Journal of Machine Learning Research 24 (2023) 1-34 Submitted 6/21; Revised 4/23; Published 6/23

On the Theoretical Equivalence of Several Trade-Off Curves Assessing Statistical Proximity

Rodrigue Siry RODRIGUE.SIRY@SAFRANGROUP.COM
Safran Electronics and Defense

Ryan Webster RYAN.WEBSTER@UNICAEN.FR
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 CAEN, FRANCE

Loic Simon✉ LOIC.SIMON@ENSICAEN.FR
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 CAEN, FRANCE

Julien Rabin JULIEN.RABIN@UNICAEN.FR
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 CAEN, FRANCE

Editor: Sivan Sabato

Abstract

The recent advent of powerful generative models has triggered the renewed development of quantitative measures to assess the proximity of two probability distributions. As the scalar Frechet Inception Distance remains popular, several methods have explored computing entire curves, which reveal the trade-off between the fidelity and variability of the first distribution with respect to the second one. Several of such variants have been proposed independently and while intuitively similar, their relationship has not yet been made explicit. In an effort to make the emerging picture of generative evaluation more clear, we propose a unification of four curves known respectively as: the Precision-Recall (PR) curve, the Lorenz curve, the Receiver Operating Characteristic (ROC) curve and a special case of Rényi divergence frontiers. In addition, we discuss possible links between PR / Lorenz curves with the derivation of domain adaptation bounds.

Keywords: trade-off curve, distributional closeness, generative modeling, domain adaptation

1. Introduction

Assessing the proximity of two probability distributions is a long standing concern in statistics. It has gained a new impetus with the advent of deep learning techniques to generate random samples from complex data distributions such as natural images. Generative models, particularly Generative Adversarial Networks (GANs), can now synthesize images with unprecedented realism. Indeed, the quality of generation has improved significantly[1] to the point where for certain datasets, human observers have difficulty discerning real and fake (Brock et al., 2018; Karras et al., 2019). As these networks see real applications, evaluation of generative networks has become essential and remains challenging. Additionally, such performance raises suspicion about memorizing or overfitting some training images. The amount of training data makes visual comparative evaluation not reliable enough. For

1. see *e.g.* https://thispersondoesnotexist.com

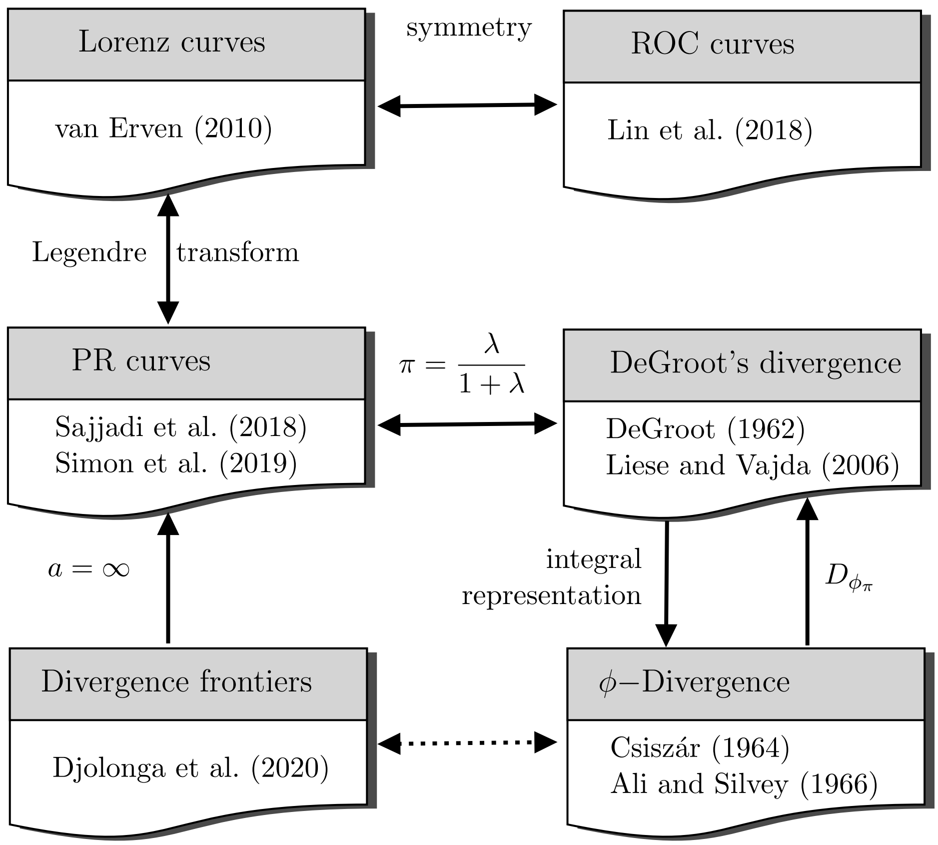

Figure 1: Overview of different works and their relationship. Note that PR and ROC curves refer here to their definition with respect to distributions and not to classifiers.

instance, the privacy and security of generative models has become paramount, including the largest ever Kaggle competition[2] to address deep fakes, or several new works addressing how generative models leak training data (Wang et al., 2019; Webster et al., 2019). Even properly determining sample quality remains challenging (Borji, 2019). The Fréchet Inception Distance (FID) (Heusel et al., 2017) was shown to correlate decently with human evaluation and remains the most popular evaluation metric, but as a scalar metric is limited when assessing model failure (Sajjadi et al., 2018). A variety of other approaches attempt to give an empirical estimation of sample quality, for instance in (Im et al., 2018), the original GAN training divergence was re-used for evaluation. Sajjadi et al. (2018) proposed computing an entire Precision-Recall curve (PR) for the generated distribution. Unlike the scalar FID, this curve distinguishes what we will refer to as the *fidelity* and *variability* of the model (Naeem et al., 2020). Fidelity evaluates whether the generated distribution produces data that are faithful to the original distribution whereas variability reflects the fact that it covers the entire distribution with the correct importance. For instance, a generator of

2. https://www.kaggle.com/c/deepfake-detection-challenge

facial images with poor diversity may only generate one gender, whereas a generator with poor fidelity contains generation artefacts.

Even if the question of estimating the similarity of two distributions has already been intensely studied in the past, it has drawn a renewed interest in recent years. Many probability distribution metrics can be used such as divergences (Rubenstein et al., 2019) or integral probability metrics (Müller, 1997; Sriperumbudur et al., 2009) to name a few. In this work, we will take special focus on the recent work of Sajjadi et al. (2018) computing a Precision-Recall curve between two distributions[3]. Several works explicitly build upon their definition and propose some extensions. For instance, Simon et al. (2019) generalizes the PR curve to arbitrary probabilities (while the work of Sajjadi et al. (2018) was restricted to a discrete settings). More practical works aim at improving empirical evaluation of fidelity and variability (Kynkäänniemi et al., 2019; Naeem et al., 2020). Djolonga et al. (2019) proposed the (Rényi) *divergence frontiers*; this alternate curve coincides with the original PR curve for discrete distributions when the Rényi exponent is infinite. Independently, a handful of alternative curves were defined to compare two distributions. For instance, *ROC curves* were proposed by Lin et al. (2018, 2017) and the *Lorenz curves* by Harremoës (2004); van Erven and Harremoës (2010).

In this work, we demonstrate that, despite their apparently independent definitions, these alternate notions are actually tightly linked with the PR curves which themselves are in fact nearly identical to the notion of DeGroot statistical information. Our main contribution is the theoretical unification between the involved curves, which is summarized in the diagram of Fig. 1. After briefly introducing standard hypotheses and notations, we recall definitions and properties of the aforementioned curves in Section 2. Relations between them are scrutinized in Section 3. In particular, we consider the link found by Djolonga et al. (2019) between PR curves and divergence frontiers (§ 3.2) for infinite Rényi exponent a, hence extending it from discrete distributions to general ones. More importantly, we show that Lorenz curves and PR curves are related through convex duality (§ 3.3). In addition, in Section 4, we explore several links with ϕ-divergences. In particular, thanks to integral representations of ϕ-divergences, we can show a reversed link between PR curves and divergence frontiers. As a side contribution, we end these notes in Section 5 by highlighting links existing between these trade-off curves and performance bounds used in the theory of domain adaptation. In particular, starting from the variational form of ϕ-divergences we extend the notion of Lorenz curves and use this extension to establish a new generalization bound.

2. Background on trade-off curves

In this section, we review several curves proposed in the literature to assess the similarity between two distributions P and Q. We simply summarize the principal definitions and useful results. Some notions are subject to minor adaptations in order to simplify the exposition of the links between the considered curves. Anytime such a revision is adopted, it shall be explicitly mentioned.

3. Note that, as we shall detail later, PR curves between distributions are different from the classical notion of PR curves in classification.

Let us start by recalling some standard notations, definitions, and results from measure theory. From now on, $(\Omega, \mathcal{A})$ represents a common measurable space, and we will denote $\mathcal{M}(\Omega)$ the set of sigma-finite signed measures, $\mathcal{M}^+(\Omega)$ the set of sigma-finite positive measures and $\mathcal{M}_p(\Omega)$ the set of probability distributions over that measurable space. The extended half real-line is denoted by $\overline{\mathbb{R}^+} = \mathbb{R}^+ \cup \{\infty\}$.

Definition 1 *Let μ, ν two signed measures. We denote by* $\mathrm{supp}(\mu)$ *the support*[4] *of μ, $|\mu|$ the total variation measure of μ, $\frac{d\mu}{d\nu}$ the Radon-Nikodym derivative of μ w.r.t. ν and $\mu \wedge \nu = \min(\mu, \nu) := \frac{1}{2}(\mu + \nu - |\mu - \nu|)$ (a.k.a the measure of largest common mass between μ and ν (Piccoli et al., 2019)). Besides, as usual, $\mu \ll \nu$ means that μ is absolutely continuous w.r.t. ν.*

2.1 Precision-recall curves

The PR curves were first proposed by Sajjadi et al. (2018) for discrete distributions and then extended to the general case by Simon et al. (2019). We follow the definition of the latter up to a minor fix[5].

Definition 2 *Let P, Q two distributions from $\mathcal{M}_p(\Omega)$. We refer to the Precision-Recall set $PRD(P, Q)$ as the set of Precision-Recall pairs $(\alpha, \beta) \in \mathbb{R}^+ \times \mathbb{R}^+$ such that*

$$\exists\, \mu \in AC(P, Q), P \geq \beta\mu, Q \geq \alpha\mu \;, \tag{1}$$

where $AC(P, Q) := \{\mu \in \mathcal{M}_p(\Omega) / \mu \ll P \text{ and } \mu \ll Q\}$.

The *precision* value α is related to the proportion of the generated distribution Q that matches the true data P, while conversely the *recall* value β is the amount of the distribution P that can be reconstructed from Q. Because of the lack of natural order on $[0, 1] \times [0, 1]$, Simon et al. (2019) has proposed to focus on the Pareto front of $\mathrm{PRD}(P, Q)$ defined as follows.

Definition 3 *The precision recall-curve $\partial PRD(P, Q)$ is the set of $(\alpha, \beta) \in PRD(P, Q)$ such that*

$$\forall (\alpha', \beta') \in PRD(P, Q), \alpha \geq \alpha' \text{ or } \beta \geq \beta'.$$

Note that this curve is clearly the Pareto front of the set:

$$\{(\kappa^*(Q|\mu), \kappa^*(P|\mu)) / \mu \in AC(P, Q)\}$$

where, following Scott et al. (2013), we define $\kappa^*(P|\mu) := \max\{\beta \in [0, 1] / \exists \nu \in \mathcal{M}_p(\Omega), P = \beta\mu + (1 - \beta)\nu\} = \inf_{\substack{A \in \mathcal{A} \\ \mu(A) > 0}} \frac{P(A)}{\mu(A)}$ (and similarly for Q).

In fact, this frontier is a curve for which Sajjadi et al. (2018) have exposed a parameterization, later generalized by Simon et al. (2019). We recall their result now.

4. Although a precise definition of the support requires a topology, we gloss over this issue because the support will not play a central role in the technical derivations.
5. There is an issue with their original definition where $(1, 0)$ and $(0, 1)$ are always in $\mathrm{PRD}(P, Q)$, while they should not when part of the mass of P is absent from Q and vice versa. Our fix consists in considering only the distributions μ that are absolutely continuous w.r.t P and Q.

Theorem 4 *Let P, Q two distributions from $\mathcal{M}_p(\Omega)$ and (α, β) non negative. Then, denoting*

$$\forall \lambda \in \overline{\mathbb{R}^+}, \begin{cases} \alpha_\lambda := ((\lambda P) \wedge Q)(\Omega) \\ \beta_\lambda := \left(P \wedge \frac{1}{\lambda} Q\right)(\Omega) \end{cases} \tag{2}$$

1. *$(\alpha, \beta) \in PRD(P, Q)$ iff $\alpha \leq \alpha_\lambda$ and $\beta \leq \beta_\lambda$ where $\lambda := \frac{\alpha}{\beta} \in \overline{\mathbb{R}^+}$.*
2. *As a result, the PR curve can be parameterized as:*

$$\partial PRD(P, Q) = \{(\alpha_\lambda, \beta_\lambda)/\lambda \in \overline{\mathbb{R}^+}\}. \tag{3}$$

In the previous theorem, and in agreement with the standard convention in measure theory, $0 \times \infty = 0$ so that $\alpha_\infty = Q(\mathrm{supp}(P))$ and $\beta_0 = P(\mathrm{supp}(Q))$.

Remark 5 *The previous parameterization reveals that the precision-recall curve is intrinsically related to the DeGroot statistical information (DeGroot, 1962) which is defined as $\Delta B_\pi(P, Q) := B_\pi(P, P) - B_\pi(P, Q)$ through the following divergence (that we will refer in short as the DeGroot divergence):*

$$B_\pi(P, Q) := [\pi P \wedge (1 - \pi) Q](\Omega) \tag{4}$$

where $\pi \in [0, 1]$ is an arbitrary prior probability. In simple words, the link with the precision-recall curve is merely a change of parameterization: $\pi = \frac{\lambda}{1+\lambda}$. This relationship will play a crucial role in one of the links made with ϕ-divergences in section 4 (see the integral representation arrow in Figure 1).

Another useful result of Simon et al. (2019) linking the frontier to likelihood ratio test is summarized now:

Theorem 6 *Let P, Q two distributions from $\mathcal{M}_p(\Omega)$. Then*

$$\forall \lambda \in \overline{\mathbb{R}^+}, \begin{cases} \alpha_\lambda = \lambda (1 - P(A_\lambda)) + Q(A_\lambda) \\ \beta_\lambda = 1 - P(A_\lambda) + \frac{Q(A_\lambda)}{\lambda} \end{cases}, \tag{5}$$

where the likelihood ratio sets are defined as

$$A_\lambda := \left\{ \frac{dQ}{d(P+Q)} \leq \lambda \frac{dP}{d(P+Q)} \right\}. \tag{6}$$

2.2 Divergence frontiers

Divergence frontiers were proposed very recently by Djolonga et al. (2019) as a generalization of precision-recall curves. Such a notion builds upon the Rényi divergence between two distributions.

Definition 7 *Let $\mu, \nu \in \mathcal{M}_p(\Omega)$ two distributions such that $\mu \ll \nu$ and $a \in \overline{\mathbb{R}^+} \setminus \{1\}$. The a-Rényi-divergence between μ and ν is defined as:*

$$D_a(\mu \parallel \nu) := \log \left(\left\| \frac{d\mu}{d\nu} \right\|_{a-1, d\mu} \right) \tag{7}$$

where the invoked norm is defined as

$$\|f\|_{a-1,d\mu} := \left(\int f^{a-1} d\mu \right)^{\frac{1}{a-1}}$$

when $a < +\infty$ and is the essential supremum norm for $a = \infty$. Besides when $a = 1$, this definition is extended by continuity and leads to the KL-divergence.

We adapt the definition of divergence frontiers from Djolonga et al. (2019).

Definition 8 (Divergence frontiers) *Let P, Q two distributions and $a \in \overline{\mathbb{R}^+}$. Then the exclusive realizable divergence region is defined*[6] *as the set:*

$$\mathcal{R}_a^{\cap}(P, Q) := \{(D_a(\mu \parallel Q), D_a(\mu \parallel P)) \,/ \mu \in AC(P, Q)\} \tag{8}$$

And the exclusive divergence frontier is defined as the (weak) Pareto front of this region, that is the set $\partial\mathcal{R}_a^{\cap}$ of couples $(\pi, \rho) \in \mathcal{R}_a^{\cap}$ such that:

$$\forall(\pi', \rho') \in \mathcal{R}_a^{\cap}, \pi \leq \pi' \text{ or } \rho \leq \rho' \,. \tag{9}$$

In the event that $\mathcal{R}_a^{\cap}(P, Q) = \emptyset$, the frontier is by convention restricted to the point $(+\infty, +\infty)$.

2.3 Lorenz and ROC curves

Lorenz curves were originally introduced by Lorenz (1905) to delineate income inequalities. In essence they highlight how much a single one dimensional distribution differs from a uniform distribution. This notion was then generalized to characterize the closeness of two arbitrary distributions by Harremoës (2004); van Erven and Harremoës (2010).

Definition 9 *Let P, Q two distributions from $\mathcal{M}_p(\Omega)$. One defines the Lorenz diagram between P and Q as*

$$LD(P, Q) = \left\{ \left(\int f dP, \int f dQ \right) /0 \leq f \leq 1 \right\}, \tag{10}$$

where the function f is required to be measurable.

Then the Lorenz curve between P and Q is defined as the lower envelop of the Lorenz diagram:

$$F_{LD}^{P,Q}(t) := \inf_{\substack{0 \leq f \leq 1 \\ \int f dP \geq t}} \int f dQ. \tag{11}$$

In absence of ambiguity on the involved distributions, we shall denote it simply $F(t)$ rather than $F_{LD}^{P,Q}(t)$. This curve is easily shown to be a monotonic and convex function.

6. In the original work, the distribution μ may range on a restricted set of distributions such as an exponential family. Besides, to avoid situations where the divergence is ill-defined, we imposed that μ is absolutely continuous w.r.t both P and Q.

If one considers in (10) only the range of indicator functions, then one recovers a subset of the Lorenz diagram, from which the Lorenz diagram can be extracted by merely taking the closed convex hull. This fact underpins the equivalence between Lorenz diagrams/curves and the seemingly different notions of Mode Collapse Region / ROC curves proposed by Lin et al. (2018). Indeed, Lin et al. (2017) shows (in Remark 6) that their Mode Collapse Region (MCR) can be obtained as the convex hull of the set of points $(P(A), Q(A))$ where A is any measurable set such that $Q(A) \geq P(A)$. As such the MCR is the upper half of the Lorenz diagram when one cuts it along the main diagonal (*i.e.* the line segment between $(0, 0)$ and $(1, 1)$). Then the authors proceed to define the ROC curve as the upper envelop of the MCR, which in turns is the symmetric transform of the lower envelop (*i.e.* the Lorenz curve) along the same diagonal. For the sake of time precedence, we shall thereafter focus solely on the Lorenz curve.

Similarly, restricting (11) to indicator functions requires convexification (more precisely Γ-regularization) to recover the Lorenz curve. In fact, because of the Neyman-Pearson lemma, one can even restrict to the indicator functions of the likelihood ratio sets A_λ, which in light of Theorem 6 underlies a subtle link with precision-recall curves that we shall detail later on.

3. Relationships between trade-off curves

Before going into further details about how the aforementioned curves relate to one another, let us state a few general facts about how they differ. One can first note that each curve is subject to specific "regularity" properties such as monotonicity, convexity and boundedness. For instance, contrary to the Lorenz curve, the PR curve does not enjoy any convexity property. Similarly, both the PR curve and the Lorenz curve are bounded within the domain $[0, 1] \times [0, 1]$, while the divergence frontiers are not bounded in general. Last both the divergence frontiers and PR curves are decreasing while the Lorenz curve is increasing.

Despite these disparities, each curve serves a similar purpose and strong links exist between them. In the next paragraph, a few simple situations are first scrutinized to describe the behavior of PR and Lorenz curves (divergence frontiers are voluntarily excluded from this preamble). Then, in § 3.2 and § 3.3, we shall establish the exact links between those curves.

3.1 Some intuition on the PR and Lorenz curves

In this preamble, we focus on the principal function of the curves under consideration: namely how they characterize the similarity between P and Q. To make our discussion more concrete, we consider an illustrative case in Fig. 2, where P and Q are two (truncated-)Gaussian mixtures. For a given curve alternative, one may consider two extreme configurations. On one hand, the perfect match between P and Q *i.e.* $P = Q$ is represented in dashed green. On the other hand, the complete discord between P and Q, denoted by $P \perp Q$ corresponds to an empty overlap of their supports (or more formally to two mutually singular distributions) and is represented in red. Then a particular instance of the considered curve will appear as an in-between case. The closer it stands to the green spot (and therefore the farther from the red spot), the more similar P and Q.

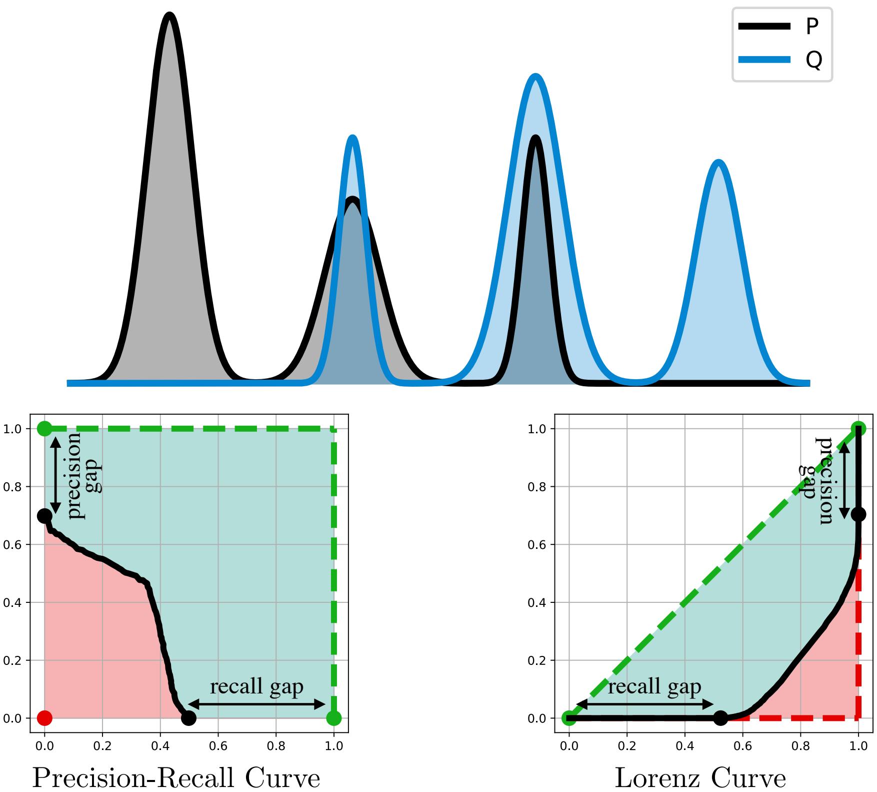

Figure 2: (top): graphical representations of two mixtures of (truncated-)Gaussians P and Q. (bottom): the corresponding alternate similarity curves. For each alternative, the green dashed curve exposes the extreme case where $P = Q$ and the red spot/curve when $P \perp Q$.

To illustrate the benefit of a trade-off curve in comparison to scalar metrics, we depict a few examples in Fig. 3. The three examples are obtained by adjusting the locations, widths and weights of the Gaussian mixture models. They corresponds to scenarios of idealistic modes of deviations between the two distributions P and Q, namely (a) pure mode dropping, (b) pure mode invention and (c) pure mode reweighting. In (a) a Gaussian component from P is missing in Q, and this translates in both curves. In the PR curve, it becomes manifest through a drop of recall that is depicted by the horizontal gap separating the curve from the dash green curve. In the Lorenz curve, this is depicted by the horizontal gap between the point where the curve becomes positive and the origin $(0, 0)$. In (b) an extra Gaussian component is present in Q and again this phenomenon is patent in the curves, but this time it is depicted by vertical gaps. In (c) P and Q present both two components centered at the same locations, but they have different mixing factors, and one of the two components is more spread in Q. In that scenario, maximal recall and maximal precision can be achieved but not at the same time. This trade-off is rendered in a way specific to each curve. In both cases, the horizontal and vertical gaps indicative of mode dropping and mode invention are null, but the curve smoothly interpolates from full recall to full

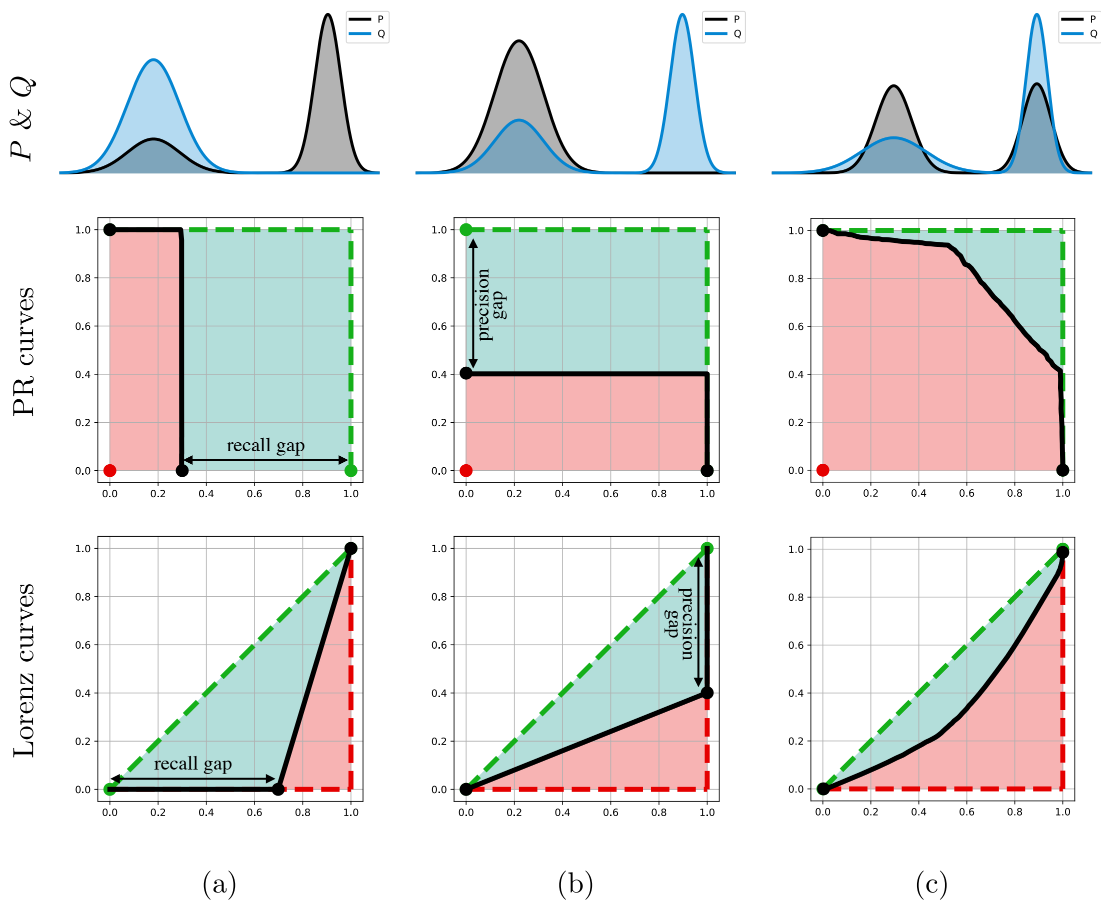

Figure 3: PR and Lorenz Curves in different scenarios: (a) pure mode dropping – (b) pure mode invention – (c) mode reweighting. Top row: the two distributions P (in black) and Q (in blue). Middle row: PR curves. Bottom row: Lorenz curves.

precision away from the green curve. The fact that one Gaussian component is identical in P and Q translates clearly in the precision-recall curve: indeed full recall can be obtained for a non-null precision (approximately 0.4 in this plot). Figure 2 is depicting an aggregate of these three extreme scenarios and both trade-off curves reflect the related phenomenons.

3.2 Precision-recall vs Divergence frontiers

In Djolonga et al. (2019), it is shown that in the case of discrete measures, the notion of divergence frontier matches precision-recall curve in the limit case where the Rényi exponent $a \to \infty$. We extend here this result to the case of general distributions. To do so, we will rely on the following technical lemma.

Lemma 10 *Let $\mu \in \mathcal{M}_p(\Omega)$ and $\nu \in \mathcal{M}_p(\Omega)$ two distributions.*

1. *If $\mu \ll \nu$ then* $\sup_{A \in \mathcal{A}} \frac{\mu(A)}{\nu(A)} = \operatorname{ess\,sup}_{d\mu} \frac{d\mu}{d\nu}$;

2. otherwise,

$$\sup_{A\in\mathcal{A}} \frac{\mu(A)}{\nu(A)} = \operatorname*{ess\,sup}_{d\mu} \frac{d\mu}{d(\mu+\nu)} \Big/ \frac{d\nu}{d(\mu+\nu)}.$$

Proof For the sake of completeness, we provide a proof of this technical result in Appendix A. ■

Theorem 11 *Let P, Q two distributions. Then,*

$$\partial PRD(P,Q) = \left\{(e^{-\pi}, e^{-\rho})/(\pi,\rho) \in \partial\mathcal{R}^{\cap}_{\infty}(P,Q)\right\}.$$

Proof From the definition of the Rényi divergence it is clear that

$$\begin{aligned} D_\infty(\mu \parallel Q) &= \log\left(\left\|\frac{d\mu}{d(\mu+Q)} \Big/ \frac{dQ}{d(\mu+Q)}\right\|_{\infty,d\mu}\right) \\ &= \log\left(\operatorname*{ess\,sup}_{d\mu} \frac{d\mu}{d(\mu+Q)} \Big/ \frac{dQ}{d(\mu+Q)}\right) \end{aligned}$$

which in turn can be expressed thanks to Lemma 10 as

$$D_\infty(\mu \parallel Q) = \log\left(\sup_{A\in\mathcal{A}} \frac{\mu(A)}{Q(A)}\right).$$

A similar derivation applies to $D_\infty(\mu \parallel P)$.

Besides, as stated at the end of Definition 3 it is clear that the precision-recall curve is obtained as the Pareto-front of the set of pairs $(\inf_{\substack{A\in\mathcal{A}\\ \mu(A)>0}} \frac{Q(A)}{\mu(A)}, \inf_{\substack{A\in\mathcal{A}\\ \mu(A)>0}} \frac{P(A)}{\mu(A)})$ where μ describes all distributions in $AC(P,Q)$. Note that because of the standard measure theory convention $\frac{0}{0} = 0$, we have that

$$\sup_{A\in\mathcal{A}} \frac{\mu(A)}{Q(A)} = \sup_{\substack{A\in\mathcal{A}\\ \mu(A)>0}} \frac{\mu(A)}{Q(A)} = 1 \Big/ \inf_{\substack{A\in\mathcal{A}\\ \mu(A)>0}} \frac{Q(A)}{\mu(A)} = e^{D_\infty(\mu\|Q)}.$$

The claimed identity easily follows. ■

3.3 Precision-recall vs Lorenz curves

The question of the relation between PR and Lorenz/ROC curves is reminiscent of the comparison of PR and ROC curves for binary classification (Davis and Goadrich, 2006). Note however that despite their name, PR-curves for distributions are not the same as the PR-curves of the likelihood ratio classifier as one might be inclined to believe. In fact, they are composed of mixed error rates of the said classifier (see Theorem 6).

In essence, PR and Lorenz curves are two ways of exposing the pairs $(P(A_\lambda), Q(A_\lambda))$. Yet, the following questions are not trivial. Given the PR-curve of P and Q, can we compute their Lorenz curve? Reciprocally, can we compute the PR-curve from the Lorenz one? If one had a more complete representation such as $(\lambda, P(A_\lambda), Q(A_\lambda))$, then one could easily compute both the PR-curve and the Lorenz curve, but in each representation, at least one datum is not explicitly known:

1. In the Lorenz curve, λ is not readily available, but we will see that it is in fact hidden in the sub-derivative of the Lorenz curve.
2. In the PR-curve, λ can be easily computed as the ratio $\frac{\alpha_\lambda}{\beta_\lambda}$ but the values of $P(A_\lambda)$ and $Q(A_\lambda)$ are mingled within α_λ so that one needs to untangle them before recovering the Lorenz curve. Note that, given a fixed λ, the system of equations given by α_λ and β_λ in Eq. (5) is always under-determined (rank 1) and therefore does not suffice on its own to recover the values of $P(A_\lambda)$ and $Q(A_\lambda)$.

We will rely on the following Lemma.

Lemma 12 *Let P, Q two distributions from $\mathcal{M}_p(\Omega)$. Then $\forall \lambda \in \overline{\mathbb{R}^+}$,*

$$\alpha_\lambda = \min_{0 \le f \le 1} \lambda \left(1 - \int f dP\right) + \int f dQ \tag{12}$$

where the functions f are measurable.

Proof See Appendix B. ■

From Lemma 12, one can draw the following link between the PR-curve and the Lorenz curve.

Theorem 13 *Let P and Q two distributions. Let $\lambda \in \mathbb{R}^+$. Consider the Lorenz Curve F defined in Eq. (11), then,*

$$F^*(\lambda) = \lambda - \alpha_\lambda \tag{13}$$

where $F^*(\lambda) = \sup_{t \in [0,1]} \lambda t - F(t)$ is the Legendre transform of F.

Proof Let $\lambda \ge 0$. Let us show that $F^*(\lambda) = \lambda - \alpha_\lambda$. Indeed , $\forall t \in [0,1]$

$$\begin{aligned}
\lambda t - F(t) =& \lambda t - \inf_{\substack{0 \le f \le 1 \\ \int f dP \ge t}} \int f dQ = \sup_{\substack{0 \le f \le 1 \\ \int f dP \ge t}} \lambda t - \int f dQ \\
\le& \sup_{\substack{0 \le f \le 1 \\ \int f dP \ge t}} \lambda \int f dP - \int f dQ \\
\le& \sup_{0 \le f \le 1} \lambda \int f dP - \int f dQ \\
=& \lambda - \alpha_\lambda \qquad \text{(thanks to Lemma 12)}
\end{aligned}$$

Which shows that $\lambda - \alpha_\lambda \ge \sup_{t \in [0,1]} \lambda t - F(t)$. Besides, letting $t_\lambda := P(A_\lambda)$

$$\begin{aligned}
\lambda - \alpha_\lambda =& \lambda - (\lambda(1 - P(A_\lambda)) + Q(A_\lambda)) \\
=& \lambda P(A_\lambda) - Q(A_\lambda) = \lambda t_\lambda - F(t_\lambda)
\end{aligned}$$

where we have used that if $t_\lambda = P(A_\lambda)$ then $F(t_\lambda) = Q(A_\lambda)$ (a result induced by the standard Neyman-Pearson lemma). Therefore, $\lambda - \alpha_\lambda = \sup_{t \in [0,1]} \lambda t - F(t) = F^*(\lambda)$. ■

Remark 14 *Theorem 13 brings many valuable prospects concerning the link between the PR and Lorenz curves.*

1. *First, since the Legendre transform is a one-to-one involution, the PR and Lorenz curves are theoretically equivalent.*

2. *Besides, letting* $t_\lambda := P(A_\lambda)$ *and relying on the Fenchel identity, one gets that* $\lambda \in \partial F(t_\lambda)$*, which theoretically provides a means to extract the missing datum as soon as one is capable of computing the subdifferential of the Lorenz curve.*

3. *More concretely, the theorem provides us a practical way to compute* α_λ *from the Lorenz curve. Indeed, given* λ*,* α_λ *can be computed by solving the following 1D convex problem:*
$$\alpha_\lambda = \lambda - F^*(\lambda) = \min_{t \in [0,1]} F(t) + \lambda(1-t)$$
One can do so efficiently thanks to the bisection method if the subdifferential of F *is available or resort to derivative-free algorithms such as the Golden Section Search method otherwise. Then* β_λ *is obtained as* $\frac{\alpha_\lambda}{\lambda}$*.*

4. *In the other way around, given* $t \in [0,1]$*, one can solve for* $F(t)$ *by considering the following 1D concave problem:*
$$\begin{aligned} F(t) = F^{**}(t) &= \sup_{\lambda \in \mathbb{R}^+} \lambda t - F^*(\lambda) \\ &= \sup_{\lambda \in \mathbb{R}^+} \alpha_\lambda + \lambda(t-1). \end{aligned}$$

4. Several links implying ϕ-divergences

In this section, we scrutinize the relationship between the trade-off curves and ϕ-divergences. To do so, we build on the link between precision and the DeGroot statistical information. In particular, as illustrated in the lower right part of Figure 1, the precision-recall curve can be cast as a parametric family of ϕ-divergence. Reciprocally, ϕ-divergences can be expressed as integral representations from the PR-curve. Note, that both facts were rediscovered recently in Verine et al. (2023). Last, based on the integral representation, we show that the Rényi divergence frontier can be recovered theoretically from the PR-curve.

4.1 A short digest on ϕ-divergences

In what follows, $\phi : \mathbb{R} \to \mathbb{R}$ is a l.s.c. convex function verifying the following assumptions:

(A0) $\mathrm{dom}(\phi) \cap]-\infty, 0[= \emptyset$ and $]0, +\infty[\subset \mathrm{dom}(\phi)$

(A1) $\phi(1) = 0$

(A2) $0 \in \partial\phi(1)$ and $\partial\phi(1)$ is symmetric around 0 (ie $\phi'_-(1) = -\phi'_+(1)$)

(A3) ϕ is strictly convex at 1

We will denote Φ_1 the set of l.s.c. convex functions that follows these four assumptions. We denote $\phi^\diamond(u) = u\phi(\frac{1}{u})$, the so-called Csizár dual of ϕ. In agreement with Liese and Vajda (2006), we use the following definition of ϕ-divergences.

Definition 15 (ϕ-divergence) *Let $\phi \in \Phi_1$ and μ and ν be two positive measures. Then,*

$$D_\phi(\mu\|\nu) := \int \phi\left(\frac{d\mu}{d\nu}\right) d\nu + \phi(0)\nu(\{\mu = 0\}) + \phi^\diamond(0)\mu(\{\nu = 0\}) \tag{14}$$

with $\phi(0) := \lim_{u\to 0} \phi(u)$ and $\phi^\diamond(0) := \lim_{u\to 0} u\phi(\frac{1}{u})$.

Remark 16 *Let us give some rationale behind the four assumptions defining Φ_1. The first part of Assumption (A0) restrict the ϕ divergence to positive measures (or same sign measures) and hence removes any subjective choice for the value of $\phi(\frac{d\mu}{d\nu})$ when $\frac{d\mu}{d\nu} < 0$. Taken together, the three other assumptions defining Φ_1 enforce that $D_\phi(\mu\|\nu) \geq 0$ with equality iff $\mu = \nu$. Indeed, using assumptions (A1) and (A2), one sees that $\forall u \in \mathbb{R}$, $\phi(u) \geq \phi(1) = 0$, and the last assumption implies that $\phi(u) = 1$ iff $u = 1$. Assumption (A2) is not always required in the literature when dealing with probability (because in this case, $\mu = P, \nu = Q$ are distributions and $D_\phi(P\|Q)$ is invariant to changes of the form $\tilde{\phi}(u) = \phi(u) - \frac{\phi'_+(1)+\phi'_-(1)}{2}(u-1)$). Besides, notice that $\phi \in \Phi_1$ iff $\phi^\diamond \in \Phi_1$ and the following symmetry relation holds: $D_\phi(\mu\|\nu) = D_{\phi^\diamond}(\nu\|\mu)$.*

Remark 17 *In addition to the Csizár dual, the Legendre-Fenchel dual of ϕ will play a key role in variational forms of D_ϕ (Nguyen et al., 2010). It is therefore interesting to comment on the impact of the assumptions $\phi \in \Phi_1$ on ϕ^*. In fact, $\phi(1) = \min_u \phi(u)$ means exactly that $\phi^*(0) = -\phi(1)$. Besides, since $\phi(1) = 0$, it means that $\phi^*(0) = 0$. Symmetrically, when $0 \in \mathrm{dom}(\phi)$, let $v_0 \in \partial\phi(0)$, then $\phi^*(v_0) = \min_v \phi^*(v)$. Besides, since ϕ is minimized at $u = 1$, then it is decreasing wherever $u < 1$ and increasing otherwise (by convexity). In particular, $\partial\phi(0) \subset [-\infty, 0]$ and therefore the minimum of ϕ^* can only be reached at a point $v_0 \leq 0$ (it is also possible that the infimum is not reached if $\partial\phi(0) = \{-\infty\}$ or if $0 \notin \mathrm{dom}(\phi)$).*

To conclude this remark, let us notice that the definition of ϕ on $]-\infty, 0]$ is irrelevant for $D_\phi(\mu\|\nu)$ (indeed, since μ and ν are assumed positive, then $\frac{d\mu}{d\nu} \geq 0$). As a result, D_ϕ is also not impacted by the values taken by $\phi^*(v)$ for $v < v_0 := \sup\{v \in \arg\min(\phi^*)\}$. In any case, since assumption (A0) imposes $\phi(\frac{d\mu}{d\nu}) = +\infty$ when $\frac{d\mu}{d\nu} < 0$ then $\phi^*(v) = \phi^*(v_0)$ whenever $v < v_0$. To simplify concrete formulae, e.g. *in Table 1, we will focus on $\phi^*(v)$ on the following restriction of its domain $\mathrm{dom}^{v_0\rightarrow}(\phi^*) := \cup_{u\geq 0}\partial\phi(u) = \mathrm{dom}(\phi^*) \cap [v_0, +\infty[$. It includes at least $]v_0, \phi'_+(+\infty)[$, and since ϕ is strictly convex at 1, $\phi'_+(\infty) > 0$ (indeed $\phi'_+(\infty) \geq \phi'(1) \geq 0$ and if by the symmetry assumption of (A2) $\phi'_+(1) = -\phi'_-(1) = 0$ and then strict convexity implies that $\phi'_+(\infty) > \phi'_+(1)$. Therefore $\mathrm{dom}(\phi^*) \supset]v_0, 0]$.*

4.2 Integral representation of an ϕ-divergence using the precision-recall curve

Considering two distributions P and Q, the very definition of $D_\phi(P\|Q)$ involves $P(\{Q = 0\}) = 1 - \beta_0$ and $Q(\{P = 0\}) = 1 - \alpha_\infty$ that is to say the gap between the best possible

Divergence	$\phi(u)$	$\phi'(u)$	$\mathrm{dom}^{v_0\rightarrow}(\phi^*)$	$\phi^{*\prime}(v)$	$\phi^*(v)$
KL	$u\log(u)-(u-1)$	$\log(u)$	$\mathbb{R}$	e^v	e^v-1
rKL	$-\log(u)+(u-1)$	$1-\frac{1}{u}$	$]-\infty,1[$	$\frac{1}{1-v}$	$-\log(1-v)$
JS	$-(u+1)\log\frac{1+u}{2}$ $+u\log u$	$\log(\frac{2u}{1+u})$	$]-\infty,\log(2)[$	$\frac{e^v}{2-e^v}$	$-\log(2-e^v)$
χ^2_{Pearson}	$(u-1)^2$	$2(u-1)$	$[-2,+\infty[$	$\frac{v}{2}+1$	$\frac{v^2}{4}+v$
Hellinger	$(\sqrt{u}-1)^2$	$1-\frac{1}{\sqrt{u}}$	$]-\infty,1[$	$\frac{1}{(1-v)^2}$	$\frac{v}{1-v}$
TV	$\lvert u-1\rvert$	$\mathrm{sign}(u-1)$	$[-1,1]$	1	v

Table 1: A few standard ϕ-divergences. For each choice of ϕ we provide the entries in the natural derivation order, namely we compute first $\phi'(u)$ then we derive the "meaningful" domain of ϕ^*, that is $\mathrm{dom}^{v_0\rightarrow}(\phi^*)$, then $\phi^{*\prime}(v)$ is determined as the inverse function of $\phi'(u)$ and $\phi^*(v)$ is determined as the anti-derivative that verifies $\phi^*(0)=0$. Note that ϕ^* and its derivative are only given on their meaningful domain.

precision/recall and their actual values. In fact, building upon known integral representations of ϕ-divergences (Liese and Vajda, 2006), the entire divergence can be expressed as an integral of weighted precision-recall gaps. Let us first reformulate the link highlighted in Remark 5 between DeGroot's statistical information and the precision-recall curve.

Proposition 18 *Let P,Q be two distributions, $\pi\in[0,1]$ and denoting by $F^1_\pi(P,Q):=\frac{2}{\frac{1}{\alpha_\lambda}+\frac{1}{\beta_\lambda}}$ the F^1-score associated with the precision-recall pair $(\alpha_\lambda,\beta_\lambda)$ where $\lambda=\frac{\pi}{1-\pi}$. Then:*

$$F^1_\pi(P,Q)=2B_\pi(P,Q)$$

As a result, the DeGroot's statistical information is related to the gaps of F^1 scores as follows:

$$\Delta B_\pi(P,Q)=\frac{1}{2}\Delta F^1_\pi(P,Q) \tag{15}$$

with $\Delta F^1_\pi(P,Q):=F^1_\pi(P,P)-F^1_\pi(P,Q)$.

Proof Let us first recall the definition of $B_\pi(P,Q):=(\pi P\wedge(1-\pi)Q)(\Omega)$ (see Remark 5) The proposition derives from a simple computation and the fact that $\alpha_\lambda=\lambda\beta_\lambda$:

$$\begin{aligned}F^1_\pi:=&\frac{2\alpha_{\frac{\pi}{1-\pi}}\beta_{\frac{\pi}{1-\pi}}}{\alpha_{\frac{\pi}{1-\pi}}+\beta_{\frac{\pi}{1-\pi}}}\\ =&\frac{2\alpha_{\frac{\pi}{1-\pi}}}{\frac{\pi}{1-\pi}+1}=2(1-\pi)\alpha_{\frac{\pi}{1-\pi}}\\ =&2(1-\pi)[\frac{\pi}{1-\pi}P\wedge Q](\Omega)=2B_\pi(P,Q)\end{aligned}$$

■

From the previous result, one can restate the Theorem 11 of Liese and Vajda (2006) in terms of precision-recall F^1 scores.

Corollary 19 *Let P, Q two distributions and $\phi \in \Phi_1$. Then,*

$$D_\phi(P\|Q) = \frac{1}{2}\int_0^1 \Delta F^1_\pi(P,Q) d\Gamma_\phi(\pi) \tag{16}$$

where $d\Gamma_\phi(\pi) := \frac{1}{\pi} d\phi'_+(\frac{1-\pi}{\pi})$ intrinsically represents the distribution of curvature of the convex function ϕ.

Proof This is a mere restatement of the result of (Liese and Vajda, 2006, Theorem 11), based on Equation 15 (in comparison, (Liese and Vajda, 2006, Theorem 11) was formulated in terms of $\Delta B_\pi(P,Q)$). ■

Remark 20 *Therefore, depending on the curvature distribution of ϕ, minimizing the ϕ-divergence between P and Q (as done in f-GAN (Nowozin et al., 2016)) is equivalent to minimizing a weighted version of the gap between the best F^1_π score and the actual score. For instance, one can consider the following family of functions ϕ_a and their associated ϕ-divergence, that we will refer to as Tsallis[7] a-divergence (Tsallis, 1988):*

$$\phi_a(u) := \frac{1}{a(a-1)}(u^a - a(u-1) - 1) \tag{17}$$

In that case, the curvature is given by $\phi''_a(u) = u^{a-2}$ and $d\Gamma_{\phi_a}(\pi) = \frac{1}{\pi^3}\phi''_a\left(\frac{1-\pi}{\pi}\right)d\pi = \frac{1}{\pi^3}\left(\frac{1-\pi}{\pi}\right)^{a-2} d\pi$. As a result, one gets

$$D_{\phi_a}(P\|Q) = \frac{1}{2}\int_0^1 \Delta F^1_\pi(P,Q)\left(\frac{1-\pi}{\pi}\right)^a \frac{1}{\pi(1-\pi)^2} d\pi$$

One can see that depending on the value of a, this objective will focus rather on regions where precision is more important $\pi \to 1$ or on those where recall is paramount $\pi \to 0$.

4.3 Divergence frontiers parameterization with ϕ-divergences

In this section, we consider again the Rényi divergence frontiers introduced by Djolonga et al. (2020). We recall that the Rényi divergence of exponent a between μ and P is denoted by $D_a(\mu \parallel P)$. Following Djolonga et al. (2020) with a minor correction, one can consider the following monotone transform of the Rényi divergence[8] and recover the previously introduced Tsallis a-divergence: $\hat{D}_a(\mu\|P) := \frac{1}{a(a-1)}\left(\exp\left((a-1)D_a(\mu\|P)\right) - 1\right) = \frac{1}{a(a-1)}\int \left(\frac{d\mu}{dP}\right)^a - 1 dP = D_{\phi_a}(\mu\|P)$. Since this is a ϕ-divergence, it might be expected

7. Note that the normalizing constant is usually set as $\frac{1}{a-1}$ for Tsallis divergences instead of $\frac{1}{a(a-1)}$. Yet the normalization chosen here presents several advantages : in particular the obtained family of divergence is well defined for all $a \in \mathbb{R}$ (instead of only $a \geq 0$) and besides the Csizár dual is given by $\phi_a^\diamond = \phi_{1-a}$. The obtained divergence was simply called α-divergence in Póczos and Schneider (2011).
8. Actually we had to amend slightly the transform used in Djolonga et al. (2020) because of a minor flaw in their exposition as well as out of personal convenience.

through the link developed in section 4.2 that this divergence frontier is determined by the precision-recall curve. Making this statement concrete requires a bit of work though, because the divergence involved here is not directly expressed between P and Q but implies an auxiliary distribution μ (in this context, μ is meant to interpolate between P and Q). The following proposition materializes the anticipated link.

Proposition 21 *Let P, Q two distributions. The above defined Tsallis divergence frontier can be obtained as the pairs: $(\hat{D}_a(\mu_\theta \| P), \hat{D}_a(\mu_\theta \| Q))$ with $\theta \in [0,1]$ and*

$$\frac{d\mu_{\theta,a}}{dP} = \frac{\left(\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a}\right)^{\frac{1}{1-a}}}{\int (\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a})^{\frac{1}{1-a}} dP} \tag{18}$$

As a result,

$$\hat{D}_a(\mu_{\theta,a} \| P) = \frac{1}{a(a-1)} \left(\frac{\int (\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a})^{\frac{a}{1-a}} dP}{\left(\int (\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a})^{\frac{1}{1-a}} dP\right)^a} - 1 \right). \tag{19}$$

Proof Note that this proposition was already presented in (Djolonga et al., 2020, Proposition 1, Proposition 3). We still provide a proof because the one presented in Djolonga et al. (2020) lacks details. The first point noted in Djolonga et al. (2020), is that contrary to the Rényi divergence, the Tsallis version is convex in both arguments, and in particular in μ. As a result, denoting by $\partial \hat{R}(P,Q)$ the Tsallis divergence frontier (defined similarly to Definition 8 using $\hat{D}_a$ instead of D_a), the bi-objective optimization associated with the Pareto-frontier $\partial \hat{R}(P,Q)$ can be solved through linear scalarization. This means that the divergence frontier for the Tsallis divergence is parametrized as

$$\partial \hat{R}(P,Q) = \{(\hat{D}_a(\mu_{\theta,a} \| P), \hat{D}_a(\mu_{\theta,a} \| Q)), \theta \in [0,1]\}$$

where

$$\mu_{\theta,a} = \underset{\mu \in \mathrm{AC}(P,Q)}{\arg\min}\ \theta \hat{D}_a(\mu \| P) + (1-\theta) D_a(\mu \| Q).$$

This is a generalized centroid problem that has been first solved in Amari (2007), as follows. Denoting, with a slight abuse of notation, $m_\theta = \frac{d\mu_{\theta,a}}{dP}$, the optmization problem becomes:

$$m_\theta = \underset{\mu = m dP \text{ s.t. } \int m dP = 1}{\arg\min}\ \theta \hat{D}_a(\mu \| P) + (1-\theta) D_a(\mu \| Q).$$

Using a Lagrange multiplier, the first order optimality condition is:

$$\int m_\theta^{a-1} (\theta + (1-\theta)\left(\frac{dP}{dQ}\right)^{a-1}) \delta m_\theta dP + \lambda \int \delta m_\theta dP = 0$$

where δm_θ is an arbitrary measurable function. This yields $m_\theta^{1-a} \propto (\theta + (1-\theta)\left(\frac{dP}{dQ}\right)^{a-1})$. Using the normalization constraint, one gets:

$$m_\theta = \frac{d\mu_{\theta,a}}{dP} = \frac{(\theta + (1-\theta)\left(\frac{dP}{dQ}\right)^{a-1})^{\frac{1}{1-a}}}{\int(\theta + (1-\theta)\left(\frac{dP}{dQ}\right)^{a-1})^{\frac{1}{1-a}} dP} = \frac{(\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a})^{\frac{1}{1-a}}}{\int(\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a})^{\frac{1}{1-a}} dP}.$$

As a result,

$$\begin{aligned}\hat{D}_a(\mu_{\theta,a}\|P) =& \frac{1}{a(a-1)} \int \left(\frac{d\mu_{\theta,a}}{dP}\right)^a - 1 dP \\ =& \frac{1}{a(a-1)} \left(\frac{\int(\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a})^{\frac{a}{1-a}} dP}{\left(\int(\theta + (1-\theta)\left(\frac{dQ}{dP}\right)^{1-a})^{\frac{1}{1-a}} dP\right)^a} - 1 \right).\end{aligned}$$

■

Remark 22 *The previous parameterization of the divergence frontier was given in (Djolonga et al., 2020, Proposition 3). It is expressed entirely in terms of ϕ-divergences between P and Q. As a result, the divergence-frontier is completely determined by the DeGroot statistical information $\Delta B_\pi(P, Q)$ which in turn are in bijection with the precision-recall curve between P and Q. From this observation, one can state that for any exponent a, the Rényi divergence frontier is entirely determined by the precision-recall curve. In other words, the precision-recall curve is always more complete a description than any Rényi divergence frontier. The same remark holds for the interpolated ϕ-divergence curves proposed in Liu et al. (2021) which, by design, are parameterized with ϕ-divergences between P and Q. That being said, the estimation from a finite sample of the precision-recall curve in a non-parametric settings might reveal more complicated than divergence frontiers. This question deserves further scrutiny, and although it is out of the scope of this article, the interesting reader can refer to (Rubenstein et al., 2019; Liu et al., 2021) and references therein.*

Remark 23 *The reader can verify that taking the limit case $a \to \infty$ for the parameterization $\mu_{\theta,a}$ tends to a single point of the precision recall curve, precisely $\mu_{\theta,\infty} = \frac{P \wedge Q}{(P \wedge Q)(\Omega)}$. In order to recover the full curve, one should use a parameterization that depends on the Rényi exponent a, by setting: $(\theta_a, 1-\theta_a) \propto (\pi^{1-a}, (1-\pi)^{1-a})$ where $\pi \in [0,1]$ is the new parameter. In that case,*

$$\frac{d\mu_{\theta_a,a}}{dP} = \frac{(\pi^{1-a} + \left((1-\pi)\frac{dQ}{dP}\right)^{1-a})^{\frac{1}{1-a}}}{\int(\pi^{1-a} + \left((1-\pi)\frac{dQ}{dP}\right)^{1-a})^{\frac{1}{1-a}} dP}$$

Then $\mu_{\theta_a,a} \to \frac{\pi P \wedge (1-\pi)Q}{(\pi P \wedge (1-\pi)Q)(\Omega)}$ as $a \to \infty$. With such a parameterization, the numerator and denominator in the expression of $\hat{D}_a(\mu_{\theta_a,a}\|P)$ are reminiscent of Arimoto divergences (see e.g. *(Liese and Vajda, 2006)).*

5. Application to generalization bounds for domain adaptation

In this part we revisit a standard bound from domain adaptation. In this setting, P and Q stand for the source and target distributions over a joint space $\Omega = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is called the sample space and $\mathcal{Y}$ is a finite set of classes referred as the label space. Given a classifier h (a.k.a an hypothesis) one would like to control the error $R_Q(h)$ in the target domain which is not directly observable (for lack of supervision) with a bound relying on the corresponding error $R_P(h)$ in the source domain (where supervision is available). The simplest bound proposed in the literature is the following (Ben-David et al., 2010):

$$R_Q(h) := \int \mathbb{1}_{h(x) \neq y} dQ(x,y) \leq R_P(h) + \|P - Q\|_{TV} \,. \tag{20}$$

In the common *covariate-shift* assumption (where the conditional distribution of labels given samples is the same in both domains), the bound can be expressed in terms of the marginal distribution over $\mathcal{X}$. It was noted directly by Ben-David et al. (2010) that the previous bound suffers from two major issues in practice. The first one is that the bound cannot be estimated efficiently from finite i.i.d samples because of its reliance on the TV norm (which involves a class of measurable sets that is not restricted enough for such purpose). The second issue lies in the fact that the bound does not imply the class of hypothesis functions where h lives (*e.g.* a class of small VC dimension or of small Rademacher capacity). As such, the bound in question is over-pessimistic, because it takes for granted that error set $\{h(x) \neq y\}$ may be spread as an arbitrary measurable set, which may be far from the case depending on restrictions applying to h. The authors then derived a more adapted bound based on the so-called $\mathcal{H}\Delta\mathcal{H}$-divergence, answering the previous two points, and many other bounds were later proposed. The interested reader may refer to Redko et al. (2020) for a very up-to-date and comprehensive review of such works.

5.1 A refined TV bound

Despite its shortcomings, we would like to revisit Eq. (20) and provide an optimized version of it which has an intuitive interpretation. We will first derive a bound, based on the Lorenz curve between P and Q and then express it in a form closer to Eq. (20) by relying on the duality between Lorenz and PR curves.

Proposition 24 *The Lorenz curve provides a lower bound for domain adaptation.*

$$R_Q(h) \leq 1 - F(1 - R_P(h)) \tag{21}$$

Proof

$$
\begin{aligned}
F(1 - R_P(h)) &= \inf_{\substack{g \text{ measurable} \\ 0 \leq g \leq 1 \\ \int g dP \geq 1 - \int \mathbb{1}_{h \neq f} dP}} \int g dQ \\
&= \inf_{\substack{g \text{ measurable} \\ 0 \leq 1-g \leq 1 \\ \int (1-g) dP \leq \int \mathbb{1}_{h \neq f} dP}} \int 1 - (1-g) dQ \\
&= 1 - \sup_{\substack{g' \text{ measurable} \\ 0 \leq g' \leq 1 \\ \int g' dP \leq \int \mathbb{1}_{h \neq f} dP}} \int g' dQ \,.
\end{aligned}
$$

Considering the particular case: $g' = \mathbb{1}_{h \neq f}$ one gets

$$
F(1 - R_P(h)) \leq 1 - \int \mathbb{1}_{h \neq f} dQ = 1 - R_Q(h) \,.
$$

■

The exact same bound can be expressed with the PR-curve parametrization thanks to Theorem 13.

Proposition 25 *We have the following bound:*

$$
R_Q(h) \leq \lambda^* R_P(h) + (1 - \alpha_{\lambda^*}) \tag{22}
$$

with $\lambda^* = \underset{\lambda \in \mathbf{R}^+}{\arg\max} \{\alpha_\lambda - \lambda R_P(h)\}$.

Proof This result follows from Theorem 13. ■

The first part of our upper-bound corresponds to the errors occurring in the common support of P and Q. There, the error rate is controlled in the source domain, and is therefore also controlled in the target domain. The amplification factor λ^* accounts for the fact that the common mass between P and Q is present in different ratios in the two domains. The second part corresponds to errors occurring in the target domain, within mass that is not present in the source domain. We do not have control over this error, and must account for it by considering the worst case where h is always wrong. As a result, the only way to keep this term under control is to make some assumptions on the class of admissible functions for h and the distribution of labels, that is to say the hypothesis class and the concept class.

5.2 Discussion

The latter form of our bound (Eq. (22)) highlights a strong tie with Eq. (20). In particular, if $\lambda^* = 1$ then, noting that $\alpha_1 = 1 - \frac{1}{2}\|P - Q\|_{TV}$, then the two bounds are almost identical. The only difference resides in a factor $\frac{1}{2}$ in our favor, which stems from the fact that we explicitly leverage the non-negativity of the error. Besides in general, 1 is not the optimal

λ^* and our bound is even tighter. This is where the reliance on a trade-off curve comes in handy: we obtain virtually one bound for each value of λ and we can pick the sharpest one. Let us consider the simple example of Fig. 2 where we can read the value of α_λ on the y-axis at the location where the PR-curve meets the line of equation $\alpha = \lambda\beta$. In this example $\alpha_1 \approx 0.38$ which means that $\|P - Q\|_{TV} = 2(1 - \alpha_1) \approx 1.24$. Therefore the bound of Eq. (20) is larger than 1 and is hence non informative. On the other hand, Eq. (5) with $\lambda = 1$ gives $R_Q(h) \leq R_P(h) + (1 - \alpha_1) \approx R_P(h) + 0.62$ which is informative as soon as $R_P(h) < 0.38$. This condition is easily met in concrete cases because $R_P(h)$ is the error in the source domain, which is under control thanks to supervision. More importantly, we can examine how the optimal trade-off between the two terms of the error bound can yield a much sharper bound. In fact, given the convexity of $1 - \alpha_\lambda$, the optimal λ is characterized by the first-order critical point condition, that is $R_P(h) \in \partial_\lambda \alpha_\lambda$. Note that it is not trivial to read the derivative of α_λ from the PR curve directly, but in this example this derivative is much larger than 1 and then larger than $R_P(h)$. This means that the optimal λ is far from 1.

Besides, as discussed earlier, our bound can be easily understood in terms of shared vs. separate mass between P and Q. Despite those advantages, one must keep in mind that it suffers from the same limitations when considering its practical estimation from finite i.i.d. samples and its pessimistic nature with regard to the actual regularity of the error set. That being said, Lorenz and PR curves draw several similarities with tools developed in domain adaptation, which calls for further scrutiny. For instance, the proof of Theorem 11 allows to express PR-curves as trade-off curves computed from weight ratios as defined in Ben-David and Urner (2012). Inspired by their work, it would be natural to restrict the class of "admissible" measurable sets in the PR-curves and get more useful bounds while retaining the notion of an optimal trade-off. A similar step could be taken on the dual representation by restricting the class of functions in the Lorenz diagram. In that way, one would leverage the restrictions on the hypothesis class and get similar bounds as many of those derived from Integral Probability Metrics (Redko et al., 2020). Nonetheless, given the proven ability of deep neural nets to overfit random labels (Zhang et al., 2016), leveraging classical notions of complexity is probably not sufficient to get bounds that are representative of the current domain adaptation problems. It seems inevitable to leverage some kind of "implicit bias" related to the optimization procedure. Doing so while relying on trade-off curves is an exciting research avenue: it is on-going and will certainly yield more practical bounds for domain adaptation. In the next section, we do derive a new bound going in that direction: more precisely we propose a generalization of Lorenz diagrams related to a variational form of ϕ-divergences and build upon this new notion to formulate a domain adaptation bound.

5.3 A bound based on generalized Lorenz diagrams associated with a ϕ-divergence

Nguyen et al. (2010) show that given a class of bounded measurable functions $\mathcal{F}$,

$$\begin{aligned} D_\phi(P\|Q) &= \int \phi\left(\frac{dP}{dQ}\right) dQ = \int \sup_{v\in\mathbb{R}} v\frac{dP}{dQ} - \phi^*(v)dQ \\ &\geq \sup_{f\in\mathcal{F}} \int f dP - \int \phi^*(f)dQ \end{aligned} \tag{23}$$

with equality iff $\partial\phi(\frac{dP}{dQ}) \cap \mathcal{F} \neq \emptyset$, which means that there exists $f \in \mathcal{F}$ s.t $f \in \partial\phi(\frac{dP}{dQ})$ (or equivalently thanks to duality $\frac{dP}{dQ} \in \partial\phi^*(f)$). Note that with standard conventions, $0 \in \partial\phi(1)$ and since convexity implies that subdifferential of ϕ are increasing, then the condition $f \in \partial\phi(\frac{dP}{dQ})$ imposes that $f \leq 0$ when $\frac{dP}{dQ} \leq 1$ and $f \geq 0$ otherwise.

This variational form can be used to create a kind of Lorenz diagram, associated with D_ϕ, as the set of pairs $\{(\int \phi^*(f)dP, \int f dQ)\}$. Then, the upper frontier of this region is characterizing the closeness of P and Q, and in particular $D_\phi(Q\|P)$ is the maximal vertical distance between the curve and the diagonal. For technical reasons (mainly ensuring a convex diagram), the definition of the Lorenz diagram is slightly more convoluted (see Figure 4 for an illustration of its construction[9]).

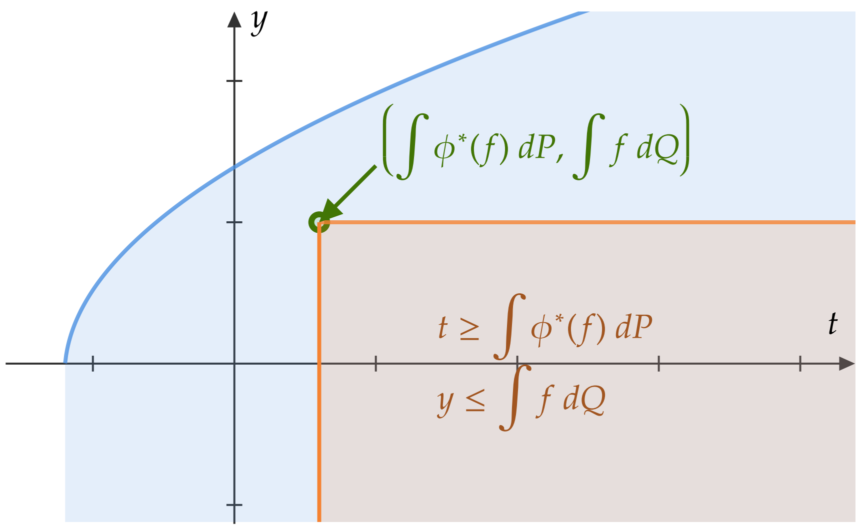

Figure 4: Generalized Lorenz diagram construction.

Definition 26 (extended Lorenz diagram) *Let $\phi \in \Phi_1$ and P, Q two distributions. The ϕ-Lorenz diagram between P and Q is defined as:*

$$LD_\phi(P,Q) = \left\{(t,y) \in \mathbb{R}^2 / \exists f : \Omega \to \mathbb{R}, t \geq \int \phi^*(f)dP, y \leq \int f dQ\right\}, \tag{24}$$

and the extended Lorenz curve is defined as its upper envelope:

$$\bar{F}_\phi(t) := \sup\{y/(t,y) \in LD_\phi(P,Q)\}. \tag{25}$$

Remark 27 *An attempt to derive a generalization bound for domain adaptation based on the variational form of ϕ-divergences has been recently published in Acuna et al. (2021). They indeed proposed a bound similar to Equation* (20) *in (Acuna et al., 2021, Lemma 1)*

9. Note that this construction can be adapted to the tighter variational formulation of D_Φ proposed in Ruderman et al. (2012) and Agrawal and Horel (2020)

for a cost function $\ell(\hat{y}; y) \in \mathrm{dom}(\phi^*)$. Denoting,

$$R_Q^\ell(h) := \int \ell(h(x); y) dQ(x, y) \tag{26}$$

their bound reads as,

$$R_Q^{\phi^* \circ \ell}(h) := \int \phi^*(\ell(h(x); y)) dQ(x, y) \leq R_P^\ell(h) + D_\phi(P \| Q) \tag{27}$$

and noting that under the standard assumption $\phi(1) = 0$ then $\phi^*(v) \geq v$, the previous bound implies that:

$$R_Q^\ell(h) \leq R_Q^{\phi^* \circ \ell}(h) \leq R_P^\ell(h) + D_\phi(P \| Q)\,. \tag{28}$$

Unfortunately, this bound is erroneous in general as can be seen by considering the family of ϕ-divergence associated with $\phi_s = s\phi$ where $s > 0$. Indeed if the bound was true, since $D_{\phi_s}(P \| Q) = sD_\phi(P \| Q)$, letting $s \to 0$, the bound would imply that $R_Q^\ell(h) \leq R_P^\ell(h)$, which is obviously wrong in general[10].

Similarly to Proposition 24, one can actually amend this bound, by using the Lorenz diagram associated to the ϕ-divergence.

Proposition 28 *Let $\phi \in \Phi_1$, $\ell(\hat{y}; y) \in \mathrm{dom}(\phi^*)$. The Lorenz curve provides the following lower bound for domain adaptation:*

$$R_Q^\ell(h) \leq R_{\lambda P}^{\phi^* \circ \ell}(h) + D_\phi(Q \| \lambda P)\,. \tag{29}$$

Proof Let us first produce a bound for the upper envelop of the Lorenz diagram. For $t \in \mathrm{dom}(\phi^*)$

$$\begin{aligned}
\bar{F}_\phi(t) &= \sup\{y \in \mathbb{R} / \exists f \in \mathrm{dom}(\phi^*), y \leq \int f dQ, t \geq \int \phi^*(f) dP\} \\
&= \sup_{f \text{ s.t. } \int \phi^*(f) dP \leq t} \int f dQ = \sup_f \inf_{\lambda \geq 0} \int f dQ - \lambda(\int \phi^*(f) dP - t) \\
&\leq \inf_{\lambda \geq 0} \sup_f \int f dQ - \lambda \left(\int \phi^*(f) dP - t \right) \\
&= \inf_{\lambda \geq 0} \lambda t + \sup_f \int f dQ - \int \phi^*(f) d(\lambda P) \\
&= \inf_{\lambda \geq 0} \lambda t + D_\phi(Q \| \lambda P)
\end{aligned} \tag{30}$$

where the inequality in the third line is trivial[11]. Note also that in the last identity, one needs to extend the variational form of D_ϕ from Nguyen et al. (2010) when applied to Q

10. The origin of the flaw is a fallacious switching between absolute values and a suppremum: $D_\phi(P \| Q) = \sup_{f \in \mathrm{dom}(\phi^*)} \int f dP - \int \phi^*(f) dQ = |\sup_{f \in \mathrm{dom}(\phi^*)} \int f dP - \int \phi^*(f) dQ| \leq \sup_{f \in \mathrm{dom}(\phi^*)} |\int f dP - \int \phi^*(f) dQ|$: the last inequality is strict in general (and in fact the RHS is often infinite), while Acuna et al. (2021) used an equality.
11. It can be interpreted in terms of primal-dual gap. And since when $t > t_{\min} := \min \phi^*(v) = -\phi(0)$, Slater's condition are easily verified (as the constant function $f = \arg\min_v \phi^*(v)$ is in the relative interior of the constraints) it is in fact an equality most of the time.

and λP (which is a positive measure but not a probability distribution in general). The demonstration of Nguyen et al. (2010) extends easily to this situation. As a result,

$$\begin{aligned} R_Q^\ell(h) := & \int \ell(h(x); y) dQ(x, y) \\ \leq & \bar{F}_\phi(R_P^{\phi^* \circ \ell}(h)) \\ \leq & \inf_{\lambda > 0} \lambda R_P^{\phi^* \circ \ell}(h) + D_\phi(Q \| \lambda P) \,. \end{aligned} \tag{31}$$

■

Remark 29 *As for the standard Lorenz diagram, this bound allows to trade the error made in the source domain for a better coverage between the target distribution Q and the non-normalized source distribution λP. When using the suboptimal choice $\lambda = 1$, one recovers a bound similar to the erroneous Equation* (28) *derived from Acuna et al. (2021):*

$$R_Q^\ell(h) \leq R_P^{\phi^* \circ \ell}(h) + D_\phi(Q \| P) \tag{32}$$

In this corrected bound, the source loss ℓ is replaced by $\phi^* \circ \ell$ which is always larger. That being said, in typical domain adaptation scenarios, ℓ would be small in average in the source domain (distribution P). Since ℓ is also positive in concrete cases, being small in average means being small with high probability under P (for instance using the Markov's inequality). Besides, since $\phi \in \Phi_1$, $0 \in \partial\phi(1)$ which by duality implies that $1 \in \partial\phi^*(0)$ and in turn shows that $\phi^*(0) = -\phi(1) = 0$ (provided $0 \in \mathrm{dom}(\phi^*)$ of course). Moreover, ϕ^* is often differentiable at 0, so its Taylor expansion is $\phi^*(v) = v + o(v)$, which implies essentially that $R_P^{\phi^* \circ \ell}(h)$ and $R_P^\ell(h)$ are similar (again if $R_P^\ell(h)$ is small).

To make such a statement more accurate, let us rely on a Taylor expansion with integral remainder. If ${\phi^*}'(0) = 1$ then one has the following Taylor expansion (see (Liese and Vajda, 2006, Theorem 1)):

$$\phi^*(v) = v + \int_0^v (v - s) d{\phi_+^*}'(s) \,.$$

In particular, if the curvature of $\phi^*(v)$ is bounded : $\forall v \in [0, 1], {\phi^*}''(v) \in [\underline{\kappa}, \bar{\kappa}]$, then one has the following estimates:

$$\frac{1}{2} \underline{\kappa} \sigma^2 \leq R_P^{\phi^* \circ \ell}(h) - R_P^\ell(h) \leq \frac{1}{2} \bar{\kappa} \sigma^2$$

where $\sigma^2 := \int \ell(h(x), y)^2 dP$ is usually quite small in concrete scenarios. Note that the previous reasoning is valid only if the curvature of ϕ^* is controlled. For instance, the counter-example based on replacing ϕ by a rescaled version $\phi_s = s\phi$ would break that assumption if $s \to 0$. Indeed, $\phi_s^*(v) = s\phi^*(\frac{v}{s})$, and then ${\phi_s^*}''(v) = \frac{1}{s}{\phi^*}''(\frac{v}{s})$.

As advocated in the domain adaptation literature Ben-David et al. (2010), to be useful, a generalization bound should :

- rely on notions that can be estimated from finite samples (this is not specific to domain adaptation)

- minimize their dependence on the distribution of labels in the target domain $Q_{Y|X}$ (especially if one aims at deriving a learning algorithm from the bound)

To do so we will rely on triangle inequalities.

Definition 30 *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function. We say that*

- $\ell \in TI$ *iff* $\forall y_a, y_b, y_c \in \mathcal{Y}, \ell(y_a; y_c) \leq \ell(y_a; y_b) + \ell(y_b; y_c)$,
- $\ell \in TI'$ *iff* $\forall y_a, y_b, y_c \in \mathcal{Y}, \ell(y_a; y_c) \leq \ell(y_a; y_b) + \ell(y_c; y_b)$.

Remark 31 *Both triangle inequalities are valid for the margin loss of Zhang et al. (2019):*

$$\ell(\hat{y}; y) = \left[1 - \frac{1}{\rho}[\delta(\hat{y}; y)]_+\right]_+ = \begin{cases} 1 & \textit{if } \arg\max(\hat{y}) \neq \arg\max(y) \\ \max(0, 1 - \frac{\delta(\hat{y};y)}{\rho}) & \textit{otherwise} \end{cases} \tag{33}$$

where $[x]_+ := \max(0, x)$ *and* $\delta(\hat{y}; y) = \frac{1}{2}\min_{k \neq k_y} \hat{y}_{k_y} - \hat{y}_k$ *is a multi-class margin between the score* $\hat{y}_{k_y}$ *attributed to the largest component of* y *(*$k_y = \arg\max_k y_k$*) and the best challenger score* $\max_{k \neq k_y} \hat{y}_k$.

The following lemma is representative of the typical ways to turn a bound such as Equation (29) into one that suits the practical purposes mentioned earlier.

Lemma 32 *Let $h, h' \in \mathcal{H}$ two hypotheses, ℓ a loss function and μ a positive measure*[12] *over pairs (x, y). Then, noting μ_X the marginal of μ w.r.t x,*

- *if* $\ell \in TI$
$$R^\ell_\mu(h) \leq R^\ell_\mu(h') + R^\ell_{\mu_X}(h; h') \tag{34}$$
where $R^\ell_{\mu_X}(h; h') = \int \ell(h(x); h'(x)) d\mu_X$.
- *Similarly, if* $\ell \in TI'$
$$R^\ell_{\mu_X}(h; h') \leq R^\ell_\mu(h) + R^\ell_\mu(h') \,. \tag{35}$$

Proof Indeed, if $\ell \in TI$

$$\begin{aligned} R^\ell_\mu(h) = &\int \ell(h(x); y) d\mu(x, y) \leq \int \ell(h(x); h'(x)) + \ell(h'(x); y) d\mu(x, y) \\ =&R^\ell_\mu(h') + R^\ell_{\mu_X}(h; h') \end{aligned} \tag{36}$$

Similarly, if $\ell \in TI'$

$$\begin{aligned} R^\ell_{\mu_X}(h; h') = &\int \ell(h(x); h'(x)) d\mu_X(x) = \int \ell(h(x); h'(x)) d\mu(x, y) \\ \leq &\int \ell(h(x); y) + \ell(h'(x); y) d\mu(x, y) \\ =&R^\ell_\mu(h) + R^\ell_\mu(h') \,. \end{aligned} \tag{37}$$

12. μ could be P or Q or even λP.

■

Following Acuna et al. (2021), and taking into account our correction to their bound based on ϕ divergences, we propose the following notion of discrepancy.

Definition 33 ($D^{\phi,\ell}_{h,\mathcal{H}}(P_X, Q_X)$) *Let P_X, Q_X two distributions over x and $\lambda > 0$. Let $\phi \in \Phi_1$, and ℓ a loss function. Let also $\mathcal{H}$ a class of hypotheses and $h \in \mathcal{H}$. Then we define,*

$$D^{\phi,\ell}_{h,\mathcal{H}}(Q_X \| \lambda P_X) := \sup_{h' \in \mathcal{H}} R^{\ell}_{Q_X}(h; h') - R^{\phi^* \circ \ell}_{\lambda P_X}(h; h') \tag{38}$$

where by extension $R^{\phi^* \circ \ell}_{\lambda P_X}(h; h') := \int \phi^*(\ell(h; h')) d\lambda P_X = \lambda R^{\phi^* \circ \ell}_{P_X}(h; h')$.

Corollary 34 *Let $\phi \in \Phi_1$. Let also $\ell \in TI$ and $\ell_\phi \in TI'$ such that $\phi^* \circ \ell \leq \ell_\phi$. Given an hypothesis $h \in \mathcal{H}$, we have the following generalization bound:*

$$R^{\ell}_{Q}(h) \leq \inf_{\lambda > 0} R^{\ell_\phi}_{\lambda P}(h) + D^{\phi,\ell}_{h,\mathcal{H}}(Q_X \| \lambda P_X) + \gamma^*_\lambda \tag{39}$$

where $\gamma^*_\lambda := \inf_{h' \in H} R^{\ell}_{Q}(h') + R^{\ell_\phi}_{\lambda P}(h')$ quantifies the adaptibility *of the task between domains P and Q.*

Proof From Lemma 32 and Definition 33, for any $h' \in \mathcal{H}$

$$\begin{aligned}
R^{\ell}_{Q}(h) \leq & R^{\ell}_{Q}(h') + \underbrace{R^{\ell}_{Q_X}(h; h')}_{\leq R^{\phi^* \circ \ell}_{\lambda P_X}(h;h') + D^{\phi,\ell}_{h,\mathcal{H}}(Q_X \| \lambda P_X)} \\
\leq & R^{\ell}_{Q}(h') + \underbrace{R^{\phi^* \circ \ell}_{\lambda P_X}(h; h')}_{\leq R^{\ell_\phi}_{\lambda P_X}(h;h') \leq R^{\ell_\phi}_{\lambda P}(h) + R^{\ell_\phi}_{\lambda P}(h')} + D^{\phi,\ell}_{h,\mathcal{H}}(Q_X \| \lambda P_X) \\
\leq & R^{\ell_\phi}_{P}(h) + D^{\phi,\ell}_{h,\mathcal{H}}(Q_X \| \lambda P_X) + (R^{\ell}_{Q}(h') + R^{\ell_\phi}_{\lambda P}(h')) \,.
\end{aligned}$$

Given that h' was chosen arbitrarily in $\mathcal{H}$ to begin with, the last term can be replaced by its infimum over $h' \in \mathcal{H}$, yielding γ^*_λ. ■

Remark 35 *The careful reader could legitimately wonder why we introduce an alternate loss ℓ_ϕ instead of using $\phi^* \circ \ell$ which appears as a natural choice. This is done to provide more flexibility on the choice of ℓ and ϕ as, when $\phi \in \Phi_1$ and $\ell \in TI$, then in general $\phi^* \circ \ell \notin TI'$. This can be verified for example for the KL-divergence and the margin loss defined in Equation* (33)*. On the contrary, since in this case $\ell(\hat{y}; y) \in [0, 1]$, then one can see that $\phi^* \circ \ell \leq \ell_\phi := \phi^*(1)\ell$, which meets the hypothesis of the theorem.*

Remark 36 *A part from the trade-off controlled by λ, the previous bound is representative of typical reasoning in domain adaptation works following the approach of Ben-David et al. (2010), as for example Ganin et al. (2016); Zhang et al. (2019); Acuna et al. (2021). Let us simplify the following argument by setting $\lambda = 1$ and note $\gamma^* = \gamma^*_1$ the adaptability term.*

This kind of bounds are in fact slightly misleading. Indeed the common algorithmic use of such bounds is the following. In the bound (and the similar ones in the literature), all terms can be efficiently evaluated from finite samples apart from γ^*. Therefore, one can consider an adversarial training scheme, where h and h' are two concurrent networks playing the following game:

$$\min_{h\in\mathcal{H}}\max_{h'\in\mathcal{H}} R_P^{\ell_\phi}(h) + R_{Q_X}^{\ell}(h;h') - R_{P_X}^{\phi^*\circ\ell}(h;h')\,.$$

This game does not lead anything useful in practical deep learning adaptation problems because usually the distributions P_X and Q_X are easily discriminated by the adversary h' whatever the choice of h. To amend this failure, previous works actually replace the game by the following one:

$$\min_{g\in\mathcal{G},h\in\mathcal{H}}\max_{h'\in\mathcal{H}} R_P^{\ell_\phi}(h\circ g) + R_{Q_X}^{\ell}(h\circ g;h'\circ g) - R_{P_X}^{\phi^*\circ\ell}(h\circ g;h'\circ g)$$

In other words, the two adversaries share a common feature extractor g which actually plays in collaboration with h and thus against h'. The rationale behind this design choice, is that one needs the ability to embed the distributions P_X and Q_X in a shared latent space $\mathcal{Z}$, where their discrimination becomes more challenging. Intuitively this embedding comes at a cost, the distributions $P_{Y|Z}$ and $Q_{Y|Z}$ can become:

- *more stochastic than their original counterpart $P_{Y|X}$ and $Q_{Y|X}$*
- *worse, they can become more inconsistent in the sense of the adaptability constant γ^**

Although this aspect was considered in Johansson et al. (2019); Zhao et al. (2019); Siry et al. (2021), to the best of our knowledge the impact of this modification was not analyzed carefully in terms of generalization bounds. In fact, to do so, one needs to amend the generalization bound as follows: $\forall g\in\mathcal{G},\forall h\in\mathcal{H}$

$$R_Q^{\ell}(h\circ g) \leq R_P^{\ell_\phi}(h\circ g) + D_{h\circ g,\mathcal{H}\circ g}^{\phi,\ell}(Q_X,P_X) + \gamma_g^* \tag{40}$$

*where we have noted $\mathcal{H}\circ g = \{h\circ g / h\in\mathcal{H}\}$ and $\gamma_g^* = \inf_{h'\in\mathcal{H}} R_Q^{\ell}(h'\circ g) + R_P^{\ell_\phi}(h'\circ g)$. The main issue is that γ_g^*, besides not being computable in practice, depends on the specific choice of the feature extractor g. It is not anymore an intrinsic measure of the adaptability of the transfer. This issue can be amplified by the fact that g is playing with h and as a result, it may tend to be biased towards performing a correct embedding of high probability areas of the source distribution and performing in an uncontrolled way on the target distribution high likelihood areas. A similar argument could of course be held against the trade-off parameter λ, but there is one subtle difference: it merely impacts the weight granted to the source distribution P. It is therefore easier to appraise its impact as well as to control it through regularization (*e.g.* favoring a smaller λ). Note that our λ variable acts in a similar fashion to the margin parameter[13] of Zhang et al. (2019) in the sense that it impacts the weight of the source risk. On the other hand, the margin parameter of Zhang et al. (2019) is introduced artificially and does not act on the discrepancy term.*

13. Beware, in Zhang et al. (2019), this parameter is denoted by γ and the adaptability term by λ which could bring confusion.

5.3.1 Practical considerations

The result presented in Corollary 34 leaves quite some room for the choice of ϕ, ℓ and ℓ_ϕ. We believe that the combined ϕ, ℓ, ℓ_ϕ configuration should follow (or remain close to) the following design principles.

(P1) $\ell \in TI$, $\ell_\phi \in TI'$ and $\phi^* \circ \ell \leq \ell_\phi$ (to ensure that our bound is valid).

(P2) ℓ should be Bayes consistent (see Steinwart (2007); Tewari and Bartlett (2007)).

(P3) ℓ_ϕ is convex and bounded below (useful from the optimization perspective).

For example, if one follows in the steps of Zhang et al. (2019) and choose for ℓ the margin loss from Equation (33) (with *e.g.* $\rho = 1$), since $\mathrm{range}(\ell) = [0, 1]$, using $\ell_\phi = \phi^*(1)\ell$, then (P1) is valid. In addition, one can show that (P2) is checked in the binary classification case, because then $\ell(\hat{y}; y) = (1 - (\hat{y}y)_+)_+$ is a piecewise linear function of the margin $\hat{y}y$ and one can compute its ψ-transform as defined in Bartlett et al. (2006) and show[14] that $\psi_\ell(\epsilon) = \epsilon$ which being invertible ensures that ℓ is Bayes consistent. Unfortunately, the extension to multi-category classification breaks the Bayes consistency (see (Tewari and Bartlett, 2007, example 1) with respect to the Crammer and Singer methodology). That being said, ℓ is an upper-bound of the $0 - 1$ loss[15], and one still enjoy the fact that controlling the ℓ-risk ensures a small $0 - 1$-risk. Last, ℓ_ϕ is bounded below but not convex. That is not a real issue given that one can replace it by a larger convex loss such as the hinge loss.

In contrast to the previous list of guiding principles, Acuna et al. (2021) merely requires that $\mathrm{range}(\ell) \subset [0, 1] \subset \mathrm{dom}(\phi^*)$. One may fail to see how this assumption is relevant, but it surely comes with some drawbacks. Mainly, $[0, 1] \cap \mathrm{dom}(\phi^*)$ always misses important parts of the domain of ϕ^*. To illustrate this point, let us consider (Acuna et al., 2021, Proposition 1), where to deepen the link between the adversarial game and the ϕ-divergence, it is required that $\exists h'$ s.t. $\forall x, \ell(h(x), h'(x)) \in \partial\phi(\frac{dP}{dQ}(x))$. Meeting this condition is infeasible if $\mathrm{range}(\ell) \subset [0, 1]$ because $\partial\phi(\frac{dP}{dQ}(x)) \subset]-\infty, 0]$ whenever $\frac{dP}{dQ}(x) < 1$. In other words, the assumption is only met when $P \geq Q$ which actually implies $P = Q$ and is of no interest since it corresponds to the absence of domain shift. Note however that despite our general principles, the only practical settings that we consider, that is to say the margin loss, suffers the same limitation since in this particular configuration, $\mathrm{range}(\ell) \subset [0, 1]$. We foresee, that this issue can be solved by refining our definition of the generalized Lorenz diagrams, using the tighter variational representation proposed in Ruderman et al. (2012) and Agrawal and Horel (2020). This representation takes the following form:

$$D_\phi(Q\|\lambda P) = \sup_f \sup_{\rho \in \mathbb{R}} \int f dQ - \int (\phi^*(f + \rho) - \rho) d(\lambda P)\,.$$

As can be seen, in this formulation the scalar $\rho \in \mathbb{R}$ allows to align the range of $f + \rho$ and the domain of ϕ^* that is to say the subdifferentials of ϕ. Besides, in the advent of deriving an adversarial algorithm from this refined variant, the variable ρ would naturally play a

14. In fact, using the notations of Bartlett et al. (2006), one can show that $H_\ell(\eta) = \min(1 - \eta, \eta)$ and $H_\ell^-(\eta) = \max(1 - \eta, \eta)$. This yields $\tilde{\psi}_\ell(\eta) = H_\ell^- - H_\ell = |2\eta - 1| > 0 \forall \eta \neq \frac{1}{2}$, yielding the Bayes consistency.

15. More precisely $\ell(\hat{y}; y) \geq \mathbb{1}_{\arg\max(\hat{y}_k) \neq \arg\max(y_k)}$

balancing role with respect to the variable λ. This extension has already been adopted for GANs (Terjék, 2021) which suggests that it is appropriate for domain adaptation adversarial approaches. It is however left for future work.

6. Conclusion

In this work, we have studied the interconnections among several trade-off curves designed to evaluate the similarity between two probability distributions, namely precision-recall curves, divergence frontiers, ROC and Lorenz curves. If one connection was known to the authors of divergence frontiers, others appear to have eluded even the authors of the implied notions. This is particularly striking for Lorenz and ROC curves which differ by mere symmetries. The interrelation between precision-recall and Lorenz curves is less direct, as it involves convex duality. That being said, it remains that the two notions are theoretically equivalent, and can be computed in practice from one another. We hope that the exposed link will foster new research avenues for evaluation curves. To begin with, while the theoretical equivalence of Lorenz and PR curves has been demonstrated, the question of their empirical estimation has yet to be examined. For instance, investigations on potentially consistent estimators need consideration, especially in the non parametric case (although this will require refining the definitions to render the estimation from finite sample feasible). In particular, exposing rates of convergence of the estimators should be a worthwhile endeavor. Similar analysis has already been carried out for scalar metrics Rubenstein et al. (2019); Sriperumbudur et al. (2009). We foresee that exploring links with divergences or integral probability metrics shall help in pursuing this undertaking for PR and Lorenz curves. Instead of considering links with IPM, we have considered the case of ϕ-divergence which have provided a fertile ground to extend the already exposed links. As a result, we have demonstrated that a general link exists between Rényi divergence frontiers of arbitrary exponent a and PR curves. Last, we have explored applications of trade-off curves to domain adaptation. More notably, we have extended the notion of Lorenz diagram, and built upon this notion to derive a novel generalization bound for domain adaptation. Several research avenues are plausible to deepen this aspect of our work.

Acknowledgments

This work is partially supported by the project ANR-19-CHIA-0017. The authors would like to thank Jalal Fadili for his helpful comments. They are also grateful to the reviewers for very constructive comments.

Appendix A. Proof of lemma 10

Proof The proof is technical and can be skipped on first reading. The second result is a simple corollary of the first since when μ is not absolutely continuous w.r.t ν then the identity is trivial: $\infty = \infty$. Let us demonstrate the first point, that is to say when $\mu \ll \nu$. We shall proceed by proving two opposite inequalities, starting from the following.

$$\sup_{A \in \mathcal{A}} \frac{\mu(A)}{\nu(A)} \leq \operatorname*{ess\,sup}_{d\mu} \frac{d\mu}{d\nu}.$$

Indeed, let $A \in \mathcal{A}$, then,

$$\begin{aligned} \frac{\mu(A)}{\nu(A)} &= \frac{\int_A \frac{d\mu}{d\nu} d\nu}{\nu(A)} = \frac{\int 1_A \frac{d\mu}{d\nu} d\nu}{\nu(A)} \\ &\leq \operatorname*{ess\,sup}_{d\nu} \frac{d\mu}{d\nu} \frac{\nu(A)}{\nu(A)} = \operatorname*{ess\,sup}_{d\nu} \frac{d\mu}{d\nu}\,. \end{aligned}$$

Note that the essential sup is w.r.t ν instead of μ. Let us then show that $\operatorname{ess\,sup}_{d\nu} \frac{d\mu}{d\nu} \leq \operatorname{ess\,sup}_{d\mu} \frac{d\mu}{d\nu}$. To do so we need to show that any upper-bound M of $\frac{d\mu}{d\nu}$ μ-a.e. is also an upper-bound ν-a.e. Let $M \geq \frac{d\mu}{d\nu}$ μ-a.e. be such an upperbound and let N the associated μ-nullset (where M may be lesser than $\frac{d\mu}{d\nu}$). If $\nu(N) = 0$ it settles it (as M is henceforth an upper-bound ν-a.e.). Otherwise, N is such that $\mu(N) = 0$ but $\nu(N) > 0$ and then necessarily $\frac{d\mu}{d\nu}\mathbb{1}_N = 0$ ν-a.e. (or else it would contradict $\mu(N) = 0$). As a result,

$$\frac{d\mu}{d\nu} \overset{\nu\text{-a.e.}}{=} \frac{d\mu}{d\nu}\mathbb{1}_{\Omega\setminus N} \leq M\mathbb{1}_{\Omega\setminus N} \leq M \ \nu\text{-a.e.}$$

Which again settles the fact that M is also an upperbound ν-a.e.. Therefore we obtain that $\operatorname{ess\,sup}_{d\nu} \frac{d\mu}{d\nu} \leq \operatorname{ess\,sup}_{d\mu} \frac{d\mu}{d\nu}$ and $\frac{\mu(A)}{\nu(A)} \leq \operatorname{ess\,sup}_{d\mu} \frac{d\mu}{d\nu}$.

For the reverse inequality, we only need to show that $\frac{d\mu}{d\nu} \leq \sup_{A\in\mathcal{A}} \frac{\mu(A)}{\nu(A)}$ μ-a.e.. Indeed, let $C := \sup_{A\in\mathcal{A}} \frac{\mu(A)}{\nu(A)}$. If $C = \infty$, the inequality is trivial. Let us then suppose that $C < \infty$. Let $E := \{\omega \in \Omega / \frac{d\mu}{d\nu} > C\}$ and let us show that $\mu(E) = 0$. One can rewrite $E = \cup_{n\in\mathbb{N}} E_n$, with $E_n := \{\omega \in \Omega / \frac{d\mu}{d\nu} \geq C + \frac{1}{n}\}$, and it suffices to show that $\mu(E_n) = 0$. We have that,

$$\begin{aligned} C\nu(E_n) \geq& \mu(E_n) = \int_{E_n} \frac{d\mu}{d\nu} d\nu \geq (C + \tfrac{1}{n}) \int_{E_n} d\nu \\ \geq& (C + \tfrac{1}{n})\nu(E_n)\,. \end{aligned}$$

For this not to be absurd, it is necessary that $\nu(E_n) = 0$ and hence $\mu(E_n) = 0$ as well. ■

Appendix B. Proof of Lemma 12

We will use the following result from Theorem 5 in Simon et al. (2019) which stipulates that the precision α_λ is optimal in the following sense:

Theorem 37 *Let P, Q two distributions from $\mathcal{M}_p(\Omega)$. Then $\forall \lambda \in \overline{\mathbb{R}^+}$,*

$$\alpha_\lambda = \min_{A\in\mathcal{A}} \lambda(1 - P(A)) + Q(A)\,. \tag{41}$$

From this theorem the lemma is an easy corollary:

Proof It suffices to show that for all measurable functions $0 \leq f \leq 1$,

$$\alpha_\lambda \leq \lambda(1 - \int f dP) + \int f dQ.$$

According to Theorem 37, this inequality holds for indicator functions. Besides it is stable under convex combinations and L^∞ limits. As a result, we can extend the inequality first to convex combinations of indicators, which is to say to all simple functions ranging in $[0, 1]$. Note that even though at first glance, simple functions ranging in $[0, 1]$ take the form of sub-convex combinations of indicators, one can leverage $\mathbb{1}_\emptyset = 0$ to express them as convex combinations. Then, taking the limit w.r.t to L^∞ convergence and using standard density results we can further extend the inequality to L^∞ functions ranging in $[0, 1]$. ■

References

David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pages 66–75. PMLR, 2021.

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. In *International Conference on Machine Learning*, pages 115–124. PMLR, 2020.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

Shun-ichi Amari. Integration of stochastic models by minimizing α-divergence. *Neural computation*, 19(10):2780–2796, 2007.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153. Springer, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Imre Csiszár. An information-theoretic inequality and its application to proofs of the ergodicity of markoff chains. *Magyar Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Morris H DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. Evaluating generative models using divergence frontiers. *arXiv preprint arXiv:1905.10768*, 2019.

Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. Precision-recall curves using information divergence frontiers. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2020.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Peter Harremoës. A new look on majorization. *Proceedings ISITA 2004, Parma, Italy*, pages 1422–1425, 2004.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Daniel Jiwoong Im, He Ma, Graham Taylor, and Kristin Branson. Quantitatively evaluating gans with divergences proposed for training. *arXiv preprint arXiv:1803.01045*, 2018.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.

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

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, pages 3929–3938, 2019.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Zinan Lin, Ashish Khetan, Giulia C. Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. *CoRR*, abs/1712.04086, 2017. URL http://arxiv.org/abs/1712.04086.

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1498–1507, 2018.

Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Divergence frontiers for generative models: Sample complexity, quantization effects, and frontier integrals. *Advances in Neural Information Processing Systems*, 34, 2021.

Max O Lorenz. Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219, 1905.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. *arXiv preprint arXiv:2002.09797*, 2020.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Benedetto Piccoli, Francesco Rossi, and Magali Tournus. A wasserstein norm for signed measures, with application to nonlocal transport equation with source term. *arXiv preprint arXiv:1910.05105*, 2019.

Barnabás Póczos and Jeff Schneider. On the estimation of alpha-divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 609–617. JMLR Workshop and Conference Proceedings, 2011.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory. *arXiv preprint arXiv:2004.11829*, 2020.

Paul Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O Tolstikhin. Practical and consistent estimation of f-divergences. In *Advances in Neural Information Processing Systems*, pages 4072–4082, 2019.

Avraham Ruderman, Mark Reid, Darío García-García, and James Petterson. Tighter variational representations of f-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664*, 2012.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5234–5243. Curran Associates, Inc., 2018.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory*, pages 489–511. PMLR, 2013.

Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5799–5808, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Rodrigue Siry, Louis Hémadou, Loïc Simon, and Frédéric Jurie. On the inductive biases of deep domain adaptation. *arXiv preprint arXiv:2109.07920*, 2021.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Dávid Terjék. Moreau-yosida f-divergences. In *International Conference on Machine Learning*, pages 10214–10224. PMLR, 2021.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(5), 2007.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

Tim van Erven and Peter Harremoës. Rényi divergence and majorization. In *2010 IEEE International Symposium on Information Theory*, pages 1335–1339. IEEE, 2010.

Alexandre Verine, Benjamin Negrevergne, Muni Sreenivas Pydi, and Yann Chevaleyre. Training normalizing flows with the precision-recall divergence. *arXiv preprint arXiv:2302.00628*, 2023.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. *arXiv preprint arXiv:1912.11035*, 2019.

Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413, 2019.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.