Learning Disconnected Manifolds: a no GAN's land

Ugo Tanielian [1 2] **Thibaut Issenhuth** [2] **Elvis Dohmatob** [2] **Jérémie Mary** [2]

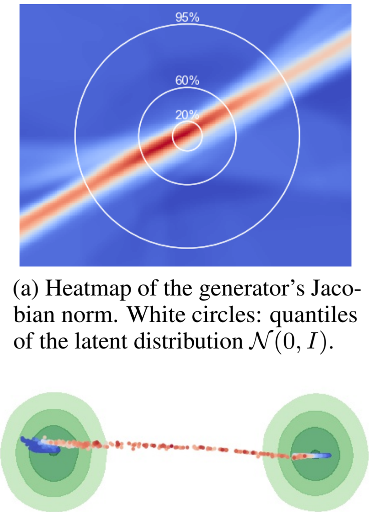
Abstract

Typical architectures of Generative Adversarial Networks make use of a unimodal latent/input distribution transformed by a continuous generator. Consequently, the modeled distribution always has connected support which is cumbersome when learning a disconnected set of manifolds. We formalize this problem by establishing a "no free lunch" theorem for the disconnected manifold learning stating an upper-bound on the precision of the targeted distribution. This is done by building on the necessary existence of a low-quality region where the generator continuously samples data between two disconnected modes. Finally, we derive a rejection sampling method based on the norm of generator's Jacobian and show its efficiency on several generators including BigGAN.

1. Introduction

GANs (Goodfellow et al., 2014) provide a very effective tool for the unsupervised learning of complex probability distributions. For example, Karras et al. (2019) generate very realistic human faces while Yu et al. (2017) match state-of-the-art text corpora generation. Despite some early theoretical results on the stability of GANs (Arjovsky & Bottou, 2017) and on their approximation and asymptotic properties (Biau et al., 2018), their training remains challenging. More specifically, GANs raise a mystery formalized by Khayatkhoei et al. (2018): *how can they fit disconnected manifolds when they are trained to continuously transform a unimodal latent distribution?* While this question remains widely open, we will show that studying it can lead to some improvements in the sampling quality of GANs.

Indeed, training a GAN with the objective of continuously transforming samples from an unimodal distribution into a disconnected requires balancing between two caveats. On one hand, the generator could just ignore all modes but one, producing a very limited variety of high quality samples: this is an extreme case of the well known mode collapse (Arjovsky & Bottou, 2017). On the other hand, the generator could cover the different modes of the target distribution and necessarily generates samples out of the real data manifold as previously explained by Khayatkhoei et al. (2018).



(a) Heatmap of the generator's Jacobian norm. White circles: quantiles of the latent distribution $\mathcal{N}(0, I)$.

(b) Green: target distribution. Coloured dots: generated samples colored w.r.t. the Jacobian Norm using same heatmap than (a).

Figure 1. Learning disconnected manifolds leads to the apparition of an area with high gradients and data sampled in between modes.

As brought to the fore by Roth et al. (2017), there is a density mis-specification between the true distribution and the model distribution. Indeed, one cannot find parameters such that the model density function is arbitrarily close to the true distribution. To solve this issue, many empirical works have proposed to over-parameterize the generative distributions, as for instance, using a mixture of generators to better fit the different target modes. Tolstikhin et al. (2017) rely on boosting while Khayatkhoei et al. (2018) force each generator to target different sub-manifolds thanks to a criterion based on mutual information. Another direction is to add complexity in the latent space using a mixture of Gaussian distributions (Gurumurthy et al., 2017).

[1] Université Paris-Sorbonne, Paris, France [2] Criteo AI Lab, France. Correspondence to: Ugo Tanielian <u.tanielian@criteo.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

To better visualize this phenomenon, we consider a simple 2D motivational example where the real data lies on two disconnected manifolds. Empirically, when learning the distribution, GANs split the Gaussian latent space into two modes, as highlighted by the separation line in red in Figure 1a. More importantly, each sample drawn inside this red area in Figure 1a is then mapped in the output space in between the two modes (see Figure 1b). For the quantitative evaluation of the presence of out-of-manifold samples, a natural metric is the Precision-Recall (PR) proposed by Sajjadi et al. (2018) and its improved version (Improved PR) (Kynkäänniemi et al., 2019). A first contribution of this paper is to formally link them. Then, taking advantage of these metrics, we lower bound the measure of this out-of-manifold region and formalize the impossibility of learning disconnected manifolds with standard GANs. We also extend this observation to the multi-class generation case and show that the volume of off-manifold areas increases with the number of covered manifolds. In the limit, this increase drives the precision to zero.

To solve this issue and increase the precision of GANs, we argue that it is possible to remove out-of-manifold samples using a truncation method. Building on the work of Arvanitidis et al. (2017) who define a Riemaniann metric that significantly improves clustering in the latent space, our truncation method is based on information conveyed by the Jacobian's norm of the generator. We empirically show that this rejection sampling scheme enables us to better fit disconnected manifolds without over-parametrizing neither the generative class of functions nor the latent distribution. Finally, in a very large high dimensional setting, we discuss the advantages of our rejection method and compare it to the truncation trick introduced by (Brock et al., 2019).

In a nutshell, our contributions are the following:

- We discuss evaluation of GANs and formally link the PR measure (Sajjadi et al., 2018) and its Improved PR version (Kynkäänniemi et al., 2019).
- We upper bound the precision of GANs with Gaussian latent distribution and formalize an impossibility result for disconnected manifolds learning.
- Using toy datasets, we illustrate the behavior of GANs when learning disconnected manifolds and derive a new truncation method based on the Jacobian's Frobenius norm of the generator. We confirm its empirical performance on state-of-the-art models and datasets.

2. Related work

Fighting mode collapse. Goodfellow et al. (2014) were the first to raise the problem of mode collapse in the learning of disconnected manifolds with GANs. They observed that when the generator is trained too long without updating the discriminator, the output distribution collapses to a few modes reducing the diversity of the samples. To tackle this issue, Salimans et al. (2016); Lin et al. (2018) suggested feeding several samples to the discriminator. Srivastava et al. (2017) proposed the use of a reconstructor network, mapping the data to the latent space to increase diversity.

In a different direction, Arjovsky & Bottou (2017) showed that training GANs using the original formulation (Goodfellow et al., 2014) leads to instability or vanishing gradients. To solve this issue, they proposed a Wasserstein GAN architecture (Arjovsky et al., 2017) where they restrict the class of discriminative functions to 1-Lipschitz functions using weight clipping. Pointing to issues with this clipping, Gulrajani et al. (2017); Miyato et al. (2018) proposed relaxed ways to enforce the Lipschitzness of the discriminator, either by using a gradient penalty or a spectral normalization. Albeit not exactly approximating the Wasserstein's distance (Petzka et al., 2018), both implementations lead to good empirical results, significantly reducing mode collapse. Building on all of these works, we will further assume that generators are now able to cover most of the modes of the target distribution, leaving us the problem of out-of-manifold samples (*a.k.a.* low-quality pictures).

Generation of disconnected manifolds. When learning complex manifolds in high dimensional spaces using deep generative models, Fefferman et al. (2016) highlighted the importance of understanding the underlying geometry. More precisely, the learning of disconnected manifold requires the introduction of disconnectedness in the model. Gurumurthy et al. (2017) used a multi-modal entry distribution, making the latent space disconnected, and showed better coverage when data is limited and diverse. Alternatively, Khayatkhoei et al. (2018) studied the learning of a mixture of generators. Using a mutual information term, they encourage each generator to focus on a different submanifold so that the mixture covers the whole support. This idea of using an ensemble of generators is also present in the work of Tolstikhin et al. (2017) and Zhong et al. (2019), though they were primarily interested in the reduction of mode collapse.

In this paper, we propose a truncation method to separate the latent space into several disjoint areas. It is a way to learn disconnected manifolds without relying on the previously introduced over-parameterization techniques. As our proposal can be applied without retraining the whole architecture, we can use it successfully on very larges nets. Close to this idea, Azadi et al. (2019) introduced a rejection strategy based on the output of the discriminator. However, this rejection sampling scheme requires the discriminator to be trained with a classification loss while our proposition can be applied to any generative models.

Evaluating GANs. The evaluation of generative models is an active area of research. Some of the proposed metrics only measure the quality of the generated samples such as the Inception score (Salimans et al., 2016) while others define distances between probability distributions. This is the case of the Frechet Inception distance (Heusel et al., 2017), the Wasserstein distance (Arjovsky et al., 2017) or kernel-based metrics (Gretton et al., 2012). The other main caveat for evaluating GANs lies in the fact that one does not have access to the true density nor the model density, prohibiting the use of any density based metrics. To solve this issue, the use of a third network that acts as an objective referee is common. For instance, the Inception score uses outputs from InceptionNet while the Fréchet Inception Distance compares statistics of InceptionNet activations. Since our work focuses on out-of-manifold samples, a natural measure is the PR measure (Sajjadi et al., 2018) and its Improved PR version (Kynkäänniemi et al., 2019), extensively discussed in the next section.

In the following, alongside precise definitions, we exhibit an upper bound on the precision of GANs with high recall (*i.e.* no mode collapse) and present a new truncation method.

3. Our approach

We start with a formal description of the framework of GANs and the relevant metrics. We later show a "no free lunch" theorem proving the necessary existence of an area in the latent space that generates out-of-manifold samples. We name this region the *no GAN's land* since any data point sampled from this area will be in the frontier in between two different modes. We claim that dealing with it requires special care. Finally, we propose a rejection sampling procedure to avoid points out of the true manifold.

3.1. Notations

In the original setting of Generative Adversarial Networks (GANs), one tries to generate data that are "similar" to samples collected from some unknown probability measure $\mu_\star$. To do so, we use a parametric family of generative distribution where each distribution is the push-forward measure of a latent distribution Z and a continuous function modeled by a neural network.

Assumption 1 (Z Gaussian)**.** *The latent distribution Z is a standard multivariate Gaussian.*

Note that for any distribution μ, S_μ refers to its support. Assumption 1 is common for GANs as in many practical applications, the random variable Z defined on a low dimensional space $\mathbb{R}^d$ is either a multivariate Gaussian. Practicioners also studied distribution or uniform distribution defined on a compact.

The measure $\mu_\star$ is defined on a subset E of $\mathbb{R}^D$ (potentially a highly dimensional space), equipped with the norm $\|\cdot\|$. The generator has the form of a parameterized class of functions from $\mathbb{R}^d$ (a space with a much lower dimension) to E, say $\mathscr{G} = \{G_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^p$ is the set of parameters describing the model. Each function G_θ thus takes input from a d-dimensional random variable Z (Z is associated with probability distribution γ) and outputs "fake" observations with distribution μ_θ. Thus, the class of probability measures $\mathscr{P} = \{\mu_\theta : \theta \in \Theta\}$ is the natural class of distributions associated with the generator, and the objective of GANs is to find inside this class of candidates the one that generates the most realistic samples, closest to the ones collected from the unknown distribution $\mu_\star$.

Assumption 2. *Let $L > 0$. The generator G_θ takes the form of a neural network whose Lipchitz constant is smaller than L, i.e. for all (z, z'), we have $\|G_\theta(z') - G_\theta(z)\| \leqslant L\|z - z'\|$.*

This is a reasonable assumption, since Virmaux & Scaman (2018) present an algorithm that upper-bounds the Lipschitz constant of deep neural networks. Initially, 1-Lipschitzness was enforced only for the discriminator by clipping the weigths (Arjovsky et al., 2017; Zhang et al., 2018), adding a gradient penalty (Gulrajani et al., 2017; Roth et al., 2017; Petzka et al., 2018), or penalizing the spectral norms (Miyato et al., 2018). Nowadays, state-of-the-art architectures for large scale generators such as SAGAN (Zhang et al., 2019) and BigGAN (Brock et al., 2019) also make use of spectral normalization for the generator.

3.2. Evaluating GANs with Precision and Recall

When learning disconnected manifolds, Srivastava et al. (2017) proved the need of measuring simultaneously the quality of the samples generated and the mode collapse. Sajjadi et al. (2018) proposed the use of a PR metric to measure the quality of GANs. The key intuition is that precision should quantify how much of the fake distribution can be generated by the true distribution while recall measures how much of the true distribution can be re-constructed by the model distribution. More formally, it is defined as follows:

Definition 1. *(Sajjadi et al., 2018) Let X, Y be two random variables. For $\alpha, \beta \in (0, 1]$, X is said to have an attainable precision α at recall β w.r.t. Y if there exists probability distributions μ, ν_X, ν_Y such that*

$$Y = \beta\mu + (1-\beta)\nu_Y \quad \textit{and} \quad X = \alpha\mu + (1-\alpha)\nu_X$$

The component ν_Y denotes the part of Y that is "missed" by X, whereas, ν_X denotes the "noise" part of X. We denote $\bar{\alpha}$ (respectively $\bar{\beta}$) the maximum attainable precision (respectively recall). Th. 1 of (Sajjadi et al., 2018) states:

$$X\big(S_Y\big) = \bar{\alpha} \quad \text{and} \quad Y\big(S_X\big) = \bar{\beta}$$

Improved PR metric. Kynkäänniemi et al. (2019) highlighted an important drawback of the PR metric proposed by Sajjadi et al. (2018): it cannot correctly interpret situations when a large numbers of samples are packed together. To better understand this situation, consider a case where the generator slightly collapses on a specific data point, i.e. there exists $x \in E, \mu_\theta(x) > 0$. We show in Appendix A that if $\mu_\star$ is a non-atomic probability measure and μ_θ is highly precise (*i.e.* $\alpha = 1$), then the recall β must be 0.

To solve these issues, Kynkäänniemi et al. (2019) proposed an *Improved Precision-Recall* (Improved PR) metric built on a nonparametric estimation of support of densities.

Definition 2. *(Kynkäänniemi et al., 2019) Let X, Y be two random variables and D_X, D_Y two finite sample datasets such that $D_X \sim X^n$ and $D_Y \sim Y^n$. For any $x \in D_X$ (respectively for any $y \in D_Y$), we consider $(x_{(1)}, \ldots, x_{(n-1)})$, the re-ordening of elements in $D_X \setminus x$ given their euclidean distance with x. For any $k \in \mathbb{N}$ and $x \in D_X$, the precision $\alpha_k^n(x)$ of point x is defined as*

$$\alpha_k^n(x) = 1 \iff \exists y \in D_Y, \|x - y\| \leqslant \|y_{(k)} - y\|$$

Similarly, the recall $\beta_k^n(y)$ of any given $y \in D_Y$ is

$$\beta_k^n(y) = 1 \iff \exists x \in D_X, \|y - x\| \leqslant \|x_{(k)} - x\|$$

Improved precision (respectively recall) are defined as the average over D_X (respectively D_Y) as follows

$$\alpha_k^n = \frac{1}{n} \sum_{x_i \in D_X} \alpha_k^n(x_i) \qquad \beta_k^n = \frac{1}{n} \sum_{y_i \in D_Y} \beta_k^n(y_i)$$

A first contribution is to formalize the link between PR and Improved PR with the following theorem:

Theorem 1. *Let X, Y two random variables with probability distributions μ and ν. Assume that both μ and ν are associated with uniformly continuous probability density functions f_μ and f_ν. Besides, there exists constants $a_1 > 0, a_2 > 0$ such that for all $x \in E$ we have $a_1 < f_{\mu_\star}(x) \leqslant a_2$ and $a_1 < f_{\mu_\theta}(x) \leqslant a_2$ for some $c > 0$. Also, (k, n) are such that $\frac{k}{\log(n)} \to +\infty$ and $\frac{k}{n} \to 0$. Then,*

$$\alpha_k^n \to \bar{\alpha} \text{ in probability} \quad \text{and} \quad \beta_k^n \to \bar{\beta} \text{ in proba.}$$

This theorem, whose proof is delayed to Appendix B, underlines the nature of the Improved PR metric: the metric compares the supports of the modeled probability distribution μ_θ and of the true distribution $\mu_\star$. This means that Improved PR is a tuple made of both maximum attainable precision $\bar{\alpha}$ and recall $\bar{\beta}$ (e.g. Theorem 1 of (Sajjadi et al., 2018)). As Improved PR is shown to have a better performance evaluating GANs sample quality, we use this metric for both the following theoretical results and experiments.

3.3. Learning disconnected manifolds

In this section, we aim to stress the difficulties of learning disconnected manifolds with standard GANs architectures. To begin with, we recall the following lemma.

Lemma 1. *Assume that Assumptions 1 and 2 are satisfied. Then, for any $\theta \in \Theta$, the support S_{μ_θ} is connected.*

There is consequently a discrepancy between the connectedness of S_{μ_θ} and the disconnectedness of $S_{\mu_\star}$. In the case where the manifold lays on two disconnected components, our next theorem exhibit a no free lunch theorem:

Theorem 2. *("No free lunch" theorem) Assume that Assumptions 1 and 2 are satisfied. Assume also that true distribution $\mu_\star$ lays on two equally measured disconnected manifolds distant from a distance $D > 0$. Then, any estimator μ_θ that samples equally in both modes must have a precision $\bar{\alpha}$ such that $\bar{\alpha} + \frac{D}{\sqrt{2\pi}L} e^{\frac{-\Phi^{-1}(\frac{\bar{\alpha}}{2})^2}{2}} \leqslant 1$, where Φ is the c.d.f. of a standard normal distribution.*

Besides, if $\bar{\alpha} \geqslant 3/4$, $\bar{\alpha} \lesssim 1 - \sqrt{\frac{2}{\pi}} W(\frac{D^2}{4L^2})$ where W is the Lambert W function.

The proof of this theorem is delayed to Appendix C. It is mainly based on the Gaussian isoperimetric inequality (Borell, 1975; Sudakov & Tsirelson, 1978) that states that among all sets of given Gaussian measure in any finite dimensional Euclidean space, half-spaces have the minimal Gaussian boundary measure. If in Fig. 1, the generator has thus learned the optimal separation, it is yet not known, to the limit of our knowledge, how to enforce such geometrical properties in the latent space.

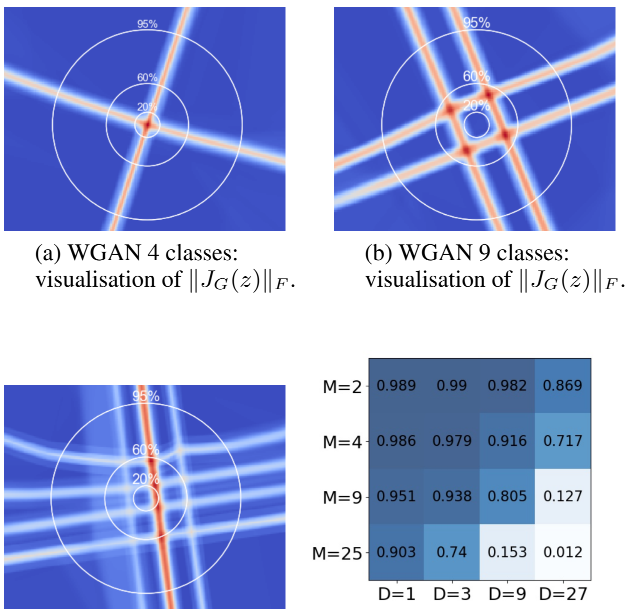
In real world applications, when the number of distinct sub-manifolds increases, we expect the volume of these boundaries to increase with respect to the number of different classes covered by the modeled distribution μ_θ. Going in this direction, we better formalize this situation, and show an extended "no free lunch theorem" by expliciting an upper-bound of the precision $\bar{\alpha}$ in this broader framework.

Assumption 3. *The true distribution $\mu_\star$ lays on M equally-measured disconnected components at least distant from some constant $D > 0$.*

This is likely to be true for datasets made of symbol designed to be highly distinguishable (*e.g.* digits in the MNIST dataset). In very high dimension, this assumption also holds for complex classes of objects appearing in many different contexts (*e.g.* the bubble class in ImageNet, see Appendix).

To better apprehend the next theorem, note A_m the pre-image in the latent space of mode m and A_m^r its r-enlargement: $A_m^r := \{z \in \mathbb{R}^d \mid \mathsf{dist(z, A_m)} \leq \mathsf{r}\}, \mathsf{r} > 0$.

Theorem 3. *(Generalized "no free lunch" theorem) Assume that Assumptions 1, 2, and 3 are satisfied, and that the pre-*



(a) WGAN 4 classes: visualisation of $\|J_G(z)\|_F$.

(b) WGAN 9 classes: visualisation of $\|J_G(z)\|_F$.

(c) WGAN 25 classes: visualisation of $\|J_G(z)\|_F$.

(d) Precision w.r.t. D (mode distance) and M (classes).

Figure 2. Illustration of Theorem 3. If the number of classes $M \to \infty$ or the distance $D \to \infty$, then the precision $\bar{\alpha} \to 0$. We provide in appendix heatmaps for more values of M.

image enlargements A^{ε}_{m}, with $\varepsilon = \frac{D}{2L}$, form a partition of the latent space with equally measured elements.

Then, any estimator μ_θ with recall $\bar{\beta} > \frac{1}{M}$ must have a precision $\bar{\alpha}$ at most $\frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x}$ where $x = \Phi^{-1}(1 - \frac{1}{\bar{\beta} M})$ and Φ is the c.d.f. of a standard normal distribution.

Theorem 3, whose proof is delayed to Appendix D, states a lower-bound the measure of samples mapped out of the true manifold. We expect our bound to be loose since no theoretical results are known, to the best of our knowledge, on the geometry of the separation that minimizes the boundary between different classes (when $M \geqslant 3$). Finding this optimal cut would be an extension of the honeycomb theorem (Hales, 2001). In Appendix D.2 we give a more technical statement of Theorem 3 without assuming equality of measure of the sets A^{ε}_{m}.

The idea of the proof is to consider the border of an individual cell with the rest of the partition. It is clear that at least half of the frontier will be inside this specific cell. Then, to get to the final result, we sum the measures of the frontiers contained inside all of the different cells. Remark that our analysis is fine enough to keep a dependency in M which translates into a maximum precision that goes to zero when M goes to the infinity and all the modes are covered. More precisely, in this scenario where all pre-images have equal measures in the latent space, one can derive the following bound, when the recall $\bar{\beta}$ is kept fixed and M increases:

$$\bar{\alpha} \overset{M\to\infty}{\leqslant} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon\sqrt{2\log(\bar{\beta} M)}} \quad \text{where } \varepsilon = \frac{D}{2L} \qquad (1)$$

For a fixed generator, this equation illustrates that the precision $\bar{\alpha}$ decreases when either the distance D (equivalently ε) or the number of classes M increases. For a given ε, $\bar{\alpha}$ converges to 0 with a speed $O(\frac{1}{(\bar{\beta}M)^{\sqrt{2\varepsilon}}})$. To better illustrate this asymptotic result, we provide results from a 2D synthetic setting. In this toy dataset, we control both the number M of disconnected manifolds and the distance D. Figure 2 clearly corroborates (1) as we can easily get the maximum precision close to 0 ($M = 25$, $D = 27$).

3.4. Jacobian-based truncation (JBT) method

The analysis of the deformation of the latent space offers a grasp on the behavior of GANs. For instance, Arvanitidis et al. (2017) propose a distance accounting for the distortions made by the generator. For any pair of points $(z_1, z_2) \sim Z^2$, the distance is defined as the length of the geodesic $d(z_1, z_2) = \int_{[0,1]} \|J_{G_\theta}(\gamma_t)\frac{d\gamma_t}{dt}\| dt$ where γ is the geodesic parameterized by $t \in [0, 1]$ and $J_{G_\theta}(z)$ denotes the Jacobian matrix of the generator at point z. Authors have shown that the use of this distance in the latent space improves clustering and interpretability. We make a similar observation that the generator's Jacobian Frobenius norm provides meaningful information.

Indeed, the frontiers highlighted in Figures 2a, 2b, and 2c correspond to areas of low precision mapped out of the true manifold: this is the *no GAN's land*. We argue that when learning disconnected manifolds, the generator tries to minimize the number of samples that do not belong to the support of the true distribution and that this can only be done by making paths steeper in the *no GAN's land*. Consequently, data points $G_\theta(z)$ with high Jacobian Frobenius norm (JFN) are more likely to be outside the true manifold. To improve the precision of generative models, we thus define a new truncation method by removing points with highest JFN.

However, note that computing the generators's JFN is expensive to compute for neural networks, since being defined as follows,

$$\|J_{G_\theta}(z)\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\partial G_\theta(z)_i}{\partial z_j}\right)^2,$$

it requires a number of backward passes equal to the output dimension. To make our truncation method tractable, we use a stochastic approximation of the Jacobian Frobenius norm based on the following result from Rifai et al. (2011):

$$\|J_{G_\theta}(z)\|^2 = \lim_{\substack{N\to\infty\\ \sigma\to 0}} \frac{1}{N}\sum_{\varepsilon_i}^{N} \frac{1}{\sigma^2}\|G_\theta(z+\varepsilon_i)-G_\theta(z)\|^2$$

where $\varepsilon_i \sim\sim \mathcal{N}(0,\sigma^2 I$ and I is the identity matrix of dimension d. The variance σ of the noise and the number of samples are used as hyper-parameters. In practice, σ in $[1\mathrm{e}{-4}; 1\mathrm{e}{-2}]$ and $N = 10$ give consistent results.

Based on the preceding analysis, we propose a new **Jacobian-based truncation** (JBT) method that rejects a certain ratio of the generated points with highest JFN. This truncation ratio is considered as an hyper-parameter for the model. We show in our experiments that our JBT can be used to to detect samples outside the real data manifold and that it consequently improves the precision of the generated distribution as measured by the Improved PR metric.

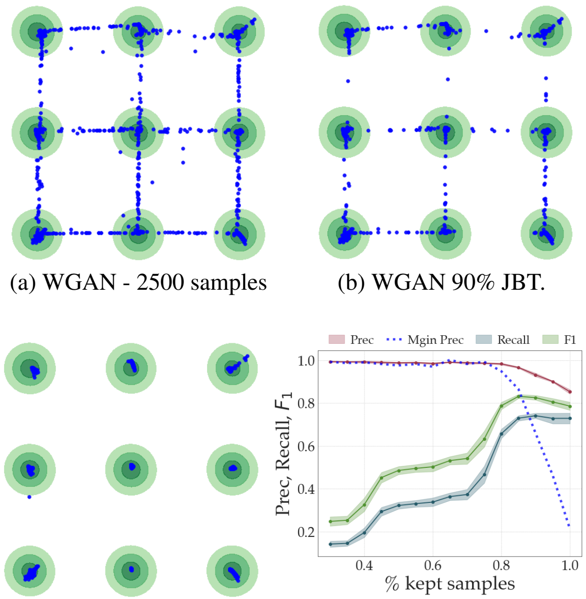
4. Experiments

In the following, we show that our truncation method, JBT, can significantly improve the performances of generative models on several models, metrics and datasets. Furthermore, we compare JBT with over-parametrization techniques specifically designed for disconnected manifold learning. We show that our truncation method reaches or surpasses their performance, while it has the benefit of not modifying the training process of GANs nor using a mixture of generators, which is computationally expensive. Finally, we confirm the efficiency of our method by applying it on top of BigGAN (Brock et al., 2019).

Except for BigGAN, for all our experiments, we use Wasserstein GAN with gradient penalty (Gulrajani et al., 2017), called WGAN for conciseness. We give in Appendix K the full details of our experimental setting. The use of WGAN is motivated by the fact that it was shown to stabilize the training and significantly reduce mode collapse (Arjovsky & Bottou, 2017). However, we want to emphasise that our method can be plugged on top of any generative model fitting disconnected components.

4.1. Evaluation metrics

To measure performances of GANs when dealing with low dimensional applications - as with synthetic datasets - we equip our space with the standard Euclidean distance. However, for high dimensional applications such as image generation, Brock et al. (2019); Kynkäänniemi et al. (2019) have shown that embedding images into a feature space with a pre-trained convolutional classifier provides more semantic information. In this setting, we consequently use the euclidean distance between the images' embeddings from a classifier. For a pair of images (a, b), we define the distance $d(a, b)$ as $d(a, b) = \|\phi(a) - \phi(b)\|_2$ where ϕ is a pre-softmax layer of a supervised classifier, trained specifically on each dataset. Doing so, they will more easily separate images sampled from the true distribution $\mu_\star$ from the ones sampled by the distribution μ_θ.



(a) WGAN - 2500 samples (b) WGAN 90% JBT.

(c) WGAN 70% JBT. (d) 97% confidence intervals .

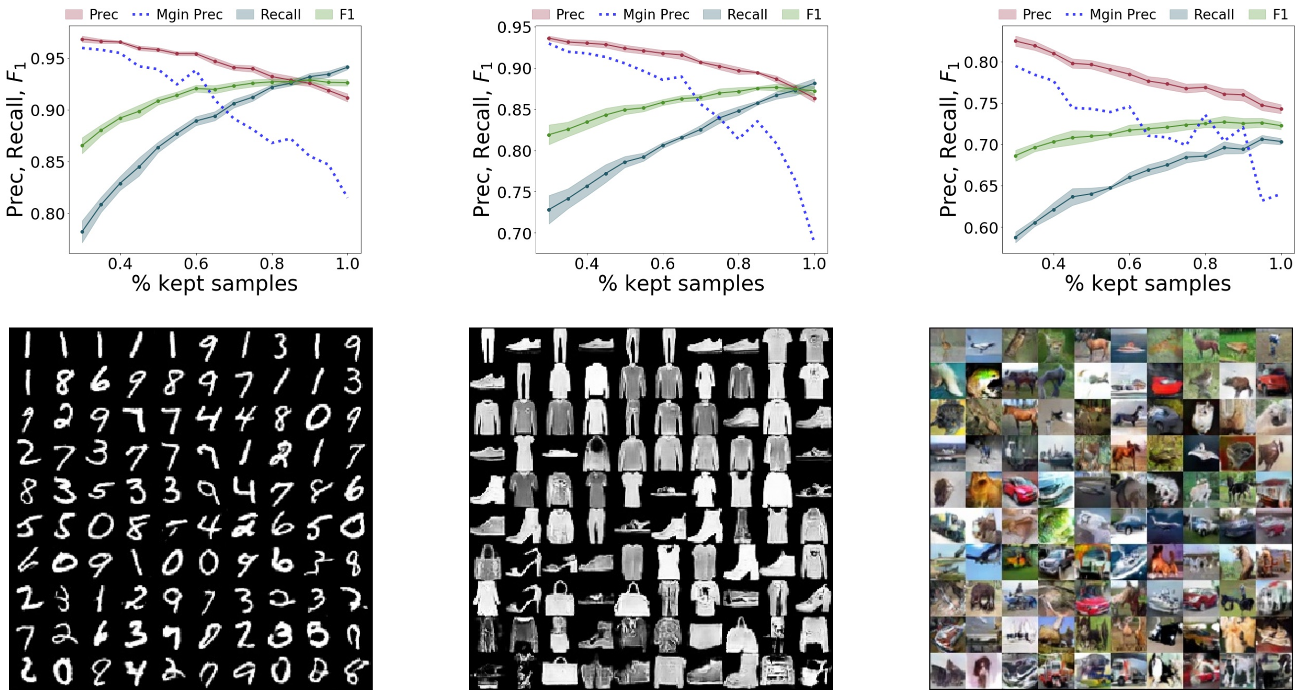
Figure 3. Mixture of 9 Gaussians in green, generated points in blue. Our truncation method (JBT) removes least precise data points as marginal precision plummets.

We compare performances using Improved PR (Kynkäänniemi et al., 2019). We also report the *Marginal Precision* which is the precision of newly added samples when increasing the ratio of kept samples. Besides, for completeness, we report FID (Heusel et al., 2017) and recall precise definitions in Appendix G. Note that FID was not computed with InceptionNet, but a classifier pre-trained on each dataset.

4.2. Synthetic dataset

We first consider the true distribution to be a 2D Gaussian mixture of 9 components. Both the generator and the discriminator are modeled with feed-forward neural networks.

Interestingly, the generator tries to minimize the sampling of off-manifolds data during training until its JFN gets saturated (see Appendix H). One way to reduce the number of off-manifold samples is to use JBT. Indeed, off-manifold data points progressively disappear when being more and more selective, as illustrated in Figure 3c. We quantitatively confirm that our truncation method (JBT) improves the precision. On Fig. 3d, we observe that keeping the 70% of lowest JFN samples leads to an almost perfect precision of the support of the generated distribution. Thus, off-manifold samples are in the 30% samples with highest JFN.



(a) MNIST dataset.

(b) F-MNIST dataset.

(c) CIFAR10 datatset.

Figure 4. For high levels of kept samples, the marginal precision plummets of newly added samples, underlining the efficiency of our truncation method (JBT). Reported confidence intervals are 97% confidence intervals. On the second row, generated samples ordered by their JFN (left to right, top to bottom). In the last row, the data points generated are blurrier and outside the true manifold.

4.3. Image datasets

We further study JBT on three different datasets: MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017) and CIFAR10 (Krizhevsky et al., 2009). Following (Khayatkhoei et al., 2018) implementation, we use a standard CNN architecture for MNIST and FashionMNIST while training a ResNet-based model for CIFAR10 (Gulrajani et al., 2017).

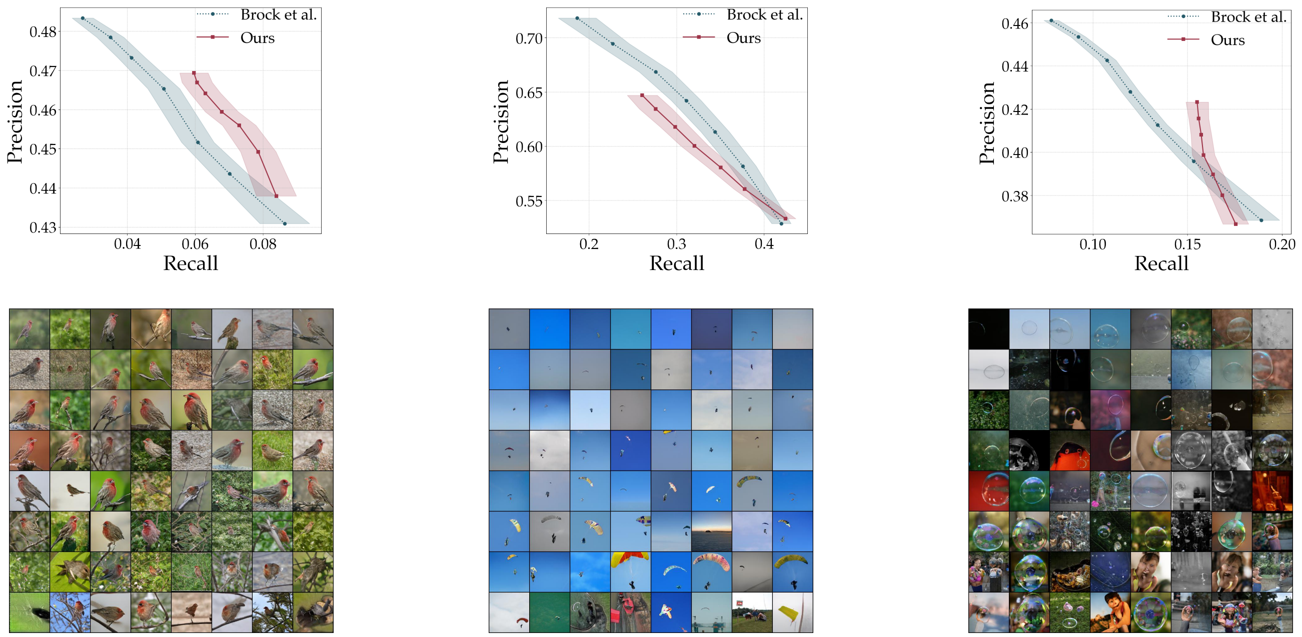
MNIST	Prec.	Rec.	FID
WGAN	91.2±0.3	**93.7**±0.5	24.3±0.3
WGAN JBT 90%	92.5±0.5	92.9±0.3	26.9±0.5
WGAN JBT 80%	**93.3**±0.3	91.8±0.4	33.1±0.3
W-Deligan	89.0±0.6	**93.6**±0.3	31.7±0.5
DMLGAN	**93.4**±0.2	92.3±0.2	**16.8**±0.4
F-MNIST			
WGAN	86.3±0.4	**88.2**±0.2	259.7±3.5
WGAN JBT 90%	88.6±0.6	86.6±0.5	**257.4**±3.0
WGAN JBT 80%	**89.8**±0.4	84.9±0.5	396.2±6.4
W-Deligan	88.5±0.3	85.3±0.6	310.9±3.1
DMLGAN	87.4±0.3	**88.1**±0.4	**253.0**±2.8

Table 1. JBT $x\%$ means we keep the $x\%$ samples with lowest Jacobian norm. Our truncation method (JBT) matches over-parameterization techniques. ± is 97% confidence interval.

Figure 4 highlights that JBT also works on high dimensional datasets as the marginal precision plummets for high truncation ratios. Furthermore, when looking at samples ranked by increasing order of their JFN, we notice that samples with highest JFN are standing in-between manifolds. For example, those are ambiguous digits resembling both a "0" and a "6" or shoes with unrealistic shapes.

To further assess the efficiency of our truncation method, we also compare its performances with two state-of-the-art over-parameterization techniques that were designed for disconnected manifold learning. First, (Gurumurthy et al., 2017) propose DeliGAN, a reparametrization trick to transform the unimodal Gaussian latent distribution into a mixture. The different mixture components are later learnt by gradient descent. For fairness, the re-parametrization trick is used on top of WGAN. Second, (Khayatkhoei et al., 2018) define DMLGAN, a mixture of generators to better learn disconnected manifolds. In this architecture, each generator is encouraged to target a different submanifold by enforcing high mutual information between generated samples and generator's ids. Keep in mind that for DeliGAN (respectively DMLGAN), the optimal number of components (respectively generators) is not known and is a hyper-parameter of the model that has to be cross-validated.

The results of the comparison are presented in Table 1. In both datasets, JBT 80 % outperforms DeliGAN and DMLGAN in terms of precision while keeping a reasonnable



(a) House finch.

(b) Parachute.

(c) Bubble.

Figure 5. On the first row, per-class precision-recall curves comparing Brock et al. (2019)'s truncation trick and our truncation method (JBT), on three ImageNet classes generated by BigGAN. We show better results on complex and disconnected classes (*e.g.* bubble). Reported confidence intervals are 97% confidence intervals. On the second row, generated samples ordered by their JFN (left to right, top to bottom). We observe a concentration of off-manifold samples for images on the bottom row, confirming the soundness of JBT.

recall. This confirms our claim that over-parameterization techniques are unnecessary. As noticed by Kynkäänniemi et al. (2019), we also observe that FID does not correlate properly with the Improved PR metric. Based on the Frechet distance, only a distance between multivariate Gaussians, we argue that FID is not suited for disconnected manifold learning as it approximates distributions with unimodal ones and looses many information.

4.4. Spurious samples rejections on BigGAN

Thanks to the simplicity of JBT, we can also apply it on top of any trained generative model. In this subsection, we use JBT to improve the precision of a pre-trained BigGAN model (Brock et al., 2019), which generates class-conditionned ImageNet (Deng et al., 2009) samples. The class-conditioning lowers the problem of off-manifold samples, since it reduces the disconnectedness in the output distribution. However, we argue that the issue can still exist on high-dimensional natural images, in particular complex classes can still be multi-modal (*e.g.* the bubble class). The bottom row in Figure 5 shows a random set of 128 images for three different classes ranked by their JFN in ascending order (left to right, top to bottom). We observe a clear concentration of spurious samples on the bottom row images.

To better assess the Jacobian based truncation method, we compare it with the truncation trick from Brock et al. (2019). This truncation trick aims to reduce the variance of the latent space distribution using truncated Gaussians. While easy and effective, this truncation has some issues: it requires to complexify the loss to enforce orthogonality in weight matrices of the network. Moreover, as explained by Brock et al. (2019) *"only 16% of models are amenable to truncation, compared to 60% when trained with Orthogonal Regularization"*. For fairness of comparison, the pre-trained network we use is optimized for their truncation method. On the opposite, JBT is simpler to apply since 100% of the tested models were amenable to the proposed truncation.

Results of this comparison are shown in the upper row of Figure 5. Our method can outperform their truncation trick on difficult classes with high intra-class variation, *e.g.* bubble and house finch. This confirms our claim that JBT can detect outliers within a class. However, one can note that their trick is particularly well suited for simpler unimodal classes, *e.g.* parachute and reaches high precision levels.

5. Conclusion

In this paper, we provide insights on the learning of disconnected manifolds with GANs. Our analysis shows the existence of an off-manifold area with low precision. We empirically show on several datasets and models that we can detect these areas and remove samples located in between two modes thanks to a newly proposed truncation method.

References

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: on the curvature of deep generative models. In *ICLR*, 2017.

Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., and Odena, A. Discriminator rejection sampling. In *International Conference on Learning Representations*, 2019.

Biau, G. and Devroye, L. *Lectures on the nearest neighbor method*. Springer, 2015.

Biau, G., Cadre, B., Sangnier, M., and Tanielian, U. Some theoretical properties of gans. *arXiv:1803.07819*, 2018.

Borell, C. The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devroye, L. and Wise, G. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38:480–488, 1980.

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, pp. 450–455, 1982.

Dudley, R. M. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002. doi: 10.1017/CBO9780511755347.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, J. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. 2014.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, pp. 723–773, 2012.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Gurumurthy, S., Kiran Sarvadevabhatla, R., and Venkatesh Babu, R. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Hales, T. C. The honeycomb conjecture. *Discrete & Computational Geometry*, pp. 1–22, 2001.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Kallenberg, O. *Foundations of modern probability*. Springer Science & Business Media, 2006.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Khayatkhoei, M., Singh, M. K., and Elgammal, A. Disconnected manifold learning for generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 7343–7353, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, pp. 3929–3938, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Ledoux, M. Isoperimetry and gaussian analysis. In *Lectures on probability theory and statistics*, pp. 165–294. Springer, 1996.

Lin, Z., Khetan, A., Fanti, G., and Oh, S. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 1498–1507, 2018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Pandeva, T. and Schubert, M. Mmgan: Generative adversarial networks for multi-modal distributions. *arXiv:1911.06663*, 2019.

Petzka, H., Fischer, A., and Lukovnikov, D. On the regularization of wasserstein GANs. In *International Conference on Learning Representations*, 2018.

Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. Higher order contractive auto-encoder. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 645–660. Springer, 2011.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems*, pp. 2018–2028, 2017.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pp. 5228–5237, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pp. 3308–3318, 2017.

Sudakov, V. N. and Tsirelson, B. S. Extremal properties of half-spaces for spherically invariant measures. *Journal of Mathematical Sciences*, pp. 9–18, 1978.

Tolstikhin, I. O., Gelly, S., Bousquet, O., Simon-Gabriel, C.-J., and Schölkopf, B. Adagan: Boosting generative models. In *Advances in Neural Information Processing Systems*, pp. 5424–5433, 2017.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pp. 3835–3844, 2018.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.

Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7354–7363, 2019.

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. On the discriminative-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

Zhong, P., Mo, Y., Xiao, C., Chen, P., and Zheng, C. Rethinking generative mode coverage: A pointwise guaranteed approach. In *Advances in Neural Information Processing Systems*, pp. 2086–2097, 2019.