DEMYSTIFYING MMD GANS

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the training and performance of generative adversarial networks using the Maximum Mean Discrepancy (MMD) as critic, termed MMD GANs. As our main theoretical contribution, we show that the natural estimator for maximum mean discrepancies yields unbiased gradient estimates, which is essential in providing a good signal to the generator. We discuss the issue of kernel choice for the MMD critic, and characterize the kernel corresponding to the energy distance used for the Cramér GAN critic. Being an intergal probability metric, the MMD benefits from training strategies recently developed for Wasserstein GANs. In experiments, the MMD GAN is able to employ a smaller critic network than the Wasserstein GAN, resulting in a simpler and faster-training algorithm with matching performance.

1 INTRODUCTION

Generative Adversarial Networks (GANs; Goodfellow et al., 2014) provide a powerful method for general-purpose generative modeling of datasets. Given examples from some distribution, a GAN attempts to learn a *generator* function, which maps from some fixed noise distribution to samples that attempt to mimic a reference or target distribution. The generator is trained to trick a *discriminator*, or *critic*, which tries to distinguish between generated and target samples.

This alternative to standard maximum likelihood approaches for training generative models has brought about a rush of interest over the past several years. Likelihoods do not necessarily correspond well to sample quality (Theis et al., 2016), and GAN-type objectives focus much more on producing plausible samples, as illustrated particularly directly by Danihelka et al. (2017). This class of models has recently led to many impressive examples of image generation (e.g. Huang et al., 2017a;b; Jin et al., 2017; Zhu et al., 2017).

GANs are, however, notoriously tricky to train (Salimans et al., 2016). This might be understood in terms of the discriminator class: Goodfellow et al. (2014) showed that, when the discriminator is trained to optimality among a rich enough function class, the generator network attempts to minimize the Jensen-Shannon divergence between the generator and target distributions. This result has been extended to general f-divergences by Nowozin et al. (2016). According to Arjovsky & Bottou (2017), however, it is likely that both the GAN and model probability measures are supported on manifolds within a larger space, as occurs for the set of images in the space of possible pixel values. These manifolds might not intersect at all, or at best intersect on sets of measure zero. In this case, the Jensen-Shannon divergence is constant, and the KL and reverse-KL divergences are infinite, meaning they provide no useful gradient for the generator to follow. This helps to explain some of the instability of GAN training.

The lack of sensitivity to distance, meaning that nearby but non-overlapping regions of high probability mass are not considered similar, is a long-recognised problem for KL divergence-based discrepancy measures (e.g. Gneiting & Raftery, 2007, Section 4.2). It is natural to address this problem using Integral Probability Metrics (IPMs; Müller, 1997): these measure the distance between probaility measures via the largest discrepancy in expectation over a class of "well behaved" witness functions. Thus, IPMs are able to signal proximity in the probability mass of the generator and reference distributions.

Arjovsky et al. (2017) proposed to use the Wasserstein distance between distributions as the discriminator, which is an integral probability metric constructed from the witness class of 1-Lipschitz functions. To implement the Wasserstein critic, weight clipping of the distriminator network was

initially used to enforce k-Lipschitz smoothness of the discriminator. Gulrajani et al. (2017) improved on this result by directly constraining the gradient of the discriminator network at points between the generator and reference samples. This new Wasserstein GAN implementation, called WGAN-GP, is more stable and easier to train.

A second integral probability metric used in GANs is the maximum mean discrepancy (MMD), for which the witness function class is a unit ball in a reproducing kernel Hilbert space (RKHS). Generative adversarial models based on minimizing the MMD were first considered by Li et al. (2015) and Dziugaite et al. (2015). These works optimized a generator to minimize the MMD with a fixed kernel, either using a generic kernel on image pixels or by modeling autoencoder representations instead of images directly. Sutherland et al. (2017) instead minimized the statistical power of an MMD-based test with a fixed kernel. Such approaches struggle with complex natural images, where pixel distances are of little value, and fixed representations can easily be tricked, as in the adversarial examples of Szegedy et al. (2014).

Adversarial training of the MMD loss is thus the obvious choice. Here the kernel MMD is defined on the output of a convolutional network, which is trained adversarially. Recent notable work has made use of the IPM representation of the MMD to employ the same witness function regularisation strategies as Arjovsky et al. (2017); Gulrajani et al. (2017), effectively corresponding to an additional constraint on the MMD function class. Without such constraints, the convolutional features are unstable and difficult to train (Sutherland et al., 2017). Li et al. (2017b) used the weight clipping strategy of Arjovsky et al., with additional constraints to ensure that the kernel distribution embeddings remained injective. In light of the observations by Gulrajani et al., however, we use a gradient constraint on the MMD witness function in the present work (see Sections 2.1 and 2.2). Bellemare et al. (2017)'s method, the Cramér GAN, also used the gradient constraint strategy of Gulrajani et al. in their discriminator network. As we discuss in Section 2.3, the Cramér GAN discriminator is related to the energy distance, which is an instance of the MMD (Sejdinovic et al., 2013), and which can therefore use a gradient constraint on the witness function. Note, however, that there are important differences between the Cramér GAN critic and the energy distance, which make it more akin to the optimization of a scoring rule: we provide further details in Appendix A. Weight clipping and gradient constraints are not the only approaches possible: variance features (Mroueh et al., 2017) and constraints (Mroueh & Sercu, 2017) can work, as can other optimization strategies (Berthelot et al., 2017; Li et al., 2017a).

Given that both the Wasserstein distance and the MMD are integral probability metrics, it is of interest to consider how they differ when used in GAN training. Bellemare et al. (2017) pointed out a particular flaw of the Wasserstein distance as an objective: that gradients of both the duality-based estimator and of the alternative estimator of Genevay et al. (2017) are biased. Optimization over biased gradients can lead to incorrect parameter values, even in expectation; Bellemare et al. provide an explicit example where this occurs. As our main theoretical contribution, we show in Section 3 that the natural maximum mean discrepancy estimator has unbiased gradients, and therefore does not suffer from this problem.

There are additional empirical advantages of the MMD-GAN over the Wasserstein GAN. Certainly we would not expect the MMD on its own to perform well on raw image data, since these data lie on a low dimensional manifold embedded in a much higher dimensional pixel space. Once the images are mapped through appropriately trained convolutional layers, however, they can follow a much simpler distribution with broader support across the mapped domain: a phenomenon also observed in autoencoders (Bengio et al., 2013). In this setting, the MMD with characteristic kernels (Sriperumbudur et al., 2010) shows strong discriminative peformance between distributions. To achieve comparable performance, a purely convolutional WGAN requires many more layers in the critic. In our experiments (Section 5), we find that MMD GANs achieve the same generator performance as WGAN-GPs with smaller discriminator networks, resulting in GANs with fewer parameters and easier training. Thus, the MMD-GAN discriminator can be understood as a hybrid model that plays to the strengths of both the initial convolutional mappings and the kernel layer that sits on top.

2 LOSSES AND WITNESS FUNCTIONS

We begin with a review of the MMD and relate it to the loss functions used by other GAN variants. Through its interpretation as an integral probability metric, we show that the gradient penalty of Gulrajani et al. (2017) applies to the MMD GAN.

2.1 MAXIMUM MEAN DISCREPANCY AND WITNESS FUNCTIONS

We consider a random variable X with probability measure P, which we associate with the generator, and a second random variable Y with probability measure Q, which we associate with the reference sample that we wish to learn. Our goal is to measure the distance from P to Q using samples drawn independently from each distribution.

The maximum mean discrepancy is a metric on probability measures (Gretton et al., 2012), which falls within the family of integral probability metrics (Müller, 1997); this family includes the Wasserstein and Kolmogorov metrics, but not for instance the KL or χ^2 divergences. Integral probability metrics make use of a class of *witness functions* to distinguish between P and Q, choosing the function with the largest discrepancy in expectation over P, Q,

$$D_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P\, f(X) - \mathbb{E}_Q\, f(Y). \tag{1}$$

The particular witness function class $\mathcal{F}$ determines the probability metric: the Wasserstein-1 metric is defined using the 1-Lipschitz functions, and the Kolmogorov metric from the functions of bounded variation 1. The loss functions of many other GAN variants also fall into this class (e.g. Mroueh et al., 2017; Mroueh & Sercu, 2017; Berthelot et al., 2017).

In this work, our witness function class $\mathcal{F}$ will be the unit ball in a reproducing kernel Hilbert space $\mathcal{H}$, with positive definite kernel $k(x, x')$. The key property of a reproducing kernel Hilbert space is the reproducing property: for all $f \in \mathcal{H}$, $f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}$. We define the *mean embedding* of the probability measure P as the element $\mu_P \in \mathcal{H}$ such that $\mathbb{E}_P\, f(X) = \langle f, \mu_P\rangle_{\mathcal{H}}$; it is given by $\mu_P = \mathbb{E}_{X\sim P}\, k(\cdot, X)$.[1]

The maximum mean discrepancy (MMD) is defined as (1) with $\mathcal{F}$ the unit ball in $\mathcal{H}$:

$$MMD(P, Q; \mathcal{H}) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \le 1} \mathbb{E}_P\, f(X) - \mathbb{E}_Q\, f(Y).$$

The witness function f^* that attains the supremum has a straightforward expression (Gretton et al., 2012, Section 2.3):

$$f^*(x) \propto \mathbb{E}_P\, k(x, X) - \mathbb{E}_Q\, k(x, Y). \tag{2}$$

Given samples $X = \{x_i\}_{i=1}^m$ drawn i.i.d. from P, and $Y = \{y_j\}_{j=1}^n$ drawn i.i.d. from Q, the empirical witness function is

$$\hat{f}(x) \propto \frac{1}{m}\sum_{i=1}^m k(x_i, x) - \frac{1}{n}\sum_{i=1}^n k(y_i, x), \tag{3}$$

and an unbiased estimator of the squared MMD is (Gretton et al., 2012, Lemma 6)

$$\frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}^m k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(x_i, y_j). \tag{4}$$

When the kernel is characteristic (Sriperumbudur et al., 2010; 2011), the embedding μ_P is injective (i.e., associated uniquely with P). Perhaps the best-known characteristic kernel is the exponentiated quadratic, also known as the Gaussian RBF, kernel,

$$k_\sigma^{rbf}(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right). \tag{5}$$

[1] This is well defined for Bochner-integrable kernels (Steinwart & Christmann, 2008, Definition A.5.20), for which $E_P\, \|k(x, \cdot)\|_{\mathcal{H}} < \infty$ for the class of probability measures P being considered. For bounded kernels, the condition always holds, but for unbounded kernels, additional conditions on the moments might apply.

Both the kernel and its derivatives decay exponentially, however, causing significant problems in high dimensions, and especially when used in gradient-based representation learning. The rational quadratic kernel

$$k_\alpha^{rq}(x, y) = \left(1 + \frac{\|x - y\|^2}{2\alpha}\right)^{-\alpha} \tag{6}$$

with $\alpha > 0$ corresponds to a scaled mixture of exponentiated quadratic kernels, with a $\mathrm{Gamma}(\alpha, 1)$ prior on the inverse lengthscale (Rasmussen & Williams, 2006, Section 4.2). This kernel will be the mainstay of our experiments, as its tail behaviour is much superior to that of the exponentiated quadratic kernel; it is also characteristic.

2.2 Witness function and gradient penalties

The MMD has been a popular choice for the role of a critic in a GAN. This idea was proposed simultaneously by Dziugaite et al. (2015) and Li et al. (2015), with numerous recent follow-up works (Sutherland et al., 2017; Liu, 2017; Li et al., 2017b; Bellemare et al., 2017). As a key strategy in these recent works, the MMD of (4) is not computed directly on the samples; rather, the samples first pass through a mapping function h, generally a convolutional network. Note that we can think of this either as the MMD with kernel k on features $h(x)$, or simply as the MMD with kernel $\kappa(x, y) = k(h(x), h(y))$. The challenge is to learn the features h so as to maximize the MMD, without causing the critic to collapse to a trivial answer early in training.

Bearing in mind that the MMD is an integral probability metric, strategies developed for training the Wasserstein GAN critic can be directly adopted for training the MMD critic. Li et al. (2017b) employed the weight clipping approach of Arjovsky et al. (2017), though they motivated it from different considerations. Gulrajani et al. (2017) found a number of issues with weight clipping, however: it oversimplifies the loss functions given standard architectures, the gradient decays exponentially as we move up the network, and it seems to require the use of slower optimizers such as RMSProp rather than standard approaches such as Adam (Kingma & Ba, 2015).

It thus seems preferable to adopt Gulrajani et al.'s proposal of regularising the critic witness (3) by constraining its gradient norm to be nearly 1 along randomly chosen convex combinations of generator and reference points, $\alpha x_i + (1 - \alpha) y_j$ for $\alpha \sim \mathrm{Uniform}(0, 1)$. This was motivated by the observation that the Wasserstein witness satisfies this property (their Lemma 1), but perhaps its main benefit is one of regularization: if the critic function becomes too flat anywhere between the samples, the generator cannot easily follow its gradient. We will thus follow this approach, as did Bellemare et al. (2017), whose model we describe next.

2.3 The energy distance and associated MMD

Liu (2017) and Bellemare et al. (2017, Section 4) proposed to use the energy distance as the critic in an adversarial network. The energy distance (Székely & Rizzo, 2004; Lyons, 2013) is a measure of divergence between two probability measures, defined as

$$D_e(P, Q) = -\frac{1}{2}\,\mathbb{E}_P\,\rho(X, X') - \frac{1}{2}\,\mathbb{E}_Q\,\rho(Y, Y') + \mathbb{E}_{PQ}\,\rho(X, Y), \tag{7}$$

where $\mathbb{E}_P\,\rho(X, X')$ is an expectation over two independent samples from the generator P (likewise, Y and Y' are independent samples from the reference Q), and $\rho(x, y)$ is a semimetric of negative type.[2] We will focus on the case $\rho(x_i, x_j) = \|X - X'\|^\beta$ for $0 < \beta \leq 2$. When $\beta = 1$, ρ is a metric.

It has been shown (Sejdinovic et al., 2013, Lemma 12) that the energy distance is an instance of the maximum mean discrepancy, where the corresponding distance-induced kernel family for the distance (7) is

$$k_{z_0}^{dist}(x, y) = \frac{1}{2}\left[\rho(x, z_0) + \rho(y, z_0) - \rho(x, y)\right] \tag{8}$$

for any choice of $z_0 \in \mathcal{X}$ (often $z_0 = 0$ is chosen to simplify notation, which corresponds to a fractional Brownian motion kernel; Sejdinovic et al., 2013, Example 15). Note that the resulting

[2] ρ must satisfy the properties of a metric besides the triangle inequality, and for all $n \geq 2$, $x_1, \ldots, x_n \in \mathcal{X}$, and $a_1, \ldots, a_n \in \mathbb{R}$ with $\sum_i a_i = 0$, $\sum_{i=1}^n \sum_{j=1}^n a_i a_j \rho(x_i, x_j) \leq 0$.

kernel is *not* translation invariant. It is characteristic (when $\beta < 2$), and the MMD is well defined, for a class of distributions $\mathcal{P}$ that satisfy a moment condition (Sejdinovic et al., 2013, Remark 21), namely that there exists $z_0 \in \mathcal{X}$ such that

$$\int \rho(z, z_0) \, \mathrm{d}P(z) < \infty \quad \forall P \in \mathcal{P}.$$

To apply the regularization strategy of Gulrajani et al. (2017) in training the critic of an adversarial network, we need to compute the form taken by the witness function (2) given the kernel (8). Bearing in mind that

$$\mathbb{E}_P \, k(x, X) = \rho(x, z_0) + \mathbb{E}_P \, \rho(X, z_0) - \mathbb{E}_P \, \rho(x, X),$$

where the second term of the above expression is constant, and substituting into (2), we have

$$\begin{aligned} f^*(x) &\propto \rho(x, z_0) - \mathbb{E}_X \, \rho(x, X) - d(x, z_0) + E_Y \rho(x, Y) + C \\ &= \mathbb{E}_Q \, \rho(x, Y) - \mathbb{E}_P \, \rho(x, X) + C. \end{aligned}$$

This is in agreement with Bellemare et al.'s function f^* (their page 5), albeit via a different argument (note that the function g^* in their footnote 4 differs from the above, and appears to be in error).

We now turn to the divergence implemented by the critic in Bellemare et al.'s Algorithm 1, which is somewhat different from the energy distance (7). The Cramér GAN witness function is defined as

$$f_c(x) = \mathbb{E}_P \, \rho(x, X) - \rho(x, 0), \tag{9}$$

which is regularized using Gulrajani et al.'s gradient constraint. The expected *surrogate loss* associated with this witness function, and used for the Cramér critic, is

$$D_c(P, Q) = \mathbb{E}_P \, \rho(X, X') + \mathbb{E}_Q \, \rho(Y, 0) - \mathbb{E}_P \, \rho(X, 0) - \mathbb{E}_{P,Q} \, \rho(X', Y). \tag{10}$$

In brief, Y' in (7) is replaced by the origin: it is explained that this is necessary for instance in the conditional case, where two independent reference samples Y, Y' are not available. Following this change, it becomes straightforward to define P and Q which are different, yet have an expected $D_c(P, Q)$ loss of zero. For example, if P is a point mass at the origin in $\mathbb{R}$, and Q is a point mass a distance t from the origin, then $P \neq Q$, and yet $D_c(P, Q) = 0$ because

$$\mathbb{E}_P \, \rho(X, X') = \mathbb{E}_P \, \rho(X, 0) = 0, \qquad \mathbb{E}_{P,Q} \, \rho(X', Y) = \mathbb{E}_Q \, \rho(Y, 0) = t.$$

Nevertheless, good empirical performance has been obtained in practice for the Cramér critic, both by Bellemare et al. (2017) and in our Section 5 experiments. Insight into this behaviour can be obtained by considering its relation to the score function associated with the energy distance. We describe this relation in Appendix A.

3 UNBIASED GRADIENTS

One of the main criticisms of the Wasserstein GAN is that its gradient is biased (Bellemare et al., 2017, Section 3), which can cause problems when performing stochastic gradient descent. By contrast, Bellemare et al. claim that the energy distance used in the Cramér GAN critic does not suffer from biased gradients. They do not, however, properly prove this claim: the essential step in the reasoning – the exchange in the order of the expectations and the derivatives – is simply assumed.[3]

We prove here that under reasonable assumptions, the expectation and gradient can be interchanged for kernels on representations given by deep networks, and hence gradients of (4) are unbiased estimates of the gradient of the true MMD.

Theorem 1. *Let $\mathbb{P}$, $\mathbb{Q}$ be distributions with moments of order $\alpha \geq 1$ (see Assumption **A**), and let k be a continuously differentiable kernel subject to some growth conditions (Assumption **B**). Let $\phi_\theta(X)$ be a feedforward neural network with parameters θ as in Definition 2; in particular, convolutional networks with ReLU-like activations are covered.*

The function $\theta \mapsto \mathbb{E}_{\mathbb{P},\mathbb{Q}}[k(\phi_\theta(X), \phi_\theta(Y))]$ is well-defined and differentiable for μ-almost all $\theta \in \mathbb{R}^m$, where μ is the Lebesgue measure. Moreover, the following formula holds:

$$\partial_\theta \, \mathbb{E}_{\mathbb{P},\mathbb{Q}}[k(\phi_\theta(X), \phi_\theta(Y))] = \mathbb{E}_{\mathbb{P},\mathbb{Q}}[\partial_\theta k(\phi_\theta(X), \phi_\theta(Y))].$$

[3] If $\hat{f}(\theta)$ is an unbiased estimator of a function $f(\theta)$, so that $\mathbb{E} \, \hat{f}(\theta) = f(\theta)$, then if we can exchange expectations and gradients, it is immediate that $\mathbb{E} \, \nabla \hat{f}(\theta) = \nabla \, \mathbb{E} \, \hat{f}(\theta) = \nabla f(\theta)$.

The proof is given in Appendix B. The main idea is to first prove the result for differentiable activation functions, using the standard dominated convergence theorem. We then consider a sequence of differentiable activation functions that converge to the activation function of interest, which is differentiable except at the origin: for example the ReLU activation. The key point is to first show that the set of parameters under which the expected gradient is undefined has zero Lebesgue measure, regardless of the probability distributions P and Q. We then prove that the expected gradients using the differentiable activations converge locally uniformly towards the expected gradient obtained using the singular activation function. We then conclude using standard results of gradients of uniformly convergent sequences of functions.

4 EVALUATION METRICS

One problem with comparing GAN models is that quantitative comparisons are difficult. Some insight can be gained by visually examining samples, but we also consider the following approaches to evaluate GAN methods.

Inception score This metric, proposed by Salimans et al. (2016), is based on the classification output $p(y \mid x)$ of the Inception model (Szegedy et al., 2016). Defined as $\exp\left(\mathbb{E}_x \operatorname{KL}(p(y \mid x) \parallel p(y))\right)$, it is highest when each image's predictive distribution has low entropy, but the marginal $p(y) = \mathbb{E}_x\, p(y \mid x)$ has high entropy. This score correlates somewhat with human judgement of sample quality on natural images, but it has some issues, particularly when applied to domains which do not represent a variety of the types of classes in ImageNet. In particular, it knows nothing about the desired distribution for the model.

FID The *Fréchet Inception Distance*, proposed by Heusel et al. (2017), avoids some of the problems of Inception by measuring the similarity of the samples' representations in the Inception architecture (at the `pool3` layer) to those of samples from the target distribution. The FID fits a Gaussian distribution to the hidden activations for each distribution and then computes the Fréchet distance, also known as the Wasserstein-2 distance, between those Gaussians. Heusel et al. show that unlike the Inception score, the FID worsens monotonically as various types of artifacts are added to images. Note, however, that the estimator of FID is biased.[4]

KID We propose a metric similar to the FID, the *Kernel Inception Distance*, to be the squared MMD between Inception representations. We use a polynomial kernel, $k(x, y) = \left(\frac{1}{d} x^\mathsf{T} y + 1\right)^3$ where d is the representation dimension, to avoid correlations with the objective of MMD GANs as well as to avoid tuning any kernel parameters.[5] Compared to the FID, the KID has several advantages. First, it does not assume a parametric form for the distribution of activations; this is particularly sensible since the representations have ReLU activations, and therefore are not only never negative, but do not even have a density: about 2% of components in Inception representations are typically exactly zero. With the cubic kernel we use here, the KID compares skewness as well as the mean and variance. Also, unlike the FID, the KID has a simple unbiased estimator.

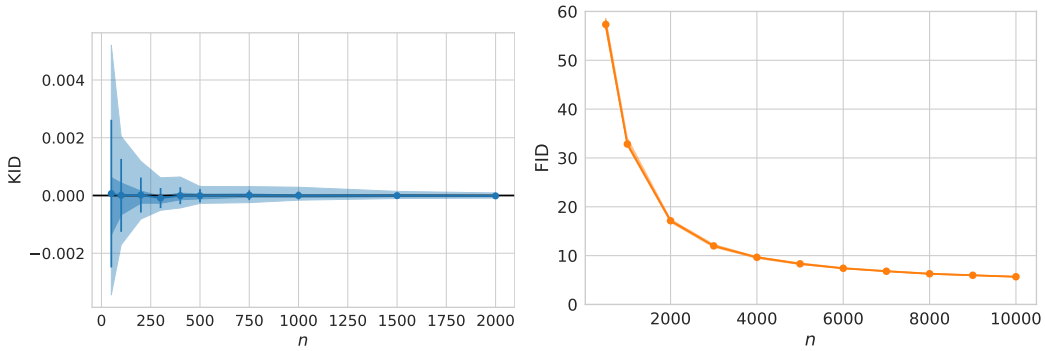
Figure 1 demonstrates the bias of the FID and the unbiasedness of the KID by comparing the CIFAR-10 train and test sets. The KID (Figure 1a) converges quickly to its presumed true value of 0; even for very small n, simple Monte Carlo estimates of the variance provide a reasonable measure of uncertainty. By contrast, the FID estimate (Figure 1b) does not behave so nicely: at $n = 2\,000$, when the KID estimator is essentially always 0, the FID estimator is still quite large. Even at $n = 10\,000$, the full size of the CIFAR test set, the FID still seems to be decreasing from its estimate of about 5.7 towards zero, showing the strong persistence of bias. More importantly, Monte Carlo estimates of the variance are extremely small even when the estimate is very far from its asymptote, so it is difficult to judge the reliability of an estimate.

For models on MNIST, we replace the Inception featurization with features from a LeNet-like convolutional classifier[6] (LeCun et al., 1998), but otherwise compute the scores in the same way.

[4] This is easily seen when the true FID is 0: here the estimator may be positive, but can never be negative.

[5] k is the default polynomial kernel in `scikit-learn` (Pedregosa et al., 2011).

[6] `github.com/tensorflow/models/blob/master/tutorials/image/mnist/convolutional.py`



(a) KID estimate based on sampling both train and test sets to size n without replacement, 100 times.

(b) FID estimate based on a with-replacement resampling of the test set to size n and the full train set. 10 samples; all standard deviations are less than 1.

Figure 1: Estimates of distances between the CIFAR-10 train and test sets. Lines show means, error bars standard deviations, dark colored regions a $\frac{2}{3}$ coverage interval of the samples, light colored regions a 95% interval. Note the differing n axes.

We also considered the diagnostic test of Arora & Zhang (2017), which estimates the approximate number of "distinct" images produced by a GAN. The amount of subjectivity in what constitutes a duplicate image, however, makes it hard to reliably compare models based on this diagnostic. Comparisons probably need to be performed both with a certain notion of duplication in mind and by a user who does not know which models are being compared, to avoid subconscious biases; we leave further exploration of this intriguing procedure to future work.

5 Experiments

We compare the quality of samples generated by MMD-GAN with various kernels with the ones obtained by WGAN-GP (Gulrajani et al., 2017) and Cramér GAN (Bellemare et al., 2017) on two standard benchmark datasets: the MNIST dataset of 28×28 handwritten digits[7] and the CIFAR-10 dataset of 32×32 photos (Krizhevsky, 2009).

For all of our experiments we used the DCGAN architecture (Radford et al., 2016) for both generator and critic. For MMD losses, we used only 16 top-layer neurons in the discriminator; more did not seem to improve performance. For the generator we used the standard number of convolutional filters (64 in the second-to-last layer); for the critic, we compared networks with 16 and 64 filters in the first convolutional later.

Networks with smaller critic run considerably faster: on our systems, the 16-filter networks typically ran at about twice the speed of the 64-filter ones. Note that the critic size is far more important to training runtime than the generator size: we update the critic 5 times for each generator step, and moreover the critic network is run on two batches each time we use it, one from P and one from Q. Given the same architecture, all models considered here run at about the same speed.

We evaluate several MMD GAN kernel functions in our experiments. The simplest is the linear kernel: $k^{dot}(x, y) = \langle x, y \rangle$, whose MMD corresponds to the distance between means (this is somewhat similar to the *feature matching* idea of Salimans et al., 2016). We also use the exponentiated quadratic (5) and rational quadratic (6) functions, with mixtures of lengthscales,

$$k^{rbf}(x, y) = \sum_{\sigma \in \Sigma} k^{rbf}_{\sigma}(x, y), \qquad k^{rq}(x, y) = \sum_{\alpha \in \mathcal{A}} k^{rq}_{\alpha}(x, y),$$

where $\Sigma = \{2, 5, 10, 20, 40, 80\}$, $\mathcal{A} = \{.2, .5, 1, 2, 5\}$. Lastly we use the distance-induced kernel k_0^{dist} of (8), using the Euclidean distance, so that the MMD is the energy distance, and the Cramér critic in (10).

[7] yann.lecun.com/exdb/mnist/

Table 1: Mean (standard deviation) of score evaluations for the MNIST models.

Model	Critic size	Inception	FID	KID
rq	16	9.11 (0.01)	4.206 (0.05)	0.005 (0.004)
rbf	16	8.98 (0.02)	8.264 (0.02)	0.011 (0.006)
dot	16	8.86 (0.02)	6.245 (0.06)	0.006 (0.004)
Cramér GAN	16	9.25 (0.02)	3.385 (0.10)	0.006 (0.005)
WGAN-GP	16	9.12 (0.02)	6.915 (0.10)	0.009 (0.004)
test set	–	9.78 (0.02)	4.305 (0.16)	0.003 (0.003)

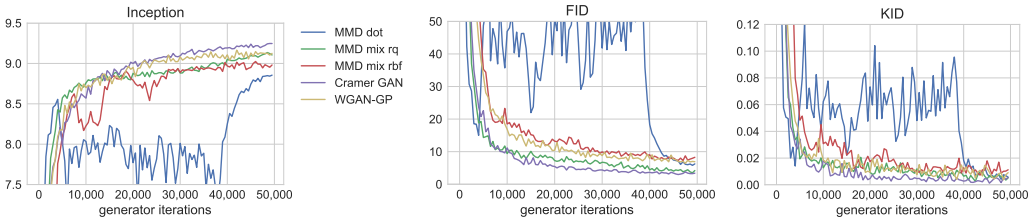


Figure 2: Score estimates over the learning process for MNIST training.

Each model was trained with a batch size of 128, and 5 discriminator updates per generator update. For CIFAR-10 we trained for $150\,000$ generator updates, while for MNIST we used $50\,000$.

MNIST After training for $50\,000$ generator iterations, all variants we tried achieved reasonable results. Table 1 shows the quantitative measures, computed on the basis of a LeNet model. All have achieved KIDs of essentially zero, and FIDs around the same as that of the test set, with Inception scores slightly lower. Figure 4, in Appendix C, shows samples.

Figure 2, however, shows the evolution of our quantitative criteria throughout the training process for several models. This shows that the linear kernel *dot* is clearly worse than the other models for most of the training process, but it makes a sharp improvement around $40\,000$ iterations and basically catches up to the other models.

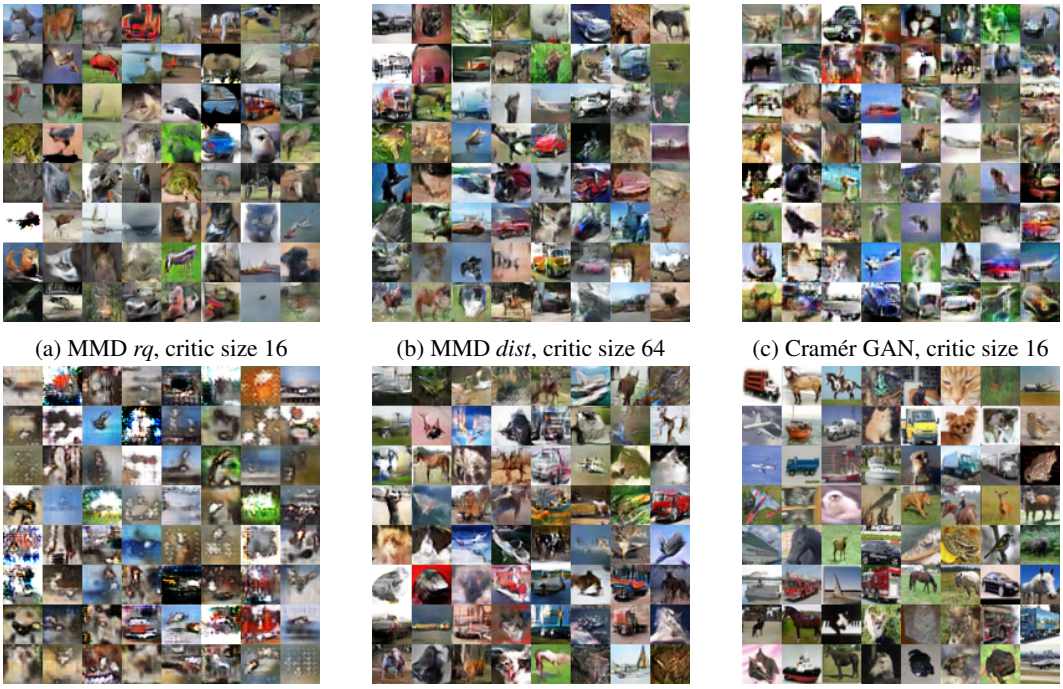
There is also some evidence that *rbf*, and perhaps WGAN-GP, converge more slowly than *rq* and Cramér GAN. Examining samples during training, we observed that *rbf* more frequently produces extremely "blurry" outputs, which can persist for a substantial amount of time before eventually resolving. This makes sense, given the very fast gradient decay of the *rbf* kernel: when generator samples are extremely far away from the reference samples, slight improvements yield very little reward for the generator, and so bad samples can stay bad for a long time. Given their otherwise similar properties, we thus recommend the use of *rq* kernels over *rbf* in MMD GANs.

CIFAR-10 Scores for various models trained on CIFAR-10 are shown in Table 2. The scores for *rq* with a small critic network approximately match those of WGAN-GP with a large critic network, at substantially reduced computational cost. Here the distance kernel, with a large critic, seemed to do somewhat worse. With a small critic, WGAN-GP and Cramér GAN both performed very poorly.

These results illustrate the benefits of using the MMD on deep convolutional feaures as a GAN critic. In this hybrid system, the initial convolutional layers map the generator and reference image distributions to a simpler representation, which is well suited to comparison via the MMD. The MMD in turn employs an infinite dimensional feature space to compare the outputs of these convolutional layers. By comparison, WGAN-GP requires a larger discriminator network to achieve similar performance. It is interesting to consider the question of kernel choice: the distance kernel and RQ kernel are both characteristic (Sriperumbudur et al., 2010), and neither suffers from the fast decay of the exponentiated quadratic kernel, yet the RQ kernel performs slightly better in our experiments. The relative merits of different kernel families for GAN training will be an interesting topic for further study.

Table 2: Mean (standard deviation) of score evaluations for the CIFAR-10 models.

Model	Critic size	Inception	FID	KID
rq	16	6.55 (0.02)	39.17 (0.13)	0.027 (0.001)
dist	64	6.03 (0.01)	46.53 (0.11)	0.034 (0.001)
Cramér GAN	16	5.07 (0.01)	69.11 (0.09)	0.055 (0.002)
WGAN-GP	16	4.63 (0.01)	88.11 (0.13)	0.063 (0.001)
WGAN-GP	64	6.53 (0.02)	37.52 (0.19)	0.026 (0.001)
test set	–	11.21 (0.13)	6.11 (0.05)	0.000 (0.000)



(a) MMD *rq*, critic size 16 (b) MMD *dist*, critic size 64 (c) Cramér GAN, critic size 16

(d) WGAN-GP, critic size 16 (e) WGAN-GP, critic size 64 (f) Test set

Figure 3: Comparison of samples from various models, as well as true samples from the test set. WGAN-GP samples with critic size 16 are quite bad, as well as being under-diverse. Cramér GAN samples with critic size 16 are more appealing to the eye, but seem to be under-diverse: note the several near-repeats in the 64 images shown here. Large-critic WGAN-GP and the two MMD GAN variants are of similar quality.

REFERENCES

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017. arXiv:1701.04862.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. arXiv:1701.07875.

S. Arora and Y. Zhang. Do GANs actually learn the distribution? an empirical study, 2017. arXiv:1706.08224.

M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer distance as a solution to biased Wasserstein gradients, 2017. arXiv:1705.10743.

Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai. Better mixing via deep representations. In *ICML*, 2013.

D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks, 2017. arXiv:1703.10717.

I. Danihelka, B. Lakshminarayanan, B. Uria, D. Wierstra, and P. Dayan. Comparison of maximum likelihood and GAN-based training of Real NVPs, 2017. arXiv:1705.05263.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015. arXiv:1505.03906.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences, 2017. arXiv:1706.00292.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *JASA*, 102 (477):359–378, 2007.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. arXiv:1406.2661.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *JMLR*, 13, 2012.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In *NIPS*, 2017. arXiv:1704.00028.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *NIPS*, 2017. arXiv:1706.08500.

R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis, 2017a. arXiv:1704.04086.

X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie. Stacked generative adversarial networks. In *CVPR*, 2017b. arXiv:1612.04357.

Y. Jin, K. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang. Towards the automatic anime characters creation with generative adversarial networks, 2017. arXiv:1708.05509.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. arXiv:1412.6980.

A. Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

C. Li, D. Alvarez-Melis, K. Xu, S. Jegelka, and S. Sra. Distributional adversarial networks, 2017a. arXiv:1706.09549.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *NIPS*, 2017b. arXiv:1705.08584.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *ICML*, 2015. arXiv:1502.02761.

Lydia Tingruo Liu. On the two-sample statistic approach to generative adversarial networks. Master's thesis, University of Princeton Senior Thesis, April 2017. URL `http://arks.princeton.edu/ark:/88435/dsp0179408079v`.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3051–3696, 2013.

Y. Mroueh and T. Sercu. Fisher GAN. In *NIPS*, 2017. arXiv:1705.09675.

Y. Mroueh, T. Sercu, and V. Goel. McGan: Mean and covariance feature matching GAN. In *ICML*, 2017. arXiv:1702.08398.

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

S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS 29*, pp. 271–279. 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. arXiv:1511.06434.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NIPS*, 2016. arXiv:1606.03498.

D. Sejdinovic, B. K. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Stastistics*, 41(5):2263–2291, 2013. arXiv:1207.6076.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010. arXiv:0907.5309.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *JMLR*, 12:2389–2410, 2011. arXiv:1003.0887.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017. arXiv:1611.04488.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. arXiv:1312.6199.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016. arXiv:1512.00567.

G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *ICLR*, 2016. arXiv:1511.01844.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. arXiv:1703.10593.

A SCORE FUNCTIONS, DIVERGENCES, AND THE CRAMÉR GAN

It is not immediately obvious how to interpret the surrogate loss (10). An insight comes from considering the *score function* associated with the energy distance, which we now briefly review (Gneiting & Raftery, 2007). A scoring rule is a function $S(P, y)$, which is the loss incurred when a forecaster makes prediction P, and the event y is observed. The *expected score* is the expectation under Q of the score,

$$S(P, Q) := \mathbb{E}_{Y \sim Q}\, S(P, Y).$$

If a score is *proper*, then the expected score obtained when $P = Q$ is greater or equal than the expected score for $P \neq Q$,

$$S(Q, Q) \geq S(P, Q).$$

A *strictly proper* scoring rule shows an equality only when P and Q agree. We can define a divergence measure based on this score,

$$D_S(P, Q) = S(Q, Q) - S(P, Q). \tag{11}$$

Bearing in mind the definition of the divergence (11), it is easy to see (Gneiting & Raftery, 2007, eq. 22) that the energy distance (7) arises from the score function

$$S(P, y) = \frac{1}{2}\, \mathbb{E}_P\, \rho(X, X') - \mathbb{E}_P\, \rho(X, y).$$

The interpretation is straightforward: the score of a reference sample y is determined by comparing its average distance to a generator sample with the average distance among independent generator samples, $\mathbb{E}_P\, \rho(X, X')$.

As discussed earlier, the Cramér GAN critic does not use the energy distance (7) directly on the samples, but first maps the samples through a function h, for instance a convolutional network; this should be chosen to maximize the discriminative performance of the critic. Writing this mapping as h, we break the energy distance down as $D_e(P, Q) = S(Q, Q) - S(P, Q)$, where

$$S(Q, Q) = -\frac{1}{2}\, \mathbb{E}_Q\, \rho(h(Y), h(Y')) \tag{12}$$

and

$$S(P, Q) = \frac{1}{2}\, \mathbb{E}_P\, \rho(h(X), h(X')) - \mathbb{E}_{PQ}\, \rho(h(X), h(Y)). \tag{13}$$

When training the discriminator, the goal is to maximize the divergence by learning h, and so both (12) and (13) change: in other words, divergence maximization is not possible without two independent samples Y, Y' from the reference distribution Q.

An alternative objective in light of the score interpretation, however, is to simply optimize the average score (13). In other words, we would find features h that make the average distance from generator to reference samples much larger than the average distance between pairs of generator samples. We no longer control the term encoding the "variability" due to Q, $\mathbb{E}_Q\, \rho(h(Y), h(Y'))$, which might therefore explode: for instance, h might cause $h(Y)$ to disperse broadly, and far from the support of P, assuming sufficient flexibility to keep $\mathbb{E}_P\, \rho(h(X), h(X'))$ under control. We can mitigate this by controlling the expected norm $\mathbb{E}_Q\, \rho(h(Y), 0)$, which has the advantage of only requiring a single sample to compute. For example, we could maximize

$$-\frac{1}{2}\, \mathbb{E}_P\, \rho(h(X), h(X')) + \mathbb{E}_{PQ}\, \rho(h(X), h(Y)) - \mathbb{E}_Q\, \rho(h(Y), 0).$$

This resembles the Cramér GAN critic (10), but the generator-to-generator distance is scaled differently, and there is an additional term: $\mathbb{E}_P\, \rho(h(X), 0)$ is being maximized, which is more difficult to interpret. An argument has been made (in personal communication with Bellemare et al.) that this last term is required if the function f_c in (9) is to be a witness of an integral probability metric (1), although the asymmetry of this witness in P vs Q needs to be analyzed further.

B PROOF OF UNBIASED GRADIENTS

In this section we prove the interversion of the expectation and the gradient in Theorem 1.

We first consider a class of "well-behaved" activation functions:

Definition 1. *An activation function will be said to be well behaved if it satisfies the following:*

- $\sigma(0) = 0$
- *σ is differentiable on $\mathbb{R}^*$*
- *The derivatives of σ are bounded $|\frac{d\sigma}{dx}| \leq L$*
- *There exist a family of activation functions $(\sigma_t)_{t>0}$ that are all differentiable on $\mathbb{R}$ and such that*
- *The family σ_t converges uniformly towards σ on $\mathbb{R}$.*
- *The family of derivatives $\sigma_t^{'}$ converges locally uniformly towards $\sigma^{'}$ on $\mathbb{R}^*$.*
- *$\sigma_t^{'}(0)$ converges towards a finite value c*

Given such conditions, the family σ_t can be extended at the value 0 by $\sigma_0 = \sigma$. By convention we denote $\sigma_0^{'}(0) = c$, although it is not the derivative of σ_0.

Definition 2. *Given an input X in $\mathbb{R}^d$, we define the feedforward neural network ϕ_θ with K hidden layers by the following:*

$$h^0 = X$$
$$h^k = \sigma(b^k + g_k(W^k)h^{k-1}) \qquad \forall k \in \{1, ..., K\}$$

Where σ is a well-behaved activation function as defined in Definition 1. b^k is a vector in $\mathbb{R}^{d_k}$ and W^k is a matrix of weights. The functions g_k are introduced to account for convolutions and are known linear transformations. the only requirement is that $g_k(W^k)$ is a d_k by d_{k-1} matrix with ($d_0 = d$). We will denote by θ the concatenation of all the parameters of the network $\theta = (b^k, W^k,)_{k=1}^K$ and we will denote the last layer h^K by $\phi_\theta(X)$, $\theta \in \mathbb{R}^m$. For simplicity we will drop the subscript for functions g.

We will first consider the case where σ is differentiable at 0. The case where σ is non differentiable will then be obtained by considering a parametric family of differentiable activation functions σ_α such that σ_α converges point-wise towards σ when α goes to 0. We will then extend the results by taking the limit when α approches 0. We have the following results:

Theorem 2. *If the activation function σ is differentiable on $\mathbb{R}$ and has bounded derivatives, then $\theta \mapsto \phi_\theta(X)$ is differentiable for all X in $\mathbb{R}^d$. Moreover, there exist continuous real valued functions defined on $\mathbb{R}^m$ $\alpha(\theta)$, $\beta(\theta)$, $a(\theta)$ and $b(\theta)$ such that:*

$$\|\phi_\theta(X)\| \leq b(\theta) + a(\theta)\|X\|$$
$$\|\partial_\theta \phi_\theta(X)\| \leq \beta(\theta) + \alpha(\theta)\|X\|$$

for all X in $\mathbb{R}^d$.

Proof. First note that σ is L-Lipchitz, therefore : $\|\sigma(X)\| \leq L\|X\|$ for any finite dimensional vector, where σ is applied point-wise and the norm used is the euclidian norm. Let now k be in $\{1, ..., K\}$, we have the following:

$$\begin{aligned}\|h^k\| &\leq \|\sigma(b^k + g(W^k)h^{k-1})| \\ &\leq L\|b^k + g(W^k)h^{k-1}| \\ &\leq L\|b^k\| + L\|g(W^k)\|_{Op}\|h^{k-1}\|\end{aligned}$$

A simple recursion then gives:

$$\|h^k\| \leq \|X\|a_k(\theta) + b_k(\theta)$$

With $a_k(\beta) = L^k \prod_{j=1}^k \|g(W^j)\|_{Op}$ and $b_k(\theta) = L^k \sum_{i=1}^n \|b^i\| \prod_{j=i+1}^n \|g(W^j)\|_{Op}$

By setting $b(\theta) = b_L(\theta)$ and $a(\theta) = a_L(\theta)$ one gets the first bound.

For the gradient, we will proceed by recursion. For $k = 0$ we trivially have $\|\partial_\theta h^0\| = 0$. Now for some k in $\{1, ...K\}$, we assume that $\|\partial_\theta h^{k-1}\| \leq \beta_{k-1}(\theta) + \alpha_{k-1}(\theta)\|X\|$. For the parameters $\bar{\theta}$ of layers higher than k we have $\|\partial_{\bar{\theta}} h^k\|^2 = 0$. Now for the parameters $\widetilde{\theta}$ of layers lower than k we get:

$$\begin{aligned}\|\partial_{\widetilde{\theta}} h^k\| &\leq L\|g(W^k)\partial_{\widetilde{\theta}} h^{k-1}\| \\ &\leq L\|g(W^k)\|_{Op}\|\partial_{\widetilde{\theta}} h^{k-1}\| \\ &\leq L\|g(W^k)\|_{Op}\|\partial_\theta h^{k-1}\|\end{aligned}$$

By recursion we get $\|\partial_{\widetilde{\theta}} h^k\| \leq L\|g(W^k)\|(\beta_{k-1}(\theta) + \alpha_{k-1}(\theta)\|X\|)$ Now for the parameters at layer k $\theta_k = (b^k, W^k)$ we also have:

$$\|\partial_{\theta_k} h^k\| \leq L\big(1 + \|h^{k-1}\|\big) \leq L(1 + b_k(\theta) + a_k(\theta)\|X\|\big]$$

Using Minkowski inequality we finally get:

$$\|\partial_{\theta_k} h^k\| \leq \alpha_k\|X\| + \beta_k(\theta)$$

with $\alpha_k = La_k(\theta) + L\|g(W^k)\|\alpha_{k-1}(\theta)$ and $\beta_k = L(1 + b_k\theta) + L\|g(W^k)\|\beta_{k-1}(\theta)$. The results follows by taking $k = K$.

□

Let now $\mathbb{P}$ and $\mathbb{Q}$ be two probability distribution over the input space $\mathbb{R}^d$ and let X and Y be two random variables following $\mathbb{P}$ and $\mathbb{Q}$ respectively. We also consider a kernel $\mathcal{K}$ defined on the feature space.

We make the following assumptions on $\mathbb{P}$, $\mathbb{Q}$ and the kernel $\mathcal{K}$ for a given $\alpha \geq 1$:

A $\mathbb{E}_{\mathbb{P}}[\|X\|^\alpha] < \infty$ and $\mathbb{E}_{\mathbb{Q}}[\|Y\|^\alpha] < \infty$

B The kernel is differentiable and satisfies the growth conditions:

$$\|\mathcal{K}(U, V)\| \leq C(\|U\|^\alpha + \|V\|^\alpha)$$
$$\|\nabla\mathcal{K}(U, V)\| \leq C(\|U\|^{\alpha-1} + \|V\|^{\alpha-1})$$

C The gradient of the kernel is M-Lipchitz for some positive M, i.e:

$$\|\nabla\mathcal{K}(X, Y) - \mathcal{K}(X^{'}, Y^{'})\| \leq M\|(X, Y) - (X^{'}, Y^{'})\|$$

Here C is a positive constant.

We have the following result:

Theorem 3. *If the activation function σ is differentiable then $\theta \mapsto \mathbb{E}_{\mathbb{P},\mathbb{Q}}[\mathcal{K}(\phi_\theta(X), \phi_\theta(Y))]$ is well-defined and differentiable for all $\theta \in \mathbb{R}^m$. Moreover, the following formula holds:*

$$\partial_\theta\, \mathbb{E}_{\mathbb{P},\mathbb{Q}}[K(\phi_\theta(X), \phi_\theta(Y))] = \mathbb{E}_{\mathbb{P},\mathbb{Q}}[\partial_\theta K(\phi_\theta(X), \phi_\theta(Y))]$$

Proof. The integrability of $\mathbb{E}_{\mathbb{P},\mathbb{Q}}[\mathcal{K}(\phi_\theta(X), \phi_\theta(Y))]$ is straightforward using the bound on the network and on the kernel. For the differentiability, we will use the dominated convergence theorem . Indeed, by the chain rule one has:

$$\partial_\theta\mathcal{K}(\phi_\theta(X), \phi_\theta(Y)) = \nabla\mathcal{K}(\phi_\theta(X), \phi_\theta(Y)) \circ (\partial_\theta\phi_\theta(X), \partial_\theta\phi_\theta(Y))$$

Taking the norm for each operator respectively, one gets:

$$\begin{aligned}\|\partial_\theta\mathcal{K}(\phi_\theta(X), \phi_\theta(Y))\| &\leq \|\nabla\mathcal{K}(\phi_\theta(X), \phi_\theta(Y))\|\|(\partial_\theta\phi_\theta(X), \partial_\theta\phi_\theta(Y))\| \\ &\leq C(\|\phi_\theta(X)\|^{\alpha-1} + \|\phi_\theta(Y)\|^{\alpha-1})(\|\partial_\theta\phi_\theta(X)\| + \|\partial_\theta\phi_\theta(Y)\|))\end{aligned}$$

We will use now the bounds on $\phi_\theta(X)$ and $\partial_\theta\phi_\theta(X)$. First recall that the functions α, β, a and b are continuous on $\mathbb{R}^m$ and considering a closed ball $B(\theta_0, \mu)$ for some θ_0 in $\mathbb{R}^m$ and $\mu > 0$, each of those functions attain a maximum value on that ball. We simply denote by b^*, a^*, β^* and α^* these maximum values and use them directly in the bounds. We also introduce a threshold $R > 0$ on the norms $\|X\|$ and $\|Y\|$. Let $G_\theta(X, Y) := \|\partial_\theta\mathcal{K}(\phi_\theta(X), \phi_\theta(Y))\|$ for convenience, we then get 4 different cases:

- If $\|X\| \leq R$ and $\|Y\| \leq R$, then:

$$G_\theta(X,Y) \leq F_1(X,Y) := 4\Big[b^* + a^* R\Big]^{\alpha-1}\Big[\beta^* + \alpha^* R\Big]$$

- If $\|X\| > R$ and $\|Y\| > R$, then:

$$G_\theta(X,Y) \leq F_2(X,Y)$$

with

$$F_2(X,Y) := \Big[\frac{b^*}{R} + a^*\Big]^{\alpha-1}\big(\|X\|^{\alpha-1} + \|Y\|^{\alpha-1}\big)(2\beta^* + \alpha^*(\|X\| + \|Y\|))$$

- If $\|X\| > R$ and $\|Y\| \leq R$, then:

$$G_\theta(X,Y) \leq F_3(X,Y)$$

with

$$F_3(X,Y) := \Big[\big(\frac{b^*}{R} + a^*\big)^{\alpha-1}\|X\|^{\alpha-1} + \big(b^* + a^* R\big)^{\alpha-1}\big)(2\beta^* + \alpha^*(\|X\| + R))$$

- If $\|Y\| > R$ and $\|X\| \leq R$, then

$$G_\theta(X,Y) \leq F_4(X,Y)$$

with

$$F_4(X,Y) = F_3(Y,X)$$

It is clear from Assumption **A** that the functions F_1, F_3 and F_4 are integrable. F_2 involves terms of the form $\|X\|^{\alpha-1}\|Y\|$ which can be bounded in expectation using Holder inequality:

$$\mathbb{E}_{\mathbb{P},\mathbb{Q}}[\|X\|^{\alpha-1}\|Y\|] \leq \Big(\mathbb{E}_{\mathbb{P}}[\|X\|^\alpha]\Big)^{1-\frac{1}{\alpha}}\Big(\mathbb{E}_{\mathbb{Q}}[\|Y\|^\alpha]\Big)^{\frac{1}{\alpha}}$$

Therefore F_2 is also integrable. Now denoting by:

$$\begin{aligned} F(X,Y) :=& F_1(X,Y)\mathbb{1}_{\|X\|\leq R,\|Y\|\leq R} + F_2(X,Y)\mathbb{1}_{\|X\|>R,\|Y\|>R} \\ &+ F_3(X,Y)\mathbb{1}_{\|X\|>R,\|Y\|\leq R} + F_4(X,Y)\mathbb{1}_{\|X\|\leq R,\|Y\|>R} \end{aligned}$$

we have that $G_\theta(X,Y)$ is upper bounded by F uniformly on the parameter ball $B(\theta_0,\eta)$. The dominated convergence applies then and we get the interversion result for all θ_0 in $\mathbb{R}^m$.

□

Now we would like to consider the case when the activation function σ is not differentiable in 0. We will consider a family of activation functions σ_t that are differentiable on $\mathbb{R}$ and that converges point-wise towards σ as t goes to 0. To keep the proof simple, we consider the case of the RELU activation function $\sigma(x) = \max(0,x)$ and the following family of approximating functions:

$$\sigma_t(x) = t(\log(1 + e^{\frac{x}{t}}) - \log 2)$$

for $t > 0$. It is easy to see why these functions converge towards σ. Moreover, the derivative of σ_t is given by:

$$\frac{d\sigma_t(x)}{dx} = 1 - \frac{1}{1 + e^{\frac{x}{t}}}$$

It is also clear that $\frac{d\sigma_t(x)}{dx}$ converge towards $\frac{d\sigma(x)}{dx}$ for all $x \in \mathbb{R}^*$. For $x = 0$, $\frac{d\sigma_t(x)}{dx} \to \frac{1}{2}$, which is a sub-gradient of σ in $x = 0$. And $\frac{d\sigma_t(x)}{dx}$ is uniformly bounded in x and t.

A key property of this family is that it is uniformly convergent and the derivatives $\frac{d\sigma_t(x)}{dx}$ are locally uniformly convergent on $\mathbb{R}^*$. In other words if C is a compact set included in $\mathbb{R}^*$ then:

$$\|\frac{d\sigma_t}{dx} - \frac{d\sigma}{dx}\|_{\infty,C} \to_{t\to 0} 0$$

where norm used is the supremum norm over the set C. We will extend the family to $t = 0$ by setting $\sigma_0 = \sigma$ and we also denote by $\sigma_0^{'} = \lim_{t\to 0} \sigma_t^{'}$. Note that $\sigma_0^{'}$ coincides with the derivative of σ on $\mathbb{R}^*$ and has a defined value in 0.

Let's define the family of functions $f_t(\theta) := \mathbb{E}_{\mathbb{P},\mathbb{Q}}[\mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y))]$ and $g_t(\theta) := \mathbb{E}_{\mathbb{P},\mathbb{Q}}[\partial_\theta \mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y))]$. Here we denoted by $\phi_\theta^{(t)}(X)$ the network with activation function σ_t.

We have the following proposition:

Lemma 1. *For a fixed θ, X and Y, then the following point-wise convergence hold:*

$$\mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y)) \to_{t\to 0} \mathcal{K}(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))$$
$$\partial_\theta \mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y)) \to_{t\to 0} \partial_\theta \mathcal{K}(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))$$

Moreover, $f_t(\theta) \to_{t\to 0} f_0(\theta)$ and $g_t(\theta) \to_{t\to 0} g_0(\theta)$

Proof. The point-wise convergence is obtained directly by taking the limit. The limits in expectation are obtained using the dominated convergence theorem and Assumptions **A** and **B** as well the growth properties of the network Theorem 2. □

Now we would like to prove that $\partial_\theta f_0(\theta) = g_0(\theta)$ for μ-almost all θ. For this to hold we need to prove first that $g_t(\theta)$ converges locally uniformly towards $g_0(\theta)$ as t goes to 0. Let's first introduce some notations. For a given $\theta_0 \in \mathbb{R}^d$ and $\eta > 0$ we define the set

$$\mathcal{N}(\theta_0, \eta) = \left\{ X \in \mathbb{R}^d \middle| \exists \theta \in B(\theta_0, \eta), k \in \{1, ..., K\}, i \in \{1, ..., d_k\} : b_i^k + g(W^k)_i h^{k-1} = 0 \right\}$$

$\mathcal{N}(\theta_0, \eta)$ is the set of X where at least one linear form in the network vanishes. This allows to introduce the set of "well-behaved" parameters:

$$\Theta_{\mathbb{P}} := \left\{ \theta \in \mathbb{R}^m \middle| (\forall \epsilon > 0)(\exists \eta > 0) : \mathbb{P}(\mathcal{N}(\theta, \eta)) \leq \epsilon \right\}$$

Note that $\mathcal{N}(\theta_0, \eta)$ is always a closed set. We have the following uniform convergence result:

Lemma 2. *For all $k \in \{0, ..., K\}$, the family of hidden layers $(h_t^k)_{t\geq 0}$ obtained using the activations functions σ_t converge uniformly on $B(\theta_0, \eta) \times \mathbb{R}^d$ towards h_0^k for every $\theta_0 \in \mathbb{R}^m$ and $\eta > 0$.*

Proof. We proceed by recursion on the layers. For $k = 0$ the statement trivially holds. Assume now that for $k-1$, $(h_t^{k-1})_{t\geq 0}$ converges uniformly on a ball $B(\theta_0, \eta) \times \mathbb{R}^d$. We have the following:

$$\begin{aligned}
\|h_t^k - h_0^k\| =& \|\sigma_t(b^k + g(W^k)h_t^{k-1}) - \sigma_0(b^k + g(W^k)h_0^{k-1})\| \\
\leq& \|\sigma_t(b^k + g(W^k)h_t^{k-1}) - \sigma_0(b^k + g(W^k)h_t^{k-1})\| \\
&+ \|\sigma_0(b^k + g(W^k)h_t^{k-1}) - \sigma_0(b^k + g(W^k)h_0^{k-1})\|.
\end{aligned}$$

For the first term we use the fact that σ_t is uniformly convergent on $\mathbb{R}$, for $\epsilon > 0$ there exists $\eta >$ such that for all $t < \eta$ we have that $\|\sigma_t(b^k + g(W^k)h_t^{k-1}) - \sigma_0(b^k + g(W^k)h_t^{k-1})\| \leq \epsilon$. For the second term we use the Lipchitz continuity of σ_0:

$$\begin{aligned}
\|\sigma_0(b^k + g(W^k)h_t^{k-1}) - \sigma_0(b^k + g(W^k)h_0^{k-1})\| &\leq L\|g(W^k)(h_t^{k-1} - h_0^{k-1})\| \\
&\leq L\|g(W^k)\|\|(h_t^{k-1} - h_0^{k-1})\|.
\end{aligned}$$

Since $\|g(W^k)\|$ is bounded on the ball $B(\theta_0, \eta)$, one gets the desired result by recursion. □

Lemma 3. *Under assumption Assumption **C**, and for all $\theta \in \Theta_{\mathbb{P}} \cap \Theta_{\mathbb{Q}}$, $g_t(\theta)$ converges locally uniformly towards $g_0(\theta)$.*

Proof. Let θ_0 be in $\Theta_{\mathbb{P}} \cap \Theta_{\mathbb{Q}}$ and let $\eta > 0$. For simplicity we drop the subscripts of the expectations. Using Assumption **C** one has:

$$\begin{aligned}
\|g_t(\theta) - g_0(\theta)\| \leq& \mathbb{E}[\|\nabla \mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y)) - \nabla \mathcal{K}(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))\|] \, \mathbb{E}[\|\partial_\theta \phi^t(X), \partial_\theta \phi^t(Y)\| \\
&+ \mathbb{E}[\|\nabla K(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))\|] \, \mathbb{E}[\|\partial_\theta \phi^t(X) - \partial_\theta \phi^0(X), \partial_\theta \phi^t(Y) - \partial_\theta \phi^0(Y)\|]
\end{aligned}$$

All the expectations that do not involve a difference between two terms can be upper bounded using Assumption **B** and the growth properties of the network in Theorem 2 over all parameters in $B(\theta_0, \eta)$. We will first control the difference $\mathbb{E}[\|\mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y)) - \mathcal{K}(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))\|]$. It can be uniformly controlled using the the uniform convergence of $\phi_(t)$ towards ϕ_0 and continuity of the gradient $\nabla\mathcal{K}$. Let A_n be a partition of the set $\mathbb{R}^d \times \mathbb{R}^d$, such that A_n are compact. The exist $N_0 > 0$ such that $\sum_{i=N_0+1}^{\infty} \mathbb{E}_{\mathbb{P},\mathbb{Q}}(A_n) < \epsilon$ On one hand we have that $\mathbb{E}[\|\mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y)) - \mathcal{K}(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))\|]$ is bounded uniformly on the ball $B(\theta_0, \eta)$ by some constant S. This comes by virtue of the growth properties of the network in Theorem 2 and of the gradient of the kernel in Assumption **B**. It comes that:

$$\sum_{n=N_0+1}^{\infty} \mathbb{E}[\|\mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y)) - \mathcal{K}(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))\| | (X,Y) \in A_n] \, \mathbb{E}_{\mathbb{P},\mathbb{Q}}(A_n) \leq \epsilon$$

On the other hand by uniform convergence of $\phi_\theta^{(t)}$ towards $\phi_\theta^{(0)}()$, there exists $\alpha > 0$ such that $\|\phi_\theta^{(t)}(X) - \phi_\theta^{(0)}(X)\| \leq \beta$ and $\|\phi_\theta^{(t)}(X) - \phi_\theta^{(0)}(X)\| \leq \beta$ for all $(X,Y) \in A_n$ for a given β. Moreover, The image U of $B(\theta_0, \eta) \times A_n$ by $(\theta, X, Y) \mapsto (\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y))$ is included in a compact set C. By continuity of the gradient $\nabla\mathcal{K}$, one deduces that it is uniformly continuous on the compact set C, therefore provided that β is small enough, one can get that:

$$\sum_{n=1}^{N_0} \mathbb{E}[\|\mathcal{K}(\phi_\theta^{(t)}(X), \phi_\theta^{(t)}(Y)) - \mathcal{K}(\phi_\theta^{(0)}(X), \phi_\theta^{(0)}(Y))\| | (X,Y) \in A_n] \, \mathbb{E}_{\mathbb{P},\mathbb{Q}}(A_n) \leq \epsilon$$

Now we will control the error:

$$\mathbb{E}[\|\partial_\theta \phi^t(X) - \partial_\theta \phi^0(X), \partial_\theta \phi^t(Y) - \partial_\theta \phi^0(Y)\|]$$

Here we will only provide a proof for the variable X, the same holds for Y. We proceed by recursion over the layers. Indeed for $k = 0$, $h^k = X$ and $\partial_\theta h^k = 0$ which define a constant family in t that is uniformly convergent. Assume now that for layer $k-1$ for every $\epsilon > 0$, there exists $\alpha_{k-1} < \alpha$ such that $E[\|\partial_\theta h_t^{k-1} - \partial_\theta h_0^{k-1}\|] \leq \epsilon$ for all $\theta \in B(\theta, \eta)$. Note that h^{k-1} depends on θ although it is not explicitly shown by the notation. We have the following decomposition:

$$\begin{aligned}
\mathbb{E}[\|\partial_\theta h_t^k - \partial_\theta h_0^k\|] =& \mathbb{E}[\|\partial_\theta \sigma_t(b^k + g(W^k)h_t^{k-1}) - \partial_\theta \sigma_0(b^k + g(W^k)h_0^{k-1})\|] \\
=& \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1}) - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})\|] \\
&+ \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1})h_t^{k-1} - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})h_0^{k-1}\|] \\
&+ \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1})g(W^k)\partial_\theta h_t^{k-1} - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})g(W^k)\partial_\theta h_0^{k-1}\|]
\end{aligned}$$

We get three terms, in the second and third terms one can add and subtract a control term that allows the use of the recursion assumption for h_t^{k-1} and $\partial_\theta h_t^{k-1}$. These terms can therefore be bounded by ϵ uniformly on the ball $B(\theta_0, \eta)$ provided that $t \leq \alpha_{k-1}$. The remaining terms are proportional to:

$$\mathcal{E}_t := \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1}) - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})\|]$$

which can further be upper-bounded by:

$$\begin{aligned}
\mathcal{E}_t \leq& \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1}) - \sigma_t^{'}(b^k + g(W^k)h_0^{k-1})\|] \\
&+ \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_0^{k-1}) - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})\|]
\end{aligned}$$

We will control now the errors:

$$\begin{aligned}
A_1 &:= \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1}) - \sigma_t^{'}(b^k + g(W^k)h_0^{k-1})\|] \\
A_2 &:= \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_0^{k-1}) - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})\|]
\end{aligned}$$

Let's consider the set $\mathcal{N}(\theta_0, \eta)$, Note that its complementary $\mathcal{N}^c(\theta_0, \eta)$ is an open set. Let D_n be a partition of this open set: $\mathcal{N}^c(\theta_0, \eta) = \cup_{n=1}^{\infty} D_n$, we assume that D_n are compact. If we restrict the parameters θ to a closed ball $\bar{B}(\theta_0, \widetilde{\eta})$ then $\bar{B}(\theta_0, \widetilde{\eta}) \times D_n$ is a compact set. We denote by V the image of $\bar{B}(\theta_0, \widetilde{\eta}) \times D_n$ by the mapping $R : (\theta, X) \mapsto b^k + g(W^k)h_0^{k-1}$. One can easily see that there exists a ball $B(0, r)$ centered on 0 with some radius r such that its intersection with V is empty. This comes from the fact that V is continuous on the compact set $\bar{B}(\theta_0, \widetilde{\eta}) \times D_n$. It follows then from the uniform convergence of h_t^{k-1} towards h_0^{k-1} that there exists an $\alpha_1 > 0$ such that for all $t \leq \alpha_1$ V_t the image of $\bar{B}(\theta_0, \widetilde{\eta}) \times D_n$ by the application $R_t : (\theta, X) \mapsto b^k + g(W^k)h_t^{k-1}$ has an empty intersection with a smaller ball $B(0, \frac{r}{2})$.

One can then safely use the fact that $\sigma_t^{'}$ is locally Lipchitz on $\mathbb{R} \setminus B(0, \frac{r}{2})$ with some constant L_r. Note that the Lipchitz constant L_r depends on r which in turn depends on the set D_n. Therefore, L_r might diverge as D_n gets closer to the critical set $\mathcal{N}(\theta_0, \eta)$. To avoid this issue, we will make sure to cut off the contribution of sets D_n that are too small anyways. Indeed there exists $N_0 > 0$ such that $\sum_{N_0+1}^{\infty} \mathbb{P}(D_n) \leq \epsilon$. One can therefore use the following decomposition for A_1:

$$
\begin{aligned}
A_1 = & \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1}) - \sigma_t^{'}(b^k + g(W^k)h_0^{k-1})\nVdash_{X \in \mathcal{N}(\theta_\prime, \eta)}\|] \\
& + \sum_{n=1}^{N_0} \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1}) - \sigma_t^{'}(b^k + g(W^k)h_0^{k-1})\nVdash_{X \in D_n}\|] \\
& + \sum_{n=N_0+1}^{\infty} \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_t^{k-1}) - \sigma_t^{'}(b^k + g(W^k)h_0^{k-1})\nVdash_{X \in D_n}\|]
\end{aligned}
$$

The third term is bounded by $2L\epsilon$ while the second term is uniformly controlled by $C(\sum_{n=1}^{N_0} L_{D_n})\epsilon$. Remains the first term. since θ_0 belongs to Θ, provided that η is small enough then $\mathbb{P}(\mathcal{N}(\theta_0, \eta)) \leq \epsilon$, the result follows using the fact that $\sigma_t^{'}$ is bounded.

The term A_2 can be controlled in exactly the same way:

$$
\begin{aligned}
A_2 = & \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_0^{k-1}) - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})\nVdash_{X \in \mathcal{N}(\theta_\prime, \eta)}\|] \\
& + \sum_{n=1}^{N_0} \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_0^{k-1}) - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})\nVdash_{X \in D_n}\|] \\
& + \sum_{n=N_0+1}^{\infty} \mathbb{E}[\|\sigma_t^{'}(b^k + g(W^k)h_0^{k-1}) - \sigma_0^{'}(b^k + g(W^k)h_0^{k-1})\nVdash_{X \in D_n}\|]
\end{aligned}
$$

Here the first and third terms are controlled by the sizes of the sets D_n and $\mathcal{N}(\theta_\prime, \eta)$, while for the second term one can simply use the locally uniform convergence of $\sigma_t^{'}$ towards $\sigma_0^{'}$ since the images of the compact sets $\bar{B}(\theta_0, \widetilde{\eta}) \times D_n$ by the mapping $R : (\theta, X) \mapsto b^k + g(W^k)h_0^{k-1}$ are far away from 0 by at least a positive number r.

We have shown that for $\epsilon > 0$ there exists $\eta > 0$ and $\alpha > 0$ such that:

$$\|g_t(\theta) - g_0(\theta)\| \leq \epsilon$$

for all $\theta \in B(\theta_0, \eta)$ and all $t \leq \alpha$, which exactly means that g_t converges locally uniformly to g_0.

□

Lemma 4. *The set $\Theta_{\mathbb{P}}^c$ has 0 Lebesgue measure.*

Proof. The key idea here is to prove that for any $\theta \in \Theta_{\mathbb{P}}^c$, and along some coordinates, there cannot be more than a countable number of neighbors of θ that are also in $\Theta_{\mathbb{P}}^c$.

One should first notice that:

$$\Theta_{\mathbb{P}}^c = \left\{\theta \in \mathbb{R}^m \middle| (\exists \epsilon > 0)\mathbb{P}(\mathcal{N}(B(\theta, 0))) > \epsilon\right\}$$

Indeed, if $\theta \in \Theta_{\mathbb{P}}^c$ then there exists $\epsilon > 0$ such that for all $\eta > 0$ we have:

$$\mathbb{P}(\mathcal{N}(B(\theta, \eta))) > \epsilon$$

We also have that $\mathbb{1}_{X \in \theta, \eta} \to \mathbb{1}_{X \in \theta,'}$ from the continuity of the network. Therefore by the dominated convergence theorem one gets that $\mathbb{P}(\mathcal{N}(B(\theta, 0))) = \lim_{\eta \to 0} \mathbb{P}(\mathcal{N}(B(\theta, \eta))) > \epsilon$. The converse is trivial.

Let θ_0 be in $\Theta_{\mathbb{P}}^c$, then by definition there exists $\epsilon > 0$ such that $\mathbb{P}(\mathcal{N}(\theta_0, 0)) > \epsilon$. This also means that there is a layer k and a coordinate i in the layer $k : i \in \{1, ..., d_k\}$ such that the set $\{X \in \mathbb{R}^d | b_i + g(W^k)_i h^{k-1} = 0\}$ has a non zero measure. Without loss of generality, one can assume that $\mathcal{N}(\theta_0, 0) = \{X \in \mathbb{R}^d | b_i + g(W^k)_i h^{k-1} = 0\}$.

Consider now the family of parameters $(\theta^h)_{h \in \mathbb{R}}$ such that all the coordinates of θ^h match those of θ_0 except at layer k and for the node i where $b_i^h = b_i + h$. It is clear that the family of sets $\mathcal{N}(\theta_h, 0)$ are jointly disjoint. That is, if $h, h^{'} \geq$ are such that $h \neq h^{'}$ then $\mathcal{N}(\theta_h, 0) \cap \mathcal{N}(\theta_h^{'}, 0) = \emptyset$. We will prove now that the intersection A between this family $(\theta^h)_{0 \leq h}$ and $\Theta_{\mathbb{P}}^c$ can be at most countable. Indeed if A was uncountable, then this would define an uncountable family of sets $(\mathcal{N}(\theta, 0))_{\theta \in A}$ with positive measures. Let's now introduce the following family of sets:

$$A_n = \cup \{\theta \in A | \mathbb{P}(\mathcal{N}(\theta, 0)) > \frac{1}{n}\}$$

Since A is uncountable, then there exists n such that A_n is infinite. Now consider the set B obtained by taking the union of all the disjoint sets $\mathcal{N}(\theta, 0$ with $\theta \in A_n$ and if $(\theta_m)_{m=1}^{\infty}$ is an injective sequence in A_n then $\mathbb{P}(B) \geq \sum_{m=1}^{\infty} \mathbb{P}(\mathcal{N}(\theta_m, 0) \geq \sum_{m=1}^{\infty} \frac{1}{n} = +\infty$.

As this can't happen, it means that the intersection set A is at most countable.

So far we have shown that while keeping all the other parameters fixed for θ, only a countable possible values of the coordinate b_i^k can be taken in $\Theta_{\mathbb{P}}^c$. One can then easily conclude using Fubini theorem:

$$\begin{aligned}
\mu(\Theta_{\mathbb{P}}^c) &= \int_{\mathbb{R}^m} \mathbb{1}_{\theta \in \Theta_{\mathbb{P}}^c} d\theta \\
&= \int_{W^1, ..., W^K} \int_{(b_j^{k'})_{(k',j) \neq (k,i)}} \Big(\int_{b_i^k} \mathbb{1}_{\theta \in \Theta_{\mathbb{P}}^c} db_i^k \Big) \prod_{1 \leq k' \leq K, 1 \leq j \leq d_{k'},} db_j^{k'} \, dW_1 ... dW_K = 0
\end{aligned}$$

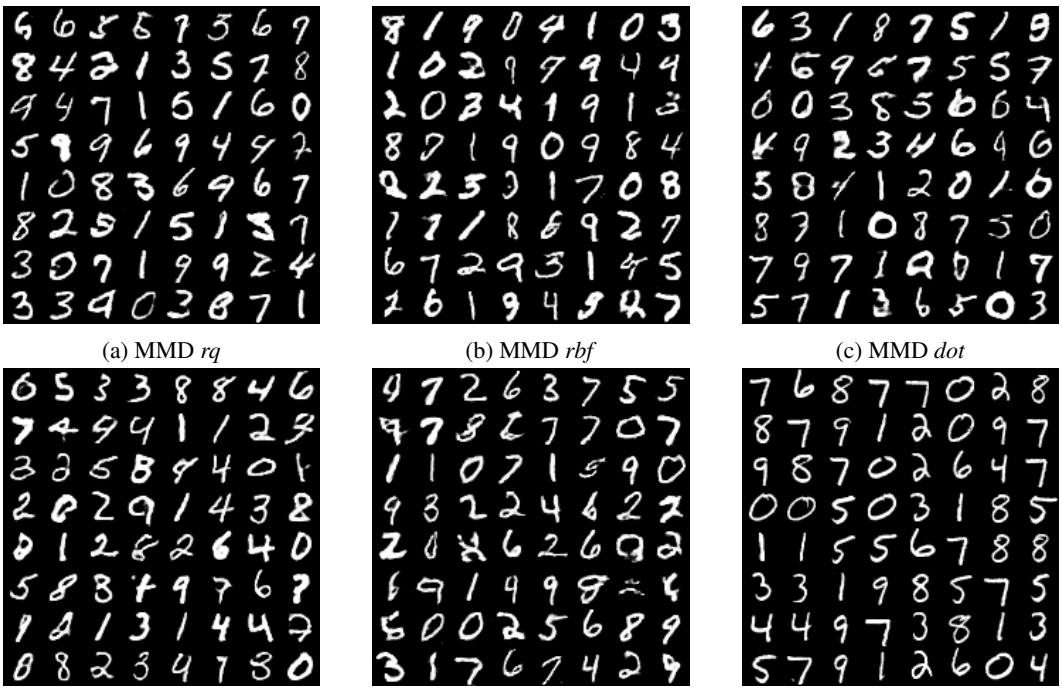
The integral over b_i^k is equal to zero which leads to the result.

□

We are now ready to prove the main result.

Proof of Theorem 1. By Theorem 3 we have that $\partial_\theta f_t(\theta) = g_t(\theta)$ for all $t > 0$ and $\theta \in \mathbb{R}^m$. We also have by Lemma 1 that $f_t(\theta)$ converges point-wise towards $f_0(\theta)$, and by Lemma 3, we know that $g_t(\theta)$ converges uniformly towards $g_0(\theta)$ for all θ be in $\Theta_{\mathbb{P}} \cap \Theta_{\mathbb{Q}}$. This allows to conclude that $f_0(\theta)$ is differentiable on $\Theta_{\mathbb{P}} \cap \Theta_{\mathbb{Q}}$ and its gradient is given by $g_0(\theta) = \mathbb{E}_{\mathbb{P},\mathbb{Q}}[\partial_\theta \mathcal{K}(\phi_\theta(X), \phi_\theta(Y))]$. Moreover, $(\Theta_{\mathbb{P}} \cap \Theta_{\mathbb{Q}})^c = \Theta_{\mathbb{P}}^c \cup \Theta_{\mathbb{Q}}^c$ has zero Lebesgue measure as it its the union of two Lebesgue 0 measure sets by Lemma 4. The conclusion follows. □

C MODEL SAMPLES



(a) MMD *rq* (b) MMD *rbf* (c) MMD *dot*

(d) Cramer GAN (e) WGAN-GP (f) MNIST test set

Figure 4: Samples from the models listed in Table 1. Rational-quadratic and Gaussian kernels obtain retain sample quality despite reduced discriminator complexity. Each of these models generates good quality samples with the standard DCGAN discriminator (critic size 64).