A Neural Tangent Kernel Perspective of GANs

Jean-Yves Franceschi, Emmanuel de Bézenac, Ibrahim Ayed, Mickaël Chen, Sylvain Lamprier, Patrick Gallinari

A Neural Tangent Kernel Perspective of GANs

Jean-Yves Franceschi*
Sorbonne Université, CNRS, LIP6,
F-75005 Paris, France
jean-yves.franceschi@lip6.fr

Emmanuel de Bézenac*
Sorbonne Université, CNRS, LIP6,
F-75005 Paris, France
emmanuel.de-bezenac@lip6.fr

Ibrahim Ayed*
Sorbonne Université, CNRS, LIP6,
F-75005 Paris, France
ThereSIS Lab, Thales, Palaiseau, France
ibrahim.ayed@lip6.fr

Mickaël Chen
Valeo.ai, Paris, France
mickael.chen@valeo.com

Sylvain Lamprier
Sorbonne Université, CNRS, LIP6,
F-75005 Paris, France
sylvain.lamprier@lip6.fr

Patrick Gallinari
Sorbonne Université, CNRS, LIP6,
F-75005 Paris, France
Criteo AI Lab, Paris, France
patrick.gallinari@lip6.fr

Abstract

Theoretical analyses for Generative Adversarial Networks (GANs) generally assume an arbitrarily large family of discriminators and do not consider the characteristics of the architectures used in practice. We show that this framework of analysis is too simplistic to properly analyze GAN training. To tackle this issue, we leverage the theory of infinite-width neural networks to model neural discriminator training for a wide range of adversarial losses via its Neural Tangent Kernel (NTK). Our analytical results show that GAN trainability primarily depends on the discriminator's architecture. We further study the discriminator for specific architectures and losses, and highlight properties providing a new understanding of GAN training. For example, we find that GANs trained with the integral probability metric loss minimize the maximum mean discrepancy with the NTK as kernel. Our conclusions demonstrate the analysis opportunities provided by the proposed framework, which paves the way for better and more principled GAN models. We release a generic GAN analysis toolkit based on our framework that supports the empirical part of our study.

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 in gaining a better understanding of the training process, resulting in a vast literature on theoretical analyses of GANs (see related work below). A large portion of them focus on studying GAN loss functions to conclude about their comparative advantages.

*Equal contribution. Authors with equal contribution are listed in a randomly chosen order.

Preprint. Under review.

Yet, empirical evaluations (Lucic et al., 2018; Kurach et al., 2019) have shown that different GAN formulations can yield approximately the same performance in terms of sample quality and stability of the training algorithm, regardless of the chosen loss. This indicates that by focusing exclusively on the formal loss function, theoretical studies might not model practical settings adequately.

In particular, the discriminator being a trained neural network is not taken into account, nor are the corresponding inductive biases which might considerably alter the generator's loss landscape. Moreover, neglecting this constraint hampers the analysis of gradient-based learning of the generator on finite training sets, since the gradient from the associated discriminator is ill-defined everywhere. These limitations thus hinder the potential of theoretical analyses to explain GAN's empirical behaviour.

In this work, leveraging the recent developments in the theory of deep learning driven by Neural Tangent Kernels (NTKs; Jacot et al., 2018), we provide a framework of analysis for GANs incorporating explicitly the discriminator's architecture which comes with several advantages.

First, we prove that, in the proposed framework, under mild conditions on its architecture and its loss, the trained discriminator has strong differentiability properties; this result holds for several GAN formulations and standard architectures, thus making the generator's learning problem well-defined. This emphasizes the role of the discriminator's architecture in GANs trainability.

We then show how our framework can be useful to derive both theoretical and empirical analyses of standard losses and architectures. We highlight for instance links between Integral Probability Metric (IPM) based GANs and the Maximum Mean Discrepancy (MMD) given by the discriminator's NTK, or the role of the ReLU activation in GAN architectures. We evaluate the adequacy and practical implications of our theoretical framework and release the corresponding Generative Adversarial Neural Tangent Kernel ToolKit GAN(TK)[2].

Related Work

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 GANs training, contrary to empirical evidence. We build up on this observation and show that under mild conditions on the discriminator's architecture, the generator's loss and gradient are actually well-defined.

Another line of work analyzes the dynamics and convergence of the generated distribution (Nagarajan & Kolter, 2017; Mescheder et al., 2017, 2018). As the studied dynamics are highly non-linear, this approach typically requires strong simplifying assumptions, e.g. restricting to linear neural networks or reducing datasets to a single datapoint. More recently, Mroueh & Nguyen (2021) proposed to study GANs using RKHSs, and Jacot et al. (2019) improve generator training by investigating checkerboard artifacts in the light of NTKs. The most advanced modelizations taking into account discriminator's parameterization are specialized to specific models (Bai et al., 2019), such as a linear one, or random feature models (Liu et al., 2017; Balaji et al., 2021). In contrast to these works, we are able to establish generally applicable results about the influence of the discriminator's architecture.

By taking into account the parameterization of discriminators for a wide range of architectures, our framework of analysis provides a more complete modelization of GANs.

Neural Tangent Kernel. 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, hereby making the dynamics of the training algorithm 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, 2020; Alemohammad et al., 2021). These prior works validate that NTKs can encapsulate the characteristics of neural network architectures, providing a solid theoretical basis to study the effect of architecture on learning problems.

While other works have studied the regularity of NTKs (Bietti & Mairal, 2019; Yang & Salman, 2019; Basri et al., 2020), as far as we know, ours is the first to state general derivability results for NTKs and infinite-width networks, as well as the first to leverage the theory of NTKs to study GANs.

2 Modeling GAN's Discriminator

We present in this section the usual GAN formulation and learning procedure, illustrate the limitations of prior analyses and introduce our framework which we develop in the remaining of the paper.

First, we 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.

2.1 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 network $f \in \mathcal{F}$ 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:

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

For instance, Goodfellow et al. (2014) originally used $a(x) = \log\big(1 - \sigma(x)\big)$, $b(x) = c(x) = -\log\big(\sigma(x)\big)$; 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 (IPMs; Müller, 1997) leveraged for example 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:

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

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 training. 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, the following sections are devoted to developing a better understanding of this gradient field when the discriminator is a neural network. In order to characterize the discriminator's gradient field, we must first study the discriminator itself.

2.2 On the Necessity of Modeling the Discriminator Parameterization

For each GAN formulation, it is customary to elucidate which loss is implemented by Equation (2), often assuming that $\mathcal{F} = L^2(\Omega)$, i.e. the discriminator can take arbitrary values. Under this assumption, the original GAN yields the Jensen-Shannon divergence between α_g and β, and LSGAN a Pearson χ^2-divergence, 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 $\hat{\alpha}$ and $\hat{\beta}$. Let $\hat{\gamma} = \frac{1}{2}\hat{\alpha} + \frac{1}{2}\hat{\beta}$ be the distribution of training samples.

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

In this setting, the Jensen-Shannon and χ^2-divergence are constant since $\hat{\alpha}$ and $\hat{\beta}$ generally do not have the same support. This is the theoretical reason given by Arjovsky & Bottou (2017) to introduce new losses, such as in WGAN (Arjovsky et al., 2017). However, this is inconsistent with empirical results showing that GANs can be trained with these losses. Actually, perhaps surprisingly, in the alternating optimization setting used in practice – as described by Equation (3) – the constancy of $\mathcal{L}_{\hat{\alpha}}$ does not imply that $\nabla_x c_{f^\star_{\hat{\alpha}}}$ in Equation (3) is zero on these points; see Section 4.2 and Appendix B.2 for further discussion on this matter. Yet, in their theoretical framework where the discriminator can take arbitrary values, this gradient field is not even defined for any loss $\mathcal{L}_{\hat{\alpha}}$.

Indeed, when the discriminator's loss $\mathcal{L}_{\hat{\alpha}}(f)$ is only computed on the empirical distribution $\hat{\gamma}$ (as it is the case for most GAN formulations), the discriminator optimization problem of Equation (2) never yields a unique optimal solution outside $\hat{\gamma}$. This is illustrated by the following straightforward result.

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

In particular, the set of solutions, if non-empty, contains non-differentiable discriminators as well as discriminators with null or non-informative gradients. This underspecification of the discriminator over Ω makes the gradient of the optimal discriminator in standard GAN analyses ill-defined.

This signifies that the loss alone does not impose any constraint on the values that $f_{\hat{\alpha}}$ takes outside $\operatorname{supp}\hat{\gamma}$, and more particularly that there are no constraints on the gradients. Therefore, an analysis beyond the loss function is necessary to precisely define the learning problem of the generator.

2.3 Modeling Inductive Biases of the Discriminator in the Infinite-Width Limit

In practice, however, the inner optimization problem of Equation (2) is not solved exactly. Instead, a proxy discriminator is computed using several steps of gradient ascent. For a learning rate ε, this results in the optimization procedure, from $i = 0$ to N:

$$\vartheta_{i+1} = \vartheta_i + \varepsilon \nabla_\vartheta \mathcal{L}_\alpha(f_{\vartheta_i}), \qquad\qquad f^\star_\alpha = f_{\vartheta_N} \tag{4}$$

In the following, we show that by modeling the discriminator as the result of a gradient ascent in a set of parameterized neural networks, the problem is no longer unspecified. To facilitate theoretical analyses of discriminator training, we consider the continuous-time equivalent of Equation (4):

$$\partial_t \vartheta_t = \nabla_\vartheta \mathcal{L}_\alpha(f_{\vartheta_t}), \tag{5}$$

which we study 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 tends to infinity, Jacot et al. (2018) showed that its so-called Neural Tangent Kernel (NTK) k_ϑ remains constant during a gradient ascent such as Equation (5), i.e. there is a limiting kernel k^∞ such that:

$$\forall \tau \in \mathbb{R}_+,\ \forall x, y \in \mathbb{R}^n,\ \forall t \in [0, \tau],\ k_{\vartheta_t}(x, y) \triangleq \partial_\vartheta f_t(x)^T \partial_\vartheta f_t(y) = k^\infty(x, y). \tag{6}$$

In particular, $k^\infty = k$ 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 for GAN algorithms in the setting of Equation (2). We discuss in details the applicability of this approximation in Appendix B.1.

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 and assumptions.

Definition 1 (Functional Gradient)**.** Whenever a functional $\mathcal{L}\colon L^2(\mu) \to \mathbb{R}$ has sufficient regularity, its gradient with respect to 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 $\psi \in L^2(\mu)$:

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

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

Definition 2 (RHKS 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:

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

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 $\mathcal{H}_k^\mu \subseteq L^2(\Omega)$.

In this infinite-width limit, Jacot et al. (2018) show that the discriminator $f_t \triangleq f_{\vartheta_t}$ trained by Equation (9) obeys the following differential equation:

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

In the following, we will rely on this differential equation to gain a better understanding of the discriminator during training, and its implications for training the generator.

3 General Analysis of The Discriminator and its Gradients

In the previous section, we highlighted the indeterminacy issues arising in common analysis settings, showing the need for a theory beyond the loss function. From this, we proposed a novel analysis framework by considering the discriminator trained using gradient descent in the NTK regime. In this section, under mild assumptions on the discriminator loss function, we prove that Equation (9) admits a unique solution for a given initial condition, thereby solving the indeterminacy issues. We then study differentiability of neural networks in this regime, a necessary condition for trainability of GANs. The results presented in this section are not specific to GANs but generalize to all neural networks trained under empirical losses of the form of Equation (2), e.g. any pointwise loss such as binary classification and regression. All results, presented in the following in the context of a discrete distribution $\hat{\gamma}$ but that generalize to other distributions, are proved in Appendix A.

3.1 Existence, Uniqueness and Characterization of the Discriminator

The following is a positive result on the existence and uniqueness of the discriminator that also characterizes the general form of the discriminator, amenable to theoretical analysis.

Assumption 3. *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 $f_\cdot \colon \mathbb{R}_+ \to L^2(\Omega)$. Moreover, the following holds:*

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

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 2.2 of determining the discriminator associated to $\hat{\gamma}$ is now resolved. It is now possible to study the discriminator in its general form thanks to Equation (10). It involves two terms: the neural network at initialization f_0, as well as the kernel operator of an integral. This integral is a function that is undefined outside $\operatorname{supp} \hat{\gamma}$, as by definition $\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s) \in L^2(\hat{\gamma})$. 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. *Under Assumptions 1 to 3, $f_t - f_0$ belongs to the RKHS $\mathcal{H}_k^{\hat{\gamma}}$ for all $t \in \mathbb{R}_+$.*

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

3.2 Differentiability of the Discriminator and its NTK

We study in this section the smoothness, i.e. infinite differentiability, of the discriminator. It mostly relies on the differentiability of the kernel k, by Equation (10). Therefore, we prove differentiability of 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) with activations that are smooth everywhere except on a closed set D of null Lebesgue measure.*

This last assumption covers in particular the sigmoid and all ReLU-like activations.

Assumption 5 (Discriminator regularity)**.** $0 \notin D$*, or linear layers have non-null bias terms.*

We first prove the differentiability of the NTK under these assumptions.

Proposition 2 (Differentiability of k)**.** *The NTK of an architecture following Assumption 4 is smooth everywhere over Ω^2 except on points of the form $(0, x)$ or $(x, 0)$. If Assumption 5 is also assumed, the NTK is then smooth everywhere.*

Remark 1. This result contradicts Bietti & Mairal (2019) about the non-Lipschitzness of the bias-free ReLU kernel, that we prove to be incorrect. We further discuss this matter in Appendix B.3.

From Proposition 2, NTKs satisfy Assumption 2. We can thus use Theorem 1 and conclude about the differentiability of f_t.

Theorem 2 (Differentiability of f_t, informal)**.** *Suppose that k is an NTK of a network following Assumption 4. Then f_t has the same regularity as k over Ω.*

Remark 2. ReLU networks with two or more layers and no bias are not differentiable at 0. However, by introducing non-zero bias, the NTK and the infinite-width discriminator become differentiable everywhere. This observation explains some experimental results in Section 5.

This result demonstrates that, for a wide range of GAN formulations, e.g. vanilla GAN and LSGAN, the optimized discriminator indeed admits gradients almost everywhere, making the gradient flow given to the generator well-defined. This supports our motivation to bring the theory closer to empirical evidence indicating that many GAN models do work in practice where their theoretical interpretation until now has been stating the opposite (Arjovsky & Bottou, 2017).

4 Fined-Grained Study for Specific Losses

Further assumptions on the loss function are needed to enhance our understanding of the discriminator in Equation (10). Hence, we restrict our study to more specific cases. Proofs are detailed in Appendix A.

4.1 The IPM as an MMD with the NTK as its Kernel

We study the case of the IPM loss for the discriminator, with the following remarkable solutions.

Proposition 3 (IPM Discriminator)**.** *Under Assumptions 1 and 2, the solutions of Equation (9) for $a = b = \mathrm{id}$ are the functions of the form $f_t = f_0 + t f^*_{\hat{\alpha}}$, where $f^*_{\hat{\alpha}}$ is the unnormalized MMD witness function, yielding:*

$$f^*_{\hat{\alpha}} = \mathbb{E}_{x\sim\hat{\alpha}}\big[k(x,\cdot)\big] - \mathbb{E}_{y\sim\hat{\beta}}\big[k(y,\cdot)\big], \qquad \mathcal{L}_{\hat{\alpha}}(f_t) = \mathcal{L}_{\hat{\alpha}}(f_0) + t \cdot \mathrm{MMD}^2_k\Big(\hat{\alpha}, \hat{\beta}\Big). \tag{11}$$

Suppose that $f_0 = 0$; this is possible with the initialization scheme of Zhang et al. (2020). In the IPM case, the loss for the optimized discriminator is then proportional to the squared MMD distance (Gretton et al., 2012) with the NTK as kernel between the empirical generated and target distributions.

This connection is especially interesting as the MMD has thoroughly been studied in the literature (Muandet et al., 2017). If k is characteristic (a hypothesis 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).

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

Moreover, as $c = \mathrm{id}$, the spatial gradient of f_t received by the generator in Equation (3) is proportional to $\nabla_x f^\star_{\hat{\alpha}}$. As following the gradient field induced by the discriminator has been shown to be a proxy to describe the evolution of the generated samples (Mroueh et al., 2019), it is possible to analyze the evolution of the generated samples in this context through the gradient flow induced by the MMD (see Section 5). To this extent, results from Arbel et al. (2019) are directly transferable, including convergence guarantees and discretization properties. This is, to the best of our knowledge, the first work considering the use of NTKs as kernels for the MMD. A more in-depth study of this use case is out of the scope of this paper, but appears relevant considering this connection with GANs and its application to our empirical analysis in Section 5.

4.2 LSGAN, Convergence, and Emergence of New Divergences

Optimality of the discriminator can be proved when assuming that its loss function is well-behaved. In particular, when it is concave and bounded from above as it is the case e.g. for vanilla GAN and LSGAN, it is possible to study the convergence of the discriminator for large training times. Consider as an example the case of LSGAN, for which Equation (9) can be solved by slightly adapting the results from Jacot et al. (2018) in the context of regression.

Proposition 4 (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:*

$$f_t = \exp\left(-4t\mathcal{T}_{k,\hat{\gamma}}\right)(f_0 - \rho) + \rho, \qquad \rho = \frac{\mathrm{d}\left(\hat{\beta} - \hat{\alpha}\right)}{\mathrm{d}\left(\hat{\beta} + \hat{\alpha}\right)}. \tag{12}$$

In the previous result, ρ is the optimum of $\mathcal{L}_{\hat{\alpha}}$ over $L^2(\hat{\gamma})$. When k is positive definite over $\hat{\gamma}$ (see Appendix B.5 for more details), f_t tends to the optimum for $\mathcal{L}_{\hat{\alpha}}$ as f_t tends to ρ over $\operatorname{supp}\hat{\gamma}$. Nonetheless, unlike the discriminator with arbitrary values of Section 2.2, f_∞ is defined over all Ω thanks to the integral operator $\mathcal{T}_{k,\hat{\gamma}}$. It is also the solution to the minimum norm interpolant problem in the RKHS (Jacot et al., 2018), therefore explaining why the discriminator tends to not overfit in scarce data regimes (see Section 5), and consequently to have bounded gradients despite large training times, assuming its NTK is well-behaved. We also prove a more detailed generalization of this result for concave bounded losses in Appendix A.4, where the same optimality conclusion holds.

Following the discussion initiated in Section 2.2, and applying it to the case of LSGAN, similarly to the Jensen-Shannon, the resulting generator loss on discrete training data is constant. However, the previous result implies that the gradients received by the generator have no a priori reason to be null; see for instance the empirical analysis of Section 5. This observation raises the question of determining the actual loss minimized by the generator, which could be retrieved by analyzing the spatial gradient that it receives from the discriminator at each step. We were able to make the connection for the IPM loss in Section 4.1 thanks to Arbel et al. (2019) who leveraged the theory of Ambrosio et al. (2008), but this connection remains to be established for other adversarial losses. Furthermore, the same problem arises for gradients obtained from incompletely trained discriminators f_t. This is beyond the scope of this paper, but we believe that our work motivates such a study of actual divergences between distributions minimized by GANs.

5 Empirical Study

In this section, we present a selection of empirical results for different losses and architectures and evaluate the adequacy and practical implications of our theoretical framework in different settings; see Appendix C for more results. All experiments were designed and performed with the proposed Generative Adversarial Neural Tangent Kernel ToolKit GAN(TK)^2, that we publicly release at `https://github.com/emited/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.,

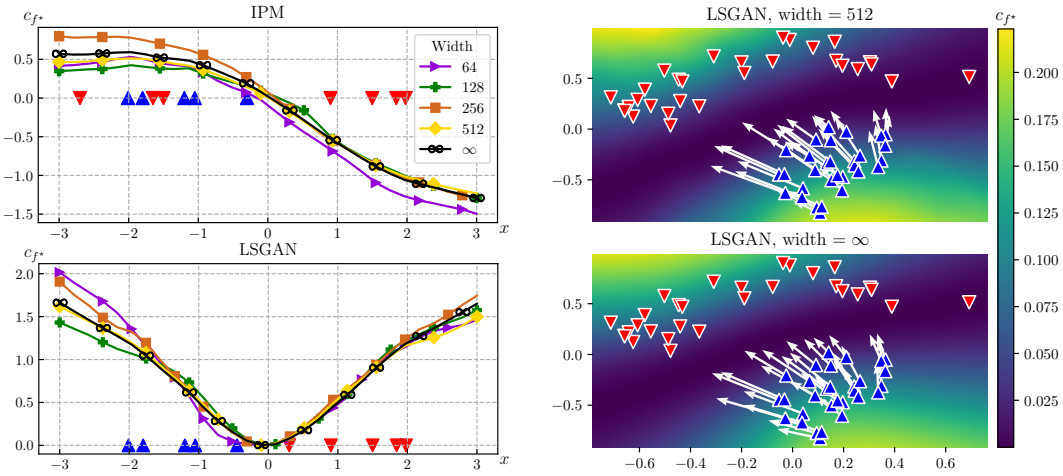

Figure 1: Values of $c_{f^\star}$ for LSGAN and IPM, where $f^\star$ is a 3-layer ReLU MLP with bias and varying width trained on the dataset represented by ▼ (fake) and ▲ (real) markers, initialized at $f_0 = 0$. The infinite-width network is trained for a time $\tau = 1$ and the finite-width networks 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.

2020), and is convenient to evaluate novel architectures and losses based on different visualizations and analyses.

Adequacy for fixed distributions. Firstly, we analyze the case where generated and target distributions are fixed. In this setting, we qualitatively study the similarity between the finite- and infinite-width regime of the discriminator and its gradients. Figure 1 shows $c_{f^\star}$ and its gradients on 1D and 2D data for LSGAN and IPM losses with a 3-layer ReLU MLP with varying widths. We find the behavior of finite-width discriminators to be close to their infinite-width counterpart for commonly used widths, and converges rapidly to the given limit as the width becomes larger.

In the following, we focus on the study of convergence of the generated distribution.

Experimental setting. We now consider a target distribution sampled from 8 Gaussians evenly distributed on a centered sphere (Figure 2), in a setup similar to that of Metz et al. (2017), Srivastava et al. (2017) and Arjovsky et al. (2017). As for the generated distribution, instead of implementing a generator network that would complexify the analysis beyond the scope of this paper, we follow Mroueh et al. (2019) and Arbel et al. (2019), and model its evolution considering a finite number of samples – initially Gaussian – moving in a direction that minimizes the loss as fast as possible, i.e. along the flow induced by the vector field $-\nabla_x c_{f^\star_\alpha}$. This setup is, in essence, similar to Equation (3); we refer to Mroueh et al. (2019) for a more formal description. For both 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 (see Figure 2 and Table 1). More precisely, we test a ReLU network with and without bias, and a "ReLU (reset)" network whose parameters are reinitialized at each generator step, corresponding to the setting of our infinite-width experiments. It is also possible to evaluate the advantages of the architecture in this setting 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 yields attractive properties (Muandet et al., 2017). Note that results for more datasets, including MNIST (LeCun et al., 1998), and architectures are also available in Appendix C.

Impact of initialization. In the analysis conducted in the previous sections, we have shown in Equation (10) that the discriminator depends on its initialization f_0 which can be non-zero and does not depend on the data. Thus, as it may bias the solution, we set $f_0 = 0$ in all the experiments, using the scheme proposed in Zhang et al. (2020). We have observed a global improvement in terms of speed of convergence of the samples.

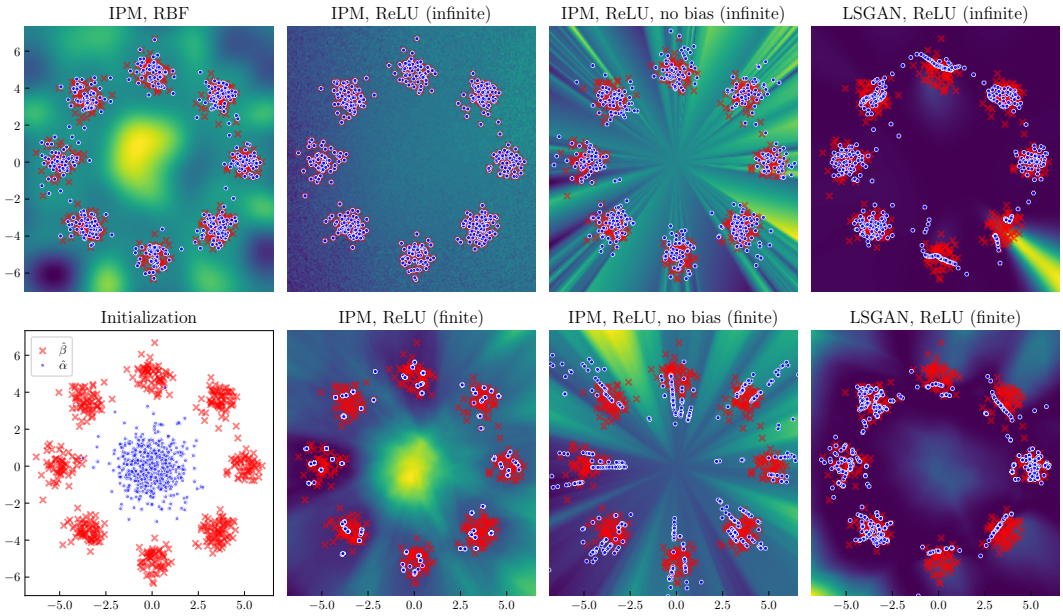

Figure 2: Generator (●) and target (×) samples for different methods. In the background, $c_{f^\star}$.

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)	ReLU (reset)
IPM (inf.)	$(2.60 \pm 0.06) \cdot 10^{-2}$	$(9.40 \pm 2.71) \cdot 10^{-7}$	$(9.70 \pm 1.88) \cdot 10^{-2}$	—
IPM	—	$(1.21 \pm 0.14) \cdot 10^{-1}$	$(1.20 \pm 0.60) \cdot 10^{0}$	$(1.97 \pm 0.31) \cdot 10^{-2}$
LSGAN (inf.)	$(4.21 \pm 0.10) \cdot 10^{-1}$	$(7.56 \pm 0.45) \cdot 10^{-2}$	$(1.27 \pm 0.01) \cdot 10^{1}$	—
LSGAN	—	$(3.07 \pm 0.68) \cdot 10^{0}$	$(7.52 \pm 0.01) \cdot 10^{0}$	$(2.14 \pm 0.59) \cdot 10^{0}$

Furthermore, we observe better results and more stable training when reinitializing the network at each step (see ReLU (reset)), which is closer to our theoretical framework. We believe that this is due to an inherent bias present in standard training of the discriminator as its parameters depend on old samples of the generator, and introduces a complex feedback loop. This surprising observation challenges the way we usually train discriminators and would constitute an interesting future investigation.

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

Impact of bias. The bias-free version of the 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 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 3.2 state that the ReLU NTK with bias is differentiable everywhere, whereas its bias-free version admits non-differentiability points, which can disrupt optimization based on its gradients: note in Figure 2 the abrupt streaks of the discriminator and its consequences on convergence.

NTK vs. RBF kernel. Finally, we observe the superiority of NTK w.r.t. to the RBF kernel. This highlights that the gradients of a ReLU network with bias are particularly well adapted to GANs. Visualizations of the gradients given by the ReLU architecture in the infinite-width limit are available

in Appendix C and further corroborate these findings. More generally, for the same reasons, we believe that the NTK of ReLU networks could be of particular interest for kernel methods requiring the computation of a spatial gradient, e.g. Stein Variational Gradient Descent (Liu & Wang, 2016).

6 Conclusion and Discussion

Contributions. Leveraging the theory of infinite-width neural networks, we proposed a framework of analysis of GANs explicitly modeling a large variety of discriminator architectures. We show that the proposed framework models more accurately GAN training compared to prior approaches by deriving properties of the trained discriminator. We demonstrate the analysis opportunities of the proposed modelization by further studying specific GAN losses and architectures, both theoretically and empirically, notably using our GAN analysis toolkit that we release publicly.

Limitations. Like all theoretical analyses, the proposed interpretation comes with its shortcomings, mostly tied to the NTK theory. In particular, this theory of infinite-width neural networks cannot in its current state model feature learning, which is an essential aspect of deep learning, although some progress have recently been made (Geiger et al., 2020; Yang & Hu, 2020). Beyond NTK-related issues, our framework does not encompass gradient penalties in GANs, since we directly tackle the issues that led to their introduction; we leave these considerations for future work.

Perspectives and broader impact. We believe that this work and its accompanying analysis toolkit will serve as a basis for more elaborate analyses – e.g. extending results to a more general setting or taking into account the generator's architecture – thus leading to more principled, better GAN models.

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, attempt such as ours at explaining generative models might also lead to ways to mitigate potential harms.

Acknowledgments and Disclosure of Funding

We would like to thank all members of the MLIA team from the LIP6 laboratory of Sorbonne Université for helpful discussions and comments.

We acknowledge financial support from 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. 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., Gigli, N., and Savaré, G. *Gradient Flows*. Birkhäuser Basel, 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. 6481–6491. 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. 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. 8139–8148. 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 over-parameterization 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 III, H. D. 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.

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. 12873–12884. 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.

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.

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):113301, November 2020.

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., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J., 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 *The 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 III, H. D. 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.

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

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

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., 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. 8570–8581. 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.

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. 2370–2378. 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. 5545–5553. Curran Associates, Inc., 2017.

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., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the 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. 1825–1835. 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.

Nagarajan, V. and Kolter, J. Z. Gradient descent GAN optimization is locally stable. 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. 5585–5595. Curran Associates, Inc., 2017.

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.

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. 3308–3318. 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.

Tancik, M., Srinivasan, 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.

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. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

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

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. PMLR, June 2019.

Supplementary Material

Table of Contents

A Proofs of Theoretical Results and Additional Results

We prove in this section all theoretical results mentioned in Sections 3 and 4. Appendix A.2 is devoted to the proof of Theorem 1, Appendix A.3 focuses on proving the differentiability results skimmed in Section 3.2, and Appendices A.4 and A.5 develop the results presented in Section 4.

A.1 Recall of Assumptions in the Paper

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

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

Assumption 3. 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) with activations that are smooth everywhere except on a closed set D of null Lebesgue measure.

Assumption 5 (Discriminator regularity)**.** $0 \notin D$, or linear layers have non-null bias terms.

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.

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:*

$$\forall u \in \mathcal{F}_{\delta T},\ \|u\| = \sup_{t \in [0, \delta T]} \|u_t\|_{L^2(\hat{\gamma})}. \tag{13}$$

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

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

$$\forall n, m, \|u_t^n - u_t^m\|_{L^2(\hat{\gamma})} \leq \|u^n - u^m\|, \tag{14}$$

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 $\epsilon > 0$, because (u^n) is Cauchy, we can choose N such that:

$$\forall n, m \geq N, \|u^n - u^m\| \leq \epsilon. \tag{15}$$

We thus have that:

$$\forall t, \forall n, m \geq N, \|u_t^n - u_t^m\|_{L^2(\hat{\gamma})} \leq \epsilon. \tag{16}$$

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

$$\forall t, \forall n \geq N, \|u_t^n - u_t^\infty\|_{L^2(\hat{\gamma})} \leq \epsilon, \tag{17}$$

which means that:

$$\forall n \geq N, \|u^n - u^\infty\| \leq \epsilon. \tag{18}$$

so that $(u^n)_n$ tends to u^∞.

Moreover, as:

$$\forall n, \|u_t^n\|_{L^2(\hat{\gamma})} \leq 1, \tag{19}$$

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:

$$\forall n, \|u_t^\infty - u_s^\infty\|_{L^2(\hat{\gamma})} \leq \|u_t^\infty - u_t^n\|_{L^2(\hat{\gamma})} + \|u_t^n - u_s^n\|_{L^2(\hat{\gamma})} + \|u_s^\infty - u_s^n\|_{L^2(\hat{\gamma})}. \tag{20}$$

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 $F \in L^2(\hat{\beta})$ with:*

$$\|F\|_{L^2(\hat{\alpha})} \leq \sqrt{2}\|F\|_{L^2(\hat{\gamma})} \text{ and } \|F\|_{L^2(\hat{\beta})} \leq \sqrt{2}\|F\|_{L^2(\hat{\gamma})}. \tag{21}$$

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

$$\|F\|^2_{L^2(\hat{\gamma})} = \frac{1}{2}\|F\|^2_{L^2(\hat{\alpha})} + \frac{1}{2}\|F\|^2_{L^2(\hat{\beta})}, \tag{22}$$

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

$$\|F\|^2_{L^2(\hat{\alpha})} = 2\|F\|_{L^2(\hat{\gamma})} - \|F\|_{L^2(\hat{\beta})} \leq 2\|F\|^2_{L^2(\hat{\gamma})} \tag{23}$$

$$\text{and } \|F\|^2_{L^2(\hat{\beta})} = 2\|F\|_{L^2(\hat{\gamma})} - \|F\|_{L^2(\hat{\alpha})} \leq 2\|F\|^2_{L^2(\hat{\gamma})}, \tag{24}$$

which allows us to conclude. □

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

Theorem 3. *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 $z \in \Omega$:

$$\left| \left(\mathcal{T}_{k,\hat{\gamma}} \left(\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) \right) - \mathcal{T}_{k,\hat{\gamma}} \left(\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(g) \right) \right)(z) \right| \leq \|k(z,\cdot)\|_{L^2(\hat{\gamma})} \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(g)\|_{L^2(\hat{\gamma})}. \tag{25}$$

Moreover, by definition:

$$\left\langle \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(g), h \right\rangle_{L^2(\hat{\gamma})} = \int (a'_f - a'_g) h \, \mathrm{d}\hat{\alpha} - \int (b'_f - b'_g) h \, \mathrm{d}\hat{\beta}, \tag{26}$$

so that:

$$\|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(g)\|^2_{L^2(\hat{\gamma})} \leq \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(g)\|_{L^2(\hat{\gamma})} (\|a'_f - a'_g\|_{L^2(\hat{\alpha})} + \|b'_f - b'_g\|_{L^2(\hat{\beta})}) \tag{27}$$

and then, along with Lemma 2:

$$\begin{aligned} \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) - \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(g)\|_{L^2(\hat{\gamma})} &\leq \|a'_f - a'_g\|_{L^2(\hat{\alpha})} + \|b'_f - b'_g\|_{L^2(\hat{\beta})} \\ &\leq \sqrt{2} \left(\|a'_f - a'_g\|_{L^2(\hat{\gamma})} + \|b'_f - b'_g\|_{L^2(\hat{\gamma})} \right). \end{aligned} \tag{28}$$

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

$$\forall x, \ |a'(f(x)) - a'(g(x))| \leq K_1 |f(x) - g(x)| \tag{29}$$

$$\text{and } \forall x, \ |b'(f(x)) - b'(g(x))| \leq K_2 |f(x) - g(x)|, \tag{30}$$

so that:

$$\|a'_f - a'_g\|_{L^2(\hat{\gamma})} \leq K_1 \|f - g\|_{L^2(\hat{\gamma})}, \qquad \|b'_f - b'_g\|_{L^2(\hat{\gamma})} \leq K_2 \|f - g\|_{L^2(\hat{\gamma})}. \tag{31}$$

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

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

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 $K = \sqrt{2}\,(K_1 + K_2)\sqrt{\int \|k(z,\cdot)\|^2_{L^2(\hat{\gamma})} \mathrm{d}\hat{\gamma}(z)}$ 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 $M = \left\| \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(0)\right) \right\|_{L^2(\hat{\gamma})}$.

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 $\mathcal{F}_{\delta T} = \mathcal{C}\left([0, \delta T], B_{L^2(\hat{\gamma})}(f_0, 1)\right)$. $\mathcal{F}_{\delta T}$ 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:

$$\forall u \in \mathcal{F}_{\delta T},\ \|u\| = \sup_{t \in [0, \delta T]} \|u_t\|_{L^2(\hat{\gamma})}. \tag{32}$$

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

$$\Phi(u)_t = f_0 + \int_0^t \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(u_s)\right) \mathrm{d}s. \tag{33}$$

We have, using Equation (B'):

$$\|\Phi(u)_t - f_0\|_{L^2(\hat{\gamma})} \leq \int_0^t K \|u_s\|_{L^2(\hat{\gamma})} + M \,\mathrm{d}s \leq (K + M)\delta T. \tag{34}$$

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

$$\forall u, v \in \mathcal{F}_{\delta T}, \|\Phi(u) - \Phi(v)\| \leq \frac{1}{2}\|u - v\|, \tag{35}$$

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'):

$$\forall t \in [0, T[, \|f_t\|_{L^2(\hat{\gamma})} \leq \|f_0\|_{L^2(\hat{\gamma})} + \int_0^t (K \|f_s\|_{L^2(\hat{\gamma})} + M) \,\mathrm{d}s, \tag{36}$$

which, using Gronwall's lemma, gives:

$$\forall t \in [0, T[, \|f_t\|_{L^2(\hat{\gamma})} \leq \|f_0\|_{L^2(\hat{\gamma})} \mathrm{e}^{KT} + \frac{M}{K}(\mathrm{e}^{KT} - 1). \tag{37}$$

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

$$\begin{aligned} \forall m \geq n, \|g^n - g^m\|_{L^2(\hat{\gamma})} &\leq \int_{T - \frac{1}{n}}^{T - \frac{1}{m}} (K \|f_s\| + M) \,\mathrm{d}s \\ &\leq \left(\frac{1}{n} - \frac{1}{m}\right)\left(\|f_0\|_{L^2(\hat{\gamma})} \mathrm{e}^{KT} + \frac{M}{K}(\mathrm{e}^{KT} - 1)\right). \end{aligned} \tag{38}$$

which shows that $(g^n)_n$ is a Cauchy sequence. $L^2(\hat{\gamma})$ being complete, we can thus consider its limit g^∞. Obviously, 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:

$$\forall t, \forall x, \tilde{f}_t(x) = f_0(x) + \int_0^t \mathcal{T}_{k,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)\Big)(x)\,\mathrm{d}s, \tag{39}$$

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

$$\mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}\left(\tilde{f}_s\right)\right) = \mathcal{T}_{k,\hat{\gamma}}\left(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}\left((\tilde{f}_s)_{|\mathrm{supp}(\hat{\gamma})}\right)\right) = \mathcal{T}_{k,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_s)\Big), \tag{40}$$

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 $\|k(z,\cdot)\|^2_{L^2(\hat{\gamma})}$ is a finite sum of terms $k(z,x_i)^2$ which shows that $\int \|k(z,\cdot)\|^2_{L^2(\hat{\gamma})}\mathrm{d}\Omega(z)$ 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 $h_{|\mathrm{supp}(\hat{\gamma})}$ coincides with f and thus with $\tilde{f}_{|\mathrm{supp}(\hat{\gamma})}$ in $L^2(\hat{\gamma})$ as we already proved uniqueness in the latter space. Thus, we have that $\|(h_s)_{|\mathrm{supp}(\hat{\gamma})} - (\tilde{f}_s)_{|\mathrm{supp}(\hat{\gamma})}\|_{L^2(\hat{\gamma})} = 0$ for any s. Now, we have:

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

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

$$\forall x, |a'(f(x))| \leq K_1|f(x)| + M_1, \tag{41}$$

where $M_1 = |a'(0)|$. We can then write, using Lemma 2 and the inequality from Equation (37):

$$\begin{aligned}
\forall s \leq T, \left\|a'\right\|_{L^2\left((f_s)_\sharp\hat{\alpha}\right)} &\leq K_1\sqrt{2}\|f_s\|_{L^2(\hat{\gamma})} + M_1 \\
&\leq K_1\sqrt{2}\left(\|f_0\|_{L^2(\hat{\gamma})}\mathrm{e}^{KT} + \frac{M}{K}(\mathrm{e}^{KT} - 1)\right) + M_1,
\end{aligned} \tag{42}$$

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 5. *Under Assumptions 1 to 3, the following integral inversion holds:*

$$f_t = f_0 + \int_0^t \mathcal{T}_{k_f,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha},\hat{\beta}}(f_s)\Big)\,\mathrm{d}s = f_0 + \mathcal{T}_{k_f,\hat{\gamma}}\left(\int_0^t \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha},\hat{\beta}}(f_s)\,\mathrm{d}s\right). \tag{43}$$

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

$$\left\langle \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f), h \right\rangle_{L^2(\hat{\gamma})} = \mathrm{d}\mathcal{L}_{\hat{\alpha}}(f)[h] = \int a'_f h \, \mathrm{d}\hat{\alpha} - \int b'_f h \, \mathrm{d}\hat{\beta}. \tag{44}$$

We can then write:

$$\|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)\|^2_{L^2(\hat{\gamma})} = \left\langle \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t), \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t) \right\rangle_{L^2(\hat{\gamma})} = \int a'_{f_t} \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t) \, \mathrm{d}\hat{\alpha} - \int b'_{f_t} \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t) \, \mathrm{d}\hat{\beta} \tag{45}$$

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

$$\begin{aligned} \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)\|^2_{L^2(\hat{\gamma})} &\leq \int |a'_{f_t}| |\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)| \mathrm{d}\hat{\alpha} + \int |b'_{f_t}| |\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)| \mathrm{d}\hat{\beta} \\ &\leq \|a'_{f_t}\|_{L^2(\hat{\alpha})} \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)\|_{L^2(\hat{\alpha})} + \|b'_{f_t}\|_{L^2(\hat{\beta})} \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)\|_{L^2(\hat{\beta})} \\ &\leq \sqrt{2} \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)\|_{L^2(\hat{\gamma})} \left[\|a'_{f_t}\|_{L^2(\hat{\alpha})} + \|b'_{f_t}\|_{L^2(\hat{\beta})} \right], \end{aligned} \tag{46}$$

which then gives us:

$$\|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_t)\|_{L^2(\hat{\gamma})} \leq \sqrt{2} \left[\|a'\|_{L^2((f_t)_\sharp \hat{\alpha})} + \|b'\|_{L^2((f_t)_\sharp \hat{\beta})} \right]. \tag{47}$$

By the Cauchy-Schwarz inequality and Equation (47), we then have for all z:

$$\begin{aligned} \int_0^t \int_x |k(z,x) \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s)(x)| \mathrm{d}\hat{\gamma}(x) \, \mathrm{d}s &\leq \int_0^t \|k(z,\cdot)\|_{L^2(\hat{\gamma})} \|\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s)\|_{L^2(\hat{\gamma})} \mathrm{d}s \\ &\leq \sqrt{2} \|k(z,\cdot)\|_{L^2(\hat{\gamma})} \int_0^t \left[\|a'\|_{L^2((f_s)_\sharp \hat{\alpha})} + \|b'\|_{L^2((f_s)_\sharp \hat{\beta})} \right] \mathrm{d}s. \end{aligned} \tag{48}$$

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

$$\begin{aligned} \int_0^t \mathcal{T}_{k_f, \hat{\gamma}} \left(\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s) \right) \mathrm{d}s &= \int_0^t \int_x k(\cdot, x) \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s)(x) \, \mathrm{d}\hat{\gamma}(x) \, \mathrm{d}s \\ &= \int_x k(\cdot, x) \left[\int_0^t \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s)(x) \, \mathrm{d}s \right] \mathrm{d}\hat{\gamma}(x) \\ &= \mathcal{T}_{k_f, \hat{\gamma}} \left(\int_0^t \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s)(x) \, \mathrm{d}s \right). \end{aligned} \tag{49}$$

□

A.3 Differentiability of Neural Tangent Kernels

Neural Tangent Kernels and the initialization of infinite-width functions being closely related to Gaussian Processes (GP), we first prove the following lemma showing the regularity of samples of a GP from the regularity of the corresponding kernel.

Lemma 4. *Let $A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, a symmetric kernel. Let V an open set such that A is $\mathcal{C}^\infty$ on $V \times V$. Then the Gaussian Process 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 $\partial^\alpha f \in C^k(V)$. This proves the lemma. □

We then tackle the differentiability of a key kernel in the theory of infinite-width neural networks (Jacot et al., 2018).

Lemma 5. *Let $A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, a symmetric, positive semi-definite kernel and $\phi : \mathbb{R} \to \mathbb{R}$. Define:*

$$\forall x, y \in \mathbb{R}^n, B(x, y) = \mathbb{E}_{f\sim\mathcal{GP}(0,A)} \left[\phi(f(x))\phi(f(y))\right]. \tag{50}$$

Moreover, suppose ϕ is $\mathcal{C}^\infty$ on an open set $O \subset \mathbb{R}$ such that $\mathbb{R} - O$ is of Lebesgue measure 0. If $0 \in O$ and A is $\mathcal{C}^\infty$ everywhere, then B is $\mathcal{C}^\infty$ everywhere. If $0 \notin O$, then for every neighbourhood V of points (x, y) such that $A(x, x) > 0 A(y, y) > 0$, if A is $\mathcal{C}^\infty$ on V, then B is $\mathcal{C}^\infty$ on V.

Proof. We have:

$$\forall x, y \in \mathbb{R}^n, B(x, y) = \mathbb{E}_{f\sim\mathcal{GP}(0,A)} \left[[\mathbf{1}_{f(x)\in O} + \mathbf{1}_{f(x)\notin O}]\phi(f(x))\phi(f(y))\right]. \tag{51}$$

Let us now study, putting $\Sigma_A^{(x,y)} = \begin{pmatrix} A(x,x) & A(x,y) \\ A(x,y) & A(y,y) \end{pmatrix}$:

$$\mathbb{E}_{f\sim\mathcal{GP}(0,A)} \left[\mathbf{1}_{f(x)\notin O}\phi(f(x))\phi(f(y))\right] = \mathbb{E}_{(z,z')\sim\mathcal{N}(0,\Sigma_A^{(x,y)})} \left[\mathbf{1}_{z\notin O}\phi(z)\phi(z')\right]. \tag{52}$$

As z is sampled from a Gaussian distribution with zero mean, even if it is degenerate, the only negligible set which can have a non null probability for z to be in, is $\{0\}$.

First case: $0 \in O$. In this case, because $\mathbb{R} - O$ is of null Lebesgue measure, we have $\mathbb{P}(z \notin O) = 0$ which means that the last expected value, which is the integral of a function a.s. null, is also of value 0. In other words, we have:

$$\forall x, y, B(x, y) = \mathbb{E}_{f\sim\mathcal{GP}(0,A)} \left[\mathbf{1}_{f(x)\in O}\phi(f(x))\phi(f(y))\right]. \tag{53}$$

Moreover, by the same reasoning as above applied to y, we also have:

$$\forall x, y, B(x, y) = \mathbb{E}_{f\sim\mathcal{GP}(0,A)} \left[\mathbf{1}_{f(x)\in O}\mathbf{1}_{f(y)\in O}\phi(f(x))\phi(f(y))\right]. \tag{54}$$

Let $x_0, y_0 \in \mathbb{R}^n$ and consider open neighbourhoods of both, V_1, V_2, with compact closures which we denote $cl(V_1)$ and $cl(V_2)$. Let us now consider a sample path f of the GP of kernel A. Lemma 4 then tells us that we can take f to be $\mathcal{C}^\infty$ in V_1 and V_2 with probability one. Let us also denote $V_1' = V_1 \cap f^{-1}(O)$ and $V_2' = V_2 \cap f^{-1}(O)$ which are open as O is open. In other words, $\phi \circ f$ is $\mathcal{C}^\infty$ on V_1' and V_2'.

Let $\alpha = (\alpha_1, ..., \alpha_n), \beta = (\beta_1, ..., \beta_n) \in \mathbb{N}^n$ such that $\sum_i \alpha_i \leq l$ and $\sum_i \beta_i \leq k$ for given k, l.

Using the usual notations for multi-indexed partial derivatives, via a multivariate Faà di Bruno formula (Leipnik & Pearce, 2007), we can write the derivative $\partial^\alpha(\phi \circ f)$ at $x \in V_1'$ as a sum of terms of the form:

$$\phi^{(j)}(f(x))g_1(x) \cdots g_N(x), \tag{55}$$

where the g_is are partial derivatives of f at x. As A is $\mathcal{C}^\infty$ everywhere, each of the g_is is thus a GP with a $\mathcal{C}^\infty$ covariance function. We can also write for all $x \in V_1'$:

$$\begin{aligned} |\phi^{(j)}(f(x))g_1(x) \cdots g_N(x)| &\leq \sup_{z\in\mathrm{cl}(V_1)} |\phi^{(j)}(f(z))g_1(z) \cdots g_N(z)| \\ &\leq \sup_{z_0\in\mathrm{cl}(V_1)} |\phi^{(j)}(f(z_0))| \sup_{z_1\in\mathrm{cl}(V_1)} |g_1(z_1)| \cdots \sup_{z_N\in\mathrm{cl}(V_1)} |g_N(z_N)|. \end{aligned} \tag{56}$$

For each i, because the covariance function of g_i is smooth over $\mathrm{cl}(V_1)$, its variance admits a maximum at in $\mathrm{cl}(V_1)$ and we take σ_i^2 the double of its value. We then know (Adler, 1990), that there is an M_i such that:

$$\forall n, \mathbb{E}_{f\sim\mathcal{GP}(0,A)} \left[\sup_{z_i\in\mathrm{cl}(V_1)} |g_N(z_i)|\right]^n \leq M_i^n \mathbb{E}|Y_i|^n, \tag{57}$$

where Y_i is a Gaussian distribution which variance is σ_i^2, the r.h.s. thus being finite.

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

$$\mathbb{E}\left[\sup_{z_0\in\mathrm{cl}(V_1)}|\phi^{(j)}(f(z_0))|\sup_{z_1\in\mathrm{cl}(V_1)}|g_1(z_1)|\cdots\sup_{z_N\in\mathrm{cl}(V_1)}|g_N(z_N)|\right]$$
$$\leq\sqrt{\mathbb{E}\left[(\sup_{z\in\mathrm{cl}(V_1)}\phi^{(j)}(f(z))|)^2\right]}\sqrt{\mathbb{E}\left[\sup_{z_1\in\mathrm{cl}(V_1)}|g_1(z_1)|^2\cdots\sup_{z_N\in\mathrm{cl}(V_1)}|g_N(z_N)|^2\right]}. \quad (58)$$

By iterated applications of the Cauchy-Schwarz inequality and using the previous arguments, we can then show that $\sup_{z_0\in\mathrm{cl}(V_1)}|\phi^{(j)}(f(z_0))|\sup_{z_1\in\mathrm{cl}(V_1)}|g_1(z_1)|\cdots\sup_{z_N\in\mathrm{cl}(V_1)}|g_N(z_N)|$ is indeed integrable.

The same reasoning applies to $\partial^\beta(\phi\circ f)$ and we can then write, by a standard corollary of the dominated convergence theorem:

$$\partial^{\alpha,\beta}B(x,y)_{|(x_0,y_0)}=\mathbb{E}_{f\sim\mathcal{GP}(0,A)}\left[\partial^\alpha(\phi\circ f)_{|x_0}\partial^\beta(\phi\circ f)_{|y_0}\right], \quad (59)$$

which shows that B is $\mathcal{C}^\infty$ on (x_0,y_0).

Second case: $0\notin O$**.** In this case, taking (x_0,y_0) such that $A(x_0,x_0)A(y_0,y_0)>0$, supposing A is $\mathcal{C}^\infty$ on a neighbourhood of (x_0,y_0) such that $A(x,x)A(y,y)>0$ on the neighbourhood, we have that $\mathbb{P}(f(x)\notin O \text{ or } f(y)\notin O)=0$ for any (x,y) in the neighbourhood as $A(x,x)>0$, $A(y,y)>0$ and O is of null Lebesgue measure. We can then prove that B is $\mathcal{C}^\infty$ on that same neighbourhood with the same reasoning as above. □

From this, we can prove the results skimmed in Section 3.2.

Proposition 6 (Proposition 2)**.** *Let k be the NTK of an architecture such as in Assumption 4. Then k is smooth on every (x,y) such that $x\neq 0$ and $y\neq 0$. Moreover, if we suppose the architecture verifies Assumption 5, then k is smooth everywhere.*

Proof. We define the following kernel:

$$\Sigma_L^\phi(x,y)=\mathbb{E}_{f\sim\mathcal{GP}(0,\Sigma_{L-1}^\phi)}\left[\phi(f(x))\phi(f(y))\right]+\beta^2, \quad (60)$$

with:

$$\Sigma_0^\phi(x,y)=x^Ty+\beta^2. \quad (61)$$

According to the definitions of Jacot et al. (2018), Arora et al. (2019) and Huang et al. (2020), the smoothness of the kernel, for an architecture of depth L is guaranteed whenever the kernels Σ_L^σ and $\Sigma_L^{\sigma'}$, where σ denotes in this proof the activation function, are smooth. Note that, in the case of ResNets, there is a slight adaptation of the formula defining Σ_L which does not change its regularity.

Under Assumption 5, if there are non-null bias terms, we have that $\Sigma_L^\phi(x,x)\geq\beta^2>0$ for every x and $\phi\in\sigma,\sigma'$ so that, by a recursion using Lemma 5 and as Σ_0^ϕ is clearly smooth, we have the smoothness of Σ_L. The same is true if the activation is smooth on 0.

Now let us consider (x,y) such that $x\neq 0$ and $y\neq 0$. Then, either σ is constant everywhere, which automatically proves smoothness, or, for $\phi\in\sigma,\sigma'$, as $\Sigma_0^\phi(x,x)\geq x^Tx>0$ and $\Sigma_1^\phi(x,x)\geq\mathbb{E}_{z\sim N(0,\Sigma_0^\phi(x,x))}[\phi(z)^2]>0$. We can continue this reasoning by recursion thus proving that $\Sigma_L^\phi(x,x)>0$. The same applies for y and we can then use Lemma 5 to prove the desired result. □

Theorem 4 (Theorem 2)**.** *Let f_t be a solution to Equation (9) under Assumptions 1 and 3 by Theorem 1, with k the NTK of a neural network and f_0 an initialization of the latter.*

Then, under Assumption 4, if $0\notin\operatorname{supp}\hat{\gamma}$,[1] f_t is smooth on any point $x\neq 0$. Under Assumption 5, f_t is smooth everywhere.

[1]Note that this is verified with probability 1 on the sampling of the dataset except if γ is concentrated on 0. Moreover, as all distributions are supposed to be compactly supported, they can always be shifted so that this is verified.

Proof. We observe that $\mathcal{T}_{k,\hat{\gamma}}(g)$ has a regularity which only depends on the regularity of $k(\cdot, x)$ for $x \in \operatorname{supp}\hat{\gamma}$: if $k(\cdot, x)$ is smooth in a certain neighbourhood V for every such x, we can bound $\partial^\alpha k(\cdot, x)$ on V for every x and then use dominated convergence to prove that $\mathcal{T}_{k,\hat{\gamma}}(g)(\cdot)$ is smooth on V. The theorem then follows from the previous results and the fact that f_0 has the same regularity as Σ_L^σ defined in the proof of the last proposition, which is the same as k, as well as the fact that $f_t - f_0 = \mathcal{T}_{k,\hat{\gamma}}\left(\int_0^t \nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f_s)\,\mathrm{d}s\right)$. □

A.4 Optimality in Concave Setting

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

A.4.1 Assumptions

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

Assumption 6. *k is positive definite over $\hat{\gamma}$.*

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

$$\begin{aligned} \mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}} \colon L^2(\hat{\gamma}) &\to L^2(\hat{\gamma}) \\ h &\mapsto \mathcal{T}_{k,\hat{\gamma}}(h)\big|_{\operatorname{supp}\hat{\gamma}} \end{aligned}, \tag{62}$$

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 7. *$\mathcal{L}_{\hat{\alpha}}$ 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 8. *$t \mapsto f_t|_{\operatorname{supp}\hat{\gamma}}$ 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.4.2 Optimality Result

Proposition 7 (Asymptotic optimality)**.** *Under Assumptions 1 to 3 and 6 to 8, f_t converges pointwise when $t \to \infty$, and:*

$$\mathcal{L}_{\hat{\alpha}}(f_t) \xrightarrow[t\to\infty]{} \mathcal{L}_{\hat{\alpha}}(y^\star), \quad f_\infty = f_0 + \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}^{-1}\left(y^\star - f_0|_{\operatorname{supp}\hat{\gamma}}\right)\right), \quad f_\infty|_{\operatorname{supp}\hat{\gamma}} = y^\star, \tag{63}$$

where we recall that:

$$y^\star = \operatorname*{arg\,max}_{y \in L^2(\hat{\gamma})} \mathcal{L}_{\hat{\alpha}}(y). \tag{64}$$

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 6. *Under Assumptions 1 to 3 and 6 to 8, $\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t) \to 0$ 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:

$$\frac{\mathrm{d}\mathcal{L}_{\hat{\alpha}}(f_t)}{\mathrm{d}t} = \left\langle \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t), \frac{\mathrm{d}\hat{f}_t}{\mathrm{d}t} \right\rangle_{L^2(\hat{\gamma})} = \left\langle \nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t), \mathcal{T}_{k,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\Big) \right\rangle_{L^2(\hat{\gamma})}, \tag{65}$$

using Equation (9). This equates to:

$$\frac{\mathrm{d}\mathcal{L}_{\hat{\alpha}}(f_t)}{\mathrm{d}t} = \left\| \mathcal{T}_{k,\hat{\gamma}}\Big(\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t)\Big) \right\|^2_{\mathcal{H}_k^{\hat{\gamma}}} \geq 0, \tag{66}$$

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 $\frac{\mathrm{d}\mathcal{L}_{\hat{\alpha}}(f_t)}{\mathrm{d}t} \to 0$ 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 8 gives that $t \mapsto f_t|_{\operatorname{supp}\hat{\gamma}}$ is uniformly continuous over $\mathbb{R}_+$. Therefore, their composition $t \mapsto \frac{\mathrm{d}\mathcal{L}_{\hat{\alpha}}(f_t)}{\mathrm{d}t}$ (from Equation (66)) is uniformly continuous over $\mathbb{R}_+$. Using Barbălat's Lemma (Farkas & Wegner, 2016), we conclude that $\frac{\mathrm{d}\mathcal{L}_{\hat{\alpha}}(f_t)}{\mathrm{d}t} \to 0$ when $t \to \infty$.

Furthermore, k is positive definite over $\hat{\gamma}$ by Assumption 6, so $\|\cdot\|_{\mathcal{H}_k^{\hat{\gamma}}}$ is actually a norm. Therefore, since $\operatorname{supp}\hat{\gamma}$ is finite, the following pointwise convergence holds:

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f_t) \xrightarrow[t\to\infty]{} 0. \tag{67}$$

□

Lemma 7. *Under Assumptions 1, 2 and 6, for all $f \in \mathcal{H}_k^{\hat{\gamma}}$, the following holds:*

$$f = \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}}\Big(f|_{\operatorname{supp}\hat{\gamma}} \Big) \right) \tag{68}$$

Proof. Since k is positive definite by Assumption 6, then $\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}$ from Equation (62) 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 $h = \mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}}\Big(f|_{\operatorname{supp}\hat{\gamma}} \Big)$. □

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 6, 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 7:

$$f_t - f_0 = \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}}\Big((f_t - f_0)\big|_{\operatorname{supp}\hat{\gamma}} \Big) \right). \tag{69}$$

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

$$\mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}}\Big((f_t - f_0)\big|_{\operatorname{supp}\hat{\gamma}} \Big) \xrightarrow[t\to\infty]{} \mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}}\Big(y^\star - f_0|_{\operatorname{supp}\hat{\gamma}} \Big). \tag{70}$$

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

$$f_t \xrightarrow[t\to\infty]{} f_0 + \mathcal{T}_{k,\hat{\gamma}}\left(\mathcal{T}_{k,\hat{\gamma}}|^{-1}_{\operatorname{supp}\hat{\gamma}}\left(y^\star - f_0|_{\operatorname{supp}\hat{\gamma}}\right)\right). \tag{71}$$

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.5 Case Studies of Discriminator Dynamics

We study in the remaining of this section the expression of the discriminators in the case of the IPM loss and LSGAN, as described in Section 4, and of the original GAN formulation.

A.5.1 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. 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 an countable othornormal 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 defined the following Radon–Nikodym derivatives with inputs in $\operatorname{supp}\hat{\gamma}$:

$$\rho = \frac{\mathrm{d}\left(\hat{\beta} - \hat{\alpha}\right)}{\mathrm{d}\left(\hat{\beta} + \hat{\alpha}\right)}, \qquad \rho_1 = \frac{\mathrm{d}\hat{\alpha}}{\mathrm{d}\hat{\gamma}}, \qquad \rho_2 = \frac{\mathrm{d}\hat{\beta}}{\mathrm{d}\hat{\gamma}}, \tag{72}$$

knowing that

$$\rho = \frac{1}{2}(\rho_2 - \rho_1), \qquad \rho_1 + \rho_2 = 2. \tag{73}$$

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

Lemma 8. *Under Assumption 3:*

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = \rho_1 \cdot \left(a' \circ f\right) - \rho_2 \cdot \left(b' \circ f\right). \tag{74}$$

Proof. We have from Equation (2):

$$\mathcal{L}_{\hat{\alpha}}(f) = \mathbb{E}_{x\sim\hat{\alpha}}\left[a_f(x)\right] - \mathbb{E}_{y\sim\hat{\beta}}\left[b_f(y)\right] = \left\langle \rho_1, a_f\right\rangle_{L^2(\hat{\gamma})} - \left\langle \rho_2, b_f\right\rangle_{L^2(\hat{\gamma})}, \tag{75}$$

hence by composition:

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 \cdot \left(a' \circ f\right) - \rho_2 \cdot \left(b' \circ f\right) = \rho_1 a'_f - \rho_2 b'_f. \tag{76}$$

□

A.5.2 LSGAN

Proposition 4. 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:

$$f_t = \exp\left(-4t\mathcal{T}_{k,\hat{\gamma}}\right)(f_0 - \rho) + \rho = f_0 + \varphi_t\left(\mathcal{T}_{k,\hat{\gamma}}\right)(f_0 - \rho), \tag{77}$$

where

$$\varphi_t\colon x \mapsto \mathrm{e}^{-4tx} - 1. \tag{78}$$

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,

$$g\colon t \mapsto g_t = f_0 + \varphi_t\big(\mathcal{T}_{k,\hat{\gamma}}\big)(f_0 - \rho) \tag{79}$$

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 8, with $a_f = -(f+1)^2$ and $b_f = -(f-1)^2$:

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a_f' - \rho_2 b_f' = -2\rho_1(f+1) - 2\rho_2(f-1) = 4\rho - 4f. \tag{80}$$

So Equation (9) equates to:

$$\partial_t f_t = 4\mathcal{T}_{k,\hat{\gamma}}(\rho - f_t). \tag{81}$$

Now let us prove that g_t is a solution to Equation (81). We have:

$$\partial_t g_t = -4\Big(\mathcal{T}_{k,\hat{\gamma}} \circ \exp\big(-4t\mathcal{T}_{k,\hat{\gamma}}\big)\Big)(f_0 - \rho) = -4\Big(\mathcal{T}_{k,\hat{\gamma}} \circ \exp\big(-4t\mathcal{T}_{k,\hat{\gamma}}\big)\Big)(f_0 - \rho). \tag{82}$$

Restricted to $\operatorname{supp}\hat{\gamma}$, we can write from Equation (79):

$$g_t = f_0 + \left(\exp\Big(-4t\,\mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}}\Big) - \mathrm{id}_{L^2(\hat{\gamma})}\right)(f_0 - \rho), \tag{83}$$

and plugging this in Equation (82):

$$\partial_t g_t = -4\mathcal{T}_{k,\hat{\gamma}}(g_t - \rho), \tag{84}$$

where we retrieve the differential equation of Equation (81). Therefore, g_t is a solution to Equation (81).

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.5.3 IPMs

Proposition 3. 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:

$$f^*_{\hat{\alpha}} = \mathbb{E}_{x\sim\hat{\alpha}}\big[k(x,\cdot)\big] - \mathbb{E}_{y\sim\hat{\beta}}\big[k(y,\cdot)\big], \qquad \mathcal{L}_{\hat{\alpha}}(f_t) = \mathcal{L}_{\hat{\alpha}}(f_0) + t\cdot \mathrm{MMD}^2_k\Big(\hat{\alpha},\hat{\beta}\Big). \tag{85}$$

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,

$$g\colon t \mapsto g_t = f_0 + tf^*_{\hat{\alpha}} \tag{86}$$

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 8, with $a_f = b_f = f$:

$$\nabla^{\hat{\gamma}}\mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a_f' - \rho_2 b_f' = -2\rho. \tag{87}$$

So Equation (9) equates to:

$$\partial_t f_t = -2\mathcal{T}_{k,\hat{\gamma}}(\rho) = 2\int_x k(\cdot,x)\rho(x)\,\mathrm{d}\hat{\gamma}(x) = \int_x k(\cdot,x)\,\mathrm{d}\hat{\alpha}(x) - \int_y k(\cdot,y)\,\mathrm{d}\hat{\beta}(y), \tag{88}$$

by definition of ρ (see Equation (72)), yielding:

$$\partial_t f_t = f^*_{\hat{\alpha}}. \tag{89}$$

Clearly, $t \mapsto g_t = f_0 + t f^*_{\hat{\alpha}}$ 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 $a = b = \mathrm{id}$:

$$\mathcal{L}_{\hat{\alpha}}(f_t) = \mathcal{L}_{\hat{\alpha}}(f_0) + t \cdot \mathcal{L}_{\hat{\alpha}}(f^*_{\hat{\alpha}}) = \mathcal{L}_{\hat{\alpha}}(f_0) + t \cdot \mathcal{L}_{\hat{\alpha}}\big(\mathcal{T}_{k,\hat{\gamma}}(-2\rho)\big). \tag{90}$$

But, from Equation (75), $\mathcal{L}_{\hat{\alpha}}(f) = \langle -2\rho, f \rangle_{L^2(\hat{\gamma})}$, hence:

$$\mathcal{L}_{\hat{\alpha}}(f_t) = \mathcal{L}_{\hat{\alpha}}(f_0) + t \cdot \big\langle -2\rho, \mathcal{T}_{k,\hat{\gamma}}(-2\rho) \big\rangle_{L^2(\hat{\gamma})} = \mathcal{L}_{\hat{\alpha}}(f_0) + t \cdot \big\| \mathcal{T}_{k,\hat{\gamma}}(-2\rho) \big\|^2_{\mathcal{H}^{\hat{\gamma}}_k}. \tag{91}$$

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

$$\partial_t f_t = \mathcal{T}_{k,\hat{\gamma}}\big(\rho_2 - 2\sigma(f)\big). \tag{92}$$

Proof. We have from Lemma 8, with $a_f = b_f = f$:

$$\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = -\rho_1 \frac{\sigma'(f)}{1 - \sigma(f)} + \rho_2 \frac{\sigma'(f)}{\sigma(f)}. \tag{93}$$

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

$$\nabla^{\hat{\gamma}} \mathcal{L}_{\hat{\alpha}}(f) = \rho_1 a'_f - \rho_2 b'_f = -\rho_1 \sigma(f) + \rho_2\big(1 - \sigma(f)\big) = \rho_2 - 2\sigma(f). \tag{94}$$

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 (92) in $\mathbb{R}_+ \to L^2(\Omega)$ with a given initialization f_0.

Let us first study Equation (92) 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:*

$$\frac{\mathrm{d}y_t}{\mathrm{d}t} = \lambda\big(r - 2\sigma(y_t)\big) \tag{95}$$

is the following set:

$$S = \left\{ y \colon t \mapsto (1 - r)\left(W\left(\mathrm{e}^{2\lambda t + C}\right) - 2\lambda t - C \right) \;\middle|\; C \in \mathbb{R} \right\}, \tag{96}$$

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 (95) 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 (95) 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 (95). Let us express the derivative of y:

$$\frac{1}{1 - r} \frac{\mathrm{d}y_t}{\mathrm{d}t} = 2\lambda \left(\mathrm{e}^{2\lambda t + C} W'\left(\mathrm{e}^{2\lambda t + C}\right) - 1 \right). \tag{97}$$

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

$$\frac{1}{1-r}\frac{\mathrm{d}y_t}{\mathrm{d}t} = 2\lambda\left(\frac{W\big(\mathrm{e}^{2\lambda t+C}\big)}{1+W\big(\mathrm{e}^{2\lambda t+C}\big)} - 1\right) = -\frac{2\lambda}{1+W\big(\mathrm{e}^{2\lambda t+C}\big)}. \tag{98}$$

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

$$\frac{1}{1-r}\frac{\mathrm{d}y_t}{\mathrm{d}t} = -\frac{2\lambda}{1+\mathrm{e}^{2\lambda t+C}\mathrm{e}^{-W\left(\mathrm{e}^{2\lambda t+C}\right)}} = -\frac{2\lambda}{1+\mathrm{e}^{(r-1)y_t}}. \tag{99}$$

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

$$\lambda\big(r-2\sigma(y_t)\big) = -\frac{2\lambda(1-r)}{1+\mathrm{e}^{(r-1)y_t}}. \tag{100}$$

Therefore:

$$\frac{\mathrm{d}y_t}{\mathrm{d}t} = \lambda\big(r-2\sigma(y_t)\big) \tag{101}$$

and y_t is a solution to Equation (95).

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 (92) 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:

$$f_t = \varphi_t(x)(\rho_2 - 1) = -\varphi_t(x)(\rho), \tag{102}$$

where

$$\varphi_t \colon x \mapsto W\Big(\mathrm{e}^{2tx+1}\Big) - 2tx - 1, \tag{103}$$

since the initial condition $y_0 = 0$ gives the constant value $C = 1$ in Equation (96). 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 (102) cannot be a solution of Equation (92). Indeed, one can prove by following essentially the same reasoning as the proof of Proposition 8 that:

$$\partial_t f_t = 2\left(\mathcal{T}_{k,\hat{\gamma}} \circ \Big(\psi_t\big(\mathcal{T}_{k,\hat{\gamma}}\big)\Big)^{-1}\right)(\rho_2 - 1), \tag{104}$$

with

$$\psi_t \colon x \mapsto 1 + W\Big(\mathrm{e}^{2tx+1}\Big) > 0. \tag{105}$$

However, this does not allow us to obtain Equation (92) since in the latter the sigmoid is taken coordinate-wise, where the exponential in Equation (104) acts on matrices.

Nonetheless, for t small enough, f_t as defined in Equation (104) should approximate the solution of Equation (92), 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 (104) is actually close to the numerical solution of Equation (92) obtained using an ODE solver. Thus, we provide here an 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 referenced 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 2. 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, 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).

There are nevertheless some limits to this 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, 2020). 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 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 focus for the sake of simplicity on 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:

$$\mathcal{C}_{f^\star_{\alpha_\theta}}(\alpha_\theta) \triangleq \mathbb{E}_{x\sim\alpha_\theta}\left[c_{f^\star_{\alpha_\theta}}(x)\right] = f^\star_{\alpha_\theta}(x_\theta)^2, \tag{106}$$

where:

$$\begin{aligned} f^\star_{\alpha_\theta} &= \operatorname*{arg\,max}_{f\in L^2\left(\frac{1}{2}\alpha_\theta+\frac{1}{2}\beta\right)} \left\{\mathcal{L}_{\alpha_\theta}(f) \triangleq \mathbb{E}_{x\sim\alpha_\theta}\left[a_f(x)\right] - \mathbb{E}_{y\sim\beta}\left[b_f(y)\right]\right\} \\ &= \operatorname*{arg\,min}_{f\in L^2\left(\frac{1}{2}\alpha_\theta+\frac{1}{2}\beta\right)} \left\{\left(f^\star_{\alpha_\theta}(x_\theta)+1\right)^2 + \left(f^\star_{\alpha_\theta}(0)-1\right)^2\right\}. \end{aligned} \tag{107}$$

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

$$\mathcal{C}_{f^\star_{\alpha_\theta}}(\alpha_\theta) = 1. \tag{108}$$

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 2.2.

B.3 Differentiability of the Bias-Free ReLU Kernel

Theorem 2 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):

$$k(x, y) = \|x\|\|y\|\kappa\left(\frac{\langle x, y\rangle}{\|x\|\|y\|}\right), \tag{109}$$

where

$$\begin{aligned}\kappa\colon [0, 1] &\to \mathbb{R} \\ u &\mapsto \frac{2}{\pi}u(\pi - \arccos u) + \frac{1}{\pi}\sqrt{1 - u^2}\end{aligned}. \tag{110}$$

More particularly, Bietti & Mairal (2019, Proposition 3) claim that $k(\cdot, y)$ is not Lipschitz on y for all y in the unit sphere. By following their proof, it amounts to prove that $k(\cdot, y)$ is not Lipschitz on y for all y in any centered sphere. This would imply that k is not differentiable for all inputs (x, y), which contradicts our result. 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 $k_0\colon x, y \mapsto \kappa\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right)$ 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 $x \neq y \in \mathbb{R}^n$, $x \neq 0$ and $y \neq 0$:

$$\frac{\partial k_0(x, y)}{\partial x} = \frac{y\|x\| - \frac{x}{\|x\|}\langle x, y\rangle}{\|x\|^2\|y\|}\kappa'(u) = \frac{1}{\|x\|\|y\|}\left(y - \langle x, y\rangle\frac{x}{\|x\|^2}\right)\kappa'(u), \tag{111}$$

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

$$\pi \cdot \kappa'(u) = \frac{u}{\sqrt{1 - u^2}} + 2(\pi - \arccos u). \tag{112}$$

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:

$$\begin{aligned}\frac{\pi}{\sqrt{2}} \cdot \frac{\partial k_0(x, y)}{\partial x} &\sim_{x\to y} \frac{1}{\|y\|^2}\left(y - \langle x, y\rangle\frac{x}{\|x\|^2}\right)\frac{\sqrt{\|x\|\|y\|}}{\sqrt{\|x\|\|y\| - \langle x, y\rangle}} \\ &\sim_{x\to y} \frac{\|x\|^2 y - \langle x, y\rangle x}{\|y\|^3\sqrt{\|x\|\|y\| - \langle x, y\rangle}} \\ &\sim_{x\to y} \frac{\|y\|^2 - \langle x, y\rangle}{\|y\|^3\sqrt{\|y\|^2 - \langle x, y\rangle}}y \xrightarrow[x\to y]{} 0,\end{aligned} \tag{113}$$

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

B.4 Integral Operator and Instance Noise

Instance noise (Kaae 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 considered 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 $\tilde{k}$.

$$\sup_{f \in L^2(\mathbb{R})} \left\{ \mathcal{L}^{\tilde{k}}_{\hat{\alpha}}(f) \triangleq \mathbb{E}_{x \sim \tilde{k} * \hat{\alpha}}\big[f(x)\big] - \mathbb{E}_{y \sim \tilde{k} * \hat{\beta}}\big[f(y)\big] - \lambda \|f\|^2_{L^2} \right\} \tag{114}$$

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

$$\nabla_f \left(\mathcal{L}^{\tilde{k}}_{\hat{\alpha}} f^{\text{IN}} - \lambda \left\| f^{\text{IN}} \right\|^2_{L^2} \right) = 0 \qquad \Leftrightarrow \qquad f^{\text{IN}} = \frac{1}{2\lambda}\Big(\tilde{k} * \hat{\alpha} - \tilde{k} * \hat{\beta} \Big). \tag{115}$$

If we now study the resolution of the optimization problem in $\mathcal{H}^{\hat{\gamma}}_k$ as in Section 4.1 with $f_0 = 0$, we find the following discriminator:

$$f_t = t\Big(\mathbb{E}_{x \sim \hat{\alpha}}\big[k(x, \cdot)\big] - \mathbb{E}_{y \sim \hat{\beta}}\big[k(y, \cdot)\big] \Big) = t\Big(\tilde{k} * \hat{\alpha} - \tilde{k} * \hat{\beta} \Big). \tag{116}$$

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 definite property equates for finite datasets to the invertibility of the mapping

$$\begin{aligned} \mathcal{T}_{k,\hat{\gamma}}\big|_{\operatorname{supp}\hat{\gamma}} \colon L^2(\hat{\gamma}) &\to L^2(\hat{\gamma}) \\ h &\mapsto \mathcal{T}_{k,\hat{\gamma}}(h)\big|_{\operatorname{supp}\hat{\gamma}} \end{aligned}, \tag{117}$$

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:

$$f = \mathcal{T}_{k,\hat{\gamma}} \circ \mathcal{T}_{k,\hat{\gamma}}\big|^{-1}_{\operatorname{supp}\hat{\gamma}} \Big(f|_{\operatorname{supp}\hat{\gamma}} \Big), \tag{118}$$

as shown in Lemma 7. Finally, as shown by Jacot et al. (2018) and in Appendix A.4, 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 asumptions 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 kind of networks, NTKs are positive definite for any training data.

Indeed, besides global convergence results (Allen-Zhu et al., 2019), prior work indicate that MLPs are universal approximators (Hornik et al., 1989; Leshno et al., 1993). This property can be linked in our

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	—	$(1.53 \pm 1.08) \cdot 10^{-2}$	$(1.64 \pm 0.19) \cdot 10^{-1}$	$(5.88 \pm 0.80) \cdot 10^{-2}$

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 4.1. 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):

$$k(x, y) = \|x\|\|y\|\kappa\left(\frac{\langle x, y\rangle}{\|x\|\|y\|}\right), \tag{119}$$

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 $\nu = \frac{1}{2}\left(\delta_x + \delta_y\right)$ with δ_z being the Dirac distribution centered on $z \in \Omega$, and where $x = 0$ and $y \in \Omega \setminus \{0\}$. Then:

$$\mathbb{E}_{z\sim\mu}k(z, \cdot) = k\left(\frac{1}{2}y, \cdot\right) = \frac{1}{2}k(y, \cdot) = \frac{1}{2}\left(k(y, \cdot) + k(x, \cdot)\right) = \mathbb{E}_{z\sim\nu}k(z, \cdot), \tag{120}$$

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

C GAN(TK)^2 and Further Empirical Analysis

We present in this section additional experimental results that complement and explain some of the results already exposed in Section 5. 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 focused 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 Other Two-Dimensional Datasets

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 5 on this datasets with more complex shapes.

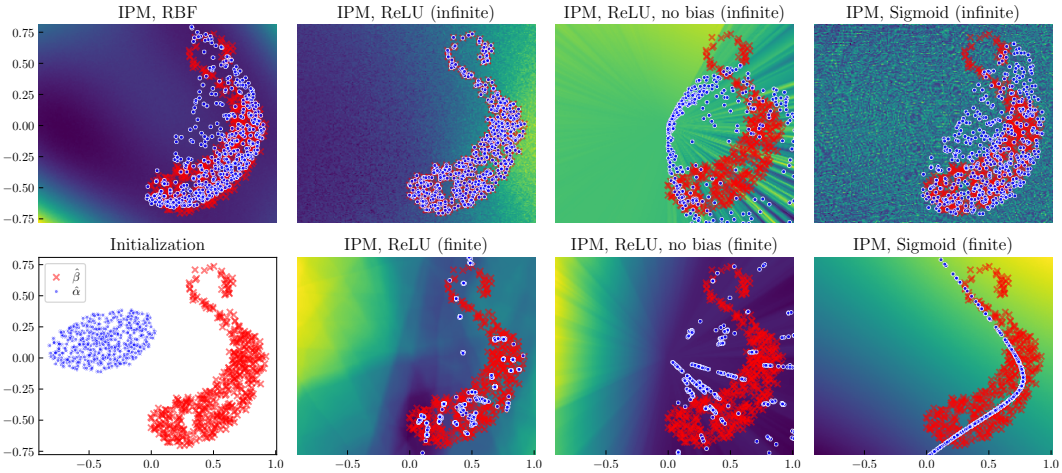

Figure 3: Generator (●) and target (×) samples for different methods applied to the Density problem. In the background, $c_{f^\star}$.

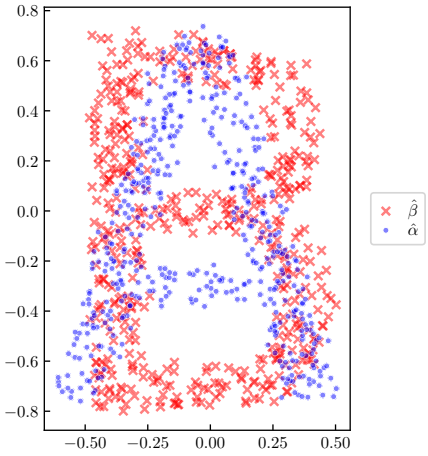

Figure 4: Initial generator (●) and target (×) samples for the AB problem.

Table 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	—	$(1.32 \pm 1.30) \cdot 10^{-1}$	$(2.57 \pm 0.73) \cdot 10^{-2}$	$(8.78 \pm 2.23) \cdot 10^{-2}$

Table 4: Sinkhorn divergence 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	—	$(3.07 \pm 0.68) \cdot 10^{0}$	$(7.52 \pm 0.01) \cdot 10^{0}$	$(7.41 \pm 0.54) \cdot 10^{0}$

(a) RBF kernel: blurry digits.

(b) ReLU: sharp digits.

(c) ReLU (no bias): mostly sharp digits with some artifacts and blurry images.

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

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. Results are given in Tables 2 to 4 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 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 images (LeCun et al., 1998) in a similar setting as the experiments on two-dimensional point clouds of the previous sections. 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 datasets, 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 Section 4 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 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:

$$f^\star_{\hat\alpha} = \sum_{x \in \operatorname{supp}\hat\gamma} h_1(x) k(x, \cdot). \tag{121}$$

Consequently,

$$\nabla f^\star_{\hat\alpha} = \sum_{x \in \operatorname{supp}\hat\gamma} h_1(x) \nabla k(x, \cdot), \qquad \nabla c_{f^\star_{\hat\alpha}} = \sum_{x \in \operatorname{supp}\hat\gamma} h_1(x) \nabla k(x, \cdot) c'\big(f^\star_{\hat\alpha}(\cdot)\big). \tag{122}$$

Dirac-GAN Setting. The latter linear combination of gradients indicates that, by examining gradients of $c_{f^\star_{\hat\alpha}}$ for pairs of $(x, y) \in (\operatorname{supp}\hat\alpha) \times \left(\operatorname{supp}\hat\beta\right)$, 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 $\hat\alpha = \delta_x$ 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 5 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 ourselves to the case where the second coordinate is positive.

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 $\operatorname{Span}\{x, y\}$, independently of the dimensionality of x, y. This means that when training the generated distribution, as in Section 5, the position of the particle x during training 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. Note 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 target.

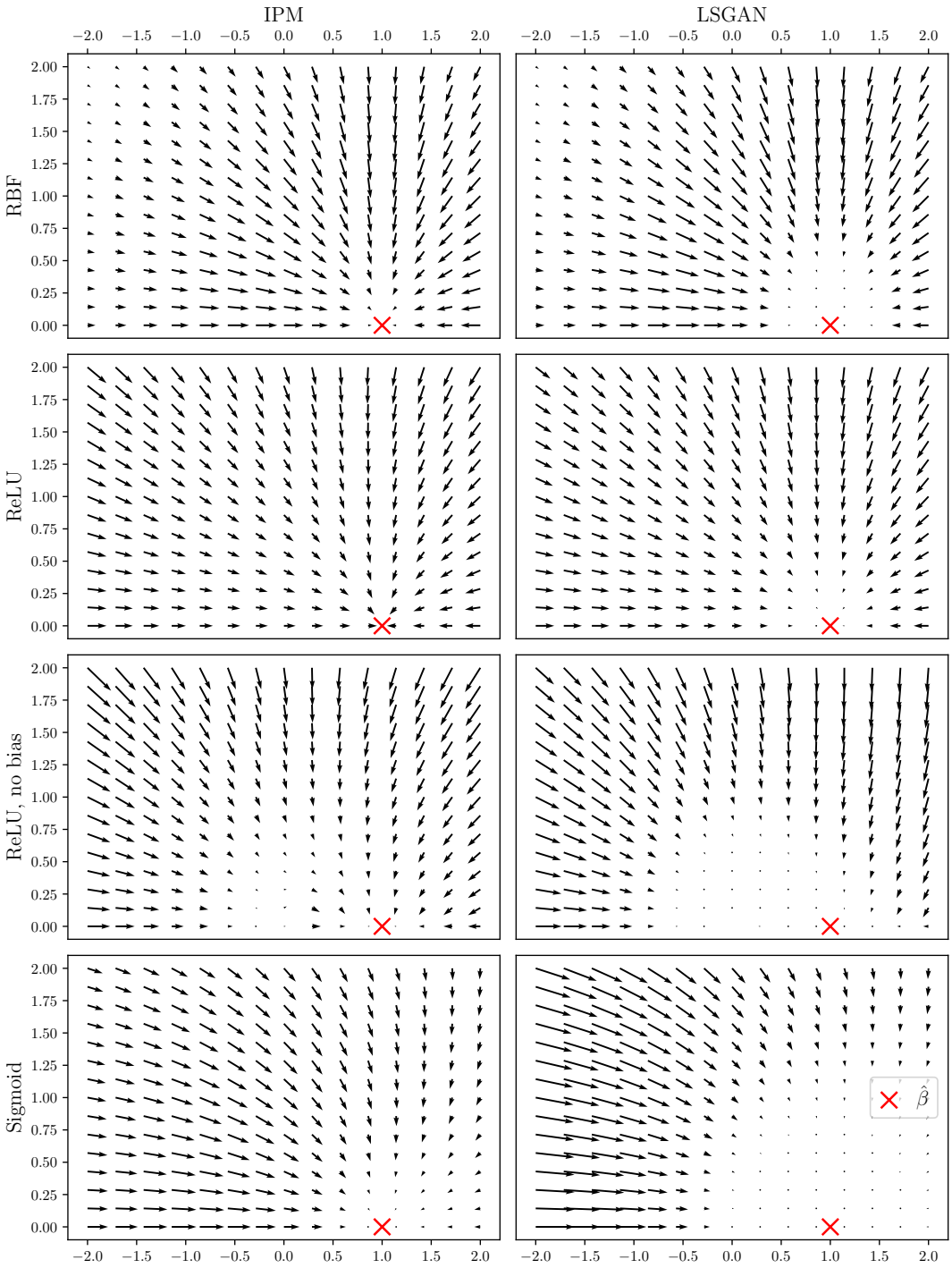

Figure 6: Gradient field $\nabla c_{f^\star_{\hat{\alpha}_x}}(x)$ received by a generated sample $x \in \mathbb{R}^2$ (i.e. $\hat{\alpha} = \hat{\alpha}_x = \delta_x$) initialized to x_0 with respect its coordinates in $\mathrm{Span}\{x_0, y\}$ where y, marked by a $\times$, is the target distribution (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.

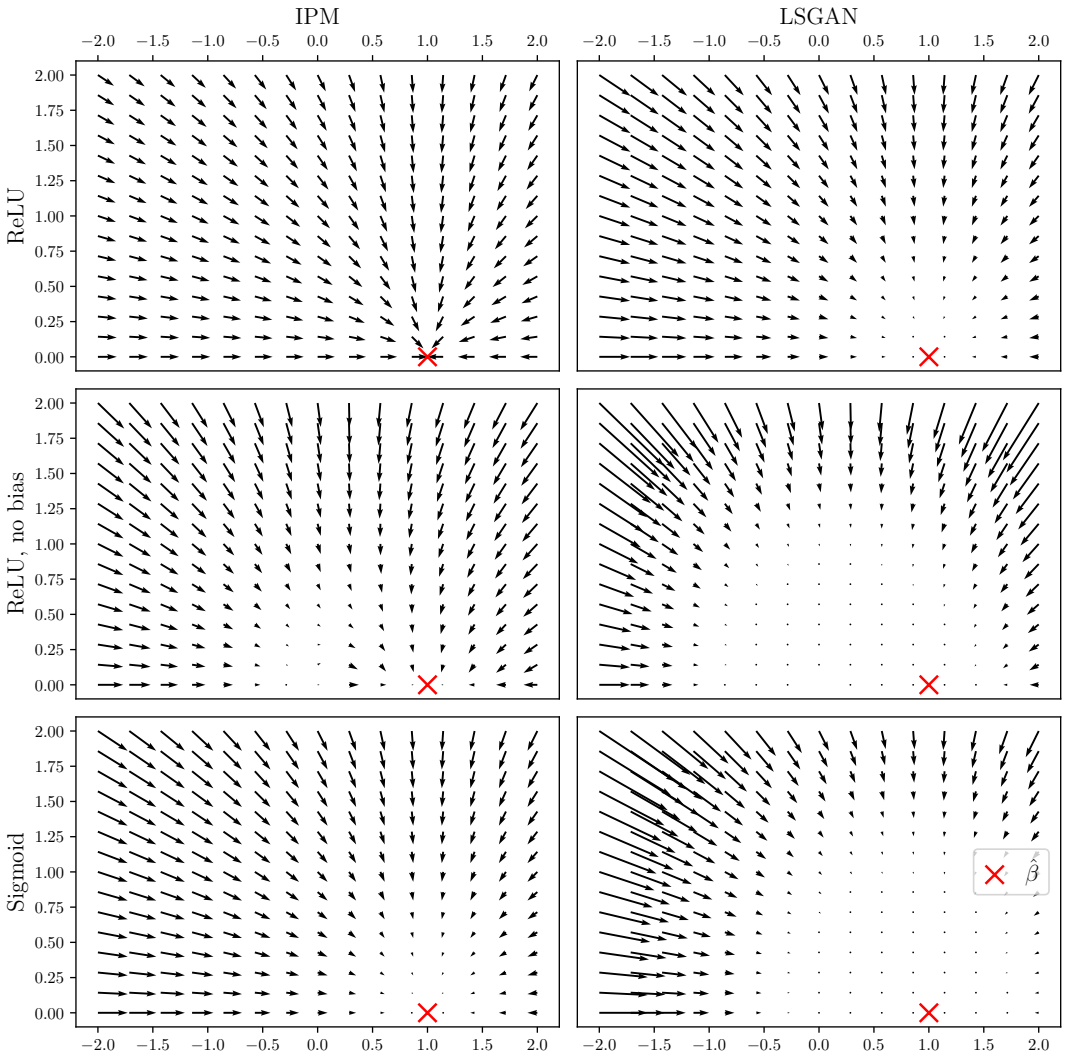

Figure 7: Same plot as Figure 6 but with underlying points $x, y \in \mathbb{R}^{512}$.

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 2 to 4). 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 3.2, 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 in 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 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 Python Python 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. Sinkhorn divergences between cloud points are computed using Geomloss (Feydy et al., 2019). 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 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)[2] and are licensed 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. MNIST (LeCun et al., 1998) is a standard dataset containing white digits over a dark frame, with no known license to the best of our knowledge.[3] We preprocess each MNIST image by extending it from 28×28 frames to 32×32 frames (by padding it with black pixels) and normalizing pixels in the $[-1, 1]$ range. For our experiments, we consider a subset of 1024 elements of MNIST, 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:

$$k(x, y) = \mathrm{e}^{\frac{\|x-y\|^2}{2n}}, \tag{123}$$

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

[2] They can be found at `https://github.com/jeanfeydy/geomloss/tree/master/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).

[3] No license is provided on the official webpage: `http://yann.lecun.com/exdb/mnist/`.

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). 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 antisymmetric initialization (Zhang et al., 2020) when using the IPM 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 4 with training time $\tau = 1$ (except for MNIST where $\tau = 1000$).

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

We multliply 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: $50\,000$.