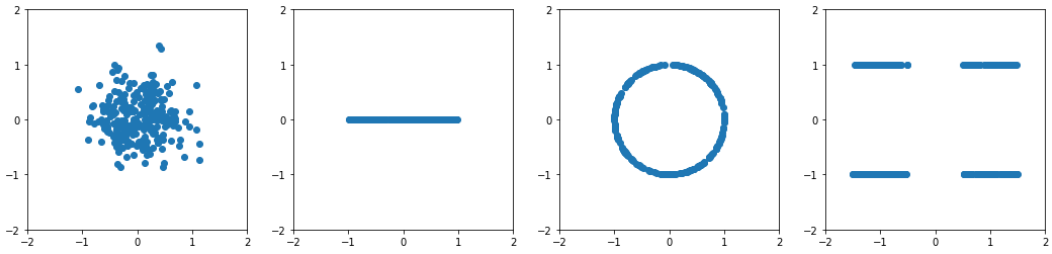
Layer	output size	filter
Fully Connected	$1024 \cdot 4 \cdot 4$	$512 \rightarrow 1024 \cdot 4 \cdot 4$
Reshape	$1024 \times 4 \times 4$	-
Resnet-Block (4x)	$1024 \times 4 \times 4$	$1024 \rightarrow 512 \rightarrow 1024$
NN-Upsampling	$1024 \times 8 \times 8$	-
Conv2D	$1024 \times 8 \times 8$	$1024 \rightarrow 1024$
Resnet-Block (4x)	$1024 \times 8 \times 8$	$1024 \rightarrow 512 \rightarrow 1024$
NN-Upsampling	$1024 \times 16 \times 16$	-
Conv2D	$512 \times 16 \times 16$	$1024 \rightarrow 512$
Resnet-Block (4x)	$512 \times 16 \times 16$	$512 \rightarrow 256 \rightarrow 512$
NN-Upsampling	$512 \times 32 \times 32$	-
Conv2D	$256 \times 32 \times 32$	$512 \rightarrow 256$
Resnet-Block (4x)	$256 \times 32 \times 32$	$256 \rightarrow 128 \rightarrow 256$
NN-Upsampling	$256 \times 64 \times 64$	-
Conv2D	$128 \times 64 \times 64$	$256 \rightarrow 128$
Resnet-Block (4x)	$128 \times 64 \times 64$	$128 \rightarrow 64 \rightarrow 128$
NN-Upsampling	$128 \times 128 \times 128$	-
Conv2D	$64 \times 128 \times 128$	$128 \rightarrow 64$
Resnet-Block (4x)	$64 \times 128 \times 128$	$64 \rightarrow 32 \rightarrow 64$
Conv2D	$3 \times 128 \times 128$	$16 \rightarrow 3$

(a) Generator architecture

Layer	output size	filter
Conv2D	$64 \times 128 \times 128$	$3 \rightarrow 64$
Resnet-Block (4x)	$64 \times 128 \times 128$	$64 \rightarrow 32 \rightarrow 64$
Conv2D	$128 \times 64 \times 64$	$64 \rightarrow 128$
Resnet-Block (4x)	$128 \times 64 \times 64$	$128 \rightarrow 64 \rightarrow 128$
Conv2D	$256 \times 32 \times 32$	$128 \rightarrow 256$
Resnet-Block (4x)	$256 \times 32 \times 32$	$256 \rightarrow 128 \rightarrow 256$
Conv2D	$512 \times 16 \times 16$	$256 \rightarrow 512$
Resnet-Block (4x)	$512 \times 16 \times 16$	$512 \rightarrow 256 \rightarrow 512$
Conv2D	$1024 \times 8 \times 8$	$512 \rightarrow 1024$
Resnet-Block (4x)	$1024 \times 8 \times 8$	$1024 \rightarrow 512 \rightarrow 1024$
Conv2D	$1024 \times 4 \times 4$	$1024 \rightarrow 1024$
Resnet-Block (4x)	$1024 \times 4 \times 4$	$1024 \rightarrow 512 \rightarrow 1024$
Reshape	$1024 \cdot 4 \cdot 4$	-
Fully Connected	$1024 \cdot 4 \cdot 4$	$1024 \cdot 4 \cdot 4 \rightarrow 1000$

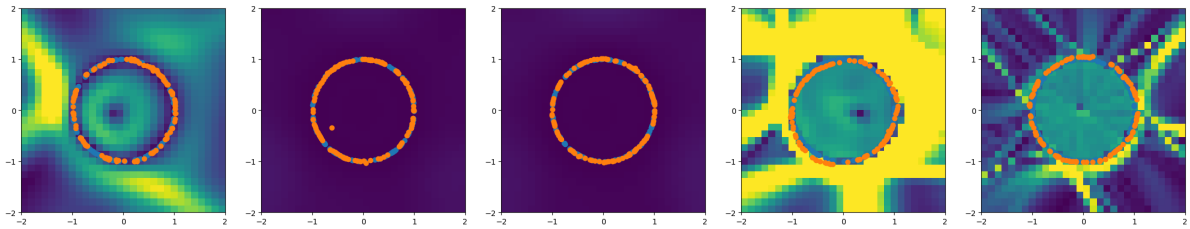
(b) Discriminator architecture

Table 3. Architectures for Imagenet-experiment.



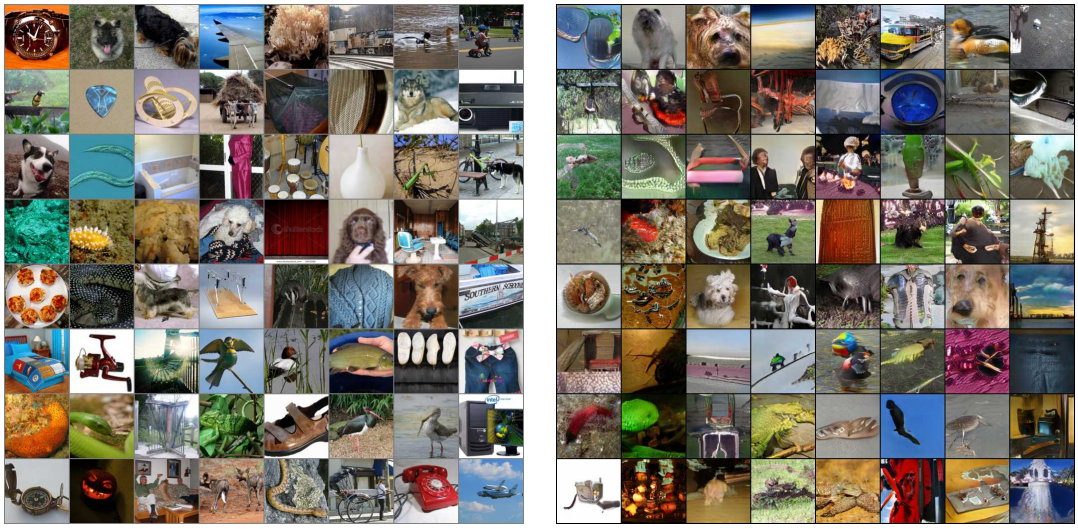
(a) 2D-Gaussian (b) Line (c) Circle (d) Four lines

Figure 8. The four 2D-data distributions on which we test the different algorithms.



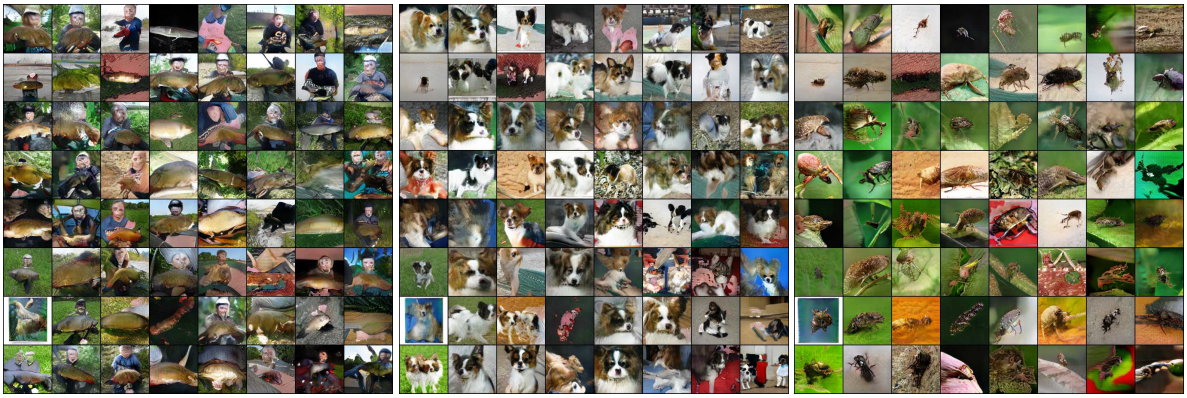
(a) unregularized (b) R_1 (c) R_2 (d) WGAN-GP-1 (e) WGAN-GP-5

Figure 9. Best solutions found by the different algorithms for learning a circle. The blue points are samples from the true data distribution, the orange points are samples from the generator distribution. The colored areas visualize the gradient magnitude of the equilibrium discriminator. We find that while the R_1- and R_2-regularizers converge to equilibrium discriminators that are 0 in a neighborhood of the true data distribution, unregularized training and WGAN-GP converge to energy solutions (Section E.1).

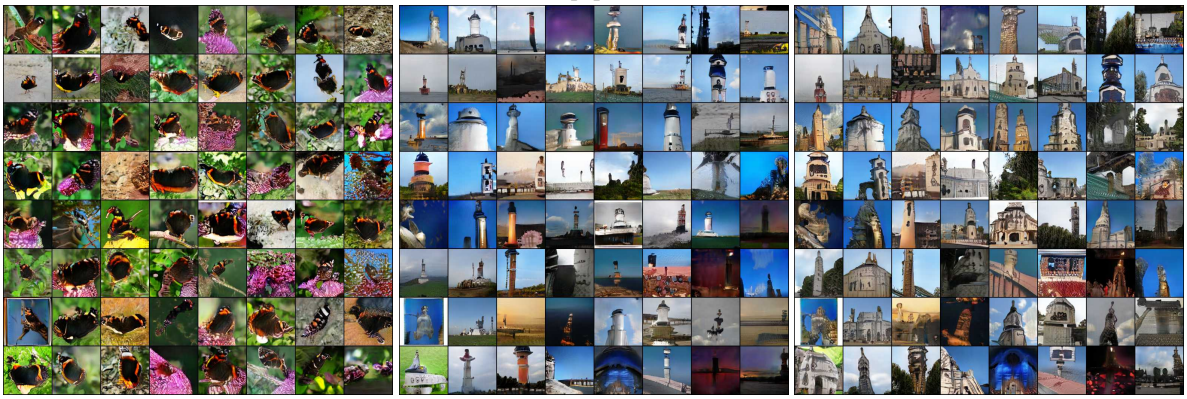


(a) Training distribution (b) Random samples

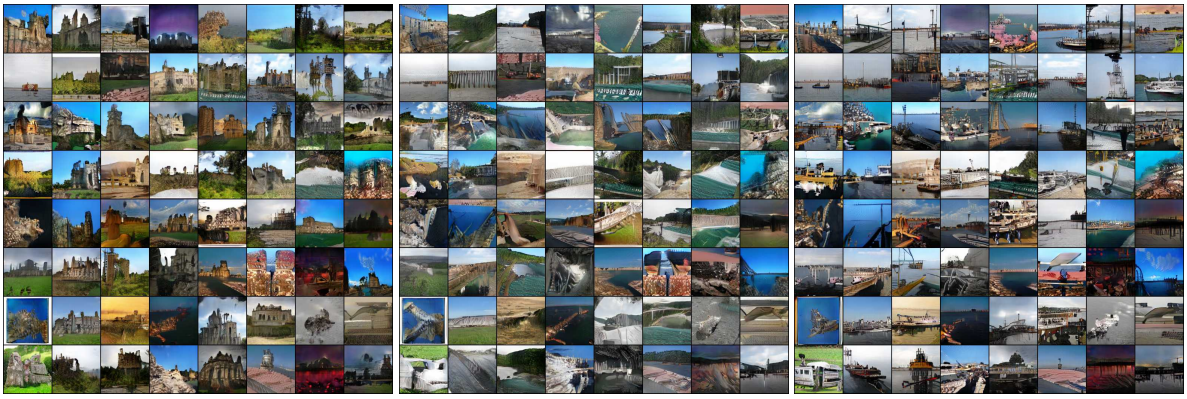
Figure 10. Unconditional results on the Imagenet dataset (Russakovsky et al., 2015) with resolution 128×128. The final inception score is 18.5 ± 0.4.



(a) tench (b) papillon (c) weevil



(d) admiral (e) lighthouse (f) bell cote



(g) castle (h) dam (i) dock

Figure 11. Class conditional results on the Imagenet dataset.

(a) home theater (b) jack-o'-lantern (c) passenger car

(d) police van (e) rugby ball (f) ski

(g) trifle (h) pizza (i) valley

Figure 12. Class conditional results on the Imagenet dataset.

Figure 13. Results on the celebA dataset (Liu et al., 2015) (256×256) for a DC-GAN (Radford et al., 2015) based architecture with additional residual connections (He et al., 2016). For both the generator and the discriminator, we do not use batch normalization.

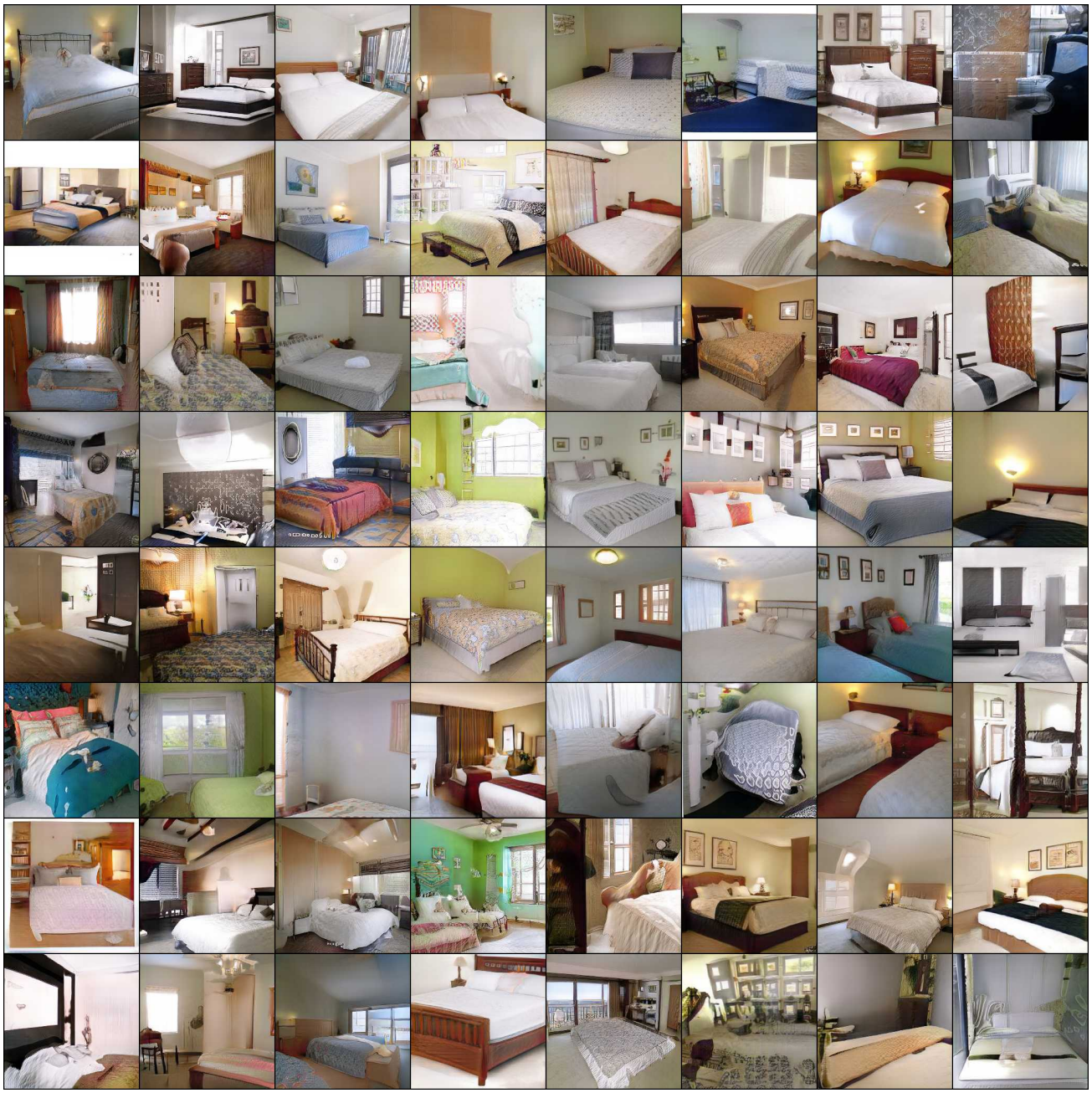


Figure 14. Results on the LSUN-bedroom dataset (Yu et al., 2015) (256×256) for a DC-GAN (Radford et al., 2015) based architecture with additional residual connections (He et al., 2016). For both the generator and the discriminator, we do not use batch normalization.

Figure 15. Results on the LSUN-church dataset (Yu et al., 2015) (256×256) for a DC-GAN (Radford et al., 2015) based architecture with additional residual connections (He et al., 2016). For both the generator and the discriminator, we do not use batch normalization.

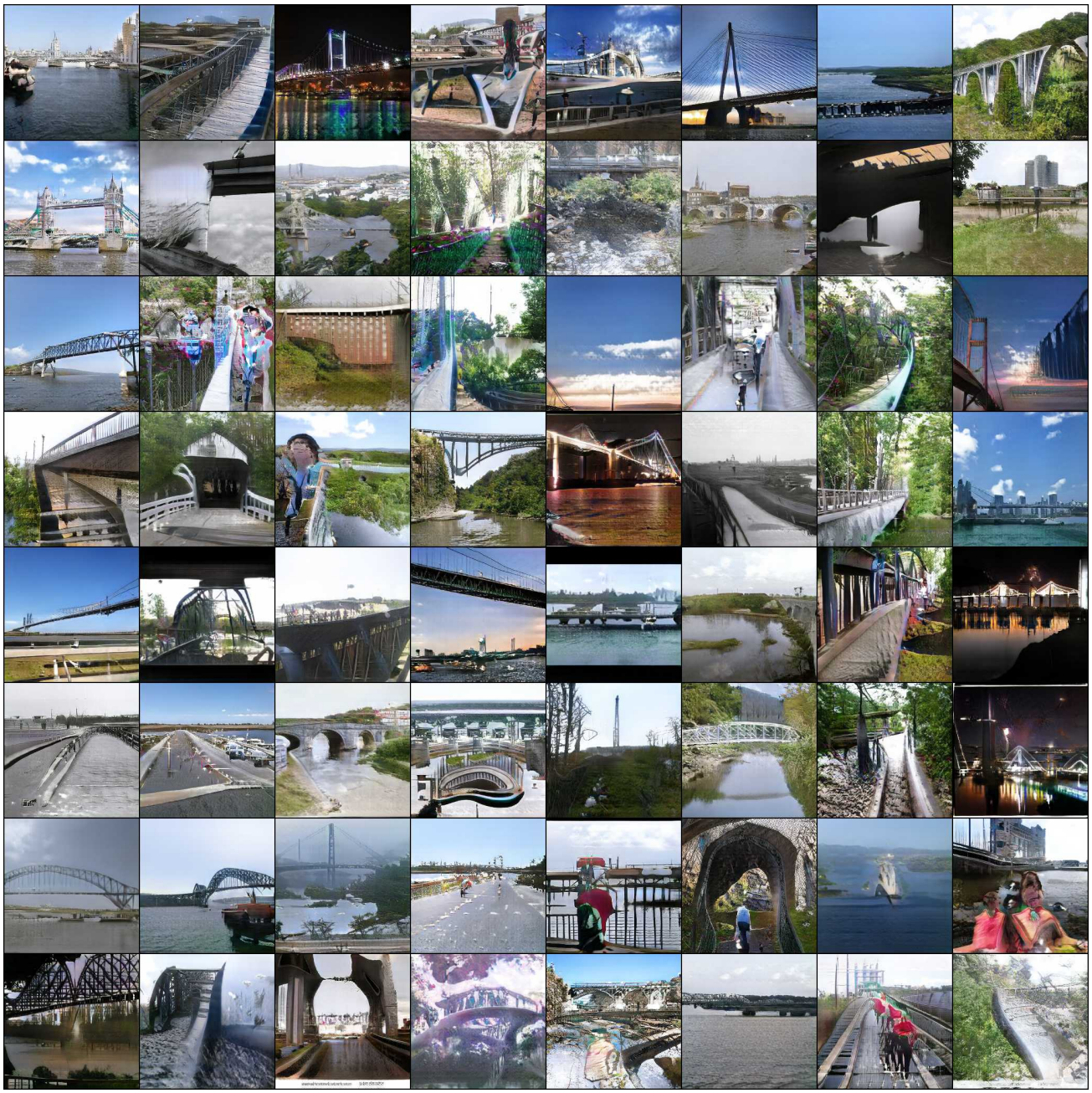


Figure 16. Results on the LSUN-bridge dataset (Yu et al., 2015) (256×256) for a DC-GAN (Radford et al., 2015) based architecture with additional residual connections (He et al., 2016). For both the generator and the discriminator, we do not use batch normalization.

Figure 17. Results on the LSUN-tower dataset (Yu et al., 2015) (256×256) for a DC-GAN (Radford et al., 2015) based architecture with additional residual connections (He et al., 2016). For both the generator and the discriminator, we do not use batch normalization.

Figure 18. Results on the celebA-HQ dataset (Karras et al., 2017) (1024×1024) for a DC-GAN (Radford et al., 2015) based architecture with additional residual connections (He et al., 2016). During the whole course of training, we directly train the full-resolution generator and discriminator end-to-end, i.e. we do not use any of the techniques described in Karras et al. (2017) to stabilize the training.

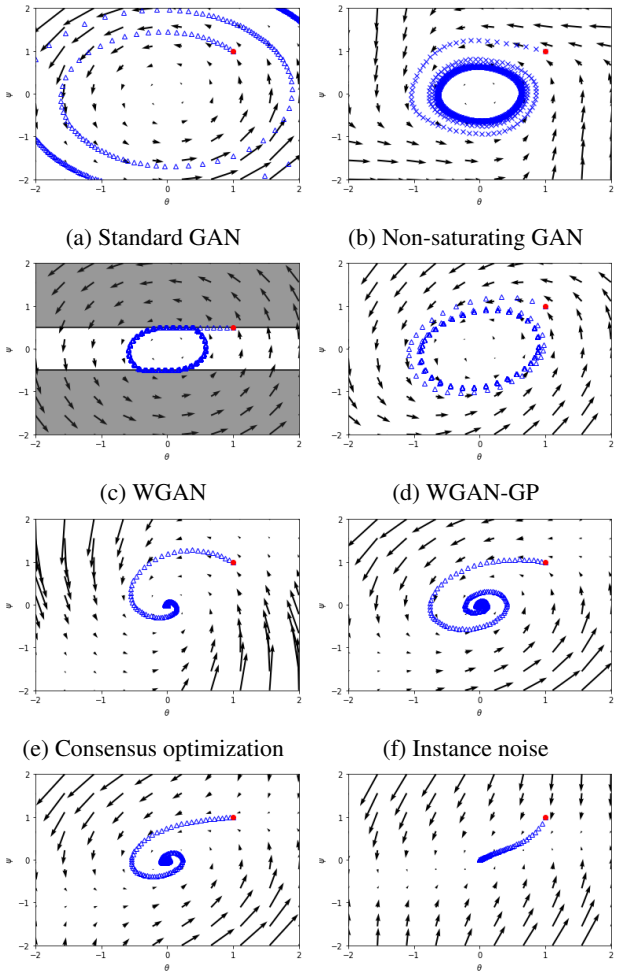
Layer	output size	filter
Fully Connected	$1024 \cdot 4 \cdot 4$	$512 \rightarrow 1024 \cdot 4 \cdot 4$
Reshape	$1024 \times 4 \times 4$	-
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
NN-Upsampling	$1024 \times 8 \times 8$	-
Resnet-Block	$1024 \times 8 \times 8$	$1024 \rightarrow 1024 \rightarrow 1024$
NN-Upsampling	$1024 \times 16 \times 16$	-
Resnet-Block	$512 \times 16 \times 16$	$1024 \rightarrow 512 \rightarrow 512$
NN-Upsampling	$512 \times 32 \times 32$	-
Resnet-Block	$256 \times 32 \times 32$	$512 \rightarrow 256 \rightarrow 256$
NN-Upsampling	$256 \times 64 \times 64$	-
Resnet-Block	$128 \times 64 \times 64$	$256 \rightarrow 128 \rightarrow 128$
NN-Upsampling	$128 \times 128 \times 128$	-
Resnet-Block	$64 \times 128 \times 128$	$128 \rightarrow 64 \rightarrow 64$
NN-Upsampling	$64 \times 256 \times 256$	-
Resnet-Block	$64 \times 256 \times 256$	$64 \rightarrow 64 \rightarrow 64$
Conv2D	$3 \times 256 \times 256$	$3 \rightarrow 3$

(a) Generator architecture

Layer	output size	filter
Conv2D	$64 \times 256 \times 256$	$3 \rightarrow 64$
Resnet-Block	$64 \times 256 \times 256$	$64 \rightarrow 64 \rightarrow 64$
Avg-Pool2D	$64 \times 128 \times 128$	-
Resnet-Block	$128 \times 128 \times 128$	$64 \rightarrow 64 \rightarrow 128$
Avg-Pool2D	$128 \times 64 \times 64$	-
Resnet-Block	$256 \times 64 \times 64$	$128 \rightarrow 128 \rightarrow 256$
Avg-Pool2D	$256 \times 32 \times 32$	-
Resnet-Block	$512 \times 32 \times 32$	$256 \rightarrow 256 \rightarrow 512$
Avg-Pool2D	$512 \times 16 \times 16$	-
Resnet-Block	$1024 \times 16 \times 16$	$512 \rightarrow 512 \rightarrow 1024$
Avg-Pool2D	$1024 \times 8 \times 8$	-
Resnet-Block	$1024 \times 8 \times 8$	$1024 \rightarrow 1024 \rightarrow 1024$
Avg-Pool2D	$1024 \times 4 \times 4$	-
Fully Connected	$1024 \cdot 4 \cdot 4$	$1024 \cdot 4 \cdot 4 \rightarrow 1000$

(b) Discriminator architecture

Table 4. Architectures for LSUN- and celebA-experiments.



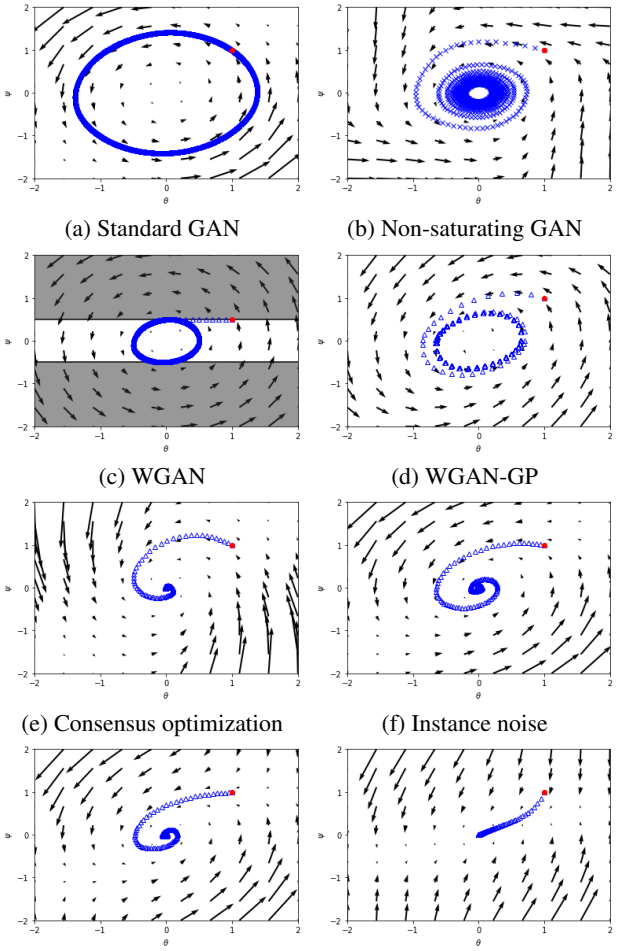
(a) Standard GAN (b) Non-saturating GAN

(c) WGAN (d) WGAN-GP

(e) Consensus optimization (f) Instance noise

(g) Gradient penalty (h) Gradient penalty (CR)

Figure 19. Convergence properties of different GAN training algorithms using simultaneous gradient descent. The shaded area in Figure 19c visualizes the set of forbidden values for the discriminator parameter ψ. The starting iterate is marked in red.



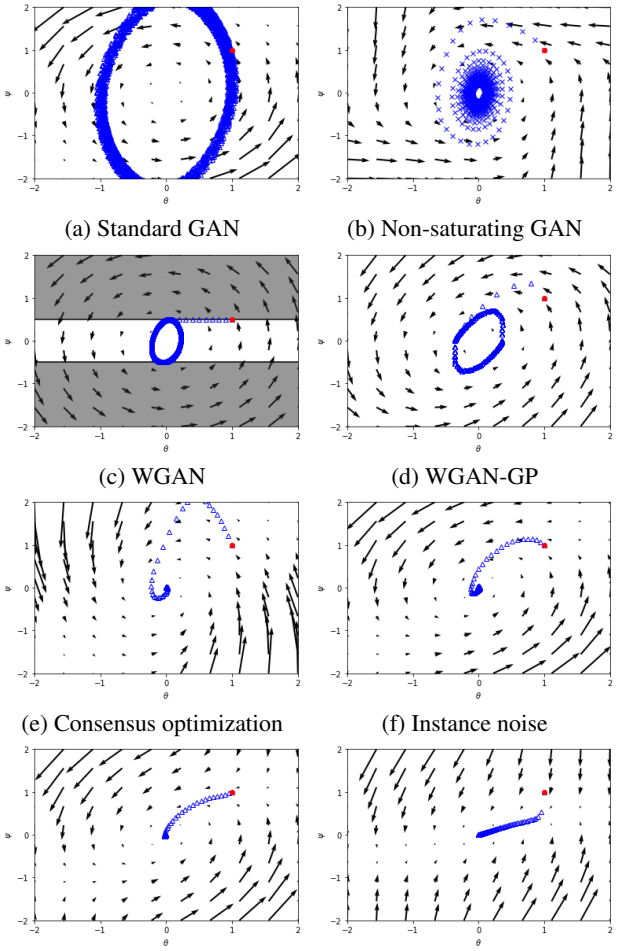
(a) Standard GAN (b) Non-saturating GAN

(c) WGAN (d) WGAN-GP

(e) Consensus optimization (f) Instance noise

(g) Gradient penalty (h) Gradient penalty (CR)

Figure 20. Convergence properties of different GAN training algorithms using alternating gradient descent with 1 discriminator update per generator update The shaded area in Figure 20c visualizes the set of forbidden values for the discriminator parameter ψ. The starting iterate is marked in red.



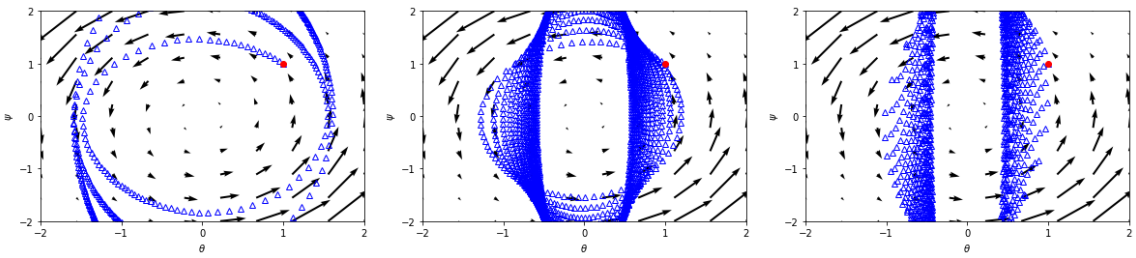
(a) Standard GAN (b) Non-saturating GAN

(c) WGAN (d) WGAN-GP

(e) Consensus optimization (f) Instance noise

(g) Gradient penalty (h) Gradient penalty (CR)

Figure 21. Convergence properties of different GAN training algorithms using alternating gradient descent with 5 discriminator updates per generator update. The shaded area in Figure 21c visualizes the set of forbidden values for the discriminator parameter ψ. The starting iterate is marked in red.



(a) SimGD (b) AltGD ($n_d = 1$) (c) AltGD ($n_d = 5$)

Figure 22. Convergence properties of our GAN using two time-scale training as proposed by Heusel et al. (2017). For the Dirac-GAN, we do not see any sign of convergence when training with two time-scales. The starting iterate is marked in red.