A. Highlighting drawbacks of the PR metric by Sajjadi et al. (2018)

Lemma 2. *Assume that the modeled distribution μ_θ slightly collapses on a specific data point, i.e. there exists $x \in E, \mu_\theta(x) > 0$. Assume also that $\mu_\star$ is a continuous probability measure and that μ_θ has a recall $\beta = 1$. Then the precision must be such that $\alpha = 0$.*

Proof. Using Definition 1, we have that there exists μ such that

$$\mu_\star = \alpha\mu + (1-\alpha)\nu_{\mu_\star} \quad \text{and} \quad \mu_\theta = \mu$$

Thus, $0 = \mu_\star(x) \geqslant \alpha\mu(x) = \alpha\mu_\theta(x)$. Which implies that $\alpha = 0$. □

B. Proof of Theorem 1

The proof of Theorem 1 relies on theoretical results from non-parametric estimation of the supports of probability distribution studied by Devroye & Wise (1980).

For the following proofs, we will require the following notation: let φ be a strictly monotonous function be such that $\lim\limits_{n\to\infty} \frac{\varphi(n)}{n} = 0$ and $\lim\limits_{n\to\infty} \frac{\varphi(n)}{\log(n)} = \infty$. We note $B(x,r) \subseteq E$, the open ball centered in x and of radius r. For a given probability distribution μ, S_μ refers to its support. We recall that for any x in a dataset D, $x_{(k)}$ denotes its k nearest neighbor in D. Finally, for a given probability distribution μ and a dataset D_μ sampled from μ^n, we note $R_{\min}$ and $R_{\max}$ the following:

$$R_{\min} = \min_{x\in E}\|x - x_{(\varphi(n))}\|, \quad R_{\max} = \max_{x\in E}\|x - x_{(\varphi(n))}\| \tag{2}$$

In the following lemma, we show asymptotic behaviours for both $R_{\min}$ and $R_{\max}$.

Lemma 3. *Let μ be a probability distribution associated with a uniformly continuous probability density function f_μ. Assume that there exists constants $a_1 > 0, a_2 > 0$ such that for all $x \in E$, we have $a_1 < f_\mu(x) \leqslant a_2$. Then,*

$$R_{min} \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.} \quad \text{and} \quad R^d_{min} \underset{n\to\infty}{\longrightarrow} \infty \text{ a.s.}$$
$$R_{max} \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.} \quad \text{and} \quad R^d_{max} \underset{n\to\infty}{\longrightarrow} \infty \text{ a.s.}$$

Proof. We will only prove that $R_{\max} \underset{n\to\infty}{\longrightarrow} 0$ a.s. and and $R^d_{\min} \underset{n\to\infty}{\longrightarrow} \infty$ a.s. as the rest follows.

The result is based on a nearest neighbor result from Biau & Devroye (2015). Considering the $\varphi(n)$ nearest neighbor density estimate $f_n^{\varphi(n)}$ based on a finite sample dataset D_μ, Theorem 4.2 states that if f_μ is uniformly continuous then:

$$\sup_{x\in E} \|f_n^{\varphi(n)}(x) - f_\mu(x)\| \to 0.$$

where $f_n^{\varphi(n)}(x) = \frac{\varphi(n)}{nV_d\|x - x_{\varphi(n)}\|^d}$ with V_d being the volume of the unit ball in $\mathbb{R}^d$.

Let $\varepsilon > 0$ such that $\varepsilon < a_1/2$. There exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, we have, almost surely, for all $x \in E$:

$$a_1 - \varepsilon \leqslant f_n^{\varphi(n)}(x) \leqslant a_2 + \varepsilon$$
$$a_1 - \varepsilon \leqslant \frac{\varphi(n)}{nV_d\|x - x_{\varphi(n)}\|^d} \leqslant a_2 + \varepsilon$$

Consequently, for all $n \geqslant N$, for all $x \in E$ almost surely:

$$\|x - x_{\varphi(n)}\| \leqslant \Big(\frac{\varphi(n)}{nV_d(a_1 - \varepsilon)}\Big)^{1/d}$$
$$\text{Thus ,}\sup_{x\in E} \|x - x_{\varphi(n)}\| \to 0 \quad \text{a.s.}$$

Also, almost surely

$$n\|x - x_{\varphi(n)}\|^d \geqslant \frac{\varphi(n)}{V_d(a_2 + \varepsilon)}$$
$$\text{Thus,} \quad \inf_{x\in E} \|x - x_{\varphi(n)}\| \to \infty \quad \text{a.s.}$$

□

Lemma 4. *Let μ, ν be two probability distributions associated with uniformly continuous probability density functions f_μ and f_ν. Assume that there exists constants $a_1 > 0, a_2 > 0$ such that for all $x \in E$, we have $a_1 < f_\mu(x) \leqslant a_2$ and $a_1 < f_\nu \leqslant a_2$. Also, let D_μ, D_ν be datasets sampled from ν^n, μ^n. If μ is an estimator for ν, then*

$$(i) \text{ for all } x \in D_\mu,\ \alpha^n_{\varphi(n)}(x) \underset{n\to\infty}{\to} \mathbb{1}_{\text{supp}(\nu)}(x) \quad \text{in proba.}$$
$$(ii) \text{ for all } y \in D_\nu,\ \beta^n_{\varphi(n)}(y) \underset{n\to\infty}{\to} \mathbb{1}_{\text{supp}(\mu)}(x) \quad \text{in proba.}$$

Proof. We will only show the result for (i), since a similar proof holds for (ii).

Thus, we want to show that

$$\text{for all } x \in D_\mu,\ \alpha^n_{\varphi(n)}(x) \underset{n\to\infty}{\to} \mathbb{1}_{supp(\nu)}(x) \quad \text{a. s.}$$

First, let's assume that $x \notin S_\nu$. Biau & Devroye (2015, Lemma 2.2) have shown that

$$\lim_{n\to\infty} \|x_{(\varphi(n))} - x\| = \inf\{\|x - y\| \mid y \in S_\nu\} \quad \text{a.s.}$$

As S_ν is a closed set - e.g. (Kallenberg, 2006) - we have

$$\lim_{n\to\infty} \|x - x_{(\varphi(n))}\| > 0 \quad \text{a.s.}$$

and

$$\text{for all } y \in D_\nu,\ \lim_{n\to\infty} \|y - y_{(\varphi(n))}\| = 0 \quad \text{a.s.}$$

Thus, $\lim\limits_{n\to\infty} \alpha^n_{\varphi(n)}(x) = 0 \quad$ a.s..

Now, let's assume that $x \in S_\nu$. Using Definition 2, the precision of a given data point x can be rewritten as follows:

$$\alpha^n_{\varphi(n)}(x) = 1 \iff \exists y \in D_\nu, x \in B(y, \|y - y_{(\varphi(n))}\|)$$

Using notation from (2), we note

$$R_{\text{min}} = \min_{y\in} \|y - y_{(\varphi(n))}\|, \;\; R_{\text{max}} = \max_{y\in E} \|y - y_{(\varphi(n))}\|.$$

It is clear that :

$$\bigcup_{y\in D_\nu} B(y, R_{\text{min}}) \subseteq S^n_\nu \subseteq \bigcup_{y\in D_\nu} B(y, R_{\text{max}}), \tag{3}$$

where $S^n_\nu = \bigcup_{y\in D_\nu} B(y, \|y - y_{(\varphi(n))}\|))$.

Besides, combining Lemma 3 with Devroye & Wise (1980, Theorem 1), we have that:

$$\nu(S_\nu \Delta \bigcup_{y\in D_\nu} B(y, R_{\text{min}})) \underset{n\to 0}{\longrightarrow} 0 \quad \text{in proba.}$$

$$\nu(S_\nu \Delta \bigcup_{y\in D_\nu} B(y, R_{\text{max}})) \underset{n\to 0}{\longrightarrow} 0 \quad \text{in proba.}$$

where Δ here refers to the symmetric difference.

Thus, using (3), it is now clear that, $\mu(S_\nu \Delta S^n_\nu) \to 0$ in probability. Finally, given $x \in S_\mu$, we have $\mu(x \in S^n_\nu) = \nu(\alpha^n_{\varphi(n)}(x) = 1) \to 1$ in probability. □

We can now finish the proof for Theorem 1. Recall that $\bar{\alpha} = \mu\big(S_\nu\big)$ and similarly, $\bar{\beta} = \nu\big(S_\mu\big)$.

Proof. We have that

$$|\alpha^n_{\varphi(n)} - \bar{\alpha}| = |\frac{1}{n}\sum_{x_i \in D_\mu} \alpha^n_{\varphi(n)}(x_i) - \int_E \mathbb{1}_{x\in S_\nu}\mu(\mathrm{d}x)|$$

Then,

$$\begin{aligned} |\alpha^n_{\varphi(n)} - \bar{\alpha}| = |\frac{1}{n}\sum_{x_i\in D_\mu} (\alpha^n_{\varphi(n)}(x_i) - \mathbb{1}_{x_i\in S_\nu}) \\ +(\frac{1}{n}\sum_{x_i\in D_\mu} \mathbb{1}_{x_i\in S_\nu} - \int_E \mathbb{1}_{x\in S_\nu}\mu(\mathrm{d}x))| \\ = |\mathbb{E}_{x_i\sim\mu_n}(\alpha^n_{\varphi(n)}(x_i) - \mathbb{1}_{x_i\in S_\nu}) \quad (4) \\ + \big(\mathbb{E}_{\mu_n}\mathbb{1}_{S_\nu} - \mathbb{E}_\mu \mathbb{1}_{S_\nu}\big)| \quad (5) \end{aligned}$$

where μ_n is the empirical distribution of μ. As μ_n converges weakly to μ almost surely (e.g. Dudley (2002, Theorem 11.4.1)) and since $\mathbb{1}_{x\in S_\nu}$ is bounded, we can bound (5) as follows:

$$\lim_{n\to\infty} \quad \mathbb{E}_{x\sim\mu_n}\mathbb{1}_{x\in\text{supp}(\mu)} - \mathbb{E}_{x\sim\mu}\mathbb{1}_{x\in\text{supp}(\mu)} = 0 \quad \text{a. s.}$$

Now, to bound (4), we use the fact that for any $x \in D_\mu$, the random variable $\alpha^n_{\varphi(n)}(x)$ converges to $\mathbb{1}_{x\in S_\nu}$ in probability (Lemma 4) and that for all $x \in D_\mu$, both $\alpha^n_{\varphi(n)}(x) \leqslant 1$ and $\mathbb{1}_{x\in S_\nu} \leqslant 1$. Consequently, using results from the weak law for triangular arrays, we have that

$$\lim_{n\to\infty} \quad \frac{1}{n}\sum_{x_i\in D_\mu} (\alpha^n_{\varphi(n)}(x_i) - \mathbb{1}_{x_i\in S_\nu}) = 0 \quad \text{in proba.}$$

Finally,

$$|\alpha^n_{\varphi(n)} - \bar{\alpha}| \underset{n\to\infty}{\to} 0 \quad \text{in proba.},$$

which proves the result. The same proof works for $\lim\limits_{k\to\infty} \beta^n_k = \bar{\beta}$. □

C. Proof of Theorem 2

This proof is based on the Gaussian isoperimetric inequality historically shown by (Borell, 1975; Sudakov & Tsirelson, 1978).

Proof. Let $\mu_\star$ be a distribution defined on E laying on two disconnected manifolds M_1 and M_2 such that $\mu_\star(M_1) = \mu_\star(M_2) = \frac{1}{2}$ and $d(M_1, M_2) = D$. Note that for any subsets $A \subseteq E$ and $B \subseteq E$, $d(A,B) := \inf\limits_{(x,y)\in A\times B} \|x - y\|$.

Let $G^{-1}_\theta(M_1)$ (respectively $G^{-1}_\theta(M_2)$ be the subset in $\mathbb{R}^d$ be the pre-images of M_1 (respectively M_2).

Consequently, we have for all $k \in [1, n]$

$$\gamma(G^{-1}_\theta(M_1)) = \mu_\theta(M_1) = \gamma(G^{-1}_\theta(M_2)) \geqslant \frac{\bar{\alpha}}{2}$$

We consider $(G^{-1}_\theta(M_1))^\varepsilon$ (respectively $(G^{-1}_\theta(M_2))^\varepsilon$) the ε enlargement of $G^{-1}_\theta(M_1)$ (respectively $G^{-1}_\theta(M_2)$ where $\varepsilon = \frac{D}{2L}$. We know that $(G^{-1}_\theta(M_1))^\varepsilon \bigcap (G^{-1}_\theta(M_2))^\varepsilon = \emptyset$.

Thus, we have that:

$$\gamma\big((G^{-1}_\theta(M_1))^\varepsilon\big) + \gamma\big((G^{-1}_\theta(M_2))^\varepsilon\big) \leqslant 1$$

Besides, by denoting Φ the function defined for any $t \in \mathbb{R}$ by $\Phi(t) = \int_{-\infty}^t \frac{\exp(-t^2/2)}{\sqrt{2\pi}} ds$, we have

$$\gamma\big((G^{-1}_\theta(M_1))^\varepsilon\big) + \gamma\big((G^{-1}_\theta(M_2))^\varepsilon\big) \geqslant 2\Phi\big(\Phi^{-1}(\frac{\alpha}{2}) + \varepsilon\big)$$

(using Theorem 1.3 from (Ledoux, 1996))

$$\geqslant \alpha + \frac{2\varepsilon}{\sqrt{2\pi}} e^{-\Phi^{-1}(\frac{\alpha}{2})^2/2}$$

(since $\Phi^{-1}(\frac{\alpha}{2}) + \varepsilon < 0$ and Φ convex on $]-\infty, 0]$)

Thus, we have that

$$\alpha + \frac{2\varepsilon}{\sqrt{2\pi}} e^{-\Phi^{-1}(\frac{\alpha}{2})^2/2} \leqslant 1$$

Thus, by noting

$$\alpha^\star = \sup\{\alpha \in [0,1] \mid \alpha + \frac{2\varepsilon}{\sqrt{2\pi}} e^{\frac{-\Phi^{-1}(\frac{\alpha}{2})^2}{2}} \leqslant 1\},$$

we have our result.

For $\alpha \geqslant 3/4$. By noting $\alpha = 1 - x$, we have

$$\begin{aligned}
\Phi^{-1}(\frac{\alpha}{2}) &= \frac{\sqrt{2\pi}x}{2} + O(x^3) \\
\text{And, } e^{\frac{-\Phi^{-1}(\frac{\alpha}{2})^2}{2}} &= e^{\frac{-\pi x^2}{4}} + O(e^{-x^4}) \\
\text{Thus, } 1 - x + \frac{2\varepsilon}{\sqrt{2\pi}} e^{\frac{-\pi x^2}{4}} + O(e^{-x^4}) &\leqslant 1 \\
\iff x &\geqslant \frac{2\varepsilon}{\sqrt{2\pi}} e^{\frac{-\pi x^2}{4}} + O(e^{-x^4}) \\
\implies x &\geqslant \sqrt{\frac{2}{\pi}} W(\epsilon^2)
\end{aligned}$$

where W is the product log function. Thus, $\alpha \leqslant 1 - \sqrt{\frac{2}{\pi}} W(\epsilon^2)$. □

As an example, in the case where $\varepsilon = 1$, we have that $W(1) \approx 0.5671$, $x > 0.4525$ and $\alpha < 0.5475$.

D. Proof of Theorem 3

D.1. Equitable setting

This result is a consequence of Prop. 1 that we will assume true in this section.

We consider that the unknown true distribution $\mu_\star$ lays on M disjoint manifolds of equal measure. As specified in Section 3, the latent distribution γ is a multivariate Gaussian defined on $\mathbb{R}^d$. For each $k \in [1, M]$, we consider in the latent space, the pre-images A_k.

It is clear that $A_1, \ldots, A_M$ are pairwise disjoint Borel subsets of $\mathbb{R}^d$. We denote $\bar{M}$, the number of classes covered by the estimator μ_θ, such that for all $i \in [1, \bar{M}]$, we have $\gamma(A_i) > 0$. We know that $\bar{M} \geqslant M\bar{\beta} > 1$.

For each $i \in [1, \bar{M}]$, we denote A_i^ε, the ε-enlargement of A_i. For any pair (i, j) it is clear that $A_i^\varepsilon \bigcap A_j^\varepsilon = 0$ where $\varepsilon = \frac{D}{2L}$ (D being the minimum distance between two sub-manifolds and L being the Lipschitz constant of the generator).

As assumed, we know that $A_i^\varepsilon, i \in [1, \bar{M}]$ partition the latent space in equal measure, consequently, we assume that

$$\sum_{i=1}^{n} \gamma(A_i^\varepsilon) = 1 \quad \text{and} \quad \gamma(A_1) = \ldots = \gamma(A_{\bar{M}}) = 1/\bar{M} \tag{6}$$

Thus, we have that

$$\bar{\alpha} = \sum_{i=1}^{\bar{M}} \gamma(A_i^\varepsilon) = 1 - \gamma(\Delta^{-\varepsilon}(A_1^\varepsilon, \ldots, A_{\bar{M}}^\varepsilon))$$

Using Proposition 1, we have

$$\begin{aligned}
\gamma(\Delta^{-\varepsilon}(A_1^\varepsilon, \ldots, A_n^\varepsilon)) &\geqslant 1 - \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x} \\
\text{Thus, } \bar{\alpha} &\leqslant \frac{1+y^2}{y^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon y}
\end{aligned}$$

where $y = \Phi^{-1}\left(1 - \max_{k \in [\bar{M}]} \gamma(A_k^\varepsilon)\right) = \Phi^{-1}(\frac{\bar{M}-1}{\bar{M}})$ and $\Phi(t) = \int_{-\infty}^{t} \frac{\exp(-t^2/2)}{\sqrt{2\pi}} ds$.

Knowing that $\bar{M} \geqslant \bar{\beta} M$ we have that

$$\Phi^{-1}(1 - \frac{1}{\bar{M}}) \geqslant \Phi^{-1}(1 - \frac{1}{\bar{\beta} M})$$

We conclude by saying that the function $x \mapsto \frac{1+x^2}{x^2} e^{-\varepsilon x}$ is decreasing for $x > 0$. Thus,

$$\bar{\alpha} \leqslant \frac{1+y^2}{y^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon y} \tag{7}$$

where $y = \Phi^{-1}(1 - \frac{1}{\bar{\beta} M})$ and $\Phi(t) = \int_{-\infty}^{t} \frac{\exp(-t^2/2)}{\sqrt{2\pi}} ds$.

For further analysis, when $\bar{M} \to \infty$, refer to subsection E and note using the result in (15) that one obtains the desired upper-bound on $\bar{\alpha}$

$$\bar{\alpha} \overset{\bar{M} \to \infty}{\leqslant} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon \sqrt{2 \log(\bar{M})}}$$

D.2. More general setting

As done previously, we denote $\bar{M}$, the number of classes covered by the estimator μ_θ, such that for all $i \in [1, \bar{M}]$, we have $\gamma(A_i) > 0$. We still assume that $\bar{M} > 1$. However, we now relax the previous assumption made in (6) and assume the milder assumption that there exists $w_1, \ldots, w_M \in [0,1]^M$ such that for all $m \in [1, M], \gamma(A_m^\varepsilon) = w_m$, $\sum_m w_m \leqslant 1$ and $\max\limits_{i \in [1,M]} w_m = w^{\text{max}} < 1$.

Consider, $A^{\complement} = \left(\bigcup_{i=1}^{\bar{M}} A_i^\varepsilon\right)^{\complement}$ and denote $w^c = \gamma(A^{\complement}) \leqslant 1 - \bar{\alpha}$. Consequently, we have

$$\sum_{i=1}^{n} \gamma(A_i^\varepsilon) + \gamma(A^{\complement}) = 1$$

$$\gamma(\Delta^{-\varepsilon}(A_1^\varepsilon, \ldots, A_M^\varepsilon, A^{\complement})) + \sum_{i=1}^{M} \gamma(A_i^\varepsilon) = 1 - \gamma(A^{\complement})$$

$$\bar{\alpha} = 1 - w^{\complement} - \gamma(\Delta^{-\varepsilon}(A_1^\varepsilon, \ldots, A_M^\varepsilon, A^{\complement}))$$

In this setting, it is clear that $A_1, \dots, A_{\bar{M}}, A^{\complement}$ is a a partition of $\mathbb{R}^d$ under the measure γ. Using, result from Proposition 1, we have

$$\gamma(\Delta^{-\varepsilon}(A_1^{\varepsilon}, \dots, A_M^{\varepsilon}, A^{\complement})) \geqslant 1 - \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x}$$

where $x = \Phi^{-1}\left(1 - \max(w^{\complement}, w^{\max})\right)$ and $\Phi(t) = \int_{-\infty}^{t} \frac{\exp(-t^2/2)}{\sqrt{2\pi}} ds$.

Finally, we have that

$$\bar{\alpha} \leqslant \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x} - w^{\complement} \tag{8}$$

In the case where $\gamma(A^{\complement}) = 0$, we find a result similar to (7).

E. Lower-bounding boundaries of partitions in a Gaussian space

Notations and preliminaries Given $\varepsilon \geq 0$ and a subset A of euclidean space $\mathbb{R}^d = (\mathbb{R}^d, \|\cdot - \cdot\|)$, let $A^{\varepsilon} := \{z \in \mathbb{R}^d \mid \mathsf{dist(z, A)} \leq \varepsilon\}$ be its ε-enlargement, where $\mathsf{dist(z, A)} := \inf_{\mathsf{z' \in A}} \|\mathsf{z' - z}\|_2$ is the distance of the point $z \in \mathbb{R}^d$ from A. Let γ be the standard Gaussian distribution in $\mathbb{R}^d$ and let $A_1, \dots, A_K$ be $K \geq 2$ pairwise disjoint Borel subsets of $\mathbb{R}^d$ whose union has unit (i.e full) Gaussian measure $\sum_{k=1}^{K} w_k = 1$, where $w_k := \gamma(A_k)$. Such a collection $\{A_1, \dots, A_K\}$ will be called an $(w_1, \dots, w_K)$-partition of standard d-dimensional Gaussian space $(\mathbb{R}^d, \gamma)$.

For each $k \in [\![K]\!]$, define the compliment $A_{-k} := \cup_{k' \neq k} A_{k'}$, and let $\partial^{-\varepsilon} A_k := \{z \in A_k \mid \mathsf{dist(z, A_{-k})} \leq \varepsilon\}$ be the *inner ε-boundary* of A_k, i.e the points of A_k which are within distance ε of some other $A_{k'}$. For every $(k, k') \in [\![K]\!]^2$ with $k' \neq k$, it is an easy exercise to show that

$$\begin{aligned} \partial^{-\varepsilon} A_k \cap \partial^{-\varepsilon} A_{k'} &= \emptyset \\ \partial^{-\varepsilon} A_k \cap A_{-k} &= \emptyset \\ A_{-k}^{\varepsilon} &= \partial^{-\varepsilon} A_k \cup A_{-k} \end{aligned} \tag{9}$$

Now, let $\Delta^{-\varepsilon}(A_1, \dots, A_K) := \cup_{k=1}^{K} \partial^{-\varepsilon} A_k$ be the union of all the inner ε-boundaries. This is $\Delta^{-\varepsilon}(A_1, \dots, A_K)$ the set of points of $\cup_{k=1}^{K} A_k$ which are on the boundary between some two distinct A_k and $A_{k'}$. We want to find a lower bound in the measure $\gamma(\Delta^{-\varepsilon}(A_1, \dots, A_K))$.

Proposition 1. *Given $K \geq 4$ and $w_1, \dots, w_K \in (0, 1/4]$ such that $\sum_{k=1}^{K} w_k = 1$, we have the bound:*

$$\inf_{A_1, \dots, A_K} \gamma(\Delta^{-\varepsilon}(A_1, \dots, A_K)) \geq 1 - \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x}$$

where the infinimum is taken over all $(w_1, \dots, w_k)$-partitions of standard Gaussian space $(\mathbb{R}^d, \gamma)$, and $x := \Phi^{-1}\left(1 - \max_{k \in [\![M]\!]} w_k\right)$.

Proof. By (9), we have the formula

$$\begin{aligned} \gamma(\Delta^{-\varepsilon}(A_1, \dots, A_K)) &= \sum_{k=1}^{K} \gamma(\partial^{-\varepsilon} A_k) \qquad (10) \\ &= \sum_{k=1}^{K} \gamma(A_{-k}^{\varepsilon}) - \gamma(A_{-k}). \qquad (11) \end{aligned}$$

Let $w_{-k} := \gamma(A_{-k}) = 1 - w_k$, and assume $w_{-k} \geq 3/4$, i.e $w_k \leq 1/4$, for all $k \in [\![K]\!]$.

For example, this condition holds in the equitable scenario where $w_k = 1/K$ for all k.

Now, by standard *Gaussian Isoperimetric Inequality* (see (Boucheron et al., 2013) for example), one has

$$\begin{aligned} \gamma(A_{-k}^{\varepsilon}) &\geq \Phi(\Phi^{-1}(\gamma(A_{-k}) + \varepsilon) \\ &= \Phi(\Phi^{-1}(1 - w_k) + \varepsilon). \end{aligned} \tag{12}$$

Using the bound $\frac{x}{1+x^2}\varphi(x) < 1 - \Phi(x) < \frac{1}{x}\varphi(x)\ \forall x > 0$ where φ is the density of the standard Gaussian law. We can further find that

$$\begin{aligned} \Phi(\Phi^{-1}(1 - w_k) + \varepsilon) &\geq 1 - w_k \frac{1 + \Phi^{-1}(1 - w_k)^2}{\Phi^{-1}(1 - w_k)^2} \times \\ &\quad e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon \Phi^{-1}(1 - w_k)} \\ &\geq 1 - w_k \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x} > 0 \end{aligned} \tag{13}$$

(since the function $x \mapsto \dfrac{1+x^2}{x^2} e^{-\varepsilon x}$ is decreasing for $x > 0$)

where $x := \min_{k \in [\![K]\!]} \Phi^{-1}(1 - w_k) = \Phi^{-1}\left(1 - \max_{k \in [\![K]\!]} w_k\right) \geq \Phi^{-1}(3/4) > 0.67$. Combining (10), (12), and (13) yields the following

$$\begin{aligned} \gamma(\Delta^{-\varepsilon}(A_1, \dots, A_K)) &\geq \sum_{k=1}^{K} \big(1 - w_k \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x} \\ &\quad - (1 - w_k))\big) \\ &= \sum_{k=1}^{K} \left(1 - \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x}\right) w_k \\ &= 1 - \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x}, \end{aligned}$$

Asymptotic analysis In the limit, it is easy to check that in the case where $\max_{k \in [\![K]\!]} w_k \longrightarrow 0$, we have that $x \longrightarrow \infty$. In this setting, we thus have $\frac{1+x^2}{x^2} \longrightarrow 1$ and can now derive the following bound:

$$\inf_{A_1, \dots, A_K} \gamma(\Delta^{-\varepsilon}(A_1, \dots, A_K)) \xrightarrow{\max_{k \in [\![K]\!]} w_k \to 0} 1 - e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x}.$$

Equitable scenario In the equitable scenario where $w_k = 1/K$ for all k, we have

$$\inf_{A_1,\ldots,A_K} \gamma(\Delta^{-\varepsilon}(A_1,\ldots,A_K)) \geqslant 1 - \frac{1+x^2}{x^2} e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon x}$$

where $x = \Phi^{-1}(1-1/K)$. When $K \geq 8$ we have:

$$\Phi^{-1}(1-1/K) \geqslant \sqrt{2\log\left(\frac{K\,(q(K)^2-1)}{\sqrt{2\pi}q(K)^3}\right)} \tag{14}$$

where $q(K) = \sqrt{2\log(\sqrt{2\pi}K)}$.

Consequently, we have when $K \to \infty$, the following behavior:

$$\gamma(\Delta^{-\varepsilon}(A_1,\ldots,A_K)) \overset{K\to\infty}{\leqslant} 1 - e^{-\frac{1}{2}\varepsilon^2} e^{-\varepsilon\sqrt{2\log(K)}} \tag{15}$$

□

Proof of the inequality (14). Set $p := 1/K$. First, for any $x > 0$, we have the following upper:

$$\int_x^\infty e^{-y^2/2}dy = \int_x^\infty \frac{y}{y}e^{-y^2/2}dy \leq \frac{1}{x}\int_x^\infty ye^{-y^2/2}dy = \frac{e^{-x^2/2}}{x}.$$

For a lower bound:

$$\int_x^\infty e^{-y^2/2}dy = \int_x^\infty \frac{y}{y}e^{-y^2/2}dy = \frac{e^{-x^2/2}}{x} - \int_x^\infty \frac{1}{y^2}e^{-y^2/2}dy$$

and

$$\int_x^\infty \frac{1}{y^2}e^{-y^2/2}dy = \int_x^\infty \frac{y}{y^3}e^{-y^2/2}dy \leq \frac{e^{-x^2/2}}{x^3}$$

and combining these gives

$$\int_x^\infty e^{-y^2/2}dy \geq \left(\frac{1}{x} - \frac{1}{x^3}\right)e^{-x^2/2}.$$

Thus

$$\frac{1}{\sqrt{2\pi}}\left(\frac{1}{x} - \frac{1}{x^3}\right)e^{-x^2/2} \leq 1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2},$$

from where

$$\frac{1}{\sqrt{2\pi}}\left(\frac{1}{\Phi^{-1}(1-p)} - \frac{1}{\Phi^{-1}(1-p)^3}\right)e^{-\Phi^{-1}(1-p)^2/2} \tag{16}$$

$$\leq p \leq \frac{1}{\sqrt{2\pi}}\frac{1}{\Phi^{-1}(1-p)}e^{-\Phi^{-1}(1-p)^2/2} \tag{17}$$

Using (17), when $\Phi^{-1}(1-p) \geq 1$ (that is $p \leqslant 0.15$ or equivalently $K \geq 8$), we have the following upper bound $\Phi^{-1}(1-p) \leqslant q(p)$ where $q(p) := \sqrt{2\log(\sqrt{2\pi}/p)}$. Then, injecting $q(p)$ in (16):

$$\frac{1}{\sqrt{2\pi}}\left(\frac{1}{q(p)} - \frac{1}{q(p)^3}\right)e^{-\Phi^{-1}(1-p)^2/2} \leq p.$$

Now when $q(p) \geq 1$ you have:

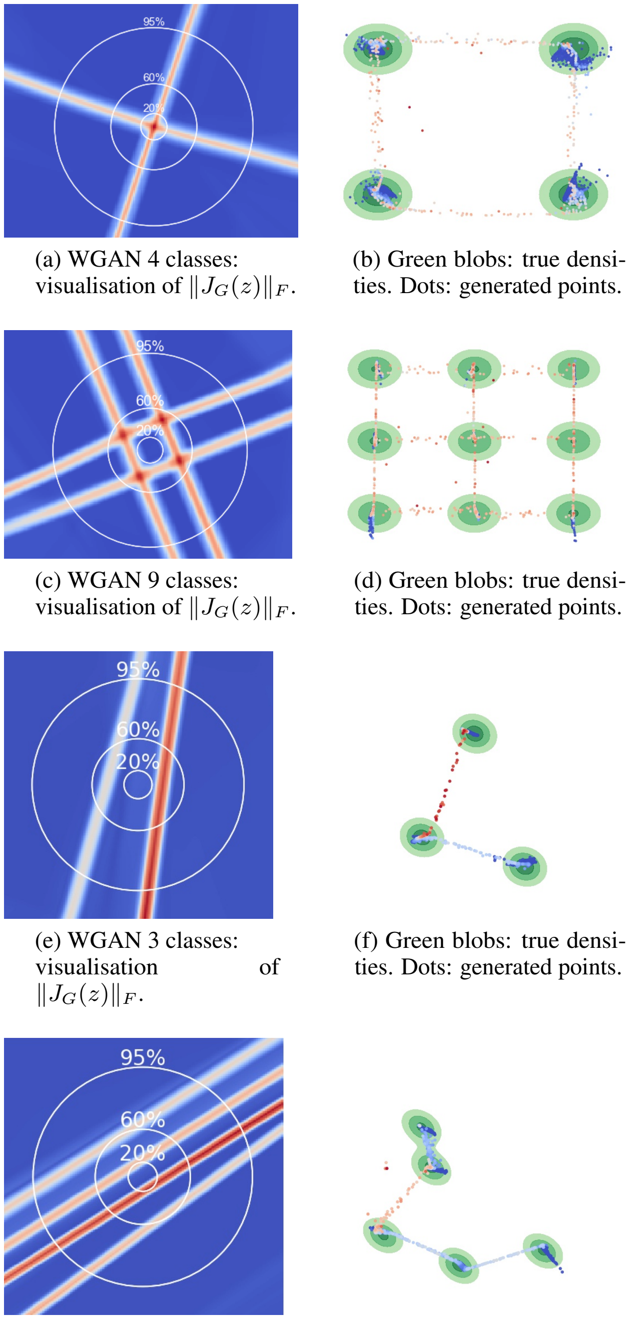
$$e^{-\Phi^{-1}(1-p)^2/2} \leq \frac{\sqrt{2\pi}pq(p)^3}{q(p)^2-1}$$

and

$$\Phi^{-1}(1-p) \geq \sqrt{2\log\left(\frac{q(p)^2-1}{\sqrt{2\pi}pq(p)^3}\right)}.$$

There is one additional requirement on p which is simply that the argument of the log should be ≥ 1 i.e. $q(p)^2 - 1 \geq \sqrt{2\pi}pq(p)^3$, which is true as soon as $K \geq 8$. □

F. Visualization of Theorem 3



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

(b) Green blobs: true densities. Dots: generated points.

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

(d) Green blobs: true densities. Dots: generated points.

(e) WGAN 3 classes: visualisation of $\|J_G(z)\|_F$.

(f) Green blobs: true densities. Dots: generated points.

(g) WGAN 5 classes: visualisation of $\|J_G(z)\|_F$.

(h) Green blobs: true densities. Dots: generated points.

Figure 6. Learning disconnected manifolds: visualization of the gradient of the generator (JFN) in the latent space and densities in the output space.

G. Definition of the different metrics used

In the sequel, we present the different metrics used in Section 4 of the paper to assess performances of GANs. We have:

- Improved Precision/Recall (PR) metric (Kynkäänniemi et al., 2019): it has been presented in Definition 2. Intuitively, Based on a k-NN estimation of the manifold of real (resp. generated) data, it assesses whether generated (resp. real) points belong in the real (resp. generated) data manifold or not. The proportion of generated (resp. real) points that are in the real (resp. generated) data manifold is the precision (resp. recall).

- the Hausdorff distance: it is defined by

$$\text{Haus}(A, B) = \max\left\{\max_{a\in A}\min_{b\in B}\|a-b\|, \max_{b\in B}\min_{a\in A}\|a-b\|\right\}$$

  Such a distance is useful to evaluate the closeness of two different supports from a metric space, but is sensitive to outliers because of the max operation. It has been recently used for theoretical purposes by Pandeva & Schubert (2019).

- the Frechet Inception distance: first proposed by Dowson & Landau (1982), the Frechet distance was applied in the setting of GANs by Heusel et al. (2017). This distance between mutlivariate Gaussians compares statistic of generated samples to real samples as follows

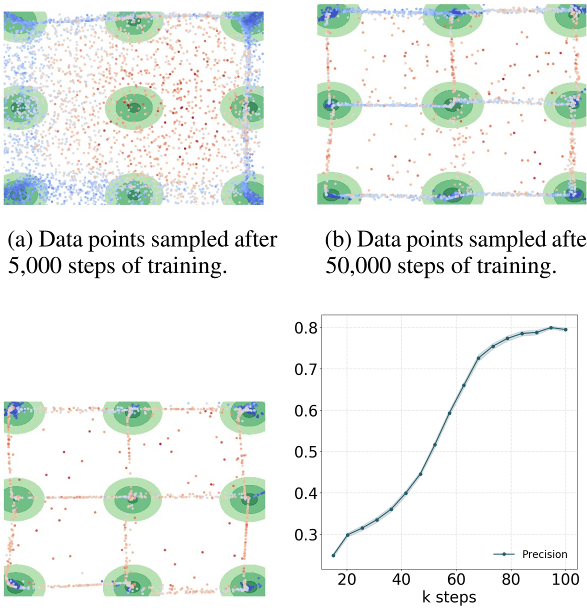
$$\text{FID} = \|\nu_\star - \nu_\theta\|^2 + Tr\big(\Sigma_\star + \Sigma_\theta + 2(\Sigma_\star\Sigma_\theta)^{\frac{1}{2}}\big)$$

  where $X_\star = \mathcal{N}(\nu_\star, \Sigma_\star)$ and $X_\theta = \mathcal{N}(\nu_\theta, \Sigma_\theta)$ are the activations of a pre-softmax layer. However, when dealing with disconnected manifolds, we argue that this distance is not well suited as it approximates the distributions with unimodal one, thus loosing many information.

The choice of such metrics is motivated by the fact that metrics measuring the performances of GANs should not rely on relative densities but should rather be point sets based metrics.

H. Saturation of a MLP neural network

In Section 4.2, we claim that the generator reduces the sampling of off-manifold data points up to a saturation point. Figure 7 below provides a visualization of this phenomenon. In this synthetic case, we learn a 9-component mixture of Gaussians using simple GANs architecture (both the generator and the discriminator are MLP with two hidden layers). The minimal distance between two modes is set to 9. We clearly see in Figure 7d that the precision saturates around 80%.



(a) Data points sampled after 5,000 steps of training.

(b) Data points sampled after 50,000 steps of training.

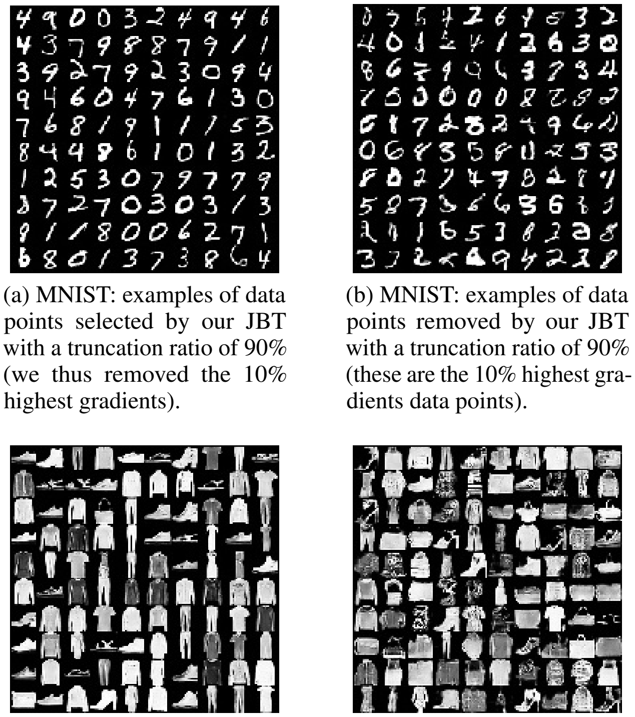
(c) Data points sampled after 100,000 steps of training.

(d) Evolution of the precision $\bar{\alpha}$ during training.

Figure 7. Learning 9 disconnected manifolds with a standard GANs architecture.

I. More results and visualizations on MNIST/F-MNIST/CIFAR10

Additionally to those in Section 4.3, we provide in Figure 9 and Table 2 supplementary results for MNIST, F-MNIST and CIFAR-10 datasets.



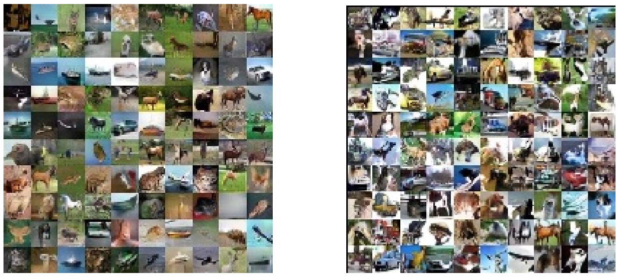
(a) MNIST: examples of data points selected by our JBT with a truncation ratio of 90% (we thus removed the 10% highest gradients).

(b) MNIST: examples of data points removed by our JBT with a truncation ratio of 90% (these are the 10% highest gradients data points).

(c) F-MNIST: examples of data points selected by our JBT with a truncation ratio of 90% (we thus removed the 10% highest gradients)..

(d) F-MNIST: examples of data points removed by our JBT with a truncation ratio of 90% (these are the 10% highest gradients data points).

Figure 8. Visualization of our truncation method on CIFAR10.



(a) CIFAR-10: examples of data points selected by our JBT with a truncation ratio of 90% (we thus removed the 10% highest gradients).

(b) MNIST: examples of data points removed by our JBT with a truncation ratio of 90% (these are the 10% highest gradients data points).

Figure 9. Visualization of our truncation method (JBT) on three real-world datasets: MNIST, F-MNIST and CIFAR-10.

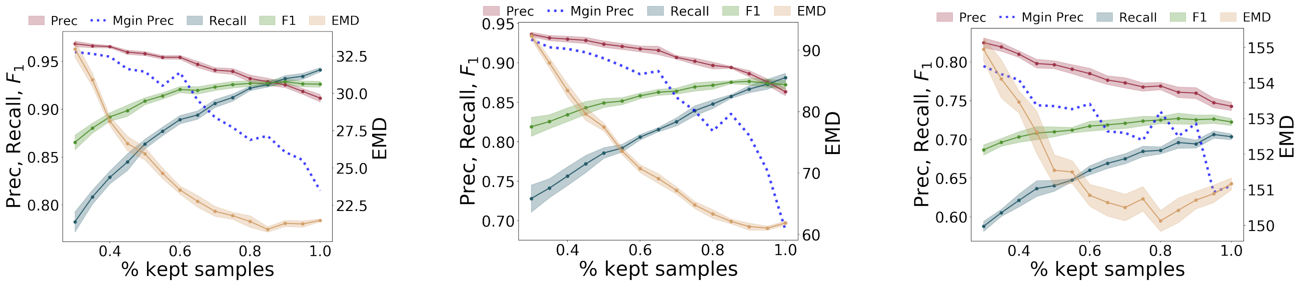


Figure 10. For high levels of kept samples, the marginal precision plummets of newly added samples, underlining the efficiency of our truncation method (JBT). Reported confidence intervals are 97% confidence intervals. On the second row, generated samples ordered by their JFN (left to right, top to bottom). In the last row, the data points generated are blurrier and outside the true manifold.

MNIST	Prec.	Rec.	F1	Haus.	FID	EMD
WGAN	91.2 ± 0.3	**93.7 ± 0.5**	**92.4 ± 0.4**	49.7 ± 0.2	24.3 ± 0.3	21.5 ± 0.1
WGAN 90% lowest JFN	92.5 ± 0.5	92.9 ± 0.3	**92.7 ± 0.4**	**48.1 ± 0.2**	26.9 ± 0.5	21.3 ± 0.2
WGAN 80% lowest JFN	**93.3 ± 0.3**	91.8 ± 0.4	**92.6 ± 0.4**	50.6 ± 0.4	33.1 ± 0.3	21.4 ± 0.4
W-Deligan	89.0 ± 0.6	**93.6 ± 0.3**	91.2 ± 0.5	50.7 ± 0.3	31.7 ± 0.5	22.4 ± 0.1
DMLGAN	**93.4 ± 0.2**	92.3 ± 0.2	**92.8 ± 0.2**	**48.2 ± 0.3**	**16.8 ± 0.4**	**20.7 ± 0.1**
Fashion-MNIST						
WGAN	86.3 ± 0.4	**88.2 ± 0.2**	**87.2 ± 0.3**	140.6 ± 0.7	259.7 ± 3.5	61.9 ± 0.3
WGAN 90% lowest JFN	88.6 ± 0.6	86.6 ± 0.5	**87.6 ± 0.5**	**138.7 ± 0.9**	**257.4 ± 3.0**	**61.3 ± 0.6**
WGAN 80% lowest JFN	**89.8 ± 0.4**	84.9 ± 0.5	**87.3 ± 0.4**	146.3 ± 1.1	396.2 ± 6.4	63.3 ± 0.7
W-Deligan	88.5 ± 0.3	85.3 ± 0.6	86.9 ± 0.4	141.7 ± 1.1	310.9 ± 3.1	**60.9 ± 0.4**
DMLGAN	87.4 ± 0.3	**88.1 ± 0.4**	**87.7 ± 0.4**	141.9 ± 1.2	**253.0 ± 2.8**	**60.9 ± 0.4**
CIFAR10						
WGAN	74.3 ± 0.5	**70.3 ± 0.4**	**72.3 ± 0.5**	334.7 ± 3.5	**634.8 ± 4.6**	151.2 ± 0.2
WGAN 90% lowest JFN	**76.0 ± 0.7**	69.4 ± 0.5	**72.5 ± 0.6**	**318.1 ± 3.7**	**631.3 ± 4.5**	150.7 ± 0.2
WGAN 80% lowest JFN	**76.9 ± 0.5**	68.6 ± 0.5	**72.5 ± 0.5**	**323.5 ± 4.0**	725.0 ± 3.5	**150.1 ± 0.3**
W-Deligan	71.5 ± 0.7	**69.8 ± 0.7**	70.6 ± 0.7	328.7 ± 2.1	727.8 ± 3.9	154.0 ± 0.3
DMLGAN	74.1 ± 0.5	65.7 ± 0.6	69.7 ± 0.6	328.6 ± 2.7	967.2 ± 4.1	152.0 ± 0.4

Table 2. Scores on MNIST and Fashion-MNIST. JFN stands for Jacobian Frobenius norm. ± is 97% confidence interval.

J. More results on BigGAN and ImageNet

In Figure 11, we show images from the Bubble class of ImageNet. It supports our claim of manifold disconectedness, even within a class, and outlines the importance of studying the learning of disconnected manifolds in generative models. Then, in Figure 12, we give more exemples from BigGAN 128x128 class-conditionned generator. We plot in the same format than in 4.4. Specifically, for different classes, we plot 128 images ranked by JFN. Here again, we see a concentration of off-manifold samples on the last row, proving the efficiency of our method. Example of classes responding particularly well to our ranking are House Finch (c), Monnarch Butterfly (i) or Wood rabbit (m). For each class, we also show an histogram of JFN based on 1024 samples. It shows that the JFN is a good indicator of the complexity of the class. For example, classes such as Cornet (q) or Football helmet (s) are very diverse and disconnected, resulting in high JFNs.

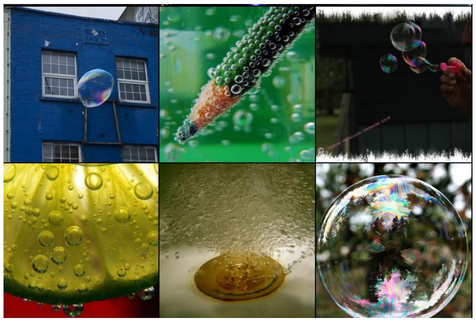
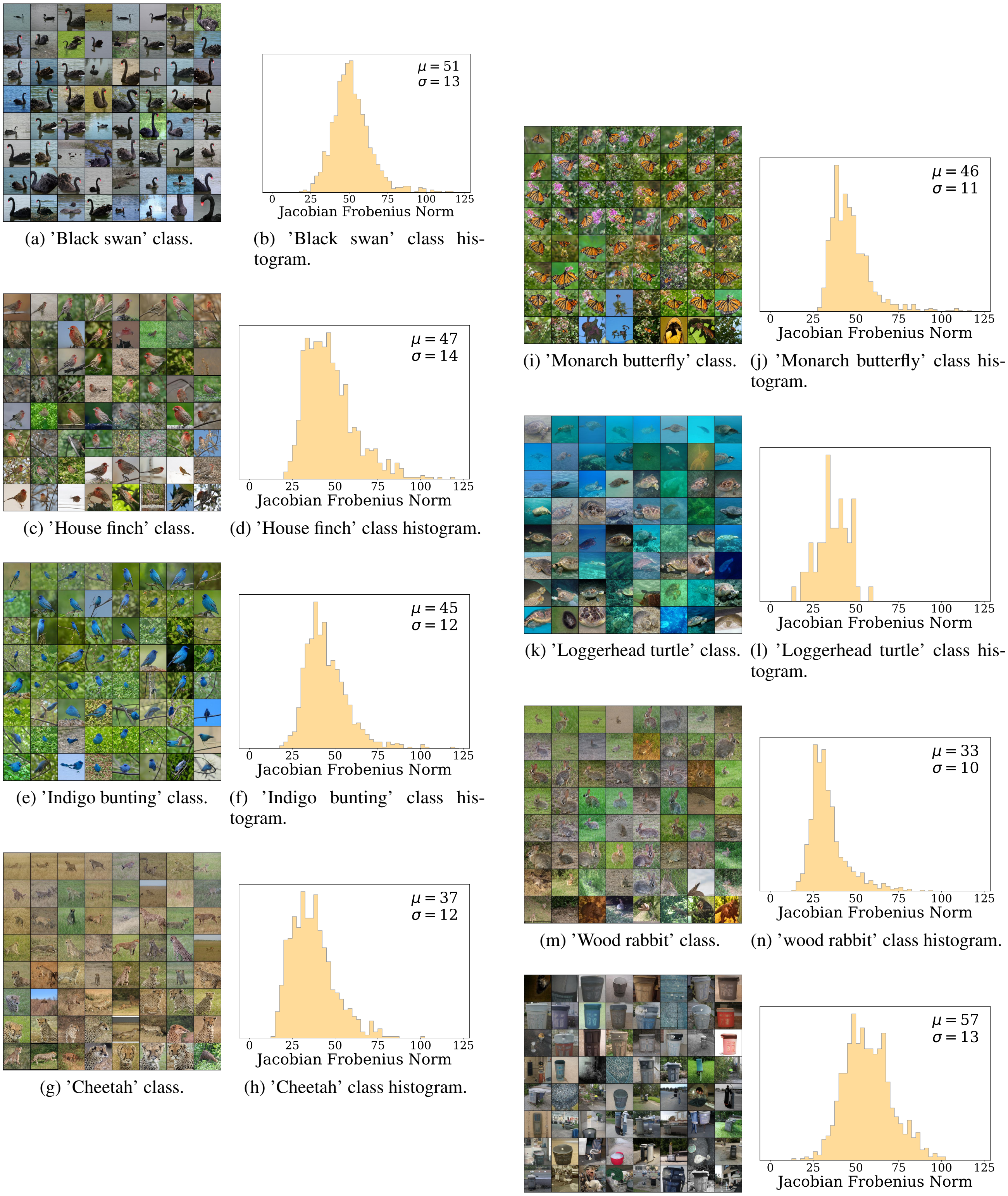


Figure 11. Images from the Bubble class of ImageNet showing that the class is complex and slightly multimodal.



(a) 'Black swan' class.

(b) 'Black swan' class histogram.

(c) 'House finch' class.

(d) 'House finch' class histogram.

(e) 'Indigo bunting' class.

(f) 'Indigo bunting' class histogram.

(g) 'Cheetah' class.

(h) 'Cheetah' class histogram.

(i) 'Monarch butterfly' class.

(j) 'Monarch butterfly' class histogram.

(k) 'Loggerhead turtle' class.

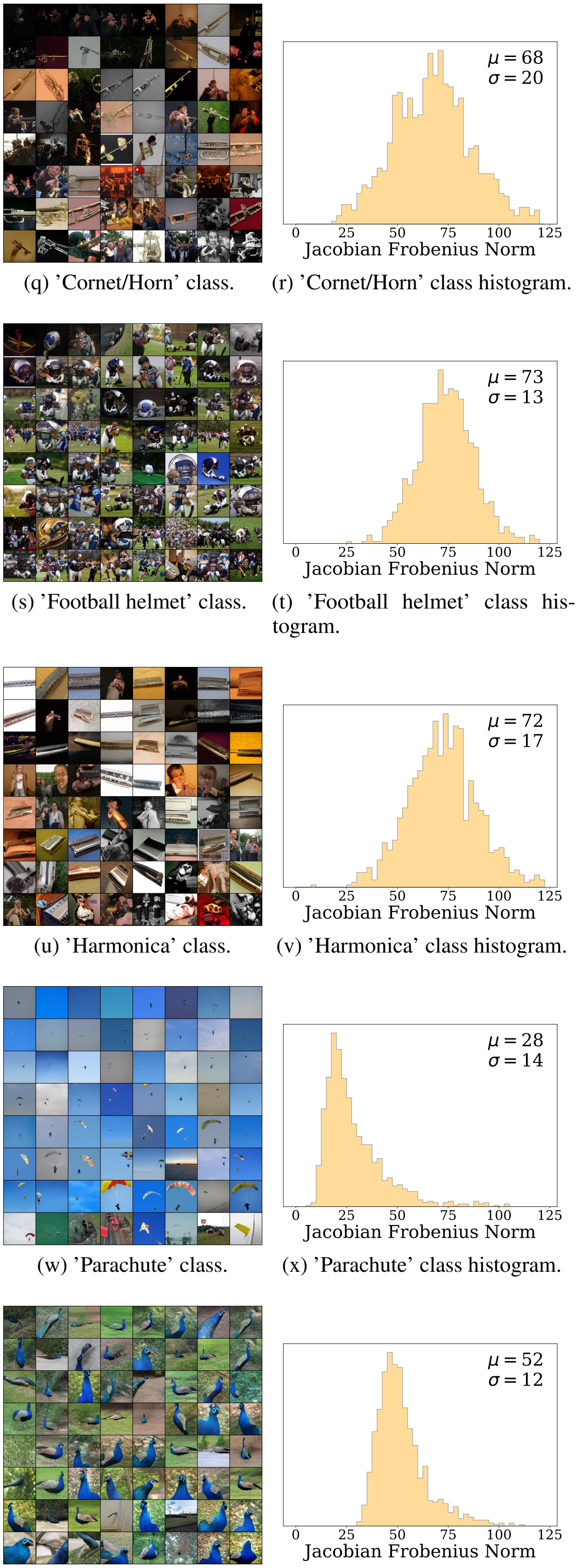
(l) 'Loggerhead turtle' class histogram.

(m) 'Wood rabbit' class.

(n) 'wood rabbit' class histogram.

(o) 'Trash can' class.

(p) 'Trash can' class histogram.



(q) 'Cornet/Horn' class.

(r) 'Cornet/Horn' class histogram.

(s) 'Football helmet' class.

(t) 'Football helmet' class histogram.

(u) 'Harmonica' class.

(v) 'Harmonica' class histogram.

(w) 'Parachute' class.

(x) 'Parachute' class histogram.

(y) 'Peacock' class.

(z) 'Peacock' class histogram.

Figure 12. For several classes with BigGAN model.

K. Network Architecture and Hyperparameters

Table 3. Models for Synthetic datasets

Operation	Feature Maps	Activation
G(z): $z \sim \mathcal{N}(0, 1)$	2	
Fully Connected - layer1	20	ReLU
Fully Connected - layer2	20	ReLU
D(x)		
Fully Connected - layer1	20	ReLU
Fully Connected - layer2	20	ReLU
Batch size	32	
Leaky ReLU slope	0.2	
Gradient Penalty weight	10	
Learning Rate	0.0002	
Optimizer	Adam: $\beta_1 = 0.5$	$\beta_2 = 0.5$

Table 4. WGAN for MNIST/Fashion MNIST

Operation	Kernel	Strides	Feature Maps	Activation
G(z): $z \sim \mathrm{N}(0, Id)$			100	
Fully Connected			$7 \times 7 \times 128$	
Convolution	3×3	1×1	$7 \times 7 \times 64$	LReLU
Convolution	3×3	1×1	$7 \times 7 \times 64$	LReLU
Nearest Up Sample			$14 \times 14 \times 64$	
Convolution	3×3	1×1	$14 \times 14 \times 32$	LReLU
Convolution	3×3	1×1	$14 \times 14 \times 32$	LReLU
Nearest Up Sample			$14 \times 14 \times 64$	
Convolution	3×3	1×1	$28 \times 28 \times 16$	LReLU
Convolution	5×5	1×1	$28 \times 28 \times 1$	Tanh
D(x)			$28 \times 28 \times 1$	
Convolution	4×4	2×2	$14 \times 14 \times 32$	LReLU
Convolution	3×3	1×1	$14 \times 14 \times 32$	LReLU
Convolution	4×4	2×2	$7 \times 7 \times 64$	LReLU
Convolution	3×3	1×1	$7 \times 7 \times 64$	LReLU
Fully Connected			1	-
Batch size	256			
Leaky ReLU slope	0.2			
Gradient Penalty weight	10			
Learning Rate	0.0002			
Optimizer	Adam	$\beta_1 : 0.5$	$\beta_2 : 0.5$	

For DeliGan, we use the same architecture and simply add 50 Gaussians for the reparametrization trick. For DMLGAN, we re-use the architecture of the authors.

Table 5. DMLGAN for MNIST/Fashion MNIST

Operation	Kernel	Strides	Feature Maps	BN	Activation
G(z): $z \sim \mathrm{N}(0, Id)$			100		
Fully Connected			$7 \times 7 \times 128$	-	
Convolution	3×3	1×1	$7 \times 7 \times 64$	-	Leaky ReLU
Convolution	3×3	1×1	$7 \times 7 \times 64$	-	Leaky ReLU
Nearest Up Sample			$14 \times 14 \times 64$	-	
Convolution	3×3	1×1	$14 \times 14 \times 32$	-	Leaky ReLU
Convolution	3×3	1×1	$14 \times 14 \times 32$	-	Leaky ReLU
Nearest Up Sample			$14 \times 14 \times 64$	-	
Convolution	3×3	1×1	$28 \times 28 \times 16$	-	Leaky ReLU
Convolution	5×5	1×1	$28 \times 28 \times 1$	-	Tanh
Encoder Q(x), Discriminator D(x)			$28 \times 28 \times 1$		
Convolution	4×4	2×2	$14 \times 14 \times 32$	-	Leaky ReLU
Convolution	3×3	1×1	$14 \times 14 \times 32$	-	Leaky ReLU
Convolution	4×4	2×2	$7 \times 7 \times 64$	-	Leaky ReLU
Convolution	3×3	1×1	$7 \times 7 \times 64$	-	Leaky ReLU
D Fully Connected			1	-	-
Q Convolution	3×3		$7 \times 7 \times 64$	Y	Leaky ReLU
Q Convolution	3×3		$7 \times 7 \times 64$	Y	Leaky ReLU
Q Fully Connected			$n_g = 10$	-	Softmax
Batch size	256				
Leaky ReLU slope	0.2				
Gradient Penalty weight	10				
Learning Rate	0.0002				
Optimizer	Adam	$\beta_1 = 0.5$	$\beta_2 = 0.5$		

Table 6. WGAN for CIFAR10, from (Gulrajani et al., 2017)

Operation	Kernel	Strides	Feature Maps	BN	Activation
G(z): $z \sim \mathrm{N}(0, Id)$			128		
Fully Connected			$4 \times 4 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$4 \times 4 \times 128$	Y	ReLU
Nearest Up Sample			$8 \times 8 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$8 \times 8 \times 128$	Y	ReLU
Nearest Up Sample			$16 \times 16 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$16 \times 16 \times 128$	Y	ReLU
Nearest Up Sample			$32 \times 32 \times 128$	-	
Convolution	3×3	1×1	$32 \times 32 \times 3$	-	Tanh
Discriminator D(x)			$32 \times 32 \times 3$		
ResBlock	$[3 \times 3] \times 2$	1×1	$32 \times 32 \times 128$	-	ReLU
AvgPool	2×2	1×1	$16 \times 16 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$16 \times 16 \times 128$	-	ReLU
AvgPool	2×2	1×1	$8 \times 8 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$8 \times 8 \times 128$	-	ReLU
ResBlock	$[3 \times 3] \times 2$	1×1	$8 \times 8 \times 128$	-	ReLU
Mean pooling (spatial-wise)	-	-	128	-	
Fully Connected			1	-	-
Batch size	64				
Gradient Penalty weight	10				
Learning Rate	0.0002				
Optimizer	Adam	$\beta_1 = 0.$	$\beta_2 = 0.9$		
Discriminator steps	5				

Table 7. DMLGAN for CIFAR10, from (Gulrajani et al., 2017)

Operation	Kernel	Strides	Feature Maps	BN	Activation
G(z): $z \sim \mathrm{N}(0, Id)$			128		
Fully Connected			$4 \times 4 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$4 \times 4 \times 128$	Y	ReLU
Nearest Up Sample			$8 \times 8 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$8 \times 8 \times 128$	Y	ReLU
Nearest Up Sample			$16 \times 16 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$16 \times 16 \times 128$	Y	ReLU
Nearest Up Sample			$32 \times 32 \times 128$	-	
Convolution	3×3	1×1	$32 \times 32 \times 3$	-	Tanh
Encoder Q(x), Discriminator D(x)			$32 \times 32 \times 3$		
ResBlock	$[3 \times 3] \times 2$	1×1	$32 \times 32 \times 128$	-	ReLU
AvgPool	2×2	1×1	$16 \times 16 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$16 \times 16 \times 128$	-	ReLU
AvgPool	2×2	1×1	$8 \times 8 \times 128$	-	
ResBlock	$[3 \times 3] \times 2$	1×1	$8 \times 8 \times 128$	-	ReLU
D ResBlock	$[3 \times 3] \times 2$	1×1	$8 \times 8 \times 128$	-	ReLU
D Mean pooling (spatial-wise)	2×2	1×1	128	-	
D Fully Connected			1	-	-
Q ResBlock	$[3 \times 3] \times 2$	1×1	$8 \times 8 \times 128$	-	ReLU
Q Mean pooling (spatial-wise)	2×2	1×1	128	-	
Q Fully Connected			$n_g = 10$	-	Softmax
Batch size	64				
Gradient Penalty weight	10				
Learning Rate	0.0002				
Optimizer	Adam	$\beta_1 = 0.$	$\beta_2 = 0.9$		
Discriminator steps	5				