

Continuous한 GAN이 항상 높은 fidelity를 의미하지 않는다.

Convergence라는 GAN의 다른 연구.

WGAN, WGAN-GP가 항상 수렴하지는 않는다.

Convergence라는 method 자체 및 원인 (instance noise and RL gradient penalty)

1. Introduction.

GAN의 stability에 관한 연구

이전 연구에서, gradient의 편평한 Jacobian의 eigenvalue가

local convergence와 stability의 관련 검사를 할 수 있다는 것을 보임.

↳ Jacobian이 평형점에서 음의 실수의 eigenvalue만 가지면, GAN은 local convergence.

↳ " 하나 혹은 더 많은 가지면, 일반적으로 local convergence가 아님

↳ 하나도 없지만, 가까운 경우 작은 $\|r\|$ 을 요구할 수 있다.

↳ 하지만 이것이 instability의 원인인지는 규명하지 않음.

논문에서는

1. unregularized GAN이 항상 local convergence하지 않는다.

2. GAN의 안정화를 위한 논의

3. Data manifold의 점이 공간에 적고나는 discriminator 때문의 instability 설명.

4. gradient penalty 설명

2. Instabilities in GAN Training.

대통 GAN의 불안정성

GAN의 local convergence는 equilibrium의 Jacobian $F_h(\theta^*, \psi^*)$ 의 spectrum을 놓고 분석.

$F_h(\theta^*, \psi^*)$ 가 결대값이 1보다 큰 eigenvalue를 가질 경우, optimal인 (θ^*, ψ^*) 이 수렴X.

1보다 작으면

- Dirac-GAN

unregularized 원 GAN은 local, globally convergence하지 않는다는 것을 간단한 반례로 보임.

- Definition 2.1.

Dirac-GAN은 다음과 같다. $p_D = \delta_0$, $D_\psi(x) = \psi \cdot x$ 로 구성.

실제 data는 0이상의 dirac의 쪽에集聚

G, D는 1개의 param으로 구성됨.

Loss는 $L(\theta, \psi) = f(\psi \theta) + f(\theta) \dots$ (C)

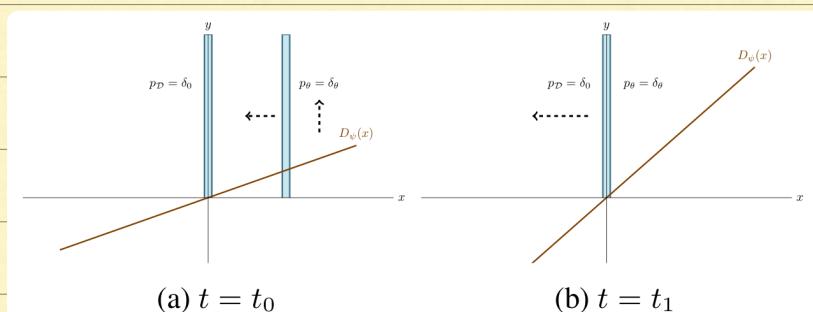


Figure 1. Visualization of the counterexample showing that gradient descent based GAN optimization is not always convergent:
(a) In the beginning, the discriminator pushes the generator towards the true data distribution and the discriminator's slope increases.
(b) When the generator reaches the target distribution, the slope of the discriminator is largest, pushing the generator away from the target distribution. This results in oscillatory training dynamics that never converge.

→ (a)에는 real 쪽으로 밀지만

어느정도 수렴 확인

D의 기울기가 커지면,

진동을 빼기 힘

즉, GAN이 수렴하지 않을

- Lemma 2.2.

(4)의 unique equilibrium point는 $\theta = \psi = 0$.

그리고 EP(equilibrium point)의 gradient의 Jacobian은 $\pm f'(0)$ 의 경계의 실수축을 가짐.

→ 선형적으로 수렴하지 않고, sublinear로 수렴됨

→ but 다음에서 사실이 아님을 보임.

- Lemma 2.3.

gradient vector field의 integral curve $v(\theta, \psi)$ 는 Nash-equilibrium으로 수렴x.

그리고, $v(\theta, \psi)$ 의 모든 integral curve $(\theta(t), \psi(t))$ 는

모든 $t \in [0, \infty]$ 에 대해 $\theta(t) \neq \psi(t) = \text{constant}$ 를 만족.

G 의 분포가 continuous이라도, 그에 대해 optimal한 D 가 continuous하지 않을 때.

$\theta = 0$ 이 아니면 초기의 D 의 param이 충족하지 않는다.

- Lemma 2.4.

simultaneous GD

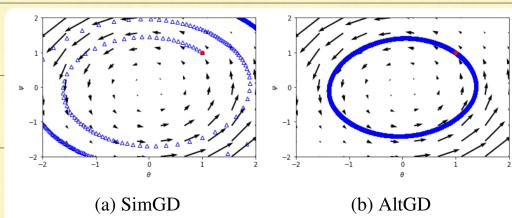
$$v(\theta, \psi) := \begin{pmatrix} -\nabla_{\theta} L(\theta, \psi) \\ \nabla_{\psi} L(\theta, \psi) \end{pmatrix}. \quad (2)$$

그리고

update operator $F_h(\theta, \psi)$ 의 Jacobian은 Nash-equilibrium

$\lambda_{1,2} = 1 \pm h f'(0)$ 의 eigenvalue를 가진다.

이 때의 문제로 EP의 진동 (unstable)



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

gradient vector field의 integral curves

→ gradient의 솔루션 θ, ψ 의 global convergence x

Figure 2. Training behavior of the Dirac-GAN. The starting iterate is marked in red.

- Lemma 2.5.

Alternating GD $F_h = F_{2,h} \circ F_{1,h}$ 를 미는

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

$$\alpha = \sqrt{n_g n_d} h f'(0) \underbrace{\text{update}}$$

$\alpha < 1$ 면 모든 eigen 은 unit

$\alpha > 1$ 면 out of circle.

→ ALEGD 는 sublinear 로 수렴.

하지만 대체로 convergence rate에서 stable한 전동. (그림 2.b)

- Where do instabilities come from?

위에서 기본적인 GAN의 학습이 항상 수렴화진 않는다는 것을 보임.

목적이 잘못되게 만드는지, 간단한 예제에서 분석

∴ 전동 등각을 Dirac-GAN 이나, 조금 복잡한 GAN이나 살펴보

Fig. 1 아래 칙면적으로 gen이 true이 떨면, true 쪽으로 빠져나가고

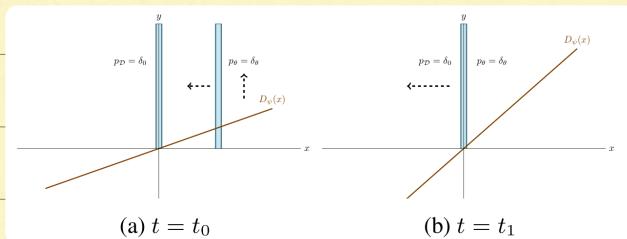
동시에 D는 slope이 커져, 공학화된다.

gen이 true이 되었으면, D의 slope이 가장 커지고,

G를 true로 만든다.

설명적으로 gen은 true로 떨어지면, D의 slope는 negative로 바뀐다.

→ 전동.



다른 방법에서 Nash 균형에서 훈련의 local behavior를 고려한다.

실际로, Nash 균형에서, true에서 slope이 0이 되도록 D를 미는 것은 없다.

(\Rightarrow) true가 경향적 초기화 되어도, D가 평행임으로 이동할 근거가 없다.

→ 평행임에서 불균형

true의 적과는 D의 gradient 흐름은 더 복잡한 곳에서도 발생.

(\Rightarrow) true에서도 D가 data manifold의 경원에 적과는 기울기를 생생으로 incentive \rightarrow 없다.

↳ 단순화 gradient가 아니다.

↳ (\Rightarrow) 초기화를 방지.

\therefore manifold가 적과다면 초기화 \times

→ true \Rightarrow 낮은 dim의 manifold의 잡동선 경우 발생

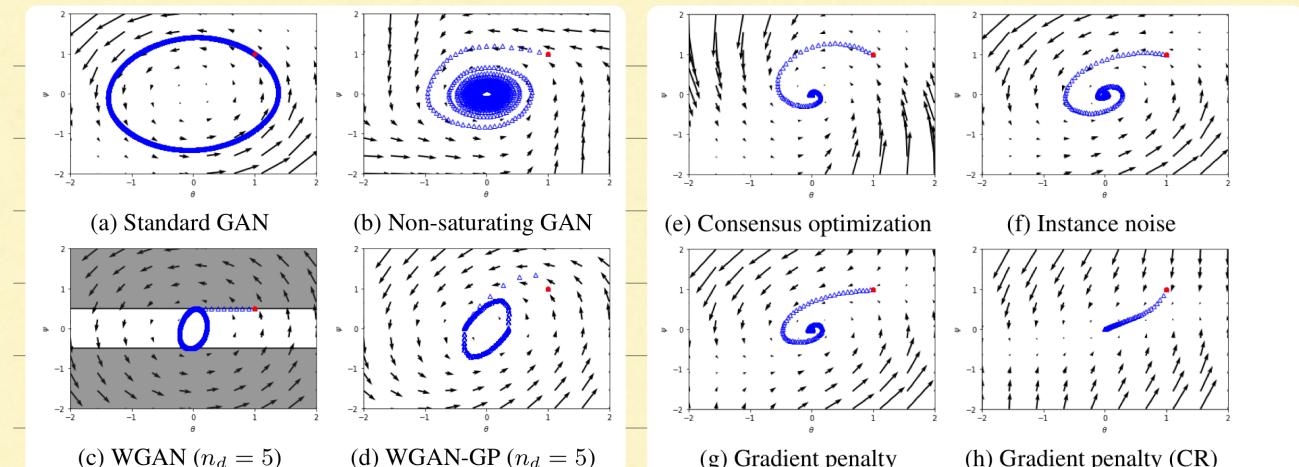


Figure 3. Convergence properties of different GAN training algorithms using alternating gradient descent with recommended number of discriminator updates per generator update ($n_d = 1$ if not noted otherwise). The shaded area in Figure 3c visualizes the set of forbidden values for the discriminator parameter ψ . The starting iterate is marked in red.

3. Regularization strategies

Unregularized GAN은 항상 Nash로 수렴하지는 않는다.

Vanilla GAN의 paper에서 언급된 non-saturated GAN은 느리지만 수렴하는 흐름

- WGAN.

GAN은 Dominate. G가 대체 best-response를 갖는 것.

data의 manifold는 전체 space의 M 보면, O이 가까운 작은 space의 M에

true의 generative distribution이 빠지지 않는 수 있는 이는 gradient divergence를 사용한다.

따라서 manifold를 정의하기 힘들 때 noise를 추가하기도 한다. (but, hyperparam)

이런 divergence를 해결하기 위해, WGAN은 JSD를 Wasserstein-Divergence와 같이 사용.

WGAN은 G가 업데이트될 때마다 D의 D의 업데이트되는 것이 아니라.

그리고 Nash 균형은 G의 학습으로써 D의 param의 discontinuous 할 수 있다.

그리고 equilibrium 균형의 gradient가 정지하지 update x.

- Lemma 3.1.

G update 및 D의 update를 하는 WGAN은 Dirac-GAN이나 일반적으로 수렴 x

TTUR을 사용하면, convergence 가능

- Instance noise.

Data의 noise를 넣어 분포를 정하기에 차리는 방식.

→ training algorithm의 convergence의 단계는 명확x.

↳ continuous의 local convergence

Dirac-GAN의 단계.

- Lemma 3.1.

noise를 넣으면, gradient vector field의 Jacobian의 eigen은

$$\lambda_{1/2} = \underbrace{f''(0)\sigma^2}_{<0 \text{ 이면}} \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}. \quad (6)$$

↳ 0 이면, Nash에서 모두 negative real-part.

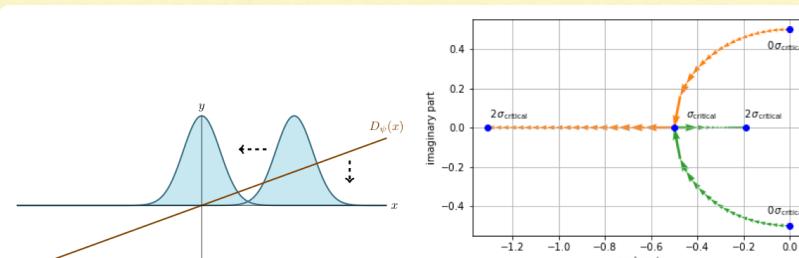
→ local convergence.

임계 noise level $\sigma_{\text{critical}}^2 = |f'(0)| / |f''(0)|$

임계 level 보다 낮으면 imaginary axis 발생. → 터전 성분

크면, 실수.

→ Critical 일 때가 제일 좋다.



(a) Example with instance noise

(b) Eigenvalues

Figure 4. Dirac-GAN with instance noise. While unregularized GAN training is inherently unstable, instance noise can stabilize it: (a) Near the Nash-equilibrium, the discriminator is pushed towards the zero discriminator. (b) As we increase the noise level σ from 0 to σ_{critical} , the real part of the eigenvalues at the equilibrium point becomes negative and the absolute value of the imaginary part becomes smaller. For noise levels bigger than σ_{critical} all eigenvalues are real-valued and GAN training hence behaves like a normal optimization problem.

- Zero-centered gradient penalties

두 distribution 사이의 f-divergence를 잘 정의하는 noise의 영감을 받아,

판별자와 DNN Zero-centered gradient penalty를 도출하는

instance noise의 local approximation을 유도.

Dirac-GAN의 Del gradient의 L-norm regularization 결과 도출

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

- Lemma 3.3.

EP의 gradient-regularized Dirac-GAN의 경우.

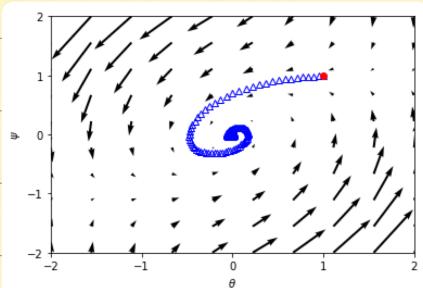
gradient vector field의 Jacobian의 eigen은

$$\lambda_{1/2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - f'(0)^2}. \quad (8) \rightarrow \text{fig 3.g.}$$

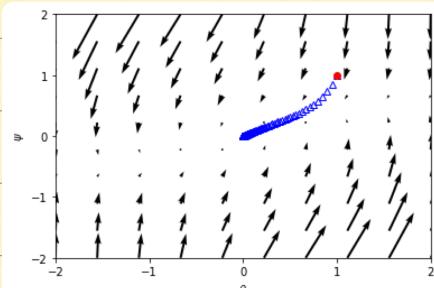
$\gamma > 0$ 일 때, negative real-part : locally convergence.

$$\gamma_{\text{critical}} = 2|f'(0)| \rightarrow \text{fig 3.h.}$$

↳ rotation component를 갖지 않음



(g) Gradient penalty



(h) Gradient penalty (CR)

4. General convergence results.

Sec. 3에서 reg 방법들의 convergence를 조사.

여기서 f-divergence를 간단히 소개 후, 이런 gradient penalty가

general true \Rightarrow locally same support \Rightarrow otherwise convergence 하는 것으로 확장.

↳ low-dim manifold 같은 현실적인 경우에도 적용됨

- Simplified gradient penalty

local stability의 zero-centered gradient penalty는

Nash equilibrium을 갖는 D가 penalty로 한다.

\rightarrow real의 대체로만 적용

$G \Rightarrow$ true를 생성하고, $D \Rightarrow$ data manifold의 0일 때,

gradient penalty는 $D \Rightarrow$ 사실상 data manifold의 직교하는 non-zero gradient를 만들지 못하기 때문

$$R_1(\psi) := \frac{\gamma}{2} \mathbb{E}_{p_D(x)} [\|\nabla D_\psi(x)\|^2]. \quad (9)$$

나 추가 가정은 x. data distribution에 영향.

$$R_2(\theta, \psi) := \frac{\gamma}{2} \mathbb{E}_{p_\theta(x)} [\|\nabla D_\psi(x)\|^2] \quad (10)$$

나 실제가 아닌 생성된 것이 문제...

- Convergence.

R_1, R_2 에 대한 reg-GAN 결과

(θ^*, ψ^*) 가 EP일 때,

Assumption 1.

$\text{Supp } P_D$ 의 어떤 local neighbor Ω 에서 $P_{\theta^*} = P_D$ 이고 $D_{\psi^*}(x) = 0$ 이다.

Assumption 2.

$f'(0) \neq 0, f''(0) < 0$ 이면.

NN의 경우 (θ^*, ψ^*) 가 빠져 있는 것도 아니기 때문에 convergence를 증명하기가 어려움

∴ reparameterization manifold M_G, M_D 정의.

$$h(\psi) := E_{p_D(x)} [|D_\psi(x)|^2 + \|\nabla_x D_\psi(x)\|^2]. \quad (11) \rightarrow \text{real의 대상 D loss}$$

$$\mathcal{M}_G := \{\theta \mid p_\theta = p_D\} \quad \mathcal{M}_D := \{\psi \mid h(\psi) = 0\}. \quad (12)$$

$$g(\theta) := E_{p_\theta(x)} [\nabla_\psi D_\psi(x)|_{\psi=\psi^*}]. \quad (13) \rightarrow \text{D의 기울기 (gen의 대상)}$$

Assumption 3.

θ^*, ψ^* 이 ϵ -ball의 $B_\epsilon(\theta^*), B_\epsilon(\psi^*)$ 일 때.

- (i) if $v \in \mathbb{R}^n$ is not in the tangent space of \mathcal{M}_D at ψ^* ,
then $\partial_v^2 h(\psi^*) \neq 0$.
- (ii) if $w \in \mathbb{R}^m$ is not in the tangent space of \mathcal{M}_G at θ^* ,
then $\partial_w g(\theta^*) \neq 0$.

M_G 는 gen의 분포가 true와 같을 때.

M_D 는 $h(D \text{ loss})$ 가 0일 때.

i) M_D 가 극복적으로 가능될 수 있다. ($loss$ function이 0이 되는 경우)

ii) D가 equil 상태의 G에도 학습 가능할 정도로 강력.

정규화된 gradient field는

$$\tilde{v}_i(\theta, \psi) := \begin{pmatrix} -\nabla_{\theta} L(\theta, \psi) \\ \nabla_{\psi} L(\theta, \psi) - \nabla_{\psi} R_i(\theta, \psi) \end{pmatrix}. \quad (14)$$

Theorem 4.1.

(θ^*, ψ^*) 이 대체 assumption 1, 2, 3을 가정하면,

충분히 작은 lr에서, \tilde{v}_i, \tilde{v}_o 에 대한 gradient descent는

(θ^*, ψ^*) 의 neighborhood에서, $M_G \times M_D$ 에 수렴

수렴 속도는 linear

→ 적어도 equilibrium에서는 잘 작동함

- Stable equilibria for unregularized GAN training

섹션 E에서 일반적인 GAN에서 안정적인 GAN 형태를 살펴 → Appendix E.

→ energy, full-rank.

하지만 고차원에서는 아직 명확하지 않음

5. Experiments.

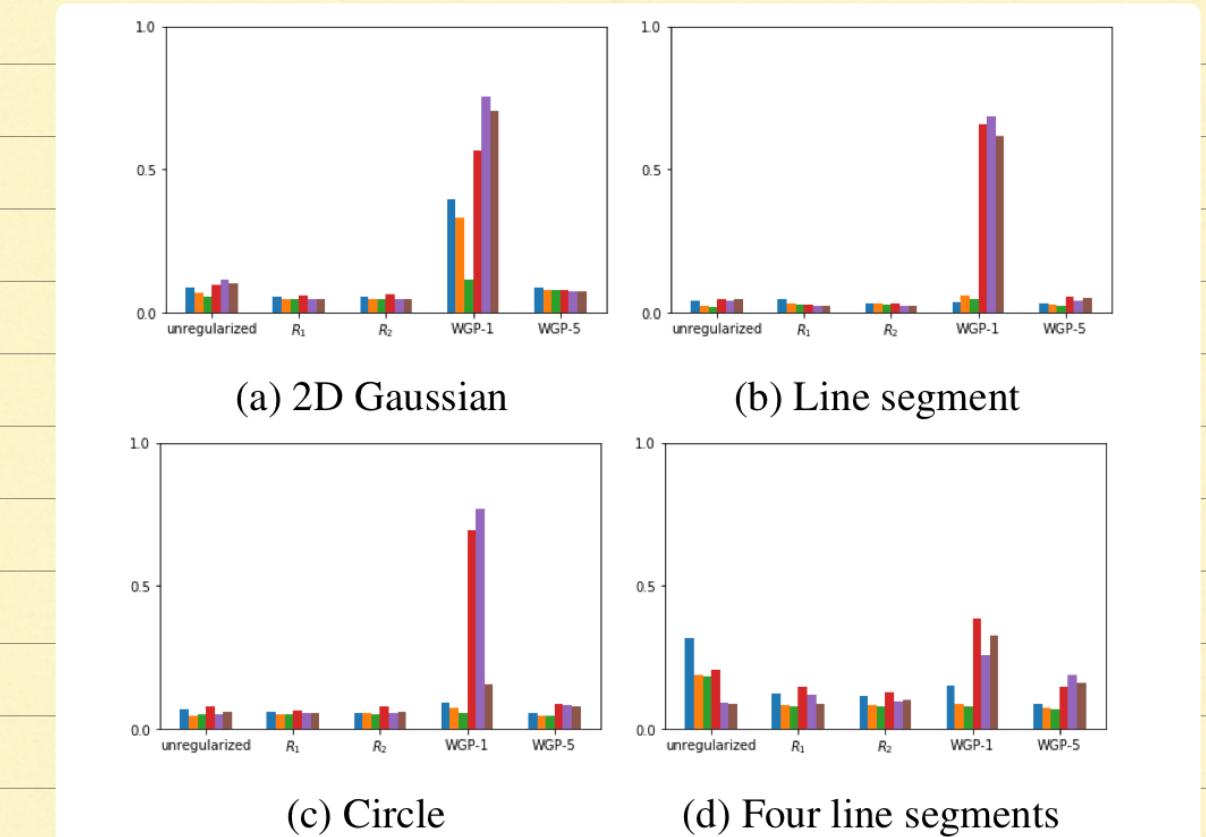


Figure 5. Wasserstein-1-distance to true data distribution for 4 different 2D-data-distributions, 6 different architectures (small bars) and 5 different training methods. Here, we abbreviate WGAN-GP with 1 and 5 discriminator update(s) per generator update as WGP-1 and WGP-5.

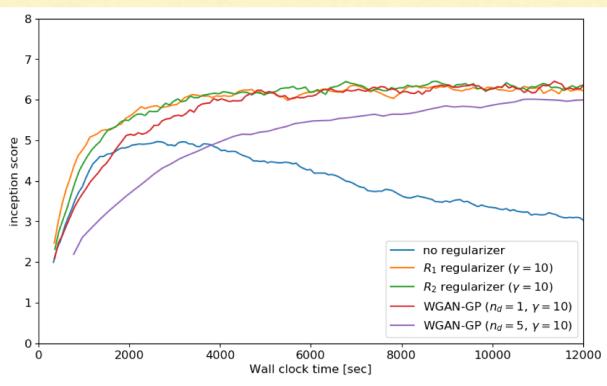


Figure 6. Inception score over time for various regularization strategies when training on CIFAR-10. While the inception score can be problematic for evaluating probabilistic models (Barratt & Sharma, 2018), it still gives a rough idea about the convergence and stability properties of different training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

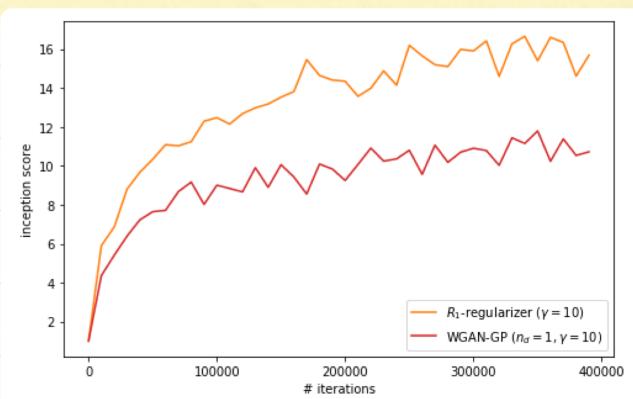


Figure 7. Inception score over the number of iterations for GAN training with R_1 - and WGAN-GP-regularization when training on Imagenet. We find that R_1 -regularization leads to higher inception scores for this dataset and GAN-architecture.

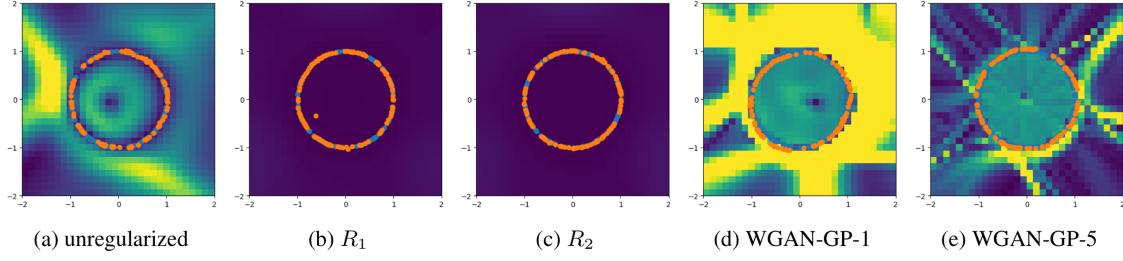


Figure 9. Best solutions found by the different algorithms for learning a circle. The blue points are samples from the true data distribution, the orange points are samples from the generator distribution. The colored areas visualize the gradient magnitude of the equilibrium discriminator. We find that 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).

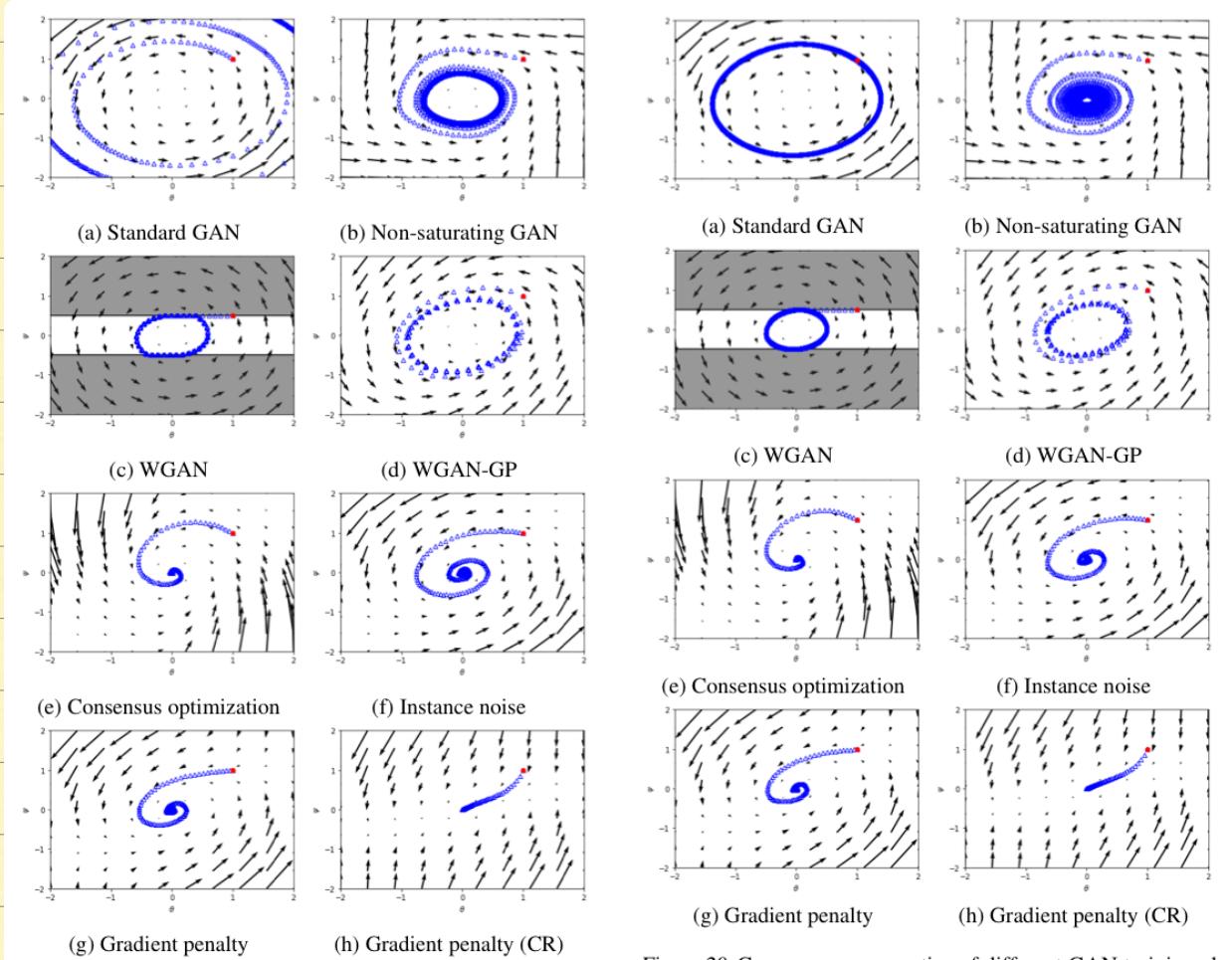


Figure 19. Convergence properties of different GAN training algorithms using simultaneous gradient descent. The shaded area in Figure 19c visualizes the set of forbidden values for the discriminator parameter ψ . The starting iterate is marked in red.

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.