Unifying and extending Precision Recall metrics for assessing generative models

Benjamin Sykes [*1] Loïc Simon [*1] Julien Rabin [*1]

Abstract

With the recent success of generative models in image and text, the evaluation of generative models has gained a lot of attention. Whereas most generative models are compared in terms of scalar values such as Fréchet Inception Distance (FID) or Inception Score (IS), in the last years (Sajjadi et al., 2018) proposed a definition of precision-recall curve to characterize the closeness of two distributions. Since then, various approaches to precision and recall have seen the light (Kynkäänniemi et al., 2019; Naeem et al., 2020; Park & Kim, 2023). They center their attention on the extreme values of precision and recall, but apart from this fact, their ties are elusive. In this paper, we unify most of these approaches under the same umbrella, relying on the work of (Simon et al., 2019). Doing so, we were able not only to recover entire curves, but also to expose the sources of the accounted pitfalls of the concerned metrics. We also provide consistency results that go well beyond the ones presented in the corresponding literature. Last, we study the different behaviors of the curves obtained experimentally.

1. Introduction

In this article, we consider metrics designed to evaluate the adequacy of a generative model to the distribution it is assumed to capture. In itself, this problem consists of evaluating the closeness of the real distribution, hereafter denoted by P and the generated one, denoted by Q. In an early period, several scalar metrics were designed such as Inception Score (Salimans et al., 2016) and the iconic Fréchet Inception Distance (Heusel et al., 2017). A rich literature completes this line of research, pointing at limitations and extensions or concurrent scalar metrics. Notwithstanding, one pitfall is shared by any such scalar metric in that they cannot account separately for two types of failures: namely for the lack of realism (a.k.a fidelity), and the lack of variability (diversity). This assessment was first carried out by (Sajjadi et al., 2018) where the authors developed a trade-off curve known as the Precision Recall Curve, which characterizes both types of flaws. Each point of the curve has two components α (a.k.a precision) and β (a.k.a recall) which in essence represent respectively the mass of P and Q that can be extracted simultaneously by selecting a subset of their common support (the formal definition will be clarified later on). This intuitive description is illustrated in Fig. 1.

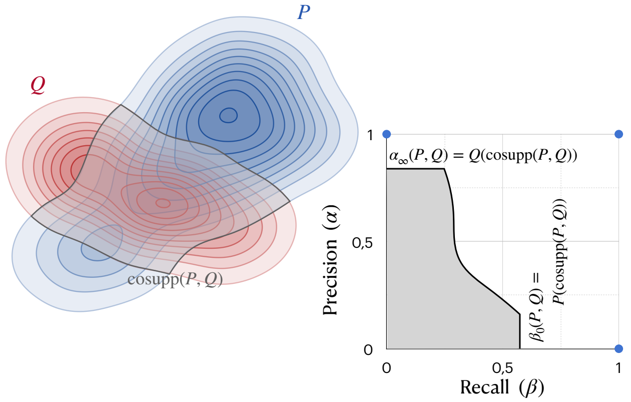


Figure 1: Right: the PR-curve is the frontier of the shaded area composed of all admissible PR pairs (β, α). In essence, these pairs represent the mass of P and Q that one can recover by selecting a subset of the common support (gray area on the left). More precisely, by selecting regions of high likelihood of P, one trades precision (α) in favor of recall (β). The extreme values $\beta_0(P,Q)$ and $\alpha_\infty(P,Q)$ embody the respective masses of the entire common support.

The authors of (Sajjadi et al., 2018) give a formal definition of their curve for general distributions, but they were able to provide a practical characterization (amenable to empirical evaluation) only in the case of discrete distributions, and relied on clustering to get an algorithmic evaluation of the curve. Followup theoretical insights were provided in (Simon et al., 2019) where an alternate characterization was exposed which extend to general distributions. More details will be given later on since it will be central in our developments. This work was also complemented by a deep theoretical analysis of the link between this curve and many other statistical notions in (Siry et al., 2023). In particular, the authors show that PR curves are in fact equivalent to Divergence Frontiers, which were developed in (Djolonga et al., 2020) in an attempt to generalize PR curves.

In parallel to these theoretical works, a handful of practical metrics have been developed such as (Kynkäänniemi et al., 2019; Naeem et al., 2020; Cheema & Urner, 2023; Khayatkhoei & AbdAlmageed, 2023; Park & Kim, 2023). All of them point to the shortcomings of earlier attempts and propose an alternative to improve the concerned aspects. A few remarks can be made already about most of these approaches. Starting from (Kynkäänniemi et al., 2019), it was argued that evaluating the extreme Precision and Recall was enough in practice, and therefore, instead of extracting a whole curve all of these variants only evaluates two scalar metrics: namely the extreme precision (denoted by $\alpha_\infty(P, Q)$) and the extreme recall ($\beta_0(P, Q)$). This choice may appear justified, but in many cases, the theoretical values of both metrics reach their saturation level (i.e. 1) even though the two distributions P and Q differ substantially. As a result, the metrics do not provide much insight into the closeness of P and Q.

Besides, the relation between the empirical estimates and the associated theoretical metrics (i.e. at the population level) is not always clear. In particular, oftentimes, experimental behaviors praised by the authors are in fact contradictory with the expected behavior of the theoretical metrics. A typical example concerns experiments where P and Q are both Gaussian distributions but where Q is shifted away from P. In that case, the theoretical metrics remain constant whatever the shift, while the empirical estimates appear to decrease with the magnitude of the shift. The authors embrace this desirable experimental observation without recognizing that it arises from the compensatory interaction of two underlying flaws: on the one hand, the two theoretical metrics are lacking, and on the other hand, their empirical estimators are not accurate. Instead, we advocate for extracting entire PR curves since this gives a complete picture of the disparity between the two distributions, and for conducting a thorough analysis of the consistency of the empirical estimator w.r.t to the population level counterparts. Note that in the literature, consistency is at best studied in the case $P = Q$.

In fact, we make the following contributions. First, we show that most previous fidelity (resp. diversity) metrics can be interpreted as estimators of $\alpha_\infty(P, Q)$ (resp. $\beta_0(P, Q)$) thanks to the binary classification point of view developed in (Simon et al., 2019). As a result, we show that they can be extended into complete curves in quite natural ways. Besides, concurrent approaches differ merely by their underlying hypothesis class (that is the family of classifiers). For instance, the Improved Precision-Recall (IPR) of (Kynkäänniemi et al., 2019) relates to kernel density estimators with adaptive bandwidth, while coverage (Naeem et al., 2020) (and its precision counterpart obtained by swapping the role of P and Q (Khayatkhoei & AbdAlmageed, 2023)) relates to knn classification[1]. Observed under this lens, any idiosyncrasy of the emerging hypothesis class (or its usage) stands out and can be readily amended. In particular, one common pitfall transpires: namely the absence of data split between training samples (used to fit the classifier) and test samples (used to evaluate precision or recall). This introduces a negative bias in the estimators as well as correlations that make the analysis of the estimator consistency unnecessarily challenging. On the contrary, by using split, which is standard practice, the bias of the estimators is trivially positive and consistency can be studied using standard techniques. We actually provide such a result for the curve associated with coverage (which goes beyond the $P = Q$ case considered originally in (Naeem et al., 2020)). We conduct experiments on toy examples similar to those promoted in previous works to re-assess the pros and cons of the PR curve estimator variants (with and without fixes). Our conclusion in this regard is that our extension to coverage has better results than the other extensions. While iPR is sensitive to outliers and is less efficient than our other extensions with the original setting, we provide amendments that correct this behavior, namely setting the value of neighbors k in kNN classifier to grow with the number of data points. With our extended metrics, we also provide some illustrations of the known cases where a generative model either creates, drops, or re-weights modes. Precision and recall curves allow to finely represent those three simultaneous behaviors at once.

2. Recap on the relevant literature

Let Ω a measurable space, and denoting $\mathcal{M}_p(\Omega)$ the set of distributions over Ω, let $P, Q \in \mathcal{M}_p(\Omega)$ (e.g. real and generated distribution).

2.1. The gist on the original PR curve notion

First the PR set between P and Q is defined as the set $\text{PRD}(P, Q)$ of non negative couples (α, β) such that, $\exists \mu \in \mathcal{M}_p(\Omega)$ verifying both $P \geq \beta\mu$ and $Q \geq \alpha\mu$. In a nutshell, the two conditions translate the fact that the ”probe” distribution μ can simultaneously “extract” some mass β from P and α from Q. Note that the PR set is included within $[0, 1]^2$ and it is a cone, meaning that $\forall(\alpha, \beta) \in \text{PRD}(P, Q)$ and $0 \leq \gamma \leq 1$ then $(\gamma\alpha, \gamma\beta) \in \text{PRD}(P, Q)$. As a result, this set is characterized by its (upper-right) Pareto frontier denoted by $\partial\text{PRD}(P, Q)$ which can be parameterized as $\partial\text{PRD}(P, Q) = \{(\alpha_\lambda, \beta_\lambda), \lambda \in \bar{\mathbb{R}}^+\}$ with

$$\begin{aligned} \alpha_\lambda &= (\lambda P \wedge Q)(\omega) \\ \beta_\lambda &= \tfrac{\alpha_\lambda}{\lambda} \end{aligned} \tag{1}$$

[1] On the contrary we will not pursue the same endeavor for (Park & Kim, 2023) since their metric does not bear any reminiscence to classical non-parametric classification literature.

where $\wedge$ is the minimum operator between two measures (see (Simon et al., 2019) for details).

This whole curve captures both extreme precision and recall values corresponding to α_∞ and β_0 which play a central role in the later literature starting from (Kynkäänniemi et al., 2019). In addition, it also describes how similarly the mass is distributed within the common support of P and Q (see (Siry et al., 2023) for details). Interestingly, this curve can be also characterized in a dual way, based on a specific two-sample classification problem (Simon et al., 2019). In short, for a sample $Z = UX + (1-U)Y \sim \frac{1}{2}(P+Q)$ (that is U is coin flip, $X \sim P$ and $Y \sim Q$), the task consists in predicting whether $U = 1$ (i.e. $Z = X \sim P$). Then,

$$\begin{aligned} \alpha_\lambda(P,Q) &= \min_{f\in\mathscr{F}} \{\lambda \cdot \mathrm{fpr}(f) + \mathrm{fnr}(f)\} \\ \beta_\lambda(P,Q) &= \min_{f\in\mathscr{F}} \left\{\mathrm{fpr}(f) + \frac{\mathrm{fnr}(f)}{\lambda}\right\} \end{aligned} \tag{2}$$

where the hypothesis class $\mathscr{F}$ is composed of all binary classifiers on Ω, and $\mathrm{fnr}(f)$ (resp. $\mathrm{fpr}(f)$) represents the false negative (resp. positive) rate of the classifier f, that is to say the probability that a sample $Y \sim Q$ was classified as a sample from P (resp. vice versa). More precisely,

$$\mathrm{fpr}(f) = \int 1 - f dP \text{ and } \mathrm{fnr}(f) = \int f dQ \tag{3}$$

2.2. Re-assessing extreme precision-recall values

In the literature, the accepted expression of the extreme precision is $\alpha_\infty(P,Q) := \lim_{\lambda\to\infty} \alpha_\lambda(P,Q) = Q(\mathrm{supp}(P))$. In fact, this identity is flimsy and requires to be amended mainly because the support of a distribution is defined up to null sets for that distribution. In the incriminated identity, the issue stems from the fact that adding a P-null set can change the Q-mass of the set, and therefore the right-hand side is not well characterized. Let us clarify a few notions.

Definition 2.1 (support and co-support)**.** Let A be a measurable subset of Ω. We say that A is a

- support of P, denoted[2] $A = \mathrm{supp}(P)$, iff $P(A^c) = 0$.
- co-support of P and Q denoted $A = \mathrm{cosupp}(P,Q)$ iff ($(P \wedge Q)(A^c) = 0$ and $\forall B \subset A$, $P(B) = 0 \Leftrightarrow Q(B) = 0$)

As the reader may notice, the second notion is characterized up to sets that are simultaneously P and Q null. More precisely, we have the following result.

Proposition 2.2. *Let P, Q two distributions. Then all co-supports of P and Q have the same Q-mass and*

$$\alpha_\infty(P,Q) = Q(\mathrm{cosupp}(P,Q))$$

Proof. (Sketch) First, if A, A' are two co-supports. Then $Q(A) = Q(A \cap A') = Q(A')$. Indeed, if $Q(A) \geq Q(A \cap A')$ then letting $B = A \setminus A' \subset A$ and $Q(B) > 0$. Yet $B \subset A'^c$ so that $P(B) \leq P(A'^c) = 0$ yielding a contradiction.

Second, let us exhibit a co-support C that verifies $Q(C) = \alpha_\infty$. In (Simon et al., 2019), it is shown that Eq. 2 can be restated as $\alpha_\infty = \min_{A \text{ s.t. } P(A^c)=0} Q(A)$. Let A^* one of the minimizers. Without further care, A^* should be merely a particular support of P but could still not be a co-support. We therefore need to filter out any part of the space that charges P but not Q (which will make it a co-support without affecting its Q-mass). To do so, we consider $C = A^* \setminus \cap_{\lambda>0}\{\lambda P > Q\}$. First, the monotone convergence theorem implies that $Q(C) = Q(A^*) = \alpha_\infty$.

It remains to show that C is indeed a co-support. Notice that $P \wedge Q(C^c) = (P \wedge Q)(A^{*c} \bigcup \cap_{\lambda>0}\{\lambda P > Q\}) \leq P(A^{*c}) + Q(\cap_{\lambda>0}\{\lambda P > Q\}) = 0$ (because of the constraint on A^* for the first summand, and by the monotone convergence theorem again for the other summand).

Besides let $B \subset C \subset (\cap_{\lambda>0}\{\lambda P > Q\})^c = \cup_{\lambda>0}\{\lambda P \leq Q\}$ so that $Q(B) = 0 \implies P(B) = 0$. Conversely, if $P(B) = 0$ let us show that $Q(B) = 0$. To do so let us reason by contradiction, by assuming that $Q(B) > 0$. Then $A = A^* \setminus B$ verifies the constraint $P(A) = 0$ and $Q(A) = Q(A^*) - Q(B)$ (because $B \subset C \subset A^*$). Then $Q(A) < Q(A^*)$ would contradict the definition of A^*. □

This first result will bring some understanding on a key difference between the approach of IPR and the one of coverage: the former being linked to the erroneous formula $Q(\mathrm{supp}(P))$ while the latter relates more to the correct one (as will be seen in the next section). Yet, there are a few caveats that apply even to the correct version of α_∞. First, it is important to realize that this metric is impacted by the tails of P and Q even if they decrease very fast. An illuminating example is the following $P = \mathcal{N}(0,1)$ and $Q = \mathcal{N}(\mu,1)$. Whatever the value of μ (be it extremely large), it remains that P and Q have full co-support and therefore $\alpha_\infty(P,Q) = 1$. The first negative impact of this observation is that α_∞ and β_0 provide a very weak characterization of the relation between P and Q. The second negative impact concerns the estimation of α_∞: namely, the tails of P and Q are elusive based on empirical samples, making this extreme precision the most challenging to evaluate. Both of these observations have gone unnoticed by the previous approaches starting from (Kynkäänniemi et al., 2019) as they purposely focused on the extreme values. In addition to those two arguments, we would like to highlight that estimating the mass of a support is not a standard topic in machine learning. As a result, taking a binary classification standpoint as in (Simon et al., 2019) will bring much more useful hindsight to design estimators correctly.

[2] This is a slight abuse of notations.

2.3. Improved PR metric and follow-up works

In this section, we describe a few published metrics related to extreme precision and recall.[3] Henceforth, we assume that one disposes of a finite set $\mathcal{X}$ of examples from P and others in $\mathcal{Y}$ sampled from Q.

IPR Proposed in (Kynkäänniemi et al., 2019), the Improved Precision Recall metric is given by

$$\hat{\alpha}^{iPR}_{\infty} := \frac{1}{\#\mathcal{Y}} \sum_{y\in\mathcal{Y}} \mathbb{1}_{\exists x\in\mathcal{X}, y\in B^{\mathcal{X}}_{kNN}(x)} \tag{4}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the observed samples from P and Q respectively, and $B^{\mathcal{X}}_{kNN}(x)$ represents the kNN ball around x computed within the set $\mathcal{X}$. This value can be interpreted as the empirical estimate of $Q(\text{supp}(P))$ where samples from $\mathcal{Y}$ are used to estimate the Q probability, and those from $\mathcal{X}$ are used to estimate the support of P as the union of kNN balls.

Coverage First proposed as an estimate of β_0 in (Naeem et al., 2020), it was also adapted to α_∞ in (Khayatkhoei & AbdAlmageed, 2023).

$$\hat{\alpha}^{cov}_{\infty} := \frac{1}{\#\mathcal{Y}} \sum_{y\in\mathcal{Y}} \mathbb{1}_{\exists x\in\mathcal{X}, x\in B^{\mathcal{Y}}_{kNN}(y)} \tag{5}$$

Note that compared to Eq. (4), the condition $y \in B^{\mathcal{X}}_{kNN}(x)$ is merely replaced by $x \in B^{\mathcal{Y}}_{kNN}(y)$. The samples x are naturally in regions of positive P-mass, so that this estimate has an interpretation in terms of $Q(\text{cosupp}(P,Q))$: samples from $\mathcal{Y}$ are again used to estimate empirical probability w.r.t Q but they are also used to estimate the support of Q (again as the union of kNN balls associated to $\mathcal{Y}$) and samples from $\mathcal{X}$ are obviously within the support of P.

EAS (Khayatkhoei & AbdAlmageed, 2023) propose to combine both previous estimates by taking their minimum

$$\hat{\alpha}^{eas}_{\infty} := \min(\hat{\alpha}^{iPR}_{\infty}, \hat{\alpha}^{cov}_{\infty}) \tag{6}$$

PRC Proposed in (Cheema & Urner, 2023) Precision Recall Cover is an extension of coverage:

$$\hat{\alpha}^{PRC}_{\infty} = \frac{1}{\#\mathcal{Y}} \sum_{y\in\mathcal{Y}} \mathbb{1}_{\#\{x\in\mathcal{X}/x\in B^{\mathcal{Y}}_{kNN}(y)\}\geq k'} \tag{7}$$

where $k' \in \mathbb{N}^*$ is an additional hyper-parameter: setting $k' = 1$ makes this estimator identical to cov.

PPR Last, Probabilistic Precision Recall was proposed in (Park & Kim, 2023):

$$\hat{\alpha}^{PPR}_{\infty} := \frac{1}{\#\mathcal{Y}} \sum_{y\in\mathcal{Y}} \left(1 - \prod_{x\in\mathcal{X}} \tau(\|y-x\|)\right) \tag{8}$$

where $\tau(d) = \max(0, 1 - \frac{d}{R})$ is a fixed bandwidth tent kernel.

3. Re-interpretation and improvements

In this section, we start by re-interpreting the iPR and cov estimators in terms of the dual theoretical expression in (2), and then build upon this new insight to both extend the PR estimates as entire trade-off curves, as well as propose new variants. Note that, in order to give a more complete picture of the state-of-the-art, we have presented three alternative metrics, namely EAS, PRC and PPR. Note however, that they all work in a similar fashion to IPR and coverage. Therefore, for the sake of clarity, we will only deal with either iPR or cov, for the remainder of the paper.

3.1. Classification interpretation

One may notice that every estimators mentioned in the previous section reads as

$$\hat{\alpha}^{M}_{\infty} = \widehat{\text{fnr}}(f^{M}_{\infty})$$

where $\widehat{\text{fnr}}$ is the empirical FNR, M is a reference to the metric approach (e.g. iPR) and f^{M}_{∞} is a classifier specific to the approach. In particular, one has $f^{iPR}_{\infty}(z) = \mathbb{1}_{\#\{x\in\mathcal{X}/y\in B_{kNN}(x)\}\geq 1}$ and $f^{cov}_{\infty}(z) = \mathbb{1}_{\#\{x\in\mathcal{X}/x\in B_{kNN}(y)\}\geq 1}$.

In both cases, by design $\widehat{\text{fpr}}(f^{M}_{\infty}) = 0$ because the classifier is equal to 1 on training samples from $\mathcal{X}$. This is reminiscent of the form of α_λ in Eq. 2 when $\lambda \to \infty$:

$$\alpha_\infty = \min_{f\in\mathscr{F} \text{ s.t. } \text{fpr}(f)=0} \text{fnr}(f)$$

In fact, one can analyze each approach M, as a mere empirical version of that equation, under a restricted hypothesis class, namely:

$$\mathscr{F}^{M} = \{f^{M}_{\gamma} / \gamma \in [0, +\infty]\}$$

Note that obviously, the hypothesis class cannot be unequivocally determined by f^{M}_{∞}. Yet, we shall see that natural families emerge for both iPR and cov. Let us describe them now.

IPR In that case, we set

$$f^{iPR}_{\gamma}(z) = \mathbb{1}_{\gamma\#\{x\in\mathcal{X}/z\in B^{\mathcal{X}}_{kNN}(x)\}\geq\#\{y\in\mathcal{Y}/z\in B^{\mathcal{Y}}_{kNN}(y)\}}$$

[3] We will focus on the estimate of $\alpha_\infty(P,Q)$ because swapping the role of P and Q entails the extreme recall β_0.

Note that this classifier is a Kernel Density Estimator (KDE) of the form

$$f_\gamma^{iPR}(z) = \mathbb{1}_{\frac{\hat{p}(z)}{\hat{q}(z)} \geq \frac{1}{\gamma}}$$

where $\hat{p}(z) \propto \sum_{x \in \mathcal{X}} \mathbb{1}_{B^{\mathcal{X}}_{kNN}(x)}(z)$ and similarly for $\hat{q}(z)$. It is therefore a KDE classifier with adaptive bandwidth.

Coverage Here we can set

$$f_\gamma^{cov}(z) = \mathbb{1}_{\gamma \# \{x \in \mathcal{X} / x \in B^{\mathcal{Y}}_{kNN}(z)\} \geq \# \{y \in \mathcal{Y} / y \in B^{\mathcal{X}}_{kNN}(z)\}}$$

This resembles to a classical kNN classifier up to a minor difference: it is more standard to use the same kNN structure for both classes, that is $B^{\mathcal{X} \cup \mathcal{Y}}_{kNN}$ rather than using a separate one per class. Interestingly, one can verify that the condition $\exists x \in \mathcal{X}$ s.t. $x \in B^{\mathcal{Y}}_{kNN}(y)$ is in fact equivalent[4] to $\exists x \in \mathcal{X}$ s.t. $x \in B^{\mathcal{X} \cup \mathcal{Y}}_{kNN}(y)$. Therefore, we may as well choose

$$f_\gamma^{knn}(z) = \mathbb{1}_{\gamma \# \{x \in \mathcal{X} / x \in B^{\mathcal{X} \cup \mathcal{Y}}_{kNN}(z)\} \geq \# \{y \in \mathcal{Y} / y \in B^{\mathcal{X} \cup \mathcal{Y}}_{kNN}(z)\}}$$

to build the hypothesis class of cov.

Note on symmetry For symmetry reasons between precision and recall, we use the previous definitions of f_γ^M for $\gamma \geq 1$ and favor a strict inequality over a loose one for $\gamma < 1$.

3.2. Extension and improvements

At this stage, it is quite easy to extend the extreme precision and recall estimators into entire curves. It suffices to use Empirical Risk Minimizer approach over the hypothesis class $\mathscr{F}^M$ for the risk arising in Eq. 2. That gives us:

$$\hat{\alpha}^M_\lambda = \min_\gamma \lambda \widehat{\text{fpr}}(f_\gamma^M) + \widehat{\text{fnr}}(f_\gamma^M) \tag{9}$$

Splitting The first improvement that calls upon us is to merely split the samples in two: one part used for fitting the classifier $\mathcal{X}^T \cup \mathcal{Y}^T$ and one part used for evaluation $\mathcal{X}^V \cup \mathcal{Y}^V$. In that case, the law of large numbers applies which is crucial for the consistency of $\hat{\alpha}^M_\lambda$.

Note however that because of splitting, it is possible that none of the classifier f_γ^M ensures a null FPR. As a result, it is possible that $\hat{\alpha}^M_\lambda > 1$. As a remedy, we always complement $\mathscr{F}^M$ with the trivial classifiers $f \equiv 1$ and $f \equiv 0$ that predict either P or Q uniformly.

Hyper-parameter k In addition to the introduction of splitting, we consider modifying either the hyper-parameter k for each approach. Concerning k, in the kNN literature (see e.g. (Devroye et al., 2013)), it is known that as the number of samples n gets bigger, k can also increase but at smaller rate (this will be a key element to ensure the consistency of the kNN estimator in Thm 3.1). We therefore consider for each approach, setting $k = \sqrt{n}$ in place of $k = 4$ similarly to previous works.

Bandwidth Besides, considering iPR, we have seen that it corresponds to an adaptive bandwidth Kernel Density Estimator (with a constant kernel). This design choice results in having bigger bandwidth around samples located at low density regions. It is responsible for the high sensibility to outliers that was pointed out in several followup works. We therefore consider as a simple alternative, a fixed bandwidth KDE that we will refer to as a Parzen classifier. In that case,

$$f_\gamma^{parzen}(z) = \mathbb{1}_{\frac{\hat{p}(z)}{\hat{q}(z)} \geq \frac{1}{\gamma}}$$

with $\hat{p}(z) \propto \sum_{x \in \mathcal{X}} \mathbb{1}_{\|x - z\| \leq \rho_{\mathcal{X}})}$ and similarly for $\hat{q}$. In comparison to iPR the bandwidth $\rho_{\mathcal{X}}$ is computed as the average knn radius[5] over the dataset $\mathcal{X}$ (and similarly for $\rho_{\mathcal{Y}}$) instead of using a specific bandwidth per sample.

3.3. Consistency analysis

The iPR approach is known to be biased in general, since when $P = Q$, it does lead to an estimate of α_∞ that can be far from the true value of 1. Indeed in such case, even when considering $n = \#\mathcal{X} = \#\mathcal{Y} \to \infty$, $\lim_{n \to \infty} \mathbb{E}[\hat{\alpha}^{iPR}_\infty]$ can be much smaller than 1 (see the Gaussian case in (Naeem et al., 2020)). On the contrary, Naeem et al. (2020) shows that when $P = Q$, coverage is consistent. By symmetry so is $\hat{\alpha}^{cov}_\infty$.

We extend the above-mentioned consistency result to the entire PR curve associated to our kNN approach and in the general case of two distributions P and Q.

Theorem 3.1. *Let $\lambda \in \bar{\mathbb{R}}^+$, $k \geq 3$ and $n = \#\mathcal{X} = \#\mathcal{Y}$. Letting $k \to \infty$ and $\frac{k}{n} \to 0$, and denoting*

$$\Gamma^*_\lambda = \arg\min_\gamma \lim_{k \to \infty, \frac{k}{n} \to 0} \mathbb{E}[\lambda \text{fpr}(f_\lambda^{kNN}) + \text{fnr}(f_\lambda^{kNN})]$$

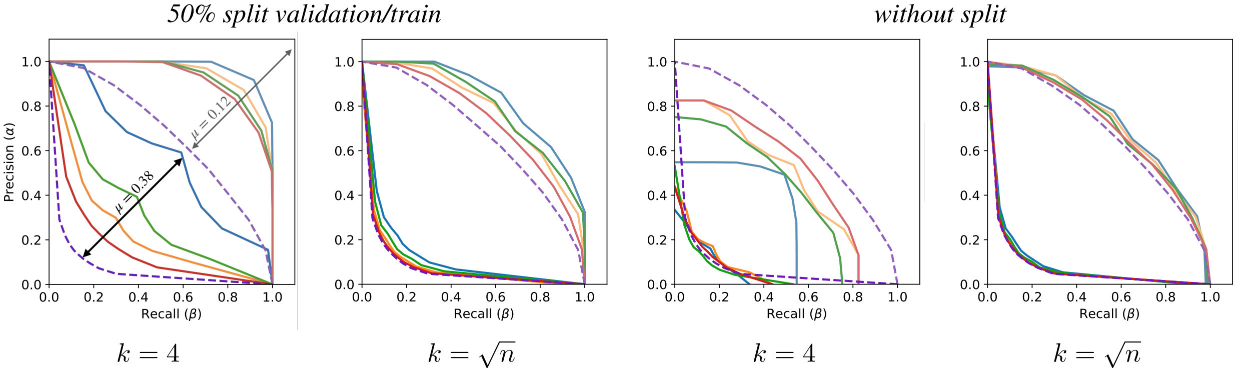
Then

1. $\lambda \in \Gamma^*_\lambda$
2. $\mathbb{E}[\hat{\alpha}^{kNN}_\lambda] \to \alpha_\lambda$ *assuming that data split was used.*

Proof (sketch). The proof is provided in Appendix A and is similar to the standard Bayes consistency results of the

[4] One direction is trivial since $B^{\mathcal{X} \cup \mathcal{Y}}_{kNN}(y) \subset B^{\mathcal{Y}}_{kNN}(y)$, the other requires a bit more reasoning: assuming $\exists x \in B^{\mathcal{Y}}_{kNN}(y)$ one may consider in particular the x closest to y and conclude that it belongs to $B^{\mathcal{X} \cup \mathcal{Y}}_{kNN}(y)$.

[5] This choice for the bandwidth is inspired by (Park & Kim, 2023) although the Parzen variant differs from their PPR estimator which does not resemble any standard classifier.



$k = 4$ $k = \sqrt{n}$ $k = 4$ $k = \sqrt{n}$

Figure 2: **Comparing two shifted Gaussians.** The Ground-Truth PR curve (**- -GT**) is compared to empirical estimates from various NN-classifiers: **–IPR**, **–KNN**, **–PARZEN**, and **–COVERAGE**. Here $P \sim \mathcal{N}(0, \mathbb{I}_d)$ and $Q \sim \mathcal{N}(\mu \mathbf{1}_d, \mathbb{I}_d)$ with $d = 64$ dimensions and $\mu = \frac{1}{\sqrt{d}} \approx .12$ or $\mu = \frac{3}{\sqrt{d}} \approx .38$. $n = 10$K points are sampled using $k = 4$ or $k = \sqrt{n}$ for NN comparison, with or without dataset validation/train split. (Curves are averaged over 10 random samples, see Appendix).

kNN classifier (see e.g. (Devroye et al., 2013)[chap 5& 6]). It is merely adapted to the fact that the risk is class weighted i.e. $R_\lambda(f) = \lambda \mathrm{fpr}(f) + \mathrm{fnr}(f)$ instead of the classical one $R(f) = \frac{1}{2}(\mathrm{fpr}(f) + \mathrm{fnr}(f))$. □

4. Experiments

In this section, we reissue a few experiments on toy datasets that were proposed in the literature. In all settings, for the different estimators under scrutiny, we use $n = 10$K samples. In each experiment, the distributions P and Q are known analytically and the ground-truth PR curve can be estimated easily because the Bayes classifier is the likelihood ratio classifier (Simon et al., 2019) $f^*_\lambda(z) = \mathbb{1}_{\frac{dQ}{dP}(z) \le \lambda}$. To obtain high accuracy ground-truth curves we resort to a large sample ($n^{GT} = 100$K) and estimate

$$\hat{\alpha}^{GT}_\lambda = \lambda \widehat{\mathrm{fpr}}(f^*_\lambda) + \widehat{\mathrm{fnr}}(f^*_\lambda)$$

Based on this PR curve, we can either evaluate the quality of an estimator visually, or use a scalar indicator to summarize the quality of the estimator. In particular, we propose to use the Jaccard index (a.k.a IoU) between the ground-truth curve and the estimator under review. This index is always smaller than 1 and the larger the better.

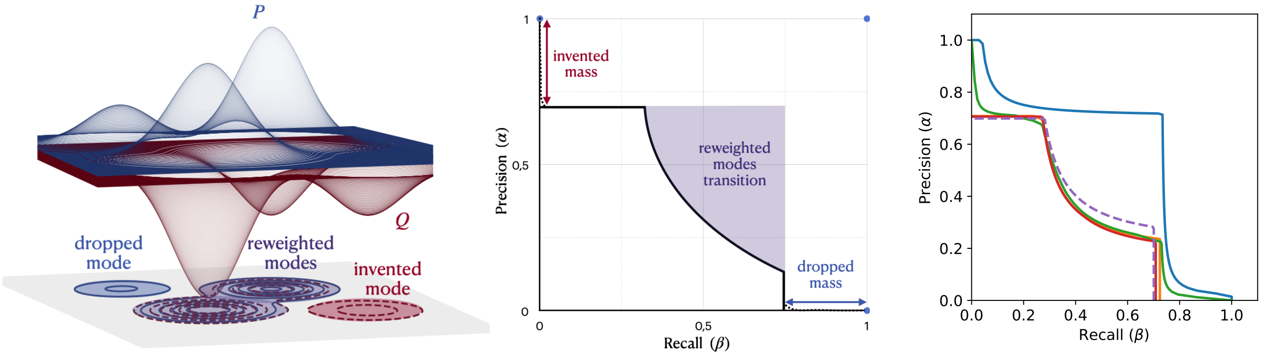
4.1. Gaussian shifts

Inspired by recent experiments made in the literature, we consider the case where P and Q are two Gaussian distributions with an increasing shift. In order to match the published experiments, we generate 10K data points from each Gaussian distribution, in $\mathbb{R}^{64}$. Even though the number of sampled points is quite high, we decide to run the experiments 100 times to have robust interpretations of the metrics. What interests us here is the behavior of the of the estimated curves in comparison to the ground-truth. In particular, unlike (Kynkäänniemi et al., 2019; Naeem et al., 2020; Park & Kim, 2023) we reaffirm that the extreme values of the curves should be equal to 1 as the supports of the two Gaussians are always the same.

We run the experiments on 5 different methods, using 4 different shifts for the fake Gaussian. We present the resulting curves for only two shifts in Fig. 2 and condense the complete experiment results in Appendix in Table 1. In these settings, the standard deviation for each experiment is lower than 10^{-2}. Whereas the maximum values for the PR curves are always at 1, the curves themselves show the lack of information to compare the two distributions. While methods from the literature, the number of nearest neighbors k used in the manifold estimations is set to 3 ((Kynkäänniemi et al., 2019)) and to 5 ((Naeem et al., 2020)), we first set $k = 4$ then, motivated by 3.1, we set $k = \sqrt{n}$. We observe that the yielded results are more convincing with $k = \sqrt{n}$, and that the difference between the split and no split scenarios is marginal. As we have theoretical guarantees in the case with split, we decided to keep the combination split and $k = \sqrt{n}$ for the following experiments (complementary results are presented in the appendix).

4.2. Gaussian mixture models

In (Luzi et al., 2023) the authors advocate that Inception features are better approximated by Gaussian Mixtures than pure Gaussians. Besides this scenario allows to illustrate the phenomenon of mode dropping (mode present in P but not in Q), mode creation (mode present in Q but not in P), and mode reweighing (shared modes between P and Q weighted differently) as illustrated in Fig. 3a. In our toy experiment, we sample points from two GMM in dimension $d = 64$, $n = 1$K samples, splitting is applied, and $k = \sqrt{n}$



(a) Left: Gaussian Mixture Models P and Q showing mode dropping (only in P), mode inventing (only Q), and mode re-weighting (in both but distributed differently). Right: Expected coarse shape of the PR-curve (solid black). Note that due to the infinite tails of the Gaussian modes, the vertical and horizontal transitions are theoretically smooth and reach 1 (dashed curve).

(b) **Comparing two Gaussian mixtures**. 3a. The Ground-Truth PR curve (- -**GT**) is compared to empirical estimates from various NN-classifiers: **–iPR**, **–kNN**, **–Parzen**, and **–Coverage**.

(see Fig. 3b). The two GMM P and Q are set as follows: $P \sim \sum_\ell p_\ell \mathcal{N}(\mu_\ell \mathbf{1}_d, \mathbb{I}_d)$ and $Q \sim \sum_\ell q_\ell \mathcal{N}(\mu_k \mathbf{1}_d, \mathbb{I}_d)$ with $d = 64$ dimensions and $\mu_\ell \in \{0, -5, 3, 5\}$. However, P and Q have different weights (p_ℓ and q_ℓ) $p_\ell \in \{0.3, 0.2, 0.5, 0\}$ $q_\ell \in \{0, 0.5, 0.2, 0.3\}$. kNN, parzen and coverage perform very well with respect to the ground-truth, while IPR overestimates the PR-curve and especially fails at catching the re-weighting transition.

4.3. Outliers

One of the main contributions of (Naeem et al., 2020) over (Kynkäänniemi et al., 2019) was its robustness to outliers. We here investigate how outliers affect our PRD curves. To do so, we simply once again define the P and Q distributions as two shifted Gaussians. Similarly to experiments in literature, we then add a single outlier x_{outlier} to the real distribution P such that $x_{\text{outlier}} = \mathbf{4}$. All the cases are considered in appendix, yet the observations are simple: as already reported in the aforementioned works, we observe that iPR classifier is indeed affected by such a perturbation for low k-NN comparison ($k = 4$) and without split. This sensitivity is especially strong near the extreme values. We also notice that this effect is easily mitigated when using a larger set $k = \sqrt{n}$. On the other hand, other PRD curves based on more robust classifiers (Parzen, Coverage and kNN) are not affected by the outlier as expected.

4.4. Impact of the ambient dimension

It is customary in the evaluation of generative models to first feed both target and generated data through a deep classifier feature space. In the protocol for computing FID, the authors use InceptionV3 where the feature space has a dimension of 2048, and use 10K samples for both real and generated data. We now test our metrics in the same Gaussian shifting setting, with samples in $\mathbb{R}^{2048}$. The experiment result is shown in Fig. 4. While we used $k = \sqrt{n}$, the results show a much larger overestimation than in the $\mathbb{R}^{64}$ setting. This illustrates the well-known curse of dimensionality.

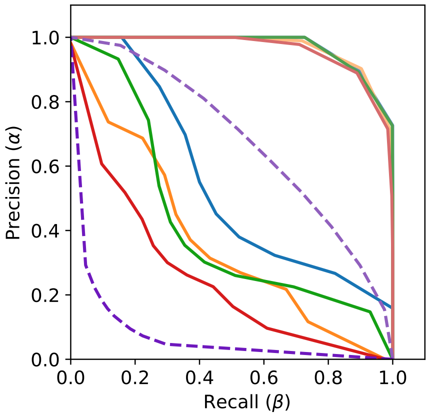


Figure 4: **PR curves in high dimension** Same experiment as in Fig. 2 (50% split, $n = 10$K, $k = \sqrt{n})$ with $d = 2048$ dimensions and $\mu \in \{\frac{1}{\sqrt{d}}, \frac{2}{\sqrt{d}}\}$.

4.5. Variability study

Additional experiments in appendix (Fig. 8) scrutinize the impact of the size sample n on the variability of the evaluation curves. They empirically illustrate the consistency of the proposed method based on robust classifiers when increasing (Thm 3.1). Empirically, using $10K$ points reduces sufficiently the variability to make comparison between curves reliable, which is in line with the standard usage for generative model evaluation (Heusel et al., 2017; Sajjadi et al., 2018).

5. Discussion and perspectives

5.1. Distilling the curve with two scalar metrics

Practitioners may enjoy summarizing the PR curve with two metrics reflecting respectively precision and recall. In other words, one would be willing to trade exhaustiveness for conciseness. This may be particularly useful to ease comparison between models. In that prospect, we have seen that extreme precision and recall are yet far from ideal. We therefore discuss two alternatives and comment on them in a scenario combining mode dropping/invention/re-weighting (e.g. Fig. 3a).

F-scores Proposed by (Sajjadi et al., 2018), the F_b score is defined as:

$$F_b = \max_{\lambda \in [0,+\infty]} \frac{1+b^2}{\frac{b^2}{\alpha_\lambda} + \frac{1}{\beta_\lambda}}.$$

When $b \to \infty$, $F_b \nearrow \alpha_\infty$ and respectively when $b \to 0$ $F_b \searrow \beta_0$, so that (Sajjadi et al., 2018) proposed to consider F_b and $F_{1/b}$ with a large value of b (namely $b = 8$). Although this aspect was not discussed in their original work, one can understand that these metrics will not be sensitive to rapidly decaying infinite tails. However, observe that the score F_b is a weighted harmonic mean that is computed from a single (optimal) point of the PR curve, which can vary dramatically without affecting the metric. As a consequence, when b is large, F_b and $F_{1/b}$ will mainly capture pure mode dropping/invention and remain superficially indicative of mode re-weighting.

PR median As another alternative, one may consider $(\alpha_{\bar{\lambda}}, \beta_{\bar{\lambda}})$ where $\bar{\lambda}$ is set so that the line $\alpha = \bar{\lambda}\beta$ cuts the region under the PR curve into two sub-parts of equal areas. In the case of pure mode dropping and mode inventions, $\alpha_{\bar{\lambda}} = \alpha_\infty$ and $\beta_{\bar{\lambda}} = \beta_0$. On the contrary, when infinite tails create quick transitions to $\alpha_\infty = 1$ like in Fig. 3a, then (alike F_b) $\alpha_{\bar{\lambda}} < 1$ will be but mildly impacted by rapid-decay tails. As opposed to F_b, the value of $\alpha_{\bar{\lambda}}$ will be largely affected by the presence of a transition due to mode re-weighting (violet transition in Fig. 3a). As a result, these metrics may be preferred when one would like to account for mode re-weighting in addition to mode dropping/invention.

An empirical assessment on the pros and cons of both alternatives would be a valuable endeavor. In particular, one could study the behavior of the said metrics with respect to hyperparameters of state-of-the-art generative models. One can for instance consider, the truncation procedure for GANs, or the guidance scale factor for diffusion models. This empirical study is left as future work.

5.2. More in-depth convergence analysis

In Theorem 3.1, we have demonstrated that the kNN estimator is universally consistent. The proof is adapted from standard Bayes consistency results on kNN classifiers. Characterizing the quality of other estimators based on standard classification schemes (e.g. Kernel Discriminant Analysis) could also be considered. Besides, characterizing, rates of convergence under regularity assumptions for P and Q is also an appealing avenue. This is left as future work, but we refer the reader to (Devroye et al., 2013) and to (Györfi & Weiss, 2021) for a recent overviews of useful results. In a more practical perspective, one may wish to choose optimally the hyper-parameters of the different estimators. In our work we have considered the following heuristics $k = \sqrt{n}$ for the kNN estimator and a fixed bandwidth $\rho_\mathcal{X}$ computed based on the average kNN radius for the Parzen classifier. A large literature does exist around those topics, see for example (Ghosh & Chaudhuri, 2004) and (Döring et al., 2018) for kNN and (Ghosh, 2006) for optimal bandwidth. It is also possible to resort to cross-validation for setting these hyper-parameters.

6. Conclusion

In this work, we have given a new perspective on recent metrics used to evaluate extreme precision and recall of generative models. Doing so we have obtained two by-products. First, we have presented a systematic way to extend the extreme values to obtain complete PR curves. Second, we built upon standard literature in non-parametric classification to improve the original approaches. In particular, we have provided a consistency result for the kNN PR-curve variant as well as several practical improvements over the original iPR and coverage metrics. We have also studied the empirical behavior of the obtained variants in the light of several toy datasets experiments.

Our main messages are the following. First, computing non extreme PR values is crucial because of essential issues in the extreme values which are related to their sensitivity to the distribution tails. Then, the curves themselves allow to describe more finely how the masses of the two distributions under comparison differ on their modes. This is useful in practice in order to tackle the case where a model generates data from the target support but with re-weighted masses. On the experimental side, it emerges that coverage is indeed better suited than iPR. Both approaches can be improved by adapting the number of neighbors k with respect to the number of available samples. If employing a data split is theoretically appealing, its empirical impact is less marked since the negative bias resulting from the lack of split can sometimes advantageously compensate the positive bias caused by the restricted hypothesis class. However, this benefit is not consistent over all experiments.

References

Cheema, F. and Urner, R. Precision recall cover: A method for assessing generative models. In *International Conference on Artificial Intelligence and Statistics*, pp. 6571–6594. PMLR, 2023.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Djolonga, J., Lucic, M., Cuturi, M., Bachem, O., Bousquet, O., and Gelly, S. Precision-recall curves using information divergence frontiers. In *International Conference on Artificial Intelligence and Statistics*, pp. 2550–2559. PMLR, 2020.

Döring, M., Györfi, L., and Walk, H. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18(227):1–16, 2018.

Ghosh, A. K. On optimum choice of k in nearest neighbor classification. *Computational Statistics & Data Analysis*, 50(11):3113–3123, 2006.

Ghosh, A. K. and Chaudhuri, P. Optimal smoothing in kernel discriminant analysis. *Statistica Sinica*, pp. 457–483, 2004.

Györfi, L. and Weiss, R. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *The Journal of Machine Learning Research*, 22 (1):6702–6726, 2021.

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.

Khayatkhoei, M. and AbdAlmageed, W. Emergent asymmetry of precision and recall for measuring fidelity and diversity of generative models in high dimensions. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

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.

Luzi, L., Marrero, C. O., Wynar, N., Baraniuk, R. G., and Henry, M. J. Evaluating generative networks using gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 279–288, 2023.

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.

Park, D. and Kim, S. Probabilistic precision and recall towards reliable evaluation of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20099–20109, 2023.

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.

Simon, L., Webster, R., and Rabin, J. Revisiting precision recall definition for generative modeling. In *International Conference on Machine Learning*, pp. 5799–5808. PMLR, 2019.

Siry, R., Webster, R., Simon, L., and Rabin, J. On the theoretical equivalence of several trade-off curves assessing statistical proximity. *Journal of Machine Learning Research*, 24(185):1–34, 2023. URL http://jmlr.org/papers/v24/21-0607.html.

A. Proof of Theorem 3.1: consistency

Proof. To establish the proof, we need only show that $R_\lambda(f_\lambda^{kNN}) \to \alpha_\lambda$ as $k \to \infty$ and $\frac{k}{n} \to 0$. This will effectively imply both items in the theorem since α_λ is the associated Bayes risk (Simon et al., 2019). To establish this limit, the first step is to show that for fixed k then $\lim_{n\to\infty} R_\lambda(f_\gamma^{kNN})$ is equal to

$$\begin{aligned} 2\lambda\mathbb{E}[\eta(Z)\mathbb{P}\{Binom(k,\eta(Z)) < \tfrac{k}{\gamma+1}|Z\}]+ \\ 2\mathbb{E}[(1-\eta(Z))\mathbb{P}\{Binom(k,\eta(Z)) > \tfrac{k}{\gamma+1}|Z\}] \end{aligned} \tag{10}$$

where $Z = UX + (1-U)Y$ with $X \sim P$, $Y \sim Q$ and U is a fair coin random variable so that $Z \sim \frac{P+Q}{2}$ and $\eta(Z) := \mathbb{P}(U=1|Z) = \frac{dP}{d(P+Q)}(Z)$ (respectively $1 - \eta(Z) := \mathbb{P}(U=0|Z) = \frac{dQ}{d(P+Q)}(Z)$). The demonstration of Eq. (10) follows the same argument as in (Devroye et al., 2013)[Thm. 5.2] (up to the occurrence of λ and γ weights) and is not repeated here for the sake of conciseness.

Now, taking $\gamma = \lambda$ we want to show that the previous expression tends to α_λ which for the recall equals $(\lambda P \wedge Q)(\Omega)$ or expressed otherwise as $2\mathbb{E}[\lambda\eta(Z) \wedge (1-\eta(Z))]$.

Eq. (10) can be reformulated as $\lim_{n\to\infty} R_\lambda(f_\lambda^{kNN}) = 2\mathbb{E}[\mu_\lambda(\eta(Z))]$ with

$$\begin{aligned} \mu_\lambda(p) =& \lambda p\mathbb{P}\{Binom(k,p) < \tfrac{k}{\lambda+1}\} \\ &+ (1-p)\mathbb{P}\{Binom(k,p > \tfrac{k}{\lambda+1}\} \end{aligned} \tag{11}$$

So that it suffices to show that $\forall p \in [0,1]$, $\mu_\lambda(p) \to \lambda p \wedge (1-p)$.

Let's proceed by cases, starting by considering $\lambda p < (1-p)$ which is also equivalent to $p < \frac{1}{\lambda+1}$. In that case we need to show that $2\mu_\lambda(p) \to \lambda p$. Denoting $q_\lambda(p) = \mathbb{P}\{Binom(k,p) > \frac{k}{\lambda+1}\}$, we have

$$\begin{aligned} \mu_\lambda(p) =& \lambda p(1-q_\lambda(p)) + (1-p)q_\lambda(p) \\ =& \lambda p + q_\lambda(p)(1-(\lambda+1)p) \end{aligned} \tag{12}$$

Using Hoeffding's inequality (i.e. $\forall t > 0, \mathbb{P}\{Binom(k,p) - kp > t\} \leq \exp(-2kt^2)$ and obtain

$$\begin{aligned} q_\lambda(p) =& \mathbb{P}\{Binom(k,p) - kp > k(\tfrac{1}{\lambda+1} - p)\} \\ \leq& \exp\left(-2k(\tfrac{1}{\lambda+1} - p)^2\right) \end{aligned} \tag{13}$$

Note that the assumption $p < \frac{1}{\lambda+1}$ is crucial to apply Hoeffding's inequality (because t needs to be positive). The right hand side converges to 0 as $k \to \infty$ because by assumption $p \neq \frac{1}{\lambda+1}$.

The case where $\lambda p > (1-p)$ (or $p > \frac{1}{\lambda+1}$) is similar and is left to the reader. In that case, we obtain $\mu_\lambda(p) \to (1-p)$. There remains the case of equality, that is $\lambda p = 1-p = \frac{1}{\lambda+1}$. In that case, even without taking the limit, one can check that $\mu_\lambda(p) = \lambda p$, which concludes the proof. □

B. Additional experimental results

Computation of PR curves The computation PR curves involves the computation of false positive rate (fpr) and false negative rate (fnr) for various classifiers f_γ^M parameterized by γ. In experiments, we consider random samples $\mathcal{X}$ and $\mathcal{Y}$ from distributions P and Q with $n = |\mathcal{X}| = |\mathcal{Y}|$.

$$\widehat{\text{fnr}}(f_\gamma) = \frac{1}{n}\sum_{y\in\mathcal{Y}} f_\gamma(y)$$

and

$$\widehat{\text{fpr}}(f_\gamma) = \frac{1}{n}\sum_{x\in\mathcal{X}} 1 - f_\gamma(x)$$

Precision $\hat{\alpha}_\lambda$, and recall $\hat{\beta}_\lambda = \frac{1}{\lambda}\alpha_\lambda$ are computed using Eq. (9), where parameters $\lambda = \text{atan}(\theta)$ and γ are both uniformly sampled with $\theta \in (0, \frac{\pi}{2})$.

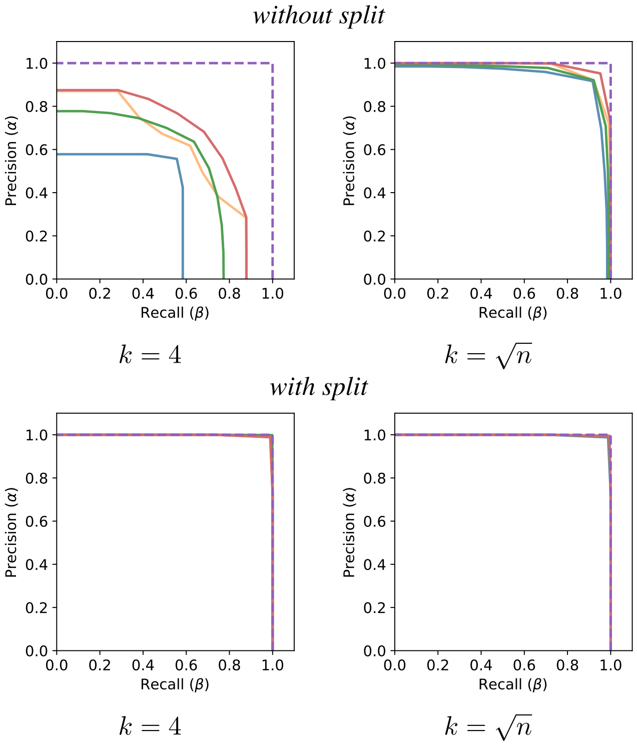


Figure 5: **Illustration of the impact of splitting for $P = Q$.** The setting is the same as Fig. 2 for a translation of $\mu = 0$ between two Gaussian in dimension $d = 64$ (curves are averaged over 3 random samples). The Ground-Truth PR curve (- -GT) is compared to empirical estimates from various NN-classifiers: –IPR, –KNN, –PARZEN, and –COVERAGE. Top reports results without splitting : as reported in the literature, estimated extremal precision and recall values are not equal to 1, contrary to the ground-truth. Bottom curves, obtained with a 50% splits, are very close to the ideal curve.

Splitting All experiments involving data splitting into training and validation sets ($\mathcal{X}^T \& \mathcal{Y}^T$ and $\mathcal{X}^V \& \mathcal{Y}^V$) are 50%:

curves computed in this setting therefore rely on $\frac{n}{2}$ data points.

Figure 5 provides a visual illustration of the practical impact of splitting the dataset into separate training and validation sets to assess precision-recall curves. We consider same experimental setting as in section 4.1 and Figure 2 with $\mu = 0$, in such a way that $P = Q$. In this specific case, the precision is equal to 1 for all recall.

Gaussians shifts Table. 1 complements section 4.1 and Figure 2 by comparing various estimated PR curves with respect to the ground-truth, using average IoU scores.

		shift μ	IPR	KNN	PARZEN	COVERAGE
with 50 % split	$k = 4$	0.12	0.69	0.71	0.72	0.73
		0.21	0.42	0.49	0.49	0.55
		0.29	0.24	0.38	0.34	0.48
		0.38	0.13	0.33	0.24	0.48
	$k = \sqrt{n}$	0.12	0.81	0.87	0.84	0.92
		0.21	0.69	0.84	0.78	0.90
		0.29	0.65	0.84	0.75	0.90
		0.38	0.63	0.84	0.75	0.93
without split	$k = 4$	0.12	0.43	0.7	0.62	0.76
		0.21	0.55	0.81	0.68	0.84
		0.29	0.62	0.79	0.68	0.77
		0.38	0.55	0.61	0.62	0.63
	$k = \sqrt{n}$	0.12	0.91	0.93	0.94	0.96
		0.21	0.88	0.93	0.92	0.97
		0.29	0.84	0.92	0.90	0.95
		0.38	0.83	0.91	0.90	0.96

Table 1: **Mean IoU scores** for shifting Gaussians. Standard deviations are $< 10^{-2}$ with $n = 10$K.

Gaussian Mixture comparison Fig. 6 complements section 4.2 and Figure 3b with additional curves for different setting (w/ and w/o splitting, $k = 4$ or $k = 100$).

Outlier Fig. 7 complements section 4.3. It shows the impact of having a single outlier (a data-point out of $n = 10$K) in one of the samples. Here only the sample $\mathcal{X}$ from P is polluted, thus mainly affecting precision.

Variability Fig. 8 complements Section 4.5 about variability. Average curves are obtained by computing the empirical mean of $N = 100$ PR curves obtained different random n-samples (with $n = 10^4$), *i.e.*

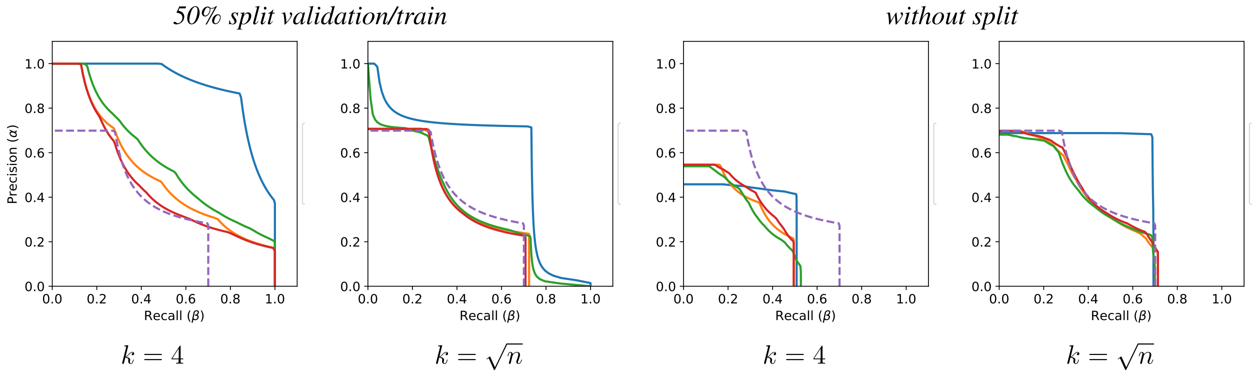
$$(\bar{\alpha}(\lambda), \bar{\beta}(\lambda)) = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i(\lambda), \beta_i(\lambda))$$

Deviation from average curves are materialized with two curves

$$(\alpha_\delta(\lambda), \beta_\delta(\lambda)) = (\bar{\alpha}(\lambda), \bar{\beta}(\lambda)) + \delta(\lambda)$$

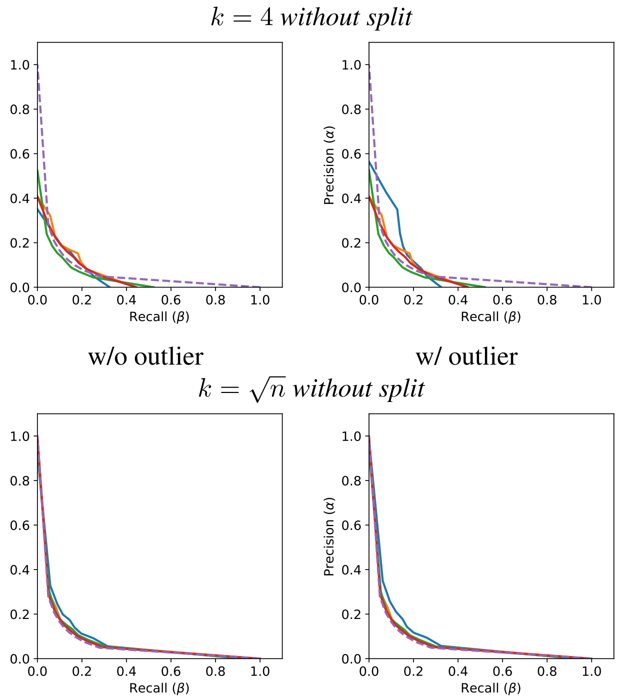
for $\delta = \pm\sigma$ with empirical estimator

$$\sigma^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left((\alpha_i(\lambda) - \bar{\alpha}(\lambda))^2, (\beta_i(\lambda) - \bar{\beta}(\lambda))^2\right).$$



$k = 4$ $k = \sqrt{n}$ $k = 4$ $k = \sqrt{n}$

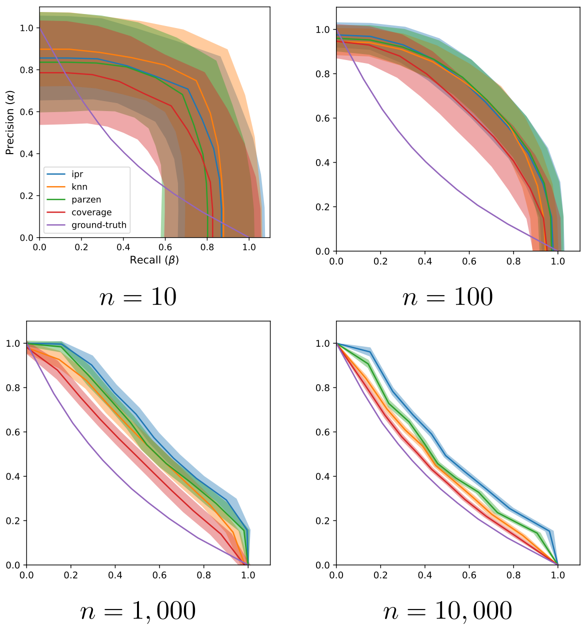
Figure 6: **Comparing two Gaussian mixtures**. This figure complements Fig. 3b. The Ground-Truth PR curve (- -GT) is compared to empirical estimates from various NN-classifiers: –iPR, –kNN, –Parzen, and –Coverage. Here P and Q are two GMMs sharing the same modes (centered at μ_k): $P \sim \sum_\ell p_\ell \mathcal{N}(\mu_\ell \mathbf{1}_d, \mathbb{I}_d)$ and $Q \sim\sim \sum_\ell q_\ell \mathcal{N}(\mu_k \mathbf{1}_d, \mathbb{I}_d)$ with $d = 64$ dimensions and $\mu_\ell \in \{0, -5, 3, 5\}$. However, P and Q have different weights (p_ℓ and q_ℓ) $p_\ell \in \{0.3, 0.2, 0.5, 0\}$ $q_\ell \in \{0, 0.5, 0.2, 0.3\}$. $n =$ 1k points are sampled and split in half between validation and train, and $k = \sqrt{n}$.



w/o outlier w/ outlier

w/o outlier w/ outlier

Figure 7: **Measuring the impact of an outlier.** On the left, the setting is the same as Fig. 2 for a translation of $\mu = 3/\sqrt{d}$ between two Gaussian in dimension $d = 64$ (without splitting nor averaging). On the right, a single outlier ($x = \mathbf{4}$) is added to the sample of P. As reported in the literature, this affects the iPR classifier, yet the PR curves are barely affected by such a perturbation.



$n = 10$ $n = 100$

$n = 1,000$ $n = 10,000$

Figure 8: **Influence of sample size n.** The setting is the same as Fig. 2 for a translation of $\mu = .21$ between two Gaussian in dimension $d = 64$ (with splitting and $k = \sqrt{n}$). Solid (respectively transparent) curves correspond to the empirical average (resp. deviations) of 100 PR curves computed from random samples. (see the text for more details).