



HAL
open science

On the expressivity of bi-Lipschitz normalizing flows

Alexandre Vérine, Benjamin Negrevergne, Fabrice Rossi, Yann Chevaleyre

► **To cite this version:**

Alexandre Vérine, Benjamin Negrevergne, Fabrice Rossi, Yann Chevaleyre. On the expressivity of bi-Lipschitz normalizing flows. ACML 2022 - 14th Asian Conference on Machine Learning,, Dec 2022, Hyderabad, India. hal-03906979

HAL Id: hal-03906979

<https://hal.science/hal-03906979v1>

Submitted on 19 Dec 2022

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

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

On the expressivity of bi-Lipschitz normalizing flows

Alexandre Vérine¹, Benjamin Negrevergne¹, Fabrice Rossi², Yann Chevaleyre¹

¹ LAMSADE, Université Paris-Dauphine

² CEREMADE, Université Paris-Dauphine

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 discuss potential remedies which include using more complex latent distributions.

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

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

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.

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 the Lipschitz constant of the mapping, on the expressivity. (Most work on the topic focus on the Lipschitz constant of the inverse mapping.) We give a new bound that depends on the Lipschitz constant of the mapping. As we will show, this new bound provide new insights and is complementary with the two other bounds we introduce. Finally, we discuss the potential remedies in the light of these new results.

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. Finally we discuss the potential remedies to the highlighted limitations in Section 4, and Section 5 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 $F : \mathcal{X} \rightarrow \mathcal{Z}$ (i.e. the *normalizing* direction) is tractable and exact and the inverse direction $F^{-1} : \mathcal{Z} \rightarrow \mathcal{X}$ (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 chose a d -dimensional Gaussian distribution Q over \mathcal{Z} (a.k.a. the latent space), and its density function $q(\mathbf{z}) = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{1}{2}\|\mathbf{z}\|_2^2}$. The choice of a normal distribution is natural since it is the most frequent latent distribution in normalizing flows. In Section 4, more complex distributions will be discussed. Then, we can define \hat{p} , the approximation of p^* , based on q and the mapping $F : \mathcal{X} \rightarrow \mathcal{Z}$, using a simple change of variable formula:

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

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

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

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 2.1. A bijective function $F : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathcal{Z} \subset \mathbb{R}^d$ 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.:

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

and

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

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

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

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

Proposition 2.1. *Jac_F satisfies for all $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$:*

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

Proposition 2.1 comes from a characterization of the Lipschitz continuity adapted to differentiable functions (Federer, 1969):

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

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

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 $\text{Lip}(g_i) \leq L < 1$. If F is composed of m residual blocks such that $F = f_m \circ \dots \circ f_1$, then the overall bi-Lipschitz constants satisfy $\text{Lip}(F) \leq (1 + L)^m$ and $\text{Lip}(F^{-1}) \leq 1/(1 - L)^m$. Alternatively, in Glow (Kingma and Dhariwal, 2018) with atomic blocks $W_i = P_i L_i (U_i + \text{diag}(s_i))$, the bi-Lipschitz constants satisfy: $\text{Lip}(F) \leq \prod_i^m \|W_i\|_2$ and $\text{Lip}(F^{-1}) \leq \prod_i^m \|W_i^{-1}\|_2$.

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 \hat{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 \hat{P} differs. We have chosen another tool to compare both distributions: the total variation (TV) distance given as :

Definition 2.2 (Total Variation Distance). *For any distribution \hat{P} and P^* , the total variation distance is defined as the maximum difference of probabilities given to a same event A :*

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

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 distance can be used to compute a lower bound on the \mathcal{D}_{KL} divergence through the Pinsker’s inequality:

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

Even if the main results of this paper are consisting in lower bounds on the TV distance, we compare our results to an existing one in terms of precision and recall. Therefore we give their definitions for generative models as they were given initially (Sajjadi et al., 2018; Kynkäänniemi et al., 2019).

Definition 2.3 (Precision α and Recall β for generative models). For $\alpha, \beta \in [0, 1]$, the distributions \hat{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 \hat{P} and P^* can be decomposed as such:

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

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

It can be interpreted as such: ν represent the part of P^* that \hat{P} correctly models, $\hat{\nu}$ is simultaneously the part of P^* that \hat{P} misses on their joint support and all the points that should not be represented by \hat{P} . Finally, ν^* cover the points of P^* that the support of \hat{P} could not reach and all the points on their joint support that \hat{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 2.2 (Maximum precision $\bar{\alpha}$ and maximum recall $\bar{\beta}$). The maximum precision and the maximum recall satisfy:

$$\bar{\alpha} = \hat{P}(\text{Supp}(P^*)) \quad \text{and} \quad \bar{\beta} = P^*(\text{Supp}(\hat{P})).$$

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. In general having upper bounds on $\bar{\alpha}$ or $\bar{\beta}$ is a stronger result than having a lower bound on the \mathcal{D}_{TV} since we have:

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

This can be expressed as :

$$\begin{aligned} \mathcal{D}_{\text{TV}}(P^*, \hat{P}) &\geq 1 - \bar{\alpha}, \\ \mathcal{D}_{\text{TV}}(P^*, \hat{P}) &\geq 1 - \bar{\beta}. \end{aligned}$$

However, as soon as the support of, respectively, the target distribution P^* or the estimated distribution \hat{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 largest 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 3.1 (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:*

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

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

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

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 2.1:

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

In other terms, the weight assigned from the Gaussian latent distribution to the subset A is bounded by $\left(\frac{L_1}{\sqrt{2\pi}} \right)^d \text{vol}(A)$. 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 the Theorem 3.1 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 3.1 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, \mathbf{x}_0} be the l_2 ball with center \mathbf{x}_0 and radius R (i.e. $B_{R, \mathbf{x}_0} = \{\mathbf{x} \in \mathcal{X} : \|\mathbf{x} - \mathbf{x}_0\|_2 \leq R\}$). 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.

Theorem 3.2 (NF with a L_1 -Lipschitz mapping F fails to capture high density balls). *Let F be L_1 -Lipschitz. Then:*

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

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

Theorem 3.2 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 ball with a high probability $P^*(B_R)$ in the data space, then the probability assigned to this ball $P^*(B_R) = Q(F(B_R))$ 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 3.2 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

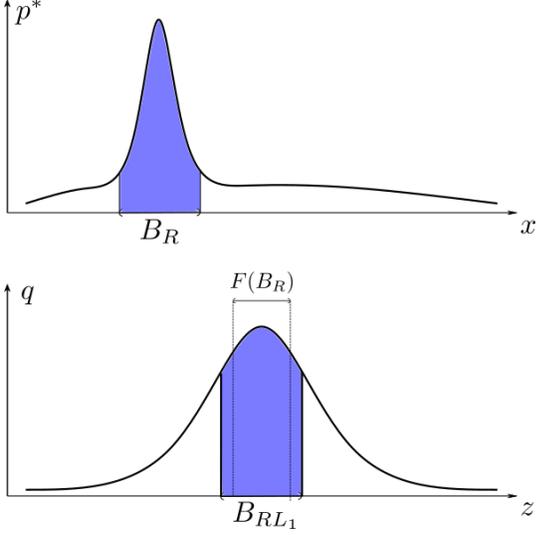


Figure 1: Example of a target distribution where theorem 3.2 applies: the subset B_R concentrates most of the weight in $P^*(B_R)$, but $\hat{P}(B_R) = Q(F(B_R))$ can only be as large as $Q(B_{RL_1})$.

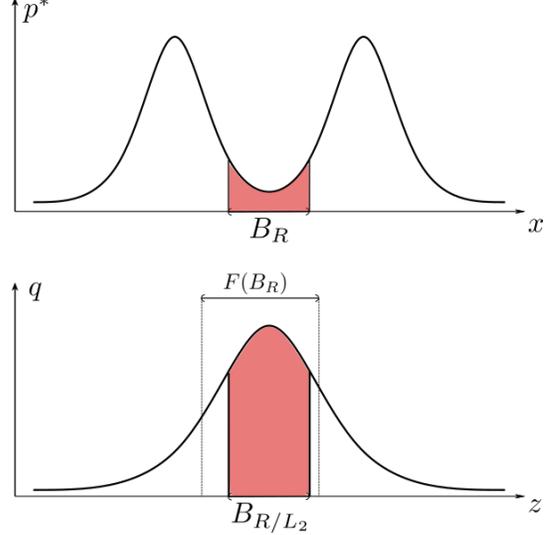


Figure 2: Example of a target distribution for which Theorem 3.4 applies: the subset B_R concentrates little weight in $P^*(B_R)$, but $\hat{P}(B_R) = Q(F(B_R))$ can only be as small as $Q(B_{R/L_2})$.

the Lipschitz constraint. For a given dense ball B_{R,\mathbf{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 function of the Gamma function Γ and the incomplete gamma function γ , therefore:

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

The Theorem 3.3 is then less open to interpretation than the Theorem 3.2 but it proposes a tight bound.

Theorem 3.3 (NF with a L_1 -Lipschitz mapping F fails to capture high density balls). *Let F be L_1 -Lipschitz. Then:*

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

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

A one dimensional representation of a pathological case for Theorems 3.2 & 3.3 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:

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

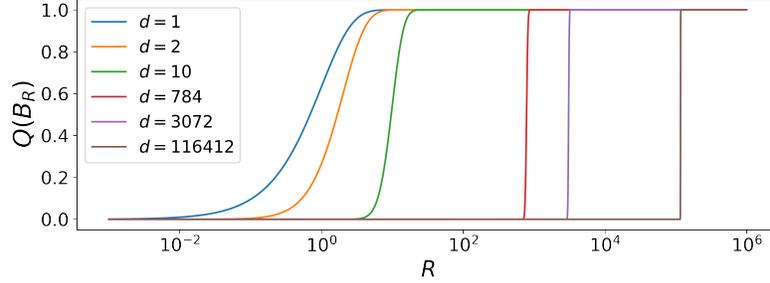


Figure 3: Representation of the Gaussian Measure of balls of radius R centered on $\mathbf{0}$. The measure is given for dimension 1, 2, 10 and then the dimensions of MNIST (Yann LeCun et al., 2010), CIFAR10 (Alex Krizhevsky, 2009) and CelebA (Liu et al., 2015)

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 3.3.

Theorem 3.4 (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:*

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

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 3.3, the probability $Q(B_{RL_1})$ 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 more increased in high dimension than in low dimension. The behavior is reversed for L_2 : from Theorem 3.4, 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 than in high dimension. The Lipschitz constant L_2 will have a greater effect in low dimension than in high dimension.

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. $\hat{P}(M_1) = \hat{P}(M_2) = 1/2$. As a consequence, $F^{-1}(0)$ is equidistant from M_1 and M_2 as illustrated in Figure 4. 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 the Theorem 3.4:

Corollary 3.4.1 (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 :*

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

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 than 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.

Note that since the TV distance is defined as the supremum on any subspace A , the approximation errors made by the network exposed by the same Theorem can be accumulated. Suppose that, in this pathological case, the balls M_1 and M_2 are dense enough that Theorem 3.3 can be applied on both subsets. It results in two different bounds on TV : $\mathcal{L}_{M_1} > 0$ and $\mathcal{L}_{M_2} > 0$. Then, since M_1 and M_2 are disjoint, we can write :

$$\begin{aligned} \mathcal{D}_{\text{TV}}(P^*, \hat{P}) &\geq |P^*(M_1 \cup M_2) - \hat{P}(M_1 \cup M_2)| \\ &\geq P^*(M_1) - \hat{P}(M_1) + P^*(M_2) - \hat{P}(M_2) \\ &\geq \mathcal{L}_{M_1} + \mathcal{L}_{M_2} \end{aligned}$$

This set up is then an appropriate pathological case to study the effect of the bi-Lipschitzness of the mapping.

4 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 μ and $\Sigma = \sigma I_d$. 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 3.3) with the expected error on subsets with low densities (Theorems 3.4) or vice-versa. In particular bounds becomes:

$$\mathcal{D}_{\text{TV}}(P^*, \hat{P}) \geq \sup_{R, \mathbf{x}_0} \left(P^*(B_{R, \mathbf{x}_0}) - \frac{\gamma\left(\frac{d}{2}, \frac{L_1^2 R^2}{2\sigma^2}\right)}{\Gamma\left(\frac{d}{2}\right)} \right),$$

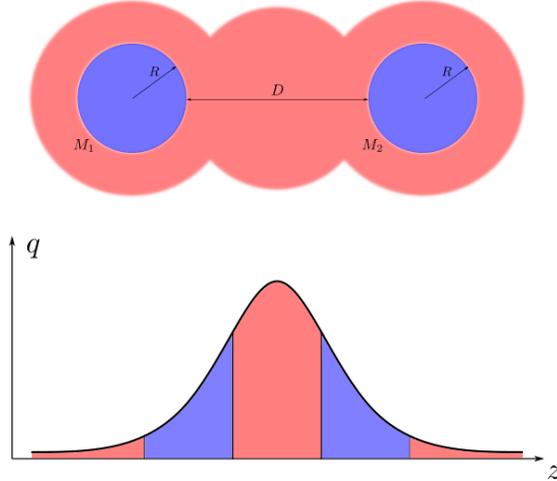


Figure 4: Experimental set up given by Tanielian et al. (2020)

and

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

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 a 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 3.4 does not hold anymore. Limitations similar to the ones highlighted in Theorem 3.2 still apply, but can be mitigated using learnable parameters.

We can trivially adapt the lower bound from Theorem 3.2 to the Gaussian Mixture with K equally distributed modes with learnable means μ_i and covariance matrices $\Sigma_i = \sigma_i I_d$:

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

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 3.3. 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 than 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 yet no efficient way 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 the Theorem 3.4, 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 3.3 & 3.4 but further investigation is required to deal with the technical issues.

5 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, of the data space can induce approximation errors. The second bound illustrates that some low density balls located between high density subsets can also result in approximations errors.

Further works could consist in completing the experimental aspects of these results. First understanding how those bound actually behave in the traditional dataset such as CIFAR10, CelebA or ImageNet. It would be interesting to develop a process to design normalizing flows based on the dataset and the Lipschitz constants of the reversible architecture.

Moreover, Theorems 3.3 & 3.4 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. One of the direction of the future work is to generalize, not only the results, but also the general framework to study the expressivity of generative lipschitz models based on the dataset.

Independently of the generative models, close attention is given on the effect the Lipschitz regularization as measure of the stability of a neural networks (Scaman and Virmaux, 2019; Combettes and Pesquet, 2020; Béthune et al., 2021), on the better generalization (Bartlett et al., 2017) or on the 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 (2009). Learning multiple layers of features from tiny images.
- Araujo, A., Negrevergne, B., Chevaleyre, Y., and Atif, J. (2020). On Lipschitz Regularization of Convolutional Layers using Toeplitz Matrix Theory. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. arXiv: 2006.08391.
- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia*. arXiv: 1701.07875.
- Ball, K. (1993). The reverse isoperimetric problem for Gaussian measure. *Discrete & Computational Geometry volume*, 10(4):411–420.
- Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, volume 30. Curran Associates, Inc.
- Behrmann, J., Grathwohl, W., Chen, R. T. Q., Duvenaud, D., and Jacobsen, J.-H. (2019). Invertible Residual Networks. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019*. arXiv: 1811.00995.
- Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R., and Jacobsen, J.-H. (2021). Understanding and Mitigating Exploding Inverses in Invertible Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1800. PMLR.

- Béthune, L., González-Sanz, A., Mamalet, F., and Serrurier, M. (2021). The Many Faces of 1-Lipschitz Neural Networks. *arXiv:2104.05097 [cs, stat]*. arXiv: 2104.05097.
- Chen, R. T. Q., Behrmann, J., Duvenaud, D., and Jacobsen, J.-H. (2020). Residual Flows for Invertible Generative Modeling. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*. arXiv: 1906.02735.
- Combettes, P. L. and Pesquet, J.-C. (2020). Lipschitz Certificates for Layered Network Structures Driven by Averaged Activation Operators. *SIAM Journal on Mathematics of Data Science, 2020*. arXiv: 1903.01014.
- Cornish, R., Caterini, A. L., Deligiannidis, G., and Doucet, A. (2021). Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows. In *Proceedings of the 37th International Conference on Machine Learning*. arXiv: 1909.13833.
- Dinh, L., Sohl-Dickstein, J., Larochelle, H., and Pascanu, R. (2020). A RAD approach to deep mixture models. In *ICLR 2019 Workshop DeepGenStruct*. arXiv: 1903.07714.
- Farnia, F., Zhang, J., and Tse, D. (2018). Generalizable Adversarial Training via Spectral Normalization. In *ICLR 2019*.
- Federer, H. (1969). *Geometric measure theory*. Berlin, heidelberg, new york, springer edition.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. In *27th Conference on Neural Information Processing Systems (NeurIPS 2014)*. arXiv: 1406.2661.
- Izmailov, P., Kirichenko, P., Finzi, M., and Wilson, A. G. (2019). Semi-Supervised Learning with Normalizing Flows. In *Proceedings of the 37th International Conference on Machine Learning*. arXiv: 1912.13025.
- Khayatkhoei, M., Elgammal, A., and Singh, M. (2019). Disconnected Manifold Learning for Generative Adversarial Networks. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*.
- Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative Flow with Invertible 1x1 Convolutions. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*, volume 31.
- Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. (2019). Improved Precision and Recall Metric for Assessing Generative Models. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*. arXiv: 1904.06991.
- Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep Learning Face Attributes in the Wild. *Proceedings of International Conference on Computer Vision (ICCV)*.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks. In *ICLR 2019*. arXiv: 1802.05957.
- Pinelis, I. (2020). Exact lower and upper bounds on the incomplete gamma function. *Mathematical Inequalities & Applications*, pages 1261–1278.
- Pires, G. G. P. F. and Figueiredo, M. A. T. (2020). Variational Mixture of Normalizing Flows. In *ESANN 2020*. arXiv: 2009.00585.
- Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). Assessing Generative Models via Precision and Recall. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*. arXiv: 1806.00035.

- Scaman, K. and Virmaux, A. (2019). Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*. arXiv: 1805.10965.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations, 2014*. arXiv: 1312.6199.
- Tanielian, U., Issenhuth, T., Dohmatob, E., and Mary, J. (2020). Learning disconnected manifolds: a no GANs land. In *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020*. arXiv: 2006.04596.
- Yann LeCun, Corinna Cortes, and Burges, C. (2010). MNIST handwritten digit database. *ATT Labs*, 2.

A Proofs

A.1 Proof of theorem 3.1

By definition we have $\widehat{P}(A) = \int_A \widehat{p}(\mathbf{x}) d\mathbf{x}$, then with the change of variable formula we obtain:

$$\begin{aligned}\widehat{P}(A) &= \int_A |\text{Jac}_F(\mathbf{x})| q(F(\mathbf{x})) d\mathbf{x} \\ &= \frac{1}{(2\pi)^{d/2}} \int_A |\text{Jac}_F(\mathbf{x})| e^{-\|F(\mathbf{x})\|^2/2} d\mathbf{x}\end{aligned}$$

As F is L_1 -Lipschitz we have $|\text{Jac}_F(\mathbf{x})| \leq L_1^d$, then

$$\begin{aligned}\widehat{P}(A) &\leq \left(\frac{L_1}{\sqrt{2\pi}}\right)^d \int_A e^{-\|F(\mathbf{x})\|_2^2} d\mathbf{x} \\ &\leq \left(\frac{L_1}{\sqrt{2\pi}}\right)^d \int_A d\mathbf{x} \\ &\leq \left(\frac{L_1}{\sqrt{2\pi}}\right)^d \text{vol}(A),\end{aligned}$$

and thus $TV(P^*, \widehat{P}) = \sup_A |P^*(A) - \widehat{P}(A)|$ implies

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

A.2 Proof of theorem 3.2

By definition of the TV distance, we have

$$\mathcal{D}_{TV}(P^*, \widehat{P}) \geq \sup_{R, \mathbf{x}_0} |P^*(B_{R, \mathbf{x}_0}) - Q(F(B_{R, \mathbf{x}_0}))|,$$

where B_{R, \mathbf{x}_0} is the ball of a radius R centered in \mathbf{x}_0 .

Then, the idea is to show that the image of a ball B_R by a L_1 -Lipschitz function is in a ball of radius $L_1 R$, and then use a reverse isoperimetric inequality to find an upper bound of the measure of a ball of a radius $L_1 R$.

Proof of $F(B_{R, \mathbf{x}_0}) \subset B_{L_1 R, F(\mathbf{x}_0)}$

First of all, for every $\mathbf{z} \in F(B_{R, \mathbf{x}_0})$, there exist $\mathbf{x} \in B_R$ such that $F^{-1}(\mathbf{z}) = \mathbf{x}$, we have:

$$\begin{aligned}\|F(F^{-1}(\mathbf{z})) - F(\mathbf{x}_0)\| &= \|F(\mathbf{x}) - F(\mathbf{x}_0)\| \\ &\leq L_1 \|\mathbf{x} - \mathbf{x}_0\| \\ &\leq L_1 R\end{aligned}$$

Upper bound of $Q(B_{L_1 R})$ This bound is extracted from the work of Ball (1993) on the Reverse Isoperimetric Inequality. First of all, it can be easily established that $Q(B_{L_1 R}(F(\mathbf{x}_0)))$ is at a maximum when $F(\mathbf{x}_0) = 0$. From now on, we will only consider $B_{L_1 R}$ the ball centered on 0. Therefore the objective is to find an upper bound on:

$$\begin{aligned}Q(B_{L_1 R}) &= \int_{\|\mathbf{z}\| < L_1 R} q(\mathbf{z}) d\mathbf{z} \\ &= \int_{\|\mathbf{z}\| < L_1 R} \frac{1}{(\sqrt{2\pi})^d} e^{-\|\mathbf{z}\|^2/2} d\mathbf{z}\end{aligned}$$

We can use the polar coordinates system to get another expression of the Gaussian measure with $S_{d-1}(r) = \frac{2\pi^{d/2}r^{d-1}}{\Gamma(d/2)}$ being the volume of the hypersphere:

$$\begin{aligned} Q(B_{L_1R}) &= \frac{1}{(2\pi)^{d/2}} \int_0^{L_1R} S_{d-1}(r) e^{-r^2/2} dr \\ &= \frac{2}{2^{d/2}\Gamma(d/2)} \int_0^{L_1R} r^{d-1} e^{-r^2/2} dr \end{aligned}$$

However $r^{d-1}e^{-r^2/2}$ has a maximum value reached for $r = \sqrt{d-1}$, we can have an upper bound:

$$\begin{aligned} Q(B_{L_1R}) &\leq \frac{2}{2^{d/2}\Gamma(d/2)} \sqrt{d-1}^{d-1} e^{-\frac{d-1}{2}} \int_0^{L_1R} dr \\ &\leq \frac{\sqrt{2}L_1R}{\Gamma(d/2)} \left(\frac{d-1}{2e}\right)^{\frac{d-1}{2}} \end{aligned}$$

Then, with the Stirling approximation of the Gamma function:

$$\begin{aligned} \frac{1}{2}\Gamma(d/2) &= \frac{1}{d}\Gamma(d/2+1) \\ &\geq \frac{\sqrt{\pi}\sqrt{d}}{d} (d/2)^{d/2} e^{-d/2} \\ &\geq \frac{\sqrt{\pi}}{2^{d/2}} d^{\frac{d-1}{2}} e^{-\frac{d}{2}} \end{aligned}$$

We obtain:

$$\begin{aligned} Q(B_{L_1R}) &\leq \frac{2}{2^{d/2}\Gamma(d/2)} (d-1)^{\frac{d-1}{2}} e^{-\frac{d-1}{2}} \\ &\leq \frac{L_1R\sqrt{e}}{\sqrt{\pi}} \left(\frac{d-1}{d}\right)^{\frac{d-1}{2}} \end{aligned}$$

Using the bound

$$\frac{1}{\sqrt{e}} < \left(\frac{d-1}{d}\right)^{\frac{d-1}{2}},$$

we have

$$Q(B_{L_1R}) < \frac{L_1R}{\sqrt{\pi}}$$

Lower Bound of the TV As soon as we have an upper bound on $Q(B_{L_1R})$, we have:

$$\begin{aligned} \mathcal{D}_{\text{TV}}(P^*, \hat{P}) &\geq \sup_{R, \mathbf{x}_0} (P^*(B_{R, \mathbf{x}_0}) - Q(F(B_{R, \mathbf{x}_0}))) \\ &\geq \sup_{R, \mathbf{x}_0} (P^*(B_{R, \mathbf{x}_0}) - Q(B_{L_1R, \mathbf{x}_0})) \\ &\geq \sup_{R, \mathbf{x}_0} \left(P^*(B_{R, \mathbf{x}_0}) - \frac{L_1R}{\sqrt{\pi}} \right) \end{aligned}$$

A.3 Proof of Theorem 3.3

Value of $Q(B_{R,0})$ By construction

$$Q(B_{R,0}) = \mathbb{P}(\|\mathbf{z}\|^2 \leq R^2),$$

when \mathbf{z} follows the standard Gaussian distribution in \mathbb{R}^d . This quantity can be computed using the cumulative distribution function of the chi-square distribution, i.e.

$$Q(B_{R,0}) = \frac{\gamma\left(\frac{d}{2}, \frac{R^2}{2}\right)}{\Gamma\left(\frac{d}{2}\right)},$$

where γ is the lower incomplete gamma function given by

$$\gamma(x, k) = \int_0^x t^{k-1} e^{-t} dt.$$

Lower Bound of the TV Since we have the closed form of the measure over a ball we can write :

$$\begin{aligned}
\mathcal{D}_{\text{TV}}(P^*, \widehat{P}) &\geq \sup_{R, \mathbf{x}_0} (P^*(B_{R, \mathbf{x}_0}) - Q(F(B_{R, \mathbf{x}_0}))) \\
&\geq \sup_{R, \mathbf{x}_0} (P^*(B_{R, \mathbf{x}_0}) - Q(B_{L_1 R, \mathbf{x}_0})) \\
&\geq \sup_{R, \mathbf{x}_0} \left(P^*(B_{R, \mathbf{x}_0}) - \frac{\gamma\left(\frac{d}{2}, \frac{L_1^2 R^2}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \right)
\end{aligned}$$

A.4 Proof of theorem 3.4

In this section, we denote $B_R = B_{R, F^{-1}(0)}$. As F^{-1} is L_2 -Lipschitz, $F^{-1}(B_{R/L_2, 0}) \subset B_R$ and thus

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

Therefore, by analogy with the proof of Theorem 3.3:

$$\begin{aligned}
\mathcal{D}_{\text{TV}}(P^*, \widehat{P}) &\geq \sup_R (Q(F(B_R)) - P^*(B_R)) \\
&\geq \sup_R (Q(B_{R/L_2}) - P^*(B_R)) \\
&\geq \sup_R \left(\frac{\gamma\left(\frac{d}{2}, \frac{R^2}{2L_2^2}\right)}{\Gamma\left(\frac{d}{2}\right)} - P^*(B_R) \right)
\end{aligned}$$

A.5 Proof of Corollary 3.4.1

Since M_1 and M_2 are separated by a distance D the ball centered on $F^{-1}(0)$ has a radius at least as big as D that we might call B_D to simplify the notation. Therefore:

$$\begin{aligned}
\bar{\alpha} &= \widehat{P}(M_1) + \widehat{P}(M_2) \\
&= 1 - \widehat{P}(M_2 \cup M_1) \\
&\leq 1 - \widehat{P}(B_D) \\
&\leq 1 - Q(F(B_D)) \\
&\leq 1 - Q(B_{D/L_2}) \\
&\leq 1 - \frac{\gamma\left(\frac{d}{2}, \frac{D^2}{2L_2^2}\right)}{\Gamma\left(\frac{d}{2}\right)}
\end{aligned}$$

And since $P^*(B_D) = 0$:

$$\begin{aligned}
\mathcal{D}_{\text{TV}}(P^*, \widehat{P}) &\geq |\widehat{P}(B_D) - P^*(B_D)| \\
&\geq \widehat{P}(B_D(F^{-1}(0))) \\
&\geq \frac{\gamma\left(\frac{d}{2}, \frac{D^2}{2L_2^2}\right)}{\Gamma\left(\frac{d}{2}\right)}
\end{aligned}$$