

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. 39th International Conference on Machine Learning, International Machine Learning Society, Jul 2022, Baltimore, MD, United States. pp.6660-6704. hal-03254591v5

HAL Id: hal-03254591
https://hal.science/hal-03254591v5

Submitted on 27 Oct 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 4.0 International License

A Neural Tangent Kernel Perspective of GANs

Jean-Yves Franceschi [*1,2] **Emmanuel de Bézenac** [*3,2] **Ibrahim Ayed** [*2,4] **Mickaël Chen** [5] **Sylvain Lamprier** [2] **Patrick Gallinari** [2,1]

Abstract

We propose a novel theoretical framework of analysis for Generative Adversarial Networks (GANs). We reveal a fundamental flaw of previous analyses which, by incorrectly modeling GANs' training scheme, are subject to ill-defined discriminator gradients. We overcome this issue which impedes a principled study of GAN training, solving it within our framework by taking into account the discriminator's architecture. To this end, we leverage the theory of infinite-width neural networks for the discriminator via its Neural Tangent Kernel. We characterize the trained discriminator for a wide range of losses and establish general differentiability properties of the network. From this, we derive new insights about the convergence of the generated distribution, advancing our understanding of GANs' training dynamics. We empirically corroborate these results via an analysis toolkit based on our framework, unveiling intuitions that are consistent with GAN practice.

1. Introduction

Generative Adversarial Networks (GANs; Goodfellow et al., 2014) have become a canonical approach to generative modeling as they produce realistic samples for numerous data types, with a plethora of variants (Wang et al., 2021). These models are notoriously difficult to train and require extensive hyperparameter tuning (Brock et al., 2019; Karras et al., 2020; Liu et al., 2021). To alleviate these shortcomings, much effort has been put into better understanding their training process, resulting in a vast literature of theoretical analyses. Many study the various GAN models, found to optimize different losses like the Jensen-Shannon (JS) divergence (Goodfellow et al., 2014) and the earth mover's distance $\mathcal{W}_1$ (Arjovsky et al., 2017), to conclude about their comparative advantages. Yet, empirical evaluations (Lucic et al., 2018; Kurach et al., 2019) showed that they can yield approximately the same performance. This indicates that such theoretical works with an exclusive focus on the GAN formulation might not properly model practical settings.

Importantly, GANs are trained in practice with alternating gradient descent-ascent of the generator and discriminator, which the vast majority of analyses do not model. Yet, this makes GAN training deviate from its formulation in prior works as a min-max problem: the networks are fixed w.r.t. to each other at each step in the former, while they depend on each other in the latter. Therefore, ignoring this ubiquitous procedure prevents those works from adequately explaining GANs' empirical behavior, as it leads to two crucial problems. Firstly, it alters the true implicitly optimized loss, which consequently differs from the widely adopted JS and $\mathcal{W}_1$. Secondly, it compels accurate frameworks to take into account the discriminator parameterization as a neural network with inductive biases influencing the generator's loss landscape, which most previous studies do not, or otherwise be subject to ill-defined discriminator gradients.

To solve these issues, we introduce the first framework of analysis for GANs modeling a wide range of discriminator architectures and GAN formulations, while encompassing alternating optimization. To this end, we leverage advances in deep learning theory driven by Neural Tangent Kernels (NTKs; Jacot et al., 2018) to model discriminator training. We develop theoretical results showing the relevance of our approach: we establish in our framework the differentiability of the discriminator, hence having well-defined gradients, by proving novel regularity results on its NTK.

This more accurate formalization enables us to derive new knowledge about the generator. We formulate the dynamics of the generated distribution via the generator's NTK and link it to gradient flows on probability spaces, thereby helping us to discover its implicitly optimized loss. We deduce in particular that, for GANs under the Integral Probability Metric (IPM), the generated distribution minimizes its Max-

[*]Equal contribution, listed in a randomly chosen order. [1]Criteo AI Lab, Paris, France [2]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France [3]Seminar for Applied Mathematics, D-MATH, ETH Zürich, Rämistrasse 101, Zürich-8092, Switzerland [4]ThereSIS Lab, Thales, Palaiseau, France [5]Valeo.ai, Paris, France. Correspondence to: Jean-Yves Franceschi <jycja.franceschi@criteo.com>, Emmanuel de Bézenac <emmanuel.debezenac@sam.math.ethz.ch>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

imum Mean Discrepancy (MMD) given by the discriminator's NTK w.r.t. the target distribution. Moreover, we release an analysis toolkit based on our framework, GAN(TK)^2, which we use to empirically validate our analysis and gather new empirical insights: for example, we study the singular performance of the ReLU activation in GAN architectures.

2. Related Work

We introduce a framework advancing GAN knowledge, supported by prior and novel contributions in the NTK theory.

Neural Tangent Kernels. NTKs were introduced by Jacot et al. (2018), who showed that a trained neural network in the infinite-width regime equates to a kernel method, thereby making its training dynamics tractable and amenable to theoretical study. This fundamental work has been followed by a thorough line of research generalizing and expanding its initial results (Arora et al., 2019; Bietti & Mairal, 2019; Lee et al., 2019; Liu et al., 2020; Sohl-Dickstein et al., 2020), developing means of computing NTKs (Novak et al., 2020; Yang, 2020), further analyzing these kernels (Fan & Wang, 2020; Bietti & Bach, 2021; Chen & Xu, 2021), studying and leveraging them in practice (Zhou et al., 2019; Arora et al., 2020; Lee et al., 2020; Littwin et al., 2020b; Tancik et al., 2020), and more broadly exploring infinite-width networks (Littwin et al., 2020a; Yang & Hu, 2021; Alemohammad et al., 2021). These prior works validate that NTKs can encapsulate the characteristics of neural network architectures, providing a solid theoretical basis to understand the effect of architecture on learning problems.

GAN theory. A first line of research, started by Goodfellow et al. (2014) and pursued by many others (Nowozin et al., 2016; Zhou et al., 2019; Sun et al., 2020), studies the loss minimized by the generator. Assuming that the discriminator is optimal and can take arbitrary values, different families of divergences can be recovered. However, as noted by Arjovsky & Bottou (2017), these divergences should be ill-suited to GAN training, contrary to empirical evidence. Our framework addresses this discrepancy, as it properly characterizes the generator's loss and gradient.

Another line of work analyzes the impact of the networks' architecture on the loss landscape of GANs. Some works, on one hand, only study the solution of the usual min-max formulation of GANs, without considering their usual optimization via alternating gradient descent-ascent (Liu et al., 2017; Bai et al., 2019; Sun et al., 2020; Biau et al., 2021; Sahiner et al., 2022). Not only are these results obtained under restrictive assumptions – by focusing on a single GAN model like WGAN, or with discriminators and generators limited to shallow, linear or random features models –, but overlooking alternating optimization hinders their ability to explain GANs' empirical behavior, as detailed in Section 3.

Some studies, on the other hand, deal with the dynamics and convergence of the generated distribution in this setting. Nonetheless, as these dynamics are highly non-linear, this approach typically requires strong simplifying assumptions: Mescheder et al. (2017) assume the existence of Nash equilibria to the considered optimization problem; Mescheder et al. (2018) reduce the generated distribution to a single datapoint; Domingo-Enrich et al. (2020) apply their zero-sum games analysis to mean-field mixtures of generators and discriminators; Balaji et al. (2021) restrict generators and discriminators to be linear or shallow networks; Yang & E (2022) only work with random feature models as discriminators and a modified WGAN loss. In contrast to these works, our framework provides a more comprehensive optimization and architecture modeling as we establish generally applicable results about the influence of the discriminator's architecture on the generator's dynamics.

GANs and NTKs. To the best of our knowledge, our contribution is the first to employ NTKs to comprehensively study GANs. Only Jacot et al. (2019) and Chu et al. (2020) have already studied GANs in the light of NTKs, but their studies had restrictive assumptions and limited scope. Jacot et al. (2019) explain, thanks to the generator's NTK, some GAN failure cases like generator collapse and identify normalization techniques to alleviate them, but without breaking down GANs' training dynamics. Chu et al. (2020) frame the generator's training dynamics for both GANs and variational autoencoders (Kingma & Welling, 2014; Rezende et al., 2014) as a Stein gradient flow under the generator's NTK like in our Section 4.4, but under a strong assumption on generator injectivity which we do not require. Moreover, both works, focusing on the generator, fail to identify the consequences of the discriminator's parameterization on the generator's dynamics via alternating optimization which, encompassed in our framework, yields in Sections 4 and 5 novel results challenging standard GAN knowledge.

Besides the generator, we thoroughly investigate for the first time in the literature the discriminator and its effect on generator optimization via its NTK. To this end, we derive novel results in NTK theory. In particular, while other works studied the regularity of NTKs (Bietti & Mairal, 2019; Yang & Salman, 2019; Basri et al., 2020), ours is, as far as we know, the first to state general differentiability results for NTKs and infinite-width networks. Furthermore, we discover the link between IPM optimization and the NTK MMD, independently of and concurrently with Cheng & Xie (2021), although in a different context: they use the NTK MMD for two-sample statistical testing, whereas we find that IPM GANs actually optimize this metric, thereby explaining the singular performance of NTKs within MMD gradient flows (Arbel et al., 2019).

3. Limits of Previous Studies

We present in this section the usual GAN formulation and illustrate the limitations of prior analyses.

First, let us introduce some notations. Let $\Omega \subseteq \mathbb{R}^n$ be a closed convex set, $\mathcal{P}(\Omega)$ the set of probability distributions over Ω, and $L^2(\mu)$ the set of square-integrable functions from the support $\operatorname{supp}\mu$ of μ to $\mathbb{R}$ with respect to measure μ, with scalar product $\langle\cdot,\cdot\rangle_{L^2(\mu)}$. If $\Lambda \subseteq \Omega$, we write $L^2(\Lambda)$ for $L^2(\lambda)$, with λ the Lebesgue measure on Λ.

3.1. Generative Adversarial Networks

GAN algorithms seek to produce samples from an unknown target distribution $\beta \in \mathcal{P}(\Omega)$. To this extent, a generator function $g \in \mathcal{G}\colon \mathbb{R}^d \to \Omega$ parameterized by θ is learned to map a latent variable $z \sim p_z$ to the space of target samples such that the generated distribution α_g and β are indistinguishable for a discriminator $f \in \mathcal{F}$ parameterized by ϑ. The generator and the discriminator are trained in an adversarial manner as they are assigned conflicting objectives.

Many GAN models consist in solving the following optimization problem, with $a, b, c\colon \mathbb{R} \to \mathbb{R}$:

$$\inf_{g\in\mathcal{G}} \left\{ \mathcal{C}_{f^\star_{\alpha_g}}(\alpha_g) \triangleq \mathbb{E}_{x\sim\alpha_g}\left[c_{f^\star_{\alpha_g}}(x)\right] \right\}, \tag{1}$$

where $c_f = c\circ f$, and $f^\star_{\alpha_g}$ is chosen to solve, or approximate, the following optimization problem:

$$\sup_{f\in\mathcal{F}} \left\{ \mathcal{L}_{\alpha_g}(f) \triangleq \mathbb{E}_{x\sim\alpha_g}\left[a_f(x)\right] - \mathbb{E}_{y\sim\beta}\left[b_f(y)\right] \right\}. \tag{2}$$

For instance, Goodfellow et al. (2014) originally used $a(x) = \log\big(1-\sigma(x)\big)$, $b(x) = c(x) = -\log\big(\sigma(x)\big)$, σ being the sigmoid function; in LSGAN (Mao et al., 2017), $a(x) = -(x+1)^2$, $b(x) = (x-1)^2$, $c(x) = x^2$; and for Integral Probability Metrics (Müller, 1997) used e.g. by Arjovsky et al. (2017), $a = b = c = \mathrm{id}$. Many more fall under this formulation (Nowozin et al., 2016; Lim & Ye, 2017).

Equation (1) is then solved using gradient descent on the generator's parameters, with at each step $j \in \mathbb{N}$:

$$\theta_{j+1} = \theta_j - \eta\mathbb{E}_{z\sim p_z}\left[\nabla_\theta g_{\theta_j}(z)^\top \nabla_x\, c_{f^\star_{\alpha_{g_{\theta_j}}}}(x)\Big|_{x=g_{\theta_j}(z)}\right]. \tag{3}$$

This is obtained via the chain rule from the generator's loss $\mathcal{C}_{f^\star_{\alpha_g}}(\alpha_g)$ in Equation (1). However, we highlight that the gradient applied in Equation (3) differs from $\nabla_\theta \mathcal{C}_{f^\star_{\alpha_g}}(\alpha_g)$: the terms taking into account the dependency of the optimal discriminator $f^\star_{\alpha_{g_\theta}}$ on the generator's parameters are discarded. This is because the discriminator is, in practice, considered to be independent of the generator in the alternating optimization between the generator and the discriminator.

Since $\nabla_x c_{f^\star_\alpha}(x) = \nabla_x f^\star_\alpha(x)\cdot c'\big(f^\star_\alpha(x)\big)$, and as highlighted e.g. by Goodfellow et al. (2014) and Arjovsky & Bottou (2017), the gradient of the discriminator plays a crucial role in the convergence of GANs. For example, if this vector field is null on the training data when $\alpha \neq \beta$, the generator's gradient is zero and convergence is impossible. For this reason, this paper is devoted to developing a better understanding of this gradient field and its consequences on generator optimization when the discriminator is a neural network. In order to characterize this gradient field, we must first study the discriminator itself.

3.2. Alternating Optimization and the Necessity of Modeling the Discriminator Parameterization

For each GAN formulation, it is customary to elucidate the true generator loss $\mathscr{C}\big(\alpha_g, \beta\big)$ implemented by Equation (2), often assuming that $\mathcal{F} = L^2(\Omega)$, i.e. the discriminator can take arbitrary values. Under this assumption, $\mathscr{C}$ would have the form of a Jensen-Shannon divergence in the original GAN and of a Pearson χ^2-divergence in LSGAN, for instance.

However, as pointed out by Arora et al. (2017), the discriminator is trained in practice with a finite number of samples: both fake and target distributions are finite mixtures of Diracs, which we respectively denote as $\hat{\alpha}_g$ and $\hat{\beta}$. Let $\hat{\gamma}_g = \frac{1}{2}\hat{\alpha}_g + \frac{1}{2}\hat{\beta}$ be the distribution of training samples.

Assumption 1 (Finite training set)**.** *$\hat{\gamma}_g \in \mathcal{P}(\Omega)$ is a finite mixture of Diracs.*

In this setting, the Jensen-Shannon and χ^2 divergences are constant since $\hat{\alpha}_g$ and $\hat{\beta}$ generally do not have the same support, which would imply that the generator could not be properly trained since it would receive null gradients. This is the theoretical reason given by Arjovsky & Bottou (2017) to introduce new losses and constraints for the discriminator such as in WGAN (Arjovsky et al., 2017). However, this is inconsistent with empirical results showing that GANs could already be trained adequately even without the latter losses and constraints (Radford et al., 2016). This entails that widely accepted theoretical frameworks miss a central ingredient in their modeling of constrained-free GANs. Uncovering the missing pieces and understanding how they affect training is one of the aims of the current work.

In fact, in the alternating optimization setting as in Equation (3), the constancy of $\mathcal{L}_{\hat{\alpha}_g}$, or even of $\mathcal{C}_{f^\star_{\alpha_g}}$, does not imply that $\nabla_x c_{f^\star_{\alpha_g}}$ in Equation (3) is zero on these points. This stems from the gradient of Equation (3) ignoring the dependency of the optimal discriminator on the generator's parameters: while $\nabla_\theta \mathcal{C}_{f^\star_{\alpha_g}}(\alpha_g)$ might be null, the gradient of Equation (3) differs and may not be zero, thereby changing the actual loss $\mathscr{C}$ optimized by the generator. This fact is unaccounted for in many prior analyses, like the ones of

Arjovsky et al. (2017) and Arora et al. (2017). We refer to Section 5.2 and Appendix B.2 for further discussion.

Furthermore, in the previous theoretical frameworks where the discriminator can take arbitrary values, this gradient field is not even defined for any loss $\mathcal{L}_{\hat{\alpha}_g}$. Indeed, when the discriminator's loss $\mathcal{L}_{\hat{\alpha}_g}(f)$ is only computed on the empirical distribution $\hat{\gamma}_g$ (as in most GAN formulations), the discriminator optimization problem of Equation (2) never yields a unique optimal solution outside $\hat{\gamma}_g$. This is illustrated by the following straightforward result.

Proposition 1 (Ill-Posed Problem in $L^2(\Omega)$)**.** *Suppose that $\mathcal{F} = L^2(\Omega)$, $\operatorname{supp}\hat{\gamma}_g \subsetneq \Omega$. Then, for all $f, h \in \mathcal{F}$ coinciding over $\operatorname{supp}\hat{\gamma}_g$, $\mathcal{L}_{\hat{\alpha}_g}(f) = \mathcal{L}_{\hat{\alpha}_g}(h)$ and Equation (2) has either no or infinitely many optimal solutions in $\mathcal{F}$, all coinciding over $\operatorname{supp}\hat{\gamma}_g$.*

In particular, the set of solutions, if non-empty, contains non-differentiable discriminators as well as discriminators with null or non-informative gradients. This signifies that the loss alone does not impose any constraint on the values that $f_{\hat{\alpha}_g}$ takes outside $\operatorname{supp}\hat{\gamma}_g$, and more particularly on its gradients. Thus, this underspecification of the discriminator over Ω makes the gradient of the optimal discriminator in standard GAN analyses ill-defined. Therefore, an analysis beyond the loss function is necessary to precisely determine the learning problem and true loss $\mathscr{C}$ of the generator implicitly defined by the discriminator under alternating optimization.

4. NTK Analysis of GANs

To tackle the aforementioned issues, we notice that, in practice, the inner optimization problem of Equation (2) is not solved exactly. Instead, using alternating optimization, a proxy neural discriminator is trained using several steps of gradient ascent for each generator update (Goodfellow, 2016). For a learning rate ε and a fixed generator g, this results in the optimization procedure, from $i = 0$ to N:

$$\vartheta_{i+1}^g = \vartheta_i^g + \varepsilon \nabla_\vartheta \mathcal{L}_{\hat{\alpha}_g}\left(f_{\vartheta_i^g}\right), \qquad f_{\hat{\alpha}_g}^\star = f_{\vartheta_N^g}. \tag{4}$$

This training of the discriminator as a neural network solves the gradient indeterminacy of the previous section, but makes a theoretical analysis of its impact unattainable. We propose to facilitate it thanks to the theory of NTKs.

We develop our framework modeling the discriminator using its NTK in Section 4.1. We confirm in Sections 4.2 and 4.3 that it is consistent by proving that the discriminator gradient is well-defined. We then leverage this accurate framework to analyze the dynamics of the generated distribution under alternating optimization via the generator's NTK in Section 4.4. We notably frame this dynamics as a gradient flow of the true generator loss $\mathscr{C}$, which we deduce to be non-increasing during training.

4.1. Modeling Inductive Biases of the Discriminator in the Infinite-Width Limit

We study the continuous-time version of Equation (4):

$$\partial_t \vartheta_t^g = \nabla_\vartheta \mathcal{L}_{\hat{\alpha}_g}\left(f_{\vartheta_t^g}\right), \tag{5}$$

which we consider in the infinite-width limit of the discriminator, making its analysis more tractable.

In the limit where the width of the hidden layers of $f_t \triangleq f_{\vartheta_t^g}$ tends to infinity, Jacot et al. (2018) showed that its so-called NTK $k_{\vartheta_t^g}$ remains constant during a gradient ascent such as Equation (5), i.e. there is a limiting kernel k such that:

$$\begin{aligned} &\forall \tau \in \mathbb{R}_+, \ \ \forall x, y \in \mathbb{R}^n, \ \ \forall t \in [0, \tau], \\ &\qquad k_{\vartheta_t^g}(x, y) \ \triangleq \ \partial_\vartheta f_t(x)^\top \partial_\vartheta f_t(y) \ = \ k(x, y). \end{aligned} \tag{6}$$

In particular, k only depends on the architecture of f and the initialization distribution of its parameters. The constancy of the NTK of f_t during gradient descent holds for many standard architectures, typically without bottleneck and ending with a linear layer (Liu et al., 2020), which is the case of most standard discriminators in the setting of Equation (2). We discuss the applicability of this approximation in Appendix B.1. We more particularly highlight that, under the same conditions, the discriminator's NTK remains constant over the whole GAN optimization process of Equation (3), and not only under a fixed generator.

Assumption 2 (Kernel)**.** *$k\colon \Omega^2 \to \mathbb{R}$ is a symmetric positive semi-definite kernel with $k \in L^2\left(\Omega^2\right)$.*

The constancy of the NTK simplifies the dynamics of training in the functional space. In order to express these dynamics, we must first introduce some preliminary definitions.

Definition 1 (Functional gradient)**.** Whenever a functional $\mathcal{L}\colon L^2(\mu) \to \mathbb{R}$ has sufficient regularity, its gradient w.r.t. μ evaluated at $f \in L^2(\mu)$ is defined in the usual way as the element $\nabla^\mu \mathcal{L}(f) \in L^2(\mu)$ such that for all $\psi \in L^2(\mu)$:

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(\mathcal{L}(f + \varepsilon\psi) - \mathcal{L}(f)\big) = \big\langle \nabla^\mu \mathcal{L}(f), \psi \big\rangle_{L^2(\mu)}. \tag{7}$$

Definition 2 (RKHS w.r.t. μ and kernel integral operator (Sriperumbudur et al., 2010))**.** If k follows Assumption 2 and $\mu \in \mathcal{P}(\Omega)$ is a finite mixture of Diracs, we define the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k^\mu$ of k with respect to μ given by the Moore–Aronszajn theorem as the linear span of functions $k(x, \cdot)$ for $x \in \operatorname{supp}\mu$. Its kernel integral operator from Mercer's theorem is defined as:

$$\mathcal{T}_{k,\mu}\colon L^2(\mu) \to \mathcal{H}_k^\mu, \quad h \mapsto \int_x k(\cdot, x) h(x) \,\mathrm{d}\mu(x). \tag{8}$$

Note that $\mathcal{T}_{k,\mu}$ generates $\mathcal{H}_k^\mu$, and elements of $\mathcal{H}_k^\mu$ are functions defined over all Ω as $\mathcal{H}_k^\mu \subseteq L^2(\Omega)$.

The results of Jacot et al. (2018) imply that the infinite-width discriminator f_t trained by Equation (5) obeys the following differential equation in-between generator updates:

$$\partial_t f_t = \mathcal{T}_{k,\hat{\gamma}_g}\left(\nabla^{\hat{\gamma}_g}\mathcal{L}_{\hat{\alpha}}(f_t)\right). \tag{9}$$

Within the alternating optimization of GANs at generator step j, f_0 would correspond to the previous discriminator step $f^\star_{\alpha_{g_{\theta_j}}} \triangleq f^j$, and $f^{j+1} = f_\tau$, with τ being the training time of the discriminator in-between generator updates.

In the following Sections 4.2 and 4.3, we rely on this differential equation to assess under mild assumptions that the proposed framework is sound w.r.t. the aforementioned gradient indeterminacy issues. We first prove that Equation (9) uniquely defines the discriminator for any initial condition. We then conclude by proving the differentiability of the resulting trained network. These results are not GAN-specific but generalize to networks trained under empirical losses like Equation (2), e.g. for classification and regression.

4.2. Existence, Uniqueness and Characterization of the Discriminator

The following is a positive result on the existence and uniqueness of the discriminator that also characterizes its general form, amenable to theoretical analysis. Presented in the context of a discrete distribution $\hat{\gamma}_g$ but generalizable to broader distributions, this result is proved in Appendix A.2.

Assumption 3 (Loss regularity)**.** *a and b from Equation (2) are differentiable with Lipschitz derivatives over $\mathbb{R}$.*

Theorem 1 (Solution of gradient descent)**.** *Under Assumptions 1 to 3, Equation (9) with initial value $f_0 \in L^2(\Omega)$ admits a unique solution $f_\cdot : \mathbb{R}_+ \to L^2(\Omega)$. Moreover, the following holds for all $t \in \mathbb{R}_+$:*

$$\begin{aligned}\forall t \in \mathbb{R}_+, f_t &= f_0 + \int_0^t \mathcal{T}_{k,\hat{\gamma}_g}\left(\nabla^{\hat{\gamma}_g}\mathcal{L}_{\hat{\alpha}_g}(f_s)\right)\mathrm{d}s \\ &= f_0 + \mathcal{T}_{k,\hat{\gamma}_g}\left(\int_0^t \nabla^{\hat{\gamma}_g}\mathcal{L}_{\hat{\alpha}_g}(f_s)\,\mathrm{d}s\right).\end{aligned} \tag{10}$$

As for any given training time t, there exists a unique $f_t \in L^2(\Omega)$, defined over all of Ω and not only the training set, the aforementioned issue in Section 3.2 of determining the discriminator associated to $\hat{\gamma}_g$ is now resolved. It is now possible to study the discriminator in its general form thanks to Equation (10). It involves two terms: the previous discriminator state $f_0 = f^j$, as well as the kernel operator of an integral. This integral is a function that is undefined outside $\operatorname{supp}\hat{\gamma}_g$, as by definition $\nabla^{\hat{\gamma}_g}\mathcal{L}_{\hat{\alpha}_g}(f_s) \in L^2(\hat{\gamma}_g)$. Fortunately, the kernel operator behaves like a smoothing operator, as it not only defines the function on all of Ω but embeds it in a highly structured space.

Corollary 1 (Training and RKHS)**.** *Under Assumptions 1 to 3, $f_t - f_0$ belongs to the RKHS $\mathcal{H}_k^{\hat{\gamma}_g}$ for all $t \in \mathbb{R}_+$.*

In our setting, this space is generated from the NTK k, which only depends on the discriminator architecture, and not on the loss function. This highlights the crucial role of the discriminator's implicit biases, and enables us to characterize its regularity for a given architecture.

4.3. Differentiability of the Discriminator and its NTK

We study in this section the smoothness, i.e. infinite differentiability, of the discriminator, which we demonstrate in Appendix A.3. It mostly relies on the differentiability of the kernel k, by Equation (10), which is obtained by characterizing the regularity of the corresponding conjugate kernel (Lee et al., 2018). Therefore, we prove the differentiability of the NTKs of standard architectures, and then conclude about the differentiability of f_t.

Assumption 4 (Discriminator architecture)**.** *The discriminator is a standard architecture (fully connected, convolutional or residual). The activation can be any standard function:* $\tanh$*, softplus, ReLU-like, sigmoid, Gaussian, etc.*

Assumption 5 (Discriminator regularity)**.** *The activation function is smooth.*

Assumption 6 (Discriminator bias)**.** *Linear layers have non-null bias terms.*

We first prove the differentiability of the NTK.

Proposition 2 (Differentiability of k)**.** *Let k be the NTK of an infinite-width network from Assumption 4. For any $y \in \Omega$, $k(\cdot, y)$ is smooth everywhere over Ω under Assumption 5, or almost everywhere if Assumption 6 holds instead.*

From Proposition 2, NTKs satisfy Assumption 2. Using Corollary 1, we thus conclude on the differentiability of f_t.

Theorem 2 (Differentiability of f_t)**.** *Suppose that k is the NTK of an infinite-width network following Assumption 4. Then f_t is smooth everywhere over Ω under Assumption 5, or almost everywhere when Assumption 6 holds instead.*

Remark 1 (Bias-free ReLU networks)**.** ReLU networks with hidden layers and no bias are not differentiable at 0. However, by introducing non-zero bias, this non-differentiability at 0 disappears in the NTK and the infinite-width discriminator. This observation explains some experimental results in Section 6. Note that Bietti & Mairal (2019) state that the bias-free ReLU kernel is not Lipschitz even outside 0. However, we find this result to be incorrect. We further discuss this matter in Appendix B.3.

This result demonstrates that, for a wide range of GANs, e.g. vanilla GAN and LSGAN, the optimized discriminator indeed admits gradients, making the gradient flow given to

the generator well-defined in our framework. This supports our motivation to bring the theory closer to the empirical evidence that many GAN models do work in practice while their theoretical interpretation until now has been stating the opposite (Arjovsky & Bottou, 2017).

4.4. Dynamics of the Generated Distribution

By ensuring the existence of $\nabla f^\star_{\hat\alpha_g}$, the previous results allow us to study Equation (3). We consider it in continuous-time like Equation (5), with training time ℓ as well as $g_\ell \triangleq g_{\theta_\ell}$ and $\alpha_\ell \triangleq \alpha_{g_\ell}$. NTKs enable us to describe the generated distribution's dynamics and uncover the true generated loss $\mathscr{C}$ in the following manner, as shown in Appendix A.4.

Proposition 3 (Dynamics of α_ℓ)**.** *Under Assumptions 4 and 5, Equation (3) is well-posed and yields in continuous-time, with k_{g_ℓ} the NTK of the generator g_ℓ:*

$$\partial_\ell g_\ell = -\mathcal{T}_{k_{g_\ell},p_z}\left(z \mapsto \nabla_x\, c_{f^\star_{\hat\alpha_{g_\ell}}}(x)\Big|_{x=g_\ell(z)}\right). \tag{11}$$

Equivalently, the following continuity equation holds for the joint distribution α^z_ℓ of $\big(z, g_\ell(z)\big)$ under $z \sim p_z$:

$$\partial_\ell \alpha^z_\ell = \nabla_x \cdot \left(\alpha^z_\ell \mathcal{T}_{k_{g_\ell},p_z}\left(z \mapsto \nabla_x\, c_{f^\star_{\hat\alpha_{g_\ell}}}(x)\Big|_{x=g_\ell(z)}\right)\right), \tag{12}$$

where α_ℓ is the marginalization of α^z_ℓ over $z \sim p_z$.

In its infinite-width limit, the generator's NTK is also constant: $k_{g_\ell} = k_g$; let us study the latter proposition under this assumption. Suppose that there exists a functional $\mathscr{C}$ over $L^2(\Omega)$ such that $c_{f^\star_{\hat\alpha}} = \partial_{\hat\alpha}\mathscr{C}(\hat\alpha)$. Standard results in gradient flows theory – see Ambrosio et al. (2008, Chapter 10) for a detailed exposition or Arbel et al. (2019, Appendix A.3) for a summary – state that $\nabla c_{f^\star_{\hat\alpha}}$ is then the strong subdifferential of $\mathscr{C}(\hat\alpha)$ for the Wasserstein geometry.

When $k_g\big(z, z'\big) = \delta_{z-z'} I_n$ with δ a Dirac centered at 0, we have $\mathcal{T}_{k_g,p_z} = \mathrm{id}$. Then, from Equation (12), α^z_ℓ follows the Wasserstein gradient flow with $\mathscr{C}$ as potential. This implies that $\mathscr{C}(\hat\alpha_\ell)$ is decreasing w.r.t. the generator's training time ℓ. In other words, the generator g is trained to minimize $\mathscr{C}\big(\hat\alpha_g\big)$. Hence, this result characterizes the implicit objective of the generator as $\mathscr{C}$ satisfying $c_{f^\star_{\hat\alpha}} = \partial_{\hat\alpha}\mathscr{C}(\hat\alpha)$.

In the general case, $\mathcal{T}_{k_g,p_z}$ introduces interactions between generated particles as a consequence of the neural parameterization of the generator. Then, Equation (12) amounts to following the same gradient flow as before, but in a Stein geometry (Duncan et al., 2019) – instead of a Wasserstein geometry – determined by the generator's integral operator, directly implying that in this case $\mathscr{C}(\hat\alpha_\ell)$ also decreases during training. This geometrical understanding opens interesting perspectives for theoretical analysis, e.g. we see that GAN training in this regime generalizes Stein variational gradient descent (Liu & Wang, 2016), with the Kullback-Leibler minimization objective between generated and target distributions being replaced with $\mathscr{C}(\hat\alpha)$.

Improving our understanding of Equation (12) is fundamental in order to elucidate the open problem of the neural generator's convergence. Our study enables us to shed light on these dynamics and highlights the necessity of pursuing the study of GANs via NTKs to obtain a more comprehensive understanding of them, which is the purpose of the rest of this paper. In particular, the non-interacting case where $\mathcal{T}_{k_g,p_z} = \mathrm{id}$ already yields particularly useful insights that we explore in Section 6. Moreover, we discuss in the following section standard GAN losses and determine the minimized functional $\mathscr{C}$ in these cases.

5. Study of Specific Losses

Armed with the previous framework, we derive in this section more fine-grained results about the optimized loss $\mathscr{C}$ for standard GAN models. Proofs are detailed in Appendix A.6.

5.1. The IPM as an NTK MMD Minimizer

We study the case of the IPM loss, with the following remarkable discriminator expression, from which we deduce the objective minimized by the generator.

Proposition 4 (IPM discriminator)**.** *Under Assumptions 1 and 2, the solutions of Equation (9) for $a = b = \mathrm{id}$ are $f_t = f_0 + t f^*_{\hat\alpha_g}$, where $f^*_{\hat\alpha_g}$ is the unnormalized MMD witness function (Gretton et al., 2012) with kernel k, yielding:*

$$\begin{aligned} f^*_{\hat\alpha_g} &= \mathbb{E}_{x\sim\hat\alpha_g}\big[k(x,\cdot)\big] - \mathbb{E}_{y\sim\hat\beta}\big[k(y,\cdot)\big], \\ \mathcal{L}_{\hat\alpha_g}(f_t) &= \mathcal{L}_{\hat\alpha_g}(f_0) + t\cdot \mathrm{MMD}^2_k\Big(\hat\alpha_g, \hat\beta\Big). \end{aligned} \tag{13}$$

The latter result signifies that the direction of the gradient given to the discriminator at each of its optimization step is optimal within the RKHS of its NTK, stemming from the linearity of the IPM loss. The connection with MMD is especially interesting as it has been thoroughly studied in the literature (Muandet et al., 2017). If k is characteristic, as discussed in Appendix B.5, then it defines a distance between distributions. Moreover, the statistical properties of the loss induced by the discriminator directly follow from those of the MMD: it is an unbiased estimator with a squared sample complexity that is independent of the dimension of the samples (Gretton et al., 2007).

Suppose that the discriminator is reinitialized at every step of the generator, with $f_0 = 0$ in Equation (9); this is possible with the initialization scheme of Zhang et al. (2020). Then, as $c = \mathrm{id}$ and from Proposition 4, $\nabla c_{f_{\hat\alpha}} = \tau \nabla f^*_{\hat\alpha_g}$, where τ is the training time of the discriminator. The latter gradient constitutes the gradient flow of the squared MMD, as shown

by Arbel et al. (2019) with convergence guarantees and discretization properties in the absence of generator. This signifies that $\mathscr{C}(\hat{\alpha}) = \tau \mathrm{MMD}_k^2\left(\hat{\alpha}_g, \hat{\beta}\right)$ (see Section 4.4).

Therefore, in the IPM case, we discover via Proposition 4 that the generator is actually trained to minimize the MMD between the empirical generated and target distributions, w.r.t. the NTK of the discriminator. This novel connection implies that prior MMD GAN convergence results, like the ones of Mroueh & Nguyen (2021) about the generator trained in such conditions, even though they were established without considering the discriminator's NTK, remarkably transfer to the general unconstrained IPM case.

We further discuss our IPM results in the following remarks.

Remark 2 (IPM and WGAN)**.** Along with a constraint on the set of functions, the IPM is involved in the earth mover's distance $\mathcal{W}_1$ (Villani, 2009) – used in WGAN and StyleGAN (Karras et al., 2019), close to the hinge loss of BigGAN (Brock et al., 2019) –, the MMD – used in MMD GAN (Li et al., 2017) –, the total variation, etc. In Proposition 4, we study the IPM with the sole constraint of having a neural discriminator. Our analysis implies that this suffices to ensure relevant gradients, given the aforementioned convergence results. This contradicts the recurring assertion that the Lipschitz constraint of WGAN (Arjovsky et al., 2017) is necessary to solve the gradient issues of prior approaches. Indeed, these issues originate from the analyses inadequacy, as shown in this work. Hence, while WGAN tackles them by changing the loss and adding a constraint, we fundamentally address them with a refined framework. A WGAN analysis, left for future work, would require combining the neural discriminator and Lipschitz constraints.

Remark 3 (Instance smoothing)**.** We show for IPMs that modeling the discriminator's architecture amounts to smoothing out the input distribution using the kernel integral operator $\mathcal{T}_{k,\hat{\gamma}_g}$ and can thus be seen as a generalization of the regularization technique for GANs called instance noise (Sønderby et al., 2017). This is discussed in Appendix B.4.

Remark 4 (Regularization by training time)**.** Proposition 4 highlights the importance of discriminator training time, which needs to be controlled to regularize its gradient magnitude. This corresponds to customary practices where the discriminator is trained for a small number of steps to avoid divergence issues, like in DCGAN (Radford et al., 2016). In the IPM case, we have, with $\|\cdot\|_{\mathcal{H}_k^{\hat{\gamma}}}$ as the RKHS semi-norm:

$$\|f_t\|_{\mathcal{H}_k^{\hat{\gamma}}}^2 \leq \|f_0\|_{\mathcal{H}_k^{\hat{\gamma}}}^2 + t^2 \left\| f^*_{\hat{\alpha}_g} \right\|_{\mathcal{H}_k^{\hat{\gamma}}}^2, \tag{14}$$

with equality when $f_0 = 0$. This provides a simple criterion to control the discriminator norm by its training time. For example, assuming $f_0 = 0$, setting $t = \left\| f^*_{\hat{\alpha}_g} \right\|^{-1}_{\mathcal{H}_k^{\hat{\gamma}}}$ recovers the MMD dual constraint of a unit-norm discriminator, i.e. that $\|f_t\|_{\mathcal{H}_k^{\hat{\gamma}}} = 1$, yielding $\mathcal{L}_{\hat{\alpha}_g}(f_t) = \mathrm{MMD}_k\left(\hat{\alpha}_g, \hat{\beta}\right)$.

5.2. LSGAN and New Divergences

Optimality of the discriminator can be proved when assuming that its loss function is well-behaved. Let us consider the case of LSGAN, for which Equation (9) can be solved by adapting the results from Jacot et al. (2018) for regression.

Proposition 5 (LSGAN discr.)**.** *Under Assumptions 1 and 2, the solutions of Equation (9) for $a = -(\mathrm{id}+1)^2$ and $b = -(\mathrm{id}-1)^2$ are defined for all $t \in \mathbb{R}_+$ as:*

$$f_t = \exp\left(-4t\mathcal{T}_{k,\hat{\gamma}_g}\right)(f_0 - \rho) + \rho, \quad \rho = \frac{\mathrm{d}\left(\hat{\beta} - \hat{\alpha}_g\right)}{\mathrm{d}\left(\hat{\beta} + \hat{\alpha}_g\right)}. \tag{15}$$

In the previous result, ρ is the optimum of $\mathcal{L}_{\hat{\alpha}_g}$ over $L^2\left(\hat{\gamma}_g\right)$. When k is positive definite over $\hat{\gamma}_g$ (see Appendix B.5), f_t tends to the optimum for $\mathcal{L}_{\hat{\alpha}_g}$ as its limit is ρ over $\operatorname{supp}\hat{\gamma}_g$. Nonetheless, unlike the discriminator with arbitrary values of Section 3.2, f_∞ is defined over all Ω thanks to the integral operator $\mathcal{T}_{k,\hat{\gamma}_g}$. It is also the solution to the minimum norm interpolant problem in the RKHS (Jacot et al., 2018), therefore explaining why the discriminator does not overfit in scarce data regimes (see Section 6), and consequently has bounded gradients despite large training times. We also prove a generalization of this optimality conclusion for concave bounded losses in Appendix A.5.

Following the discussion initiated in Section 3.2 and applying it to LSGAN using Proposition 5, similarly to the Jensen-Shannon, the resulting generator loss on discrete training data is constant when the discriminator is optimal. However, the gradients received by the generator are not necessarily null, e.g. in the empirical analysis of Section 6. This is because the learning problem of the generator induced by the discriminator makes the generator minimize another loss $\mathscr{C}$, as explained in Section 4.4. This raises the question of determining $\mathscr{C}$ for LSGAN and other standard losses. Furthermore, the same problem arises in the case of incompletely trained discriminators f_t. Unlike the IPM case for which the results of Arbel et al. (2019) who leveraged the theory of Ambrosio et al. (2008) led to a remarkable solution, this connection remains to be established for other adversarial losses. We leave this as future work.

6. Empirical Study

We present a selection of empirical results for different losses and architectures to show the relevance of our framework, with more insights in Appendix C, by evaluating its adequacy and practical implications on GAN convergence. All experiments are performed with the proposed Generative Adversarial Neural Tangent Kernel ToolKit GAN(TK)^2 that we release at `https://github.com/emited/`

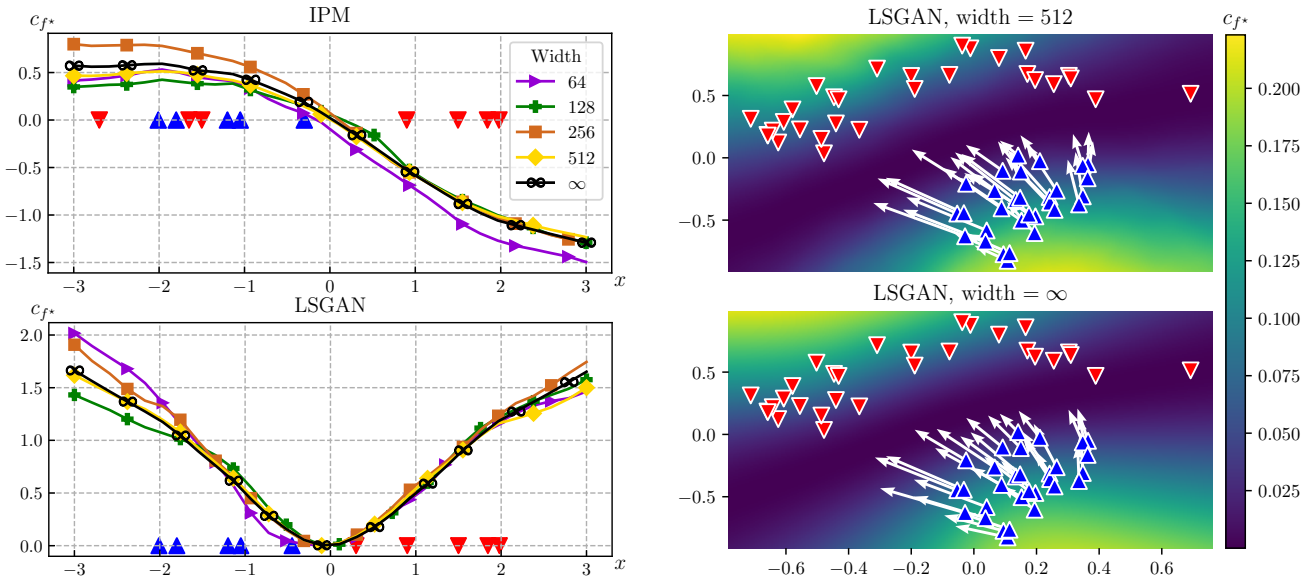


Figure 1. Values of $c_{f^\star}$ for LSGAN and IPM, where $f^\star$ is a 3-layer ReLU MLP with bias and varying width trained on the dataset represented by ▼ (real) and ▲ (fake) markers, initialized at $f_0 = 0$. The infinite-width network is trained for a time $\tau = 1$ and the finite-width networks using 10 gradient descent steps with learning rate $\varepsilon = 0.1$, to make training times correspond. The gradients $\nabla_x c_{f^\star}$ are shown with white arrows on the two-dimensional plots for the fake distribution.

`gantk2` in the hope that the community leverages and expands it for principled GAN analyses. It is based on the JAX Neural Tangents library (Novak et al., 2020), and is convenient to evaluate architectures and losses based on different visualizations and analyses.

For the sake of efficiency and for these experiments only, we choose $f_0 = 0$ using the antisymmetrical initialization (Zhang et al., 2020). Indeed, in the analytical computations of the infinite-width regime, taking into account all previous discriminator states for each generator step is computationally infeasible. This choice also allows us to ignore residual gradients from the initialization, which introduce noise in the optimization process.

Adequacy for fixed distributions. We first study the case where generated and target distributions are fixed. In this setting, we qualitatively study the similarity between the finite- and infinite-width regimes of the discriminator. Figure 1 shows $c_{f^\star}$ and its gradients on one- and two-dimensional data for LSGAN and IPM losses with a ReLU MLP with 3 hidden layers of varying widths. We find the behavior of finite-width discriminators to be close to their infinite-width counterpart for standard widths, and converges rapidly to the given limit as the width becomes larger.

In the rest of this section, we focus on the study of convergence of the generated distribution.

Experimental setting. We consider a target distribution sampled from 8 Gaussians evenly distributed on a centered sphere (cf. Figure 2), in a setup similar to that of Metz et al. (2017), Srivastava et al. (2017) and Arjovsky et al. (2017). We alleviate the complexity of the analysis by following Equation (12) with $\mathcal{T}_{k_{g_\ell}, p_z} = \mathrm{id}$, similarly to Mroueh et al. (2019) and Arbel et al. (2019), thereby modeling the generator's evolution by considering a finite number of samples, initially Gaussian. For IPM and LSGAN losses, we evaluate the convergence of the generated distributions for a discriminator with ReLU activations in the finite- and infinite-width regime, either with or without bias. We also comparatively evaluate the advantages of this architecture by considering the case where the infinite-width loss is not given by an NTK, but by the popular Radial Basis Function (RBF) kernel, which is characteristic and presents attractive properties (Muandet et al., 2017). We refer to Figure 2 for qualitative results and Table 1 in Appendix C for a numerical evaluation. Note that similar results for more datasets, including MNIST and CelebA, and architectures are available in Appendix C.

Adequacy. We observe that correlated performances between the finite- and infinite-width regimes, ReLU networks being considerably better in the latter. Remarkably, for the infinite-width IPM, generated and target distributions perfectly match. This can be explained by the high capacity of infinite-width networks; it has already been shown that NTKs benefit from low-data regimes (Arora et al., 2020).

Impact of bias. The bias-free discriminator performs worse than with bias, for both regimes and both losses. This is in line with findings of e.g. Basri et al. (2020), and can be

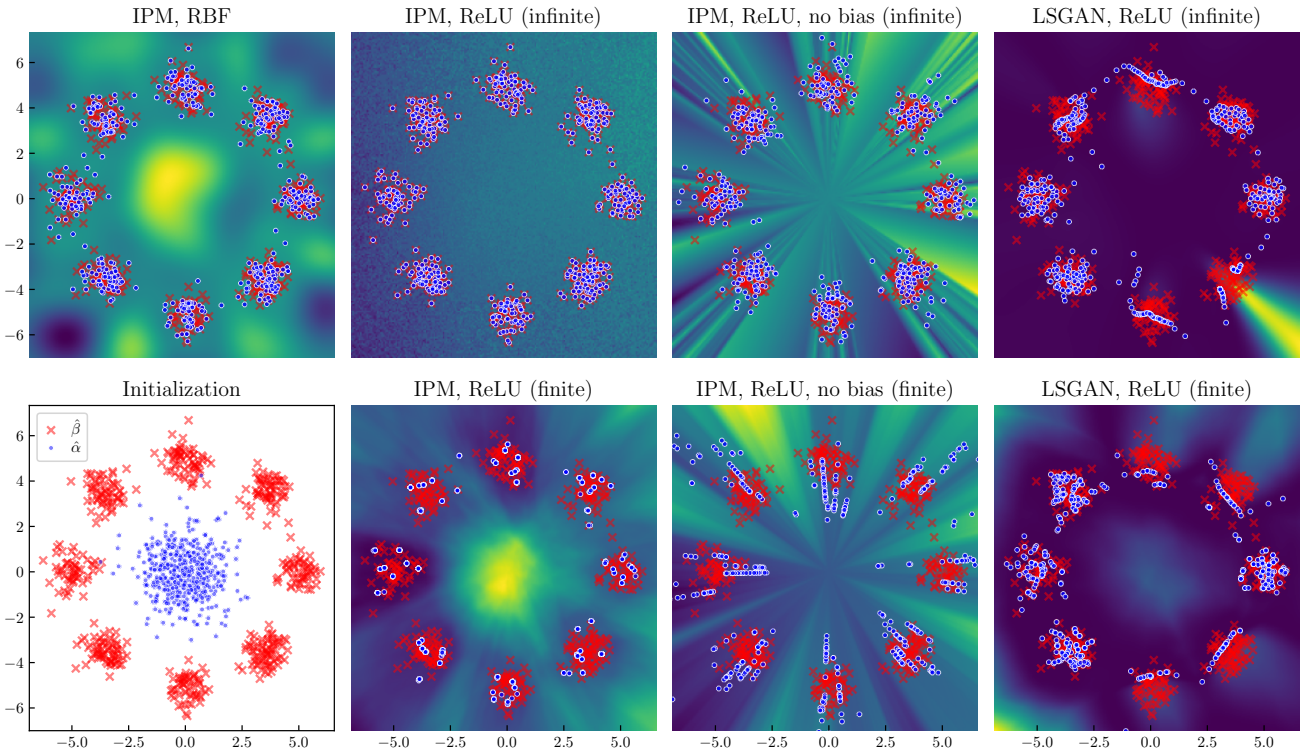


Figure 2. Generator (●) and target (×) samples for different methods. In the background, $c_{f^\star}$.

explained in our theoretical framework by comparing their NTKs. Indeed, the NTK of a bias-free ReLU network is not characteristic, whereas its bias counterpart was proven to present powerful approximation properties (Ji et al., 2020). Furthermore, results of Section 4.3 state that the ReLU NTK with bias is differentiable at 0, whereas its bias-free version is not, which can disrupt optimization based on its gradients: note in Figure 2 the abrupt streaks of the discriminator directed towards 0 and their consequences on convergence.

NTK vs. RBF. We observe the superiority of NTKs over the RBF kernel. This highlights that the gradients of a ReLU network with bias are particularly well adapted to GANs. Visualizations of these gradients in the infinite-width limit are available in Appendix C.4 and further corroborate these findings. More generally, we believe that the NTK of ReLU networks could be of particular interest for kernel methods requiring the computation of a spatial gradient, like Stein variational gradient descent (Liu & Wang, 2016).

7. Conclusion

Leveraging the theory of infinite-width neural networks, we propose a framework of analysis for GANs explicitly modeling a large variety of discriminator architectures under the alternating optimization setting. We show that the proposed framework more accurately models GAN training compared to prior approaches by deriving properties of the trained discriminator. We demonstrate the analysis opportunities of the proposed modeling by studying the generated distribution that we find to follow a gradient flow on probability spaces minimizing some functional that we characterize. We further study the latter for specific GAN losses and architectures, both theoretically and empirically, notably using our public GAN analysis toolkit. We believe that this work will serve as a basis for more elaborate analyses, thus leading to more principled, better GAN models.

Acknowledgements

We would like to thank all members of the MLIA team from the ISIR laboratory of Sorbonne Université for helpful discussions and comments.

We acknowledge financial support from the DEEPNUM ANR project (ANR-21-CE23-0017-02), the ETH Foundations of Data Science, and the European Union's Horizon 2020 research and innovation programme under grant agreement 825619 (AI4EU). This work was granted access to the HPC resources of IDRIS under allocations 2020-AD011011360 and 2021-AD011011360R1 made by GENCI (Grand Equipement National de Calcul Intensif). Patrick Gallinari is additionally funded by the 2019 ANR AI Chairs program via the DL4CLIM project.

References

Adler, R. J. *The Geometry Of Random Fields*. Society for Industrial and Applied Mathematics, December 1981.

Adler, R. J. An introduction to continuity, extrema, and related topics for general gaussian processes. *Lecture Notes-Monograph Series*, 12:i–155, 1990.

Alemohammad, S., Wang, Z., Balestriero, R., and Baraniuk, R. G. The recurrent neural tangent kernel. In *International Conference on Learning Representations*, 2021.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, June 2019.

Ambrosio, L. and Crippa, G. Continuity equations and ODE flows with non-smooth velocity. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 144 (6):1191–1244, 2014.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient Flows*. Birkhäuser Basel, Basel, Switzerland, 2008.

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 6484–6494. Curran Associates, Inc., 2019.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, August 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (GANs). In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 224–232. PMLR, August 2017.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 8141–8150. Curran Associates, Inc., 2019.

Arora, S., Du, S. S., Li, Z., Salakhutdinov, R., Wang, R., and Yu, D. Harnessing the power of infinitely wide deep nets on small-data tasks. In *International Conference on Learning Representations*, 2020.

Bai, Y., Ma, T., and Risteski, A. Approximability of discriminators implies diversity in GANs. In *International Conference on Learning Representations*, 2019.

Balaji, Y., Sajedi, M., Kalibhat, N. M., Ding, M., Stöger, D., Soltanolkotabi, M., and Feizi, S. Understanding overparameterization in generative adversarial networks. In *International Conference on Learning Representations*, 2021.

Basri, R., Galun, M., Geifman, A., Jacobs, D., Kasten, Y., and Kritchman, S. Frequency bias in neural networks for input of non-uniform density. In Daumé, III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 685–694. PMLR, July 2020.

Biau, G., Sangnier, M., and Tanielian, U. Some theoretical insights into wasserstein GANs. *Journal of Machine Learning Research*, 22(119):1–45, 2021.

Bietti, A. and Bach, F. Deep equals shallow for ReLU networks in kernel regimes. In *International Conference on Learning Representations*, 2021.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 12893–12904. Curran Associates, Inc., 2019.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Chen, L. and Xu, S. Deep neural tangent kernel and Laplace kernel have the same RKHS. In *International Conference on Learning Representations*, 2021.

Cheng, X. and Xie, Y. Neural tangent kernel maximum mean discrepancy. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Wortman Vaughan, J. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6658–6670. Curran Associates, Inc., 2021.

Chu, C., Minami, K., and Fukumizu, K. The equivalence between Stein variational gradient descent and black-box variational inference. *arXiv preprint arXiv:2004.01822*, 2020.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. On the Lambert W function. *Advances in Computational Mathematics*, 5(1):329–359, December 1996.

Corless, R. M., Ding, H., Higham, N. J., and Jeffrey, D. J. The solution of $S\exp(S) = A$ is not always the Lambert W function of A. In *Proceedings of the 2007 International Symposium on Symbolic and Algebraic Computation*, ISSAC '07, pp. 116–121, New York, NY, USA, 2007. Association for Computing Machinery.

Domingo-Enrich, C., Jelassi, S., Mensch, A., Rotskoff, G., and Bruna, J. A mean-field analysis of two-player zero-sum games. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20215–20226. Curran Associates, Inc., 2020.

Duncan, A., Nüsken, N., and Szpruch, L. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

Fan, Z. and Wang, Z. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7710–7721. Curran Associates, Inc., 2020.

Farkas, B. and Wegner, S.-A. Variations on Barbălat's lemma. *The American Mathematical Monthly*, 123(8): 825–830, 2016.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. Interpolating between optimal transport and MMD using Sinkhorn divergences. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690. PMLR, April 2019.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11), November 2020.

Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680. Curran Associates, Inc., 2014.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J. C., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19, pp. 513–520. MIT Press, 2007.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

Higham, N. J. *Functions of matrices: theory and computation*. Society for Industrial and Applied Mathematics, 2008.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. Infinite attention: NNGP and NTK for deep attention networks. In Daumé, III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4376–4386. PMLR, July 2020.

Huang, K., Wang, Y., Tao, M., and Zhao, T. Why do deep residual networks generalize better than deep feedforward networks? — a neural tangent kernel perspective. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2698–2709. Curran Associates, Inc., 2020.

Iacono, R. and Boyd, J. P. New approximations to the principal real-valued branch of the Lambert W-function. *Advances in Computational Mathematics*, 43(6):1403–1436, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 8580–8589. Curran Associates, Inc., 2018.

Jacot, A., Gabriel, F., Ged, F., and Hongler, C. Order and chaos: NTK views on DNN normalization, checkerboard and boundary artifacts. *arXiv preprint arXiv:1907.05715*, 2019.

Jain, N., Olmo, A., Sengupta, S., Manikonda, L., and Kambhampati, S. Imperfect imaGANation: Implications of GANs exacerbating biases on facial data augmentation and Snapchat selfie lenses. *arXiv preprint arXiv:2001.09528*, 2020.

Ji, Z., Telgarsky, M., and Xian, R. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2020.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, June 2019.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, June 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Kurach, K., Lucic, M., Zhai, X., Michalski, M., and Gelly, S. A large-scale study on regularization and normalization in GANs. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3581–3590. PMLR, June 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 8572–8583. Curran Associates, Inc., 2019.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: an empirical study. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15156–15172. Curran Associates, Inc., 2020.

Leipnik, R. B. and Pearce, C. E. M. The multivariate Faà di Bruno formula and multivariate Taylor expansions with explicit integral remainder term. *The ANZIAM Journal*, 48(3):327–341, 2007.

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards deeper understanding of moment matching network. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 2200–2210. Curran Associates, Inc., 2017.

Lim, J. H. and Ye, J. C. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017.

Littwin, E., Galanti, T., Wolf, L., and Yang, G. On infinite-width hypernetworks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13226–13237. Curran Associates, Inc., 2020a.

Littwin, E., Myara, B., Sabah, S., Susskind, J., Zhai, S., and Golan, O. Collegial ensembles. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18738–18748. Curran Associates, Inc., 2020b.

Liu, C., Zhu, L., and Belkin, M. On the linearity of large non-linear models: when and why the tangent kernel is constant. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15954–15964. Curran Associates, Inc., 2020.

Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., and Mallya, A. Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 109(5):839–862, 2021.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 2378–2386. Curran Associates, Inc., 2016.

Liu, S., Bousquet, O., and Chaudhuri, K. Approximation and convergence properties of generative adversarial learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 5551–5559. Curran Associates, Inc., 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, December 2015.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are GANs created equal? a large-scale study. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 698–707. Curran Associates, Inc., 2018.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821, October 2017.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of GANs. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 1823–1833. Curran Associates, Inc., 2017.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3481–3490. PMLR, July 2018.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Mroueh, Y. and Nguyen, T. On the convergence of gradient descent in GANs: MMD GAN as a gradient flow. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1720–1728. PMLR, April 2021.

Mroueh, Y., Sercu, T., and Raj, A. Sobolev descent. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2976–2985. PMLR, April 2019.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1–2):1–141, 2017.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29 (2):429–443, 1997.

Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural Tangents: Fast and easy infinite neural networks in Python. In *International Conference on Learning Representations*, 2020. URL https://github.com/google/neural-tangents.

Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 271–279. Curran Associates, Inc., 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Beijing, China, June 2014. PMLR.

Sahiner, A., Ergen, T., Ozturkler, B., Bartan, B., Pauly, J. M., Mardani, M., and Pilanci, M. Hidden convexity of wasserstein GANs: Interpretable generative models with closed-form solutions. In *International Conference on Learning Representations*, 2022.

Scheuerer, M. *A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis*. PhD thesis, Georg-August-Universität Göttingen, October 2009. URL https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-0006-B3D5-1.

Sohl-Dickstein, J., Novak, R., Schoenholz, S. S., and Lee, J. On the infinite width limit of neural networks with a standard parameterization. *arXiv preprint arXiv:2001.07301*, 2020.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. Hilbert space

embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. VEEGAN: Reducing mode collapse in GANs using implicit variational learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 3310–3320. Curran Associates, Inc., 2017.

Steinwart, I. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, November 2001.

Sun, R., Fang, T., and Schwing, A. Towards a better global loss landscape of GANs. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10186–10198. Curran Associates, Inc., 2020.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised MAP inference for image super-resolution. In *International Conference on Learning Representations*, 2017.

Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547. Curran Associates, Inc., 2020.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. DeepFakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

Villani, C. *The Wasserstein distances*, pp. 93–111. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, Berlin - Heidelberg, Germany, 2009.

Wang, Z., She, Q., and Ward, T. E. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys*, 54(2), April 2021.

Yang, G. Tensor programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

Yang, G. and Hu, E. J. Tensor programs iv: Feature learning in infinite-width neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, July 2021.

Yang, G. and Salman, H. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

Yang, H. and E, W. Generalization error of GAN from the discriminator's perspective. *Research in the Mathematical Sciences*, 9(8), 2022.

Zhang, Y., Xu, Z.-Q. J., Luo, T., and Ma, Z. A type of generalization error induced by initialization in deep neural networks. In Lu, J. and Ward, R. (eds.), *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 144–164, Princeton University, Princeton, NJ, USA, July 2020. PMLR.

Zhou, Z., Liang, J., Song, Y., Yu, L., Wang, H., Zhang, W., Yu, Y., and Zhang, Z. Lipschitz generative adversarial nets. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7584–7593, Long Beach, California, USA, June 2019. PMLR.

Appendix

Table of Contents

A Proofs of Theoretical Results and Additional Results **16**
A.1 Recall of Assumptions in the Paper . . . 16
A.2 On the Solutions of Equation (9) . . . 16
A.3 Differentiability of Infinite-Width Networks and their NTKs . . . 21
A.4 Dynamics of the Generated Distribution . . . 26
A.5 Optimality in Concave Setting . . . 27
A.6 Case Studies of Discriminator Dynamics . . . 29

B Discussions and Remarks **33**
B.1 From Finite to Infinite-Width Networks . . . 33
B.2 Loss of the Generator and its Gradient . . . 34
B.3 Differentiability of the Bias-Free ReLU Kernel . . . 35
B.4 Integral Operator and Instance Noise . . . 35
B.5 Positive Definite NTKs . . . 36
B.6 Societal Impact . . . 37

C GAN(TK)^2 and Further Empirical Analyses **37**
C.1 Two-Dimensional Datasets . . . 37
C.2 ReLU vs. Sigmoid Activations . . . 37
C.3 Qualitative MNIST and CelebA Experiment . . . 40
C.4 Visualizing the Gradient Field Induced by the Discriminator . . . 40

D Experimental Details **43**
D.1 GAN(TK)^2 Specifications and Computing Resources . . . 43
D.2 Datasets . . . 43
D.3 Parameters . . . 44

In the course of this appendix, we drop the subscript g for $\hat{\gamma}_g$, $\hat{\alpha}_g$ and other notations when the dependency on a fixed generator g is clear and indicated in the main paper, for the sake of clarity.

A. Proofs of Theoretical Results and Additional Results

We prove in this section all theoretical results mentioned in Sections 4 and 5. Appendix A.2 is devoted to the proof of Theorem 1, Appendix A.3 focuses on proving the differentiability results skimmed in Section 4.3, Appendix A.4 contains the demonstration of Proposition 3, and Appendices A.5 and A.6 develop the results presented in Section 5.

We will need in the course of these proofs the following standard definition. For any measurable function T and measure μ, $T_\sharp\mu$ denotes the push-forward measure which is defined as $T_\sharp\mu(B) = \mu\big(T^{-1}(B)\big)$, for any measurable set B.

A.1. Recall of Assumptions in the Paper

Assumption 1 (Finite training set)**.** $\hat{\gamma} \in \mathcal{P}(\Omega)$ is a finite mixture of Diracs.

Assumption 2 (Kernel)**.** $k\colon \Omega^2 \to \mathbb{R}$ is a symmetric positive semi-definite kernel with $k \in L^2\big(\Omega^2\big)$.

Assumption 3 (Loss regularity)**.** a and b from Equation (2) are differentiable with Lipschitz derivatives over $\mathbb{R}$.

Assumption 4 (Discriminator architecture)**.** The discriminator is a standard architecture (fully connected, convolutional or residual). Any activation ϕ in the network satisfies the following properties:

- ϕ is smooth everywhere except on a finite set D;
- for all $j \in \mathbb{N}$, there exist scalars $\lambda_1^{(j)}$ and $\lambda_2^{(j)}$ such that:

$$\forall x \in \mathbb{R} \setminus D, \left|\phi^{(j)}(x)\right| \leq \lambda_1^{(j)}|x| + \lambda_2^{(j)}, \tag{16}$$

  where $\phi^{(j)}$ is the j-th derivative of ϕ.

Assumption 5 (Discriminator regularity)**.** $D = \emptyset$, i.e. ϕ is smooth.

Assumption 6 (Discriminator bias)**.** Linear layers have non-null bias terms. Moreover, for all $x, y \in \mathbb{R}$ such that $x \neq y$, the following holds:

$$\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\phi(x\varepsilon)^2 \neq \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\phi(y\varepsilon)^2. \tag{17}$$

Remark 5 (Typical activations)**.** Assumptions 4 to 6 cover multiple standard activation functions, including $\tanh$, softplus, ReLU, leaky ReLU and sigmoid.

A.2. On the Solutions of Equation (9)

The methods used in this section are adaptations to our setting of standard methods of proof. In particular, they can be easily adapted to slightly different contexts, the main ingredient being the structure of the kernel integral operator. Moreover, it is also worth noting that, although we relied on Assumption 1 for $\hat{\gamma}$, the results are essentially unchanged if we take a compactly supported measure γ instead.

We decompose the proof into several intermediate results. Theorem 3 and Proposition 6, stated and demonstrated in this section, correspond when combined to Theorem 1.

Let us first prove the following two intermediate lemmas.

Lemma 1. *Let $\delta T > 0$ and $\mathcal{F}_{\delta T} = \mathcal{C}\Big([0, \delta T], B_{L^2(\hat{\gamma})}(f_0, 1)\Big)$ endowed with the norm:*

$$\forall u \in \mathcal{F}_{\delta T},\ \|u\| = \sup_{t\in[0,\delta T]} \|u_t\|_{L^2(\hat{\gamma})}. \tag{18}$$

Then $\mathcal{F}_{\delta T}$ is complete.

Proof. Let $(u^n)_n$ be a Cauchy sequence in $\mathcal{F}_{\delta T}$. For a fixed $t \in [0, \delta T]$:

$$\forall n, m, \|u_t^n - u_t^m\|_{L^2(\hat{\gamma})} \leq \|u^n - u^m\|, \tag{19}$$

which shows that $(u_t^n)_n$ is a Cauchy sequence in $L^2(\hat{\gamma})$. $L^2(\hat{\gamma})$ being complete, $(u_t^n)_n$ converges to a $u_t^\infty \in L^2(\hat{\gamma})$. Moreover, for $\varepsilon > 0$, because (u^n) is Cauchy, we can choose N such that:

$$\forall n, m \geq N, \|u^n - u^m\| \leq \varepsilon. \tag{20}$$

We thus have that:

$$\forall t, \forall n, m \geq N, \|u_t^n - u_t^m\|_{L^2(\hat{\gamma})} \leq \varepsilon. \tag{21}$$

Then, by taking m to ∞, by continuity of the $L^2(\hat{\gamma})$ norm:

$$\forall t, \forall n \geq N, \|u_t^n - u_t^\infty\|_{L^2(\hat{\gamma})} \leq \varepsilon, \tag{22}$$

which means that:

$$\forall n \geq N, \|u^n - u^\infty\| \leq \varepsilon. \tag{23}$$

so that $(u^n)_n$ tends to u^∞.

Moreover, as:

$$\forall n, \|u_t^n\|_{L^2(\hat{\gamma})} \leq 1, \tag{24}$$

we have that $\|u_t^\infty\|_{L^2(\hat{\gamma})} \leq 1$.

Finally, let us consider $s, t \in [0, \delta T]$. We have that:

$$\forall n, \|u_t^\infty - u_s^\infty\|_{L^2(\hat{\gamma})} \leq \|u_t^\infty - u_t^n\|_{L^2(\hat{\gamma})} + \|u_t^n - u_s^n\|_{L^2(\hat{\gamma})} + \|u_s^\infty - u_s^n\|_{L^2(\hat{\gamma})}. \tag{25}$$

The first and the third terms can then be taken as small as needed by definition of u^∞ by taking n high enough, while the second can be made to tend to 0 as t tends to s by continuity of u^n. This proves the continuity of u^∞ and shows that $u^\infty \in \mathcal{F}_{\delta T}$. □

Lemma 2. *For any $F \in L^2(\hat{\gamma})$, we have that $F \in L^2(\hat{\alpha})$ and $F \in L^2\left(\hat{\beta}\right)$ with:*

$$\|F\|_{L^2(\hat{\alpha})} \leq \sqrt{2}\|F\|_{L^2(\hat{\gamma})} \text{ and } \|F\|_{L^2\left(\hat{\beta}\right)} \leq \sqrt{2}\|F\|_{L^2(\hat{\gamma})}. \tag{26}$$

Proof. For any $F \in L^2(\hat{\gamma})$, we have that

$$\|F\|^2_{L^2(\hat{\gamma})} = \frac{1}{2}\|F\|^2_{L^2(\hat{\alpha})} + \frac{1}{2}\|F\|^2_{L^2\left(\hat{\beta}\right)}, \tag{27}$$

so that $F \in L^2(\hat{\alpha})$ and $F \in L^2\left(\hat{\beta}\right)$ with:

$$\|F\|^2_{L^2(\hat{\alpha})} = 2\|F\|^2_{L^2(\hat{\gamma})} - \|F\|^2_{L^2\left(\hat{\beta}\right)} \leq 2\|F\|^2_{L^2(\hat{\gamma})}, \quad \|F\|^2_{L^2\left(\hat{\beta}\right)} = 2\|F\|_{L^2(\hat{\gamma})} - \|F\|_{L^2(\hat{\alpha})} \leq 2\|F\|^2_{L^2(\hat{\gamma})}, \tag{28}$$

which allows us to conclude. □

From this, we can prove the existence and uniqueness of the initial value problem from Equation (9).

Theorem 3 (Existence and Uniqueness)**.** *Under Assumptions 1 to 3, Equation (9) with initial value f_0 admits a unique solution $f_\cdot : \mathbb{R}_+ \to L^2(\Omega)$.*

Proof.

A few inequalities. We start this proof by proving a few inequalities.

Let $f, g \in L^2(\hat{\gamma})$. We have, by the Cauchy-Schwarz inequality, for all $z \in \Omega$:

$$\left|\left(\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f)\right) - \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g)\right)\right)(z)\right| \leq \left\|k(z,\cdot)\right\|_{L^2(\hat{\gamma})}\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g)\right\|_{L^2(\hat{\gamma})}. \tag{29}$$

Moreover, by definition:

$$\left\langle \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g), h\right\rangle_{L^2(\hat{\gamma})} = \int \left(a'_f - a'_g\right) h \,\mathrm{d}\hat{\alpha} - \int \left(b'_f - b'_g\right) h \,\mathrm{d}\hat{\beta}, \tag{30}$$

so that:

$$\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g)\right\|^2_{L^2(\hat{\gamma})} \leq \left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g)\right\|_{L^2(\hat{\gamma})}\left(\left\|a'_f - a'_g\right\|_{L^2(\hat{\alpha})} + \left\|b'_f - b'_g\right\|_{L^2\left(\hat{\beta}\right)}\right), \tag{31}$$

and then, along with Lemma 2:

$$\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g)\right\|_{L^2(\hat{\gamma})} \leq \left\|a'_f - a'_g\right\|_{L^2(\hat{\alpha})} + \left\|b'_f - b'_g\right\|_{L^2\left(\hat{\beta}\right)} \leq \sqrt{2}\left(\left\|a'_f - a'_g\right\|_{L^2(\hat{\gamma})} + \left\|b'_f - b'_g\right\|_{L^2(\hat{\gamma})}\right). \tag{32}$$

By Assumption 3, we know that a' and b' are Lipschitz with constants that we denote K_1 and K_2. We can then write for all x:

$$\left|a'\big(f(x)\big) - a'\big(g(x)\big)\right| \leq K_1\left|f(x) - g(x)\right|, \qquad \left|b'\big(f(x)\big) - b'\big(g(x)\big)\right| \leq K_2\left|f(x) - g(x)\right|, \tag{33}$$

so that:

$$\left\|a'_f - a'_g\right\|_{L^2(\hat{\gamma})} \leq K_1\|f - g\|_{L^2(\hat{\gamma})}, \qquad \left\|b'_f - b'_g\right\|_{L^2(\hat{\gamma})} \leq K_2\|f - g\|_{L^2(\hat{\gamma})}. \tag{34}$$

Finally, we can now write, for all $z \in \Omega$:

$$\left|\left(\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f)\right) - \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g)\right)\right)(z)\right| \leq \sqrt{2}(K_1 + K_2)\|f - g\|_{L^2(\hat{\gamma})}\left\|k(z,\cdot)\right\|_{L^2(\hat{\gamma})}, \tag{A}$$

and then:

$$\left\|\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f)\right) - \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(g)\right)\right\|_{L^2(\hat{\gamma})} \leq K\|f - g\|_{L^2(\hat{\gamma})}, \tag{B}$$

where $K = \sqrt{2}(K_1 + K_2)\sqrt{\int \left\|k(z,\cdot)\right\|^2_{L^2(\hat{\gamma})} \mathrm{d}\hat{\gamma}(z)}$ is finite as a finite sum of finite terms from Assumptions 1 and 2. In particular, putting $g = 0$ and using the triangular inequality also gives us:

$$\left\|\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f)\right)\right\|_{L^2(\hat{\gamma})} \leq K\|f\|_{L^2(\hat{\gamma})} + M, \tag{B'}$$

where $M = \left\|\mathcal{T}_{k,\hat{\gamma}}\big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(0)\big)\right\|_{L^2(\hat{\gamma})}$.

Existence and uniqueness in $L^2(\hat{\gamma})$. We now adapt the standard fixed point proof to prove existence and uniqueness of a solution to the studied equation in $L^2(\hat{\gamma})$.

We consider the family of spaces $\mathcal{F}_{\delta T} = \mathcal{C}\Big([0, \delta T], B_{L^2(\hat{\gamma})}(f_0, 1)\Big)$. $\mathcal{F}_{\delta T}$ is defined, for $\delta T > 0$, as the space of continuous functions from $[0, \delta T]$ to the closed ball of radius 1 centered around f_0 in $L^2(\hat{\gamma})$ which we endow with the norm:

$$\forall u \in \mathcal{F}_{\delta T},\ \|u\| = \sup_{t \in [0, \delta T]} \|u_t\|_{L^2(\hat{\gamma})}. \tag{35}$$

We now define the application Φ where $\Phi(u)$ is defined as, for any $u \in \mathcal{F}_{\delta T}$:

$$\Phi(u)_t = f_0 + \int_0^t \mathcal{T}_{k,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(u_s)\Big)\,\mathrm{d}s. \tag{36}$$

We have, using Equation (B'):

$$\big\|\Phi(u)_t - f_0\big\|_{L^2(\hat{\gamma})} \leq \int_0^t \Big(K\|u_s\|_{L^2(\hat{\gamma})} + M\Big)\,\mathrm{d}s \leq (K+M)\delta T. \tag{37}$$

Thus, taking $\delta T = \big(2(K+M)\big)^{-1}$ makes Φ an application from $\mathcal{F}_{\delta T}$ into itself. Moreover, we have:

$$\forall u, v \in \mathcal{F}_{\delta T}, \big\|\Phi(u) - \Phi(v)\big\| \leq \frac{1}{2}\|u - v\|, \tag{38}$$

which means that Φ is a contraction of $\mathcal{F}_{\delta T}$. Lemma 1 and the Banach-Picard theorem then tell us that Φ has a unique fixed point in $\mathcal{F}_{\delta T}$. It is then obvious that such a fixed point is a solution of Equation (9) over $[0, \delta T]$.

Let us now consider the maximal $T > 0$ such that a solution f_t of Equation (9) is defined over $[0, T)$. We have, using Equation (B'):

$$\forall t \in [0,T), \|f_t\|_{L^2(\hat{\gamma})} \leq \|f_0\|_{L^2(\hat{\gamma})} + \int_0^t \Big(\|f_s\|_{L^2(\hat{\gamma})} + M\Big)\,\mathrm{d}s, \tag{39}$$

which, using Grönwall's lemma, gives:

$$\forall t \in [0,T), \|f_t\|_{L^2(\hat{\gamma})} \leq \|f_0\|_{L^2(\hat{\gamma})}\mathrm{e}^{KT} + \frac{M}{K}\Big(\mathrm{e}^{KT} - 1\Big). \tag{40}$$

Define $g^n = f_{T-\frac{1}{n}}$. We have, again using Equation (B'):

$$\forall m \geq n, \|g^n - g^m\|_{L^2(\hat{\gamma})} \leq \int_{T-\frac{1}{n}}^{T-\frac{1}{m}} (K\|f_s\| + M)\,\mathrm{d}s \leq \left(\frac{1}{n} - \frac{1}{m}\right)\left(\|f_0\|_{L^2(\hat{\gamma})}\mathrm{e}^{KT} + \frac{M}{K}\Big(\mathrm{e}^{KT} - 1\Big)\right), \tag{41}$$

which shows that $(g^n)_n$ is a Cauchy sequence. $L^2(\hat{\gamma})$ being complete, we can thus consider its limit g^∞. Clearly, f_t tends to g^∞ in $L^2(\hat{\gamma})$. By considering the initial value problem associated with Equation (9) starting from g^∞, we can thus extend the solution f_t to $[0, T + \delta T)$, thus contradicting the maximality of T, which proves that the solution can be extended to $\mathbb{R}_+$.

Existence and uniqueness in $L^2(\Omega)$. We now conclude the proof by extending the previous solution to $L^2(\Omega)$. We keep the same notations as above and, in particular, f is the unique solution of Equation (9) with initial value f_0.

Let us define $\tilde{f}$ as:

$$\forall t, \forall x, \tilde{f}_t(x) = f_0(x) + \int_0^t \mathcal{T}_{k,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)\Big)(x)\,\mathrm{d}s, \tag{42}$$

where the r.h.s. only depends on f and is thus well-defined. By remarking that $\tilde{f}$ is equal to f on $\operatorname{supp}\hat{\gamma}$ and that, for every s,

$$\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}\left(\tilde{f}_s\right)\right) = \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}\left(\tilde{f}_s\Big|_{\operatorname{supp}\hat{\gamma}}\right)\right) = \mathcal{T}_{k,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)\Big), \tag{43}$$

we see that $\tilde{f}$ is solution to Equation (9). Moreover, from Assumption 2, we know that, for any $z \in \Omega$, $\int k(z,x)^2\,\mathrm{d}\Omega(x)$ is finite and, from Assumption 1, that $\big\|k(z,\cdot)\big\|^2_{L^2(\hat{\gamma})}$ is a finite sum of terms $k(z,x_i)^2$ which shows that $\int \big\|k(z,\cdot)\big\|^2_{L^2(\hat{\gamma})}\,\mathrm{d}\Omega(z)$ is finite, again from Assumption 2. We can then say that $\tilde{f}_s \in L^2(\Omega)$ for any s by using the above with Equation (A) taken for $g = 0$.

Finally, suppose h is a solution to Equation (9) with initial value f_0. We know that $h|_{\operatorname{supp}\hat{\gamma}}$ coincides with f and thus with $\tilde{f}\Big|_{\operatorname{supp}\hat{\gamma}}$ in $L^2(\hat{\gamma})$ as we already proved uniqueness in the latter space. Thus, we have that $\left\|h_s|_{\operatorname{supp}\hat{\gamma}} - \tilde{f}_s\Big|_{\operatorname{supp}\hat{\gamma}}\right\|_{L^2(\hat{\gamma})} = 0$

for any s. Now, we have:

$$\begin{aligned}\forall z \in \Omega, \forall s, &\left| \left(\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(h_s)\right) - \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}\left(\tilde{f}_s\right)\right)\right)(z)\right| \\ &= \left| \left(\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}\left(h_s|_{\operatorname{supp}\hat{\gamma}}\right)\right) - \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}\left(\tilde{f}_s\Big|_{\operatorname{supp}\hat{\gamma}}\right)\right)\right)(z)\right| \leq 0,\end{aligned} \tag{44}$$

by Equation (A). This shows that $\partial_t\left(\tilde{f} - h\right) = 0$ and, given that $h_0 = \tilde{f}_0 = f_0$, we have $h = \tilde{f}$ which concludes the proof.

□

There only remains to prove for Theorem 1 the inversion between the integral over time and the integral operator. We first prove an intermediate lemma and then conclude with the proof of the inversion.

Lemma 3. *Under Assumptions 1 to 3,* $\int_0^T \left(\left\|a'\right\|_{L^2\left((f_s)_\sharp\hat{\alpha}\right)} + \left\|b'\right\|_{L^2\left((f_s)_\sharp\hat{\beta}\right)}\right) \mathrm{d}s$ *is finite for any* $T > 0$.

Proof. Let $T > 0$. We have, by Assumption 3 and the triangular inequality:

$$\forall x, \left|a'\big(f(x)\big)\right| \leq K_1\left|f(x)\right| + M_1, \tag{45}$$

where $M_1 = \left|a'(0)\right|$. We can then write, using Lemma 2 and the inequality from Equation (40):

$$\forall s \leq T, \left\|a'\right\|_{L^2\left((f_s)_\sharp\hat{\alpha}\right)} \leq K_1\sqrt{2}\|f_s\|_{L^2(\hat{\gamma})} + M_1 \leq K_1\sqrt{2}\left(\|f_0\|_{L^2(\hat{\gamma})}\mathrm{e}^{KT} + \frac{M}{K}\left(\mathrm{e}^{KT} - 1\right)\right) + M_1, \tag{46}$$

the latter being constant in s and thus integrable on $[0, T]$. We can then bound $\left\|b'\right\|_{L^2\left((f_s)_\sharp\hat{\beta}\right)}$ similarly, which concludes the proof. □

Proposition 6 (Integral inversion)**.** *Under Assumptions 1 to 3, the following integral inversion holds:*

$$f_t = f_0 + \int_0^t \mathcal{T}_{k_f,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha},\hat{\beta}}(f_s)\right)\mathrm{d}s = f_0 + \mathcal{T}_{k_f,\hat{\gamma}}\left(\int_0^t \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha},\hat{\beta}}(f_s)\,\mathrm{d}s\right). \tag{47}$$

Proof. By definition, a straightforward computation gives, for any function $h \in L^2(\hat{\gamma})$:

$$\left\langle \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f), h\right\rangle_{L^2(\hat{\gamma})} = \mathrm{d}\mathcal{L}_{\hat{\alpha}}(f)[h] = \int a'_f h\,\mathrm{d}\hat{\alpha} - \int b'_f h\,\mathrm{d}\hat{\beta}. \tag{48}$$

We can then write:

$$\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right\|^2_{L^2(\hat{\gamma})} = \left\langle \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t), \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right\rangle_{L^2(\hat{\gamma})} = \int a'_{f_t}\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\,\mathrm{d}\hat{\alpha} - \int b'_{f_t}\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\,\mathrm{d}\hat{\beta}, \tag{49}$$

so that, with the Cauchy-Schwarz inequality and Lemma 2:

$$\begin{aligned}\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right\|^2_{L^2(\hat{\gamma})} &\leq \int \left|a'_{f_t}\right|\left|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right|\mathrm{d}\hat{\alpha} + \int \left|b'_{f_t}\right|\left|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right|\mathrm{d}\hat{\beta} \\ &\leq \left\|a'_{f_t}\right\|_{L^2(\hat{\alpha})}\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right\|_{L^2(\hat{\alpha})} + \left\|b'_{f_t}\right\|_{L^2(\hat{\beta})}\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right\|_{L^2(\hat{\beta})} \\ &\leq \sqrt{2}\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right\|_{L^2(\hat{\gamma})}\left[\left\|a'_{f_t}\right\|_{L^2(\hat{\alpha})} + \left\|b'_{f_t}\right\|_{L^2(\hat{\beta})}\right],\end{aligned} \tag{50}$$

which then gives us:

$$\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right\|_{L^2(\hat{\gamma})} \leq \sqrt{2}\left[\left\|a'\right\|_{L^2\left((f_t)_\sharp\hat{\alpha}\right)} + \left\|b'\right\|_{L^2\left((f_t)_\sharp\hat{\beta}\right)}\right]. \tag{51}$$

By the Cauchy-Schwarz inequality and Equation (51), we then have for all z:

$$\begin{aligned}\int_0^t \int_x \left|k(z,x)\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)(x)\right| \mathrm{d}\hat{\gamma}(x)\,\mathrm{d}s &\leq \int_0^t \left\|k(z,\cdot)\right\|_{L^2(\hat{\gamma})}\left\|\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)\right\|_{L^2(\hat{\gamma})}\mathrm{d}s \\ &\leq \sqrt{2}\left\|k(z,\cdot)\right\|_{L^2(\hat{\gamma})}\int_0^t\left[\left\|a'\right\|_{L^2\left((f_s)_\sharp\hat{\alpha}\right)} + \left\|b'\right\|_{L^2\left((f_s)_\sharp\hat{\beta}\right)}\right]\mathrm{d}s.\end{aligned} \tag{52}$$

The latter being finite by Lemma 3, we can now use Fubini's theorem to conclude that:

$$\begin{aligned}\int_0^t \mathcal{T}_{k_f,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)\right)\mathrm{d}s &= \int_0^t\int_x k(\cdot,x)\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)(x)\,\mathrm{d}\hat{\gamma}(x)\,\mathrm{d}s \\ &= \int_x k(\cdot,x)\left[\int_0^t \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)(x)\,\mathrm{d}s\right]\mathrm{d}\hat{\gamma}(x) \\ &= \mathcal{T}_{k_f,\hat{\gamma}}\left(\int_0^t \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)(x)\,\mathrm{d}s\right).\end{aligned} \tag{53}$$

□

A.3. Differentiability of Infinite-Width Networks and their NTKs

Given Theorem 1, establishing the desired differentiability of f_t can be done by separately proving similar results on both $f_t - f_0$ and f_0.

In both cases, this involves the differentiability of the following activation kernel $\mathcal{K}_\phi(A)$ given another differentiable kernel A:

$$\mathcal{K}_\phi(A)\colon x, y \mapsto \mathbb{E}_{f\sim\mathcal{GP}(0,A)}\Big[\phi\big(f(x)\big)\phi\big(f(y)\big)\Big], \tag{54}$$

where $\mathcal{GP}(0,A)$ is a univariate centered Gaussian Process (GP) with covariance function A. Indeed, the kernel-transforming operator $\mathcal{K}_\phi$ is central in the recursive computation of the neural network conjugate kernel sss which determines the NTK (involved in $f_t - f_0 \in \mathcal{H}_k^{\hat{\gamma}_g}$) as well as the behavior of the network at initialization (which follows a GP with the conjugate kernel as covariance).

Hence, our proof of Theorem 2 relies on the preservation of kernel smoothness through $\mathcal{K}_\phi$, proved in Appendix A.3.1, which ensures the smoothness of the conjugate kernel, the NTK and, in turn, of f_t as addressed in Appendix A.3.2 which concludes the overall proof.

Before developing these two main steps, we first need to state the following lemma showing the regularity of samples of a GP from the regularity of the corresponding kernel.

Lemma 4 (GP regularity)**.** *Let $A\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a symmetric kernel. Let V an open set such that A is $\mathcal{C}^\infty$ on $V \times V$. Then the GP induced by the kernel A has a.s. $\mathcal{C}^\infty$ sample paths on V.*

Proof. Because A is $\mathcal{C}^\infty$ on $V \times V$, we know, from Theorem 2.2.2 of Adler (1981) for example, that the corresponding GP f is mean-square smooth on V. If we take α a k-th order multi-index, we also know, again from Adler (1981), that $\partial^\alpha f$ is also a GP with covariance kernel $\partial^\alpha A$. As A is $\mathcal{C}^\infty$, $\partial^\alpha A$ then is differentiable and $\partial^\alpha f$ has partial derivatives which are mean-square continuous. Then, by the Corollary 5.3.12 of Scheuerer (2009), we can say that $\partial^\alpha f$ has continuous sample paths a.s. which means that $f \in \mathcal{C}^k(V)$. This proves the lemma. □

A.3.1. $\mathcal{K}_\phi$ Preserves Kernel Differentiability

Given the definition of $\mathcal{K}_\phi(A)$ in Equation (54), we choose to prove its differentiability via the dominated convergence theorem and Leibniz integral rule. This requires to derive separate proofs depending on whether ϕ is smooth everywhere or almost everywhere.

The former case allows us to apply strong GP regularity results leading to $\mathcal{K}_\phi$ preserving kernel smoothness without additional hypothesis in Lemma 5. The latter case requires a careful decomposition of the expectation of Equation (54) via two-dimensional Gaussian sampling to circumvent the non-differentiability points of ϕ, yielding additional constraints on kernels A for $\mathcal{K}_\phi$ to preserve their smoothness in Lemma 6; these constraints are typically verified in the case of neural networks with bias (cf. Appendix A.3.2).

In any case, we emphasize that these differentiability constraints may not be tight and are only sufficient conditions ensuring the smoothness of $\mathcal{K}_\phi(A)$.

Lemma 5 ($\mathcal{K}_\phi$ with smooth ϕ)**.** *Let $A\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a symmetric positive semi-definite kernel and $\phi\colon \mathbb{R} \to \mathbb{R}$. We suppose that ϕ is an activation function following Assumptions 4 and 5; in particular, ϕ is smooth.*

Let $y \in \mathbb{R}^n$ and U be an open subset of $\mathbb{R}^n$ such that $x \mapsto A(x,x)$ and $x \mapsto A(x,y)$ are infinitely differentiable over U. Then, $x \mapsto \mathcal{K}_\phi(A)(x,x)$ and $x \mapsto \mathcal{K}_\phi(A)(x,y)$ are infinitely differentiable over U as well.

Proof. In order to prove the smoothness results over the open set U, it suffices to prove them on any open bounded subset of U. Let then $V \subseteq U$ be an open bounded set. Without loss of generality, we can assume that its closure $\operatorname{cl} V$ is also included in U.

We define B_1 and B_2 from Equation (54) as follows, for all $x \in V$:

$$B_1(x) \triangleq \mathcal{K}_\phi(A)(x,y) = \mathbb{E}_{f\sim\mathcal{GP}(0,A)}\Big[\phi\big(f(x)\big)\phi\big(f(y)\big)\Big], \quad B_2(x) \triangleq \mathcal{K}_\phi(A)(x,x) = \mathbb{E}_{f\sim\mathcal{GP}(0,A)}\Big[\phi\big(f(x)\big)^2\Big]. \tag{55}$$

In the previous expressions, Lemma 4 tells us that we can take f to be $\mathcal{C}^\infty$ over $\operatorname{cl} V$ with probability one. Hence, B_1 and B_2 are expectations of smooth functions over V. We seek to apply the dominated convergence theorem to prove that B_1 and B_2 are, in turn, smooth over V. To this end, we prove in the following the integrability of the derivatives of their integrands.

Let $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$. Using the usual notations for multi-indexed partial derivatives, via a multivariate Faà di Bruno formula (Leipnik & Pearce, 2007), we can write the derivatives $\partial^\alpha(\psi \circ f)$ at $x \in V$ for $\psi \in \{\phi, \phi^2\}$ as a weighted sum of terms of the form:

$$\psi^{(j)}\big(f(x)\big) g_1(x) \cdots g_N(x), \tag{56}$$

where the g_is are partial derivatives of f at x. As A is $\mathcal{C}^\infty$ over V, each of the g_is is thus a GP with a $\mathcal{C}^\infty$ covariance function by Lemma 4. We can also write for all $x \in V$:

$$\begin{aligned}\Big|\psi^{(j)}\big(f(x)\big) g_1(x) \cdots g_N(x)\Big| &\leq \sup_{z\in\operatorname{cl} V}\Big|\psi^{(j)}\big(f(z)\big) g_1(z) \cdots g_N(z)\Big| \\ &\leq \sup_{z_0\in\operatorname{cl} V}\Big|\psi^{(j)}\big(f(z_0)\big)\Big| \sup_{z_1\in\operatorname{cl} V}\big|g_1(z_1)\big| \cdots \sup_{z_N\in\operatorname{cl} V}\big|g_N(z_N)\big|.\end{aligned} \tag{57}$$

For each i, because the covariance function of g_i is smooth over the compact set $\operatorname{cl} V$, its variance admits a maximum in $\operatorname{cl} V$ and we take σ_i^2 the double of its value. We then know from Adler (1990), that there is an M_i such that:

$$\forall m \in \mathbb{N}, \mathbb{E}_{f\sim\mathcal{GP}(0,A)}\left[\sup_{z_i\in\operatorname{cl} V}\big|g_i(z_i)\big|^m\right] \leq M_i^m \mathbb{E}|Y_i|^m, \tag{58}$$

where Y_i is a Gaussian distribution which variance is σ_i^2, the right-hand side thus being finite.

We also have, by Assumption 4 from Appendix A.1, that:

$$\sup_{z\in\operatorname{cl} V}\Big|\phi^{(j)}\big(f(z)\big)\Big|^2 \leq \sup_{z\in\operatorname{cl} V}\Big(\lambda_1^{(j)}\big|f(z)\big| + \lambda_2^{(j)}\Big)^2, \tag{59}$$

which is shown to be integrable over f by the same arguments as for the g_is. Moreover, the Faà di Bruno formula decomposes $\psi^{(j)}$ when $\psi = \phi^2$ as a weighted sum of terms of the form $\phi^{(l)}\phi^{(l')}$ with $l, l' \in \mathbb{N}$. Therefore, thanks to similar arguments, for any $\psi \in \{\phi, \phi^2\}$:

$$\mathbb{E}_{f\sim\mathcal{GP}(0,A)}\left[\sup_{z\in\operatorname{cl} V}\Big|\psi^{(j)}\big(f(z)\big)\Big|^2\right] < \infty. \tag{60}$$

Now, by using the Cauchy-Schwarz inequality, we have that:

$$\begin{aligned}&\mathbb{E}\left[\sup_{z_0\in\operatorname{cl} V}\left|\psi^{(j)}\big(f(z_0)\big)\right|\sup_{z_1\in\operatorname{cl} V}\left|g_1(z_1)\right|\cdots\sup_{z_N\in\operatorname{cl} V}\left|g_N(z_N)\right|\right]\\&\leq\sqrt{\mathbb{E}\left[\sup_{z_0\in\operatorname{cl} V}\left|\psi^{(j)}\big(f(z_0)\big)\right|^2\right]}\sqrt{\mathbb{E}\left[\sup_{z_1\in\operatorname{cl} V}\left|g_1(z_1)\right|^2\cdots\sup_{z_N\in\operatorname{cl} V}\left|g_N(z_N)\right|^2\right]}.\end{aligned}\tag{61}$$

By iterated applications of the Cauchy-Schwarz inequality and using the previous arguments, we can then show that:

$$\sup_{z_0\in\operatorname{cl} V}\left|\psi^{(j)}\big(f(z_0)\big)\right|\sup_{z_1\in\operatorname{cl} V}\left|g_1(z_1)\right|\cdots\sup_{z_N\in\operatorname{cl} V}\left|g_N(z_N)\right|\tag{62}$$

is integrable over f. Additionally, note that by the same arguments for the case of $\psi=\phi$, a multiplication by $\phi\big(f(y)\big)$ preserves this integrability.

We can then write for all $x\in V$, by a standard corollary of the dominated convergence theorem:

$$\partial^\alpha B_1(x)=\mathbb{E}_{f\sim\mathcal{GP}(0,A)}\Big[\partial^\alpha\left(\phi\circ f\right)\big|_x\phi\big(f(y)\big)\Big],\qquad\partial^\alpha B_2(x)=\mathbb{E}_{f\sim\mathcal{GP}(0,A)}\left[\partial^\alpha\left(\phi^2\circ f\right)\Big|_x\right],\tag{63}$$

which shows that B_1 and B_2 are $\mathcal{C}^\infty$ over V. This in turn means that B_1 and B_2 are $\mathcal{C}^\infty$ over U. □

Lemma 6 ($\mathcal{K}_\phi$ with piecewise smooth ϕ)**.** *Let $A\colon\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ be a symmetric positive semi-definite kernel and $\phi\colon\mathbb{R}\to\mathbb{R}$. We suppose that ϕ is an activation function following Assumptions 4 and 6 (cf. Appendix A.1). Let us define the matrix $\Sigma_A^{x,y}$ as:*

$$\Sigma_A^{x,y}\triangleq\begin{pmatrix}A(x,x)&A(x,y)\\A(x,y)&A(y,y)\end{pmatrix}.\tag{64}$$

Let $y\in\mathbb{R}^n$ and U be an open subset of $\mathbb{R}^n$ such that $x\mapsto A(x,x)$ and $x\mapsto A(x,y)$ are infinitely differentiable over U. Then, $x\mapsto\mathcal{K}_\phi(A)(x,x)$ and $x\mapsto\mathcal{K}_\phi(A)(x,y)$ are infinitely differentiable over $U'\triangleq\big\{x\in U\;\big|\;\Sigma_A^{x,y}$ is invertible$\big\}$.

Proof. Since $\det\Sigma_A^{x,y}$ is smooth over U and $U'=\big\{x\in U\;\big|\;\det\Sigma_A^{x,y}>0\big\}$, U' is an open subset of U. Hence, similarly to the proof of Lemma 5, it suffices to prove the smoothness of B_1 and B_2 defined in Equation (55) on any open bounded subset of U'. Let then $V\subseteq\mathbb{R}^n$ be an open bounded set such that $\operatorname{cl} V\subseteq U'$. Note that $\det\Sigma_A^{x,y}>0$ implies that $A(x,x)>0$ and $A(y,y)>0$.

We will conduct in the following the proof that B_1 is smooth over V. Like in the proof of Lemma 5, the smoothness of B_2 follows the same reasoning with little adaptation; in particular, it relies on the fact that $A(x,x)>0$ for all $x\in U'$, making its square root smooth over $\operatorname{cl} V$.

Since the dominated convergence theorem cannot be directly applied from Equation (55) because of ϕ's potential non-differentiability points D, let us decompose its expression for all $x\in U'$:

$$\begin{aligned}B_1(x)&=\mathbb{E}_{f\sim\mathcal{GP}(0,A)}\Big[\phi\big(f(x)\big)\phi\big(f(y)\big)\Big]=\mathbb{E}_{(z,z')\sim\mathcal{N}\left((0,0),\Sigma_A^{x,y}\right)}\Big[\phi(z)\phi\big(z'\big)\Big]&(65)\\&=\mathbb{E}_{z'\sim\mathcal{N}\left(0,A(y,y)\right)}\left[\phi\big(z'\big)\mathbb{E}_{z\sim\mathcal{N}\left(\frac{A(x,y)}{A(y,y)}z',A(x,x)-\frac{A(x,y)^2}{A(y,y)}\right)}\big[\phi(z)\big]\right]&(66)\\&=\mathbb{E}_{z'\sim\mathcal{N}\left(0,A(y,y)\right)}\Big[\phi\big(z'\big)h\big(z',x\big)\Big],&(67)\end{aligned}$$

where h is defined as:

$$h\big(z',x\big)\triangleq\int_{-\infty}^{+\infty}\phi(z)\cdot\frac{1}{\sigma(x)\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu(x)z'}{\sigma(x)}\right)^2}\,\mathrm{d}z,\qquad\mu(x)=\frac{A(x,y)}{A(y,y)},\qquad\sigma(x)=\sqrt{\frac{\det\Sigma_A^{x,y}}{A(y,y)}}.\tag{68}$$

Now, if $D = \{c_1, \dots, c_L\}$ with $L \in \mathbb{N}$ and $c_1 < \cdots < c_L$, the c_ls constitute the non-differentiability points of ϕ; we can then decompose the integration of ϕ in Equation (68) as a sum of $L+1$ integrals with differentiable integrands, using $c_0 = -\infty$ and $c_{L+1} = +\infty$:

$$h(\varepsilon, x) = \frac{1}{\sqrt{2\pi}} \sum_{l=0}^{L} \int_{c_l}^{c_{l+1}} \frac{\phi(z)}{\sigma(x)} \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu(x)z'}{\sigma(x)}\right)^2} \,\mathrm{d}z. \tag{69}$$

Therefore, it remains to show the smoothness of all applications $B_{1,l}$ for $l \in [\![0, L]\!]$ defined as:

$$B_{1,j}(x) = \mathbb{E}_{z' \sim \mathcal{N}\left(0, A(y,y)\right)} \left[\int_{c_l}^{c_{l+1}} \frac{\phi\left(z'\right)\phi(z)}{\sigma(x)\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu(x)z'}{\sigma(x)}\right)^2} \,\mathrm{d}z \right]. \tag{70}$$

The rest of this proof unfolds similarly to the one of Lemma 5. Indeed, the integrand of Equation (70) is smooth over $\operatorname{cl} V$. There remains to show that all derivatives of this integrand are dominated by an integrable function of z and z'. Consider the following integrand:

$$\iota\left(z, z', x\right) = \frac{\phi\left(z'\right)\phi(z)}{\sigma(x)\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu(x)z'}{\sigma(x)}\right)^2}. \tag{71}$$

By applying the multivariate Faà di Bruno formula and noticing that σ and μ are smooth over the closed set $\operatorname{cl} V$, we know that the derivatives of $\iota\left(z, z', x\right)$ with respect to x for any derivation order are weighted sums of terms of the form:

$$z^k z'^{k'} \kappa(x) \frac{\phi\left(z'\right)\phi(z)}{\sigma(x)\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu(x)z'}{\sigma(x)}\right)^2}, \tag{72}$$

where κ is a smooth function over $\operatorname{cl} V$ and $k, k' \in \mathbb{N}$. Moreover, because σ, μ and κ are smooth over the closed set $\operatorname{cl} V$ with positive values for σ, there are constants a_1, a_2 and a_3 such that:

$$\left| z^k z'^{k'} \kappa(x) \frac{\phi\left(z'\right)\phi(z)}{\sigma(x)\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu(x)z'}{\sigma(x)}\right)^2} \right| \leq \left| z^k z'^{k'} \phi\left(z'\right)\phi(z) \right| a_3 \mathrm{e}^{-\frac{1}{2}\left(\frac{z-a_1 z'}{a_2}\right)^2}, \tag{73}$$

which is integrable over z via Assumption 4 and Equation (16). Finally, let us notice that for some constants b_1, b_2 and b_3:

$$\int_{c_l}^{c_{l+1}} \left| z^k z'^{k'} \phi\left(z'\right)\phi(z) \right| a_3 \mathrm{e}^{-\frac{1}{2}\left(\frac{z-a_1 z'}{a_2}\right)^2} \leq b_1 \mathbb{E}_{z \sim \mathcal{N}(b_2 z', b_3)} \left| z^k z'^{k'} \phi\left(z'\right)\phi(z) \right|, \tag{74}$$

which is also integrable with respect to $\mathbb{E}_{z' \sim \mathcal{N}\left(0, A(y,y)\right)}$ by similar arguments (see also the integrability of Equation (58) in Lemma 5). This concludes the proof of integrability required to apply the dominated convergence theorem, allowing us to conclude about the smoothness of all $B_{1,j}$ and, in turn, of B_1 over U'. □

Remark 6 (Relaxed condition for smoothness)**.** The invertibility condition of Lemma 6 is actually stronger than needed: it suffices to assume that the rank of $\Sigma_A^{x,y}$ remains constant in a neighborhood of x.

A.3.2. Differentiability of Conjugate Kernels, NTKs and Discriminators

From the previous lemmas, we can then prove the results of Section 4.3. We start by demonstrating the smoothness of the conjugate kernel for dense networks, and conclude in consequence about the smoothness of the NTK and trained network.

Lemma 7 (Differentiability of the conjugate kernel)**.** *Let k_{c} be the conjugate kernel (Lee et al., 2018) of an infinite-width dense non-residual architecture such as in Assumption 4. For any $y \in \mathbb{R}^n$, the following holds for $A \in \left\{k_{\mathrm{c}}, \mathcal{K}_{\phi'}(k_{\mathrm{c}})\right\}$:*

- *if Assumption 5 holds, then $x \mapsto A(x,x)$ and $x \mapsto A(x,y)$ are smooth everywhere over $\mathbb{R}^n$;*
- *if Assumption 6 holds, then $x \mapsto A(x,x)$ and $x \mapsto A(x,y)$ are smooth over an open set whose complement has null Lebesgue measure.*

Proof. We define the following kernel:

$$C_L^\phi(x,y) = \mathbb{E}_{f\sim\mathcal{GP}\left(0,C_{L-1}^\phi\right)}\Big[\phi\big(f(x)\big)\phi\big(f(y)\big)\Big] + \beta^2 = \mathcal{K}_\phi\Big(C_{L-1}^\phi\Big) + \beta^2, \tag{75}$$

with:

$$C_0^\phi(x,y) = \frac{1}{n}x^\top y + \beta^2. \tag{76}$$

We have that $k_\mathrm{c} = C_L^\phi$, with L being the number of hidden layers in the network. Therefore, Lemma 5 ensures the smoothness result under Assumption 5.

Let us now consider Assumption 6 (cf. the detailed assumption in Appendix A.1); in particular, $\beta > 0$. We prove by induction over L in the following that:

- $B_1 \colon x \mapsto C_L^\phi(x,y)$ is smooth over $U = \big\{x \in \mathbb{R}^n \;\big|\; \|x\| \neq \|y\|\big\}$;
- $B_2 \colon x \mapsto C_L^\phi(x,x)$ is smooth;
- for all $x, x' \in \mathbb{R}^n$ with $\|x\| \neq \big\|x'\big\|$, $B_2(x) \neq B_2\big(x'\big)$.

The result is immediate for $L = 0$. We now suppose that it holds for some $L \in \mathbb{N}$ and prove that it also holds for $L+1$ hidden layers. Let us express B_2:

$$B_2(x) = \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\left[\phi\left(\varepsilon\sqrt{C_L^\phi(x,x)+\beta^2}\right)^2\right]. \tag{77}$$

Using Lemma 6 and Remark 6, the fact that $\beta > 0$ and the induction hypothesis ensures that B_2 is smooth. Moreover, Assumption 6, in particular Equation (17), allows us to assert that $\|x\| \neq \big\|x'\big\|$ implies $B_2(x) \neq B_2\big(x'\big)$.

Finally, in order to apply Lemma 6 to prove the smoothness of B_1 over U, there remains to show that the following matrix is invertible:

$$\Sigma_\beta^{x,y} \triangleq \begin{pmatrix} C_L^\phi(x,x)+\beta^2 & C_L^\phi(x,y)+\beta^2 \\ C_L^\phi(x,y)+\beta^2 & C_L^\phi(y,y)+\beta^2 \end{pmatrix}. \tag{78}$$

Let us compute its determinant:

$$\begin{aligned} \det\Sigma_\beta^{x,y} &= \Big(C_L^\phi(x,x)+\beta^2\Big)\Big(C_L^\phi(y,y)+\beta^2\Big) - \Big(C_L^\phi(x,y)+\beta^2\Big)^2 \\ &= \det\Sigma_0^{x,y} + \beta^2\Big(C_L^\phi(x,x) + C_L^\phi(y,y) - 2C_L^\phi(x,y)\Big). \end{aligned} \tag{79}$$

C_L^ϕ is a symmetric positive semi-definite kernel, thus:

$$\det\Sigma_\beta^{x,y} - \det\Sigma_0^{x,y} = \beta^2 \cdot \begin{pmatrix}1 & -1\end{pmatrix}\Sigma_0^{x,y}\begin{pmatrix}1\\-1\end{pmatrix} \geq 0. \tag{80}$$

Hence, if $\det\Sigma_0^{x,y} > 0$, then $\det\Sigma_\beta^{x,y} > 0$. Besides, if $\det\Sigma_0^{x,y} = 0$, then:

$$\det\Sigma_\beta^{x,y} = \beta^2\Big(\sqrt{B_2(x)} - \sqrt{B_2(y)}\Big)^2 > 0, \tag{81}$$

for all $x \in U$. This proves that B_1 is indeed smooth over U, and concludes the induction.

Note that U is indeed an open set whose complement in $\mathbb{R}^n$ has null Lebesgue measure. Overall, the result is thus proved for $A = k_\mathrm{c}$; a similar reasoning using the previous induction result also transfers the result to $A = \mathcal{K}_{\phi'}(k_\mathrm{c})$. □

Proposition 2 (Differentiability of k)**.** Let k be the NTK of an infinite-width architecture following Assumption 4. For any $y \in \mathbb{R}^n$:

- if Assumption 5 holds, then $k(\cdot, y)$ is smooth everywhere over $\mathbb{R}^n$;
- if Assumption 6 holds, then $k(\cdot, y)$ is smooth almost everywhere over $\mathbb{R}^n$, in particular over an open set whose complement has null Lebesgue measure.

Proof. According to the definitions of Jacot et al. (2018), Arora et al. (2019) and Huang et al. (2020), the smoothness of the kernel is guaranteed whenever the conjugate kernel k_c and its transform $\mathcal{K}_{\phi'}(k_c)$ are smooth; the result of Lemma 7 then applies. In the case of residual networks, there is a slight adaptation of the formula which does not change its regularity. Regarding convolutional networks, their conjugate kernels and NTKs involve finite combinations of such dense conjugate kernels and NTKs, thereby preserving their smoothness almost everywhere. □

Theorem 2 (Differentiability of f_t)**.** Let f_t be a solution to Equation (9) under Assumptions 1 and 3 by Theorem 1, with k the NTK of an infinite-width neural network and f_0 an initialization of the latter.

Then, under Assumptions 4 and 5, f_t is smooth everywhere. Under Assumptions 4 and 6, f_t is smooth almost everywhere, in particular over an open set whose complement has null Lebesgue measure.

Proof. From Theorem 1, we have:

$$f_t - f_0 = \mathcal{T}_{k,\hat{\gamma}}\left(\int_0^t \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s)\,\mathrm{d}s\right). \tag{82}$$

We observe that $\mathcal{T}_{k,\hat{\gamma}}(h)$ has, for any $h \in L^2(\hat{\gamma})$, a regularity which only depends on the regularity of $k(\cdot, y)$ for $y \in \operatorname{supp}\hat{\gamma}$. Indeed, if $k(\cdot, y)$ is smooth in a certain neighborhood V for every such y, we can bound $\partial^\alpha k(\cdot, y)$ over V for every y and any multi-index α and then use dominated convergence to prove that $\mathcal{T}_{k,\hat{\gamma}}(h)(\cdot)$ is smooth over V. Therefore, the regularity of $k(\cdot, y)$ transfers to $f_t - f_0$. Given Proposition 2, there remains to prove the same result for f_0.

The theorem then follows from the fact that f_0 has the same regularity as its conjugate kernel k_c thanks to Lemma 4 because f_0 is a sample from the GP with kernel k_c. Lemma 7 shows the smoothness almost everywhere over an open set of applications $x \mapsto k_c(x, y)$; to apply Lemma 4 and concludes this proof, this result must be generalized to prove the smoothness of k_c with respect to both its inputs. This can be done by generalizing the proofs of Lemmas 5 and 6 to show the smoothness of kernels with respect to both x and y, with the same arguments than for x alone. □

Remark 7. In the previous theorem, f_0 is considered to be the initialization of the network. However, we highlight that, without loss of generality, this theorem encompasses the change of training distribution $\hat{\gamma}$ during GAN training. Indeed, as explained in Section 4.1, f_0 after j steps of generator training can actually be decomposed as, for some $h_k \in L^2(\hat{\gamma}_k)$, $k \in [\![1, j]\!]$:

$$f_0 = f^0 + \sum_{k=1}^{j} \mathcal{T}_{k,\hat{\gamma}_k}(h_k), \tag{83}$$

by taking into account the updates of the discriminators over the whole GAN optimization process. The proof of Theorem 2 can then be applied similarly in this case by showing the differentiability of $f_0 - f^0$ on the one hand and of f^0, being the initialization of the discriminator at the very beginning of GAN training, on the other hand.

A.4. Dynamics of the Generated Distribution

We derive in this proposition the differential equation governing the dynamics of the generated distribution.

Proposition 3 (Dynamics of α_ℓ)**.** Under Assumptions 4 and 5, Equation (3) is well-posed. Let us consider its continuous-time version with discriminators trained on discrete distributions as described above:

$$\partial_\ell \theta_\ell = -\mathbb{E}_{z\sim p_z}\left[\nabla_\theta g_\ell(z)^\top \nabla_x\, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x)\Big|_{x=g_\ell(z)}\right]. \tag{84}$$

This yields, with k_{g_ℓ} the NTK of the generator g_ℓ:

$$\partial_\ell g_\ell = -\mathcal{T}_{k_{g_\ell},p_z}\left(z \mapsto \nabla_x\, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x)\Big|_{x=g_\ell(z)}\right). \tag{85}$$

Equivalently, the following continuity equation holds for the joint distribution $\alpha_\ell^z \triangleq (\mathrm{id}, g_\ell)_\sharp p_z$:

$$\partial_\ell \alpha_\ell^z = -\nabla_x \cdot \left(\alpha_\ell^z \mathcal{T}_{k_{g_\ell}, p_z} \left(z \mapsto \nabla_x \, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x) \Big|_{x=g_\ell(z)} \right) \right). \tag{86}$$

Proof. Assumptions 4 and 5 ensure, via Proposition 2 and Theorem 2 that the trained discriminator is differentiable everywhere at all times, whatever the state of the generator. Therefore, Equation (3) is well-posed.

By following Mroueh et al. (2019, Equation (5))'s reasoning on a similar equation, Equation (84) yields the following generator dynamics for all inputs $z \in \mathbb{R}^d$:

$$\partial_\ell g_\ell(z) = -\mathbb{E}_{z' \sim p_z} \left[\nabla_{\theta_\ell} g_\ell(z)^\top \nabla_{\theta_\ell} g_\ell\big(z'\big) \nabla_x \, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x) \Big|_{x=g_\ell(z')} \right]. \tag{87}$$

We recognize the NTK k_{g_ℓ} of the generator as:

$$k_{g_\ell}\big(z, z'\big) \triangleq \nabla_{\theta_\ell} g_\ell(z)^\top \nabla_{\theta_\ell} g_\ell\big(z'\big). \tag{88}$$

From this, we obtain the dynamics of the generator:

$$\partial_\ell g_\ell = -\mathcal{T}_{k_{g_\ell}, p_z} \left(z \mapsto \nabla_x \, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x) \Big|_{x=g_\ell(z)} \right). \tag{89}$$

In other words, the transported particles $\big(z, g_\ell(z)\big)$ have trajectories X_ℓ which are solutions of the Ordinary Differential Equation (ODE):

$$\frac{\mathrm{d}X_\ell}{\mathrm{d}\ell} = \big(0, v_\ell(X_\ell)\big), \tag{90}$$

where:

$$v_\ell = -\mathcal{T}_{k_{g_\ell}, p_z} \left(z \mapsto \nabla_x \, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x) \Big|_{x=g_\ell(z)} \right). \tag{91}$$

Then, because $\alpha_\ell^z \triangleq (\mathrm{id}, g_\ell)_\sharp p_z$ is the induced transported density, following Ambrosio & Crippa (2014), whenever the ODE above is well-defined and has unique solutions (which is necessarily the case for any trained g), α_ℓ^z verifies the continuity equation with the velocity field v_ℓ:

$$\begin{aligned} \partial_\ell \alpha_\ell^z &= -\nabla_{z,x} \cdot \left(\alpha_\ell^z \left(0, -\mathcal{T}_{k_{g_\ell}, p_z} \left(z \mapsto \nabla_x \, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x) \Big|_{x=g_\ell(z)} \right) \right) \right) \\ &= \nabla_x \cdot \left(\alpha_\ell^z \mathcal{T}_{k_{g_\ell}, p_z} \left(z \mapsto \nabla_x \, c_{f^\star_{\hat{\alpha}_{g_\ell}}}(x) \Big|_{x=g_\ell(z)} \right) \right). \end{aligned} \tag{92}$$

This yields the desired result. □

A.5. Optimality in Concave Setting

We derive an optimality result for concave bounded loss functions of the discriminator and positive definite kernels.

A.5.1. Assumptions

We first assume that the NTK is positive definite over the training dataset.

Assumption 7 (Positive definite kernel)**.** *k is positive definite over $\hat{\gamma}$.*

This positive definiteness property equates for finite datasets to the invertibility of the mapping

$$\begin{aligned} \mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}} \colon L^2(\hat{\gamma}) &\to L^2(\hat{\gamma}) \\ h &\mapsto \mathcal{T}_{k,\hat{\gamma}}(h)\big|_{\operatorname{supp}\hat{\gamma}} \end{aligned}, \tag{93}$$

that can be seen as a multiplication by the invertible Gram matrix of k over $\hat{\gamma}$. We further discuss this hypothesis in Appendix B.5.

We also assume the following properties on the discriminator loss function.

Assumption 8 (Concave loss)**.** *$\mathcal{L}_{\hat{\alpha}}$ is concave and bounded from above, and its supremum is reached on a unique point $y^\star$ in $L^2(\hat{\gamma})$.*

Moreover, we need for the sake of the proof a uniform continuity assumption on the solution to Equation (9).

Assumption 9 (Solution continuity)**.** *$t \mapsto f_t|_{\operatorname{supp}\hat{\gamma}}$ is uniformly continuous over $\mathbb{R}_+$.*

Note that these assumptions are verified in the case of LSGAN, which is the typical application of the optimality results that we prove in the following.

A.5.2. Optimality Result

Proposition 7 (Asymptotic optimality)**.** *Under Assumptions 1 to 3 and 7 to 9, f_t converges pointwise when $t \to \infty$, and:*

$$\mathcal{L}_{\hat{\alpha}}(f_t) \xrightarrow[t\to\infty]{} \mathcal{L}_{\hat{\alpha}}(y^\star), \qquad f_\infty = f_0 + \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}}\left(y^\star - f_0|_{\operatorname{supp}\hat{\gamma}}\right)\right), \qquad f_\infty|_{\operatorname{supp}\hat{\gamma}} = y^\star, \tag{94}$$

where we recall that:

$$y^\star = \operatorname*{arg\,max}_{y \in L^2(\hat{\gamma})} \mathcal{L}_{\hat{\alpha}}(y). \tag{95}$$

This result ensures that, for concave losses such as LSGAN, the optimum for $\mathcal{L}_{\hat{\alpha}}$ in $L^2(\Omega)$ is reached for infinite training times by neural network training in the infinite-width regime when the NTK of the discriminator is positive definite. However, this also provides the expression of the optimal network outside $\operatorname{supp}\hat{\gamma}$ thanks to the smoothing of $\hat{\gamma}$.

In order to prove this proposition, we need the following intermediate results: the first one about the functional gradient of $\mathcal{L}_{\hat{\alpha}}$ on the solution f_t; the second one about a direct application of positive definite kernels showing that one can retrieve $f \in \mathcal{H}_k^{\hat{\gamma}}$ over all Ω from its restriction to $\operatorname{supp}\hat{\gamma}$.

Lemma 8. *Under Assumptions 1 to 3 and 7 to 9, $\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t) \to 0$ when $t \to \infty$. Since $\operatorname{supp}\hat{\gamma}$ is finite, this limit can be interpreted pointwise.*

Proof. Assumptions 1 to 3 ensure the existence and uniqueness of f_t, by Theorem 1.

$t \mapsto \hat{f}_t \triangleq f_t|_{\operatorname{supp}\hat{\gamma}}$ and $\mathcal{L}_{\hat{\alpha}}$ being differentiable, $t \mapsto \mathcal{L}_{\hat{\alpha}}(f_t)$ is differentiable, and:

$$\partial_t \mathcal{L}_{\hat{\alpha}}(f_t) = \left\langle \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t), \partial_t \hat{f}_t \right\rangle_{L^2(\hat{\gamma})} = \left\langle \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t), \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right) \right\rangle_{L^2(\hat{\gamma})}, \tag{96}$$

using Equation (9). This equates to:

$$\partial_t \mathcal{L}_{\hat{\alpha}}(f_t) = \left\| \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\right) \right\|^2_{\mathcal{H}_k^{\hat{\gamma}}} \geq 0, \tag{97}$$

where $\|\cdot\|_{\mathcal{H}_k^{\hat{\gamma}}}$ is the semi-norm associated to the RKHS $\mathcal{H}_k^{\hat{\gamma}}$. Note that this semi-norm is dependent on the restriction of its input to $\operatorname{supp}\hat{\gamma}$ only. Therefore, $t \mapsto \mathcal{L}_{\hat{\alpha}}(f_t)$ is increasing. Since $\mathcal{L}_{\hat{\alpha}}$ is bounded from above, $t \mapsto \mathcal{L}_{\hat{\alpha}}(f_t)$ admits a limit when $t \to \infty$.

We now aim at proving from the latter fact that $\partial_t \mathcal{L}_{\hat{\alpha}}(f_t) \to 0$ when $t \to \infty$. We notice that $\|\cdot\|^2_{\mathcal{H}_k^{\hat{\gamma}}}$ is uniformly continuous over $L^2(\hat{\gamma})$ since $\operatorname{supp}\hat{\gamma}$ is finite, $\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}$ is uniformly continuous over $L^2(\hat{\gamma})$ since a' and b' are Lipschitz-continuous,

$\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}$ is uniformly continuous as it amounts to a finite matrix multiplication, and Assumption 9 gives that $t \mapsto f_t|_{\operatorname{supp}\hat{\gamma}}$ is uniformly continuous over $\mathbb{R}_+$. Therefore, their composition $t \mapsto \partial_t \mathcal{L}_{\hat{\alpha}}(f_t)$ (from Equation (97)) is uniformly continuous over $\mathbb{R}_+$. Using Barbălat's Lemma (Farkas & Wegner, 2016), we conclude that $\partial_t \mathcal{L}_{\hat{\alpha}}(f_t) \to 0$ when $t \to \infty$.

Furthermore, k is positive definite over $\hat{\gamma}$ by Assumption 7, so $\|\cdot\|_{\mathcal{H}_k^{\hat{\gamma}}}$ is actually a norm. Therefore, since $\operatorname{supp}\hat{\gamma}$ is finite, the following pointwise convergence holds:

$$\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t) \xrightarrow[t\to\infty]{} 0. \tag{98}$$

□

Lemma 9 ($\mathcal{H}_k^{\hat{\gamma}}$ determined by $\operatorname{supp}\hat{\gamma}$)**.** *Under Assumptions 1, 2 and 7, for all $f \in \mathcal{H}_k^{\hat{\gamma}}$, the following holds:*

$$f = \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}\left(f|_{\operatorname{supp}\hat{\gamma}}\right)\right). \tag{99}$$

Proof. Since k is positive definite by Assumption 7, then $\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}$ from Equation (93) is invertible. Let $f \in \mathcal{H}_k^{\hat{\gamma}}$. Then, by definition of the RKHS in Definition 2, there exists $h \in L^2(\hat{\gamma})$ such that $f = \mathcal{T}_{k,\hat{\gamma}}(h)$. In particular, $f|_{\operatorname{supp}\hat{\gamma}} = \mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}(h)$, hence $h = \mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}\left(f|_{\operatorname{supp}\hat{\gamma}}\right)$. □

We can now prove the desired proposition.

Proof of Proposition 7. Let us first show that f_t converges to the optimum $y^\star$ in $L^2(\hat{\gamma})$. By applying Lemma 8, we know that $\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t) \to 0$ when $t \to \infty$. Given that the supremum of the differentiable concave function $\mathcal{L}_{\hat{\alpha}} \colon L^2(\hat{\gamma}) \to \mathbb{R}$ is achieved at a unique point $y^\star \in L^2(\hat{\gamma})$ with finite $\operatorname{supp}\hat{\gamma}$, then the latter convergence result implies that $\hat{f}_t \triangleq f_t|_{\operatorname{supp}\hat{\gamma}}$ converges pointwise to $y^\star$ when $t \to \infty$.

Given this convergence in $L^2(\hat{\gamma})$, we can deduce convergence on the whole domain Ω by noticing that $f_t - f_0 \in \mathcal{H}_k^{\hat{\gamma}}$, from Corollary 1. Thus, using Lemma 9:

$$f_t - f_0 = \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}\left((f_t - f_0)\big|_{\operatorname{supp}\hat{\gamma}}\right)\right). \tag{100}$$

Again, since $\operatorname{supp}\hat{\gamma}$ is finite, and $\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}$ can be expressed as a matrix multiplication, the fact that f_t converges to $y^\star$ over $\operatorname{supp}\hat{\gamma}$ implies that:

$$\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}\left((f_t - f_0)\big|_{\operatorname{supp}\hat{\gamma}}\right) \xrightarrow[t\to\infty]{} \mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}\left(y^\star - f_0|_{\operatorname{supp}\hat{\gamma}}\right). \tag{101}$$

Finally, using the definition of the integral operator in Definition 2, the latter convergence implies the following desired pointwise convergence:

$$f_t \xrightarrow[t\to\infty]{} f_0 + \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}\left(y^\star - f_0|_{\operatorname{supp}\hat{\gamma}}\right)\right). \tag{102}$$

We showed at the beginning of this proof that f_t converges to the optimum $y^\star$ in $L^2(\hat{\gamma})$, so $\mathcal{L}_{\hat{\alpha}}(f_t) \to \mathcal{L}_{\hat{\alpha}}(y^\star)$ by continuity of $\mathcal{L}_{\hat{\alpha}}$ as claimed in the proposition. □

A.6. Case Studies of Discriminator Dynamics

We study in the rest of this section the expression of the discriminators in the case of the IPM loss and LSGAN, as described in Section 5, and of the original GAN formulation.

A.6.1. PRELIMINARIES

We first need to introduce some definitions.

The presented solutions to Equation (9) leverage a notion of functions of linear operators, similarly to functions of matrices (Higham, 2008). We define such functions in the simplified case of non-negative symmetric compact operators with a finite number of eigenvalues, such as $\mathcal{T}_{k,\hat{\gamma}}$.

Definition 3 (Linear operator)**.** Let $\mathcal{A}\colon L^2(\hat{\gamma}) \to L^2(\Omega)$ be a non-negative symmetric compact linear operator with a finite number of eigenvalues, for which the spectral theorem guarantees the existence of a countable orthonormal basis of eigenfunctions with non-negative eigenvalues. If $\varphi\colon \mathbb{R}_+ \to \mathbb{R}$, we define $\varphi(\mathcal{A})$ as the linear operator with the same eigenspaces as $\mathcal{A}$, with their respective eigenvalues mapped by φ; in other words, if λ is an eigenvalue of $\mathcal{A}$, then $\varphi(\mathcal{A})$ admits the eigenvalue $\varphi(\lambda)$ with the same eigenspace.

In the case where $\mathcal{A}$ is a matrix, this amounts to diagonalizing $\mathcal{A}$ and transforming its diagonalization elementwise using φ. Note that $\mathcal{T}_{k,\hat{\gamma}}$ has a finite number of eigenvalues since it is generated by a finite linear combination of linear operators (see Definition 2).

We also need to define the following Radon–Nikodym derivatives with inputs in $\operatorname{supp}\hat{\gamma}$:

$$\rho = \frac{\mathrm{d}\left(\hat{\beta} - \hat{\alpha}\right)}{\mathrm{d}\left(\hat{\beta} + \hat{\alpha}\right)}, \qquad \rho_1 = \frac{\mathrm{d}\hat{\alpha}}{\mathrm{d}\hat{\gamma}}, \qquad \rho_2 = \frac{\mathrm{d}\hat{\beta}}{\mathrm{d}\hat{\gamma}}, \tag{103}$$

knowing that

$$\rho = \frac{1}{2}(\rho_2 - \rho_1), \qquad \rho_1 + \rho_2 = 2. \tag{104}$$

These functions help us to compute the functional gradient of $\mathcal{L}_{\hat{\alpha}}$, as follows.

Lemma 10 (Loss derivative)**.** *Under Assumption 3:*

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = \rho_1 \cdot \big(a' \circ f\big) - \rho_2 \cdot \big(b' \circ f\big). \tag{105}$$

Proof. We have from Equation (2):

$$\mathcal{L}_{\hat{\alpha}}(f) = \mathbb{E}_{x\sim\hat{\alpha}}\big[a_f(x)\big] - \mathbb{E}_{y\sim\hat{\beta}}\big[b_f(y)\big] = \big\langle \rho_1, a_f \big\rangle_{L^2(\hat{\gamma})} - \big\langle \rho_2, b_f \big\rangle_{L^2(\hat{\gamma})}, \tag{106}$$

hence by composition:

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 \cdot \big(a' \circ f\big) - \rho_2 \cdot \big(b' \circ f\big) = \rho_1 a'_f - \rho_2 b'_f. \tag{107}$$

□

A.6.2. LSGAN

Proposition 5 (LSGAN discriminator)**.** Under Assumptions 1 and 2, the solutions of Equation (9) for $a = -(\mathrm{id} + 1)^2$ and $b = -(\mathrm{id} - 1)^2$ are the functions defined for all $t \in \mathbb{R}_+$ as:

$$f_t = \exp\big(-4t\mathcal{T}_{k,\hat{\gamma}}\big)(f_0 - \rho) + \rho = f_0 + \varphi_t\big(\mathcal{T}_{k,\hat{\gamma}}\big)(f_0 - \rho), \tag{108}$$

where:

$$\varphi_t\colon x \mapsto \mathrm{e}^{-4tx} - 1. \tag{109}$$

Proof. Assumptions 1 and 2 are already assumed and Assumption 3 holds for the given a and b in LSGAN. Thus, Theorem 1 applies, and there exists a unique solution $t \mapsto f_t$ to Equation (9) over $\mathbb{R}_+$ in $L^2(\Omega)$ for a given initial condition f_0. Therefore, there remains to prove that, for a given initial condition f_0,

$$g\colon t \mapsto g_t = f_0 + \varphi_t\big(\mathcal{T}_{k,\hat{\gamma}}\big)(f_0 - \rho) \tag{110}$$

is a solution to Equation (9) with $g_0 = f_0$ and $g_t \in L^2(\Omega)$ for all $t \in \mathbb{R}_+$.

Let us first express the gradient of $\mathcal{L}_{\hat{\alpha}}$. We have from Lemma 10, with $a_f = -(f+1)^2$ and $b_f = -(f-1)^2$:

$$\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = -2\rho_1(f+1) - 2\rho_2(f-1) = 4\rho - 4f. \tag{111}$$

So Equation (9) equates to:

$$\partial_t f_t = 4\mathcal{T}_{k,\hat{\gamma}}(\rho - f_t). \tag{112}$$

Now let us prove that g_t is a solution to Equation (112). We have:

$$\partial_t g_t = -4\Big(\mathcal{T}_{k,\hat{\gamma}} \circ \exp\big(-4t\mathcal{T}_{k,\hat{\gamma}}\big)\Big)(f_0 - \rho) = -4\Big(\mathcal{T}_{k,\hat{\gamma}} \circ \exp\big(-4t\mathcal{T}_{k,\hat{\gamma}}\big)\Big)(f_0 - \rho). \tag{113}$$

Restricted to $\operatorname{supp}\hat{\gamma}$, we can write from Equation (110):

$$g_t = f_0 + \left(\exp\Big(-4t\,\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}\Big) - \mathrm{id}_{L^2(\hat{\gamma})}\right)(f_0 - \rho), \tag{114}$$

and plugging this in Equation (113):

$$\partial_t g_t = -4\mathcal{T}_{k,\hat{\gamma}}(g_t - \rho), \tag{115}$$

where we retrieve the differential equation of Equation (112). Therefore, g_t is a solution to Equation (112).

It is clear that $g_0 = f_0$. Moreover, $\mathcal{T}_{k,\hat{\gamma}}$ being decomposable in a finite orthonormal basis of elements of operators over $L^2(\Omega)$, its exponential has values in $L^2(\Omega)$ as well, making g_t belong to $L^2(\Omega)$ for all t. With this, the proof is complete. □

A.6.3. IPMs

Proposition 4 (IPM discriminator)**.** Under Assumptions 1 and 2, the solutions of Equation (9) for $a = b = \mathrm{id}$ are the functions of the form $f_t = f_0 + tf^*_{\hat{\alpha}}$, where $f^*_{\hat{\alpha}}$ is the unnormalized MMD witness function, yielding:

$$f^*_{\hat{\alpha}} = \mathbb{E}_{x\sim\hat{\alpha}}\big[k(x,\cdot)\big] - \mathbb{E}_{y\sim\hat{\beta}}\big[k(y,\cdot)\big], \qquad \mathcal{L}_{\hat{\alpha}}(f_t) = \mathcal{L}_{\hat{\alpha}}(f_0) + t\cdot \mathrm{MMD}_k^2\Big(\hat{\alpha},\hat{\beta}\Big). \tag{116}$$

Proof. Assumptions 1 and 2 are already assumed and Assumption 3 holds for the given a and b of the IPM loss. Thus, Theorem 1 applies, and there exists a unique solution $t \mapsto f_t$ to Equation (9) over $\mathbb{R}_+$ in $L^2(\Omega)$ for a given initial condition f_0. Therefore, in order to find the solution of Equation (9), there remains to prove that, for a given initial condition f_0,

$$g\colon t \mapsto g_t = f_0 + tf^*_{\hat{\alpha}} \tag{117}$$

is a solution to Equation (9) with $g_0 = f_0$ and $g_t \in L^2(\Omega)$ for all $t \in \mathbb{R}_+$.

Let us first express the gradient of $\mathcal{L}_{\hat{\alpha}}$. We have from Lemma 10, with $a_f = b_f = f$:

$$\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = -2\rho. \tag{118}$$

So Equation (9) equates to:

$$\partial_t f_t = -2\mathcal{T}_{k,\hat{\gamma}}(\rho) = 2\int_x k(\cdot,x)\rho(x)\,\mathrm{d}\hat{\gamma}(x) = \int_x k(\cdot,x)\,\mathrm{d}\hat{\alpha}(x) - \int_y k(\cdot,y)\,\mathrm{d}\hat{\beta}(y), \tag{119}$$

by definition of ρ (see Equation (103)), yielding:

$$\partial_t f_t = f^*_{\hat{\alpha}}. \tag{120}$$

Clearly, $t \mapsto g_t = f_0 + tf^*_{\hat{\alpha}}$ is a solution of the latter equation, $g_0 = f_0$ and $g_t \in L^2(\Omega)$ given that $\operatorname{supp}\hat{\gamma}$ is finite and $k \in L^2\big(\Omega^2\big)$ by assumption. The set of solutions for the IPM loss is thus characterized.

Finally, let us compute $\mathcal{L}_{\hat{\alpha}}(f_t)$. By linearity of $\mathcal{L}_{\hat{\alpha}}$ for $a = b = \mathrm{id}$:

$$\mathcal{L}_{\hat{\alpha}}(f_t) = \mathcal{L}_{\hat{\alpha}}(f_0) + t\cdot\mathcal{L}_{\hat{\alpha}}(f^*_{\hat{\alpha}}) = \mathcal{L}_{\hat{\alpha}}(f_0) + t\cdot\mathcal{L}_{\hat{\alpha}}\big(\mathcal{T}_{k,\hat{\gamma}}(-2\rho)\big). \tag{121}$$

But, from Equation (106), $\mathcal{L}_{\hat{\alpha}}(f) = \langle -2\rho, f\rangle_{L^2(\hat{\gamma})}$, hence:

$$\mathcal{L}_{\hat{\alpha}}(f_t) = \mathcal{L}_{\hat{\alpha}}(f_0) + t\cdot\big\langle -2\rho, \mathcal{T}_{k,\hat{\gamma}}(-2\rho)\big\rangle_{L^2(\hat{\gamma})} = \mathcal{L}_{\hat{\alpha}}(f_0) + t\cdot\big\|\mathcal{T}_{k,\hat{\gamma}}(-2\rho)\big\|^2_{\mathcal{H}^{\hat{\gamma}}_k}. \tag{122}$$

By noticing that $\mathcal{T}_{k,\hat{\gamma}}(-2\rho) = f^*_{\hat{\alpha}}$ and that $\big\|f^*_{\hat{\alpha}}\big\|_{\mathcal{H}^{\hat{\gamma}}_k} = \mathrm{MMD}_k\Big(\hat{\alpha},\hat{\beta}\Big)$ since $f^*_{\hat{\alpha}}$ is the unnormalized MMD witness function, the expression of $\mathcal{L}_{\hat{\alpha}}(f_t)$ in the proposition is obtained. □

A.6.4. VANILLA GAN

Unfortunately, finding the solutions to Equation (9) in the case of the original GAN formulation, i.e. $a = \log(1-\sigma)$ and $b = -\log\sigma$, remains to the extent of our knowledge an open problem. We provide in the rest of this section some leads that might prove useful for more advanced analyses.

Let us first determine the expression of Equation (9) for vanilla GAN.

Lemma 11. *For $a = \log(1-\sigma)$ and $b = -\log\sigma$, Equation (9) equates to:*

$$\partial_t f_t = \mathcal{T}_{k,\hat{\gamma}}\big(\rho_2 - 2\sigma(f)\big). \tag{123}$$

Proof. We have from Lemma 10, with $a_f = b_f = f$:

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = -\rho_1 \frac{\sigma'(f)}{1-\sigma(f)} + \rho_2 \frac{\sigma'(f)}{\sigma(f)}. \tag{124}$$

By noticing that $\sigma'(f) = \sigma(f)\big(1-\sigma(f)\big)$, we obtain:

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = -\rho_1\sigma(f) + \rho_2\big(1-\sigma(f)\big) = \rho_2 - 2\sigma(f). \tag{125}$$

By plugging the latter expression in Equation (9), the desired result is achieved. □

Note that Assumption 3 holds for these choices of a and b. Therefore, under Assumptions 1 and 2, there exists a unique solution to Equation (123) in $\mathbb{R}_+ \to L^2(\Omega)$ with a given initialization f_0.

Let us first study Equation (123) in the simplified case of a one-dimensional ordinary differential equation.

Proposition 8. *Let $r \in \{0, 2\}$ and $\lambda \in \mathbb{R}$. The set of differentiable solutions over $\mathbb{R}$ to this ordinary differential equation:*

$$\partial_t y_t = \lambda\big(r - 2\sigma(y_t)\big) \tag{126}$$

is the following set:

$$S = \left\{ y\colon t \mapsto (1-r)\left(W\left(\mathrm{e}^{2\lambda t + C}\right) - 2\lambda t - C\right) \;\middle|\; C \in \mathbb{R} \right\}, \tag{127}$$

where W the is principal branch of the Lambert W function (Corless et al., 1996).

Proof. The theorem of Cauchy-Lipschitz ensures that there exists a unique global solution to Equation (126) for a given initial condition $y_0 \in \mathbb{R}$. Therefore, we only need to show that all elements of S are solutions of Equation (126) and that they can cover any initial condition.

Let us first prove that $y\colon t \mapsto (1-r)\Big(W\big(\mathrm{e}^{2\lambda t + C}\big) - 2\lambda t - C\Big)$ is a solution of Equation (126). Let us express the derivative of y:

$$\frac{1}{1-r}\partial_t y_t = 2\lambda\left(\mathrm{e}^{2\lambda t + C} W'\left(\mathrm{e}^{2\lambda t + C}\right) - 1\right). \tag{128}$$

$W'(z) = \frac{W(z)}{z\left(1+W(z)\right)}$, so:

$$\frac{1}{1-r}\partial_t y_t = 2\lambda\left(\frac{W\left(\mathrm{e}^{2\lambda t + C}\right)}{1 + W\left(\mathrm{e}^{2\lambda t + C}\right)} - 1\right) = -\frac{2\lambda}{1 + W\left(\mathrm{e}^{2\lambda t + C}\right)}. \tag{129}$$

Moreover, $W(z) = z\mathrm{e}^{-W(z)}$, and with $r - 1 \in \{1, -1\}$:

$$\frac{1}{1-r}\partial_t y_t = -\frac{2\lambda}{1 + \mathrm{e}^{2\lambda t + C}\mathrm{e}^{-W\left(\mathrm{e}^{2\lambda t + C}\right)}} = -\frac{2\lambda}{1 + \mathrm{e}^{(r-1)y_t}}. \tag{130}$$

Finally, we notice that, since $r \in \{0, 2\}$:

$$\lambda\big(r - 2\sigma(y_t)\big) = -\frac{2\lambda(1-r)}{1 + \mathrm{e}^{(r-1)y_t}}. \tag{131}$$

Therefore:

$$\partial_t y_t = \lambda\big(r - 2\sigma(y_t)\big) \tag{132}$$

and y_t is a solution to Equation (126).

Since $y_0 = (1-r)\Big(W\big(\mathrm{e}^C\big) - C\Big)$ and $z \mapsto W(\mathrm{e}^z) - z$ can be proven to be bijective over $\mathbb{R}$, the elements of S can cover any initial condition. With this, the result is proved. □

Suppose that $f_0 = 0$ in Equation (123) and that ρ_2 has values in $\{0, 2\}$ – i.e. $\hat{\alpha}$ and $\hat{\beta}$ have disjoint supports (which is the typical case for distributions with finite support). From Proposition 8, a candidate solution would be:

$$f_t = \varphi_t(x)(\rho_2 - 1) = -\varphi_t(x)(\rho), \tag{133}$$

where:

$$\varphi_t \colon x \mapsto W\Big(\mathrm{e}^{2tx+1}\Big) - 2tx - 1, \tag{134}$$

since the initial condition $y_0 = 0$ gives the constant value $C = 1$ in Equation (127). Note that the Lambert W function of a symmetric linear operator is well-defined, all the more so as we choose the principal branch of the Lambert function in our case; see the work of Corless et al. (2007) for more details. Note also that the estimation of $W(\mathrm{e}^z)$ is actually numerically stable using approximations from Iacono & Boyd (2017).

However, Equation (133) cannot be a solution of Equation (123). Indeed, one can prove by following essentially the same reasoning as the proof of Proposition 8 that:

$$\partial_t f_t = 2\Big(\mathcal{T}_{k,\hat{\gamma}} \circ \Big(\psi_t\big(\mathcal{T}_{k,\hat{\gamma}}\big)\Big)^{-1}\Big)(\rho_2 - 1), \tag{135}$$

with:

$$\psi_t \colon x \mapsto 1 + W\Big(\mathrm{e}^{2tx+1}\Big) > 0. \tag{136}$$

However, this does not allow us to obtain Equation (123) since in the latter the sigmoid is taken coordinate-wise, where the exponential in Equation (135) acts on matrices.

Nonetheless, for t small enough, f_t as defined in Equation (135) should approximate the solution of Equation (123), since sigmoid is approximately linear around 0 and $f_t \approx 0$ when t is small enough. We find in practice that for reasonable values of t, e.g. $t \leq 5$, the approximate solution of Equation (135) is actually close to the numerical solution of Equation (123) obtained using an ODE solver. Thus, we provide here a candidate approximate expression for the discriminator in the setting of the original GAN formulation – i.e., for binary classifiers. We leave for future work a more in-depth study of this case.

B. Discussions and Remarks

We develop in this section some remarks and explanations on the topics that are broached in the main paper.

B.1. From Finite to Infinite-Width Networks

The constancy of the neural tangent kernel during training when the width of the network becomes increasingly large is broadly applicable. As summarized by Liu et al. (2020), typical neural networks with the building blocks of multilayer perceptrons and convolutional neural networks comply with this property, as long as they end with a linear layer and they do not have any bottleneck – indeed, this constancy needs the minimum internal width to grow unbounded (Arora et al., 2019). This includes, for example, residual convolutional neural networks (He et al., 2016). The requirement of a final linear activation can be circumvented by transferring this activation into the loss function, as we did for the original GAN formulation in Section 3. This makes our framework encompass a wide range of discriminator architectures.

Indeed, many building blocks of state-of-the-art discriminators can be studied in this infinite-width regime with a constant NTK, as highlighted by the exhaustiveness of the Neural Tangents library (Novak et al., 2020). Assumptions about the used activation functions are mild and include many standard activations such as ReLU, sigmoid and $\tanh$. Beyond fully connected linear layers and convolutions, NTK constancy also affect typical operations such as self-attention (Hron et al., 2020), layer normalization and batch normalization (Yang, 2020). This variety of networks affected by the constancy of the NTK supports the generality of our approach, as it includes powerful discriminator architectures such as BigGAN (Brock et al., 2019).

We highlight that the NTK of the discriminator remains constant throughout the whole GAN optimization process, and not only under a fixed generator. Indeed, if it remains constant in-between generator updates, then it also remains constant when the generator changes. This is because, for a finite training time, the constancy of the NTK solely depends on the network architecture and initialization, regardless of the training loss which may change in the course of training without affecting the NTK.

There are nevertheless some limits to the NTK approximation, as we are not aware of works studying the application of the infinite-width regime to some operations such as spectral normalization, and networks in the regime of a constant NTK cannot perform feature learning as they are equivalent to kernel methods (Geiger et al., 2020; Yang & Hu, 2021). However, this framework remains general and constitutes the most advanced attempt at theoretically modeling the discriminator's architecture in GANs.

B.2. Loss of the Generator and its Gradient

We highlight in this section the importance of taking into account alternating optimization and discriminator gradients in the optimization of the generator. Let us focus on an example similar to the one of Arjovsky et al. (2017, Example 1) and choose as β a single Dirac centered at 0 and as $\alpha_g = \alpha_\theta$ single Dirac centered at $x_\theta = \theta$ (the generator parameters being the coordinates of the generated point). Let us study for the sake of simplicity the case of LSGAN since it is a recurring example in this work, but a similar reasoning can be done for other GAN instances.

In the theoretical min-max formulation of GANs considered by Arjovsky et al. (2017), the generator is trained to minimize the following quantity:

$$\mathcal{C}_{f^\star_{\alpha_\theta}}(\alpha_\theta) \triangleq \mathbb{E}_{x\sim\alpha_\theta}\left[c_{f^\star_{\alpha_\theta}}(x)\right] = f^\star_{\alpha_\theta}(x_\theta)^2, \tag{137}$$

where:

$$\begin{aligned} f^\star_{\alpha_\theta} &= \operatorname*{arg\,max}_{f\in L^2\left(\frac{1}{2}\alpha_\theta+\frac{1}{2}\beta\right)} \left\{\mathcal{L}_{\alpha_\theta}(f) \triangleq \mathbb{E}_{x\sim\alpha_\theta}\left[a_f(x)\right] - \mathbb{E}_{y\sim\beta}\left[b_f(y)\right]\right\} \\ &= \operatorname*{arg\,min}_{f\in L^2\left(\frac{1}{2}\alpha_\theta+\frac{1}{2}\beta\right)} \left\{\left(f^\star_{\alpha_\theta}(x_\theta)+1\right)^2 + \left(f^\star_{\alpha_\theta}(0)-1\right)^2\right\}. \end{aligned} \tag{138}$$

Consequently, $f^\star_{\alpha_\theta}(0) = 1$ and $f^\star_{\alpha_\theta}(x_\theta) = -1$ when $x_\theta \neq 0$, thus in this case:

$$\mathcal{C}_{f^\star_{\alpha_\theta}}(\alpha_\theta) = 1. \tag{139}$$

This constancy of the generator loss would make it impossible to be learned by gradient descent, as pointed out by Arjovsky et al. (2017).

However, the setting does not correspond to the actual optimization process used in practice and represented by Equation (3). We do have $\nabla_\theta \mathcal{C}_{f^\star_{\alpha_\theta}}(\alpha_\theta) = 0$ when $x_\theta \neq 0$, but the generator never uses this gradient in standard GAN optimization. Indeed, this gradient takes into account the dependency of the optimal discriminator $f^\star_{\alpha_\theta}$ in the generator parameters, since the optimal discriminator depends on the generated distribution. Yet, in practice and with few exceptions such as Unrolled GANs (Metz et al., 2017) and as done in Equation (3), this dependency is ignored when computing the gradient of the generator, because of the alternating optimization setting – where the discriminator is trained in-between generator's updates. Therefore, despite being constant on the training data, this loss can yield non-zero gradients to the generator. However, this requires the gradient of $f^\star_{\alpha_\theta}$ to be defined, which is the issue addressed in Section 3.2.

B.3. Differentiability of the Bias-Free ReLU Kernel

Remark 1 contradicts the results of Bietti & Mairal (2019) on the regularity of the NTK of a bias-free ReLU MLP with one hidden layer, which can be expressed as follows (up to a constant scaling the matrix multiplication in linear layers):

$$k(x,y) = \|x\|\|y\|\kappa\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right), \tag{140}$$

where:

$$\begin{aligned}\kappa\colon [0,1] &\to \mathbb{R}\\ u &\mapsto \frac{2}{\pi}u(\pi - \arccos u) + \frac{1}{\pi}\sqrt{1-u^2}\,.\end{aligned} \tag{141}$$

More particularly, Bietti & Mairal (2019, Proposition 3) claim that $k(\cdot, y)$ is not Lipschitz around y for all y in the unit sphere. By following their proof, it amounts to prove that $k(\cdot, y)$ is not Lipschitz around y for all y in any centered sphere. We highlight that this also contradicts empirical evidence, as we did observe the Lipschitzness of such NTK in practice using the Neural Tangents library (Novak et al., 2020).

We believe that the mistake in the proof of Bietti & Mairal (2019) lies in the confusion between functions κ and $k_0\colon x, y \mapsto \kappa\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right)$, which have different geometries. Their proof relies on the fact that κ is indeed non-Lipschitz in the neighborhood of $u = 1$. However, this does not imply that k_0 is not Lipschitz, or not derivable. We can prove that it is actually at least locally Lipschitz.

Indeed, let us compute the following derivative for $x \neq y \in \mathbb{R}^n \setminus \{0\}$:

$$\frac{\partial k_0(x,y)}{\partial x} = \frac{y\|x\| - \frac{x}{\|x\|}\langle x,y\rangle}{\|x\|^2\|y\|}\kappa'(u) = \frac{1}{\|x\|\|y\|}\left(y - \langle x,y\rangle\frac{x}{\|x\|^2}\right)\kappa'(u), \tag{142}$$

where $u = \frac{\langle x,y\rangle}{\|x\|\|y\|}$ and:

$$\pi\cdot\kappa'(u) = \frac{u}{\sqrt{1-u^2}} + 2(\pi - \arccos u). \tag{143}$$

Note that $\kappa'(u) \sim_{u\to 1^-} \frac{\pi u}{\sqrt{1-u^2}} \sim_{u\to 1^-} \frac{\pi}{\sqrt{2}\sqrt{1-u}}$. Therefore:

$$\begin{aligned}\frac{\pi}{\sqrt{2}}\cdot\frac{\partial k_0(x,y)}{\partial x} &\sim_{x\to y} \frac{1}{\|y\|^2}\left(y - \langle x,y\rangle\frac{x}{\|x\|^2}\right)\frac{\sqrt{\|x\|\|y\|}}{\sqrt{\|x\|\|y\| - \langle x,y\rangle}}\\ &\sim_{x\to y} \frac{\|x\|^2 y - \langle x,y\rangle x}{\|y\|^3\sqrt{\|x\|\|y\| - \langle x,y\rangle}}\\ &\sim_{x\to y} \frac{\|y\|^2 - \langle x,y\rangle}{\|y\|^3\sqrt{\|y\|^2 - \langle x,y\rangle}}y \xrightarrow[x\to y]{} 0,\end{aligned} \tag{144}$$

which proves that k_0 is actually Lipschitz around points (y, y), as well as differentiable, and confirms our remark.

B.4. Integral Operator and Instance Noise

Instance noise (Sønderby et al., 2017) consists in adding random Gaussian noise to the input and target samples. This amounts to convolving the data distributions with a Gaussian density, which will have the effect of smoothing the discriminator. In the following, for the case of IPM losses, we link instance noise with our framework, showing that smoothing of the data distributions already occurs via the NTK kernel, stemming from the fact that the discriminator is a neural network trained with gradient descent.

More specifically, it can be shown that if k is an RBF kernel, the optimal discriminators in both case are the same. This is based on the fact that the density of a convolution of an empirical measure $\hat{\mu} = \frac{1}{N}\sum_i \delta_{x_i}$, where δ_z is the Dirac distribution centered on z, and a Gaussian density $\tilde{k}$ with associated RBF kernel k can be written as $\tilde{k} * \hat{\mu} = \frac{1}{N}\sum_i k(x_i, \cdot)$.

Let us consider the following regularized discriminator optimization problem in $L^2(\mathbb{R})$ smoothed from $L^2(\Omega)$ with instance noise, i.e. convolving $\hat{\alpha}$ and $\hat{\beta}$ with $\tilde{k}$.

$$\sup_{f \in L^2(\mathbb{R})} \left\{ \mathcal{L}^{\tilde{k}}_{\hat{\alpha}}(f) \triangleq \mathbb{E}_{x \sim \tilde{k} * \hat{\alpha}}\big[f(x)\big] - \mathbb{E}_{y \sim \tilde{k} * \hat{\beta}}\big[f(y)\big] - \lambda \|f\|^2_{L^2} \right\} \tag{145}$$

The optimum f^{IN} can be found by taking the gradient:

$$\nabla_f \left(\mathcal{L}^{\tilde{k}}_{\hat{\alpha}}\left(f^{\text{IN}}\right) - \lambda \left\| f^{\text{IN}} \right\|^2_{L^2} \right) = 0 \qquad \Leftrightarrow \qquad f^{\text{IN}} = \frac{1}{2\lambda}\Big(\tilde{k} * \hat{\alpha} - \tilde{k} * \hat{\beta} \Big). \tag{146}$$

If we now study the resolution of the optimization problem in $\mathcal{H}^{\hat{\gamma}}_k$ as in Section 5.1 with $f_0 = 0$, we find the following discriminator:

$$f_t = t\Big(\mathbb{E}_{x \sim \hat{\alpha}}\big[k(x, \cdot)\big] - \mathbb{E}_{y \sim \hat{\beta}}\big[k(y, \cdot)\big] \Big) = t\Big(\tilde{k} * \hat{\alpha} - \tilde{k} * \hat{\beta} \Big). \tag{147}$$

Therefore, we have that $f^{\text{IN}} \propto f_t$, i.e. instance noise and regularization by neural networks obtain the same smoothed solution.

This analysis was done using the example of an RBF kernel, but it also holds for stationary kernels, i.e. $k(x, y) = \tilde{k}(x - y)$, which can be used to convolve measures. We remind that this is relevant, given that NTKs are stationary over spheres (Jacot et al., 2018; Yang & Salman, 2019), around where data can be concentrated in high dimensions.

B.5. Positive Definite NTKs

Optimality results in the theory of NTKs usually rely on the assumption that the considered NTK k is positive definite over the training dataset $\hat{\gamma}$ (Jacot et al., 2018; Zhang et al., 2020). This property offers several theoretical advantages.

Indeed, this gives sufficient representational power to its RKHS to include the optimal solution over $\hat{\gamma}$. Moreover, this positive definiteness property equates for finite datasets to the invertibility of the mapping

$$\begin{aligned} \mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}} \colon L^2(\hat{\gamma}) &\to L^2(\hat{\gamma}) \\ h &\mapsto \mathcal{T}_{k,\hat{\gamma}}(h)\big|_{\operatorname{supp}\hat{\gamma}} \end{aligned}, \tag{148}$$

that can be seen as a multiplication by the invertible Gram matrix of k over $\hat{\gamma}$. From this, one can retrieve the expression of $f \in \mathcal{H}^{\hat{\gamma}}_k$ from its restriction $f|_{\operatorname{supp}\hat{\gamma}}$ to $\operatorname{supp}\hat{\gamma}$ in the following way:

$$f = \mathcal{T}_{k,\hat{\gamma}} \circ \mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}} \Big(f|_{\operatorname{supp}\hat{\gamma}} \Big), \tag{149}$$

as shown in Lemma 9. Finally, as shown by Jacot et al. (2018) and in Appendix A.5, this makes the discriminator loss function strictly increase during training.

One may wonder whether this assumption is reasonable for NTKs. Jacot et al. (2018) proved that it indeed holds for NTKs of non-shallow MLPs with non-polynomial activations if data is supported on the unit sphere, supported by the fact that the NTK is stationary over the unit sphere. Others, such as Fan & Wang (2020), have observed positive definiteness of the NTK subject to specific assumptions on the networks and data. We are not aware of more general results of this kind. However, one may conjecture that, at least for specific kinds of networks, NTKs are positive definite for any training data.

Indeed, besides global convergence results (Allen-Zhu et al., 2019), prior work indicates that MLPs are universal approximators (Hornik et al., 1989; Leshno et al., 1993). This property can be linked in our context to universal kernels (Steinwart, 2001), which are guaranteed to be positive definite over any training data (Sriperumbudur et al., 2011). Universality is linked to the density of the kernel RKHS in the space of continuous functions. In the case of NTKs, previously cited approximation properties can be interpreted as signs of expressive RKHSs, and thus support the hypothesis of universal NTKs. Furthermore, beyond positive definiteness, universal kernels are also characteristic (Sriperumbudur et al., 2011), which is interesting when they are used to compute MMDs, as we do in Section 5.1. Note that for the standard case of ReLU MLPs, Ji et al. (2020) showed universal approximation results in the infinite-width regime, and works such as the one of Chen & Xu (2021) observed that their RKHS is close to the one of the Laplace kernel, which is positive definite.

Bias-free ReLU NTKs are not characteristic. As already noted by Leshno et al. (1993), the presence of bias is important when it comes to representational power of MLPs. We can retrieve this observation in our framework. In the case of a ReLU shallow network with one hidden layer and without bias, Bietti & Mairal (2019) determine its associated NTK as follows (up to a constant scaling the matrix multiplication in linear layers):

$$k(x,y) = \|x\|\|y\|\kappa\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right), \tag{150}$$

with in particular $k(x,0) = 0$ for all $x \in \Omega$; suppose that $0 \in \Omega$. This expression of the kernel implies that k is not positive definite for all datasets: take for example $x = 0$ and $y \in \Omega \setminus \{0\}$; then the Gram matrix of k has a null row, hence k is not strictly positive definite over $\{x, y\}$. Another consequence is that k is not characteristic. Indeed, take probability distributions $\mu = \delta_{\frac{y}{2}}$ and $\nu = \frac{1}{2}\big(\delta_x + \delta_y\big)$ with δ_z being the Dirac distribution centered on $z \in \Omega$, and where $x = 0$ and $y \in \Omega \setminus \{0\}$. Then:

$$\mathbb{E}_{z\sim\mu}k(z,\cdot) = k\left(\frac{1}{2}y,\cdot\right) = \frac{1}{2}k(y,\cdot) = \frac{1}{2}\big(k(y,\cdot) + k(x,\cdot)\big) = \mathbb{E}_{z\sim\nu}k(z,\cdot), \tag{151}$$

i.e., kernel embeddings of μ and $\nu \neq \mu$ are identical, making k not characteristic by definition.

B.6. Societal Impact

As our work is mainly theoretical and does not deal with real-world data, it does not have direct broader negative impact on the society. However, the practical perspectives that it opens constitute an object of interrogation. Indeed, the developments of performant generative models can be the source of harmful manipulation (Tolosana et al., 2020) and reproduction of existing biases in databases (Jain et al., 2020), especially as GANs are still misunderstood. While such negative effects should be considered, attempts such as ours at explaining generative models might also lead to ways to mitigate potential harms by paving the way for more principled GAN models.

C. GAN(TK)^2 and Further Empirical Analyses

We present in this section additional experimental results that complement and explain some of the results already exposed in Section 6. All these experiments were conducted using the proposed general toolkit GAN(TK)^2.

We focus in this article on particular experiments for the sake of clarity and as an illustration of the potential of analysis of our framework, but GAN(TK)^2 is a general-purpose toolkit centered around the infinite-width of the discriminator and could be leveraged for an even more extensive empirical analysis. We specifically focus on the IPM and LSGAN losses for the discriminator since they are the two losses for which we know the analytic behavior of the discriminator in the infinite-width limit, but other losses can be studied as well in GAN(TK)^2. We leave a large-scale empirical study of our framework, which is out of the scope of this paper, for future work.

C.1. Two-Dimensional Datasets

We provide in Table 1 numerical results corresponding to the experiments described in Section 6 on the 8 Gaussians dataset.

We present additional experimental results on two other two-dimensional problems, Density and AB; see, respectively, Figures 3 and 4. Numerical results are detailed in Tables 2 and 3. We globally retrieve the same conclusions that we developed in Section 6 on these datasets with more complex shapes.

C.2. ReLU vs. Sigmoid Activations

We additionally introduce a new baseline for the 8 Gaussians, Density and AB problems, where we replace the ReLU activation in the discriminator by a sigmoid-like activation $\tilde{\sigma}$, that we abbreviate to sigmoid in this experimental study for readability purposes. We choose $\tilde{\sigma}$ instead of the actual sigmoid σ for computational reasons, since $\tilde{\sigma}$, contrary to σ, allows for analytic computations of NTKs in the Neural Tangents library (Novak et al., 2020). $\tilde{\sigma}$ is defined in the latter using the error function erf scaled in order to minimize a squared loss with respect to σ over $[-5, 5]$, with the following expression:

$$\tilde{\sigma}\colon x \mapsto \frac{1}{2}\left(\mathrm{erf}\left(\frac{x}{2.402\,056\,353\,171\,979\,6}\right) + 1\right). \tag{152}$$

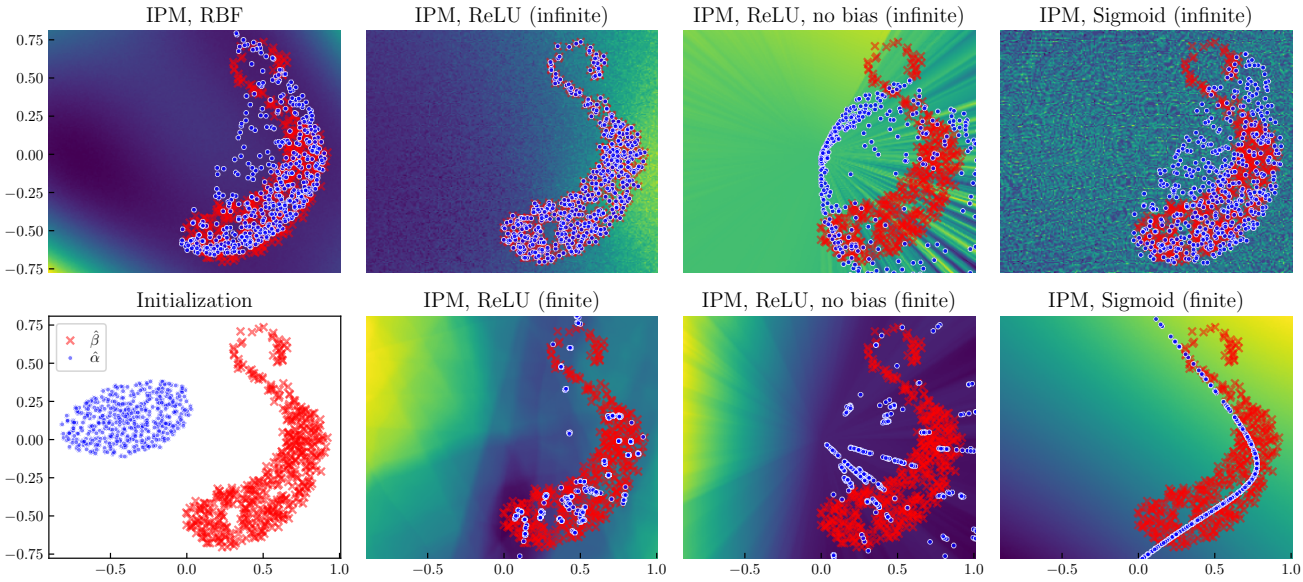


Figure 3. Generator (●) and target (×) samples for different methods applied to the Density problem. In the background, $c_{f^\star}$.

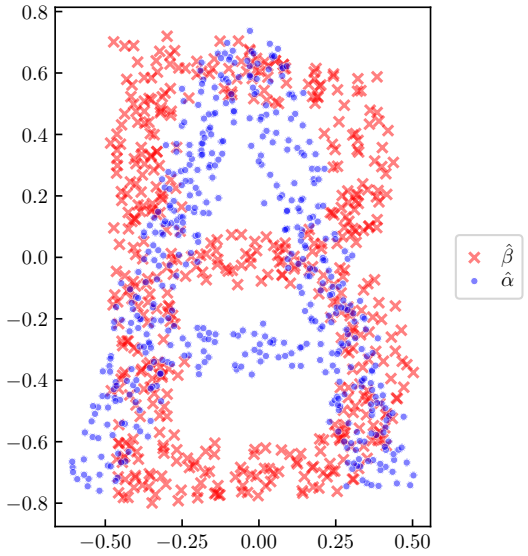


Figure 4. Initial generator (●) and target (×) samples for the AB problem.

Table 1. Sinkhorn divergence (Feydy et al., 2019, lower is better, similar to $\mathcal{W}_2$) averaged over three runs between the final generated distribution and the target dataset for the 8 Gaussians problem.

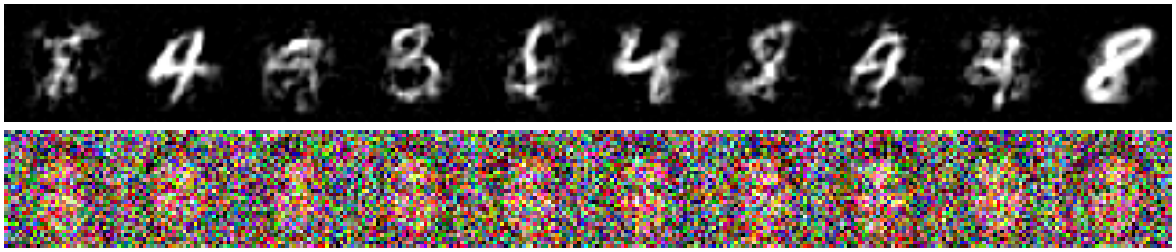
Loss	RBF kernel	ReLU	ReLU (no bias)	Sigmoid
IPM (inf.)	$(2.60 \pm 0.06) \cdot 10^{-2}$	$(9.40 \pm 2.71) \cdot 10^{-7}$	$(9.70 \pm 1.88) \cdot 10^{-2}$	$(8.40 \pm 0.02) \cdot 10^{-2}$
IPM	—	$(1.21 \pm 0.14) \cdot 10^{-1}$	$(1.20 \pm 0.60) \cdot 10^{0}$	$(7.40 \pm 1.30) \cdot 10^{-1}$
LSGAN (inf.)	$(4.21 \pm 0.10) \cdot 10^{-1}$	$(7.56 \pm 0.45) \cdot 10^{-2}$	$(1.27 \pm 0.01) \cdot 10^{1}$	$(7.35 \pm 0.11) \cdot 10^{0}$
LSGAN	—	$(3.07 \pm 0.68) \cdot 10^{0}$	$(7.52 \pm 0.01) \cdot 10^{0}$	$(7.41 \pm 0.54) \cdot 10^{0}$

Table 2. Sinkhorn divergence averaged over three runs between the final generated distribution and the target dataset for the Density problem.

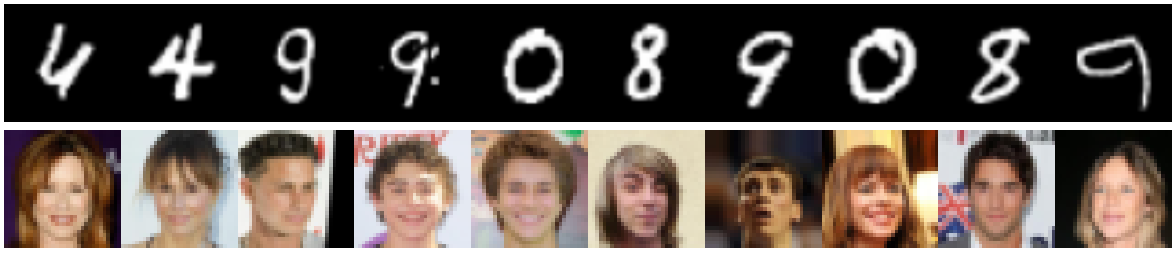
Loss	RBF kernel	ReLU	ReLU (no bias)	Sigmoid
IPM (inf.)	$(2.37 \pm 0.32) \cdot 10^{-3}$	$(3.34 \pm 0.49) \cdot 10^{-9}$	$(7.34 \pm 0.34) \cdot 10^{-2}$	$(6.25 \pm 0.31) \cdot 10^{-3}$
IPM	—	$(5.02 \pm 1.19) \cdot 10^{-3}$	$(9.25 \pm 0.30) \cdot 10^{-2}$	$(3.06 \pm 0.57) \cdot 10^{-2}$
LSGAN (inf.)	$(7.53 \pm 0.59) \cdot 10^{-3}$	$(1.49 \pm 0.11) \cdot 10^{-3}$	$(2.80 \pm 0.03) \cdot 10^{-1}$	$(2.21 \pm 0.01) \cdot 10^{-1}$
LSGAN	—	$(1.53 \pm 1.08) \cdot 10^{-2}$	$(1.64 \pm 0.19) \cdot 10^{-1}$	$(5.88 \pm 0.80) \cdot 10^{-2}$

Table 3. Sinkhorn divergence averaged over three runs between the final generated distribution and the target dataset for the AB problem.

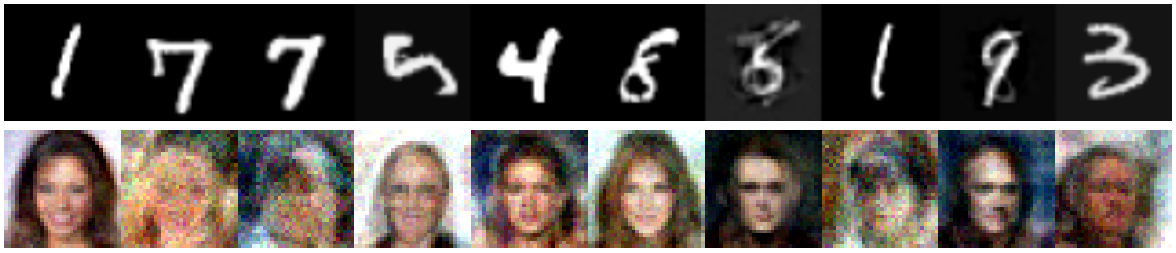
Loss	RBF kernel	ReLU	ReLU (no bias)	Sigmoid
IPM (inf.)	$(4.65 \pm 0.82) \cdot 10^{-3}$	$(2.64 \pm 2.13) \cdot 10^{-9}$	$(6.11 \pm 0.19) \cdot 10^{-3}$	$(5.69 \pm 0.38) \cdot 10^{-3}$
IPM	—	$(2.75 \pm 0.20) \cdot 10^{-3}$	$(3.65 \pm 1.44) \cdot 10^{-2}$	$(1.25 \pm 0.32) \cdot 10^{-2}$
LSGAN (inf.)	$(1.13 \pm 0.05) \cdot 10^{-2}$	$(8.63 \pm 2.24) \cdot 10^{-3}$	$(1.02 \pm 0.40) \cdot 10^{-1}$	$(1.40 \pm 0.06) \cdot 10^{-2}$
LSGAN	—	$(1.32 \pm 1.30) \cdot 10^{-1}$	$(2.57 \pm 0.73) \cdot 10^{-2}$	$(8.78 \pm 2.23) \cdot 10^{-2}$



(a) RBF kernel: blurry digits on MNIST, prohibitively noisy images on CelebA.



(b) ReLU: sharp digits on MNIST, high-quality images on CelebA.



(c) ReLU (no bias): mostly sharp digits with some artifacts and blurry images on MNIST, blurry and noisy images on CelebA.

Figure 5. Uncurated samples from the results of the descent of a set of 1024 particles over a subset of 1024 elements of MNIST and CelebA, starting from a standard Gaussian. Training is done using the IPM loss in the infinite-width kernel setting.

Results are given in Tables 1 to 3 and an illustration is available in Figure 3. We observe that the sigmoid baseline is consistently outperformed by the RBF kernel and ReLU activation (with bias) for all regimes and losses. This is in accordance with common experimental practice, where internal sigmoid activations are found less effective than ReLU because of the potential activation saturation that they can induce.

We provide a qualitative explanation to this underperformance of sigmoid via our framework in Appendix C.4.

C.3. Qualitative MNIST and CelebA Experiment

An experimental analysis of our framework on complex image datasets is out the scope of our study – we leave it for future work. Nonetheless, we present an experiment on MNIST (LeCun et al., 1998) and CelebA (Liu et al., 2015) images in a similar setting as the experiments on two-dimensional point clouds of the previous sections. For each dataset, we make a point cloud $\hat{\alpha}$, initialized to a standard Gaussian, move towards a subset of the MNIST dataset following the gradients of the IPM loss in the infinite-width regime. Qualitative results are presented in Figure 5.

We notice, similarly to the two-dimensional experiments, that the ReLU network with bias outperforms its bias-free counterpart and a standard RBF kernel in terms of sample quality. The difference between the RBF kernel and ReLU NTK is even more flagrant in this complex high-dimensional setting, as the RBF kernel is unable to produce accurate samples.

C.4. Visualizing the Gradient Field Induced by the Discriminator

We raise in Sections 4.4 and 5 the open problem of studying the convergence of the generated distribution towards the target distribution with respect to the gradients of the discriminator. We aim in this section at qualitatively studying these gradients in a simplified case that could shed some light on the more general setting and explain some of our experimental results. These gradient fields can be plotted using the provided GAN(TK)^2 toolkit.

C.4.1. Setting

Since we study gradients of the discriminator expressed in Equation (10), we assume that $f_0 = 0$ – for instance, using the anti-symmetrical initialization Zhang et al. (2020) – in order to ignore residual gradients from the initialization.

By Theorem 1, for any loss and any training time, the discriminator can be expressed as $f^\star_{\hat{\alpha}} = \mathcal{T}_{k,\hat{\gamma}}(h_0)$, for some $h_0 \in L^2(\hat{\gamma})$. Thus, there exists $h_1 \in L^2(\hat{\gamma})$ such that:

$$f^\star_{\hat{\alpha}} = \sum_{x \in \operatorname{supp} \hat{\gamma}} h_1(x) k(x, \cdot). \tag{153}$$

Consequently,

$$\nabla f^\star_{\hat{\alpha}} = \sum_{x \in \operatorname{supp} \hat{\gamma}} h_1(x) \nabla k(x, \cdot), \qquad \nabla c_{f^\star_{\hat{\alpha}}} = \sum_{x \in \operatorname{supp} \hat{\gamma}} h_1(x) \nabla k(x, \cdot) c'\big(f^\star_{\hat{\alpha}}(\cdot)\big). \tag{154}$$

Dirac-GAN setting. The latter linear combination of gradients indicates that, by examining gradients of $c_{f^\star_{\hat{\alpha}}}$ for pairs of $(x, y) \in \operatorname{supp} \hat{\alpha} \times \operatorname{supp} \hat{\beta}$, one could already develop potentially valid intuitions that can hold even when multiple points are considered. This is especially the case for the IPM loss, as h_0, h_1 have a simple form: $h_1(x) = 1$ if $x \in \operatorname{supp} \hat{\alpha}$ and $h_1(y) = -1$ if $y \in \operatorname{supp} \hat{\alpha}$ (assuming points from $\hat{\alpha}$ and $\hat{\beta}$ are uniformly weighted); moreover, note that $c'\big(f^\star_{\hat{\alpha}}(\cdot)\big) = 1$. Thus, we study here $\nabla c_{f^\star_{\hat{\alpha}}}$ when $\hat{\alpha}$ and $\hat{\beta}$ are only comprised of one point, i.e. the setting of Dirac GAN (Mescheder et al., 2018), with $\hat{\alpha} = \delta_x \triangleq \hat{\alpha}_x$ and $\hat{\beta} = \delta_y$.

Visualizing high-dimensional inputs. Unfortunately, the gradient field is difficult to visualize when the samples live in a high-dimensional space. Interestingly, the NTK $k(x, y)$ for any architecture starting with a fully connected layer only depends on $\|x\|$, $\|y\|$ and $\langle x, y \rangle$ (Yang & Salman, 2019), and therefore all the information of $\nabla c_{f^\star_{\hat{\alpha}}}$ is contained in $\operatorname{Span}\{x, y\}$. From this, we show in Figures 6 and 7 the gradient field $\nabla c_{f^\star_{\hat{\alpha}}}$ in the two-dimensional space $\operatorname{Span}\{x, y\}$ for different architectures and losses in the infinite-width regime described in Section 6 and in this section. Figure 6 corresponds to two-dimensional $x, y \in \mathbb{R}^2$, and Figure 7 to high-dimensional $x, y \in \mathbb{R}^{512}$. Note that in the plots, the gradient field is symmetric w.r.t. the horizontal axis and for this reason we have restricted the illustration to the case where the second coordinate is positive.

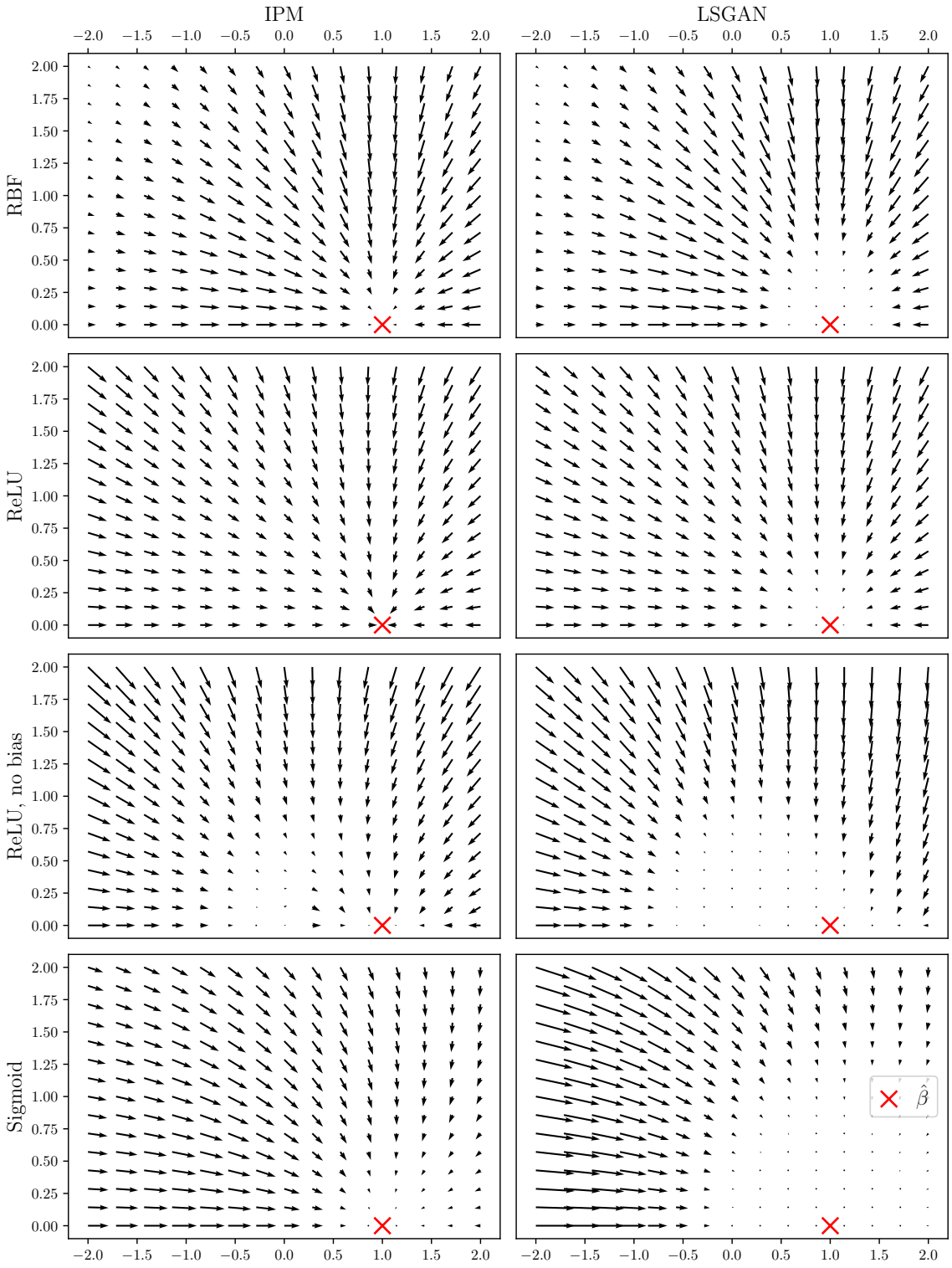


Figure 6. Gradient field $\nabla c_{f^\star_{\hat{\alpha}_x}}(x)$ received by a generated sample $x \in \mathbb{R}^2$ (i.e. $\hat{\alpha} = \hat{\alpha}_x = \delta_x$) initialized to x_0 with respect to its coordinates in $\mathrm{Span}\{x_0, y\}$ where y, marked by a $\times$, is the target distribution (i.e. $\hat{\beta} = \delta_y$), with $\|y\| = 1$. Arrows correspond to the movement of x in $\mathrm{Span}\{x_0, y\}$ following $\nabla c_{f^\star_{\hat{\alpha}_x}}(x)$, for different losses and networks; scales are specific for each pair of loss and network. The ideal case is the convergence of x along this gradient field towards the target y. Note that in the chosen orthonormal coordinate system, without loss of generality, y has coordinate $(1, 0)$; moreover, the gradient field is symmetrical with respect to the horizontal axis.

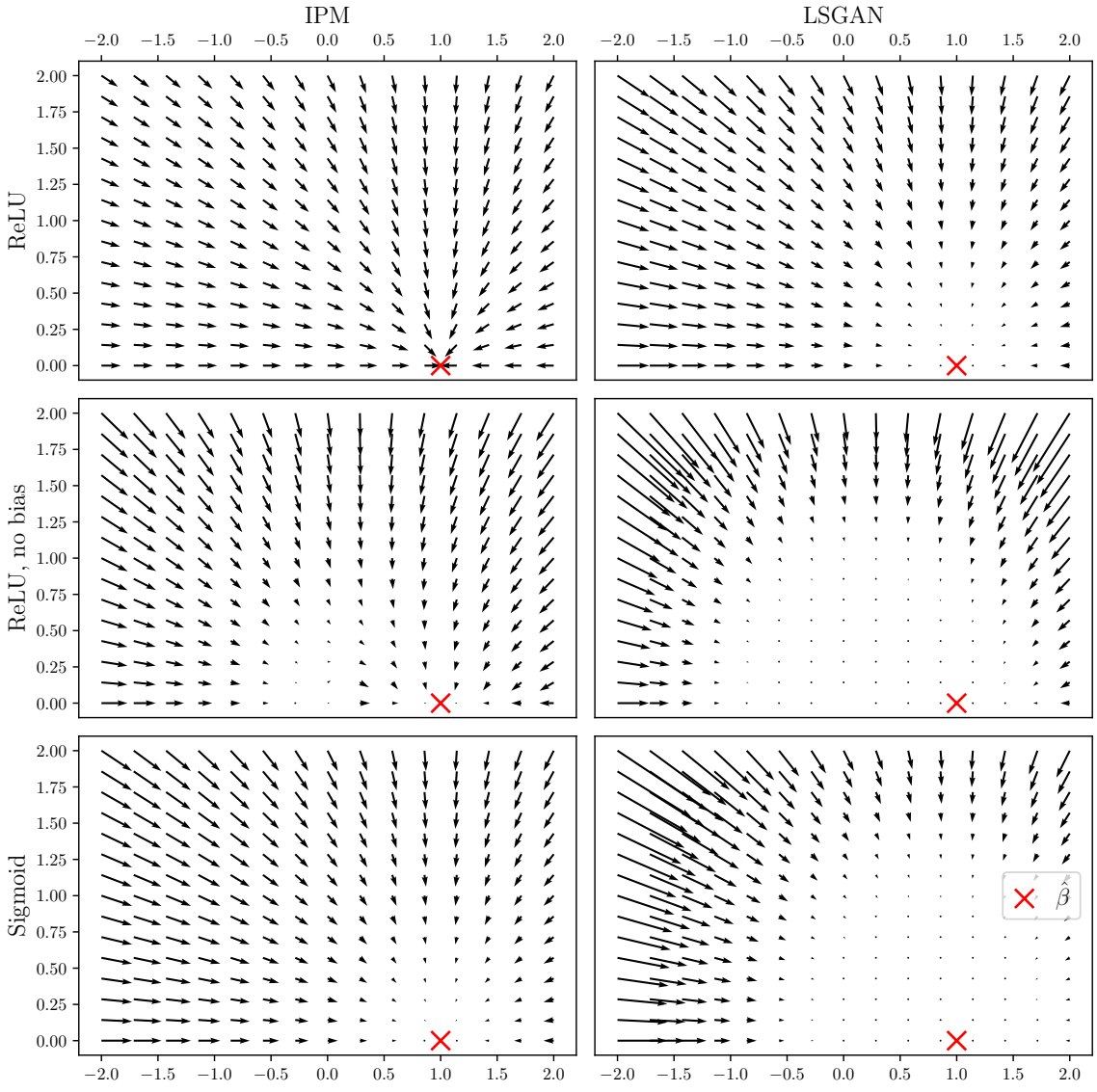


Figure 7. Same plot as Figure 6 but with underlying points $x, y \in \mathbb{R}^{512}$.

Convergence of the gradient flow. In the last paragraph, we have seen that the gradient field in the Dirac-GAN setting lives in the two-dimensional $\mathrm{Span}\{x, y\}$, independently of the dimensionality of x, y. This means that when training the generated distribution, as in Section 6, the position of the particle x always remains in this two-dimensional space, and hence (non-)convergence in this setting can be easily checked by studying this gradient field. This is what we do in the following, for different architectures and losses.

C.4.2. Qualitative Analysis of the Gradient Field

x is far from y. When generated outputs are far away from the target, it is essential that their gradient has a large enough magnitude in order to pull these points towards the target. The behavior of the gradients for distant points can be observed in the plots. For ReLU networks, for both losses, the gradients for distant points seem to be well behaved and large enough. Note that in the IPM case, the magnitude of the gradients is even larger when x is further away from y. This is not the case for the RBF kernel when the variance parameter is too small, as the magnitude of the gradient becomes prohibitively small. We highlight that we selected a large variance parameter in order to avoid such a behavior, but diminishing magnitudes can still be observed. Note that choosing an overly large variance may also have a negative impact on the points that are closer to the target.

x is close to y. A particularity of the NTK of ReLU discriminators with bias that arises from this study is that the gradients vanish more slowly when the generated x tends to the target y, compared to NTKs of ReLU without bias and sigmoid networks, and to the RBF kernel. We hypothesize that this is also another distinguishing feature that helps the generated distribution to converge more easily to the target distribution, especially when they are not far apart. On the contrary, this gradient vanishes more rapidly for NTKs of ReLU without bias and sigmoid networks, compared to the RBF kernel. This can explain the worse performance of such NTKs compared to the RBF kernel in our experiments (see Tables 1 to 3). Note that this phenomenon is even more pronounced in high-dimensional spaces such as in Figure 7.

x is close to 0. Finally, we highlight gradient vanishing and instabilities around the origin for ReLU networks without bias. This is related to its differentiability issues at the origin exposed in Section 4.3, and to its lack of representational power discussed in Appendix B.5. This can also be retrieved on larger scale experiments of Figures 2 and 3 where the origin is the source of instabilities in the descent.

Sigmoid network. It is also possible to evaluate the properties of the discriminator's gradient for architectures that are not used in practice, such as networks with the sigmoid activation. Figures 2 and 3 provide a clear explanation: as stated above, the magnitudes of the gradients become too small when $x \to y$, and heavily depend on the direction from which x approaches y. Ideally, the induced gradient flow should be insensitive to the direction in order for the convergence to be reliable and robust, which seems to be the case for ReLU networks.

D. Experimental Details

We detail in this section the experimental parameters needed to reproduce our experiments.

D.1. GAN(TK)^2 Specifications and Computing Resources

GAN(TK)^2 is implemented in Python (tested on versions 3.8.1 and 3.9.2) and based on JAX (Bradbury et al., 2018) for tensor computations and Neural Tangents (Novak et al., 2020) for NTKs. We refer to the code released at `https://github.com/emited/gantk2` for detailed specifications and instructions.

All experiments presented in this paper were run on Nvidia GPUs (Nvidia Titan RTX – 24GB of VRAM – with CUDA 11.2 as well as Nvidia Titan V – 12GB – and Nvidia GeForce RTX 2080 Ti – 11 GB – with CUDA 10.2). All two-dimensional experiments require only a few minutes of computations on a single GPU. Experiments on MNIST and CelebA were run using simultaneously four GPUs for parallel computations, for at most a couple of hours.

D.2. Datasets

8 Gaussians. The target distribution is composed of 8 Gaussians with their means being evenly distributed on the centered sphere of radius 5, and each with a standard deviation of 0.5. The input fake distribution is drawn at initialization from a standard normal distribution $\mathcal{N}(0, 1)$. We sample in our experiments 500 points from each distribution at each run to build

$\hat{\alpha}$ and $\hat{\beta}$.

AB and Density. These two datasets are taken from the Geomloss library examples (Feydy et al., 2019)[1] and are distributed under the MIT license. To sample a point from a distribution based on these greyscale images files, we sample a pixel (considered to lie in $[-1, 1]^2$) in the image from a distribution where each pixel probability is proportional to the darkness of this pixel, and then apply a Gaussian noise centered at the chosen pixel coordinates with a standard deviation equal to the inverse of the image size. We sample in our experiments 500 points from each distribution at each run to build $\hat{\alpha}$ and $\hat{\beta}$.

MNIST and CelebA. We preprocess each MNIST image (LeCun et al., 1998) by extending it from 28×28 frames to 32×32 frames (by padding it with black pixels). CelebA images (Liu et al., 2015) are downsampled from a size of 178×218 to 32×39 and then center-cropped to 32×32.

For both datasets, we normalize pixels in the $[-1, 1]$ range. For our experiments, we consider a subset of 1024 elements of each dataset, which are randomly sampled for each run.

D.3. Parameters

Sinkhorn divergence. The Sinkhorn divergence is computed using the Geomloss library (Feydy et al., 2019), with a blur parameter of 0.001 and a scaling of 0.95, making it close to the Wasserstein $\mathcal{W}_2$ distance.

RBF kernel. The RBF kernel used in our experiments is the following:

$$k(x, y) = \mathrm{e}^{\frac{\|x-y\|^2}{2n}}, \tag{155}$$

where n is the dimension of x and y, i.e. the dimension of the data.

Architecture. We used for the neural networks of our experiments the standard NTK paramaterization (Jacot et al., 2018), with a scaling factor of 1 for matrix multiplications and, when bias in enabled, a multiplicative constant of 1 for biases (except for sigmoid where this bias factor is lowered to 0.2 to avoid saturating the sigmoid, and for CelebA where it is equal to 4). All considered networks are composed of 3 hidden layers and end with a linear layer. In the finite-width case, the width of these hidden layers is 128. We additionally use antisymmetrical initialization (Zhang et al., 2020), except for the finite-width LSGAN loss.

Discriminator optimization. Discriminators in the finite-width regime are trained using full-batch gradient descent without momentum, with one step per update to the distributions and the following learning rates ε:

- for the IPM loss: $\varepsilon = 0.01$;
- for the IPM loss with reset and LSGAN: $\varepsilon = 0.1$.

In the infinite-width limit, we use the analytic expression derived in Section 5 with training time $\tau = 1$ (except for MNIST and CelebA where $\tau = 1000$) and $f_0 = 0$ (through the initialization of Zhang et al. (2020)) to avoid the computational cost of accumulating discriminators' analytic expressions across the generator's optimization steps.

Point cloud descent. The multiplicative constant η over the gradient applied to each datapoint for two-dimensional problems is chosen as follows:

- for the IPM loss in the infinite-width regime: $\eta = 1000$;
- for the IPM loss in the finite-width regime: $\eta = 100$;
- for the IPM loss in the finite-width regime and discriminator reset: $\eta = 1000$;

[1] They can be downloaded at `https://github.com/jeanfeydy/geomloss/tree/main/geomloss/examples/optimal_transport/data`: AB corresponds to files `A.png` (source) and `B.png` (target), and Density corresponds to files `density_a.png` (source) and `density_a.png` (target).

- for LSGAN in the infinite-width regime: $\eta = 1000$;
- for LSGAN in the finite-width regime: $\eta = 1$.

We multiply η by 1000 when using sigmoid activations, because of the low magnitude of the gradients it provides. We choose for MNIST $\eta = 100$.

Training is performed for the following number of iterations:

- for 8 Gaussians: $20\,000$;
- for Density and AB: $10\,000$;
- for MNIST: $50\,000$.