On the expressivity of bi-Lipschitz normalizing flows

Alexandre Verine ALEXANDRE.VERINE@DAUPHINE.PSL.EU
Benjamin Negrevergne BENJAMIN.NEGREVERGNE@DAUPHINE.PSL.EU
Yann Chevaleyre YANN.CHEVALEYRE@DAUPHINE.PSL.EU
LAMSADE, Université Paris-Dauphine, PSL, Paris, France

Fabrice Rossi ROSSI@CEREMADE.DAUPHINE.FR
CEREMADE, Université Paris-Dauphine, PSL, Paris, France

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

An invertible function is *bi-Lipschitz* if both the function and its inverse have bounded Lipschitz constants. Most state-of-the-art Normalizing Flows are bi-Lipschitz by design or by training to limit numerical errors (among other things). In this paper, we discuss the expressivity of bi-Lipschitz Normalizing Flows and identify several target distributions that are difficult to approximate using such models. Then, we characterize the expressivity of bi-Lipschitz Normalizing Flows by giving several lower bounds on the Total Variation distance between these particularly unfavorable distributions and their best possible approximation. Finally, we show how to use the bounds to adjust the training parameters, and discuss potential remedies.

1. Introduction

A weak smoothness condition for a function F, beyond continuity, can be enforced by requesting F to be L-Lipschitz, that is to verify

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \quad \|F(\boldsymbol{x}_1) - F(\boldsymbol{x}_2)\|_2 \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2,$$

where L is the Lipschitz constant of F.

A number of recent publications have demonstrated the benefits of constructing machine learning models with a small Lipschitz constant. First, models with a small Lipschitz constant have been linked with better generalization capabilities, both in terms of true risk (Bartlett et al., 2017), and adversarial risk (Farnia et al., 2018). In addition, models with a small Lipschitz constants are more stable during training (Miyato et al., 2018), and are less prone to numerical errors, a property which is particularly important in the context of invertible neural networks and normalizing flows (Behrmann et al., 2021).

Unfortunately, enforcing a small Lipschitz constant, either by design, or using regularization during training, can impede the ability of a model to fit the data distribution. Based on this observation, several researchers have studied the limitations of neural networks with bounded Lipschitz constants. In particular, Tanielian et al. (2020) were able to identify a family of target distributions with disconnected support that cannot be fitted with a Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) with a bounded Lipschitz constant. For the particular case of normalizing flows, Cornish et al. (2021) were able to

demonstrate that there exist pairs of latent and target distributions with particular topological conditions on their support, that require the model to have an unbounded Lipschitz constant.

While the universality and the consistency of Normalizing Flow have been investigated (Kong and Chaudhuri, 2020; Zhang et al., 2020; Teshima et al., 2020; Koehler et al., 2021), the numerical stability, i.e. the Lipschitz constraints are set aside. Only Kong and Chaudhuri (2021) derive a clear link between a Lipschitz constant and the universality as a Maximum Mean Discrepancy of a subclass of Normalzing Flow.

In this paper we focus on characterizing the impact of Lipschitz constraints on the expressivity of normalizing flows. More precisely, we discuss the impact of the Lipschitz constant on the Total Variation distance (TV) between the approximated distribution and the target distribution. We give several lower bounds on the TV distance which (unlike previous works), do not assume any hypothesis on the support of the target distribution, and are thus applicable to any learning settings. Furthermore, since Normalizing Flows are often not only Lipschitz, but *bi-Lipschitz*, (meaning that both the inverse mapping function and the mapping itself have bounded Lipschitz constant), we also study the impact of the Lipschitz constant of the mapping, on the expressivity. (Most work on the topic focus on the Lipschitz constant of the *inverse* mapping.) Building on this analysis, we give an additional bound that depends on the Lipschitz constant of the mapping. Then, we use our bounds to exhibit practical limitations of Lipschitz constrained models on real world datasets such as *CIFAR10*, and we discuss the potential remedies in the light of these new results. Finally, we show how the bound can help choosing the depth and the Lipschitz constraint required for a particular dataset.

To recap, our contribution is condensed to Theorems 8 and 9. We show that, with no specific hypothesis on the support, a normalizing flow will fail to perfectly approximate/generate if one the density conditions is met by the target distribution. The strength of the contribution is that we show that the Lipschitz continuity can also affect flows when the support are not disconnected. However, the limitation of the Theorems is that it can be hard in practice to check the density/sparsity properties of a high dimensional distribution.

Outline of the paper The rest of this paper is organized as follows. Section 2 reviews normalizing flows as well as total variation distance and the precision/recall for generative networks. The main results are presented in Section 3: a first general bound is presented in the Section 3.1, the two main theorems are presented and proved in the Section 3.2. We draw a link between the main results and the previous related work in the Section 3.3. In Section 4, we make a expreimental analysis of the bounds. Finally we discuss the potential remedies to the highlighted limitations in Section 5, and Section 6 concludes the paper and gives some directions of future works. In the Supplementary Material, formal proofs of the different results are presented.

2. Background

2.1. Normalizing Flow

A *normalizing flow* is an invertible density model in which both density estimation and sampling can be done efficiently. In short, training a normalizing flow consists in learning

an invertible mapping between a data space $\mathcal{X}$ and a latent space $\mathcal{Z}$. Typically, the forward direction $F : \mathcal{X} \to \mathcal{Z}$ (i.e. the *normalizing* direction) is tractable and exact and the inverse direction $F^{-1} : \mathcal{Z} \to \mathcal{X}$ (i.e. the *generative* direction) either has a closed form, or can be approximated using an iterative algorithm.

Suppose that P^* is the true data distribution over $\mathcal{X}$, and that P^* admits a density function denoted p^* that we wish to approximate. We first choose a d-dimensional Gaussian distribution Q over $\mathcal{Z}$ (a.k.a. the latent space), and its density function $q(\boldsymbol{z}) = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{1}{2}\|\boldsymbol{z}\|_2^2}$. The choice of a normal distribution is natural since it is the most frequent latent distribution in normalizing flows. In Section 5, more complex distributions will be discussed. Then, we can define $\hat{p}$, the approximation of p^*, based on q and the mapping $F : \mathcal{X} \to \mathcal{Z}$, using a simple change of variable formula:

$$\forall \boldsymbol{x} \in \mathcal{X}, \quad \hat{p}(\boldsymbol{x}) = |\det \mathrm{Jac}_F(\boldsymbol{x})|\ q(F(\boldsymbol{x})). \tag{1}$$

Note that the estimated probability $\widehat{P}(A)$ of any event $A \subseteq \mathcal{X}$ can be computed as follows:

$$\widehat{P}(A) = Q(F(A)) = \int_{F(A)} q(\boldsymbol{z}) d\boldsymbol{z}.$$

As seen in Equation 1, performing density estimation requires computing the determinant of the Jacobian matrix which can be large in practice, thus most normalizing flows have been specifically designed to make this computation efficient.

2.2. Bi-Lipschitz Normalizing Flows

In this paper, we focus on bi-Lipschitz normalizing flows, which is a mapping F whose Lipschitz constants are bounded in both directions. More specifically, we define the bi-Lipschitz property as follows.

Definition 1 *A bijective function $F : \mathcal{X} \subset \mathbb{R}^d \to \mathcal{Z} \subset \mathbb{R}^d$ is said to be (L_1, L_2)-bi-Lipschitz if F is L_1-Lipschitz and its inverse F^{-1} is L_2-Lipschitz, i.e.:*

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \quad \|F(\boldsymbol{x}_1) - F(\boldsymbol{x}_2)\|_2 \leq L_1 \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2,$$

and

$$\forall \boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}, \quad \|F^{-1}(\boldsymbol{z}_1) - F^{-1}(\boldsymbol{z}_2)\|_2 \leq L_2 \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2.$$

Alternatively, since the mapping F is bijective, the bi-Lipschitz continuity can be expressed over F only as:

$$\frac{1}{L_2}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leq \|F(\boldsymbol{x}_1) - F(\boldsymbol{x}_2)\|_2 \leq L_1 \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2.$$

However, enforcing the bi-Lipschitz continuity of F results in a bounded determinant for the Jacobian matrix:

Proposition 2 Jac_F *satisfies for all $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$:*

$$\frac{1}{L_2^d} \leq |\det\ \mathrm{Jac}_F(\boldsymbol{x})| \leq L_1^d.$$

Proposition 2 comes from a characterization of the Lipschitz continuity adapted to differentiable functions (Federer, 1969):

$$\forall \boldsymbol{x} \in \mathcal{X},\ \|\mathrm{Jac}_F(\boldsymbol{x})\|_2 \leq L_1.$$

Therefore, through the spectral decomposition of the jacobian matrix, we can show that $\forall \boldsymbol{x} \in \mathcal{X},\ |\det \mathrm{Jac}_F(\boldsymbol{x})| \leq L_1^d$. Then, since F is bi-Lipschitz, the same inequality can be expressed for F^{-1}: $\forall \boldsymbol{z} \in \mathcal{Z},\ |\det \mathrm{Jac}_{F^{-1}}(\boldsymbol{z})| \leq L_2^d$ and thus $\forall \boldsymbol{x} \in \mathcal{X},\ |\det \mathrm{Jac}_F(x)| \geq 1/L_2^d$.

As we will show in the rest of this paper, this can limit the expressivity of normalizing flows. This is relevant, because many normalizing flows are bi-Lipschitz in practice, for example, the i-ResNet (Behrmann et al., 2019) and the Residual Flow (Chen et al., 2020) are both based on residual atomic blocks $f_i = I_d + g_i$. Their invertibility is ensured by the Lipschitz constant $\mathrm{Lip}(g_i) \leq L < 1$. If F is composed of m residual blocks such that $F = f_m \circ \cdots \circ f_1$, then the overall bi-Lipschitz constants satisfy $\mathrm{Lip}(F) \leq (1+L)^m$ and $\mathrm{Lip}(F^{-1}) \leq 1/(1-L)^m$. Alternatively, in Glow (Kingma and Dhariwal, 2018) with atomic blocks $W_i = P_i L_i (U_i + \mathrm{diag}(s_i))$, the bi-Lipschitz constants satisfy: $\mathrm{Lip}(F) \leq \prod_i^m \|W_i\|_2$ and $\mathrm{Lip}(F^{-1}) \leq \prod_i^m \|W_i^{-1}\|_2$.

Notice that the bi-Lipschitzness constraints on either the function or its Jacobian determinant can frequently be weakened by increasing the depth of the network but, by doing so, the stability of the inverse can be affected (Behrmann et al., 2021).

2.3. Assessing the expressivity

Our goal is to understand how the bi-Lipschitz property affects the approximation ability of the network. To do so, we will compare the true data distribution P^* and its density p^* with the learned distribution $\widehat{P}$ and its density $\hat{p}$.

In previous works, Tanielian et al. (2020) use the maximum precision to evaluate how the true distribution P^* and the generated distribution $\widehat{P}$ differs. We have chosen another tool to compare both distributions: the total variation (TV) distance given as :

Definition 3 (Total Variation Distance) *For any distribution $\widehat{P}$ and P^*, the total variation distance is defined as the maximum difference of probabilities given to a same event A:*

$$\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) = \sup_A |P^*(A) - \widehat{P}(A)|.$$

This choice has been made for several reasons. It is adequate to highlight the Lipschitz constraints of the mapping. It has a close connection with the precision and the recall, and yet, the TV is more general in terms of support of target distribution as explained in the followings. Finally, the TV can be used to compute a lower bound on the Kullback-Leibler divergence $\mathcal{D}_{\mathrm{KL}}$ through the Pinsker's inequality:

$$2\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P})^2 \leq \mathcal{D}_{\mathrm{KL}}(P^* \| \widehat{P}).$$

Even if the main results of this paper are consisting in lower bounds on the TV distance, we translate of result in terms of precision and recall. Therefore we provide their definitions as they were given initially (Sajjadi et al., 2018; Kynkäänniemi et al., 2019).

Definition 4 (Precision α and Recall β for generative models) *For $\alpha, \beta \in [0,1]$, the distributions $\widehat{P}$ is said to have a precision α at recall β with respect to P^* if there exist the distributions $\nu, \hat{\nu}, \nu^*$, such that $\widehat{P}$ and P^* can be decomposed as such:*

$$\widehat{P} = \alpha\nu + (1-\alpha)\hat{\nu} \quad \textit{and} \quad P^* = \beta\nu + (1-\beta)\nu^*.$$

The distribution ν defined on $\mathrm{Supp}(\widehat{P}) \cup \mathrm{Supp}(P^*)$ while $\mathrm{Supp}(\hat{\nu}) = \mathrm{Supp}(\widehat{P})$ and $\mathrm{Supp}(P^*) = \mathrm{Supp}(\nu^*)$

It can be interpreted as such: ν represent the part of P^* that $\widehat{P}$ correctly models, $\hat{\nu}$ is simultaneously the part of P^* that $\widehat{P}$ misses on their joint support and all the points that should not be represented by $\widehat{P}$. Finally, ν^* cover the points of P^* that the support of $\widehat{P}$ could not reach and all the points on their joint support that $\widehat{P}$ misestimated.

Among all the potential decompositions, i.e. the pairs (α, β), the focus is set on the maximum precision $\bar{\alpha}$ and the maximum recall $\bar{\beta}$.

Proposition 5 (Maximum precision $\bar{\alpha}$ and maximum recall $\bar{\beta}$) *The maximum precision and the maximum recall satisfy:*

$$\bar{\alpha} = \widehat{P}(\mathrm{Supp}(P^*)) \quad \textit{and} \quad \bar{\beta} = P^*(\mathrm{Supp}(\widehat{P})).$$

The results given by Tanielian et al. (2020) is an upper bound on the maximum precision for a Lipschitz F^{-1} and for a particular target distribution. Having upper bounds on $\bar{\alpha}$ or $\bar{\beta}$ is stronger than lower bounds on the $\mathcal{D}_{\mathrm{TV}}$ since:

$$\begin{aligned}\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq& |P^*(\mathrm{Supp}(P^*)) - \widehat{P}(\mathrm{Supp}(P^*))|,\\ \mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq& |P^*(\mathrm{Supp}(\widehat{P})) - \widehat{P}(\mathrm{Supp}(\widehat{P}))|.\end{aligned}$$

Or equivalently : $\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq \max\left(1-\bar{\alpha}, 1-\bar{\beta}\right)$

However, as soon as the support of, respectively, the target distribution P^* or the estimated distribution $\widehat{P}$ covers $\mathcal{X}$, we have respectively $\bar{\alpha} = 1$ or $\bar{\beta} = 1$. Thus, the maximum precision/recall become irrelevant for assessing the expressivity of the normalizing flow.

3. Lower bounds on the TV distance

The general idea is to look for subsets of the dataset in the data space $\mathcal{X}$ that may be particularly difficult to fit with a Lipschitz constrained mapping function. Intuitively, the Lipschitz constraints limit the ability of normalizing flows to contract or to expand the latent space, so we focus our analysis on very dense subsets or very sparse subsets of the data space that will likely be the most difficult to fit.

We focus first on dense subsets with arbitrary shape, and we are able to derive a positive lower bound on the TV that depends on the volume of the largest dense subset. Then, we show how to compute more specific but stronger results when considering subsets with a ball shape. First an intuitive but loose bound is derived. We then discuss two new tight bounds that are based on dense and sparse ball shaped subsets of the data space.

3.1. A first bound based on the most dense subset of $\mathcal{X}$

The first theorem is a lower bound on the TV distance between the learned distribution and the target distribution in a general setting. Intuitively, the idea is to find a subset A with an arbitrary shape that is sufficiently concentrated so that the Lipschitz constrained mapping can not concentrate enough weight from the Gaussian distribution onto this subset.

Theorem 6 (L_1-Lipschitz mappings fail to capture high density subset) *Let F be L_1-Lipschitz and $\eta_A = \frac{P^*(A)}{\text{vol}(A)}$ be the average density over any subset $A \subset \mathbb{R}^d$. Then:*

$$\mathcal{D}_{\text{TV}}(P^*, \widehat{P}) \geq \sup_A \text{vol}(A) \left(\eta_A - \left(\frac{L_1}{\sqrt{2\pi}} \right)^d \right).$$

Therefore, if there is a subset A that satisfies $\eta_A > \left(\frac{L_1}{\sqrt{2\pi}}\right)^d$, then the TV is necessarily strictly positive.

Theorem 6 results from the definition of the estimated distribution $\widehat{P}$ and the change of variable. Indeed for an arbitrary subset A of $\mathcal{X}$:

$$\widehat{P}(A) = \int_A \hat{p}(\boldsymbol{x}) d\boldsymbol{x} = \int_A |\text{Jac}_F(\boldsymbol{x})| q(F(\boldsymbol{x})) d\boldsymbol{x}.$$

Then, since q is the density function of the normal distribution it is upper bounded by $1/\sqrt{2\pi}^d$, and with the upper bound of the determinant of the jacobian matrix given in Proposition 2 :

$$\widehat{P}(A) \leq \left(\frac{L_1}{\sqrt{2\pi}} \right)^d \int_A d\boldsymbol{x} = \left(\frac{L_1}{\sqrt{2\pi}} \right)^d \text{vol}(A).$$

In other terms, the weight assigned from the Gaussian latent distribution to the subset A is bounded by $\left(L_1/\sqrt{2\pi}\right)^d \text{vol}(A)$. Consequently if there is a dense subset for which $P^*(A) = \eta_A \text{vol}(A)$ is high enough, the TV will be strictly positive. More precisely, the total variation is greater than the difference made by the most dense subset A. The detailed proof of Theorem 6 is given in Appendix A.1.

The main advantage of this formulation is to be applied to any subset of the data space, but at the expense of a loose bound on the TV.

3.2. Bounds based on dense and sparse balls

The bound in Theorem 6 can be further improved by making assumptions on the structure of the subset A. We choose to focus on l_2 balls instead of arbitrary subsets.

Let $B_{R,\boldsymbol{x}_0}$ be the l_2 ball with center $\boldsymbol{x}_0$ and radius R (i.e. $B_{R,\boldsymbol{x}_0} = \{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq R\}$). Then we can show that both high density balls and low density ones are difficult to fit properly, the former because of the Lipschitz constraint of F, the latter because of the Lipschitz constraint of F^{-1}.

We first consider high density balls.

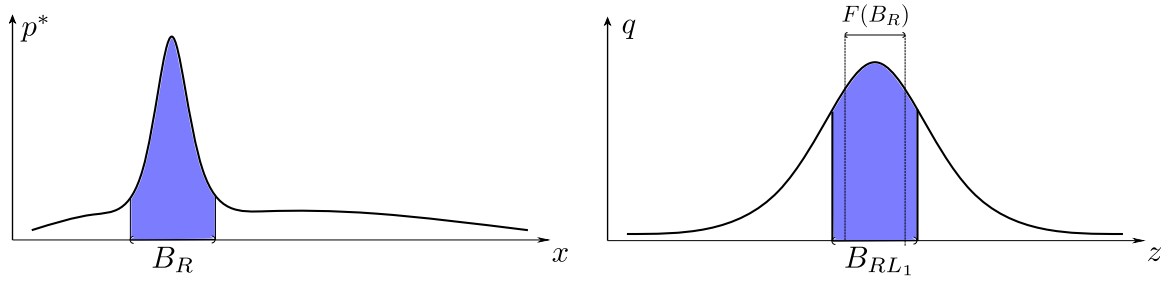

Figure 1: Example of a target distribution where theorem 7 applies: the subset B_R concentrates most of the weight in P^*, but $\widehat{P}(B_R) = Q(F(B_R))$ can only be as large as $Q(B_{RL_1})$.

Theorem 7 (NF with a L_1-Lipschitz mapping F fails to capture high density balls) *Let F be L_1-Lipschitz. Then:*

$$\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq \sup_{R, \boldsymbol{x}_0} \left(P^*(B_{R,\boldsymbol{x}_0}) - \frac{RL_1}{\sqrt{\pi}} \right).$$

Therefore, if we find a ball for which the true measure satisfies $\frac{P^*(B_{R,\boldsymbol{x}_0})}{R} > \frac{L_1}{\sqrt{\pi}}$, then the TV is necessarily strictly positive.

Theorem 7 highlights the effect of the L_1 Lipschitz constraint of the forward mapping F. The image of a ball B_R by the mapping F is constrained in a ball:

$$F(B_R) \subset B_{L_1 R}.$$

Thus, if we consider a ball with a high probability $P^*(B_R)$ in the data space, then the probability assigned to this ball $\hat{P}(B_R) = Q(F(B_R))$ is at most $Q(B_{RL_1})$ in the latent space and is upper bounded by $RL_1/\sqrt{\pi}$ (Ball, 1993). The lower bound of the TV in Theorem 7 is given with a linear relation with radius R and the Lipschitz constant L_1 in dimension d. It serves as an intuitive representation of the limitations induced by the Lipschitz constraint. For a given dense ball $B_{R,\boldsymbol{x}_0}$, the smaller R is, the greater L_1 should be to insure that the normalizing flow can properly map the Gaussian distribution onto P^*. The bounds mainly serve for interpretation purpose as we can compute a tighter bound on the TV distance. The closed form of Gaussian measure of a ball $B_{R,\mathbf{0}}$ is given by a function of the Gamma function Γ and the incomplete gamma function γ, therefore:

$$Q(B_{RL_1,\boldsymbol{x}_0}) \leq Q(B_{RL_1,\mathbf{0}}) = \gamma\left(\frac{d}{2}, \frac{L_1^2 R^2}{2}\right) / \Gamma\left(\frac{d}{2}\right).$$

The Theorem 8 is then less easy to interpret than Theorem 7 but it proposes a tighter bound.

Theorem 8 (NF with a L_1-Lipschitz mapping F fails to capture high density balls) *Let F be L_1-Lipschitz. Then:*

$$\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq \sup_{R, \boldsymbol{x}_0} \left(P^*(B_{R,\boldsymbol{x}_0}) - \frac{\gamma\left(\frac{d}{2}, \frac{L_1^2 R^2}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \right).$$

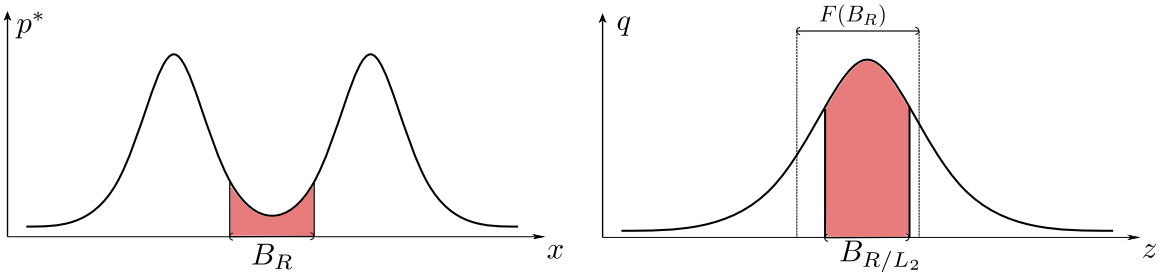

Figure 2: Example of a target distribution for which Theorem 9 applies: the subset B_R concentrates little weight in P^*, but $\widehat{P}(B_R) = Q(F(B_R))$ can only be as small as $Q(B_R/L_2)$.

Therefore, if we find a ball for which the true measure satisfies $P^*(B_{R,\boldsymbol{x}_0}) > \gamma(\frac{d}{2}, \frac{L_1^2 R^2}{2})/\Gamma(\frac{d}{2})$, *then the TV is necessarily strictly positive.*

A one dimensional representation of a pathological case for Theorems 7 & 8 is shown on Figure 1. In other words no ball with a density high enough in the data space can be expanded sufficiently to have a matching probability in the latent space.

Conversely, the mapping being bi-Lipschitz, it can not contract arbitrarily:

$$B_{R/L_2} \subset F(B_R).$$

If there is a low density zone mapped on the maximum of the Gaussian density, then the Normalizing Flow cannot reduce enough the probability of the corresponding zone:

$$\widehat{P}(B_R) \geq \widehat{P}(F^{-1}(B_R)) = Q(B_{R/L_2,0}).$$

Notice that the assumption of a low density zone is strong but fairly reasonable. For instance, one can observe a multi-modal density with fairly well separated modes. If the modes are roughly equiprobable, we expect a mapping to assign those modes in balanced way around the mode of the Gaussian distribution in the latent space. Therefore, the low density ball is mapped on a zone wider than the ball B_{R/L_2} and consequently the Gaussian measure associated is lower bounded by $Q(B_{R/L_2})$ as illustrated in the one dimensional example on Figure 2. Despite the lower bounds established by Pinelis (2020), there is no reasonably interpretable bounds, therefore we use the closed-form as in Theorem 8.

Theorem 9 (NF with L_2-Lipschitz inverse mappings F^{-1} fail to capture low density balls) *Let F^{-1} be L_2-Lipschitz. We consider the balls centered on $F^{-1}(0)$, we have the lower bound:*

$$\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq \sup_R \left(\frac{\gamma\left(\frac{d}{2}, \frac{R^2}{2L_2^2}\right)}{\Gamma\left(\frac{d}{2}\right)} - P^*(B_{R,F^{-1}(0)}) \right).$$

Therefore, if we find a ball for which the true measure satisfies $P^*(B_{R,F^{-1}(0)}) < \frac{\gamma(d/2, R^2/2L_2^2)}{\Gamma(d/2)}$, *then the TV is necessarily strictly positive.*

All formal proofs are detailed in Appendix A.2 & A.4. Numerical illustrations of the behavior of the closed form bound are given in Figure 3. Note that for Theorem 8, the probability

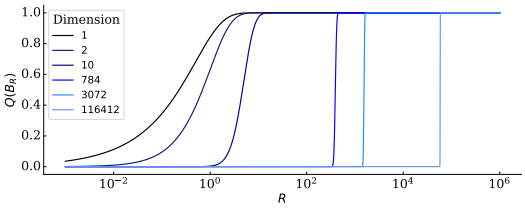

Figure 3: Representation of the Gaussian Measure of balls of radius R centered on $\mathbf{0}$. The measure is given for dimension 1, 2, 10 and then the dimensions of *MNIST* (Yann LeCun et al., 2010), *CIFAR10* (Alex Krizhevsky, 2009) and *CelebA* (Liu et al., 2015)

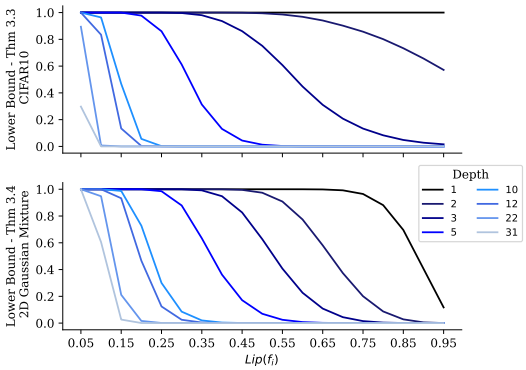

Figure 4: Lower bound of Theorem 8 for *CIFAR10* (top) and Theorem 9 for a 2D *Gaussian mixture* of Figure 5 (bottom).

$Q(B_{RL_1})$ needs to be as large as possible which, given a radius RL_1, will be harder while the dimension increases. In other terms, the Lipschitz constant L_1 will have to be consequently increased more in high dimension than in low dimension. The behavior is reversed for L_2: from Theorem 9, we see that the probability $Q(B_{R/L_2})$ should be as small as possible. Then, given a radius R/L_2, it will be harder in low dimension dimension than in high dimension. The Lipschitz constant L_2 will have a greater effect in low dimension than in high dimension as we will see in Section 4.2

3.3. Comparison to related work

A related set up is used in the work of Tanielian et al. (2020). The authors consider two disconnected subsets M_1 and M_2 separated by a distance D, with equal probabilities in the latent space, i.e. $\widehat{P}(M_1) = \widehat{P}(M_1) = 1/2$. As a consequence, $F^{-1}(0)$ is equidistant from M_1 and M_2 as illustrated in Figure 5. The original work assesses the learning abilities of their generative model, a L_2-Lipschitz GAN (Goodfellow et al., 2014; Arjovsky et al., 2017), with a definition of precision and recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019). The authors propose an upper bound of the maximum recall based on the cumulative distribution function of the 1-dimensional normal distribution $\Phi(t) = \int_{-\infty}^{t} \frac{\exp(-r^2/2)}{2\pi} dr$:

$$\bar{\alpha} + \frac{2D}{L_2} e^{-\Phi^{-1}(\bar{\alpha}/2)^2} \leq 1.$$

Our method offers another bound on the maximum precision:

$$\bar{\alpha} \leq 1 - \frac{\gamma\left(\frac{d}{2}, \frac{D^2}{2L_2^2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

The main advantage of our bound is that it can be directly computed whereas the bound given by Tanielian et al. (2020) is not explicit. The advantage of their bound is that it does not depend on the dimension. The detailed proof can be found in Appendix A.5. From the link between the divergence and the maximum precision, we can derive a lower bound on the TV as a particular case of Theorem 9:

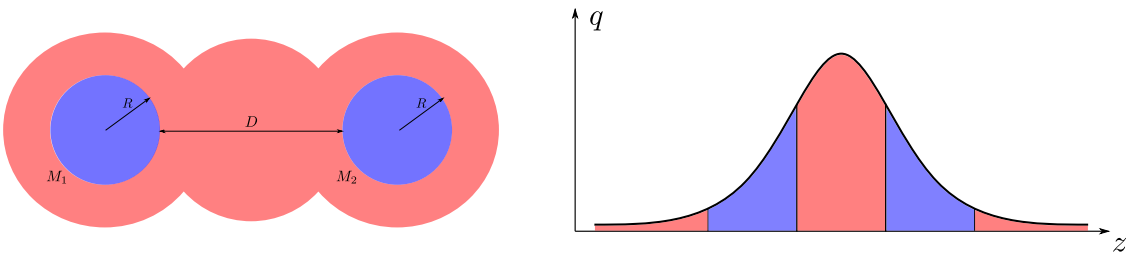

Figure 5: Experimental set up given by Tanielian et al. (2020)

Corollary 10 (NF with L_2-Lipschitz inverse mapping) *If F^{-1} is L_2-Lipschitz, then we have a lower bound on the TV distance based on the distance D between M_1 and M_2:*

$$\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq \gamma\left(\frac{d}{2}, \frac{D^2}{2L_2^2}\right) / \Gamma\left(\frac{d}{2}\right).$$

The main benefit of using maximum precision to assess the quality of the mapping is that it is well fitted to be used with the Gaussian Isoperimetric Inequality and therefore gives results that do not depend on the dimension d. The benefit of using TV distance is that we can now compute the bounds for distributions with arbitrary support. This covers more cases that the work by Cornish et al. (2021) which only discusses the limits of the normalizing flows when the supports of the two distributions are *not* homeomorphic. Instead, the bounds we introduce in the present paper can be used to discuss the limitations of normalizing flows, even when dealing with two distributions with homeomorphic supports. Let us consider for instance, the case where a normalizing flow tries to map a 1-dimensional Gaussian distribution with an excessively low or high variance onto the standard normal distribution: the supports are homeomorphic, the maximum precision and recall are both 1, but our method can be used to derive a strictly positive lower bound of the TV.

4. Experiments

Informally, our objective is to discover how deep and how constrained a network should be to fit a given dataset. To do so, we can compute the bounds for given theoretical Lipschitz constants. If the lower bounds are greater than zero, then the Lipschitz property will be limitating, and the settings should be adjusted.

We focus on the Residual Flow (Chen et al., 2020), for which the theoretical Lipschitz constant derivation is straightforward. Two parameters will affect the two global Lipschitz constants L_1 and L_2 : the depth m and the Lipschitz constant of the layers $L = \mathrm{Lip}(f_i) < 1$ with $L_1 \leq (1+L)^m$ and $L_2 \leq 1/(1-L)^m$. We compute the bounds for the two one-dimensional examples from Figure 1 and 2, for the pathological case from Tanielian et al. (2020) in Figure 5, for the *8 Gaussians* dataset and the *Circles* dataset shown in Figure B.1 in Appendix and finally, for *MNIST* (Yann LeCun et al., 2010) and *CIFAR10* (Alex Krizhevsky, 2009).

4.1. Bound of Theorem 8

The lower bound on the TV is based on the supremum over every ball in the dataset. To enumerate every candidate ball and find the supremum, we first consider every example x in the dataset as a candidate center and the distance to every other points in the dataset as a candidate radius R. Then we can compute an approximation of the true measure of the ball $P^*(B_{R,x})$ with the empirical measure (Kloeckner, 2018). The results are presented in Figure 4 and 6.

In Figure 4 (top), we can observe that if the network has a Lipschitz constant $L < 0.5$ (usually required for stable training) it should be at least 10 layers deep. In contrast, a shallow network (1, 2 or 3 layers) with high Lipschitz constant will fail to capture the target distribution.

In Figure 6 we can see that for *1D Gaussian*, the Residual Flow should be at least 7 layers deep with an high atomic lipschitz constant of 0.95. For *2D Circles*, the inner circle is too dense to be properly mapped by the network. Since the inner circle represents half of the datapoints, the bound is at most 0.5. Finally, we can see that the bound is limitating even on the real world datasets *MNIST* and *CIFAR10*. Moreover, for shallow and highly constrained networks, the support of the learned distribution will not even intersect the support of the target distribution, as a consequence, TV (i.e. approximation error) will maximum. Instead, when the depth or the atomic Lipschitz constant are increased, it results in a greater global constant and therefore in a reduction of the lower bound down to 0.

4.2. Bound of Theorem 9

The lower bound in Theorem 9 relies on the supremum computed over all the balls centered on $F^{-1}(0)$, which can only be computed after the normalizing flow has been trained. Thus the bound cannot be used to adjust the training parameters a priori as we did in the previous section. However, the center of the probability mass is an intrinsic property of the dataset and we observe in practice that $F^{-1}(0)$ is often mapped onto the same point in the data space. For example, all the normalizing flows trained on the *8 Gaussians* dataset will map the center of the Gaussian in the latent space onto the center of the 8 Gaussians in the data space. Building on this observation, we can compute $F^{-1}(0)$ analytically or train a first normalizing flow to find an estimate of $F^{-1}(0)$, and then compute the bound and adjust the parameters of the normalizing flow based on the instantiated bound. For the one or two dimensional datasets, $F^{-1}(0)$ is clear, for *MNIST* and *CIFAR10*, $F^{-1}(0)$ is not obvious and can be seen in Figure C.2 and C.3 from the Appendix C.

The value of the bound computed for the different datasets are shown in Figure 4 and 7. As we can see, this bound does not highlight any limitation for *MNIST* or *CIFAR10*, (as discussed in Section 3.2. However, as we will see in Section 5, one potential remedies introduces a tradeoff between the two bounds, and this bound may end up being limitating when it is used together with the other bounds. For the other two datasets, we can observe that for stable configuration with low L, the network needs to be at least 15 layers deep to learn the pathological case illustrated in Figure 5.

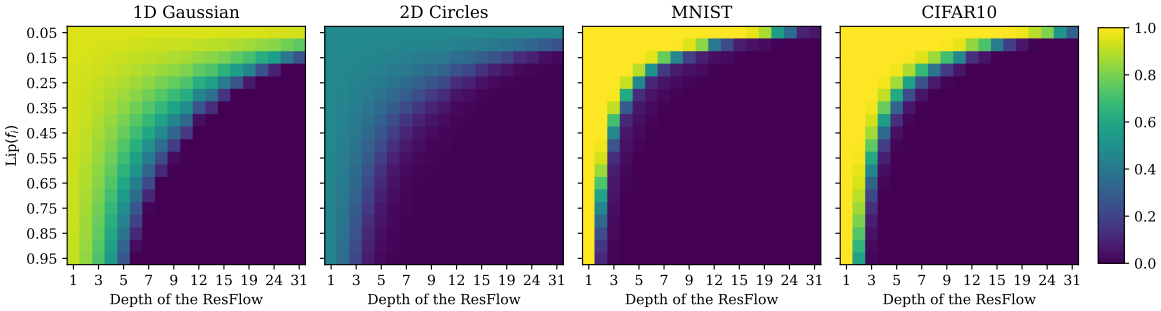

Figure 6: Empirical values of the lower bound on the TV based on Theorem 8.

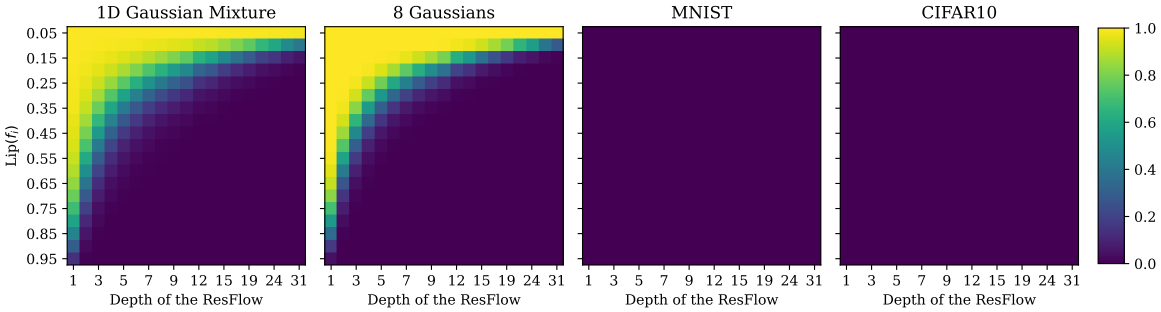

Figure 7: Empirical values of the lower bound on the TV based on Theorem 9.

5. Discussion on the potential remedies

As mentioned earlier, increasing the Lipschitz constants of the entire network (for example, by adding extra layers) may impact invertibility and stability during training (Behrmann et al., 2021), and thus is not a suitable approach to improve the expressivity.

Alternatively, one can consider learning the parameters of the latent Gaussian distribution $\boldsymbol{\mu}$ and $\Sigma = \sigma^2 I_d$. However, this is equivalent to changing the Lipschitz constants of F from (L_1, L_2) to $(\frac{L_1}{\sigma}, L_2\sigma)$, thus this results in trading off the expected error on very dense subsets (Theorem 8) with the expected error on subsets with low densities (Theorems 9) or vice-versa. A proof of this statement is given in Appendix A.6. Increasing, reducing or learning the variance will indeed increase one bound and decrease the other one. In other words this can lead to a better approximation for some particular data distributions, but it does not generally improve the expressivity of the normalizing flow.

To improve expressivity, a Gaussian Mixture latent distribution can also be considered. Indeed, Khayatkhoei et al. (2019) and Izmailov et al. (2019) have shown that such distributions can learn disconnected manifolds. When the latent distribution is a Gaussian Mixture, Theorem 9 does not hold anymore. Limitations similar to the ones highlighted in Theorem 7 still apply, but can be mitigated using learnable parameters.

We can trivially adapt the lower bound from Theorem 7 to the Gaussian Mixture with K equally distributed modes with learnable means $\boldsymbol{\mu}_i$ and covariance matrices $\Sigma_i = \sigma_i^2 I_d$:

$$\mathcal{D}_{\mathrm{TV}}(P^*, \widehat{P}) \geq \sup_{R, \boldsymbol{x}_0} \left(P^*(B_{R,\boldsymbol{x}_0}) - \frac{1}{K} \frac{\gamma\left(\frac{d}{2}, \frac{L_1^2 R^2}{2\sigma_i^2}\right)}{\Gamma\left(\frac{d}{2}\right)} \right).$$

As we can see here, the lower bound depends on the inverse of the number of modes K in the mixture and the variance σ_i. Thus, this approach can solve the limitations highlighted in Theorem 8. The lower bound increases with K but the learnable σ_i can compensate this augmentation. However, learning a Gaussian Mixture suffers from the same issues than another method that we would like to mention: the Variational Mixture of Normalizing Flow (Pires and Figueiredo, 2020). They train set of K different normalizing flows and a neural network with a K-class softmax output to set the mixture. They encountered strong training difficulties to learn the mixture and, as for the Gaussian Mixture, there is no efficient way yet to learn the hyperparameter K (Izmailov et al., 2019). A closely related method by Dinh et al. (2020) consists in a discrete partition of the data space $\mathcal{X}$ into K distinct subsets. K different normalizing flows would be trained on each of those partitions. This methods also suffers from training difficulties and the lack of evidence on how to set the number of partitions. To avoid a discrete set of normalizing flows, (Cornish et al., 2021) propose a continuous set on normalizing flows, a promising method to tackle the limitations highlighted by Theorem 9, but it does not preserve an exact likelihood and requires a complex training process.

Overall, there exist some theoretical potential remedies to both of the limitations highlighted by Theorems 8 & 9 but further investigation is required to to deal with the technical issues.

6. Conclusion

We have established that the bi-Lipschitz constraints reduce the expressivity of Normalizing flows. When the dataset meets some particular conditions such as a high density zone or a low density zone between two high density zones, the reduced expressivity fails to capture the real distribution of the dataset. More specifically, we have brought to light two particular lower bounds of the total variation distance between the target distribution and the approximated one. The first bound illustrates that dense subset, and especially dense balls in the data space can induce high approximation errors. The second bound illustrates that some low density balls located between high density subsets can also result in approximations errors. Finally, we showed that a more complex latent distribution, such as a Gaussian Mixtures could solve the highlighted limitations.

Moreover, Theorems 8 & 9 are based on the bi-Lipschitz constraints of the network and on general properties met by the dataset. Some similar bounds could be derived for non invertible structures that satisfy local bi-Lipschitzness. A possible direction for future work is to generalize, not only the results, but also the framework to study the expressivity of generative lipschitz models based on the dataset.

Independently of the generative models, close attention is given to the effect of Lipschitz regularization on the stability of a neural networks (Scaman and Virmaux, 2019; Combettes and Pesquet, 2020; Béthune et al., 2021), on the generalization capability of the network (Bartlett et al., 2017) or on its adversarial robustness (Szegedy et al., 2014; Araujo et al., 2020). The theoretical framework presented in this work could be transposed to more general applications linked to Lipschitz regularization.

References

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Alexandre Araujo, Benjamin Negrevergne, Yann Chevaleyre, and Jamal Atif. On Lipschitz Regularization of Convolutional Layers using Toeplitz Matrix Theory. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, November 2020. URL http://arxiv.org/abs/2006.08391. arXiv: 2006.08391.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia,*, December 2017. URL http://arxiv.org/abs/1701.07875. arXiv: 1701.07875.

Keith Ball. The reverse isoperimetric problem for Gaussian measure. *Discrete & Computational Geometry volume*, 10(4):411–420, December 1993. ISSN 1432-0444. URL https://doi.org/10.1007/BF02573986.

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *30th Conference on Neural Information Processing Systems (NeurIPS 2017)*, December 2017. arXiv: 1706.08498.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible Residual Networks. In *Proceedings of the 36 th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019*, May 2019. arXiv: 1811.00995.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Joern-Henrik Jacobsen. Understanding and Mitigating Exploding Inverses in Invertible Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1800. PMLR, March 2021. URL http://proceedings.mlr.press/v130/behrmann21a.html.

Louis Béthune, Alberto González-Sanz, Franck Mamalet, and Mathieu Serrurier. The Many Faces of 1-Lipschitz Neural Networks. *arXiv:2104.05097 [cs, stat]*, May 2021. URL http://arxiv.org/abs/2104.05097. arXiv: 2104.05097.

Ricky T. Q. Chen, Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Residual Flows for Invertible Generative Modeling. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.*, July 2020. arXiv: 1906.02735.

Patrick L. Combettes and Jean-Christophe Pesquet. Lipschitz Certificates for Layered Network Structures Driven by Averaged Activation Operators. *SIAM Jour- nal on Mathematics of Data Science, 2020.*, June 2020. URL http://arxiv.org/abs/1903.01014. arXiv: 1903.01014.

Rob Cornish, Anthony L. Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows. In *Proceedings of the 37th International Conference on Machine Learning*, April 2021. URL http://arxiv.org/abs/1909.13833. arXiv: 1909.13833.

Laurent Dinh, Jascha Sohl-Dickstein, Hugo Larochelle, and Razvan Pascanu. A RAD approach to deep mixture models. In *ICLR 2019 Workshop DeepGenStruct*, August 2020. URL http://arxiv.org/abs/1903.07714. arXiv: 1903.07714.

Farzan Farnia, Jesse Zhang, and David Tse. Generalizable Adversarial Training via Spectral Normalization. In *ICLR 2019*, September 2018. URL https://openreview.net/forum?id=Hyx4knR9Ym.

Herbert Federer. *Geometric measure theory*. Berlin, heidelberg, new york, springer edition, 1969.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *27th Conference on Neural Information Processing Systems (NeurIPS 2014)*, June 2014. URL http://arxiv.org/abs/1406.2661. arXiv: 1406.2661.

Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-Supervised Learning with Normalizing Flows. *arXiv:1912.13025 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1912.13025. arXiv: 1912.13025.

Mahyar Khayatkhoei, Ahmed Elgammal, and Maneesh Singh. Disconnected Manifold Learning for Generative Adversarial Networks. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*, January 2019. URL http://arxiv.org/abs/1806.00880.

Durk P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.*, volume 31, 2018.

Benoît Kloeckner. Empirical measures: regularity is a counter-curse to dimensionality. *arXiv:1802.04038 [math]*, February 2018. URL http://arxiv.org/abs/1802.04038. arXiv: 1802.04038.

Frederic Koehler, Viraj Mehta, and Andrej Risteski. Representational aspects of depth and conditioning in normalizing flows. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5628–5636. PMLR, July 2021. URL https://proceedings.mlr.press/v139/koehler21a.html. ISSN: 2640-3498.

Zhifeng Kong and Kamalika Chaudhuri. The Expressive Power of a Class of Normalizing Flow Models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 3599–3609. PMLR, June 2020. URL https://proceedings.mlr.press/v108/kong20a.html. ISSN: 2640-3498.

Zhifeng Kong and Kamalika Chaudhuri. Universal Approximation of Residual Flows in Maximum Mean Discrepancy. arXiv, June 2021. URL http://arxiv.org/abs/2103.05793. Number: arXiv:2103.05793 arXiv:2103.05793 [cs, stat].

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *33rd Conference*

on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada., October 2019. arXiv: 1904.06991.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. In *ICLR 2019*, February 2018. URL http://arxiv.org/abs/1802.05957. arXiv: 1802.05957.

Iosif Pinelis. Exact lower and upper bounds on the incomplete gamma function. *Mathematical Inequalities & Applications*, pages 1261–1278, 2020. doi: 10.7153/mia-2020-23-95.

Guilherme G. P. Freitas Pires and Mário A. T. Figueiredo. Variational Mixture of Normalizing Flows. *arXiv:2009.00585 [cs, stat]*, September 2020. URL http://arxiv.org/abs/2009.00585. arXiv: 2009.00585.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing Generative Models via Precision and Recall. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*, October 2018. URL http://arxiv.org/abs/1806.00035. arXiv: 1806.00035.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.*, October 2019. URL http://arxiv.org/abs/1805.10965. arXiv: 1805.10965.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Represen- tations, 2014.*, February 2014. URL http://arxiv.org/abs/1312.6199. arXiv: 1312.6199.

Ugo Tanielian, Thibaut Issenhuth, Elvis Dohmatob, and Jeremie Mary. Learning disconnected manifolds: a no GANs land. In *Proceedings of the 37 th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020*, December 2020. arXiv: 2006.04596.

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators. In *Advances in Neural Information Processing Systems*, volume 33, pages 3362–3373. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2290a7385ed77cc5592dc2153229f082-Abstract.html.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs*, 2, 2010. URL http://yann.lecun.com/exdb/mnist.

Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation Capabilities of Neural ODEs and Invertible Residual Networks. arXiv, February 2020. URL http://arxiv.org/abs/1907.12998. Number: arXiv:1907.12998 arXiv:1907.12998 [cs, stat].