On the expressivity of bi-Lipschitz normalizing flows

Anonymous Authors[1]

Abstract

An invertible function is *bi-Lipschitz* if both the function and its inverse have bounded Lipschitz constants. Nowadays, most Normalizing Flows are bi-Lipschitz by design or by training to limit numerical errors (among other things). In this paper, we discuss the expressivity bi-Lipschitz of 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 divergence 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 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 constraint are more stable during training, 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 constant. In particular Tanielian et al. (2020) was able identify a family of target distributions with disconnected support that cannot be fitted with a GAN with a bounded Lipschitz constant.

In this paper we focus on the impact of the Lipschitz constraint on normalizing flows. When they are invertible, normalizing flows are often not just Lipschitz, but *bi-Lipschitz*, meaning that both the mapping function and its inverse have bounded Lipschitz constant. For example, Affine Coupling, Additive Coupling, neural ODE, Invertible 1×1 Convolution and Residual Networks are bi-Lipschitz by design. Other types of normalizing flows, can also be trained to be bi-Lipschitz, in order to avoid exploding inverses (Behrmann et al., 2021). We study the expressivity of normalizing flows with bounded Lipschitz constant and discuss the impact of the bi-Lipschitz constant on the total variation divergence. More precisely, we give several lower bounds on the total variation divergence between the generated distribution and the target distribution, in some (particularly unfavorable) training settings.

2. Background

An *invertible normalizing flow* or normalizing flow for short, is a an invertible density model in which both density estimation and sampling can be done efficiently. Normalizing flows (NF) learn an invertible mapping between a latent space $\mathcal{Z}$ and a data space $\mathcal{X}$. Typically, the forward direction $F : \mathcal{X} \to \mathcal{Z}$ (i.e. the *normalizing* direction) is tractable and exact whereas 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 interative algorithm.

Let P^* be the true data distribution over $\mathcal{X}$ that we wish to approximate. We assume this distribution admits a density p^*. Let $q(\boldsymbol{z}) = \frac{1}{(\sqrt{2\pi})^d} e^{-\|\boldsymbol{z}\|^2/2} d\boldsymbol{z}$ be a gaussian density function and Q the associated measure. If we name P_θ the learned data distribution, and p_θ its density, we can perform the density estimate efficiently using a simple change of variable:

$$\forall x \in \mathcal{X}, \quad p_\theta(\boldsymbol{x}) = |\mathrm{Jac}_F(\boldsymbol{x})| q(F(\boldsymbol{x})) \tag{1}$$

The learned distribution P_θ can there be expressed with the Gaussian measure over any subset $A \subseteq \mathcal{X}$:

$$P_\theta(A) = Q(F(A)) = \int_{F(A)} q(\boldsymbol{z}) d\boldsymbol{z}$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by INNF+ 2021. Do not distribute.

For this to be useful in practice, the Jacobian (more specifically its determinant) must be computed efficiently. The relation between both distributions requires the determinant as given in Equation 1. To satisfies this property, Normalizing Flows usually have particular structures and consequently most of the common blocks are bi-Lipschitz.

2.1. Bi-Lipschitz Normalizing Flows

In this paper, we consider bi-Lipschitz normalizing flows. We define the bi-Lipschitz property as follows.

Definition 2.1. *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:*

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

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

The bi-Lipschitz continuity implies some intrinsic properties satisfied by normalizing flows. Indeed, if a function can not expend infinitely, its inverse function can not contract infinitely. Therefore, not only the contractive power is limited but also the determinant of the jacobian is bounded.

Proposition 2.1. *If a normalizing flow F is (L_1, L_2)-bi-Lipschitz, then for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, it satisfies:*

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

Moreover, the determinant of its Jacobian Matrix Jac_F satisfies for all $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$: $1_{\frac{}{L_2^d \leq |\mathrm{Jac}_F(\boldsymbol{x})| \leq L_1^d}}$

Notice that this bounding is often used as an alternate bi-Lipschitz definition.

For instance, 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 & Dhariwal, 2018) with atomic blocks $W_i = P_i L_i (U_i + \mathrm{diag}(s_i))$, the bi-Lipschitz constants statisfy: $\mathrm{Lip}(F) \leq \prod_i^m \|W_i\|_2$ and $\mathrm{Lip}(F^{-1}) \leq \prod_i^m \|W_i^{-1}\|_2$. Consequently, the bi-Lipschitzness constraints on either the function or its Jacobian determinant can be released by increasing the depth of the network but, by doing so, the stability of the inverse can be affected (Behrmann et al., 2021).

2.2. Assessing the learning abilities

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 P_θ and its density p_θ.

To evaluate how the true distribution P^* and the generated distribution P_θ differ, we use the Total Variation Distance, whose definition is recalled here:

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

3. Lower Bounds on the TV Distance

3.1. A bound on bi-Lipschitz normalizing flow for any subset A

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 an arbitrary subset A that is sufficiently concentrated so that the Lipschitz contrained mapping can not concentrate enough weight form the Gaussian distribution onto this subset.

Theorem 3.1 (bi-Lipschitz mappings fail to capture high density subset)**.** *Let F be (L_1, L_2)-bi-Lipschitz and $\eta_A = \frac{P^*(A)}{\mathrm{vol}(A)}$ be the average density over any subset $A \subset \mathbb{R}^d$. Then:*

$$\mathcal{D}_{\mathrm{TV}}(P^*, P_\theta) \geq \sup_A \mathrm{vol}(A) \left(\eta_A - \left(\frac{L_1}{4L_2\sqrt{2\pi}} \right)^d \right)$$

Therefore, if there is a subset A that satisfies $\eta_A > \left(\frac{L_1}{4L_2\sqrt{2\pi}}\right)^d$, then the TV is necessarily strictly positive.

The proof of this Theorem is given in appendix A.1.

Remark that the bound in Theorem 3.1 depends on both Lipschitz constraints L_1 and L_2. If a subset A is found to be very dense, the mapping will not be able to expand the given volume of A to match the lower density of the Gaussian density because of L_1. On the other hand, the point with the highest density within A will be matched with the highest point on the Gaussian density but all its neighbourhood has to me moved by a factor of $1/L_2$. The main advantage of this formulation is to apply to any subset of the data space, but at the expense of a loose bound on the TV.

3.2. Bounds for specific subset $B_{R,\boldsymbol{x}}$

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 $B_{R,\boldsymbol{x}_0}$ 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.

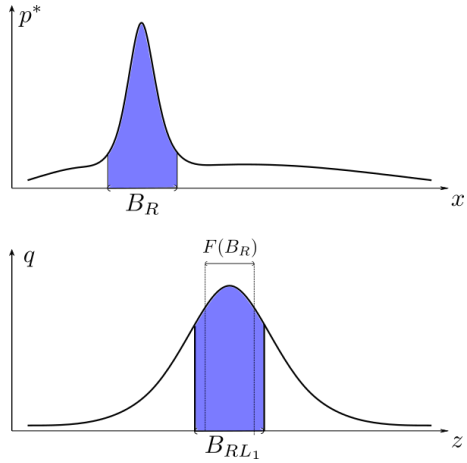

Figure 1. Example of a target distribution where theorem 3.2 applies: the subset B_R concentrates most of the weight in p^*, but $q(F(B_R))$ can only be as big as $q(B_{R_{L_1}})$.

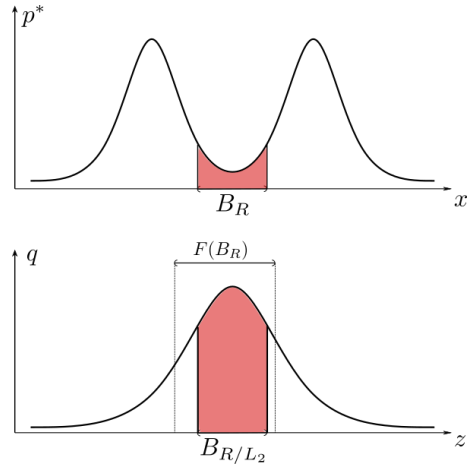

Figure 2. Example of a target distribution for which Theorem 3.3 applies: the subset B_R concentrates little weight in p^*, but $q(F(B_R))$ can only be as small as $q(B_R/L_2)$.

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

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

$$\mathcal{D}_{\mathrm{TV}}(P^*, P_\theta) > \sup_{R, \boldsymbol{x}_0} \left(P^*(B_{R,\boldsymbol{x}_0}) - 4d^{1/4} RL_1 \right)$$

Therefore, if we find a ball for which the true measure satisfies $\frac{P^*(B_{R,\boldsymbol{x}_0})}{R} > \frac{L_1}{\sqrt{\pi}}$ or $\frac{P^*(B_{R,\boldsymbol{x}_0})}{R} > 4d^{1/4}L_1$, then the TV is necessarily strictly positive.

The theorem 3.2 highlights the effect of the Lipschitz constraint of the forward mapping F. If a ball has a high probability in the data space $P^*(B_R)$, then the probability assigned to this ball is at most $Q(B_{RL_1})$ in the latent space and is upper bounded by $RL_1/\sqrt{\pi}$ (Ball, 1993). A one dimensional representation of a pathological case for theorem 3.2 is shown on figure 1. In other words no ball with a high enough density in data space can be expended sufficiently to have a matching probability in the latent space. Note that we could use a closed-form of $Q(B_{RL_1})$ but it is less open to interpretation than the approximations we have made.

Conversely, the mapping being bi-Lipschitz, the mapping can not contract arbitrarily. 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. Notice that the assumption of a low density zone is strong but fairly reasonable for instance, one can observe a multimodal 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 on 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 that is expressed with the Gamma function Γ and the incomplete gamma function γ. The numerical approximations of the closed form are given in figure 3. We can observe that the higher the dimension is, the larger the l_2 distance between two modes can be.

Theorem 3.3 (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:*

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

Therefore, if we find a ball for which the 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.

Both formal proofs are detailed in appendix A.2 and A.3.

3.3. Comparison to related work

A related set up is used in (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. $P_\theta(M_1) = P_\theta(M_1) = 1/2$. As a consequence, $F^{-1}(0)$ is equidistant from M_1 and M_2 as illustrated in Figure 4.

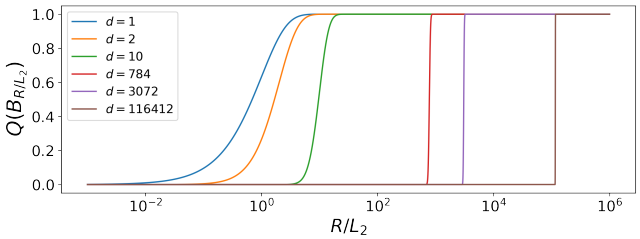

Figure 3. Representation of the Gaussian Measure of balls of radius R/L_2 centered on 0. The measure is given for dimension 1, 2, 10 and then the dimensions of MNIST (Yann LeCun et al., 2010), CIFAR10 (Alex Krizhevsky, 2009) and CelebA (Liu et al., 2015)

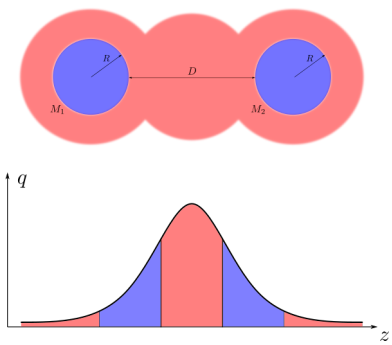

Figure 4. Experimental set up given by (Tanielian et al., 2020)

Corollary 3.3.1 (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:*

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

Note that, the TV distance being defined as the sup on any subspace A, we can accumulate the failures made by the network, and therefore take into account the error when the two manifolds M_1 and M_2 are too dense. This set up is then a appropriate pathological case to study the effect of the bi-Lipschitzness of the mapping.

The original work assesses the learning abilities of their generative model, a GAN (Goodfellow et al., 2014), with a definition of precision and recall given by (Sajjadi et al., 2018) and improved by (Kynkäänniemi et al., 2019). The main advantage of this metric is that is well fitted to be used with the Gaussian Isoperimetric Inequality and therefore gives a result independent from the dimension. By using, the TV distance or any distance for that matter, it can be applied on distributions with any support. The details of the Precision and Recall and the comparison between both methods can be found in appendix B and the proof of corollary 3.3.1 is in appendix A.4.

4. Potential remedies

To mitigate the limitations induced by the two Lipschitz constants, one may suggest to simply increase the aggregated Lipschitz constants. For most structures of Normalizing Flow, one solution would be to increase the number of layers of the network. However, according to (Behrmann et al., 2021), this method affects the invertible stability of the normalizing flow. The solution might be held in more complex latent distribution. The first idea would be to use a Gaussian distribution with learnable parameters $\boldsymbol{\mu}$ and $\Sigma = \mathrm{diag}(\sigma_i)$ but this only results in a trade-off between the failures highlighted by the theorems 3.2 and 3.3. Since a learnable standard deviation σ is equivalent to a reparametrization trick, the consequence would be shifting the Lipschitz constant (L_1, L_2) toward $(\frac{L_1}{\sigma}, L_2\sigma)$. Then, it could be interesting when only one failure is met under the condition that the stability may not be affection by a larger range within the latent space.

To solve this, a Gaussian Mixture (GM) can be used to learn disconnected manifolds (Khayatkhoei et al., 2019; Izmailov et al., 2019). Since each mode can be assigned to a different manifold or high density zones, the maximum density of the gaussian density will not be mapped to the low density space between the two density peaks. Consequently, if theorem 3.3 has no reason to hold, we can express the lower bound of theorem 3.2 with a GM with K equally distributed modes with learnable mean $\boldsymbol{\mu}_j$ and diagonal covariance matrix $\Sigma_j = \mathrm{diag}(\sigma_{ji})$:

$$\mathcal{D}_{\mathrm{TV}}(P^*, P_\theta) \geq \sup_{R, \boldsymbol{x}_0} \left(P^*(B_{R,\boldsymbol{x}_0}) - \frac{1}{K} \frac{RL_1}{\prod_i \sigma_{ji}^d \sqrt{\pi}} \right)$$

Therefore, the GM could solve the issue set by split subset of the data space under two conditions: the number of modes K should be reasonably low such that the measure mapped onto potential dense subset can be large enough, and most importantly, great attention should be given to the experimental issue linked to the training with learnable σ_{ij} and K.

5. Conclusion

We have established 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. To compensate, this lack of learning ability of the mapping, a more complex, i.e. expressive latent distribution can be implemented. However, this method suffers from training difficulties and should be further studied.

References

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

Ball, K. 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.

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

Behrmann, J., Grathwohl, W., Chen, R. T. Q., Duvenaud, D., and Jacobsen, J.-H. 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.

Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R., and Jacobsen, J.-H. Understanding and Mitigating Exploding Inverses in Invertible Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1792–1800. PMLR, March 2021. URL http://proceedings.mlr.press/v130/behrmann21a.html.

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

Farnia, F., Zhang, J., and Tse, D. Generalizable Adversarial Training via Spectral Normalization. September 2018. URL https://openreview.net/forum?id=Hyx4knR9Ym.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 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.

Izmailov, P., Kirichenko, P., Finzi, M., and Wilson, A. G. Semi-Supervised Learning with Normalizing Flows. In *Proceedings of the 37th International Conference on Machine Learning*, December 2019. arXiv: 1912.13025.

Khayatkhoei, M., Elgammal, A., and Singh, M. 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.

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

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. 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.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the Wild. *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. 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.

Tanielian, U., Issenhuth, T., Dohmatob, E., and Mary, J. 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.

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

A. Proofs

A.1. Proof of theorem 3.1

By definition we have $P_\theta(A) = Q(F(A)) = \int_{F(A)} q(\boldsymbol{z})d\boldsymbol{z}$, then with the change of variable formula we obtain:

$$\begin{aligned} P_\theta(A) &= \int_A |\mathrm{Jac}_F(\boldsymbol{x})| q(F(\boldsymbol{x}))d\boldsymbol{x} \\ &= \frac{1}{(2\pi)^{d/2}} \int_A |\mathrm{Jac}_F(\boldsymbol{x})| e^{-\|F(\boldsymbol{x})\|^2/2} d\boldsymbol{x} \end{aligned}$$

As F^{-1} is L_2-Lipschitz, F satisfies:

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \quad \frac{1}{L_2}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \leq \|F(\boldsymbol{x}_1) - F(\boldsymbol{x}_2)\|,$$

and in particular, we have:

$$\forall \boldsymbol{x} \in \mathcal{X}, \quad \frac{1}{L_2}\|\boldsymbol{x}_1 - F^{-1}(0)\| \leq \|F(\boldsymbol{x})\|.$$

Consequently $\forall \boldsymbol{x} \in \mathcal{X}$

$$\begin{aligned} q(F(\boldsymbol{x})) &= \frac{1}{(2\pi)^{d/2}} e^{-\|F(\boldsymbol{x})\|^2/2}, \\ &\leq \frac{1}{(2\pi)^{d/2}} e^{-\|\boldsymbol{x}/L_2 - F^{-1}(0)/L_2\|^2/2}, \\ &\leq \frac{1}{(2\pi)^{d/2}} e^{-\|T(\boldsymbol{x})\|^2}, \end{aligned}$$

where T is the affine mapping given by

$$T(\boldsymbol{x}) = \frac{\boldsymbol{x} - F^{-1}(0)}{4L_2}.$$

As F is L_1-Lipschitz we have $|\mathrm{Jac}_F(\boldsymbol{x})| < L_1^d$ and thus

$$\begin{aligned} P_\theta(A) &\leq \left(\frac{L_1}{\sqrt{2\pi}}\right)^d \int_A e^{-\|T(\boldsymbol{x})\|_2^2} d\boldsymbol{x} \\ &\leq \left(\frac{L_1}{\sqrt{2\pi}}\right)^d \int_{T(A)} \frac{1}{(4L_2)^d} d\boldsymbol{x} \\ &\leq \left(\frac{L_1}{4L_2\sqrt{2\pi}}\right)^d \mathrm{vol}(A), \end{aligned}$$

and thus $TV(P^*, P_\theta) = \sup_A |P^*(A) - P_\theta(A)|$ implies

$$TV(P^*, P_\theta) \geq \sup_A \left(P^*(A) - \left(\frac{L_1}{4L_2\sqrt{2\pi}}\right)^d \mathrm{vol}(A) \right)$$

A.2. Proof of theorem 3.2

By definition of the TV distance, we have

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

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

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 the find an upper bound of the measure of a ball of a radius $L_1 R$.

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

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

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

Upper bound of $Q(B_{L_1 R})$ First of all, it can be easily establish that $Q(B_{L_1 R}(F(\boldsymbol{x}_0)))$ is at a maximum when $F(\boldsymbol{x}_0) = 0$. 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:

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

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

$$\begin{aligned} Q(B_{L_1 R}) &= \frac{1}{(2\pi)^{d/2}} \int_0^{L_1 R} S_{d-1}(r) e^{-r^2/2} dr \\ &= \frac{2}{2^{d/2}\Gamma(d/2)} \int_0^{L_1 R} r^{d-1} e^{-r^2/2} dr \end{aligned}$$

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

$$\begin{aligned} Q(B_{L_1 R}) &\leq \frac{2}{2^{d/2}\Gamma(d/2)} \sqrt{d-1}^{d-1} e^{-\frac{d-1}{2}} \int_0^{L_1 R} dr \\ &\leq \frac{\sqrt{2} L_1 R}{\Gamma(d/2)} \left(\frac{d-1}{2e}\right)^{\frac{d-1}{2}} \end{aligned}$$

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:

$$\begin{aligned} Q(B_{L_1 R}) &\leq \frac{2}{2^{d/2}\Gamma(d/2)} (d-1)^{\frac{d-1}{2}} e^{-\frac{d-1}{2}} \\ &\leq \frac{L_1 R\sqrt{e}}{\sqrt{\pi}} \left(\frac{d-1}{d}\right)^{\frac{d-1}{2}} \end{aligned}$$

Using the bound

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

we have

$$Q(B_{L_1 R}) < \frac{L_1 R}{\sqrt{\pi}}$$

Otherwise, an inequality from (Ball, 1993) give for the hypersphere in $\mathbb{R}^{d-1}$, an upper bound of the Gaussian Measure over a convex set C in $\mathbb{R}^d$:

$$\forall d \geq 2, \int_{\partial C} q < 4d^{1/4}$$

Therefore, for a Ball $B_{L_1 R}$:

$$\forall d \geq 2, \quad Q(B_{L_1 R}) < 4d^{1/4} L_1 R$$

Lower Bound of the TV As soon as we have an upper bound on $Q(B_{L_1 R})$, we have:

$$\mathcal{D}_{\mathrm{TV}}(P^*, P_\theta) \geq \sup_{R, \boldsymbol{x}_0} \left(P^*(B_{R,\boldsymbol{x}_0}) - \frac{L_1 R}{\sqrt{\pi}} \right)$$

With the second upper bound, the theorem can be formulated as:

$$\mathcal{D}_{\mathrm{TV}}(\mu^*, \mu_\theta) > \sup_{R, \boldsymbol{x}_0} \left(\mu^*(B_{R,\boldsymbol{x}_0}) - 4d^{1/4} R L_1 \right)$$

A.3. Proof of theorem 3.3

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

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

By construction

$$Q(B_{R/L_2,0}) = \mathbb{P}\left(\|\boldsymbol{z}\|^2 \leq \frac{R^2}{L_2^2} \right),$$

when $\boldsymbol{z}$ 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.

$$Q(B_{R/L_2,0}) = \frac{\gamma(\frac{d}{2}, \frac{R^2}{2L_2^2})}{\Gamma(\frac{d}{2})},$$

where γ is the lower incomplete gamma function given by

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

A.4. Proof of Corollary 3.3.1

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:

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

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

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

B. The link with the Precision and Recall for generative models

B.1. Definitions of the precision and the recall

The precision and the recall are defined as such :

Definition B.1. *For $\alpha, \beta \in [0, 1]$, the distributions P_θ is said to have a precision α at recall β with respect to P^* if there exist the distributions ν, ν_θ, ν^*, such that P_θ and P^* can be decomposed as such :*

$$P_\theta = \alpha\nu + (1-\alpha)\nu_\theta \quad \textit{and} \quad P^* = \beta\nu + (1-\beta)\nu^*$$

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

It can be interpreted as such in (Sajjadi et al., 2018): ν represent the part of P^* that P_θ correctly models, ν_θ is simultaneously the part of P^* that P_θ misses on their joint support and all the points that should not be represented by P_θ. Finally, ν^* cover the points of P^* that P_θ could not model and the difference between ν and ν_θ on their joint support. 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 B.1. *The maximum precision and the maximum recall satisfy :*

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

An improved version of the precision and the recall are defined in (Kynkäänniemi et al., 2019).

B.2. The link between the maximum precision, the maximum recall and the TV distance

The link between the TV and the maximum precision $\bar{\alpha}$ and the maximum recall $\bar{\beta}$ is :

$$\mathcal{D}_{\mathrm{TV}}(P^*, P_\theta) \geq |P^*(\mathrm{Supp}(P^*)) - P_\theta(\mathrm{Supp}(P^*))| = 1 - \bar{\alpha}$$

$$\mathcal{D}_{\mathrm{TV}}(P^*, P_\theta) \geq |P^*(\mathrm{Supp}(P_\theta)) - P_\theta(\mathrm{Supp}(P_\theta))| = 1 - \bar{\beta}$$

B.3. Corollary 3.3.1 given in terms of maximum precision

Corollary B.0.1 (F^{-1} L_2-Lipschitz)**.** *If F^{-1} is L_2-Lipschitz, therefore we have an upper bound on the maximum precision based on the distance D between M_1 and*

M_2:

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

This result is to be compared with the actual upper bound on the maximum precision:

$$\bar{\alpha} + \frac{2D}{L_2} e^{-\Phi^{-1}(\bar{\alpha}/2)^2} \leq 1 \text{ where } \Phi(t) = \int_{-\infty}^{t} \frac{\exp(-r^2/2)}{2\pi} dr$$

Our result may depend on the dimension but an upper bound of the precision can be directly computed.