TopP&R: Robust Support Estimation Approach for Evaluating Fidelity and Diversity in Generative Models

Anonymous Authors[1]

Abstract

We propose a robust and reliable evaluation metric for generative models by introducing topological and statistical treatments for rigorous support estimation. Existing metrics, such as Inception Score (IS), Fréchet Inception Distance (FID), and the variants of `Precision` and `Recall` (`P&R`), heavily rely on supports that are estimated from sample features. However, the reliability of their estimation has not been seriously discussed (and overlooked) even though the quality of the evaluation entirely depends on it. In this paper, we propose Topological Precision and Recall (`TopP&R`, pronounced "topper"), which provides a systematic approach to estimating supports, retaining only topologically and statistically important features with a certain level of confidence. This not only makes `TopP&R` strong for noisy features, but also provides statistical consistency. Our theoretical and experimental results show that `TopP&R` is robust to outliers and non-independent and identically distributed (Non-IID) perturbations, while accurately capturing the true trend of change in samples. To the best of our knowledge, this is the first evaluation metric focused on the robust estimation of the support and provides its statistical consistency under noise.

1. Introduction

In keeping with the remarkable improvements of deep generative models (Karras et al., 2019; 2020; 2021; Brock et al., 2018; Ho et al., 2020; Kingma & Welling, 2013; Sauer et al., 2022; 2021; Kang & Park, 2020), evaluation metrics that can well measure the performance of generative models have also been continuously developed (Salimans et al., 2016; Heusel et al., 2017; Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020). For instance, Inception Score (IS) (Salimans et al., 2016) measures the Kullback-Leibler divergence between the real and fake sample distributions. Fréchet Inception Score (FID) (Heusel et al., 2017) calculates the distance between the real and fake supports using the estimated mean and variance under the multi-Gaussian assumption. The original Precision and Recall (Sajjadi et al., 2018) and its variants (Kynkäänniemi et al., 2019; Naeem et al., 2020) measure fidelity and diversity separately, where fidelity is about how closely the generated samples resemble real samples, while diversity is about whether a generative model can generate samples that are as diverse as real samples.

Considering the eminent progress of deep generative models based on these existing metrics, some may question why we need another evaluation study. In this paper, we argue that we need more reliable evaluation metrics now precisely, because deep generative models have reached sufficient maturity and evaluation metrics are saturated (Table 8 in (Kang et al., 2022)). Even more, it has been recently reported that even the most widely used evaluation metric, FID, sometimes doesn't match with the expected perceptual quality, fidelity, and diversity, which means the metrics are not always working properly (Kynkäänniemi et al., 2022).

In addition, existing metrics are vulnerable to the existence of noise because all of them rely on the assumption that real data is clean. However, in practice, real data often contain lots of artifacts, such as mislabeled data and adversarial examples (Pleiss et al., 2020; Li et al., 2022), which can cause the overestimation of the data distribution in the evaluation pipeline. This error seriously perturbs the scores, leading to a false impression of improvement when developing generative models. See Appendix E.2 for possible scenarios. Thus, to provide more accurate and comprehensive ideas for improvements and to illuminate a new direction in the generative field, we need a more robust and reliable evaluation metric.

An ideal evaluation metric must capture the real signal of the data, while being robust to noise. Note that there is an inherent tension in developing metrics that meets these goals. On one hand, the metric should be sensitive enough so that

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

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

it can capture real signals lurking in data. On the other hand, it must ignore noises that hide the signal. However, sensitive metrics are inevitably susceptible to noise to some extent. To address this, one needs a systematic way to answer the following three questions: 1) What is signal and what is noise? 2) How do we draw a line between them? 3) How certain are we on the result?

One solution can be to use the idea of topological data analysis (TDA) (Carlsson, 2009) and statistical inference. TDA is a recent and emerging field of data science that relies on topological tools to infer relevant features for possibly complex data. A key object in TDA is persistent homology, which observes how long each topological feature would survive over varying resolutions and provides a measure to quantify its significance; *i.e.*, if some features persist longer than others over varying resolutions, we consider them as topological signal and vice versa as noise.

In this paper, we combine these ideas to estimate supports more robustly and overcome various issues of previous metrics such as unboundedness, inconsistency, etc. Our main contributions are as follows: we establish 1) a systematic approach to estimate supports via Kernel Density Estimator (KDE) derived under topological conditions; 2) a new metric that is robust to outliers while reliably detecting the change of distributions on various scenarios; 3) a consistency guarantee with robustness under very weak assumptions that is suitable for high dimensional data; 4) a combination of a noise framework and a statistical inference in TDA. Our code is available at `TopP&R`-ICML.

2. Background

To lay the foundation for our method and theoretical analysis, we first explain the previous evaluation method `Precision` and `Recall` (`P&R`). Then, we introduce the main idea of persistent homology and its confidence estimation techniques that bring the benefit of using topological and statistical tools for addressing uncertainty in samples. For the sake of space constraint and to streamline the discussion of our main idea, we only provide a brief overview of the concepts that are relevant to this work and refer the reader to Appendix A or (Edelsbrunner & Harer, 2010; Chazal & Michel, 2021; Wasserman, 2018; Hatcher, 2002) for further details on TDA.

Notation. For any x and $r > 0$, we use the notation $\mathcal{B}_d(x, r) = \{y : d(y, x) < r\}$ be the open ball in distance d of radius r. We also write $\mathcal{B}(x, r)$ when d is understood from context. For a distribution P on $\mathbb{R}^d$, we let $\mathrm{supp}(P) := \{x \in \mathbb{R}^d : P(\mathcal{B}(x, r)) > 0 \text{ for all } r > 0\}$ be the support of P. Throughout the paper, we refer to $\mathrm{supp(P)}$ as support of P, or simply support, or manifold, but we don't necessarily require the (geometrical) manifold structure on $\mathrm{supp(P)}$. For a kernel function $K : \mathbb{R}^d \to \mathbb{R}$, a dataset $\mathcal{X} = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ and bandwidth $h > 0$, we let the kernel density estimator (KDE) as $\hat{p}_h(x) := \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$, and we let the average KDE as $p_h := \mathbb{E}[\hat{p}_h]$. We denote by P, Q the probability distributions in $\mathbb{R}^d$ of real data and generated samples, respectively. And we use $\mathcal{X} = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{Y_1, \ldots, Y_m\} \subset \mathbb{R}^d$ for real data and generated samples possibly with noise, respectively.

Precision and Recall. There exist two aspects of generative samples' quality; fidelity and diversity. Fidelity refers to the degree to which the generated samples resemble the real ones. Diversity, on the other hand, measures whether the generated samples cover the full variability of the real samples. Sajjadi et al. (2018) was the first to propose assessing these two aspects separately via `Precision` and `Recall` (`P&R`). In the ideal case where we have full access to the probability distributions P and Q, $\mathrm{precision}_P(Q) := Q(\mathrm{supp}(P))$, $\mathrm{recall}_Q(P) := P(\mathrm{supp}(Q))$, which correspond to the max `Precision` and max `Recall` in (Sajjadi et al., 2018), respectively. In practice, however, because P and Q are usually unknown, `P&R` is computed based on the estimated supports of the real and fake distributions, and several variants have followed up proposing different ways of estimating the supports (Kynkäänniemi et al., 2019; Naeem et al., 2020).

Persistent homology and diagram. Persistent homology (PH) is a tool in computational topology that measures the topological invariants (homological features) of data that persist across multiple scales, and is represented in the persistence diagram. Formally, let a *filtration* be a collection of subspaces $\mathcal{F} = \{\mathcal{F}_\delta \subset \mathbb{R}^d\}_{\delta \in \mathbb{R}}$ with $\delta_1 \leq \delta_2$ implying $\mathcal{F}_{\delta_1} \subset \mathcal{F}_{\delta_2}$. Typically, filtration is defined through a function f related to data. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we consider its sublevel filtration $\{f^{-1}(-\infty, \delta]\}_{\delta \in \mathbb{R}}$ or a superlevel filtration $\{f^{-1}[\delta, \infty)\}_{\delta \in \mathbb{R}}$. **For a filtration $\mathcal{F}$ and for each nonnegative k, we track when k-dimensional homological features** (e.g., 0-dimension: connected component, 1-dimension: loop, 2-dimension: cavity, ...) appear and disappear in the filtration. As increasing or decreasing δ in the filtration $\{\mathcal{F}_\delta\}$, if a homological feature appears at $\mathcal{F}_b$ and disappears at $\mathcal{F}_d$, then we say that it is born at b and dies at d. By considering these pairs $\{(b, d)\}$ as points in the plane $(\mathbb{R} \cup \{\pm\infty\})^2$, we obtain a *persistence diagram*. From this, a homological feature with a longer life length, $d - b$, can be treated as a significant feature in the data set, and a homological feature with a shorter life length as a topological noise, which lies near the diagonal line $\{(\delta, \delta) : \delta \in \mathbb{R}\}$(Figure 1 (b)).

Confidence band estimation. Statistical inference has recently been developed for TDA (Chazal et al., 2013; 2015; Fasy et al., 2014). TDA consists of features reflecting

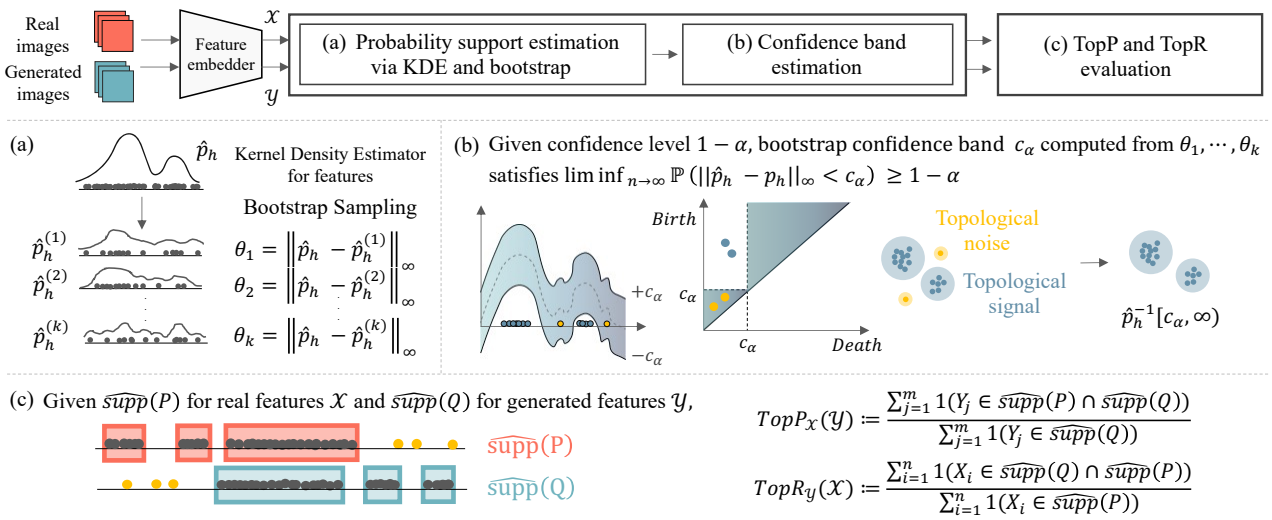


Figure 1: **Illustration of the proposed evaluation pipeline.** The proposed metric `TopP&R` is defined in three steps: (a) Confidence band estimation in section 2, (b) Robust support estimation, and (c) Evaluation via `TopP&R` in section 3.

topological characteristics of data, and it is of question to systematically distinguish features that are indeed from geometrical structures and features that are insignificant or due to noise. To *statistically* separate topologically significant features from topological noise, we employ confidence band estimation. Given the significance level α, let confidence band c_α be the bootstrap bandwidth of $\|\hat{p}_h - p_h\|_\infty$, computed as in Algorithm 1. Then it satisfies $\liminf_{n\to\infty} \mathbb{P}\left(\|\hat{p}_h - p_h\|_\infty < c_\alpha\right) \geq 1-\alpha$, as in Proposition C.2 in Appendix C. This confidence band can simultaneously determine significant topological features while filtering out noise features. We use $c_\mathcal{X}$ and $c_\mathcal{Y}$ to denote the confidence band defined under significance level α according to the datasets $\mathcal{X}$ and $\mathcal{Y}$. In later sections, we use these tools to provide a more rigorous way of scoring samples based on the confidence level we set.

3. Robust support estimation for reliable evaluation

Current evaluation metrics for generative models typically rely on strong regularity conditions. For example, they assume samples are well-curated without outliers or adversarial perturbation, real or generative models have bounded densities, etc. However, practical scenarios are wild: both real and generated samples can be corrupted with noise from various sources, and the real data can be very sparsely distributed without density. In this work, we consider more general and practical situations, wherein both real and generated samples can have noises that come from sampling procedure, remained uncertainty due to data or model, etc. See Appendix E.2 for more on our motivations.

3.1. Overview of our metric

We design our metric with the philosophy of evaluating performance very conservatively. Our metric is based on only topologically significant data structures, also with statistical confidence above a certain level. Toward this, we apply KDE as function f to define a filtration, which allows us to approximate the support with data through $\{f^{-1}[\delta,\infty)\}_{\delta\in\mathbb{R}}$. Since the significance of data comprising the support is determined by the life length of homological features, we calculate c_α that enables us to systematically separate short/long lifetimes of homological features. We then estimate the supports with topologically significant data structure via superlevel set $f^{-1}[c_\alpha,\infty)$ and finally, we evaluate fidelity and diversity with the estimated supports. We have collectively named this process `TopP&R`. By its nature, `TopP&R` is bounded and yields consistent performance under various conditions such as noisy data points, outliers, and even with long-tailed data distribution. To compensate for the limitation of the curse of dimensionality that KDE has in high dimensions, we use a method of dimensionality reduction towards a random projection of embedded features onto 32 dimensions prior to calculating `TopP&R`. The random projection is characterized by preserving the information about distances and homological features defined in the higher dimensional spaces by Johnson Linenstrauss Lemma (Johnson et al., 1986).

3.2. Topological precision and recall

To facilitate our discussion, we rewrite the precision as $\text{precision}_P(\mathcal{Y}) = Q\left(\text{supp}(P)\cap\text{supp}(Q)\right)/Q\left(\text{supp}(Q)\right)$

and define the precision of data points as

$$\text{precision}_P(\mathcal{Y}) := \frac{\sum_{j=1}^m 1\left(Y_j \in \text{supp}(P) \cap \text{supp}(Q)\right)}{\sum_{j=1}^m 1\left(Y_j \in \text{supp}(Q)\right)},$$

which is just replacing the distribution Q by the empirical distribution $\frac{1}{m}\sum_{j=1}^m \delta_{Y_j}$ of Y in the precision. Similarly,

$$\text{recall}_Q(\mathcal{X}) := \frac{\sum_{i=1}^n 1\left(X_i \in \text{supp}(Q) \cap \text{supp}(P)\right)}{\sum_{i=1}^n 1\left(X_i \in \text{supp}(P)\right)}.$$

In practice, $\text{supp}(P)$ and $\text{supp}(Q)$ are not known a priori and need to be estimated, and these estimates should be robust to noise since we allow it now. For this, we use the KDE $\hat{p}_{h_n}(x) := \frac{1}{nh_n^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)$ of $\mathcal{X}$ and the bootstrap bandwidth $c_{\mathcal{X}}$ of $\|\hat{p}_{h_n} - p_{h_n}\|_\infty$, where $h_n > 0$ and a significance level $\alpha \in (0,1)$ (Section 2). Then, we estimate the support of P by the superlevel set at $c_{\mathcal{X}}$[1] as $\widehat{\text{supp}}(P) = \hat{p}_{h_n}^{-1}[c_{\mathcal{X}}, \infty)$, which allows to filter out noise whose KDE values are likely to be small. Similarly, the support of Q is estimated: $\widehat{\text{supp}}(Q) = \hat{q}_{h_m}^{-1}[c_{\mathcal{Y}}, \infty)$, where $\hat{q}_{h_m}(x) := \frac{1}{mh_m^d}\sum_{j=1}^m K\left(\frac{x-Y_j}{h_m}\right)$ is the KDE of $\mathcal{Y}$ and $c_{\mathcal{Y}}$ is the bootstrap bandwidth of $\|\hat{q}_{h_m} - q_{h_m}\|_\infty$

For the robust estimates of the precision, we apply the support estimates to $\text{precision}_P(\mathcal{Y})$ and $\text{recall}_Q(\mathcal{X})$ and define the topological precision and recall (TopP&R) as

$$\begin{aligned}\texttt{TopP}_{\mathcal{X}}(\mathcal{Y}) &:= \frac{\sum_{j=1}^m 1\left(Y_j \in \widehat{\text{supp}}(P) \cap \widehat{\text{supp}}(Q)\right)}{\sum_{j=1}^m 1\left(Y_j \in \widehat{\text{supp}}(Q)\right)} \\ &= \frac{\sum_{j=1}^m 1\left(\hat{p}_{h_n}(Y_j) > c_{\mathcal{X}},\ \hat{q}_{h_m}(Y_j) > c_{\mathcal{Y}}\right)}{\sum_{j=1}^m 1\left(\hat{q}_{h_m}(Y_j) > c_{\mathcal{Y}}\right)}, \\ \texttt{TopR}_{\mathcal{Y}}(\mathcal{X}) &:= \frac{\sum_{i=1}^n 1\left(\hat{q}_{h_m}(X_i) > c_{\mathcal{Y}},\ \hat{p}_{h_n}(X_i) > c_{\mathcal{X}}\right)}{\sum_{i=1}^n 1\left(\hat{p}_{h_n}(X_i) > c_{\mathcal{X}}\right)}.\end{aligned}$$

The kernel bandwidths h_n and h_m are hyperparameters, and we provide guidelines to select the optimal bandwidths h_n and h_m in practice (See Appendix F.4).

3.3. Bandwidth estimation using bootstrapping

Using the bootstrap bandwidth $c_{\mathcal{X}}$ as the threshold is the key part of our estimator (TopP&R) for robustly estimating $\text{supp}(P)$. As we have seen in Section 2, the bootstrap bandwidth $c_{\mathcal{X}}$ filters out the topological noise in topological data analysis. Analogously, using $c_{\mathcal{X}}$ allows to robustly estimate $\text{supp}(P)$. When X_i is an outlier, its KDE value $\hat{p}_h(X_i)$ is likely to be small as well as the values at the connected component generated by X_i. So those components from outliers are likely to be removed in the estimated support $\hat{p}_h^{-1}[c_{\mathcal{X}}, \infty)$. Higher dimensional homological noises are also removed. Hence, the estimated support denoises topological noise and robustly estimates $\text{supp}(P)$. See Appendix B for a more detailed explanation.

Now that we are only left with topological features of high confidence, this allows us to draw analogies to confidence intervals in statistical analysis, where the uncertainty of the samples is treated by setting the level of confidence. In the next section, we show that TopP&R not only gives a more reliable evaluation score for generated samples but also has good theoretical properties.

4. Consistency with robustness of TopP&R

The key properties of TopP&R is consistency with robustness. The consistency ensures that, the precision and the recall we compute from the *data* approaches the precision and the recall from the *distribution* as we have more samples. The consistency allows to investigate the precision and recall of full distributions only with access to finite sampled data. TopP&R achieves consistency with robustness, that is, the consistency holds with the data possibly corrupted by noise. This is due to the robust estimation of supports with KDE with confidence bands.

We demonstrate the statistical model for both data and noise. Let P, Q, $\mathcal{X}$, $\mathcal{Y}$ be as in Section 2, and let $\mathcal{X}^0, \mathcal{Y}^0$ be real data and generated data without noise. $\mathcal{X}, \mathcal{Y}, \mathcal{X}^0, \mathcal{Y}^0$ are understood as multisets, *i.e.*, elements can be repeated. We first assume that the uncorrupted data are IID.

Assumption 1. *The data $\mathcal{X}^0 = \{X_1^0, \ldots, X_n^0\}$ and $\mathcal{Y}^0 = \{Y_1^0, \ldots, Y_m^0\}$ are IID from P and Q, respectively.*

In practice, the data is often corrupted with noise. We consider the adversarial noise, where some fraction of data are replaced with arbitrary point cloud data.

Assumption 2. *Let $\{\rho_k\}_{k\in\mathbb{N}}$ be a sequence of nonnegative real numbers. Then the observed data $\mathcal{X}$ and $\mathcal{Y}$ satisfies $\left|\mathcal{X}\backslash\mathcal{X}^0\right| = n\rho_n$ and $\left|\mathcal{Y}\backslash\mathcal{Y}^0\right| = m\rho_m$.*

In the adversarial model, we control the level of noise by the fraction ρ, but do not assume other conditions such as IID or boundedness, to make our noise model very general and challenging.

For distributions and kernel functions, we assume weak condition, detailed in Assumption A1 and A2 in Appendix C. Under the data and the noise models, TopP&R achieves consistency with robustness. That is, the estimated precision and recall is asymptotically correct with high probability even if up to a portion of $1/\sqrt{n}$ or $1/\sqrt{m}$ are replaced by adversarial noise. This is due to the robust estimation of the support with the kernel density estimator with the confidence band of the persistent homology.

Proposition 4.1. *Suppose Assumption 1, 2, A1, A2 hold. Suppose $\alpha \to 0$, $h_n \to 0$, $nh_n \to \infty$, $nh_n^{-d}\rho_n^2 \to 0$, and*

[1] The computation of c_α and its practical interpretation is described in Algorithm 1.

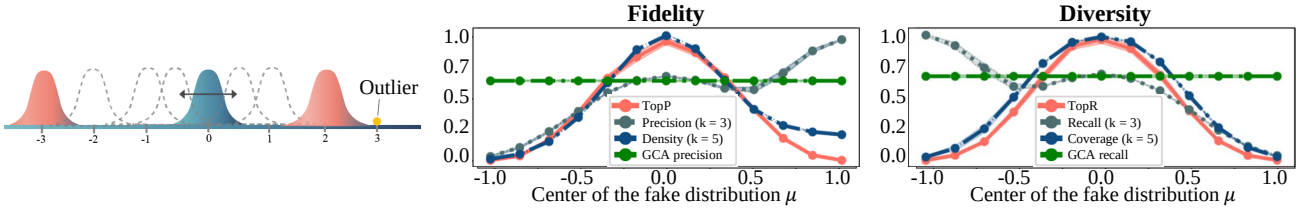
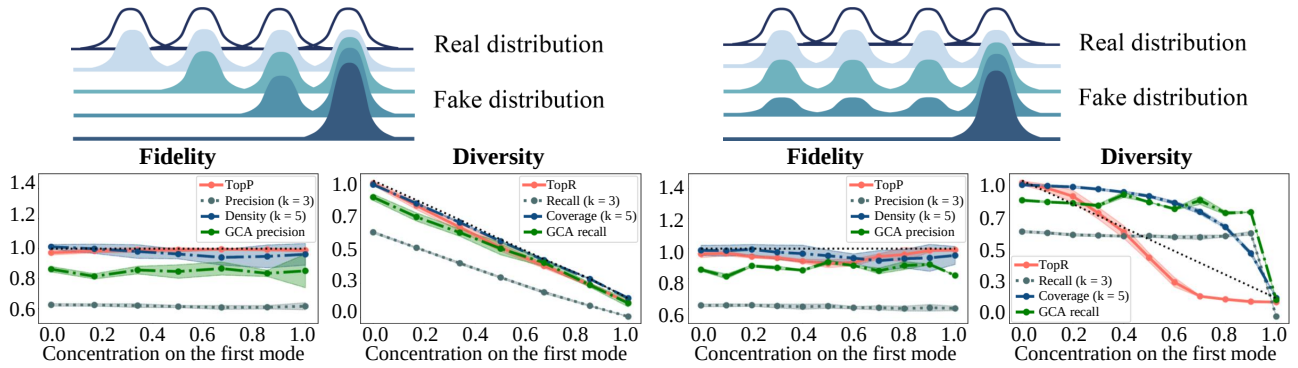


Figure 2: Behaviors of evaluation metrics for outliers on real and fake distribution. The horizontal axis corresponds to the value of μ.



(a) Sequential mode dropping **(b) Simultaneous mode dropping**

Figure 3: Behaviors of evaluation metrics for (a) sequential and (b) simultaneous mode-drop scenarios. The horizontal axis shows the concentration ratio on the distribution centered at $\mu = 0$.

similar relations hold for h_m, ρ_m. Then,

$$|\mathrm{TopP}_{\mathcal{X}}(\mathcal{Y}) - \mathrm{precision}_P(\mathcal{Y})| = O_{\mathbb{P}}\left(Q(B_{n,m}) + \rho_m\right),$$
$$\left|\mathrm{TopR}_{\mathcal{Y}}(\mathcal{X}) - \mathrm{recall}_Q(\mathcal{X})\right| = O_{\mathbb{P}}\left(P(A_{n,m}) + \rho_n\right),$$

for fixed sequence of sets $\{A_{n,m}\}_{n,m\in\mathbb{N}}, \{B_{n,m}\}_{n,m\in\mathbb{N}}$ with $P(A_{n,m}) \to 0$ and $Q(B_{n,m}) \to 0$ as $n, m \to \infty$.

Theorem 4.2. *Under the same condition as in Proposition 4.1,*

$$|\mathrm{TopP}_{\mathcal{X}}(\mathcal{Y}) - \mathrm{precision}_P(Q)| = O_{\mathbb{P}}\left(Q(B_{n,m}) + \rho_m\right),$$
$$\left|\mathrm{TopR}_{\mathcal{Y}}(\mathcal{X}) - \mathrm{recall}_Q(P)\right| = O_{\mathbb{P}}\left(P(A_{n,m}) + \rho_n\right).$$

Since $P(A_{n,m}) \to 0$ and $Q(B_{n,m}) \to 0$, this implies consistencies of `TopP&R`. In fact, additionally under minor probabilistic and geometrical assumptions, $P(A_{n,m})$ and $Q(B_{n,m})$ are of order $h_m + h_n$.

Lemma 4.3. *Under the same condition as in Proposition 4.1 and additionally under Assumption A3, A4, $P(A_{n,m}) = O(h_n + h_m)$ and $Q(B_{n,m}) = O(h_n + h_m)$.*

Remark 4.4. Consistency guarantees from Proposition 4.1 and Theorem 4.2 are in principle due to the uniform convergence of KDE over varying bandwidth h_n (Proposition C.2). Once we replace estimating the support with KDE by k-NN or something else, we wouldn't have consistency guarantees. Hence, using the KDE is an essential part for the theoretical guarantees of `TopP&R`.

Our theoretical results in Proposition 4.1 and Theorem 4.2 are novel and important in several perspectives. These results are among the first theoretical guarantees for evaluation metrics for generative models as far as we are aware of. Also, as in Remark C.1, assumptions are very weak and suitable for high dimensional data. Also, robustness to adversarial noise is provably guaranteed.

5. Experiments

A good evaluation metric must correctly capture the changes of the underlying data distribution. To examine the performance of evaluation metrics, we carefully select a set of experiments for sanity checks. With toy and real image data, we check 1) how well the metric captures the true trend of underlying data distributions and 2) how well the metric resists perturbations applied to samples.

5.1. Comparison with existing metrics

In the experiments, we compare `TopP&R` with existing methods to verify the specific characteristic of our metric. For the comparison we consider Improved `Precision` and `Recall` (`P&R`) (Kynkäänniemi et al., 2019),

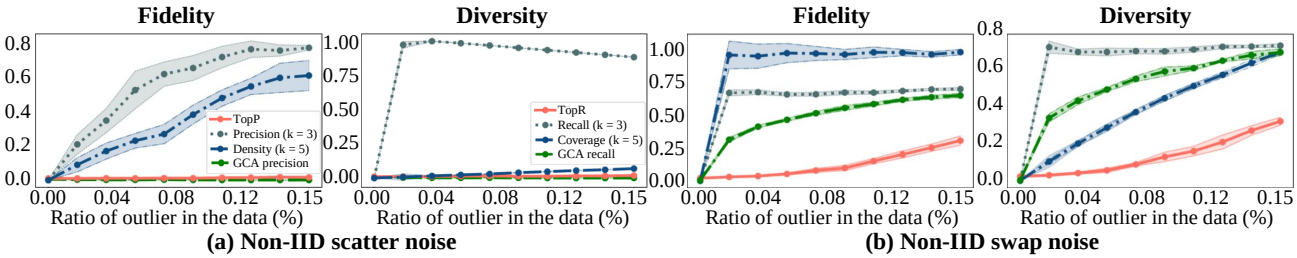


Figure 4: Behaviors of evaluation metrics on Non-IID perturbations. We replace a certain percentage of real and fake data (a) with random uniform noise and (b) by switching some data between real and fake.

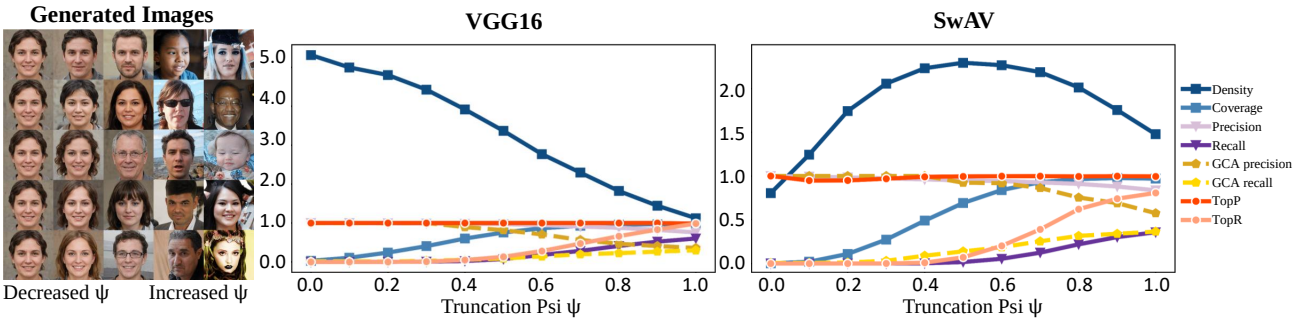


Figure 5: Behaviors of metrics with truncation trick. The horizontal axis corresponds to the value of ψ denoting the increased diversity. The images are generated via StyleGAN2 with FFHQ dataset (Karras et al., 2019).

`Density` and `Coverage` (`D&C`) (Naeem et al., 2020), Geometric Component Analysis (GCA) (Poklukar et al., 2021), Manifold Topology Divergence (MTD) (Barannikov et al., 2021). GCA is a method that forms a connected component for each P and Q and defines a `precision` and `recall` using the number of vertices and edges. Since this method has the property of evaluating the topological similarity of P and Q, there is a part that has a different perspective than existing `P&R` variants. MTD is a method of measuring the distance between two distributions by simultaneously filtering both the P and Q distributions and using the sum of lifetimes of homologies. The shaded area of the figures denotes the ± 1 standard deviation for ten trials. For all the experiments, linear random projection to 32 dimensions is additionally used for `TopP&R`. For more experimental details, please refer to Appendix F.1.

5.2. Sanity checks with toy data

Following (Naeem et al., 2020), we first examine how well the metric reflects the trend of $\mathcal{Y}$ moving away from $\mathcal{X}$ and whether it is suitable for finding mode-drop phenomena. In addition to these, we newly design several experiments that can highlight `TopP&R`'s favorable theoretical properties of consistency with robustness in various scenarios.

5.2.1. Shifting the Generated Feature Manifold

We generate samples from $\mathcal{X} \sim \mathcal{N}(\mathbf{0}, I)$ and $\mathcal{Y} \sim \mathcal{N}(\mu\mathbf{1}, I)$ in $\mathbb{R}^{64}$, where $\mathbf{1}$ is a vector of ones and I is an identity matrix. We examine how each metric responds to shifting $\mathcal{Y}$ with $\mu \in [-1, 1]$ while there are outliers at $\mathbf{3} \in \mathbb{R}^{64}$ for both $\mathcal{X}$ and $\mathcal{Y}$ (Figure 2). From the result, GCA does not work well in situations where the topological properties of data don't change. Also, we find that both improved `P&R` and `D&C` behave pathologically when there are outliers. Since these methods are based on the k-nearest neighbor algorithm, they inevitably overestimate the underlying support when there are outliers. For example, when $\mu < 0.5$, `Recall` returns a high-diversity score, even though the true supports of $\mathcal{X}$ and $\mathcal{Y}$ are actually far apart. In addition, `P&R` does not reach 1 in high dimensions even when $\mathcal{X} = \mathcal{Y}$. `D&C` (Naeem et al., 2020) shows better results than `P&R` because it always uses $\mathcal{X}$ (the real data distribution) as a reference point, which usually has fewer outliers than $\mathcal{Y}$ (the fake data distribution). However, there is no guarantee that this will be the case in practice (Pleiss et al., 2020; Li et al., 2022). When there is an outlier in $\mathcal{X}$, `D&C` also returns an incorrect high-fidelity score at $\mu > 0.5$. On the other hand, `TopP&R` shows a stable trend unaffected by outliers, demonstrating its robustness.

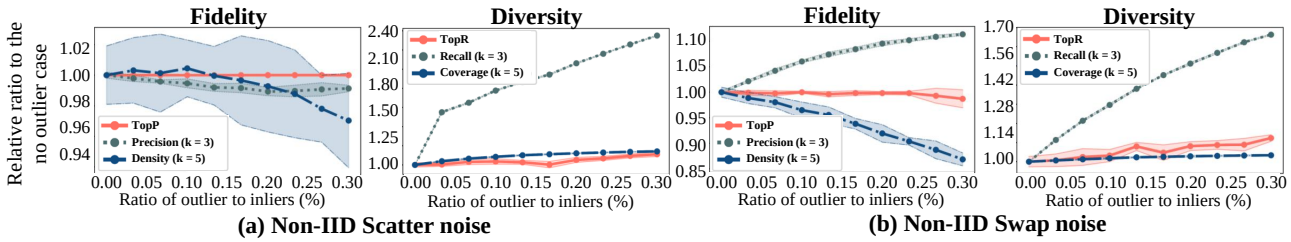


Figure 6: Comparison of evaluation metrics on Non-IID perturbations using FFHQ dataset. We replaced certain ratio of $\mathcal{X}$ and $\mathcal{Y}$ (a) with outliers and (b) by exchanging features.

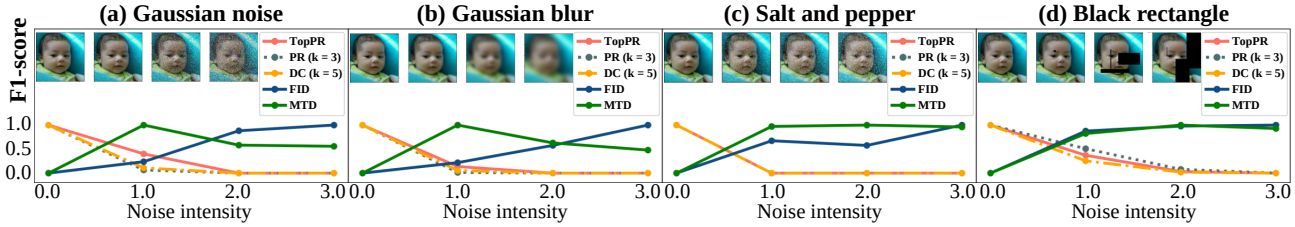


Figure 7: Verification of whether TopP&R can make an accurate quantitative assessment of noisy image features. Gaussian noise, Gaussian blur, salt and pepper, and black rectangle noise are added to the FFHQ images and embedded with T4096.

5.2.2. Dropping modes

For this experiment, we consider the mixture of Gaussians with seven modes in $\mathbb{R}^{64}$. We simulate mode-drop phenomena by gradually dropping all but one mode from the fake distribution $\mathcal{Y}$ that is initially identical to $\mathcal{X}$ (Figure 3). As in the illustration of mode-drop experiment, when the number of samples in a particular mode decreases, we kept the number of samples in $\mathcal{X}$ constant so that the same amount of decreased samples are supplemented to the first mode which leads fidelity to be fixed to 1. From the result, We observe that the values of `Precision` fail to saturate, *i.e.*, mainly smaller than 1, and the `Density` fluctuates to a value greater than 1 indicating their instability and unboundedness. In terms of diversity, `Recall` and GCA Recall does not respond to the simultaneous mode drop, nor does the improved metric `Coverage` show a fast decay as the reference line. Compared to these methods, `TopP` performs well, being held at the upperbound of 1 in sequential mode drop, and `TopR` also decreases closest to the reference line in simultaneous mode drop.

5.2.3. Tolerance to Non-IID perturbations

Robustness to perturbations is another important aspect we should consider when designing a metric. Here, we test whether `TopP&R` behaves stably under two variants of noise cases; 1) **scatter noise**: replacing X_i and Y_j with uniformly distributed noise and 2) **swap noise**: swapping the position between X_i and Y_j. These two cases all correspond to the adversarial noise model of Assumption 2. We set $\mathcal{X} \sim \mathcal{N}(\mu = 0, I) \in \mathbb{R}^{64}$ and $\mathcal{Y} \sim \mathcal{N}(\mu = 1, I) \in \mathbb{R}^{64}$ where $\mu = 1$, and thus an ideal evaluation metric must return zero for both fidelity and diversity. In the result, while the GCA precision is relatively robust to the scatter noise, GCA recall tends to be sensitive to the swap noise. In both cases, we find that `P&R` and `D&C` are more sensitive while `TopP&R` remains relatively stable until the noise ratio reaches 15% of the total data, which is a clear example of the weakness of existing metrics to perturbation (Figure 4).

5.3. Sanity check with Real data

Now that we have verified the metrics on toy data, we move on to real data. Just like in the toy experiments, we concentrate on how the metrics behave in extreme situations, such as outliers, mode-drop phenomena, perceptual distortions, etc. We also test different image embedders, including pretrained VGG16 (Simonyan & Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), and SwAV (Morozov et al., 2020).

5.3.1. Resolving fidelity and diversity

To test whether `TopP&R` responds appropriately to the change in the underlying distributions in real scenarios, we test the metric on the generated images of StyleGAN2 (Karras et al., 2020) using the truncation trick (Karras et al., 2019). As shown in Figure 5, every time the distribution is transformed by ψ, `TopP&R` responds well and shows consistent behavior across different embedders with bounded scores in $[0, 1]$, which are important virtues

Table 1: Generative models ranked by FID and F1-scores based on TopP&R, D&C, and P&R, respectively. The $\mathcal{X}$ and $\mathcal{Y}$ are embedded with InceptionV3, VGG16, and SwAV. The number inside the parenthesis denotes the rank based on each metric.

	Model	StyleGAN2	ReACGAN	BigGAN	PDGAN	ACGAN	WGAN-GP
InceptionV3	**FID** (↓)	3.78 (1)	3.87 (2)	4.16 (3)	31.54 (4)	33.39 (5)	107.68 (6)
	TopP&R (↑)	0.9769 (1)	0.8457 (2)	0.7751 (3)	0.7339 (4)	0.6951 (5)	0.0163 (6)
	D&C (↑)	0.9626 (2)	0.9409 (3)	1.1562 (1)	0.4383 (4)	0.3883 (5)	0.1913 (6)
	P&R (↑)	0.6232 (1)	0.3320 (2)	0.3278 (3)	0.1801 (4)	0.0986 (5)	0.0604 (6)
	MTD (↓)	2.3380 (3)	2.2687 (2)	1.4473 (1)	7.0188 (4)	8.0728 (5)	11.498 (6)
VGG16	**TopP&R** (↑)	0.9754 (1)	0.5727 (3)	0.7556 (2)	0.4021 (4)	0.3463 (5)	0.0011 (6)
	D&C (↑)	0.9831 (3)	1.0484 (1)	0.9701 (4)	0.9872 (2)	0.8971 (5)	0.6372 (6)
	P&R (↑)	0.6861 (1)	0.1915 (3)	0.3526 (2)	0.0379 (4)	0.0195 (5)	0.0001 (6)
	MTD (↓)	25.757 (4)	25.826 (3)	34.755 (5)	24.586 (2)	23.318 (1)	41.346 (6)
SwAV	**TopP&R** (↑)	0.9093 (1)	0.3568 (3)	0.5578 (2)	0.1592 (4)	0.1065 (5)	0.0003 (6)
	D&C (↑)	1.0732 (1)	0.9492 (3)	1.0419 (2)	0.6328 (4)	0.4565 (5)	0.0721 (6)
	P&R (↑)	0.5623 (1)	0.0901 (3)	0.1459 (2)	0.0025 (4)	0.0000 (6)	0.0002 (5)
	MTD (↓)	1.1098 (1)	1.5512 (3)	1.3280 (2)	1.8302 (4)	2.2982 (5)	4.9378 (6)

as an evaluation metric. On the other hand, Density gives unbounded scores (fidelity > 1) and shows inconsistent trend depending on the embedder. Because Density is not capped in value, it is difficult to interpret the score and know exactly which value denotes the best performance (*e.g.*, in our case, the best performance is when TopP&R $= 1$). Since TopP&R pays more attention to the consistent behavior of a model by examining what the model primarily generates, rather than relying on the entire sample, which contains results by chance, the fact that TopP is kept at 1.0 means that StyleGAN2 produces high-quality images most of the time. Thus, this behavior ("TopP remains constant") does not mean that TopP is inferior to regular precision for checking the trade-off between fidelity and diversity, but rather reveals its property focusing on different perspectives than the others.

5.3.2. Dropping modes in CIFAR-10

We conduct an additional mode-drop experiment to verify TopP&R's actual sensitiveness on the real data set CIFAR-10 (Krizhevsky et al., 2009). The performance of each metric (Figure A2) is measured with the identical data while simultaneously dropping the modes of nine classes of CIFAR-10. Since the number of the images dropped in each step is identical, the trend of ground truth diversity should linearly decrease. Here, P&R metric captures the simultaneous mode drop better than D&C because this time random drop of the modes has reduced the area of the estimated fake manifold. On the other hand, TopP&R best captures the true trend of decreasing diversity on average, consistent with the toy result in Figure 3. In addition, we perform the experiments on a dataset with long-tailed distribution and find that TopP&R captures the trend well even when there are minority sets (Appendix G.2). This again shows the reliability of TopP&R.

5.3.3. Robustness to perturbations

To demonstrate the robustness of our metric against the adversarial noise model of Assumption 2, we test both scatter-noise and swap noise scenarios with real data. In the experiment, following Kynkäänniemi et al. (2019), we first classify inliers and outliers that are generated by StyleGAN (Karras et al., 2019). For scatter noise we add the outliers to the inliers and for swap noise we swap the real FFHQ images with generated images. Under these specific noise conditions, Precision shows similar or even better robustness than Density (Figure 6). On the other hand, Coverage is more robust than Recall. In both cases, TopP&R shows the best performance, resistant to noise.

5.3.4. Sensitiveness to the noise intensity

One of the advantages of FID (Heusel et al., 2017) is that it is good at estimating the degrees of distortion applied to the images. Similarly, we check whether the F1-score based

on TopP&R provides a reasonable evaluation according to different noise levels. As illustrated in Figure 7, $\mathcal{X}$ and $\mathcal{Y}$ are sets of reference FFHQ features and noisy FFHQ features, respectively. The experimental results show that TopP&R actually reflects well the different degrees of distortion added to the images while a similar topology-based method MTD shows inconsistent behavior to the distortions.

5.3.5. Ranking between generative models

One of the major caveats with two-score metrics is that it is difficult to rank between different models; *e.g.*, which model is better? High fidelity with low diversity? Or low fidelity with high diversity? In the case of traditional precision and recall, this problem could be solved by using F1-score, which is the harmonic mean of fidelity and diversity. However, unlike the traditional ones, F1-scores based on P&R or D&C do not provide a reliable or stable score due to their inherent instability and unboundedness. Thanks to its stability and robustness to various perturbations, we find that the TopP&R-based F1-score offers consistent ranking with FID under various embedding networks (Table 1)[2]. To quantitatively compare between the similarity of rankings across varying embedders by different metrics, we have computed Hamming Distance (HD) (Appendix F.5) where lower HD indicates more similarity. TopP&R, P&R, D&C, and MTD have HDs of 1.33, 2.66, 3.0, and 3.33, respectively. From this, TopP&R provides the most consistent ranking across varying embedders (consistent to Section 5.3.1).

6. Conclusions

Many works have been proposed recently to assess the fidelity and diversity of generative models. However, none of them has focused on the accurate estimation of support even though it is one of the key components in the entire evaluation pipeline. In this paper, we proposed topological precision and recall (TopP&R) that provides a systematical fix by robustly estimating the support with both topological and statistical treatments. Our theoretical and experimental results showed that TopP&R serves as a robust and reliable evaluation metric under various embeddings and noise conditions, including mode drop, outliers, and Non-IID perturbations.

References

Barannikov, S., Trofimov, I., Sotnikov, G., Trimbach, E., Korotin, A., Filippov, A., and Burnaev, E. Manifold topology divergence: a framework for comparing data manifolds. *Advances in Neural Information Processing Systems*, 34, 2021.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Carlsson, G. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009. ISSN 0273-0979. doi: 10.1090/S0273-0979-09-01249-X. URL https://doi.org/10.1090/S0273-0979-09-01249-X.

Chazal, F. and Michel, B. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers Artif. Intell.*, 4:667963, 2021. doi: 10.3389/frai.2021.667963. URL https://doi.org/10.3389/frai.2021.667963.

Chazal, F., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. On the bootstrap for persistence diagrams and landscapes. *Modelirovanie i Analiz Informacionnyh Sistem*, 20, 11 2013. doi: 10.18255/1818-1015-2013-6-111-120.

Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. Stochastic convergence of persistence landscapes and silhouettes. *J. Comput. Geom.*, 6(2):140–161, 2015.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., Rinaldo, A., and Wasserman, L. Robust topological inference: Distance to a measure and kernel distance. *The Journal of Machine Learning Research*, 18(1):5845–5884, 2017.

Edelsbrunner, H. and Harer, J. L. *Computational topology*. American Mathematical Society, Providence, RI, 2010. ISBN 978-0-8218-4925-5. doi: 10.1090/mbk/069. URL https://doi.org/10.1090/mbk/069. An introduction.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1252. URL https://doi.org/10.1214/14-AOS1252.

Federer, H. Curvature measures. *Transactions of the American Mathematical Society*, 93:418–491, 1959. ISSN 0002-9947. doi: 10.2307/1993504. URL https://doi.org/10.2307/1993504.

[2]All GAN models used in the experiment follow the settings in StudioGAN PyTorch-StudioGAN is an open-source library under the MIT license (MIT), which are under the NVIDIA source code license.

Hamming, R. W. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

Hatcher, A. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-79160-X; 0-521-79540-0.

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. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Johnson, W. B., Lindenstrauss, J., and Schechtman, G. Extensions of lipschitz maps into banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, 1986.

Kang, M. and Park, J. Contragan: Contrastive learning for conditional image generation. *Advances in Neural Information Processing Systems*, 33:21357–21369, 2020.

Kang, M., Shin, J., and Park, J. Studiogan: A taxonomy and benchmark of gans for image synthesis. *arXiv preprint arXiv:2206.09479*, 2022.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Kim, J., Shin, J., Rinaldo, A., and Wasserman, L. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. 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. 3398–3407. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kim19e.html.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kosorok, M. R. *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York, 2008. ISBN 978-0-387-74977-8. doi: 10.1007/978-0-387-74978-5. URL https://doi.org/10.1007/978-0-387-74978-5.

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. *Advances in Neural Information Processing Systems*, 32, 2019.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of imagenet classes in fréchet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

Li, Z., Wu, R., and Gan, T. Study on image data cleaning method of early esophageal cancer based on vgg_nin neural network. *Scientific Reports*, 12(1):1–10, 2022.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 eighth ieee international conference on data mining*, pp. 413–422. IEEE, 2008.

Morozov, S., Voynov, A., and Babenko, A. On self-supervised image representations for gan evaluation. In *International Conference on Learning Representations*, 2020.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

Neumann, M. H. Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Ann. Statist.*, 26(5):2014–2048, 1998. ISSN 0090-5364. doi: 10.1214/aos/1024691367. URL https://doi.org/10.1214/aos/1024691367.

Pleiss, G., Zhang, T., Elenberg, E., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

Poklukar, P., Varava, A., and Kragic, D. Geomca: Geometric evaluation of data representations. In *International Conference on Machine Learning*, pp. 8588–8598. PMLR, 2021.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 31, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Sauer, A., Chitta, K., Müller, J., and Geiger, A. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34, 2021.

Sauer, A., Schwarz, K., and Geiger, A. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv preprint arXiv:2202.00273*, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Terrell, G. R. and Scott, D. W. Variable kernel density estimation. *The Annals of Statistics*, pp. 1236–1265, 1992.

Thäle, C. 50 years sets with positive reach—a survey. *Surv. Math. Appl.*, 3:123–165, 2008. ISSN 1843-7265. doi: 10.1007/s11590-008-0097-2. URL https://doi.org/10.1007/s11590-008-0097-2.

van der Vaart, A. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN 9780521784504. URL https://books.google.fr/books?id=UEuQEM5RjWgC.

Wagner, H., Chen, C., and Vuçini, E. Efficient computation of persistent homology for cubical data. In *Topological methods in data analysis and visualization II*, pp. 91–106. Springer, 2012.

Wasserman, L. Topological data analysis. *Annu. Rev. Stat. Appl.*, 5:501–535, 2018. ISSN 2326-8298. doi: 10.1146/annurev-statistics-031017-100045. URL https://doi.org/10.1146/annurev-statistics-031017-100045.

Appendix

A. More Background on Topological Data Analysis

Topological data analysis (TDA) (Carlsson, 2009) is a recent and emerging field of data science that relies on topological tools to infer relevant features for possibly complex data. A key object in TDA is persistent homology, which quantifies salient topological features of data by observing them in multi-resolutions.

A.1. Persistent Homology

Persistent homnology. *Persistent homology* is a multiscale approach to represent the topological features. For a filtration $\mathcal{F}$ and for each $k \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, the associated k-th persistent homology $PH_k\mathcal{F}$ is a collection of k-th dimensional homologies $\{H_k\mathcal{F}_\delta\}_{\delta\in\mathbb{R}}$ equipped with homomorphisms $\{\imath_k^{a,b} : H_k\mathcal{F}_a \to H_k\mathcal{F}_b\}_{a\le b}$ induced by the inclusion $\mathcal{F}_a \subset \mathcal{F}_b$.

Persistence diagram. For the k-th persistent homology $PH_k\mathcal{F}$, the set of filtration levels at which a specific homology appears is always an interval $[b, d) \subset [-\infty, \infty]$. The corresponding k-th persistence diagram is a multiset of points $(\mathbb{R} \cup \{\infty\})^2$, consisting of all pairs (b, d) where $[b, d)$ is the interval of filtration values for which a specific homology appears in $PH_k\mathcal{F}$.

A.2. Statistical Inference of Persistent Homology

As discussed above, a homological feature with a long life-length is important information in topology while the homology with a short life-length can be treated as non-significant information or noise. The confidence band estimator provides the confidence set from the features that only includes topologically and statistically significant (statistically considered as elements in the population set) under a certain level of confidence. And to build a confidence set, we first need to endow a metric on the space of persistence diagrams.

Bottleneck distance. The most fundamental metric to measure the distance between two persistence diagrams is the *bottleneck distance*.

Definition A.1. The *bottleneck* distance between two persistence diagrams D_1 and D_2 is defined by

$$d_B(D_1, D_2) = \inf_{\gamma\in\Gamma} \sup_{p\in D_1} \|p - \gamma(p)\|_\infty,$$

where the set Γ consists of all the bijections $\gamma : D_1 \cup Diag \to D_2 \cup Diag$, and $Diag$ is the diagonal $\{(x, x) : x \in \mathbb{R}\} \subset \mathbb{R}^2$ with infinite multiplicity.

One way of constructing the confidence set uses the superlevel filtration of the kernel density estimator and the bootstrap confidence band. Let $\mathcal{X} = \{X_1, X_2, ..., X_n\}$ as given points cloud, then the probability for the distribution of points can be estimated via KDE defined as following: $\hat{p}_h(x) := \frac{1}{nh^d}\sum_{i=1}^n K(\frac{x-X_i}{h})$ where h is the bandwidth and d as a dimension of the space. We compute $\hat{p}_h$ and $\hat{p}_h^*$, which are the KDE of $\mathcal{X}$ and the KDE of bootstrapped samples $\mathcal{X}^*$, respectively. Now, given the significance level α and $h > 0$, let confidence band $q_\mathcal{X}$ be bootstrap bandwidth of a Gaussian Empirical Process (van der Vaart, 2000; Kosorok, 2008), $\sqrt{n}||\hat{p}_h - \hat{p}_h^*||_\infty$. Then it satisfies $P(\sqrt{n}||\hat{p}_h - p_h||_\infty < q_\mathcal{X}) \ge 1 - \alpha$, as in Proposition C.2 in Section C. Then $\mathcal{B}_{d_B}(\hat{\mathcal{P}}_h, c_\mathcal{X})$, the ball of persistent homology centered at $\hat{\mathcal{P}}_h$ and radius $c_\mathcal{X} = q_\mathcal{X}/\sqrt{n}$ in the bottleneck distance d_B, is a valid confidence set as $\liminf_{n\to\infty} \mathbb{P}\left(\mathcal{P} \in \mathcal{B}_{d_B}(\hat{\mathcal{P}}_h, c_\mathcal{X})\right) \ge 1 - \alpha$. This confidence set has further interpretation that in the persistence diagram, homological features that are above twice the radius $2c_\mathcal{X}$ from the diagonal are simultaneously statistically significant.

B. Denoising topological features from outliers

Using the bootstrap bandwidth $c_\mathcal{X}$ as the threshold is the key part of our estimators `TopP&R` for robustly estimating $\mathrm{supp}(P)$. When the level set $\hat{p}_h^{-1}[c_\mathcal{X}, \infty)$ is used, the homology of $\hat{p}_h^{-1}[c_\mathcal{X}, \infty)$ consists of homological features whose $(\mathrm{birth}) \ge c_\mathcal{X}$ and $(\mathrm{death}) \le c_\mathcal{X}$, which are the homological features in skyblue area in Figure A1. In this example, we consider three types of homological noise, though there can be many more corresponding to different homological dimensions.

- There can be a 0-dimensional homological noise of $(\mathrm{birth}) < c_\mathcal{X}$ and $(\mathrm{death}) < c_\mathcal{X}$, which is the red point in the persistence diagram of Figure A1. This noise corresponds to the orange connected component on the left. As in the

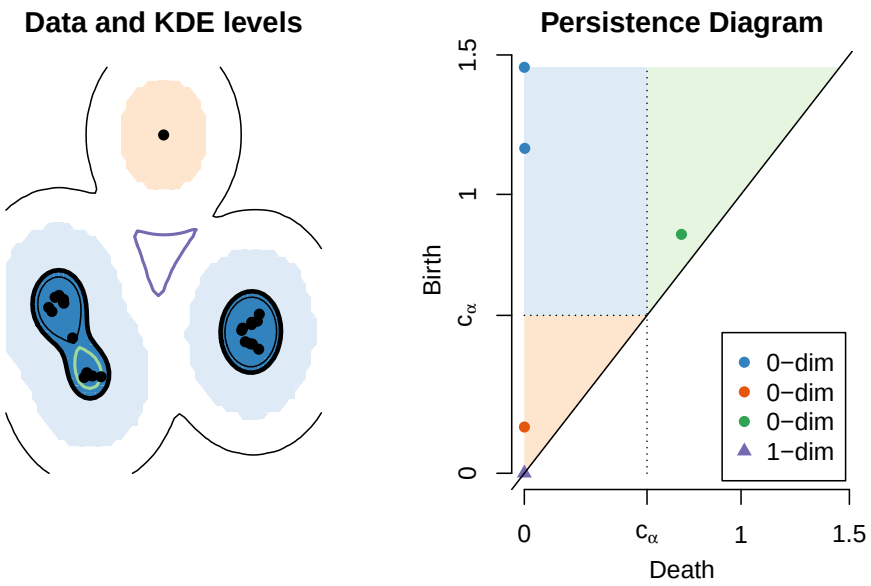


Figure A1: To robustly estimate the support, we use the bootstrap bandwidth c_α to filter out topological noise (orange) and keep topological signal (skyblue). Then `TopP&R` is computed on this support.

figure, this type of homological noise usually corresponds to outliers.

- There can be a 0-dimensional homological noise of $(\text{birth}) > c_{\mathcal{X}}$ and $(\text{death}) > c_{\mathcal{X}}$, which is the green point in the persistence diagram of Figure A1. This noise corresponds to the connected component surrounded by the green line on the left. As in the figure, this type of homological noise lies within the estimated support, not like the other two.
- There can be a 1-dimensional homological noise of $(\text{birth}) < c_{\mathcal{X}}$ and $(\text{death}) < c_{\mathcal{X}}$, which is the purple point in the persistence diagram of Figure A1. This noise corresponds to the purple loop on the left.

These homological noises satisfy either their $(\text{birth}) < c_{\mathcal{X}}$ and $(\text{death}) < c_{\mathcal{X}}$ or their $(\text{birth}) > c_{\mathcal{X}}$ and $(\text{death}) > c_{\mathcal{X}}$ simultaneously with high probability, so those homological noises are removed in the estimated support $\hat{p}_h^{-1}[c_{\mathcal{X}}, \infty)$, which is the blue area in the left and the skyblue area in the right in Figure A1.

We would like to further emphasize that homological noises are not restricted to 0-dimension lying outside the estimated support (red point in the persistence diagram of Figure A1). 0-dimensional homological noise inside the estimated support (green point in the persistence diagram of Figure A1), 1-dimensional homological noise can also arise, and the bootstrap bandwidth $c_{\mathcal{X}}$ allows to simultaneously filter them.

C. Assumptions on distributions and kernels

For distributions, we assume that the order of probability volume decay $P(\mathcal{B}(x, r))$ is at least r^d.

Assumption A1. *For all $x \in \text{supp}(P)$ and $y \in \text{supp}(Q)$,*

$$\liminf_{r \to 0} \frac{P\left(\mathcal{B}(x, r)\right)}{r^d} > 0, \qquad \liminf_{r \to 0} \frac{Q\left(\mathcal{B}(y, r)\right)}{r^d} > 0.$$

Remark C.1. Assumption A1 is analogous to Assumption 2 of Kim et al. (2019), but is weaker since the condition is pointwise on each $x \in \mathbb{R}^d$. And this condition is much weaker than assuming a density on $\mathbb{R}^d$: for example, a distribution supported on a low-dimensional manifold satisfies Assumption A1. This provides a framework suitable for high dimensional data, since many times high dimensional data lies on a low dimensional structure hence its density on $\mathbb{R}^d$ cannot exist. See Kim et al. (2019) for more detailed discussion.

For kernel functions, we assume the following regularity conditions:

Assumption A2. *Let $K : \mathbb{R}^d \to \mathbb{R}$ be a nonnegative function with $\|K\|_1 = 1$, $\|K\|_\infty, \|K\|_2 < \infty$, and satisfy the following:*

(1) $K(0) > 0$.

(2) K has a compact support.

(3) K is Lipschitz continuous and of second order.

Assumption A2 allows to build a valid bootstrap confidence band for kernel density estimator (KDE). See Theorem 12 of (Fasy et al., 2014) or Theorem 3.4 of (Neumann, 1998)

Proposition C.2 (Theorem 3.4 of (Neumann, 1998))**.** *Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be IID from a distribution P. For $h > 0$, let $\hat{p}_h, \hat{p}_h^*$ be kernel density estimator for $\mathcal{X}$ and its bootstrap $\mathcal{X}^*$, respectively, and for $\alpha \in (0, 1)$, let $c_\mathcal{X}$ be the α bootstrap quantile from $\sqrt{nh^d}\|\hat{p}_h - \hat{p}_h^*\|_\infty$. For $h_n \to 0$,*

$$\mathbb{P}\left(\sqrt{nh_n^d}\|\hat{p}_{h_n} - p_{h_n}\|_\infty > c_\mathcal{X}\right) = \alpha + \left(\frac{\log n}{nh_n^d}\right)^{\frac{4+d}{4+2d}}.$$

Assumption A1, A2 ensures that, when the bandwidth $h_n \to 0$, average KDEs are bounded away from 0.

Lemma C.3. *Let P be a distribution satisfying Assumption A1. Suppose K is a nonnegative function satisfying $K(0) > 0$ and continuous at 0. Suppose $\{h_n\}_{n\in\mathbb{N}}$ with $h_n \geq 0$ and $h_n \to 0$ is given. Then for all $x \in \mathrm{supp}(P)$,*

$$\liminf_n p_{h_n}(x) > 0.$$

Proof. Since $K(0) > 0$ and K is continuous at 0, there is $r_0 > 0$ such that for all $y \in B(0, r_0)$, $K(y) \geq \frac{1}{2}K(0) > 0$. And hence

$$\begin{aligned} p_h(x) &= \int \frac{1}{h^d} K\left(\frac{x-y}{h}\right) dP(y) \geq \int \frac{K(0)}{2h^d} 1\left(\frac{x-y}{h} \in B(0, r_0)\right) dP(y) \\ &\geq \frac{K(0)}{2h^d} P\left(B(x, r_0 h)\right). \end{aligned}$$

Hence as $h_n \to 0$,

$$\liminf_n p_{h_n}(x) > 0.$$

□

Before specifying the rate of convergence, we introduce the concept of reach. First introduced by (Federer, 1959), the reach is a quantity expressing the degree of geometric regularity of a set. Given a closed subset $A \subset \mathbb{R}^d$, the medial axis of A, denoted by $\mathrm{Med}(A)$, is the subset of $\mathbb{R}^d$ consisting of all the points that have at least two nearest neighbors on A.

$$\mathrm{Med}(A) = \left\{x \in \mathbb{R}^d \setminus A \colon \exists q_1 \neq q_2 \in A, ||q_1 - x|| = ||q_2 - x|| = d(x, A)\right\},$$

where $d(x, A) = \inf_{q\in A} ||q - x||$ denotes the distance from a generic point $x \in \mathbb{R}^d$ to A, The reach of A is then defined as the minimal distance from A to $\mathrm{Med}(A)$.

Definition C.4. The reach of a closed subset $A \subset \mathbb{R}^d$ is defined as

$$\mathrm{reach}(A) = \inf_{q\in A} d\left(q, \mathrm{Med}(A)\right) = \inf_{q\in A, x\in\mathrm{Med}(A)} ||q - x||.$$

Now, for specifying the rate of convergence, we first assume that distributions have densities away from 0 and ∞.

Assumption A3. *P and Q have Lebesgue densities p and q that, there exists $0 < p_{\min} \le p_{\max} < \infty$, $0 < q_{\min} \le q_{\max} < \infty$ with for all $x \in \mathrm{supp}(P)$ and $y \in \mathrm{supp}(Q)$,*

$$p_{\min} \le p(x) \le p_{\max}, \qquad q_{\min} \le q(x) \le q_{\max}.$$

We also assume weak geometric assumptions on the support of the distributions $\mathrm{supp}(P)$ and $\mathrm{supp}(Q)$, being bounded and having positive reach.

Assumption A4. *We assume $\mathrm{supp}(P)$ and $\mathrm{supp}(Q)$ are bounded. And the support of P and Q have positive reach, i.e. $\mathrm{reach}(\mathrm{supp}(P)) > 0$ and $\mathrm{reach}(\mathrm{supp}(Q)) > 0$.*

Sets with positive reach ensure that the volume of its tubular neighborhood grows in polynomial order, which is Theorem 26 from (Thäle, 2008) and originally from (Federer, 1959).

Proposition C.5. *Let A be a set with $\mathrm{reach}(A) > 0$. Let $A_r := \{x \in \mathbb{R}^d : d(x, A) \le r\}$. Then for $r < \mathrm{reach}(A)$, there exists $a_0, \dots, a_d \in \mathbb{R} \cup \{\infty\}$ that satisfies*

$$\mathcal{H}^d(A_r) = \sum_{k=0}^{d-1} a_k r^k.$$

D. Details and Proofs for Section 4

Let $\tilde{p}_h$ be the KDE on $\mathcal{X}^0$. For a finite set $\mathcal{X}$, we use the notation $c_{\mathcal{X},\alpha}$ for α-bootstrap quantile satisfying $\mathbb{P}\left(\left\|\hat{p}_{\mathcal{X},h_n} - \hat{p}_{\mathcal{X}^b,h_n}\right\|_\infty > c_{\mathcal{X},\alpha} | \mathcal{X}\right) = 1 - \alpha$, where $\mathcal{X}^b$ is the bootstrap sample from $\mathcal{X}$. For a distribution P, we use the notation $c_{P,\alpha}$ for α-quantile satisfying $\mathbb{P}\left(\|\hat{p}_{h_n} - p_{h_n}\|_\infty > c_{P,\alpha}\right) = 1 - \alpha$, where $\hat{p}_{h_n}$ is kernel density estimator of IID samples from P. Hence when $\mathcal{X}$ is not IID samples from P, the relation of Proposition C.2 may not hold.

Lemma D.1. *(1) Under Assumption 1, 2 and A2,*

$$\|\hat{p}_h - \tilde{p}_h\|_\infty \le \frac{\rho_n \|K\|_\infty}{h^d}.$$

(2) Under Assumption 1, 2 and A2,

$$c_{\mathcal{X}^0,\alpha+\delta} - O\left(\rho_n + \sqrt{\frac{\rho_n \log(1/\delta)}{nh^{2d}}}\right) \le c_{\mathcal{X},\alpha} \le c_{\mathcal{X}^0,\alpha-\delta} + O\left(\rho_n + \sqrt{\frac{\rho_n \log(1/\delta)}{nh^{2d}}}\right).$$

(3) Suppose Assumption 1, 2, A2 hold, and suppose $nh_n^{-d}\rho_n^2 \to 0$. Then with probability $1 - \alpha - 2\delta$,

$$\|\hat{p}_h - p_h\|_\infty < c_{\mathcal{X},\alpha} \le c_{P,\alpha-\delta}.$$

Proof. (1)

First, note that

$$\hat{p}_h - \tilde{p}_h = \frac{1}{nh^d} \sum_{i=1}^n \left(K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x - X_i^0}{h}\right)\right).$$

Then under Assumption A2,

$$\begin{aligned}\|\hat{p}_h - \tilde{p}_h\|_\infty &\le \frac{1}{nh^d} \sum_{i=1}^n \left\| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x - X_i^0}{h}\right)\right\|_\infty \\ &\le \frac{1}{nh^d} \sum_{i=1}^n \|K\|_\infty I\left(X_i \ne X_i^0\right).\end{aligned}$$

Then from Assumption 2, $\sum_{i=1}^n I\left(X_i \ne X_i^0\right) \le n\rho_n$, and hence

$$\|\hat{p}_h - \tilde{p}_h\|_\infty \le \frac{\|K\|_\infty \rho_n}{h^d}.$$

(2)

Let $\mathcal{X}_b$, $\mathcal{X}_b^0$ be bootstrapped samples of $\mathcal{X}$, $\mathcal{X}^0$ with the same sampling with replacement process. Let $\hat{p}_h^b, \tilde{p}_h^b$ be KDE of $\mathcal{X}_b$ and $\mathcal{X}_b^0$, respectively. And, note that

$$\left|\left\|\hat{p}_h - \hat{p}_h^b\right\|_\infty - \left\|\tilde{p}_h - \tilde{p}_h^b\right\|_\infty\right| \le \left\|\hat{p}_h - \tilde{p}_h\right\|_\infty + \left\|\hat{p}_h^b - \tilde{p}_h^b\right\|_\infty.$$

Let L_b be the number of elements where $\mathcal{X}_b$ and $\mathcal{X}_b^0$ differ, i.e., $L_b = \left|\mathcal{X}_b \backslash \mathcal{X}_b^0\right| = \left|\mathcal{X}_b^0 \backslash \mathcal{X}_b\right|$, then $L_b \sim \text{Binomial}(n, \rho_n)$, and

$$\left\|\hat{p}_h^b - \tilde{p}_h^b\right\|_\infty \le \frac{\|K\|_\infty L_b}{nh^d}.$$

And hence,

$$\left|\left\|\hat{p}_h - \hat{p}_h^b\right\|_\infty - \left\|\tilde{p}_h - \tilde{p}_h^b\right\|_\infty\right| \le \frac{\|K\|_\infty (n\rho_n + L_b)}{nh^d}.$$

Then by using subgaussian tail bound, with probability $1-\delta$,

$$\left|\left\|\hat{p}_h - \hat{p}_h^b\right\|_\infty - \left\|\tilde{p}_h - \tilde{p}_h^b\right\|_\infty\right| \le O\left(\rho_n + \sqrt{\frac{\rho_n \log(1/\delta)}{nh^{2d}}}\right).$$

Hence this implies

$$c_{\mathcal{X}^0,\alpha+\delta} - O\left(\frac{\rho_n}{h^d} + \sqrt{\frac{\rho_n \log(1/\delta)}{nh^{2d}}}\right) \le c_{\mathcal{X},\alpha} \le c_{\mathcal{X}^0,\alpha-\delta} + O\left(\frac{\rho_n}{h^d} + \sqrt{\frac{\rho_n \log(1/\delta)}{nh^{2d}}}\right).$$

(3)

Since $\mathcal{X}^0$ is IID samples from P, with probability $1-\alpha-2\delta$,

$$\left\|\tilde{p}_h - p_h\right\|_\infty < c_{P,\alpha+2\delta} \le c_{\mathcal{X}^0,\alpha+2\delta} + O\left(\frac{1}{nh^d}\right).$$

Now, note that $c_{\mathcal{X}^0,\alpha} = \Theta\left(\sqrt{\frac{\log(1/\alpha)}{nh^d}}\right)$, and hence

$$c_{\mathcal{X}^0,\alpha} - c_{\mathcal{X}^0,\alpha+\delta} = \Theta\left(\sqrt{\frac{\log(1/\alpha)}{nh^d}} - \sqrt{\frac{\log(1/(\alpha+\delta))}{nh^d}}\right) \ge \Omega\left(\frac{\log((\alpha+\delta)/\alpha)}{\sqrt{nh^d}}\right).$$

Then under Assumption 2, since $nh_n^{-d}\rho_n^2 = o(1)$ and $h_n^{-d}\rho_n = o(1)$,

$$\begin{aligned}
\left\|\hat{p}_h - p_h\right\| &\le \left\|\tilde{p}_h - p_h\right\|_\infty + \left\|\hat{p}_h - \tilde{p}_h\right\|_\infty \\
&< c_{\mathcal{X}^0,\alpha+2\delta} + O\left(\frac{1}{nh^d}\right) \\
&\le c_{\mathcal{X}^0,\alpha+\delta} - O\left(\frac{\rho_n}{h^d} + \sqrt{\frac{\rho_n \log(1/\delta)}{nh^{2d}}}\right) \\
&\le c_{\mathcal{X},\alpha} \\
&\le c_{\mathcal{X}^0,\alpha-\delta} + O\left(\frac{\rho_n}{h^d} + \sqrt{\frac{\rho_n \log(1/\delta)}{nh^{2d}}}\right) \\
&\le c_{P,\alpha-2\delta}.
\end{aligned}$$

□

Corollary D.2. *Suppose Assumption 1, 2, A2 hold.*

(1) Suppose $nh_n^{-d}\rho_n^2 \to 0$*. Then with probability* $1-\alpha-2\delta$*,*

$$p_{h_n}^{-1}[2c_{P,\alpha-2\delta},\infty) \subset \hat{p}_{h_n}^{-1}[c_{\mathcal{X},\alpha},\infty) \subset \mathrm{supp}(P_{h_n}).$$

(2) Suppose $mh_m^{-d}\rho_m^2 \to 0$*. Then with probability* $1-\alpha-2\delta$*,*

$$q_{h_m}^{-1}[2c_{Q,\alpha-2\delta},\infty) \subset \hat{q}_{h_m}^{-1}[c_{\mathcal{Y},\alpha},\infty) \subset \mathrm{supp}(Q_{h_m}).$$

Proof. (1) From Lemma D.1, $\|\hat{p}_h - p_h\| < c_{\mathcal{X},\alpha} \le c_{P,\alpha-2\delta}$. This implies

$$p_{h_n}^{-1}[2c_{P,\alpha-2\delta},\infty) \subset \hat{p}_{h_n}^{-1}[c_{\mathcal{X},\alpha},\infty) \subset \mathrm{supp}(P_{h_n}).$$

(2) can be proven similarly to (1).

□

Claim D.3. For a nonnegative measure μ and sets A, B, C, D,

$$\mu(A\cap B) - \mu(C\cap D) \le \mu(A\backslash C) + \mu(B\backslash D).$$

Proof.

$$\begin{aligned}\mu(A\cap B) - \mu(C\cap D) &\le \mu((A\cap B)\backslash(C\cap D)) = \mu((A\cap B)\cap(C^{\complement}\cup D^{\complement}))\\ &= \mu((A\cap B)\cap C^{\complement}) \cup (A\cap B)\cap D^{\complement}))\\ &\le \mu((A\cap B)\backslash C) + \mu(A\cap B)\backslash D)\\ &\le \mu(A\backslash C) + \mu(B\backslash D).\end{aligned}$$

□

From here, let P_n and Q_m be the empirical measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively, i.e., $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_m = \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}$.

Lemma D.4. *Suppose Assumption 1, 2 hold. Let* $A \subset \mathbb{R}^d$*. Then with probability* $1-\delta$*,*

$$|P_n - P|(A) \le \rho_n + \sqrt{\frac{\log(2/\delta)}{2n}},$$

and in particular,

$$P_n(A) \le P(A) + \rho_n + \sqrt{\frac{\log(2/\delta)}{n}}.$$

Proof. Let P_n^0 be the empirical measure on $\mathcal{X}^0$, i.e., $P_n^0 = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^0}$. By using Hoeffding's inbequality ,

$$\mathbb{P}\left(\left|\left|P_n^0 - P\right|(A)\right| \ge t\right) \le 2\exp\left(-2nt^2\right),$$

and hence with probability probability $1-\delta$,

$$\left|P_n^0 - P\right|(A) \le \sqrt{\frac{\log(2/\delta)}{2n}}.$$

And $\left|P_n - P_n^0\right|(A)$ is expanded as

$$\left|P_n - P_n^0\right|(A) = \frac{1}{n}\sum_{i=1}^n \left|I(X_i\in A) - I(X_i^0\in A)\right|.$$

Under Assumption 2 , $\sum_{i=1}^{n} I\left(X_i \neq X_i^0\right) \leq n\rho_n$, and hence

$$\begin{aligned}\left|P_n - P_n^0\right|(A) &= \frac{1}{n}\sum_{i=1}^{n}\left|I(X_i \in A) - I(X_i^0 \in A)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \neq X_i^0\right) = \rho_n.\end{aligned}$$

Therefore, with probability $1-\delta$,

$$|P_n - P|(A) \leq \rho_n + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

□

Claim D.5. Suppose Assumption 1, 2, A1, A2 hold.

(1) With probability $1-\alpha-8\delta$,

$$\begin{aligned}&\left|Q_m\left(\hat{p}_{h_n}^{-1}[c_{\mathcal{X}},\infty) \cap \hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right) - Q_m\left(\mathrm{supp}(P) \cap \mathrm{supp}(Q)\right)\right| \\ &\leq C\left(Q\left(\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right) \cup q_{h_m}^{-1}(0,2c_Q) \cup \mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right) + \rho_m + \sqrt{\frac{\log(1/\delta)}{m}}\right).\end{aligned}$$

(2) With probability $1-\alpha-8\delta$,

$$\left|Q_m\left(\hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right) - Q_m\left(\mathrm{supp}(Q)\right)\right| \leq C\left(Q\left(q_{h_m}^{-1}(0,2c_Q)\right) + \rho_m + \sqrt{\frac{\log(1/\delta)}{m}}\right).$$

(3) With probability $1-\alpha-8\delta$,

$$\begin{aligned}&\left|Q_m\left(\hat{p}_{h_n}^{-1}[c_{\mathcal{X}},\infty) \cap \hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right) - Q\left(\mathrm{supp}(P)\right)\right| \\ &\leq C\left(Q\left(\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right) \cup q_{h_m}^{-1}(0,2c_Q) \cup \mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right) + \rho_m + \sqrt{\frac{\log(1/\delta)}{m}}\right).\end{aligned}$$

(4) With probability $1-\alpha-8\delta$,

$$\left|Q_m\left(\hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right) - 1\right| \leq C\left(Q\left(q_{h_m}^{-1}(0,2c_Q)\right) + \rho_m + \sqrt{\frac{\log(1/\delta)}{m}}\right).$$

Proof. (1)

From Lemma D.2, with probability $1-\alpha-4\delta$,

$$\begin{aligned}Q_m\left(p_{h_n}^{-1}[2c_P,\infty) \cap q_{h_m}^{-1}[2c_Q,\infty)\right) &\leq Q_m\left(\hat{p}_{h_n}^{-1}[c_{\mathcal{X}},\infty) \cap \hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right) \\ &\leq Q_m\left(\mathrm{supp}(P_{h_n}) \cap \mathrm{supp}(Q_{h_m})\right).\end{aligned}$$

Then from the first inequality, combining with Claim D.3 gives

$$\begin{aligned}&Q_m\left(\hat{p}_{h_n}^{-1}[c_{\mathcal{X}},\infty) \cap \hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right) - Q_m\left(\mathrm{supp}(P) \cap \mathrm{supp}(Q)\right) \\ &\geq Q_m\left(p_{h_n}^{-1}[2c_P,\infty) \cap q_{h_m}^{-1}[2c_Q,\infty)\right) - Q_m\left(\mathrm{supp}(P) \cap \mathrm{supp}(Q)\right) \\ &\geq -\left(Q_m\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right) + Q_m\left(\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)\right)\right).\end{aligned}$$

And from the second inequality, combining with Claim D.3 gives

$$\begin{aligned}
&Q_m\left(\hat{p}_{h_n}^{-1}[c_{\mathcal{X}},\infty)\cap\hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right)-Q_m\left(\mathrm{supp}(P)\cap\mathrm{supp}(Q)\right)\\
&\le Q_m\left(\mathrm{supp}(P_{h_n})\cap\mathrm{supp}(Q_{h_m})\right)-Q_m\left(\mathrm{supp}(P)\cap\mathrm{supp}(Q)\right)\\
&\le Q_m\left(\mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right)+Q_m\left(\mathrm{supp}(Q_{h_m})\backslash\mathrm{supp}(Q)\right).
\end{aligned}$$

And hence

$$\begin{aligned}
&\left|Q_m\left(\hat{p}_{h_n}^{-1}[c_{\mathcal{X}},\infty)\cap\hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right)-Q_m\left(\mathrm{supp}(P)\cap\mathrm{supp}(Q)\right)\right|\\
&\le\max\left\{Q_m\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right)+Q_m\left(\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)\right)\right.\\
&\qquad\left.,\,Q_m\left(\mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right)+Q_m\left(\mathrm{supp}(Q_{h_m})\backslash\mathrm{supp}(Q)\right)\right\}.
\end{aligned}$$

Now from Lemma D.4, with probability $1-\delta$,

$$Q_m\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right)\le Q\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right)+\rho_m+\sqrt{\frac{\log(2/\delta)}{2m}}.$$

And similarly, with probabilty $1-\delta$,

$$\begin{aligned}
Q_m\left(\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)\right)&\le Q\left(\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)\right)+\rho_m+\sqrt{\frac{\log(2/\delta)}{2m}}\\
&=Q\left(q_{h_m}^{-1}(0,\infty)\right)+\rho_m+\sqrt{\frac{\log(2/\delta)}{2m}}.
\end{aligned}$$

And similaly, with probability $1-\delta$,

$$Q_m\left(\mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right)\le Q\left(\mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right)+\rho_m+\sqrt{\frac{\log(2/\delta)}{2m}}.$$

And similarly, with probability $1-\delta$,

$$\begin{aligned}
Q_m\left(\mathrm{supp}(Q_{h_m})\backslash\mathrm{supp}(Q)\right)&\le Q\left(\mathrm{supp}(Q_{h_m})\backslash\mathrm{supp}(Q)\right)+\rho_m+\sqrt{\frac{\log(2/\delta)}{2m}}\\
&=\rho_m+\sqrt{\frac{\log(2/\delta)}{2m}}.
\end{aligned}$$

Hence by putting these altogether, with probability $1-\alpha-8\delta$,

$$\begin{aligned}
&\left|Q_m\left(\hat{p}_{h_n}^{-1}[c_{\mathcal{X}},\infty)\cap\hat{q}_{h_m}^{-1}[c_{\mathcal{Y}},\infty)\right)-Q_m\left(\mathrm{supp}(P)\cap\mathrm{supp}(Q)\right)\right|\\
&\le C\left(Q\left(\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right)\cup q_{h_m}^{-1}(0,2c_Q)\cup\mathrm{supp}(P_{h_m})\backslash\mathrm{supp}(P)\right)+\rho_m+\sqrt{\frac{\log(1/\delta)}{m}}\right).
\end{aligned}$$

And by letting $B_{n,m}:=\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\right)\cup q_{h_m}^{-1}(0,2c_Q)\cup\mathrm{supp}(P_{h_m})\backslash\mathrm{supp}(P)$, note that for all $x\in\mathrm{supp}(P)$, $\liminf_n p_{h_n}(x)>0$, so $p_{h_n}(x)>2c_P$ for large enough n. And hence

$$\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\to\emptyset.$$

And similar argument holds for $\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)$, so $\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)\to\emptyset$ as well.

Now, note that from Lemma C.3 implies that for all $x\in\mathrm{supp}(P)$, $\liminf_n p_{h_n}(x)>0$, so $p_{h_n}(x)>2c_P$ for large enough n. And hence

$$\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P,\infty)\to\emptyset.$$

And similar argument holds for $\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)$, so $\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q,\infty)\to\emptyset$ as well.

Also, since K has compact support, for any $x \notin \mathrm{supp}(P)$, $x \notin \mathrm{supp}(P_{h_n})$ once $h_n < d(x, \mathrm{supp}(P))$. Hence $\mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P) \to \emptyset$, and similarly $\mathrm{supp}(Q_{h_m})\backslash\mathrm{supp}(Q) \to \emptyset$ as well. Hence as $n, m \to \infty$,

$$Q(B_{n,m}) \to 0$$

(2)

This can be done similarly to (1).

(3)

Lemma D.4 gives that with probability $1-\delta$,

$$|Q_m\left(\mathrm{supp}(P) \cap \mathrm{supp}(Q)\right) - Q\left(\mathrm{supp}(P)\right)| \le \rho_m + \sqrt{\frac{\log(2/\delta)}{2m}}.$$

Hence combining with (1) gives the desired result.

(4)

Lemma D.4 gives that with probability $1-\delta$,

$$|Q_m\left(\mathrm{supp}(Q)\right) - 1| \le \rho_m + \sqrt{\frac{\log(2/\delta)}{2m}}.$$

Hence combining with (2) gives the desired result.

□

Below we state a more formal version of Proposition 4.1.

Proposition D.6. *Suppose Assumption 1,2,A1,A2 hold. Suppose $\alpha \to 0$, $h_n \to 0$, $nh_n \to \infty$, $nh_n^{-d}\rho_n^2 \to 0$, and similar relations hold for h_m, ρ_m. Then there exists some constant $C > 0$ not depending on anything else such that*

$$|\mathrm{TopP}_{\mathcal{X}}(\mathcal{Y}) - \mathrm{precision}_P(\mathcal{Y})| \le C\left(Q(B_{n,m}) + \rho_m + \sqrt{\frac{\log(1/\delta)}{m}}\right),$$

$$\left|\mathrm{TopR}_{\mathcal{Y}}(\mathcal{X}) - \mathrm{recall}_Q(\mathcal{X})\right| \le C\left(P(A_{n,m}) + \rho_n + \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where

$$A_{n,m} := \left(\mathrm{supp}(Q)\backslash q_{h_m}^{-1}[2c_Q, \infty)\right) \cup p_{h_n}^{-1}(0, 2c_P) \cup \mathrm{supp}(Q_{h_m})\backslash\mathrm{supp}(Q),$$
$$B_{n,m} := \left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P, \infty)\right) \cup q_{h_m}^{-1}(0, 2c_Q) \cup \mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P).$$

And as $n, m \to \infty$, $P(A_{n,m}) \to 0$ and $Q(B_{n,m}) \to 0$ hold.

Proof of Proposition D.6. Now this is a combination of Claim D.5 (1) (2).

□

Similarly we state a more formal version of Theorem 4.2.

Theorem D.7. *Suppose Assumption 1,2,A1,A2 hold. Suppose $\alpha \to 0$, $h_n \to 0$, $nh_n \to \infty$, $nh_n^{-d}\rho_n^2 \to 0$, and similar relations hold for h_m, ρ_m. Then there exists some constant $C > 0$ not depending on anything else such that*

$$|\mathrm{TopP}_{\mathcal{X}}(\mathcal{Y}) - \mathrm{precision}_P(Q)| \le C\left(Q(B_{n,m}) + \rho_m + \sqrt{\frac{\log(1/\delta)}{m}}\right),$$

$$\left|\mathrm{TopR}_{\mathcal{Y}}(\mathcal{X}) - \mathrm{recall}_Q(P)\right| \le C\left(P(A_{n,m}) + \rho_n + \sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $A_{n,m}$and $B_{n,m}$ are the same as in Proposition D.6. Again as $n, m \to \infty$, $P(A_{n,m}) \to 0$ and $Q(B_{n,m}) \to 0$ hold.

Proof of Theorem D.7. Now this is a combination of Claim D.5 (3) (4).

□

Proof of Lemma 4.3. Note that $B_{n,m}$ is defined as

$$B_{n,m} = \left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P, \infty)\right) \cup q_{h_m}^{-1}(0, 2c_Q) \cup \mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P),$$

and hence $Q(B_{n,m})$ can be upper bounded as

$$Q(B_{n,m}) \leq Q\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P, \infty)\right) + Q\left(q_{h_m}^{-1}(0, 2c_Q)\right) + Q\left(\mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right).$$

Suppose $\mathrm{supp}(K) \subset B_{\mathbb{R}^d}(0,1)$ for convenience, and let $A_{-h_n} := \left\{x \in \mathrm{supp}(P) : d(x, \mathbb{R}^d\backslash\mathrm{supp}(P)) \geq h_n\right\}$. Then from Assumption A3, for all $x \in A_{-h_n}$, $p_{h_n}(x) \geq p_{\min}$ holds. Hence for large enough N such that for $n \geq N$, $c_P < p_{\min}$, then $p_{h_n}(x) \geq p_{\min}$ for all $x \in A_{-h_n}$, and hence

$$Q\left(\mathrm{supp}(P)\backslash p_{h_n}^{-1}[2c_P, \infty)\right) \leq Q\left(\mathrm{supp}(P)\backslash A_{-h_n}\right).$$

Also, $\mathrm{supp(P)}$ being bounded implies that all the coefficients $a_0, \ldots, a_{d-1}$ in Proposition C.5 for $\mathrm{supp}(P)$ are in fact finite. Then $q \leq q_{\max}$ from Assumption A3 and Proposition C.5 implies

$$Q\left(\mathrm{supp}(P)\backslash A_{-h_n}\right) = O(h_n).$$

And $Q\left(q_{h_m}^{-1}(0, 2c_Q)\right) = O(h_m)$ follows with a similar argument. Finally, Proposition C.5 implies

$$Q\left(\mathrm{supp}(P_{h_n})\backslash\mathrm{supp}(P)\right) = O(h_n).$$

Hence altogether,

$$Q(B_n) = O(h_n + h_m).$$

□

E. Philosophy of our metric & Practical Scenarios

E.1. Philosophy of our metric

All evaluation metrics have different resolutions and properties. The philosophy of our metric is to propose a reliable evaluation metric based on statistically and topologically certain things. In a real situation, the data often contains outliers or noise, and these data points come from a variety of sources (e.g. human error, feature embedding network). These errors may play as outliers, which leads to overestimation of the data distribution, which in turn leads to a false impression of improvement when developing generative models. As discussed in section 2., `P&R` and its variants have different ways to estimate the supports, which overlook the possible presence and effect of noise. Naeem et al. (2020) revealed that previous support estimation approaches may inaccurately estimate the true support under noisy conditions, and partially solved this problem by proposing to only use the real support. However, this change goes beyond the natural definition of precision and recall, and results in losing some beneficial properties like boundedness. Moreover, this is still a temporary solution since real data can also contain outliers. We propose a solution to the existing problem by minimizing the effects of noise, using topologically significant data points and retaining the natural definition of precision and recall.

E.2. Practical Scenarios

From this perspective, we present two examples of realistic situations where outliers exist in the data and filtering out them can have a significant impact on proper model analysis and evaluation. With real data, there are many cases where outliers are introduced into the data due to human error (Pleiss et al., 2020; Li et al., 2022). Taking the simplest MNIST as an example, suppose our task of interest is to generate 4. Since image number 7 is included in data set number 4 due to incorrect labeling (see Figure 1 of (Pleiss et al., 2020)), the support of the real data in the feature space can be overestimated by such outliers, leading to an unfair evaluation of generative models (as in Section 5.2.2 and 5.3.3); That is, the sample generated with weird noise may be in the overestimated support, and existing metrics without taking into account the reliability of the support could not penalize this, giving a good score to a poorly performing generator.

A similar but different example is when noise or distortions in the captured data (unfortunately) behave adversely on the feature embedding network used by the current evaluation metrics (as in Section 5.2.2 and 5.3.3); e.g., visually it is the number 7, but it is mismapped near the feature space where there are usually 4 and becomes an outlier. Then the same problem as above may occur. Note that in these simple cases, where the definition of outliers is obvious with enough data, one could easily examine the data and exclude outliers a priori to train a generative model. In the case of more complicated problems such as the medical field (Li et al., 2022), however, it is often not clear how outliers are to be defined. Moreover, because data are often scarce, even outliers are very useful and valuable in practice for training models and extracting features, making it difficult to filter outliers in advance and decide not to use them.

On the other hand, we also provide an example where it is very important to filter out outliers in the generator sample and then evaluate them. To evaluate the generator, samples are generated by sampling from the preset latent space (typically Gaussian). Even after training is complete and the generators' outputs are generally fine, there's a latent area where generators aren't fully trained. Note that latent space sampling may contain samples from regions that the generator does not cover well during training ("unfortunate outlier"). When unfortunate outliers are included, the existing evaluation metrics may underestimate or overestimate the generator's performance than its general performance. (To get around this, it is necessary to try this evaluation several times to statistically stabilize it, but this requires a lot of computation and becomes impractical, especially when the latent space dimension is high.)

Especially considering the evaluation scenario in the middle of training, the above situation is likely to occur due to frequent evaluation, which can interrupt training or lead to wrong conclusions. On the other hand, we can expect that our metric will be more robust against the above problem since it pays more attention to the core (samples that form topologically meaningful structures) generation performance of the model.

F. Experimental Details

F.1. Implementation details of embedding

We summarize the detailed information of our embedding networks implemented for the experiments. In Figure 2, 3, 4, 5, A2, and 6, `P&R` and `D&C` are computed from the features of ImageNet pre-trained VGG16 (fc2 layer), and `TopP&R` is computed from features placed in $\mathcal{R}^{32}$ with additional random linear projection. In the experiment in Figure 5, the SwAV embedder is additionally considered. We implement ImageNet pre-trained InceptionV3 (fc layer), VGG16 (fc2 layer), and SwAV as embedding networks with random linear projection to 32 dimensional feature space to compare the ranking of GANs in Table 1.

F.2. Implementation details of confidence band estimator

Algorithm 1 Confidence Band Estimator

KDE: kernel density estimator
h: kernel bandwidth parameter
k: number of repeats
$\hat{\theta}$: set of difference

Given $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$
$\hat{p}_h = KDE(\mathcal{X})$
for *iteration* $= 1, 2, \ldots, k$ **do**
  # Define $\mathcal{X}^*$ with bootstrap sampling
  $\mathcal{X}^*$= random sample n times with repeat from $\mathcal{X}$
  # $\hat{p}_h^*$ replaces population density
  $\hat{p}_h^* = KDE(\mathcal{X}^*)$
  # compute $\hat{\theta}$ with bootstrap samples
  Append $\hat{\theta}$ with $\sqrt{n}||\hat{p}_h - \hat{p}_h^*||_\infty$
end for

# grid search for the confidence band
for $q \in [min(\hat{\theta}), max(\hat{\theta})]$ **do**
  count = 0
  for *element* $\in \hat{\theta}$ **do**
    if *element* $> q$ **then**
      count $=$ *count* $+ 1$
    end if
  end for
  # define the band threshold
  if *count*$/k \approx \alpha$ **then**
    $q_\alpha = q$
  end if
end for
# define estimated confidence band
$c_\alpha = q_\alpha/\sqrt{n}$

F.3. Choice of confidence level

For the confidence level α, we would like to point out that α is not the usual hyperparameter to be tuned: It has a statistical interpretation of the probability or the level of confidence to allow error, noise, etc. The most popular choices are $\alpha = 0.1, 0.05, 0.01$, leading to 90%, 95%, 99% confidence. We used $\alpha = 0.1$ throughout our experiments.

F.4. Estimation of Bandwidth parameter

As we discussed in section 2, since `TopP&R` estimates the manifold through KDE with kernel bandwidth parameter h, we need to approximate it. The estimation techniques for h are as follows: (**a**) a method of selecting h that maximizes the survival time ($S(h)$) or the number of significant homological features ($N(h)$) based on information obtained about persistent homology using the filtration method, (**b**) a method using the median of the k-nearest neighboring distances between features obtained by the balloon estimator (for more details, please refer to (Chazal et al., 2017), (Wagner et al., 2012), and (Terrell & Scott, 1992)). Note that, the bandwidth h for all the experiments in this paper are estimated via Balloon Bandwidth Estimator.

For **(a)**, following the notation in Section A, let the ith homological feature of persistent diagram be (b_i, d_i), then we define its life length as $l_i(h) = d_i - b_i$ at kernel bandwidth h. With confidence band $c_\alpha(h)$, we select h that maximizes one of the following two quantities:

$$N(h) = \#\{i : \ l_i(h) > c_\alpha(h)\}, \ S(h) = \sum_i [l_i(h) - c_\alpha(h)]_+.$$

Note that, we denote the confidence band c_α as $c_\alpha(h)$ considering the kernel bandwidth parameter h of KDE in Algorithm 1.

For **(b)**, the balloon bandwidth estimator is defined as bellow:

Algorithm 2 Balloon Bandwidth Estimator

```
# h: kernel bandwidth
# KND: kth nearest distance
# idx: index
# sort: sort in ascending order

Given 𝒳 = {X_1, X_2, ..., X_n}
for idx = 1, 2, ..., n do
  # Compute L2 distance between X_idx and 𝒳
  d(X_idx, 𝒳) = ||X_idx − 𝒳||_2^2,
  i.e., {d(X_idx, X_1), d(X_idx, X_2), d(X_idx, X_3), ...}
  # Calculate kth nearest neighbor distance
  KND_idx = sort(d(X_idx, 𝒳))[k]
end for
Given KND = {KND_1, KND_2, ..., KND_n}
# Define the estimated bandwidth ĥ
ĥ = median(KND)
```

F.5. Mean hamming distance

Hamming distance (HD) (Hamming, 1950) counts the number of items with different ranks between A and B, then measures how much proportion differs in the overall order, i.e. for $A_i \in A$ and $B_i \in B$, $HD(A, B) := \sum_{i=1}^{n} 1\{k : A_i \neq B_i\}/n$ where n is the number of items in list A or B, and k is the number of differently ordered elements. The **mean HD** is calculated as follows to measure the average distances of three ordered lists: Given three ordered lists $A, B,$ and C, $\bar{HD} = (HD(A, B) + HD(A, C) + HD(B, C))/3$.

F.6. Explicit values of bandwidth parameter

Since our metric adaptively reacts to the given samples of P and Q, we have two hs per experiment. For example, in the translation experiment (Figure 2), there are 13 steps in total, and each time we estimate h for P and Q, resulting in a total of 26 hs. To show them all at a glance, we have listed all values in one place. We will also provide the code that can reproduce the results in our experiments upon acceptance.

μ	$\leftarrow$ shift												shift $\rightarrow$
real h	8.79	8.60	8.58	8.67	8.63	8.71	8.66	8.73	8.54	8.63	8.54	8.75	8.78
fake h	8.46	8.86	8.60	8.49	8.44	8.80	8.55	8.63	8.74	8.61	8.58	8.83	8.60

Table A1: Bandwidth parameters h of distribution shift in Section 5.2.1.

Steps	0	1	2	3	4	5	6
real h	10.5	10.6	10.5	10.7	10.5	10.8	10.6
fake h	11.0	10.2	10.0	9.82	9.57	9.29	8.78

Table A2: Bandwidth parameters h of sequential mode drop in Section 5.2.2.

Steps	0	1	2	3	4	5	6	7	8	9	10
real h	10.5	10.6	10.6	10.8	10.3	10.6	10.7	10.6	10.4	10.4	10.6
fake h	10.6	10.5	10.7	10.3	10.4	10.0	9.88	9.31	9.19	9.07	8.79

Table A3: Bandwidth parameters h of simultaneous mode drop in Section 5.2.2.

Steps		0	1	2	3	4	5	6	7	8	9
Scatter	real h	8.76	8.71	8.78	8.59	8.85	8.71	8.61	8.97	9.00	8.95
	fake h	8.61	8.76	8.63	8.93	8.46	8.72	8.67	8.68	9.07	8.88
Swap	real h	8.65	8.44	8.76	8.62	8.55	8.75	8.96	8.59	8.53	8.74
	fake h	8.52	8.55	8.68	8.97	8.87	8.94	8.84	9.05	8.94	8.90

Table A4: Bandwidth parameters h of non-IID noise perturbation in Section 5.2.3.

Ψ	$\Psi \downarrow$										$\Psi \uparrow$
real h	6.26	6.41	6.31	6.23	6.06	6.24	6.20	6.61	6.39	6.28	6.26
fake h	1.04	1.67	2.36	2.90	3.41	3.93	4.31	4.90	5.51	5.60	6.46

Table A5: Bandwidth parameters h of FFHQ truncation trick in Section 5.3.1.

Steps	0	1	2	3	4	5	6	7	8	9
real h	7.09	6.95	6.96	6.92	7.02	6.81	6.86	6.88	6.78	6.67
fake h	6.87	6.75	6.88	6.64	6.40	6.64	6.41	6.44	6.62	6.04

Table A6: Bandwidth parameters h of CIFAR10 sequential mode drop in Section 5.3.2.

Steps	0	1	2	3	4	5	6	7	8	9	10
real h	6.78	6.95	7.05	6.62	6.85	6.51	6.94	6.91	6.80	7.00	6.78
fake h	6.75	6.69	7.14	6.68	6.75	6.84	6.61	6.68	6.50	6.32	6.17

Table A7: Bandwidth parameters h of CIFAR10 simultaneous mode drop in Section 5.3.2.

Steps		0	1	2	3	4	5	6	7	8	9
Scatter	real h	7.24	7.03	7.01	7.45	7.38	7.66	7.46	7.36	7.79	7.66
	fake h	5.41	5.77	5.67	5.49	5.71	5.65	5.82	6.08	5.81	6.02
Swap	real h	7.25	7.20	6.99	6.67	7.07	6.96	7.03	6.94	6.83	6.86
	fake h	6.87	6.63	6.60	6.91	6.74	6.77	6.92	6.83	6.83	6.82

Table A8: Bandwidth parameters h of FFHQ non-IID noise perturbation in Section 5.3.3.

	Noise intensity	0	1	2	3
Gaussian noise	real h	6.52	6.27	6.41	6.49
	fake h	6.32	6.15	4.06	2.64
Gaussian blur	real h	6.08	6.26	6.26	6.47
	fake h	6.12	5.13	4.60	3.40
Black rectangle	real h	6.06	6.28	6.49	6.44
	fake h	6.36	6.66	6.31	5.93

Table A9: Bandwidth parameters h of FFHQ noise addition in Section 5.3.4.

		StyleGAN2	ReACGAN	BigGAN	PDGAN	ACGAN	WGAN-GP
InceptionV3	real h	2.605	2.629	2.581	2.587	2.615	2.606
	fake h	2.226	2.494	2.226	2.639	2.573	2.063
VGG16	real h	7.732	8.021	8.175	7.792	7.845	8.036
	fake h	7.880	6.839	7.178	6.174	6.839	3.615
SwAV	real h	0.774	0.765	0.785	0.780	0.780	0.822
	fake h	0.740	0.667	0.6746	0.619	0.627	0.502

Table A10: Bandwidth parameters h of GAN ranking in Section 5.3.5.

G. Additional Experiments

G.1. Trucation trick

ψ is a parameter for the truncation trick and is first introduced in (Brock et al., 2018) and (Karras et al., 2019). We followed the approach in (Brock et al., 2018) and (Karras et al., 2019). GANs generate images using the noise input z, which follows the standard normal distribution $\mathcal{N}(0, I)$ or uniform distribution $\mathcal{U}(-1, 1)$. Suppose GAN inadvertently samples noise outside of distribution, then it is less likely to sample the image from the high density area of the image distribution $p(z)$ defined in the latent space of GAN, which leads to generate an image with artifacts. The truncation trick takes this into account and uses the following truncated distribution. Let f be the mapping from the input to the latent space. Let $w = f(z)$, and $\bar{w} = \mathbb{E}[f(z)]$, where z is either from $\mathcal{N}(0, I)$ or $\mathcal{U}(-1, 1)$. Then we use $w' = \bar{w} + \psi(w - \bar{w})$ as a truncated latent vector. If the value of ψ increases, then the degree of truncation decreases which makes images have greater diversity but possibly lower fidelity.

G.2. Survivability of minority sets in the long-tailed distribution

	Before mode drop	**After mode drop**
Proportion of survival (before / after filtering)	100% → 59%	100% → 57%
TopP&R (Fidelity / Diversity)	0.99 / 0.96	0.99 / 0.87 (9 p.p ↓)
P&R (Fidelity / Diversity)	0.73 / 0.73	0.74 / 0.70 (3 p.p ↓)
D&C (Fidelity / Diversity)	0.99 / 0.96	1.02 / 0.94 (2 p.p ↓)

Table A11: Proportion of surviving minority samples in the long-tailed distribution after the noise exclusion with confidence band $c_{\mathcal{X}}$. The p.p. indicates the percentage points.

An important point to check in our proposed metric is the possibility that a small part of the total data (i.e., minority sets), but containing important information, can be ignored by the confidence band. We emphasize that since our metric takes topological features into account, even minority sets are not filtered conditioned that they have topologically significant structures. We assume that signals or data that are minority sets have topological structures, but outliers exist far apart and lack a topological structure in general.

To test this, we experimented with CIFAR10, which has 5,000 samples per class. We simulate a dataset with the majority set of six classes (2,000 samples per class, 12,000 total) and the minority set of four classes (500 samples per class, 2,000 total), and an ideal generator that exactly mimics the full data distribution. As shown in the Table A11, the samples in the

minority set remained after the filtering process, meaning that the samples were sufficient to form a significant structure. Both D&C and TopP&R successfully evaluate the distribution for the ideal generator. To check whether our metric reacts to the change in the distribution even with this harsh setting, we also carried out the mode decay experiment. We dropped the samples of the minority set from 500 to 100 per class, which can be interpreted as an 11.3% decrease in diversity relative to the full distribution (Given (1) ratio of the number of samples between majority and minority sets = $12,000 : 2,000 = 6 : 1$ and (2) 80% decrease in samples per minority class, the true decay in the diversity is calculated as $\frac{1}{(1+6)} \times 0.8 = 11.3\%$ with respect to the enitre samples). Here, recall and coverage react somewhat less sensitively with their reduced diversities as 3 p.p. and 2 p.p., respectively, while TopR reacted most similarly (9 p.p.) to the ideal value. In summary, TopP&R shows much more sensitiveness to the changes in data distribution like mode decay. Thus, once the minority set has survived the filtering process, our metric is likely to be much more responsive than existing methods.

G.3. Sequential and simultaneous mode dropping with CIFAR-10

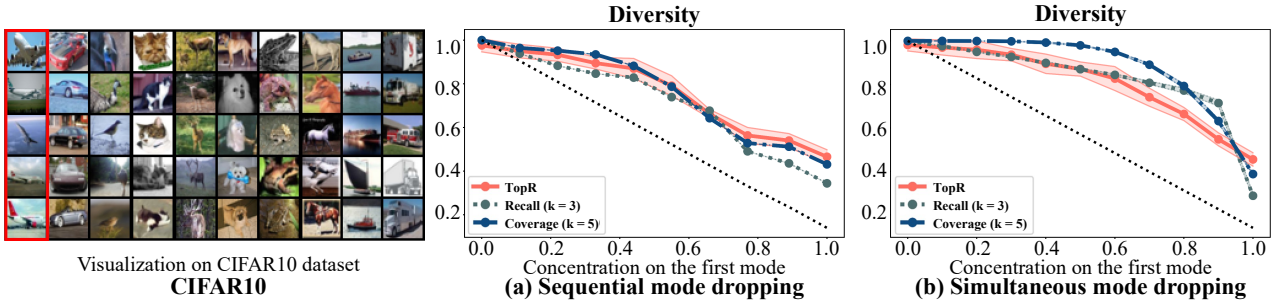


Figure A2: Comparison of evaluation metrics under sequential and simultaneous mode dropping scenario with CIFAR-10.

G.4. Experiment on Non-IID perturbation with outlier removal methods

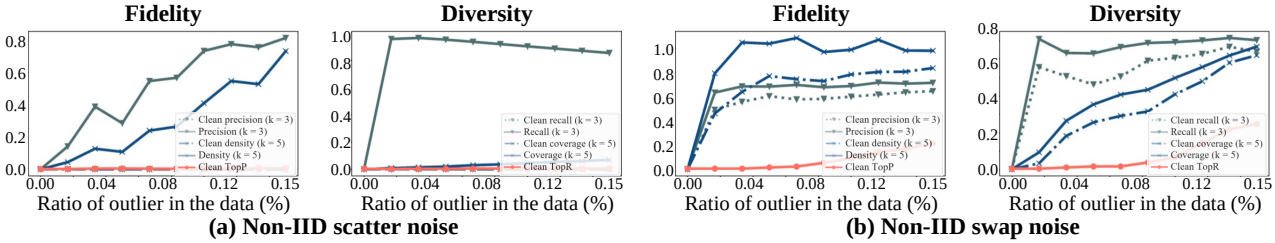


Figure A3: Behaviors of evaluation metrics on Non-IID perturbations. The dotted line in the graph shows the performance of metrics after the removal of outliers using the IF. We use "Clean" as a prefix to denote the evaluation after IF.

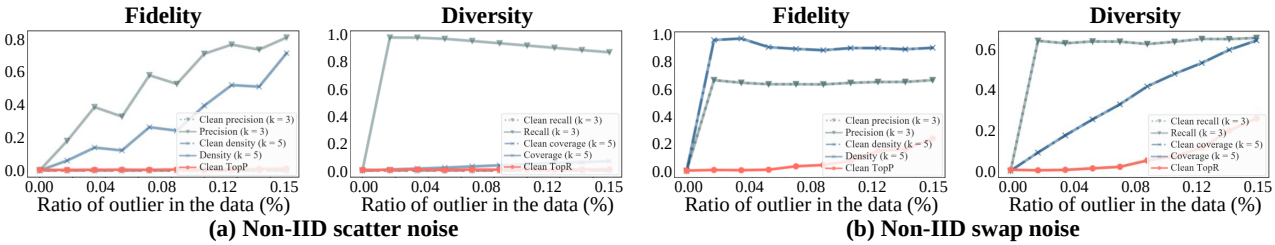


Figure A4: Behaviors of evaluation metrics on Non-IID perturbations. The dotted line in the graph shows the performance of metrics after the removal of outliers using the LOF. We use "Clean" as a prefix to denote the evaluation after LOF.

We compared the performance of metrics on a denoised dataset using Local Outlier Factor (LOF) (Breunig et al., 2000) and Isolation Forest (IF) (Liu et al., 2008). From the result, P&R and D&C still do not provide a stable evaluation, while TopP&R shows the most consistent evaluation for the two types of noise without changing its trend. Since TopP&R can discriminate topologically significant signals by estimating the confidence band (using KDE filtration function), there is

no need to arbitrarily set cutoff thresholds for each dataset. On the other hand, in order to distinguish inliers from specific datasets using LOF and IF methods, a threshold specific to the dataset must be arbitrarily set each time, so there is a clear constraint that different results are expected each time for each parameter setting (which is set by the user). Through this, it is confirmed that our evaluation metric guarantees the most consistent scoring based on the topologically significant signals. Since the removal of outliers in the `TopP&R` is part of the support estimation process, it is not appropriate to replace our unified process with the LOF or IF. In detail, `TopP&R` removes outliers through a clear threshold called confidence band which is defined by KDE, and this process simultaneously defines super-level set as the estimated support. If this process is separated, the topological properties and interpretations of estimated support also disappear. Therefore, it is not practical to remove outliers by other methods when calculating `TopP&R`.