Convergence Analysis of GAN training procedures

Reza Pishkoo   Kasra Khoshjoo

Sharif University of Technology

presentation for HDP course
July 8, 2023

Presentation Overview

Generative Adversarial Networks

- latent variable: $Z \sim P_Z(z)$
- generator : $G_\theta(z)$
- discriminator : $D_\psi(x)$
- target function:

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{P_Z}\left[f(D_\psi(G_\theta(Z)))\right] + \mathbb{E}_{X \sim P_D}\left[f(-D_\psi(X))\right]$$

- zero-sum game

$$\min_\theta \max_\psi \mathcal{L}(\theta, \psi)$$

Iteration Procedures

$$(\theta, \psi)^{(k)} = F(\theta^{(k-1)}, \psi^{(k-1)})$$

- simultaneous gradient descent : $F_h(\theta, \psi) = (\theta, \psi) + hv(\theta, \psi)$ where $v(\theta, \psi) = \begin{bmatrix} -\nabla_\theta \mathcal{L} \\ \nabla_\psi \mathcal{L} \end{bmatrix}$
- alternating gradient descent : $F_h = F_{2,h} \circ F_{1,h}$
- continuous(ODE solver) : $\frac{d(\theta,\psi)}{dt} = v(\theta, \psi)$

Uniqueness of Nash-Equilibrium

Theorem (unique Nash-equilibrium [Goodfellow, 2014])

given that G_θ, D_ψ are powerful enough to approximate any real valued function, there is a unique Nash-equilibrium where :

$$G_\theta(Z) \sim P_D$$

Convergence Theory [Mescheder, 2017]

Theorem (sufficient condition for local convergence [Mescheder, 2017])

given iteration function F at point $\bar{x}$:

- *is stationary, i.e. $F(\bar{x}) = \bar{x}$*
- *has a Jacobian matrix with eigenvalues smaller than. i.e. $|\lambda_i(F'(\bar{x}))| < 1$*

for every initial point x_0 in some neighbourhood of $\bar{x}$, the sequence $F^{(k)}(x_0)$ converges to $\bar{x}$. i.e. $\|F^{(k)}(x_0) - \bar{x}\| \in O(|\lambda_{max}(\bar{x})|^k)$

in practice we have :

$$F'(\theta, \psi) = I + hv'(\theta, \psi)$$

proposition (sufficient condition for convergence [Mescheder, 2017])

there exists $h > 0$ where eigenvalues of $I + hv'(\theta, \psi)$ are within the unit ball if the real part of all eigenvalues of $v'(\theta, \psi)$ are negative.

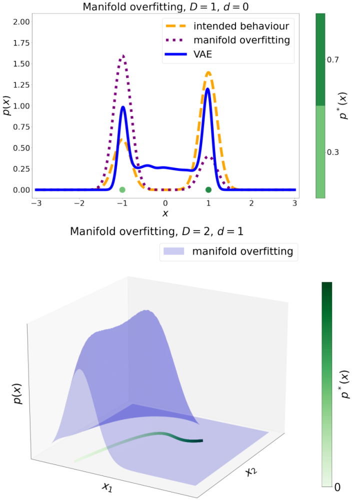
Which Training Methods for GANs do actually Converge?[Mescheder, 2018]

The Manifold Hypothesis

Example (low dimensional dist. [Arjovsky, 2014])

divergance between the distribution of $(0, Z)$ and (θ, Z) where $Z \sim Uni(0,1)$ is :

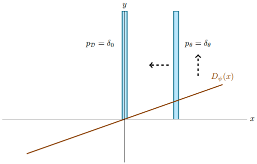
- $W_1(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$
- $\mathrm{JS}(\mathbb{P}_0, \mathbb{P}_\theta) = \log 2 \cdot \mathbf{1}_{\{\theta \neq 0\}}$
- $D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_\theta) = \mathbb{I}_{\{0\}}(\theta)$



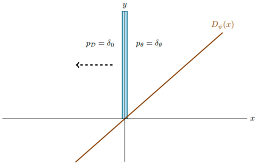
a toy data distribution

Definition (Dirac-GAN [Mescheder, 2018])

The Dirac-GAN consists of a (univariate) generator distribution $p_\theta = \delta_\theta$ and a linear discriminator $D(x) = \psi \cdot x$. The true data distribution P_D is given by a Dirac-distribution concentrated at 0.



(a) $t = t_0$



(b) $t = t_1$

a toy data distribution

the target function of Dirac-GAN is:

$$L(\theta, \psi) = f(\psi \cdot \theta) + f(0) \quad ; \quad f(x) = -\log(1 + \exp(-x))$$

the unique Nash-equilibrium is at $\theta = \psi = 0$.
the eigenvalues of Jacobian matrix of $v(\theta, \psi)$ are $\pm f'(0)i$, real values of which are zero.

a toy data distribution

Lemma (integral curves do not converge)

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 $\theta(t)^2 + \psi(t)^2 = \text{const}$ for all $t \in [0, \infty)$.

Lemma (simultaneous gradient does not converge)

For simultaneous gradient descent, the Jacobian of the update operator $F_h(\theta, \psi)$ has eigenvalues $\lambda_{\{1,2\}} = 1 \pm hf'(0)i$ 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.

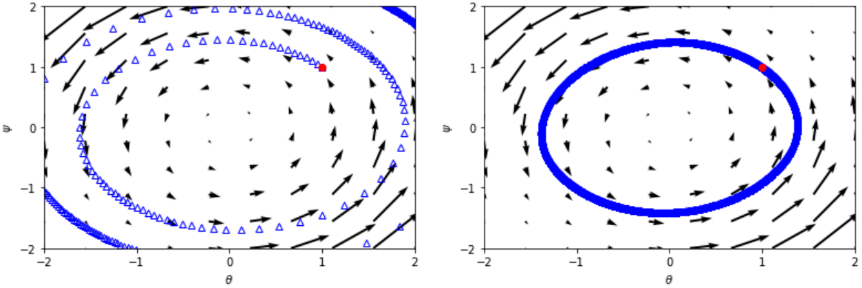
Lemma (alternating gradient descent does not converge)

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

$$\lambda_{1,2} = 1 - \frac{\alpha^2}{2} \pm \sqrt{\left(1 - \frac{\alpha^2}{2}\right)^2 - 1} \quad ; \quad \alpha = \sqrt{n_g n_d} h f'(0)$$

which implies that eigenvalues cannot be strictly inside the unit circle.

a toy data distribution



(a) SimGD (b) AltGD

Theorem (Convergence for continuous distributions[Nagarajan, 2017])

under some suitable assumptions - gradient descent based GAN optimization is locally convergent for absolutely continuous distributions

Creating convergent procedures

Additive Noise

- adding Independent gaussian noise to data points will make both distributions continuous, thus, making the gradient descent convergent.
- in the Dirac-GAN case:

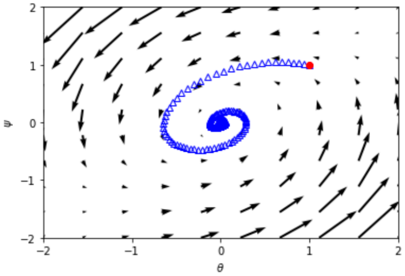
Lemma

When using Gaussian instance noise with standard deviation σ, the eigenvalues of the Jacobian of thegradient vector field are given by

$$\lambda_{1,2} = f^{''}(0)\sigma^2 \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}$$

- if $f^{''}(0) < 0$ both eigenvalues have negative real parts which means additive noise leads to convergence in the Dirac-GAN.

Additive Noise



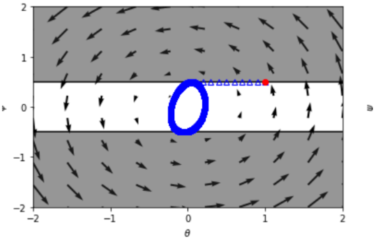
(f) Instance noise

Wasserstein GAN

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{P_Z}\left[D_\psi(G_\theta(Z))\right] - \mathbb{E}_{X \sim P_D}\left[D_\psi(X)\right]; \quad D_\psi \text{ is 1-Lipschitz}$$

- To make the divergence continuous with respect to the parameters of the generator, Wasserstein GANs (WGANs) replace the Jensen-Shannon divergence used in the original derivation of GANs with the Wasserstein-divergence
- alternating and simultaneous gradient descent for Wasserstein divergance does not converge in the Dirac-GAN problem.

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



(c) WGAN ($n_d = 5$)

regularizing the target function for Wasserstein GAN can make it convergent(Wasserstein GAN with Quadratic Transport Cost [Liu, 2019]).

Regularizing Wasserstein GAN [Liu, 2019]

- consider the linear program:

$$\begin{cases} \max\limits_{H_i, H_j} & \frac{1}{m}\sum H_j - \frac{1}{m}\sum H_i \\ \text{subject to} & |H_i - H_j| \leq \frac{1}{2}\|x_i - y_j\|^2 \end{cases}$$

- consider the optimal transport problem:

$$\sigma(j) = \operatorname*{argmin}_i \ \frac{1}{2}\|x_i - y_j\|^2 + H_i^* - H_j^*$$

- define losses:

$$\min_\psi \frac{1}{2}\left(\frac{1}{m}\sum D_\psi(y_j) - \frac{1}{n}H_j^*\right)^2 + \frac{1}{2}\cdot\frac{1}{n}\sum\left(D_\psi(x_i) - H_i^*\right)^2$$
$$\underbrace{+\frac{\lambda}{2}\mathbb{E}_{X\sim G_\theta(Z)}\left[\left(\|\nabla_x D_\psi(X)\| - \|y_{\sigma^*(X)} - X\|\right)^2\right]}_{\text{gradient regulrization}}$$

$$\min_\theta \mathcal{L}(\theta) = -\frac{1}{n}\sum D_\psi(G_\theta(Z_i))$$

Regularizing Wasserstein GAN [Liu, 2019]

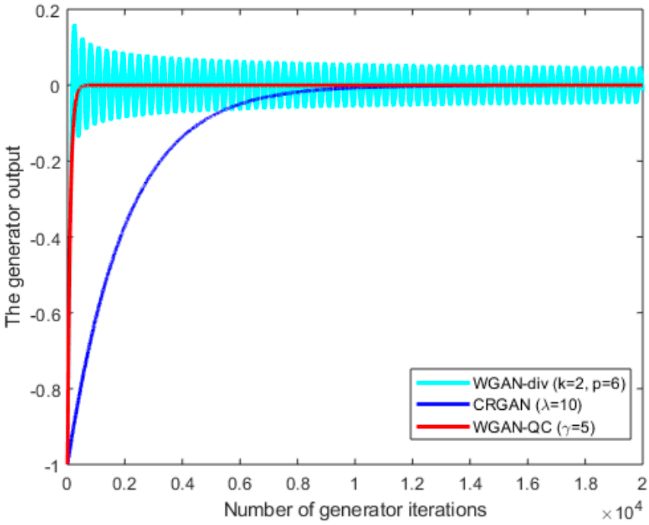
nframe it can be shown that :

$$-v^{'}(\theta^*, \psi^*) = \begin{bmatrix} M_{DD} + M_R & M_{GD} \\ 0 & 0 \end{bmatrix}$$

where :

$$\begin{aligned} M_R &= \lambda \cdot \mathbb{E}_{P_Z}\left[\nabla_{\theta,\psi} D_{\psi^*}(G_\theta(Z)) \nabla_{\theta,\psi} D_{\psi^*}(G_\theta(Z))^T\right] \\ M_{DD} &= \mathbb{E}_{y \sim P_D}\left[\nabla_\psi D_{\psi^*}(y)\right] \mathbb{E}_{y \sim P_D}\left[\nabla_\psi D_{\psi^*}(y)\right]^T \\ &\quad + \mathbb{E}_{y \sim P_D}\left[\nabla_\psi D_{\psi^*}(y) \nabla_\psi D_{\psi^*}(y)^T\right] \\ M_{GD} &= -\mathbb{E}_{Z \sim P_Z}\left[\nabla^2 D_{\psi^*}(G_\theta(Z)) \nabla_\theta G_{\theta^*}(Z)^T\right] \end{aligned}$$

Convergence WGAN-QC in Dirac-GAN



References I

Mescheder et al.(2017)
The Numerics of GANs
NIPS

Mescheder et al.(2018)
Which Training Methods for GANs do actually Converge?
ICML

Goodfellow et al.(2014)
Generative Adversarial Networks
NIPS

Arjovsky et al.(2014)
Wasserstein GAN
PMLR

Nagarajan et al.(2017)
Gradient descent GAN optimization is locally stable.
NIPS

Liu et al.(2019)
Wasserstein GAN with Quadratic Transport Cost
CVF