A NEURAL TANGENT KERNEL PERSPECTIVE OF GANS

ICML 2022 – July 17th to 23rd, 2022

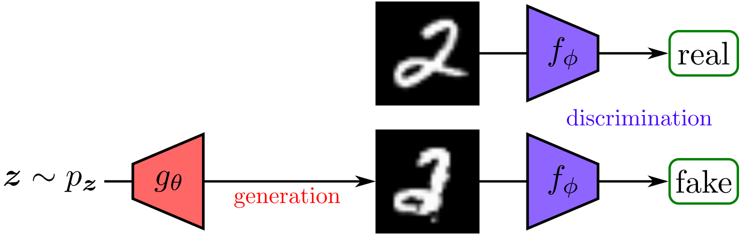
***J.-Y. Franceschi*,[1,2] *E. de Bézenac*,[3,2] I. Ayed,[2,4]**
M. Chen,[5] S. Lamprier,[2] P. Gallinari[2,1]

[1]Criteo AI Lab, Paris, France
[2]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
[3]Seminar for Applied Mathematics, D-MATH, ETH Zürich, Zürich-8092, Switzerland
[4]ThereSIS Lab, Thales, Palaiseau, France [5]Valeo.ai, Paris, France

We solve fundamental flaws of GAN analyses via a theoretical framework based on NTKs.

Principle

- The generator g_θ generates a distribution α_θ, with target β.
- g_θ is trained in competition with a discriminator f_ϕ.
- g_θ and f_ϕ have conflicting objectives:
  - f aims at distinguishing between fake and target samples;
  - g should make fake and target samples indistinguishable for f.



- This is typically framed as, for some loss $\mathcal{L}$:

$$\inf_{\theta} \sup_{\phi} \mathcal{L}\big(g_\theta, f_\phi\big).$$

Textbook GANs: Min-Max

- This is typically framed as, for some loss $\mathcal{L}$:

$$\inf_{\theta} \sup_{\phi} \mathcal{L}(g_\theta, f_\phi).$$

- Many analyses solve the inner optimization problem and find that for some loss $\mathscr{C}$ and optimal $f_{\phi^\star_\theta}$:

$$\inf_{\theta} \sup_{\phi} \mathcal{L}(g_\theta, f_\phi) = \inf_{\theta} \mathcal{L}\left(g_\theta, f_{\phi^\star_\theta}\right) \approx \inf_{\theta} \mathscr{C}(\alpha_\theta, \beta).$$

  - In vanilla GAN, $\mathscr{C}$ is a Jensen-Shannon (JS) divergence.
  - In WGAN, $\mathscr{C}$ is the earth mover's distance $\mathcal{W}_1$.

SCIENCES SORBONNE UNIVERSITÉ

- This is typically framed as, for some loss $\mathcal{L}$:

$$\inf_\theta \sup_\phi \mathcal{L}(g_\theta, f_\phi).$$

- Many analyses solve the inner optimization problem and find that for some loss $\mathscr{C}$ and optimal $f_{\phi^\star_\theta}$:

$$\inf_\theta \sup_\phi \mathcal{L}(g_\theta, f_\phi) = \inf_\theta \mathcal{L}\left(g_\theta, f_{\phi^\star_\theta}\right) \approx \inf_\theta \mathscr{C}(\alpha_\theta, \beta).$$

  - In vanilla GAN, $\mathscr{C}$ is a Jensen-Shannon (JS) divergence.
  - In WGAN, $\mathscr{C}$ is the earth mover's distance $\mathcal{W}_1$.
- Gradient received by g_θ:

$$\nabla_\theta \mathcal{L}\left(g_\theta, f_{\phi^\star_\theta}\right).$$

GANs in Practice: Alternating Optimization

- In practice, GANs are iteratively optimized as follows:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(g_\theta, f_\phi);$$
$$\phi \leftarrow \phi + \lambda \nabla_\phi \mathcal{L}(g_\theta, f_\phi).$$

- f_ϕ and g_θ are considered to be independent of each other.

GANs in Practice: Alternating Optimization

- In practice, GANs are iteratively optimized as follows:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(g_\theta, f_\phi);$$
$$\phi \leftarrow \phi + \lambda \nabla_\phi \mathcal{L}(g_\theta, f_\phi).$$

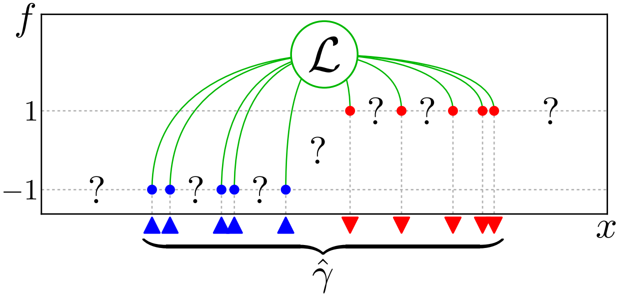
- f_ϕ and g_θ are considered to be independent of each other.
- Gradient received by g_θ:

$$\cancel{\nabla_\theta \mathcal{L}\left(g_\theta, f_{\phi^\star_\theta}\right)} \quad \Rightarrow \quad \nabla_\theta \mathcal{L}(g_\theta, f_\phi).$$

Consequence

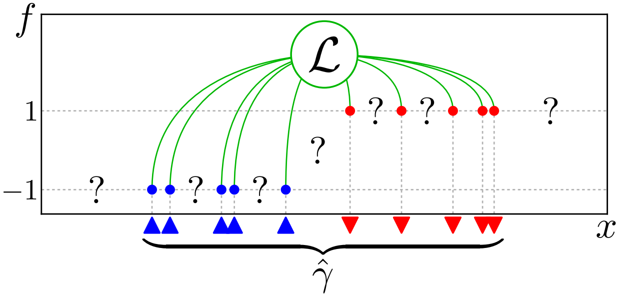
Altering the gradient changes the loss $\mathscr{C}$ minimized by the generator.

SCIENCES SORBONNE UNIVERSITÉ



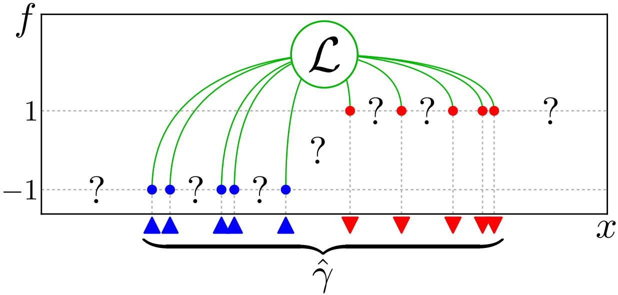
In an Alternating Optimization setting:

- Computing gradient of generator requires ∇f (chain rule).



In an Alternating Optimization setting:

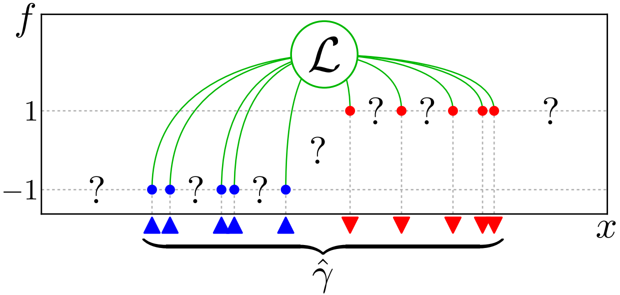
- Computing gradient of generator requires ∇f (chain rule).
- Without any assumption on the structure of f, as loss $\mathcal{L}$ is only defined on training points, ∇f is not defined.



In an Alternating Optimization setting:

- Computing gradient of generator requires ∇f (chain rule).
- Without any assumption on the structure of f, as loss $\mathcal{L}$ is only defined on training points, ∇f is not defined.
- The gradient of the generator is thus also ill-defined.

Ill-defined Gradient Issues



In an Alternating Optimization setting:

- Computing gradient of generator requires ∇f (chain rule).
- Without any assumption on the structure of f, as loss $\mathcal{L}$ is only defined on training points, ∇f is not defined.
- The gradient of the generator is thus also ill-defined.
- *Need to take into account structure of f.*

SCIENCES SORBONNE UNIVERSITÉ

Problem

Most prior analyses fail to model practical GAN settings, leading to:

- be unable to determine the true loss $\mathscr{C}$;
- ill-defined gradient issues.

Our Work

We propose a *finer-grained* framework solving these issues, modeling the discriminator's architecture along with alternating optimization.

SCIENCES SORBONNE UNIVERSITÉ

Infinite-Width NTK Framework

- We consider the NNs in the NTK regime (Jacot et al., 2018).
- Allows theoretical analysis of evolution of NNs during training.

Infinite-Width NTK Framework

- We consider the NNs in the NTK regime (Jacot et al., 2018).
- Allows theoretical analysis of evolution of NNs during training.

Theorem (Smoothness of the discriminator, Informal)

The discriminator trained with gradient descent is infinitely differentiable (almost) everywhere.

- Gradients of both the discriminator and generator well defined.

We analyze evolution of generated distribution α_θ during training:

- Follows *Stein gradient flow* w.r.t. loss $\mathscr{C}$ (Duncan et al., 2019);
- $\mathscr{C}$ is automatically non-increasing during adversarial training;
- $\mathscr{C}$ can be analyzed theoretically; in particular:

We analyze evolution of generated distribution α_θ during training:

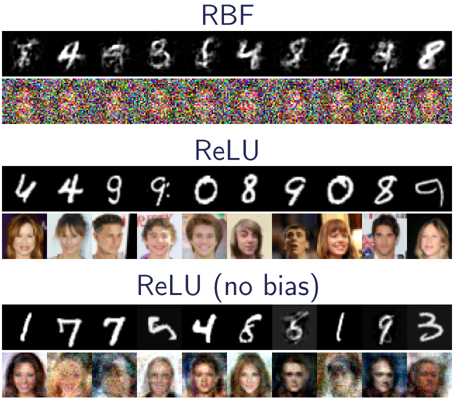
- Follows *Stein gradient flow* w.r.t. loss $\mathscr{C}$ (Duncan et al., 2019);
- $\mathscr{C}$ is automatically non-increasing during adversarial training;
- $\mathscr{C}$ can be analyzed theoretically; in particular:

GAN Loss for IPMs

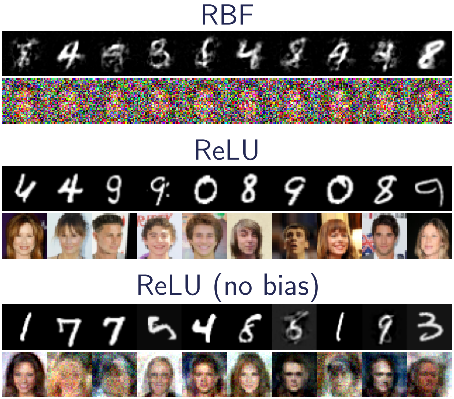
For the IPM loss, $\mathscr{C}$ is the squared MMD with the NTK as kernel:

$$\mathscr{C}(\alpha_\theta, \beta) = \mathrm{MMD}_k^2(\alpha_\theta, \beta).$$

- More results of this type in the paper!



- We conduct empirical analysis,
- Yields insights into GAN training,



- We conduct empirical analysis,
- Yields insights into GAN training,

Experimental Framework

Code: `https://github.com/emited/gantk2`.