A Neural Tangent Kernel Perspective of GANs
Jean-Yves Franceschi, Emmanuel de Bézenac, Ibrahim Ayed, Mickaël Chen, Sylvain Lamprier, Patrick Gallinari

To cite this version:
Jean-Yves Franceschi, Emmanuel de Bézenac, Ibrahim Ayed, Mickaël Chen, Sylvain Lamprier, et al.. A Neural Tangent Kernel Perspective of GANs. Thirty-ninth International Conference on Machine Learning, Jul 2022, Baltimore, MD, United States. hal-03716574

HAL Id: hal-03716574
https://hal.science/hal-03716574v1
Submitted on 7 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License
A Neural Tangent Kernel Perspective of GANs

J.-Y. Franceschi,*1,2 E. de Bèzenac,*3,2 I. Ayed,*2,4 M. Chen,5 S. Lamprici,² P. Gallinari²,1

Criteo AI Lab Sorbonne Université, CNRS, ISIR, SAM, D-MATH, ETH Zürich
Thales SfS Lab, Thales

Many analyses cannot explain GAN training as they fail to take into account alternating optimization and the architecture and implicit biases of the discriminator.

We propose a theoretical framework solving these issues using the theory of Neural Tangent Kernels.

We deduce new insights about the flow and convergence of the generated distribution during training.

**Generative Adversarial Networks**

**A Background on NTKs**

The Neural Tangent Kernel: For a neural network \( f \) with parameters \( \theta \), its NTK \( k_{\theta}(x,y) \) is defined as:

\[
k_{\theta}(x,y) = \nabla_{\theta} f(x) \cdot \nabla_{\theta} f(y).
\]

In the infinite-width limit of \( f \), during training:

\[
b_{\theta}(x,y) = \nabla_{\theta} f(x) \cdot \nabla_{\theta} f(y).
\]

The Kernel Integral Operator and RKHS:

\[
T_{\alpha} : L^2(\mathbb{R}) \rightarrow H, \quad \alpha \mapsto \left\{ \int x \cdot \phi_{\alpha}(x) \, dx \right\}
\]

where \( H \subset L^2(\mathbb{R}) \) is the RKHS of kernel generated by \( \gamma \).

**Discriminator Inner Loop**

We consider the NNs in the NTK regime. This enables a theoretical study of their evolution w.r.t. training time \( \tau \):

\[
\frac{\partial f}{\partial \tau} = T_{\alpha}(\nabla \mathcal{L}(f, g)).
\]

**Discriminator Structure**

Under mild assumptions, \( f \) is uniquely defined and:

\[
\forall t \in [0, \tau], \quad f_t = f_0 + T_{\alpha}(\int \nabla \mathcal{L}(f, g) \, dx) = f_0 + T_{\alpha} L_{\tau}.
\]

\( T_{\alpha} L_{\tau} \) smooths out gradients over the whole input space by sending them into \( H_{\alpha} \).

\( H_{\alpha} \) depends on discriminator architecture.

**Differentiability of the Discriminator:**

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

The spatial gradient of the discriminator \( \nabla f; \alpha \) is well-defined.

Underlying NTK Regularity Results

To prove the above results, we establish novel regularity results on NTKs. Given, for the network \( f \):

- a standard architecture (fully connected, convolutional, residual, etc.);
- a standard activation function (tanh, softplus, ReLU, LeakyReLU, sigmoid, Gaussian, etc.).

We prove that the NTK \( k \) is:

- smooth almost everywhere if the network has non-null bias terms.
- smooth everywhere if the activation is smooth.

These results, obtained from similar regularity results on the conjugate kernel \( f^* \), then transfer to \( f \).

**Resulting Convergence Results**

Our finer-grained framework allows us to derive novel convergence insights, with highlighted results below.

**Gradient Flow of Generated Distribution:**

\[
\partial_{\tau} g = -\nabla_{\tau} \mathcal{L}(g),
\]

\[
-\nabla_{\tau} \mathcal{L}(g) = \frac{1}{\lambda} T_{\alpha}(g) \cdot (z - \nabla_{\tau} \mathcal{L}(g).)
\]

In the non-interacting case, i.e. \( T_{\alpha} = I \), this corresponds to a Wasserstein gradient flow:

\[
\partial_{\tau} g = -\nabla_{\tau} \mathcal{L}(g),
\]

In the general case, this is a gradient flow in a Stein geometry defined by the generator's NTK \( k_{\theta} \).

\( \mathcal{L}(g) \) is automatically decreasing via this gradient flow, as fast as possible (locally).

**IPM GANs \((a - b - i)\) and NTK MMD:**

We find \( f_t = f_0 + T_{\alpha}(\int \mathbb{E}_a[\mathcal{L}(x)] - \mathbb{E}_b[\mathcal{L}(x)]) \), hence, if \( f_0 = 0 \):

\[
\mathcal{L}(g) = \mathcal{L}(f_0) \propto MMD_{\alpha}(a, b).
\]

**Empirical Study**

- We assess the adequacy of our framework by observing how close finite and infinite-width regimes are.
- We study the convergence of GANs on empirical distributions in the non-interacting case \( T_{\alpha} = I \).
- We discover the singular performance of ReLU architectures for generative modeling and explain it by studying generator gradients with our framework.