A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions

Guillaume Staerman[1] **Pavlo Mozharovskyi**[1] **Pierre Colombo**[1,2]

Stéphan Clémencon[1] **Florence d'Alché-Buc**[1]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris

[2] IBM GBS France

Abstract

The design of a metric between probability distributions is a longstanding problem motivated by numerous applications in machine learning. Focusing on continuous probability distributions in the Euclidean space $\mathbb{R}^d$, we introduce a novel pseudo-metric between probability distributions by leveraging the extension of univariate quantiles to multivariate spaces. Data depth is a nonparametric statistical tool that measures the centrality of any element $x \in \mathbb{R}^d$ with respect to (w.r.t.) a probability distribution or a dataset. It is a natural median-oriented extension of the cumulative distribution function (cdf) to the multivariate case. Thus, its upper-level sets—the depth-trimmed regions—give rise to a definition of multivariate quantiles. The new pseudo-metric relies on the average of the Hausdorff distance between the depth-based quantile regions w.r.t. each distribution. Its good behavior w.r.t. major transformation groups, as well as its ability to factor out translations, are depicted. Robustness, an appealing feature of this pseudo-metric, is studied through the finite sample breakdown point. Moreover, we propose an efficient approximation method with linear time complexity w.r.t. the size of the dataset and its dimension. The quality of this approximation, as well as the performance of the proposed approach, are illustrated in numerical experiments.

1 INTRODUCTION

Metrics or pseudo-metrics between probability distributions have attracted a long-standing interest in information theory (Kullback, 1959; Rényi, 1961; Csiszàr, 1963; Stummer and Vajda, 2012), probability theory and statistics (Billingsley, 1999; Sriperumbudur et al., 2012; Panaretos and Zemel, 2019; Rachev, 1991). While they serve many purposes in machine learning (Cha and Srihari, 2002; MacKay, 2003), they are of crucial importance in automatic evaluation of natural language generation (see e.g. Kusner et al., 2015; Zhang et al., 2019), especially when leveraging deep contextualized embeddings such as the popular BERT (Devlin et al., 2018). Yet designing a measure to compare two probability distributions is a challenging research field. This is certainly due to the inherent difficulty in capturing in a single measure typical desired properties such as: (i) metric or pseudo metric properties, (ii) invariance under specific geometric transformations, (iii) efficient computation, and (iv) robustness to contamination.

One can find in the literature a vast collection of discrepancies between probability distributions that rely on different principles. The f-divergences (Csiszàr, 1963) are defined as the weighted average by a well-chosen function f of the odds ratio between the two distributions. They are widely used in statistical inference but they are by design ill-defined when the supports of both distributions do not overlap, which appears to be a significant limitation in many applications. IPMs (Sriperumbudur et al., 2012) are based on a variational definition of the metric, i.e. the maximum difference in expectation for both distributions calculated over a class of measurable functions and give rise to various metrics (Maximum Mean Discrepancy (MMD), Dudley's metric, L_1-Wassertein Distance) depending on the choice of this class. However, except in the case of MMD, which appears to enjoy a closed-form solution, the variational definition raises issues in computation. From the side of Optimal transport (OT) (see Villani, 2003; Peyré and Cuturi, 2019), the L_p-Wasserstein distance is based on a

ground metric able to take into account the geometry of the space on which the distributions are defined. Its ability to handle non-overlapping support and appealing theoretical properties make OT a powerful tool, mainly when applied to generative models (Arjovsky et al., 2017) or automatic text evaluation (Zhao et al., 2019; Clark et al., 2019; Colombo et al., 2021a).

In this work, we adopt another angle. Focusing on continuous probability distributions in the Euclidean space $\mathbb{R}^d$, we propose to consider a new metric between probability distributions by leveraging the extension of univariate quantiles to multivariate spaces. The notion of quantile function is an interesting ground to build a comparison between two probability measures as illustrated by the closed-form of the Wasserstein distance defined over $\mathbb{R}$. However, given the lack of natural ordering on $\mathbb{R}^d$ as soon as $d > 1$, extending the concept of univariate quantiles to the multivariate case raises a real challenge. Many extensions have been proposed in the literature, such as minimum volume sets (Einhmahl and Mason, 1992), spatial quantiles (Koltchinskii and Dudley, 1996) or data depth (Tukey, 1975). The latter offers different ways of ordering multivariate data w.r.t a probability distribution. Precisely, *data depths* are non parametric statistics that determine the centrality of any element $x \in \mathbb{R}^d$ w.r.t. a probability measure. They provide a multivariate ordering based on topological properties of the distribution, allowing it to be characterized by its location, scale or shape (see e.g. Mosler, 2013 or Staerman, 2022 for a review). Several data depths were subsequently proposed such as convex hull peeling depth (Barnett, 1976), simplicial depth (Liu, 1990), Oja depth (Oja, 1983) or zonoid depth (Koshevoy and Mosler, 1997) differing in their properties and applications. With a substantial body of literature devoted to its computation, recent advances allow for fast exact (Pokotylo et al., 2019) and approximate (Dyckerhoff et al., 2021) computation of several depth notions. The desirable properties of data depth, such as affine invariance, continuity w.r.t. its arguments, and robustness (Zuo and Serfling, 2000) make it an important tool in many fields. Today, in its variety of notions and applications, data depth constitutes a versatile methodology (Mosler and Mozharovskyi, 2021) that has been successfully employed in a variety of machine learning tasks such as regression (Rousseeuw and Hubert, 1999; Hallin et al., 2010), classification (Li et al., 2012; Lange et al., 2014), anomaly detection (Serfling, 2006; Rousseeuw and Hubert, 2018; Staerman et al., 2019, 2020, 2022) and clustering (Jörnsten, 2004).

This paper presents a new discrepancy measure between probability distributions, well-defined for non-overlapping supports, that leverages the interesting features of data depths. This measure is studied through the lens of the previously stated properties, yielding the contributions listed below.

Contributions:

- A new discrepancy measure between probability distributions involving the upper-level sets of data depth is introduced. We show that this measure defines a pseudo-metric in general. Its good behavior regarding major transformation groups, as well as its ability to factor out translations, are depicted. Its robustness is investigated through the concept of finite sample breakdown point.
- An efficient approximation of the depth-trimmed regions-based pseudo-metric is proposed for convex depth functions such as halfspace, projection and integrated rank-weighted depths. This approximation relies on a nice feature of the Hausdorff distance when computed between convex bodies.
- The behavior of this algorithm regarding its parameters is studied through numerical experiments, which also highlight the by-design robustness of the depth-trimmed regions based pseudo-metric. Applications to robust clustering of images and automatic evaluation of natural language generation (NLG) show the benefits of this approach when benchmarked with state-of-the-art probability metrics.

2 BACKGROUND ON DATA DEPTH

In this section, we recall the concept of statistical *data depth* function and its attractive theoretical properties for clarity. Here and throughout, the space of all continuous probability measures on $\mathbb{R}^d$ with $d \in \mathbb{N}^*$ is denoted by $\mathcal{M}_1(\mathbb{R}^d)$. By $g_\sharp$ we denote the push-forward operator of the function g. Introduced by Tukey (1975), the concept of data depth initially extends the notion of median to the multivariate setting. In other words, it measures the centrality of any element $x \in \mathbb{R}^d$ w.r.t. a probability distribution (respectively, a dataset). Formally, a data depth is defined as follows:

$$D : \begin{array}{ccc} \mathbb{R}^d \times \mathcal{M}_1(\mathbb{R}^d) & \longrightarrow & [0,1]\,, \\ (x,\ \rho) & \longmapsto & D(x,\ \rho). \end{array} \tag{1}$$

We denote by $D(x, \rho)$ (or $D_\rho(x)$ for brevity) the depth of $x \in \mathbb{R}^d$ w.r.t. $\rho \in \mathcal{M}_1(\mathbb{R}^d)$. The higher $D(x, \rho)$, the deeper it is in ρ. The depth-induced median of ρ is then defined by the set attaining $\sup_{x \in \mathbb{R}^d} D(x, \rho)$. Since data depth naturally and in a nonparametric way defines a pre-order on $\mathbb{R}^d$ w.r.t. a probability distribution, it can be seen as a centrality-based alternative to the cumulative distribution function (cdf) for multivariate data. For any $\alpha \in [0, 1]$, the associated α-depth region of a depth function is defined as its upper-level set:

$$D_\rho^\alpha = \left\{x \in \mathbb{R}^d,\ D_\rho(x) \geq \alpha\right\}.$$

It follows that depth regions are nested, i.e. $D_\rho^{\alpha'} \subseteq D_\rho^\alpha$ for any $\alpha < \alpha'$. These depth regions generalize the notion of quantiles to a multivariate distribution.

A depth function's relevance to capturing information about a distribution relies on the statistical properties it satisfies. Such properties have been thoroughly investigated in Liu (1990); Zuo and Serfling (2000) and Dyckerhoff (2004) with slightly different sets of axioms (or postulates) to be satisfied by a proper depth function. In this paper, we restrict to *convex depth functions* (Dyckerhoff, 2004) mainly motivated by recent algorithmic developments including theoretical results (Nagy et al., 2020) as well as implementation guidelines (Dyckerhoff et al., 2021).

The general formulation (1) opens the door to various possible definitions. While these differ in theoretical and practically related properties such as robustness or computational complexity (see Mosler and Mozharovskyi, 2021 for a detailed discussion), several postulates have been developed throughout the recent decades the "good" depth function should satisfy. Formally, a function D is called a *convex depth function* if it satisfies the following postulates:

D1 (AFFINE INVARIANCE) $D(g(x), g_\sharp\rho) = D(x, \rho)$ holds for $g : x \in \mathbb{R}^d \mapsto Ax+b$ with any non-singular matrix $A \in \mathbb{R}^{d\times d}$ and any vector $b \in \mathbb{R}^d$.

D2 (VANISHING AT INFINITY) $\lim_{||x||\to\infty} D_\rho(x) = 0$.

D3 (UPPER SEMICONTINUITY) $\{x \in \mathbb{R}^d \;\; D_\rho(x) < \alpha\}$ is an open set for every $\alpha \in (0, 1]$.

D4 (QUASICONCAVITY) For every $\lambda \in [0, 1]$ and $x, y \in \mathbb{R}^d$, $D_\rho(\lambda x + (1 - \lambda)y) \geq \min\{D_\rho(x), D_\rho(y)\}$.

While (**D1**) is useful in applications providing independence w.r.t. measurement units and coordinate system, (**D2**) and (**D3**) appear as natural properties since data depth is a (center-outward) generalization of cdf. Limit values vanish due to median-oriented construction. (**D4**) allows to preserve the original center-outward ordering goal of data depth and induces convexity of the depth regions. Furthermore, it is easy to see that (**D1–D4**) respectively yield properties of affine equivariance, boundedness, closedness and convexity of the central regions D_ρ^α (Dyckerhoff, 2004). Thanks to (**D2–D4**), if $\alpha > 0$, non-empty regions associated to convex depth functions are convex bodies (compact convex set in $\mathbb{R}^d$).

Below we recall three convex depth functions satisfying (**D1–D4**) that will be used throughout the paper: halfspace depth (Tukey, 1975), which is probably the most studied in the literature, projection depth (Liu, 1992), and the (affine-invariant) integrated rank-weighted (AI-IRW) depth (Ramsay et al., 2019; Staerman et al., 2021b). For this, let $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$ and X a random variable defined on a certain probability space $(\Omega, \mathcal{A}, \mathbb{P})$ that takes values in $\mathcal{X} \subset \mathbb{R}^d$ following distribution ρ. The halfspace depth of a given $x \in \mathbb{R}^d$ w.r.t. ρ is defined as the smallest probability mass that can be contained in a closed halfspace containing x:

$$HD_\rho(x) = \inf_{u\in\mathbb{S}^{d-1}} \mathbb{P}\left(\langle u, X\rangle \leq \langle u, x\rangle\right).$$

Projection depth, being a monotone transform of the Stahel-Donoho outlyingness (Donoho and Gasko, 1992; Stahel, 1981), is defined as follows:

$$PD_\rho(x) = \left(1 + \sup_{u\in\mathbb{S}^{d-1}} \frac{|\langle u, x\rangle - \text{med}(\langle u, X\rangle)|}{\text{MAD}(\langle u, X\rangle)}\right)^{-1},$$

where med and MAD stand for the univariate median and median absolute deviation from the median, respectively. The affine-invariant integrated rank-weighted (AI-IRW) depth of a given $x \in \mathbb{R}^d$ relative to a square-integrable random vector X with probability distribution ρ on $\mathbb{R}^d$ and positive definite variance-covariance matrix Σ, named $AD_\rho(x)$, is given by:

$$\mathbb{E}\left[\min\left\{\mathbb{P}\left(\langle V, X\rangle \leq \langle V, x\rangle\right), \mathbb{P}\left(\langle V, X\rangle > \langle V, x\rangle\right)\right\}\right], \tag{2}$$

where $V = \Sigma^{-\top/2}U/||\Sigma^{-\top/2}U||$ and U is uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.

Remark 2.1. *Data depth functions have connections with the density function in particular cases. Indeed, for elliptical distributions, the level sets of any data depth satisfying (**D1–D4**) are concentric ellipsoids with the same center, and orientation as the density level sets (Liu and Singh, 1993). The density is a local measure assigning the score of an element as the probability mass in an infinitesimal neighborhood. In contrast, data depths are global measures of ordering taking into account the whole distribution to assign a score to an element and are thus not equivalent to the density for general distributions. However, they provide interesting alternatives in many applications, such as anomaly detection (see e.g. Staerman et al., 2021b). For example, the density will assign a zero score to every $x \in \mathbb{R}^d$ far from a concentrated group of observations regardless of the distance. At the same time, the projection depth described above will be able to rank these "outliers" depending on how it moves away from them.*

3 A PSEUDO-METRIC BASED ON DEPTH-TRIMMED REGIONS

In this section, we introduce the depth-based pseudo-metric and study its properties. We consider depth regions pos-

sessing the same probability mass to compare depth regions from different probability distributions fairly. Following Paindaveine and bever (2013), we denote by $\alpha : (\beta, \rho) \in [0,1] \times \mathcal{M}_1(\mathbb{R}^d) \longmapsto \alpha(\beta, \rho) \in [0,1]$ the highest level such that the probability mass of the depth-trimmed region at this level is at least β. Precisely, for any pair $(\beta, \rho) \in [0,1) \times \mathcal{M}_1(\mathbb{R}^d)$:

$$\alpha(\beta, \rho) = \sup\{\gamma \in [0,1] : \quad \rho\left(D_\rho^\gamma\right) > \beta\}. \tag{3}$$

In the remainder of this paper, when the quantity $\alpha(\beta, \rho)$ will be associated with depth regions of ρ, the second argument of the function $\alpha(\cdot, \cdot)$ will be omitted, for notation simplicity. It is worth mentioning that $D_\rho^{\alpha(\beta')} \subseteq D_\rho^{\alpha(\beta)}$ for any $\beta > \beta'$, since $\beta \mapsto \alpha(\beta, \rho)$ is a monotone decreasing function. Thus, $D_\rho^{\alpha(\beta)}$ is the smallest depth region with probability larger than or equal to β and can be defined in an identical way as:

$$D_\rho^{\alpha(\beta)} = \bigcap_{\gamma \in \Gamma_\rho(\beta)} D_\rho^\gamma,$$

where $\Gamma_\rho(\beta) = \{\zeta \in [0,1] : \quad \rho\left(D_\rho^\zeta\right) > \beta\}$. The strict inequalities in (3) and in the definition of $\Gamma_\rho(\beta)$ eliminate cases where the supremum does not exist. Indeed, when $\beta = 0$, the depth region is then an infinitesimal set with a probability higher than zero. To the best of our knowledge, the supremum exists (without necessarily being unique) in the case of the halfspace depth (Rousseeuw and Rutz, 1999) and the projection depth (Zuo, 2003) under mild assumptions. Still, no results have been derived for AI-IRW depth yet. The set $\{D_\rho^{\alpha(\beta)},\ \beta \in [0, 1-\varepsilon],\ \varepsilon \in (0,1]\}$ where each region probability mass is equal to β then defines quantile regions of ρ.

Let μ, ν be two absolutely continuous probability measures (w.r.t. the Lebesgue measure) on $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ respectively. Denote by $d_{\mathcal{H}}(A, B)$ the Hausdorff distance between the sets A and B. The pseudo-metric between probability distributions μ and ν based on the depth-trimmed regions is defined as follows:

Definition 3.1. *Let $\varepsilon \in (0,1]$ and $p \in (0, \infty)$, for all pairs (μ, ν) in $\mathcal{M}_1(\mathcal{X}) \times \mathcal{M}_1(\mathcal{Y})$, the depth-trimmed regions ($DR_{p,\varepsilon}$) discrepancy measure between μ and ν is defined as*

$$DR_{p,\varepsilon}^p(\mu, \nu) = \int_0^{1-\varepsilon} d_{\mathcal{H}}\left(D_\mu^{\alpha(\beta)}, D_\nu^{\alpha(\beta)}\right)^p \, \mathrm{d}\beta. \tag{4}$$

Our discrepancy measure relies on the Hausdorff distance averaged over depth-trimmed regions with the same probability mass w.r.t. each distribution. Properties (**D2–D3**) ensure that for every $0 \leq \beta < 1$, $D_\mu^{\alpha(\beta)}$ is a non-empty compact subset of $\mathbb{R}^d$ leading to a well-defined discrepancy measure. Observe that the parameter ε can be considered as a robustness tuning parameter. Indeed, choosing higher ε amounts to ignore the larger upper-level sets of data depth function, i.e. the tails of the distributions.

Remark 3.2. *Data depths provide robustness to (4) together with the ε-trimming. Indeed, data depths such as the three previously introduced in Section 2 exhibit attractive robustness properties. The asymptotic breakdown point of the halfspace and the integrated rank-weighted medians are higher than $1/(d+1)$. In contrast, the projection median is known to have a breakdown point equal to $1/2$ (Donoho and Gasko, 1992; Ramsay et al., 2019).*

Remark 3.3. *When $d = 1$, the L_p-Wasserstein distance enjoys an explicit expression involving quantile and distribution functions. Let $X^1 \sim \mu_1$, $Y^1 \sim \nu_1$ be two random variables where μ_1, ν_1 are univariate probability distributions. Denoting by $F_{X^1}^{-1}$ the quantile function of X^1, the L_p-Wasserstein distance can be written as*

$$W_p^p(\mu_1, \nu_1) = \int_0^1 |F_{X^1}^{-1}(q) - F_{Y^1}^{-1}(q)|^p \, \mathrm{d}q. \tag{5}$$

Since data depth and its central regions are extensions of cdf and quantiles to dimension $d > 1$, $DR_{p,\varepsilon}$ is then a possible (center-outward) generalization of (5) to higher dimensions. When $DR_{p,\varepsilon}$ is associated with the halfspace depth, a simple calculus (see Lemma A.3 in the Appendix for mathematical details) leads to

$$DR_{p,\varepsilon}^p(\mu_1, \nu_1) = 2 \int_{\varepsilon/2}^{1/2} \max\Big\{|F_{X^1}^{-1}(q) - F_{Y^1}^{-1}(q)|^p, \\ |F_{X^1}^{-1}(1-q) - F_{Y^1}^{-1}(1-q)|^p\Big\} \, \mathrm{d}q.$$

Thus, $W_p^p(\mu_1, \nu_1) \leq \lim_{\varepsilon \to 0} DR_{p,\varepsilon}^p(\mu_1, \nu_1)$ in general where the equality holds for symmetric distributions.

3.1 Metric properties

We now investigate to which extent the proposed discrepancy measure satisfies the metric axioms. As a first go, we show that $DR_{p,\varepsilon}$ fulfills most conditions. However, it does not define distance in general.

Proposition 3.4 (METRIC PROPERTIES)**.** *For any convex data depth, $DR_{p,\varepsilon}$ is positive, symmetric and satisfies triangular inequality but the entailment $DR_{p,\varepsilon}(\mu, \nu) = 0 \implies \mu = \nu$ does not hold in general.*

Thus, $DR_{p,\varepsilon}$ defines a pseudo-metric rather than a distance. Based on distance, the proposed discrepancy measure preserves isometry invariance, as stated in the following proposition.

Proposition 3.5 (ISOMETRY INVARIANCE)**.** *Let $A \in \mathbb{R}^{d \times d}$ be a non-singular matrix and $b \in \mathbb{R}^d$. Define the isometry*

mapping $g : x \in \mathbb{R}^d \mapsto Ax + b$ *with* $AA^\top = I_d$*, then it holds:*

$$DR_{p,\varepsilon}(g_\sharp \mu, g_\sharp \nu) = DR_{p,\varepsilon}(\mu, \nu),$$

where $g_\sharp \mu$ *is the push-forward of* μ *by* g*. In particular, it ensures invariance of* $DR_{p,\varepsilon}$ *under translations and rotations.*

Although formulas (4) and (5) are based on the same spirit, there are no apparent reasons why the proposed pseudo-metric should have the same behavior as the Wasserstein distance. It is the purpose of Proposition 3.6 to investigate the ability to factor out translations, for $DR_{2,\varepsilon}$ associated with the halfspace depth, giving a positive answer for the case of two Gaussian distributions with equal covariance matrices.

Proposition 3.6 (TRANSLATION CHARACTERIZATION)**.** *Consider* X, Y *two random variables following* $\mu \in \mathcal{M}_1(\mathcal{X})$ *and* $\nu \in \mathcal{M}_1(\mathcal{Y})$ *with expectations* $\mathbf{m}_1, \mathbf{m}_2$ *and variance-covariance matrices* $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ *respectively. Denoting by* μ^*, ν^* *the centered versions of* μ, ν*, it holds:*

$$\left| DR^2_{2,\varepsilon}(\mu, \nu) - DR^2_{2,\varepsilon}(\mu^*, \nu^*) - ||\mathbf{m}_1 - \mathbf{m}_2||^2 \right| \leq 2\, DR_{1,\varepsilon}(\mu^*, \nu^*)||\mathbf{m}_1 - \mathbf{m}_2||.$$

Now, let $\mu \sim \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ *and* $\nu \sim \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$*. Then it holds:*

$$\left| DR_{1,\varepsilon}(\mu, \nu) - ||\mathbf{m}_1 - \mathbf{m}_2|| \right| \leq C_\varepsilon \sup_{u \in \mathbb{S}^{d-1}} \left| \sqrt{u^\top \mathbf{\Sigma}_1 u} - \sqrt{u^\top \mathbf{\Sigma}_2 u} \right|,$$

where $C_\varepsilon = \int_0^{1-\varepsilon} \left| \Phi^{-1}(1 - \alpha(\beta)) \right| \mathrm{d}\beta$ *with* Φ *the cdf of the univariate standard Gaussian distribution.*

Following Proposition 3.6: when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, one has $DR_{2,\varepsilon}(\mu, \nu) = DR_{1,\varepsilon}(\mu, \nu) = ||\mathbf{m}_1 - \mathbf{m}_2||$ for any $\mu \sim \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1)$ and $\nu \sim \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$ providing a closed-form expression in the Gaussian case.

3.2 Robustness

In this part, we explore the robustness of the proposed distance, associated with the halfspace depth, given the finite sample breakdown point (BP; Donoho, 1982; Donoho and Hubert, 1983). This notion investigates the smallest contamination fraction under which the estimation breaks down in the worst case. Considering a sample $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ composed of i.i.d. observations drawn from a distribution μ with empirical measure $\hat{\mu}_n = (1/n) \sum_{i=1}^n \delta_{X_i}$, the finite sample breakdown point of $DR_{p,\varepsilon}$ w.r.t. $\mathcal{S}_n$, denoted by $BP(DR_{p,\varepsilon}, \mathcal{S}_n)$ is defined as

$$\min \left\{ \frac{o}{n+o} : \sup_{Z_1, \ldots, Z_o} DR_{p,\varepsilon}(\hat{\mu}_{n+o}, \hat{\mu}_n) = +\infty \right\},$$

where $\hat{\mu}_{n+o} = \frac{1}{n+o} \left(\sum_{i=1}^n \delta_{X_i} + \sum_{j=1}^o \delta_{Z_j} \right)$ is the "concatenate" empirical measure between $X_1, \ldots, X_n$ and the contamination sample $Z_1, \ldots, Z_o$ with $o \in \mathbb{N}^*$. It is well known that the extremal regions of the halfspace depth are not robust while its central regions are rather stable under contamination (Donoho and Gasko, 1992). Fortunately, by construction, the parameter ε allows us to ignore these extremal depth regions and thus ensure the robustness of the depth-trimmed regions distance. Based on the results of Donoho and Gasko (1992) and Nagy and Dvořák (2021), the following proposition provides a lower bound on the finite sample breakdown point of $DR_{p,\varepsilon}$, which highlights the robustness of the proposed distance (as well as its dependence on ε).

Proposition 3.7 (BREAKDOWN POINT)**.** *For the halfspace depth function, for any* $\beta \in [0, 1 - \varepsilon]$ *such that* $\alpha(\beta, \hat{\mu}_n) < \alpha_{\max}(\hat{\mu}_n)$*, it holds:*

$$BP(DR_{p,\varepsilon}, \mathcal{S}_n) \geq \begin{cases} \frac{\lceil n\alpha(1-\varepsilon,\hat{\mu}_n)/(1-\alpha(1-\varepsilon,\hat{\mu}_n)) \rceil}{n + \lceil n\alpha(1-\varepsilon,\hat{\mu}_n)/(1-\alpha(1-\varepsilon,\hat{\mu}_n)) \rceil} & \text{if } \alpha(1-\varepsilon, \hat{\mu}_n) \leq \frac{\alpha_{\max}(\hat{\mu}_n)}{1+\alpha_{\max}(\hat{\mu}_n)}, \\ \frac{\alpha_{\max}(\hat{\mu}_n)}{1+\alpha_{\max}(\hat{\mu}_n)} & \text{otherwise,} \end{cases}$$

where $\alpha_{\max}(\hat{\mu}_n) = \max_{x \in \mathbb{R}^d} HD_{\hat{\mu}_n}(x)$.

Thus, at least a proportion $\alpha(1-\varepsilon, \hat{\mu}_n)/(1 - \alpha(1-\varepsilon, \hat{\mu}_n))$ of outliers must be added to break down $DR_{p,\varepsilon}$ when considering larger regions, while central regions are robust independently of ε. For two datasets, $DR_{p,\varepsilon}$ breaks down if depth regions for at least one of the datasets do. The breakdown point is then the minimum between the breakdown points of each dataset. However, the breakdown point considers the worst case, i.e. the supremum over all possible contaminations, and is often pessimistic. Indeed the proposed pseudo-metric can handle more outliers in certain cases, as experimentally illustrated in Section 5.

4 EFFICIENT APPROXIMATE COMPUTATION

Exact computation of $DR_{p,\varepsilon}$ can appear time-consuming due to the high time complexity of the algorithms that calculate depth-trimmed regions (c.f. Liu and Zuo, 2014 and Liu et al., 2019a for projection and halfspace depths, respectively) rapidly growing with dimension. However, we design a universal approximate algorithm that achieves (log-) linear time complexity in n. Since properties (**D2–D4**) ensure that depth regions are convex bodies in $\mathbb{R}^d$, they can be characterized by their support functions defined by $h_{\mathcal{K}}(u) = \sup\{\langle x, u \rangle,\ x \in \mathcal{K}\}$ for any $u \in \mathbb{S}^{d-1}$ where $\mathcal{K}$ is a convex compact of $\mathbb{R}^d$. Following Schneider (1993), for two (convex) regions $D_\mu^{\alpha(\beta)}$ and $D_\nu^{\alpha(\beta)}$, the Hausdorff distance between them can be calculated as:

$$d_{\mathcal{H}}(D_\mu^{\alpha(\beta)}, D_\nu^{\alpha(\beta)}) = \sup_{u\in\mathbb{S}^{d-1}} \left|h_{D_\mu^{\alpha(\beta)}}(u) - h_{D_\nu^{\alpha(\beta)}}(u)\right|.$$

As we shall see in Section 5, mutual approximation of $h_{D^{\alpha(\beta)}}(u)$ by points from the sample and of sup by taking maximum over a finite set of directions allows for stable estimation quality. Recently, motivated by their numerous applications, many algorithms have been developed for the (exact and approximate) computation of data depths; see, e.g., Section 5 of Mosler and Mozharovskyi (2021) for a recent overview. Depths satisfying the projection property (which also include halfspace and projection depth, see Dyckerhoff (2004)) can be approximated by taking minimum over univariate depths; see e.g. Rousseeuw and Struyf (1998); Chen et al. (2013); Liu and Zuo (2014), Nagy et al. (2020) for theoretical guarantees, and Dyckerhoff et al. (2021) for an experimental validation. The case of AI-IRW is easier since the expectation in Equation (2) can be approximated through Monte-Carlo approximation.

Let $\mathbf{X}, \mathbf{Y}$ be two samples $\mathbf{X} = \{X_1, \dots, X_n\}$ and $\mathbf{Y} = \{Y_1, \dots, Y_m\}$ from μ, ν. When calculating approximated depth of sample points $D^{\mathbf{X}} \triangleq \{D(X_i, \hat{\mu}_n)\}_{i=1}^n$ (respectively $D^{\mathbf{Y}}$), a matrix $\mathbf{M}^{\mathbf{X}} \in \mathbb{R}^{n\times K}$ (respectively $\mathbf{M}^{\mathbf{Y}} \in \mathbb{R}^{m\times K}$) of projections of sample points on (a common) set of $K \in \mathbb{N}^*$ directions (with its element $\mathbf{M}^{\mathbf{X}}_{i,k} = \langle u_k, X_i\rangle$ for some $u_k \sim \mathcal{U}(\mathbb{S}^{d-1})$, where $\mathcal{U}(\cdot)$ is the uniform probability distribution) can be obtained as a side product. More precisely, $D^{\mathbf{X}}, D^{\mathbf{Y}}, \mathbf{M}^{\mathbf{X}}, \mathbf{M}^{\mathbf{Y}}$ are used in Algorithm 1, which implements the MC-approximation of the integral in (4). Particular cases of approximation algorithms for the halfspace depth, the projection depth and the AI-IRW depth are recalled in Section C in the Appendix. Time complexity of Algorithm 1 is $O\big(K(\Omega_\cdot(n \vee m, d) \vee n_\alpha(n \vee m))\big)$, where $\Omega_\cdot(\cdot,\cdot)$ stands for the complete complexity of computing univariate depths—in projections on u—for all points of the sample. As a byproduct, projections on u can be saved to be reused after for the approximation of $h_{D^{\alpha(\beta)}}(u)$. E.g., for the halfspace depth $\Omega_{hsp}(n,d) = O\big(n(d \vee \log n)\big)$ composed of projection of the data onto u, ordering them, and passing to record the depths (see e.g. Mozharovskyi et al., 2015). For the projection depth, $\Omega_{prj}(n,d) = O(nd)$, where after projecting the data onto u, univariate median and MAD can be computed with complexity $O(n)$ (see e.g. Liu and Zuo, 2014). For the AI-IRW depth, $\Omega_{aiirw}(n,d) = O(d^3 \vee n(d \vee \log(n))$ since it involves the computation of the square root of the precision matrix. However, $O(d^3)$ may be improved, which depends on the algorithm employed for computing the inverse of the covariance matrix (Staerman et al., 2021b). In comparison with popular distances, fixing $n = m$, the Wasserstein distance is of order $O(n^2(d \vee n))$ with approximations in $O(n^2 d)$ for Sinkhorn (Cuturi et al., 2013) and in $O(Kn(d \vee \log(n)))$ for the Sliced-Wasserstein distance (Rabin et al., 2012); the MMD (Gretton et al., 2007) is of order $O(n^2 d)$. For example, the computational complexity of $DR_{p,\varepsilon}$ with the projection depth is only of $O(Kn(d \vee n_\alpha))$ and thus competes with the fastest (max) sliced-Wasserstein distance.

Algorithm 1 Approximation of $DR_{p,\varepsilon}$

Initialization: $\mathbf{X}$,$\mathbf{Y}$,n_α,K

$H = 0$; compute $D^{\mathbf{X}}, D^{\mathbf{Y}}, \mathbf{M}^{\mathbf{X}}, \mathbf{M}^{\mathbf{Y}}$
for $\ell = 1, \dots, n_\alpha$ **do**
  Draw $\beta_\ell \sim \mathcal{U}([0, 1-\varepsilon])$
  Compute $\hat{\alpha}_\ell(\cdot) := \hat{\alpha}(\beta_\ell, \cdot)$
  Determine points inside $\alpha_\ell(\cdot)$-regions: $\mathcal{I}_\ell^{\mathbf{X}} = \{i : D_i^{\mathbf{X}} > \hat{\alpha}_\ell(\mathbf{X})\}$; $\mathcal{I}_\ell^{\mathbf{Y}} = \{j : D_j^{\mathbf{Y}} > \hat{\alpha}_\ell(\mathbf{Y})\}$
  for $k = 1, \dots, K$ **do**
    Compute approximation of support functions: $h_k^{\mathbf{X}} = \max \mathbf{M}^{\mathbf{X}}_{\mathcal{I}_\ell^{\mathbf{X}},k}$; $h_k^{\mathbf{Y}} = \max \mathbf{M}^{\mathbf{Y}}_{\mathcal{I}_\ell^{\mathbf{Y}},k}$
  end for
  Increase cumulative Hausdorff distance: $H \mathrel{+}= \max_{k\le K} |h_k^{\mathbf{X}} - h_k^{\mathbf{Y}}|^p$
end for
Output: $\widehat{DR}_{p,\varepsilon} = (H/n_\alpha)^{1/p}$

5 NUMERICAL EXPERIMENTS

In this section, we first measure the quality of the approximation introduced in Section 4 and explore its dependency on the number of projections. Further, we present two studies on robustness of the proposed pseudo-metric $DR_{p,\varepsilon}$ to outliers. On synthetic datasets, we investigate how $DR_{p,\varepsilon}$ behaves under the presence of outliers using two different settings. On a real image dataset extracted from Fashion-MNIST where images are seen as bags of pixels, we evaluate the robustness of spectral clustering based on $DR_{p,\varepsilon}$. Finally, we analyze the relevance of using $DR_{p,\varepsilon}$ as an evaluation metric in natural language generation to compare the empirical distributions of words of a pair of texts. Where applicable, we include state-of-the-art methods for comparison. Due to space limitations, experiments on the influence of the parameters n_α and ε, as well as on the statistical rates, are deferred to the Appendix section.

Approximation error in terms of the number of projections. Proposition 3.6 allows to derive a closed form expression for $DR_{2,\varepsilon}(\mu,\nu)$ when μ, ν are Gaussian distributions with the same variance-covariance matrix. In order to investigate the quality of the approximation on light-tailed and heavy-tailed distributions, we focus on computing $DR_{p,\varepsilon}$ (with $p = 2$, $\varepsilon = 0.3$, $n_\alpha = 20$ and using the halfspace depth) for varying number of random projections K between a sample of 1000 points stemming from $\mu \sim \mathcal{N}(\mathbf{0}_d, I_d)$ for $d = 5$ and two different samples. These two samples are constructed from 1000 observations stemming from *Gaussian* and symmetrical *Cauchy* distributions, both with a center equal to $\mathbf{7}_d$. Comparison with the approximation of max

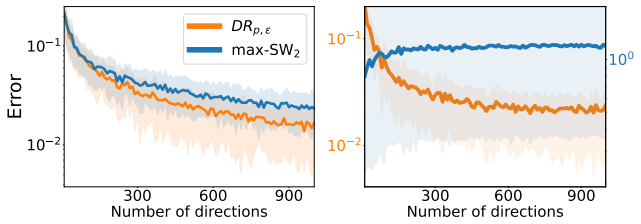

Figure 1: Relative approximation error (averaged over 100 runs) of $DR_{p,\varepsilon}$ and the max Sliced-Wassserstein for *Gaussian* (left) and *Cauchy* (right) sample with dimension $d = 5$ for differing numbers of approximating directions.

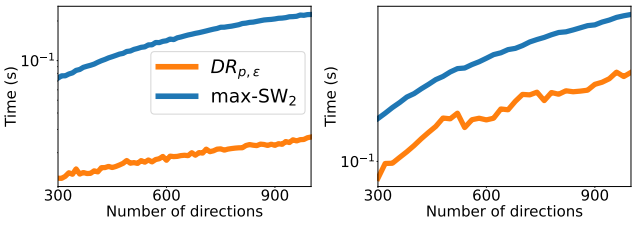

Figure 2: Computation time (averaged over 100 runs) of $DR_{p,\varepsilon}$ and the max Sliced-Wassserstein for *Gaussian* with $n = 100$, $d = 5$ (left) and $n = 1000$, $d = 50$ (right) for differing numbers of approximating directions.

Sliced-Wasserstein (max-SW; see e.g. Kolouri et al., 2019), which shares the same closed-form as $DR_{2,\varepsilon}$, is also provided. Denoting by max-$\widehat{\text{SW}}$ the Monte-Carlo approximation of the max-SW, the relative approximation errors, i.e., $(\widehat{DR}_{p,\varepsilon} - ||\mathbf{7}_d||_2)/||\mathbf{7}_d||_2$ and $(\text{max-}\widehat{\text{SW}} - ||\mathbf{7}_d||_2)/||\mathbf{7}_d||_2$, are computed investigating both the quality of the approximation and the robustness of these discrepancy measures. Results that report the averaged approximation error and the 25-75% empirical quantile intervals are depicted in Figure 1. They show that $DR_{p,\varepsilon}$ possesses the same behavior as max-SW when considering *Gaussians* while it behaves advantageously for *Cauchy* distribution. Computation times are depicted in Figure 2, highlighting a constant-multiple improvement compared to the max-SW, which is already computationally fast.

Robustness to outliers. We analyze the robustness of $DR_{p,\varepsilon}$ by measuring its ability to overcome outliers (its robustness regarding the influence of the parameter ε are given in the Section D.4 in the Appendix). In this benchmark, we naturally include existing robust extensions of the Wasserstein distance: Subspace Robust Wasserstein (SRW; Paty and Cuturi, 2019) searching for a maximal distance on lower-dimensional subspaces, ROBOT (Mukherjee et al., 2020) and RUOT (Balaji et al., 2020) being robust modifications of the unbalanced optimal transport (Chizat et al., 2018). Medians-of-Means Wasserstein (MoMW; Staerman et al., 2021a) that replaces the empirical means in the Kantorivich duality formulae by the robust mean estimator MoM (see e.g. Lecué and Lerasle, 2020; Laforgue et al., 2021), is not employed due to high computational burden. Further, for completeness, we add the standard Wasserstein distance (W) and its approximation, the Sliced-Wasserstein (Sliced-W; Rabin et al., 2012) distance, with the same number of projections ($K = 1000$) as $DR_{p,\varepsilon}$. Since the scales of the compared methods differ, *relative error* is used as a performance metric, i.e., the ratio of the absolute difference of the computed distance with and without anomalies divided by the latter. Two settings for a pair of distributions are addressed: (a) *Fragmented hypercube* precedently studied in Paty and Cuturi (2019), where the source distribution is uniform in the hypercube $[-1, 1]^2$ and the target distribution is transformed from the source via the map $T : x \mapsto x + 2\text{sign}(x)$ where $sign(.)$ is taken elementwisely. Outliers are drawn uniformly from $[-4, 4]^2$. (b) Two multivariate standard *Gaussian* distributions, one shifted by $\mathbf{10}_2$, with outliers drawn uniformly from $[-10, 20]^2$. Our analysis is conducted over 500 sampled points from the distributions described above.

To investigate the robustness of $DR_{p,\varepsilon}$, we consider the following values of ε: $0.1, 0.2, 0.3$ computed with the projection depth. Thus, data depths are computed on source and target distributions such that 10%, 20%, 30% of data with lower depth values w.r.t. each distribution are not used in computation of $DR_{p,0.1}, DR_{p,0.2}, DR_{p,0.3}$, respectively. Figure 3, which plots the relative error depending on the portion of outliers varying up to 20%, illustrates advantageous behavior of $DR_{p,\varepsilon}$ (for $\varepsilon = 0.1, 0.2, 0.3$) for reasonable (starting with $\approx 2.5\%$) contamination. It also confirms the pessimism of the breakdown point provided in Proposition 3.7 since $DR_{p,0.1}$ (represented by the blue curve) shows robustness to at least 20 % of outliers.

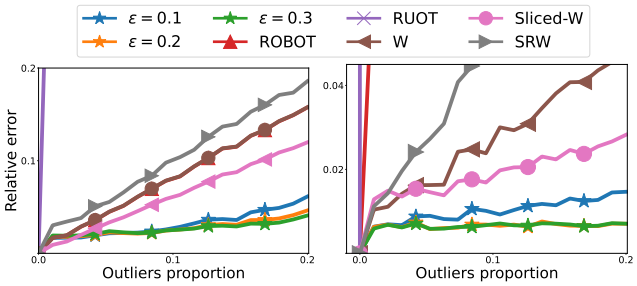

Figure 3: Relative error (averaged over 100 runs) of different distances for increasing outliers proportion on *fragmented hypercube* (left) and *Gaussian* (right) data.

(Robust) Clustering on bags of pixels. We demonstrate the relevance of the proposed pseudo-metric through an application to (robust) clustering. To that end, we perform spectral clustering (Shi and Malik, 2000) on two datasets derived from Fashion-MNIST (FM). Each grayscale image is seen as a bag of pixels (Jebara, 2003), i.e. as an empirical probability distribution over a 3-dimensional space (the two

first dimensions indicate the pixel position and the third one, its intensity). The first dataset (FM) is constructed by taking the 100 first images in each class of the Fashion-MNIST dataset. The second dataset (Cont. FM), considered contaminated, is designed by introducing white patches on the left corner of 50 images drawn uniformly in the first dataset, which yields 5% of contamination. We benchmark $DR_{p,\varepsilon}$ (using the projection depth) setting $p = 2$ and $\varepsilon = 0.1$ with the Wasserstein (W), the Sliced-Wasserstein (Sliced-W) and the Maximum Mean Discrepancy (MMD; Gretton et al., 2007) distances. $DR_{p,\varepsilon}$ and the Sliced-Wasserstein are approximated by Monte-Carlo using 100 directions while the MMD distance is computed using a Gaussian kernel with a bandwidth equal to 1. As a baseline method, spectral clustering is also applied to images considered as vectors using Euclidean distance. Standard parameters of the `scikit-learn` spectral clustering implementation are employed with a number of clusters fixed to 10. Performances of the benchmarked metrics are assessed by measuring the normalized mutual information (NMI; Shannon, 1948) and the adjusted rank index (ARI; Hubert and Arabie, 1985), which are standard clustering evaluation measures when the ground truth class labels are available. Results presented in Table 1 show that for both cases, i.e. with or without contamination, spectral clustering based on $DR_{p,\varepsilon}$ outperforms spectral clustering based on the other metrics.

	FM		Cont. FM	
	NMI	ARI	NMI	ARI
$DR_{p,\varepsilon}$	**0.58**	**0.43**	**0.55**	**0.42**
W	0.50	0.35	0.48	0.30
Sliced-W	0.55	0.39	0.47	0.33
MMD	0.54	0.37	0.50	0.36
Euclidean	0.50	0.32	0.48	0.30

Table 1: Spectral clustering performances.

Automatic evaluation of natural language generation (NLG). Collecting human annotations to evaluate NLG systems is both expensive and time-consuming. Thus, automatically assessing the similarity between two texts is highly interesting for the NLP community (Specia et al., 2010). This task aims to build an evaluation metric that achieves a high correlation with the score given by a human annotator. String-based metrics (i.e. that compare the string representations of texts) such as BLEU (Papineni et al., 2002), METEOR (MET.; Banerjee and Lavie, 2005), ROUGE (Lin, 2004), TER (Snover et al., 2006), have been outperformed in many tasks by embedding-based metrics, i.e., that rely on continuous representations (Devlin et al., 2019). Embedding-based metrics (e.g BertScore (BertS; Zhang et al., 2019) and MoverScore (MoverS; Zhao et al., 2019) that are now the state-of-the-art domain, compare input and reference texts both represented as probability distributions and are both constructed similarly. The first step relies on a deep contextualized encoder (BERT in our case, see Devlin et al., 2019) that maps texts into elements of a finite-dimensional space. Each text corresponds to a collection of words, where each word is represented by an element in $\mathbb{R}^d$, where d is fixed by the encoder. The second step involves using a function that measures the similarity between the embedded texts.

	Correctness			Data Coverage			Relevance		
	r	τ	ρ	r	τ	ρ	r	τ	ρ
$DR_{p,\varepsilon}$	**89.4**	**80.0**	**92.6**	**84.2**	58.3	72.3	**86.2**	62.7	72.9
W	86.2	73.0	86.7	80.4	45.3	62.3	83.8	51.3	67.6
Sliced-W	86.1	73.0	85.8	80.9	45.5	60.0	82.0	51.3	68.2
MMD	25.4	71.7	8.3	19.1	45.3	10.0	26.1	51.3	15.0
BertS	85.5	73.3	83.4	74.7	53.3	68.2	83.3	65.0	**79.4**
MoverS	84.1	73.3	84.1	78.7	53.3	66.2	82.1	65.0	77.4
BLEU	77.6	60.0	66.3	55.7	36.6	50.2	63.0	51.6	65.2
ROUGE	80.6	65.0	65.0	76.5	**60.3**	**76.3**	64.3	56.7	69.2
MET.	86.5	70.0	66.3	77.3	46.6	50.2	82.1	58.6	65.2
TER	79.6	58.0	78.3	69.7	38.0	58.2	75.0	**77.6**	70.2

Table 2: Absolute correlation at the system level with three human judgment criteria. The best overall results are indicated in bold, best results in their group are underlined.

We follow previous BERT-based metrics and evaluate performances of $DR_{p,\varepsilon}$ (with $p = 2$, $\varepsilon = 0.01$ and using the AI-IRW depth) on two different NLG tasks namely: data2text generation (using the WebNLG 2020 dataset Ferreira et al., 2020) and summarization. For the sake of place, summarization results and additional experimental details are reported in Section E in the Appendix. For WebNLG, we follow standard methods to assess the performance of NLG metrics (see e.g. Zhao et al., 2019). We compute the correlation with the following annotation scores: *correctness*, *data coverage*, and *relevance*. We report in Table 2 correlation results on the WebNLG task using Pearson (r), Spearman (ρ) and Kendall (τ) correlation coefficients. When performing a fair comparison between metrics, i.e. when $DR_{p,\varepsilon}$, W, Sliced-W, MMD are directly used on the output of BERT, we observe that $DR_{p,\varepsilon}$ achieves the best results on all configurations. It is worth noting that $DR_{p,\varepsilon}$ also compares favorably against existing state-of-the-art NLG methods in many different scenarios and shows promising results.

6 DISCUSSION

Leveraging the notion of statistical data depth function, a novel pseudo-metric between multivariate probability distributions—that meets the aforementioned requirements—was introduced. The developed framework exhibits inherent versatility due to numerous data depth variants. The linear approximation algorithm and the robustness property make $DR_{p,\varepsilon}$ a promising tool for a large spectrum of applications

beyond clustering and NLG, e.g. in generative adversarial networks (GANs) or information retrieval. Moreover, recent works extending the notion of data depth to further types of data such as functional and time-series data (Nieto-Reyes and Battey, 2016; Gijbels and Nagy, 2017), directional (or spherical) data (Ley et al., 2014), random matrices (Paindaveine and Van Bever, 2018), curves (or paths) data (Lafaye et al., 2020), and random sets (Cascos et al., 2021) shall allow for the use of the proposed pseudo-metric for a wide range of applications.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Balaji, Y., Chellappa, R., and Feizi, S. (2020). Robust optimal transport with applications in generative modeling and domain adaptation. *arXiv preprint arXiv:2010.05862*.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the royal society A*.

Bhandari, M., Gour, P., Ashfaq, A., Liu, P., and Neubig, G. (2020). Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

Billingsley, P. (1999). *Convergence of probability measures (2nd ed.)*. John Wiley & Sons.

Brunel, V.-E. (2019). Concentration of the empirical level sets of tukey's halfspace depth. *Probability Theory and Related Fields*, 173(3):1165–1196.

Cascos, I., Li, Q., and Molchanov, I. (2021). Depth and outliers for samples of sets and random sets distributions. *Australian & New Zealand Journal of Statistics*, 63(1):55–82.

Cha, S. and Srihari, S. N. (2002). On measuring the distance between histograms. *Pattern Recognit.*, 35(6):1355–1370.

Chapuis, E., Colombo, P., Labeau, M., and Clave, C. (2021). Code-switched inspired losses for generic spoken dialog representations. *arXiv preprint arXiv:2108.12465*.

Chapuis, E., Colombo, P., Manica, M., Labeau, M., and Clavel, C. (2020). Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*.

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Chen, D., Morin, P., and Wagner, U. (2013). Absolute approximation of tukey depth: Theory and experiments. *Computational Geometry*, 46(5):566 – 573.

Chen, Y.-C. and Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). Unbalanced optimal transport: dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090 – 3123.

Clark, E., Celikyilmaz, A., and Smith, N. A. (2019). Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Colombo, P., Chapuis, E., Manica, M., Vignon, E., Varni, G., and Clavel, C. (2020). Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7594–7601.

Colombo, P., Staerman, G., Clavel, C., and Piantanida, P. (2021a). Automatic text evaluation through the lens of wasserstein barycenters. *arXiv preprint arXiv:2108.12463*.

Colombo, P., Witon, W., Modi, A., Kennedy, J., and Kapadia, M. (2019). Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*.

Colombo, P., Yang, C., Varni, G., and Clavel, C. (2021b). Beam search with bidirectional strategies for neural response generation. *arXiv preprint arXiv:2110.03389*.

Csiszàr, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den bewis der ergodizität von markhoffschen kette. *Magyer Tud. Akad. Mat. Kutato Int. Koezl*, 8:85–108.

Cuturi, M., Teboul, O., and Vert, J.-P. (2013). Sinkhorn distances: Lightspeed computation of optimal transportation. In *Advances in Neural Information Processing Systems*.

Dang, H. T. and Owczarzak, K. (2008). Overview of the tac 2008 update summarization task. In *Proceedings of the Text Analysis Conference (TAC)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Donoho, D. L. (1982). Breakdown properties of location estimators. *P.h.D., qualifying paper, Dept. Statistics, Hardvard University*.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on half space depth and projected outlyingness. *The Annals of Statistics*, 20:1803–1827.

Donoho, D. L. and Hubert, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich Lehman*, pages 157–184.

Dyckerhoff, R. (2004). Data depth satisfying the projection property. *Allgemeines Statistisches Archiv*, 88(2):163–190.

Dyckerhoff, R., Mozharovskyi, P., and Nagy, S. (2021). Approximate computation of projection depths. *Computational Statistics and Data Analysis*, 157:107166.

Einhmahl, J. H. and Mason, D. M. (1992). Generalized quantile process. *The annals of statistics*, 20(2):1062–1078.

Ferreira, T., Gardent, C., Ilinykh, N., van der Lee, C., Mille, S., Moussallem, D., and Shimorina, A. (2020). The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*.

Ferreira, T. C., Moussallem, D., Krahmer, E., and Wubben, S. (2018). Enriching the webnlg corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176.

Garcia, A., Colombo, P., Essid, S., d'Alché Buc, F., and Clavel, C. (2019). From the token to the review: A hierarchical multimodal approach to opinion mining. *arXiv preprint arXiv:1908.11216*.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017). Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*.

Gehrmann, S., Deng, Y., and Rush, A. (2018). Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Gijbels, I. and Nagy, S. (2017). On a general definition of depth for functional data. *Statistical Science*, 32(4):630–639.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2007). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*.

Hallin, M., Paindaveine, D., and Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From l1 optimization to halfspace depth. *Ann. Statist.*, 38(2):635–669.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Jebara, T. (2003). Images as bags of pixels. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 265–272.

Jörnsten, R. (2004). Clustering and classification based on the l1 data depth. *Journal of Multivariate Analysis*, 90(1):67 – 89.

Kedzie, C., McKeown, K., and Daume III, H. (2018). Content selection in deep learning models of summarization. *arXiv preprint arXiv:1810.12343*.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Kolouri, S., Nadjahi, K., Umut, S., Badeau, R., and Rohde K., G. (2019). Generalized sliced wasserstein distance. In *Advances Neural Information Processing Systems*.

Koltchinskii, V. I. and Dudley, R. M. (1996). On spatial quantiles. *Unpublished manuscript*.

Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25(5):1998–2017.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Lafaye, P., Mozharovskyi, P., and Vimond, M. (2020). Depth for curve data and applications. *Journal of the American Statistical Association*, pages 1–17. in press.

Laforgue, P., Staerman, G., and Clémençon, S. (2021). Generalization bounds in the presence of outliers: a median-of-means study. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5937–5947.

Lange, T., Mosler, K., and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers*, 55(1):49–69.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT evaluation using block movements. In *11th Conference of the EACL*.

Leusch, G., Ueffing, N., Ney, H., et al. (2003). A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of Mt Summit IX*, pages 240–247.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Ley, C., Sabbah, C., and Verdebout, T. (2014). A new concept of quantiles for directional data and the angular Mahalanobis depth. *Electronic Journal of Statistics*, 8(1):795–816.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). Dd-classifier: Nonparametric classification procedure based on dd-plot. *JASA*, 107(498):737–753.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414.

Liu, R. Y. (1992). *Data Depth and Multivariate Rank Tests*, page 279–294. North-Holland, Amsterdam.

Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260.

Liu, X., Mosler, K., and Mozharovskyi, P. (2019a). Fast computation of tukey trimmed regions and median in dimension $p > 2$. *Journal of Computational and Graphical Statistics*, 28(3):682–697.

Liu, X. and Zuo, Y. (2014). Computing projection depth and its associated estimators. *Statistics and Computing*, 24(1):51–63.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Mairesse, F., Gasic, M., Jurcicek, F., Keizer, S., Thomson, B., Yu, K., and Young, S. (2010). Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561.

McNamee, P. and Dang, H. T. (2009). Overview of the tac 2009 knowledge base population track. In *Proceedings of the Text Analysis Conference (TAC)*, volume 17, pages 111–113.

Melamed, I. D., Green, R., and Turian, J. (2003). Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pages 61–63.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Mosler, K. (2013). Depth statistics. *Robustness and complex data structures*.

Mosler, K. and Mozharovskyi, P. (2021). Choosing among notions of depth for multivariate data. *Statistical Science*. In press.

Mozharovskyi, P., Mosler, K., and Lange, T. (2015). Classifying real-world data with the $DD\alpha$-procedure. *Advances in Data Analysis and Classification*, 9(3):287–314.

Mukherjee, D., Guha, A., Solomon, J., Sun, Y., and Yurochkin, M. (2020). Outlier-robust optimal transport. *arXiv preprint arXiv:2012.07363*.

Nagy, S. (2019). Halfspace depth does not characterize probability distributions. *Statistical Papers*, 26(3):1135–1139.

Nagy, S. and Dvořák, J. (2021). Illumination depth. *Journal of Computational and Graphical Statistics*, 30(1):78–90.

Nagy, S., Dyckerhoff, R., and Mozharovskyi, P. (2020). Uniform convergence rates for the approximated halfspace and projection depth. *Electronic Journal of Statistics*, 14(2):3939–3975.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es.

Nenkova, A. and Passonneau, R. J. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Nieto-Reyes, A. and Battey, H. (2016). A topologically valid definition of depth for functional data. *Statistical Science*, 31(1):61–79.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*.

Paindaveine, D. and bever, G. V. (2013). From depth to local depth: A focus on centrality. *Journal of the American Statistical Association*, 108(503):1105–1119.

Paindaveine, D. and Van Bever, G. (2018). Halfspace depths for scatter, concentration and shape matrices. *The Annals of Statistics*, 46(6B):3276–3307.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Paty, F.-P. and Cuturi, M. (2019). Subspace robust Wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5072–5081.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 EMNLP (EMNLP)*, pages 1532–1543. ACL.

Perez-Beltrachini, L., Sayed, R., and Gardent, C. (2016). Building rdf content for data-to-text generation. In *The 26th International Conference on Computational Linguistics (COLING 2016)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pokotylo, O., Mozharovskyi, P., and Dyckerhoff, R. (2019). Depth and depth-based classification with R-Package ddalpha. *Journal of Statistical Software, Articles*, 91(5):1–46.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012). Wasserstein barycenter and its application to texture mixing. In Bruckstein, A. M., ter Haar Romeny, B. M., Bronstein, A. M., and Bronstein, M. M., editors, *Scale Space and Variational Methods in Computer Vision*, pages 435–446, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rachev, S. (1991). *Probability Metrics and the Stability of Stochastic Models*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Ramsay, K., Durocher, S., and Leblanc, A. (2019). Integrated rank-weighted depth. *Journal of Multivariate Analysis*, 173:51–69.

Rankel, P. A., Conroy, J., Dang, H. T., and Nenkova, A. (2013). A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Association for Computational Linguistics (ACL)*, pages 131–136.

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94(446):388–402.

Rousseeuw, P. J. and Hubert, M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, 8(2):1236.

Rousseeuw, P. J. and Rutz, I. (1999). The depth function of a population distribution. *Metrika*, 49(3):213–244.

Rousseeuw, P. J. and Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8(3):193–203.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, Berkeley, Calif. University of California Press.

Schneider, R. (1993). *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press, Cambridge.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 22(08):888–905.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Specia, L., Raj, D., and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550 – 1599.

Staerman, G. (2022). *Functional anomaly detection and robust estimation*. PhD thesis, Institut polytechnique de Paris.

Staerman, G., Adjakossa, E., Mozharovskyi, P., Hofer, V., Gupta, J. S., and Clémençon, S. (2022). Functional

anomaly detection: a benchmark study. *arXiv preprint arXiv:2201.05115*.

Staerman, G., Laforgue, P., Mozharovskyi, P., and d'Alché Buc, F. (2021a). When ot meets mom: Robust estimation of wasserstein distance. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 136–144.

Staerman, G., Mozharovskyi, P., and Clémençon, S. (2020). The area of the convex hull of sampled curves: a robust functional statistical depth measure. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 570–579.

Staerman, G., Mozharovskyi, P., Clémençon, S., and d'Alché Buc, F. (2019). Functional isolation forest. In *Proceedings of The 11th Asian Conference on Machine Learning*.

Staerman, G., Mozharovskyi, P., and Clémençon, S. (2021b). Affine-invariant integrated rank-weighted depth: Definition, properties and finite sample analysis. *arXiv preprint arXiv:2106.11068*.

Stahel, W. A. (1981). Breakdown of covariance estimators. Technical report, Fachgruppe für Statistik, ETH, Zürich.

Stanchev, P., Wang, W., and Ney, H. (2019). Eed: Extended edit distance measure for machine translation. In *Proceedings of the Fourth WMT (Volume 2: Shared Task Papers, Day 1)*, pages 514–520.

Stummer, W. and Vajda, I. (2012). On bregman distances and divergences of probability measures. *IEEE Transactions on Information Theory*, 58(3):1277 – 1288.

Tukey, J. W. (1975). Mathematics and the picturing of data. In James, R., editor, *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531. Canadian Mathematical Congress.

Villani, C. (2003). *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, New York.

Wang, D., Liu, P., Zheng, Y., Qiu, X., and Huang, X. (2020). Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Wang, W., Peter, J.-T., Rosendahl, H., and Ney, H. (2016). Character: Translation edit rate on character level. In *Proceedings of the First WMT: Volume 2, Shared Task Papers*, pages 505–510.

Wen, T.-H., Gašić, M., Mrkšić, N., Su, P.-H., Vandyke, D., and Young, S. (2015). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Witon, W., Colombo, P., Modi, A., and Kapadia, M. (2018). Disney at iest 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yoon, W., Yeo, Y. S., Jeong, M., Yi, B.-J., and Kang, J. (2020). Learning by semantic similarity makes abstractive summarization better. *arXiv preprint arXiv:2002.07767*.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2020). Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

Zhong, M., Liu, P., Wang, D., Qiu, X., and Huang, X. (2019). Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., and Zhao, T. (2018). Neural document summarization by jointly learning to score and select sentences. *arXiv preprint arXiv:1807.02305*.

Zuo (2003). Projected based depth functions and associated medians. *The annals of statistics*, 31(5):1460–1490.

Zuo, B. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482.

This Appendix is organized as follows:

- Appendix A contains additional notations as well as useful preliminary results.
- Appendix B contains the proofs of the propositions/theorems provided in the paper.
- Appendix C contains approximation algorithms to compute halfspace/projection/AI-IRW depth.
- Appendix D contains additional synthetic experiments.
- Appendix E contains details on experimental settings of NLP applications.

A PRELIMINARY RESULTS

First, we introduce additional notations and recall some lemmas, used in the subsequent proofs.

A.1 Hausdorff Distance

The Hausdorff distance between two bounded subspaces $\mathcal{K}_1, \mathcal{K}_2$ of $\mathbb{R}^d$ is defined as

$$d_{\mathcal{H}}(\mathcal{K}_1, \mathcal{K}_2) = \max\left\{ \sup_{x\in\mathcal{K}_1} \inf_{y\in\mathcal{K}_2} ||x-y||, \sup_{y\in\mathcal{K}_2} \inf_{x\in\mathcal{K}_1} ||x-y|| \right\}.$$

Furthermore, if $\mathcal{K}_1$ and $\mathcal{K}_2$ are convex bodies (i.e. non empty compact convex sets), the Hausdorff distance can be reformulated with support functions of $\mathcal{K}_1, \mathcal{K}_2$:

$$d_{\mathcal{H}}(\mathcal{K}_1, \mathcal{K}_2) = \sup_{u\in\mathbb{S}^{d-1}} \left| h_{\mathcal{K}_1}(u) - h_{\mathcal{K}_2}(u) \right|,$$

where $h_{\mathcal{K}_1}(u) = \sup\{\langle u, x\rangle, \quad x \in \mathcal{K}_1\}$.

A.2 Quantile regions

Let $u \in \mathbb{S}^{d-1}$ and $X \sim \mu$ where $\mu \in \mathcal{M}_1(\mathcal{X})$ with $\mathcal{X} \subset \mathbb{R}^d$. We define the $(1-\beta)$ directional quantile of a distribution μ in the direction u as

$$q^{1-\beta}_{\mu,u} = \inf\left\{t \in \mathbb{R} : \quad \mathbb{P}\left(\langle u, X\rangle \le t\right) \ge 1-\beta\right\} \tag{6}$$

and the upper $(1-\beta)$ quantile set of μ

$$Q^{1-\beta}_{\mu} = \left\{x \in \mathbb{R}^d : \quad \langle u, x\rangle \le q^{1-\beta}_{\mu,u}, \quad \forall\, u \in \mathbb{S}^{d-1}\right\}. \tag{7}$$

A.3 Auxiliary results

We now recall useful results, so as to characterize the halfspace depth regions.

Lemma A.1 (Brunel, 2019, Lemma 1)**.** *Let $\mu \in \mathcal{M}_1(\mathcal{X})$, for any $\beta \in (0,1)$, it holds: $D^{\beta}_{\mu} = Q^{1-\beta}_{\mu}$.*

Lemma A.2 (Brunel, 2019, Proposition 1)**.** *Let $\mu \in \mathcal{M}_1(\mathcal{X})$ with a $(1-\beta)$ directional quantile $q^{1-\beta}_{\mu,u}$ for any $u \in \mathbb{S}^{d-1}$. Assume that $u \mapsto q^{1-\beta}_{\mu,u}$ are sublinear, i.e., $q^{1-\beta}_{\mu,u+\lambda v} \le q^{1-\beta}_{\mu,u} + \lambda\, q^{1-\beta}_{\mu,v}, \quad \forall\ \lambda > 0$. Then for any $u \in \mathbb{S}^{d-1}$, it holds $h_{Q^{1-\beta}_{\mu,u}}(u) = q^{1-\beta}_{\mu,u}$.*

Lemma A.3. *Let $d = 1$ and $X^1 \sim \mu_1$, $Y^1 \sim \nu_1$ be two random variables where μ_1, ν_1 are univariate probability distributions. Denoting by $F^{-1}_{X^1}$ the quantile function of X^1, then the depth-trimmed region based pseudo-metric (associated with the halfspace depth) is defined as*

$$DR^{p}_{p,\varepsilon}(\mu_1, \nu_1) = 2\int_{\varepsilon/2}^{1/2} \max\left\{ |F^{-1}_{X^1}(q) - F^{-1}_{Y^1}(q)|^p,\ |F^{-1}_{X^1}(1-q) - F^{-1}_{Y^1}(1-q)|^p \right\} \mathrm{d}q.$$

Proof. In dimension one, the halfspace depth of any $t \in \mathbb{R}$ w.r.t. μ_1 and ν_1 boils down to

$$D(t, \mu_1) = \min\Big\{F_{X^1}(t), 1 - F_{X^1}(t)\Big\} \quad \text{and} \quad D(t, \nu_1) = \min\Big\{F_{Y^1}(t), 1 - F_{Y^1}(t)\Big\},$$

and for any $\gamma \in [0, 1]$, its upper-level sets to intervals

$$D^\gamma_{\mu_1} = [F^{-1}_{X^1}(\gamma),\ F^{-1}_{X^1}(1-\gamma)] \quad \text{and} \quad D^\gamma_{\nu_1} = [F^{-1}_{Y^1}(\gamma),\ F^{-1}_{Y^1}(1-\gamma)]. \tag{8}$$

Now, the quantile function $\alpha(\beta, .)$ can be explicitly derived as function of $\beta \in [0, 1]$:

$$\begin{aligned}
\alpha(\beta, \mu_1) &= \sup\Big\{\gamma \in [0,1] :\ \ \mu_1\Big([F^{-1}_{X^1}(\gamma),\ F^{-1}_{X^1}(1-\gamma)]\Big) \geq \beta\Big\} \\
&= \sup\Big\{\gamma \in [0,1] :\ \ 1 - 2\gamma \geq \beta\Big\} \\
&= \frac{1-\beta}{2}.
\end{aligned}$$

Following the same reasoning, it holds $\alpha(\beta, \nu_1) = \frac{1-\beta}{2}$. Further, by change of variables

$$\int_0^{1-\varepsilon} d_{\mathcal{H}}\left(D^{(1-\beta)/2}_{\mu_1}, D^{(1-\beta)/2}_{\nu_1}\right)^p \,\mathrm{d}\beta = 2\int_{\varepsilon/2}^{1/2} d_{\mathcal{H}}\left(D^q_{\mu_1}, D^q_{\nu_1}\right)^p \,\mathrm{d}q.$$

Combining Equation (8) and the Hausdorff distance definition recalled in Section A.1 lead to the result.

□

B TECHNICAL PROOFS

We now prove the main results stated in the paper.

B.1 Proof of Proposition 3.4

For any $0 \leq \beta \leq 1 - \varepsilon$ with $\varepsilon \in (0, 1]$, and any $\mu \in \mathcal{M}_1(\mathcal{X})$, $\nu \in \mathcal{M}_1(\mathcal{Y})$, $D^{\alpha(\beta)}_\mu, D^{\alpha(\beta)}_\nu$ are non-empty compact subsets of $\mathbb{R}^d$ due to the properties (**D2-D3**). The Hausdorff distance $d_{\mathcal{H}}$, recalled in Section A.1, is known to be a distance on the space of non-empty compact sets which implies that $DR_{p,\varepsilon}$ satisfies positivity, symmetry and the triangle inequality (thanks to Minkowski inequality). If $\mu = \nu$ then $D^{\alpha(\beta)}_\mu = D^{\alpha(\beta)}_\nu, \quad \forall\ \beta \in [0, 1-\varepsilon]$ which leads to $DR_{p,\varepsilon}(\mu, \nu) = 0$. The reverse is not true. $DR_{p,\varepsilon}(\mu, \nu) = 0$ implies $D^{\alpha(\beta)}_\mu = D^{\alpha(\beta)}_\nu, \quad \forall\ \beta \in [0, 1-\varepsilon]$ that not leads to $\mu = \nu$. Indeed, convex depth regions do not characterize probability distributions in general (see Nagy, 2019 for the halfspace depth) that would be the first step in order to prove the previous entailment.

B.2 Proof of Proposition 3.5

Let $A \in \mathbb{R}^{d\times d}$ be a non-singular matrix and $b \in \mathbb{R}^d$ such that $g : x \mapsto Ax + b$. Then, it holds:

$$\begin{aligned}
DR^p_{p,\varepsilon}(g_\sharp\mu, g_\sharp\nu) &= \int_0^{1-\varepsilon} \Big[d_{\mathcal{H}}(D^{\alpha(\beta)}_{g_\sharp\mu}, D^{\alpha(\beta)}_{g_\sharp\nu})\Big]^p \,\mathrm{d}\beta \\
&\overset{(i)}{=} \int_0^{1-\varepsilon} \Big[d_{\mathcal{H}}(AD^{\alpha(\beta)}_\mu + b, AD^{\alpha(\beta)}_\nu + b)\Big]^p \,\mathrm{d}\beta,
\end{aligned} \tag{9}$$

where (i) holds because any data depth satisfies (**D1**) by definition. Furthermore,

$$\begin{aligned} d_{\mathcal{H}}(AD_{\mu}^{\alpha(\beta)}+b, AD_{\nu}^{\alpha(\beta)}+b) &= \max\left\{ \sup_{x\in D_{\mu}^{\alpha(\beta)}} \inf_{y\in D_{\nu}^{\alpha(\beta)}} ||Ax-Ay||, \sup_{y\in D_{\nu}^{\alpha(\beta)}} \inf_{x\in D_{\mu}^{\alpha(\beta)}} ||Ax-Ay|| \right\} \\ &\overset{(ii)}{=} \max\left\{ \sup_{x\in D_{\mu}^{\alpha(\beta)}} \inf_{y\in D_{\nu}^{\alpha(\beta)}} ||x-y||, \sup_{y\in D_{\nu}^{\alpha(\beta)}} \inf_{x\in D_{\mu}^{\alpha(\beta)}} ||x-y|| \right\} \\ &= d_{\mathcal{H}}(D_{\mu}^{\alpha(\beta)}, D_{\nu}^{\alpha(\beta)}), \end{aligned}$$

where (ii) holds by virtue of hypothesis $AA^{\top}=I_d$. Replacing it in (9) yields the desired results.

B.3 Proof of Proposition 3.6

First assertion. Denote Z_1, Z_2 two random variables following μ^*, ν^* respectively. Assume that X, Y, Z_1, Z_2 are defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For any $x\in\mathbb{R}^d$ and $\beta\in[0, 1-\varepsilon]$,

$$\begin{aligned} x\in D_{\mu}^{\alpha(\beta)} \iff HD_{\mu}(x)\geq\alpha(\beta) &\iff \forall\, u\in\mathbb{S}^{d-1}, \quad \mathbb{P}\left(\langle u, X\rangle \leq \langle u, x\rangle\right) \geq \alpha(\beta) \\ &\iff \forall\, u\in\mathbb{S}^{d-1}, \quad \mathbb{P}\left(\langle u, Z_1+\mathbf{m}_1\rangle \leq \langle u, x\rangle\right) \geq \alpha(\beta) \\ &\iff \forall\, u\in\mathbb{S}^{d-1}, \quad \mathbb{P}\left(\langle u, Z_1\rangle \leq \langle u, x-\mathbf{m}_1\rangle\right) \geq \alpha(\beta) \\ &\iff x-\mathbf{m}_1\in D_{\mu^*}^{\alpha(\beta)} \end{aligned}$$

The same reasoning holds for ν and ν^*. Following this, for any $\beta\in[0,\ 1-\varepsilon]$ and $u\in\mathbb{S}^{d-1}$, it holds:

$$h_{D_{\mu}^{\alpha(\beta)}}(u) = h_{D_{\mu^*}^{\alpha(\beta)}}(u) - \langle u, \mathbf{m}_1\rangle \qquad \text{and} \qquad h_{D_{\nu}^{\alpha(\beta)}}(u) = h_{D_{\nu^*}^{\alpha(\beta)}}(u) - \langle u, \mathbf{m}_2\rangle$$

Thus it holds:

$$\begin{aligned} DR_{2,\varepsilon}^2(\mu,\nu) &= \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D_{\mu^*}^{\alpha(\beta)}}(u) - \langle u, \mathbf{m}_1\rangle - h_{D_{\nu^*}^{\alpha(\beta)}}(u) + \langle u, \mathbf{m}_2\rangle \right|^2 \mathrm{d}\beta \\ &\leq \sup_{u\in\mathbb{S}^{d-1}} \left|\langle u, \mathbf{m}_1-\mathbf{m}_2\rangle\right|^2 + \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D_{\mu^*}^{\alpha(\beta)}}(u) - h_{D_{\nu^*}^{\alpha(\beta)}}(u)\right|^2 \mathrm{d}\beta \\ &\quad + 2\sup_{u\in\mathbb{S}^{d-1}} \left|\langle u, \mathbf{m}_1-\mathbf{m}_2\rangle\right| \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D_{\mu^*}^{\alpha(\beta)}}(u) - h_{D_{\nu^*}^{\alpha(\beta)}}(u)\right| \mathrm{d}\beta \\ &= ||\mathbf{m}_1-\mathbf{m}_2||^2 + DR_{2,\varepsilon}^2(\mu^*,\nu^*) + 2||\mathbf{m}_1-\mathbf{m}_2|| DR_{1,\varepsilon}(\mu^*,\nu^*). \end{aligned} \tag{10}$$

On the other side, we have:

$$\begin{aligned} DR_{2,\varepsilon}^2(\mu,\nu) &\geq \sup_{u\in\mathbb{S}^{d-1}} \left|\langle u, \mathbf{m}_1-\mathbf{m}_2\rangle\right|^2 + \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D_{\mu^*}^{\alpha(\beta)}}(u) - h_{D_{\nu^*}^{\alpha}}(u)\right|^2 \mathrm{d}\beta \\ &\quad - 2\sup_{u\in\mathbb{S}^{d-1}} \left|\langle u, \mathbf{m}_1-\mathbf{m}_2\rangle\right| \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D_{\mu^*}^{\alpha(\beta)}}(u) - h_{D_{\nu^*}^{\alpha(\beta)}}(u)\right| \mathrm{d}\beta \\ &= ||\mathbf{m}_1-\mathbf{m}_2||^2 + DR_{2,\varepsilon}^2(\mu^*,\nu^*) - 2||\mathbf{m}_1-\mathbf{m}_2|| DR_{1,\varepsilon}(\mu^*,\nu^*). \end{aligned} \tag{11}$$

Combining (10) and (11) lead to the desired result.

Second assertion. For any $u\in\mathbb{S}^{d-1}$, the $(1-\alpha(\beta))$ quantiles of random variables $\langle u, X\rangle$ and $\langle u, Y\rangle$ such that $\langle u, X\rangle \sim \mathcal{N}(\langle u, \mathbf{m}_1\rangle, u^{\top}\mathbf{\Sigma}_1 u)$ and $\langle u, Y\rangle \sim \mathcal{N}(\langle u, \mathbf{m}_2\rangle, u^{\top}\mathbf{\Sigma}_2 u)$ are defined by

$$q_{\mu,u}^{1-\alpha(\beta)} = \langle u, \mathbf{m}_1 \rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{u^\top \mathbf{\Sigma}_1 u} \qquad q_{\nu,u}^{1-\alpha(\beta)} = \langle u, \mathbf{m}_2 \rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{u^\top \mathbf{\Sigma}_2 u},$$

where Φ is the cumulative distribution function of the univariate standard Gaussian distribution. Now, to apply Lemma A.2, it is sufficient to prove that directional quantiles are sublinear. It holds using subadditivity of the square root function. Indeed, for any $u, v \in \mathbb{S}^{d-1}$ and $\lambda > 0$, we have:

$$\begin{aligned}
\langle u + \lambda v, \mathbf{m}_1 \rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{(u+\lambda v)^\top \mathbf{\Sigma}_1 (u+\lambda v)} &= \langle u, \mathbf{m}_1 \rangle + \lambda \langle v, \mathbf{m}_1 \rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{(u+\lambda v)^\top \mathbf{\Sigma}_1 (u+\lambda v)} \\
&\leq \langle u, \mathbf{m}_1 \rangle + \lambda \langle v, \mathbf{m}_1 \rangle + \Phi^{-1}(1-\alpha(\beta)) \left[\sqrt{u^\top \mathbf{\Sigma}_1 u} + \lambda \sqrt{v^\top \mathbf{\Sigma}_1 v}\right] \\
&= q_{\mu,u}^{1-\alpha(\beta)} + \lambda\, q_{\mu,v}^{1-\alpha(\beta)}.
\end{aligned}$$

The same reasoning holds for ν. Applying Lemma A.1 and Lemma A.2, for any $u \in \mathbb{S}^{d-1}$, we have $h_{D_\mu^{\alpha(\beta)}}(u) = q_{\mu,u}^{1-\alpha(\beta)}$ and $h_{D_\nu^{\alpha(\beta)}}(u) = q_{\nu,u}^{1-\alpha(\beta)}$. It follows:

$$\begin{aligned}
DR_{1,\varepsilon}(\mu,\nu) &= \int_0^{1-\varepsilon} d_{\mathcal{H}}\left(D_\mu^{\alpha(\beta)}, D_\nu^{\alpha(\beta)}\right) \mathrm{d}\beta = \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D_\mu^{\alpha(\beta)}}(u) - h_{D_\nu^{\alpha(\beta)}}(u) \right| \mathrm{d}\beta \\
&= \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| \langle u, \mathbf{m}_1 - \mathbf{m}_2 \rangle + \Phi^{-1}(1-\alpha(\beta)) \left[\sqrt{u^\top \mathbf{\Sigma}_1 u} - \sqrt{u^\top \mathbf{\Sigma}_2 u}\right] \right| d\beta \\
&\leq ||\mathbf{m}_1 - \mathbf{m}_2|| + \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| \Phi^{-1}(1-\alpha(\beta)) \left[\sqrt{u^\top \mathbf{\Sigma}_1 u} - \sqrt{u^\top \mathbf{\Sigma}_2 u}\right] \right| \mathrm{d}\beta \\
&= ||\mathbf{m}_1 - \mathbf{m}_2|| + C_\varepsilon \sup_{u\in\mathbb{S}^{d-1}} \left| \sqrt{u^\top \mathbf{\Sigma}_1 u} - \sqrt{u^\top \mathbf{\Sigma}_2 u} \right|,
\end{aligned}$$

with $C_\varepsilon = \int_0^{1-\varepsilon} \left|\Phi^{-1}(1-\alpha(\beta))\right| \mathrm{d}\beta$. The lower bound is obtained by means the same reasoning. Notice that

$$||\mathbf{m}_1 - \mathbf{m}_2|| = \sup_{u\in\mathbb{S}^{d-1}} \left| \langle u, \mathbf{m}_1 - \mathbf{m}_2 \rangle \right| = \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| \langle u, \mathbf{m}_1 - \mathbf{m}_2 \rangle \right| \mathrm{d}\beta.$$

Introducing $h_{D_\mu^{\alpha(\beta)}}(u), h_{D_\nu^{\alpha(\beta)}}(u)$ and using triangular inequality, subadditivity of the supremum and linearity of the integral, we obtain:

$$||\mathbf{m}_1 - \mathbf{m}_2|| \leq DR_{1,\varepsilon}(\mu,\nu) + C_\varepsilon \sup_{u\in\mathbb{S}^{d-1}} \left| \sqrt{u^\top \mathbf{\Sigma}_1 u} - \sqrt{u^\top \mathbf{\Sigma}_2 u} \right|,$$

which ends the proof.

B.4 Proof of Proposition 3.7

For $DR_{p,\varepsilon}$ to break down at $\mathcal{S}_n$, it needs to have at least one trimmed-region that breaks down. Then the breakdown point of $DR_{p,\varepsilon}$ is higher than the minimum of the breakdown point of each region. Indeed, we have

$$\begin{aligned}
BP(DR_{p,\varepsilon}, \mathcal{S}_n) &= \min\left\{ \frac{o}{n+o} : \sup_{Z_1,\dots,Z_o} DR_{p,\varepsilon}\left(\hat{\mu}_{n+o}, \hat{\mu}_n\right) = +\infty \right\} \\
&\geq \min_{\beta\in[0,1-\varepsilon]} \min\left\{ \frac{o}{n+o} : \sup_{Z_1,\dots,Z_o} d_{\mathcal{H}}\left(D_{\hat{\mu}_{n+o}}^{\alpha(\beta,\hat{\mu}_{n+o})}, D_{\hat{\mu}_n}^{\alpha(\beta,\hat{\mu}_n)}\right) = +\infty \right\} \\
&= \min_{\beta\in[0,1-\varepsilon]} BP(D_{\hat{\mu}_n}^{\alpha(\beta,\hat{\mu}_n)}, \mathcal{S}_n).
\end{aligned}$$

Now applying Lemma 3.1 in Donoho and Gasko (1992) and Theorem 4 in Nagy and Dvořák (2021), a lower bound of the breakdown point of each halfspace region, for every $\beta \in [0, 1-\varepsilon]$, is given by

$$BP(D_{\hat{\mu}_n}^{\alpha(\beta,\hat{\mu}_n)}, \mathcal{S}_n) \geq \begin{cases} \frac{\lceil n\alpha(1-\varepsilon,\hat{\mu}_n)/(1-\alpha(1-\varepsilon,\hat{\mu}_n))\rceil}{n+\lceil n\alpha(1-\varepsilon,\hat{\mu}_n)/(1-\alpha(1-\varepsilon,\hat{\mu}_n))\rceil} & \text{if } \alpha(1-\varepsilon,\hat{\mu}_n) \leq \frac{\alpha_{\max}(\hat{\mu}_n)}{1+\alpha_{\max}(\hat{\mu}_n)}, \\ \frac{\alpha_{\max}(\hat{\mu}_n)}{1+\alpha_{\max}(\hat{\mu}_n)} & \text{otherwise,} \end{cases}$$

where $\alpha_{\max}(\hat{\mu}_n) = \max\limits_{x\in\mathbb{R}^d} \ HD_{\hat{\mu}_n}(x)$.

C APPROXIMATION ALGORITHMS

In this part, we display the approximation algorithms of the halfspace depth (see Algorithm 2), the projection depth (see Algorithm 3) and the AI-IRW depth (see Algorithm 4, proposed in Staerman et al., 2021b) used in the first step of the Algorithm 1.

Algorithm 2 Approximation of the halfspace depth

Initialization: $\mathbf{X} \in \mathbb{R}^{n\times d}$, K.

1: Construct $\mathbf{U} \in \mathbb{R}^{d\times K}$ by sampling uniformly K vectors $U_1, \dots, U_K$ in $\mathbb{S}^{d-1}$
2: Compute $\mathbf{M} = \mathbf{X}\mathbf{U}$
3: Compute the rank value $\sigma(i,k)$, the rank of index i in $\mathbf{M}_{:,k}$ for every $i \leq n$ and $k \leq K$
4: Set $D_i = \min\limits_{k\leq K} \ \sigma(i,k)$ for every $i \leq n$

Output: $D, \mathbf{M}$

Algorithm 3 Approximation of the projection depth

Initialization: $\mathbf{X} \in \mathbb{R}^{n\times d}$, K.

1: Construct $\mathbf{U} \in \mathbb{R}^{d\times K}$ by sampling uniformly K vectors $U_1, \dots, U_K$ in $\mathbb{S}^{d-1}$
2: Compute $\mathbf{M} = \mathbf{X}\mathbf{U}$
3: Find $\mathbf{M}_{\text{med},k}$ the median value of $\mathbf{M}_{:,k}$, $\forall\ k \leq K$
4: Compute $\text{MAD}_k = \text{median}\{\left|\mathbf{M}_{i,k} - \mathbf{M}_{\text{med},k}\right|,\ i \leq n\}$ for $k \leq K$
5: Compute $\mathbf{V}$ s.t. $\mathbf{V}_{i,k} = \left|\mathbf{M}_{i,k} - \mathbf{M}_{\text{med},k}\right|/\text{MAD}_k$
6: Set $D_i = \min\limits_{k\leq K} \ 1/(1+\mathbf{V}_{i,k})$ for every $i \leq n$

Output: $D, \mathbf{M}$

Algorithm 4 Approximation of the AI-IRW depth

Initialization: $\mathbf{X} \in \mathbb{R}^{n\times d}$, K.

1: Construct $\mathbf{U} \in \mathbb{R}^{d\times K}$ by sampling uniformly K vectors $U_1, \dots, U_K$ in $\mathbb{S}^{d-1}$
2: Compute $\widehat{\Sigma}$ using any estimator
3: Perform Cholesky or SVD on $\widehat{\Sigma}$ to obtain $\widehat{\Sigma}^{-1/2}$
4: Compute $\mathbf{V} = \widehat{\Sigma}^{-1/2}\mathbf{U}/||\widehat{\Sigma}^{-1/2}\mathbf{U}||$
5: Compute $\mathbf{M} = \mathbf{X}\mathbf{V}$
6: Compute the rank value $\sigma(i,k)$, the rank of index i in $\mathbf{M}_{:,k}$ for every $i \leq n$ and $k \leq K$
7: Set $D_i = \frac{1}{K}\sum_{k=1}^{K} \ \sigma(i,k)$ for every $i \leq n$

Output: $D, \mathbf{M}$

D ADDITIONAL EXPERIMENTS

D.1 Illustration of data depth contours

Figure 4, which plots a family of AI-IRW (using MCD estimator) depth induced trimmed-contours for a dataset contaminated with outliers, illustrates its robustness.

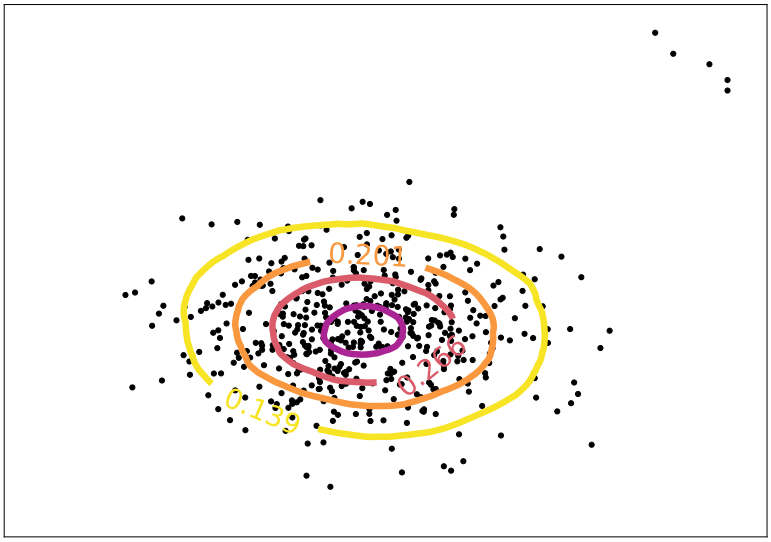

Figure 4: AI-IRW depth contours for a bivariate sample contaminated with outliers.

D.2 Illustration of the depth trimmed-regions based pseudo-Metric

Figure 5, which plots a family of (approximated) AI-IRW depth induced trimmed-regions for two datasets contaminated with outliers, illustrates the key idea of our proposed pseudo-metric.

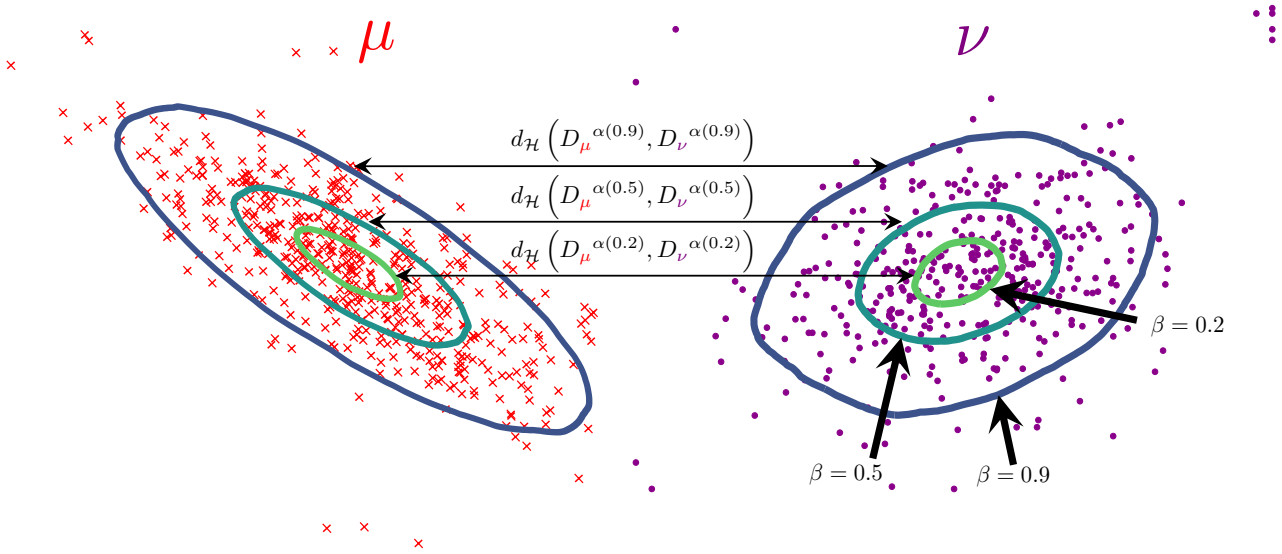

Figure 5: Illustration of the principle of the depth trimmed-regions based pseudo-metric.

D.3 Empirical analysis of statistical rates

Deriving theoretical finite-sample analysis may appear to be challenging for the proposed pseudo-metric. Thus, we numerically investigate the statistical convergence speed of $DR_{2,\varepsilon}$. To that end, we simulate two samples $\mathbf{X}$ and $\mathbf{Y}$ from two standard Gaussian distributions in dimension two with varying sample sizes. We compute the $DR_{2,\varepsilon}$ between $\mathbf{X}$ and $\mathbf{Y}$ with $n_\alpha \in \{5, 20, 100\}$ using the halfspace and the projection depths. Our proposed metric is computed with a high number of directions $K = 10000$ to isolate the statistical error. We report the estimation error (averaged over 10 runs, the true value of $DR_{2,\varepsilon}$ being equal to zero) in Figure 6. The experiment suggests that the statistical rates should be in $O(n^{1/4})$.

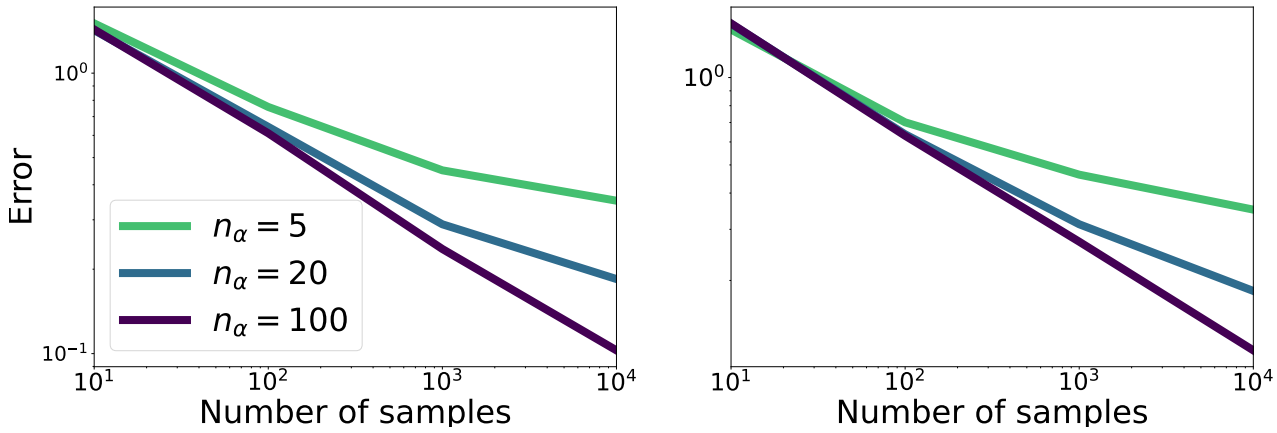

Figure 6: Empirical analysis of statistical convergence rates. Resulting error of the proposed pseudo-metric when increasing the sample size using the projection depth (left) and the halfspace depth (right) for various n_α parameters.

D.4 The influence of the parameter ε

The parameter ε plays the role of the robust tuning parameter of $DR_{2,\varepsilon}$. In this part, we complete our theoretical results provided in Section 3.2. We assess the robustness of our pseudo-metric making varying the parameter ε. Precisely, we simulate two normal samples $\mathbf{X}$ and $\mathbf{Y}$ from two standard Gaussian distributions in dimension two with a sample size of 10000. From that, we construct abnormal samples with a proportion of anomalies equal to $\{1\%, 10\%, 20\%\}$. To that end, we choose a proportion of normal samples and replace their first (for $\mathbf{X}$) and second (for $\mathbf{Y}$) coordinates as follows: $X_{\text{anom}} = 30 + 50Z$ and $Y_{\text{anom}} = -30 - 50Z$ where Z follows a uniform distribution on $[0, 1]$; leading to points far from the normal distributions. Thus, we compute $DR_{2,\varepsilon}$ with both robust and non-robust data depths, i.e. the projection and halfspace depths between $\mathbf{X}$ and $\mathbf{Y}$ being used as a benchmark. Further, we compute $DR_{2,\varepsilon}$ between abnormal samples and report mean error (comparing values obtained between normal samples and values obtained between abnormal samples; averaged over ten runs) on Figure 7. First, when computing with a robust depth function, we can see that the robustness of the proposed pseudo-metric relies directly on the parameter ε. This is shown by the presence of an elbow when the parameter ε reaches the level of the proportion of anomalies. In contrast, we can see that for a non-robust depth function such as the halfspace depth, our proposed pseudo-metric becomes non-robust once the abnormal proportion is higher than 1%, leading to a poorly robust depth. This experiment then confirms our theoretical results on the Breakdown Point of $DR_{p,\varepsilon}$ displayed in Propostion 3.7. The parameter ε provides robustness to our pseudo-metric when combined with a robust depth function.

D.5 The choice of the parameter n_α

Proposition 3.6 allows to derive a closed form expression for $DR_{2,\varepsilon}(\mu, \nu)$ when μ, ν are Gaussian distributions with the same variance-covariance matrix. In order to investigate the quality of the approximation on light-tailed and heavy-tailed distributions, we focus on computing $DR_{2,0.1}$ (with $K = 500$) for varying number of n_α between a sample of 1000 points stemming from $\mu \sim \mathcal{N}(\mathbf{0}_d, \Sigma)$ for $d \in \{2, 3, 10\}$, Σ drawn from the Wishart distribution (with parameters (d, I_d)) on the space of definite matrices and three different samples (which yields nine settings). These three samples are constructed from 1000 observations stemming from elliptically symmetric *Cauchy*, *Student-t_2* and *Gaussian* distributions all centered at $\mathbf{7}_d$. Results that report the averaged approximation error and the 25-75% empirical quantile intervals are depicted in Figure 8.

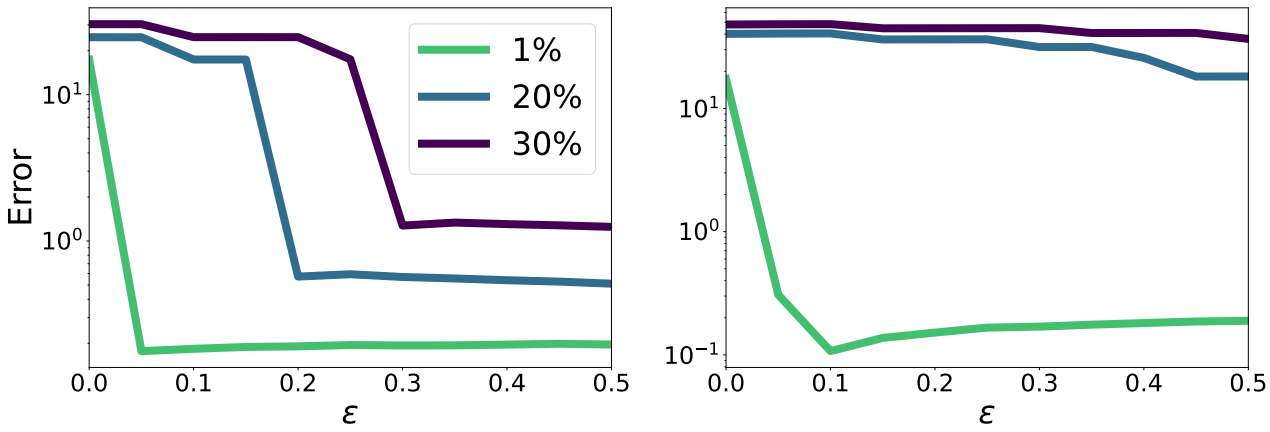

Figure 7: Influence of the parameter ε on the robustness of the proposed pseudo-metric with a robust depth function (the projection depth, left) and a non-robust one (the halfspace depth, right) for various proportion of anomalies.

They show that $DR_{p,\varepsilon}$ converges slowly for *Cauchy* with growing n_α, while it converges with small n_α for *Gaussian* and *Student-*t_2 distributions.

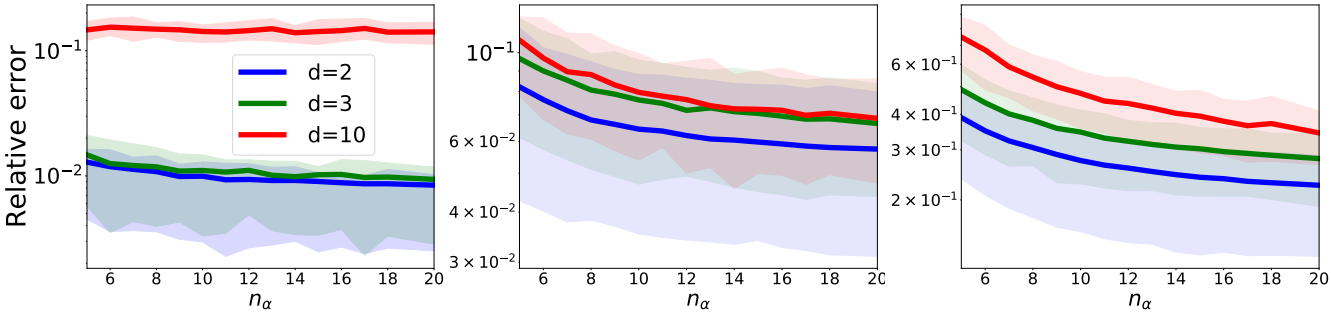

Figure 8: Relative approximation error (averaged over 100 repetitions, y-axis in log scale) of $DR_{p,\varepsilon}$ for elliptically symmetric *Cauchy* (left), *Student-*t_2 (middle) and *Gaussian* (right) distributions for differing numbers of n_α.

D.6 Robustness to outliers

Datasets on which experiments regarding "Robustness to outliers" in Section 5 have been performed are displayed in Figure 9.

E APPLICATIONS TO NLP

In this section, we gather details on experimental settings and additional results on the automatic evaluation of natural language generation (NLG).

E.1 Extended related works on automatic evaluation of NLG

Many metrics have been recently introduced for the automatic evaluation of text generation. In this work, we rely on untrained metrics. These metrics can be grouped into two categories: string-based metrics that depend on the string representation of the input texts to compute the similarity score and embedding-based metrics that rely on a continuous representation of the texts.

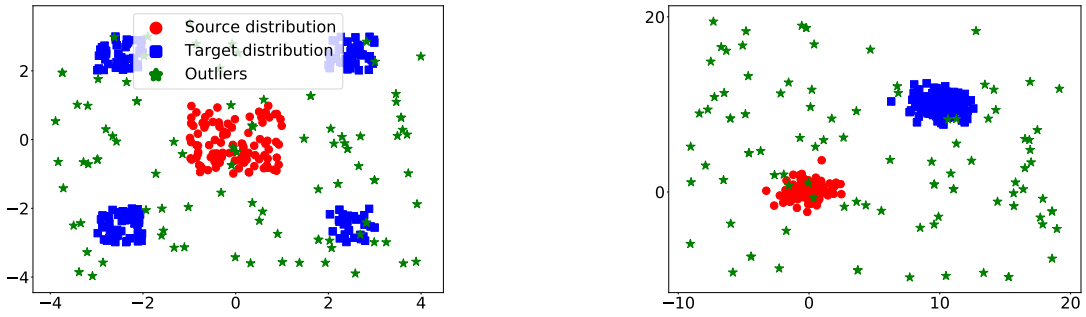

Figure 9: datasets related to robustness experiments depicted in Section 5 with 20% of outliers for *fragmented hypercube* (left) and *Gaussian* (right).

String matching metrics can be divided into two categories: N-gram matching and edit distance-based metrics. Perhaps the most used N-gram matching metrics are BLEU, ROUGE and METEOR. Edit distance-based metrics (e.g. TER; Snover et al., 2006) measure the distance as the number of basic operations such as 'edit'/'delete'/'insert'. Variants of TER include CHARACTERE (Wang et al., 2016), CDER (Leusch et al., 2006), EED (Stanchev et al., 2019). String-based metrics fail to produce meaningful scores in the case of paraphrases, especially if no common n-grams are found between the candidate and the reference text.

The second category of untrained metrics (namely embedding-based metrics) achieves state-of-the-art performance in many NLG evaluation tasks and has been introduced to address the issues mentioned above. Originally introduced for the widely used words embedding (Garcia et al., 2019; Colombo et al., 2019, 2020, 2021b) such as Word2Vec (Mikolov et al., 2013) or Glove (Pennington et al., 2014), this class of metrics has leveraged recently introduced contextualized word representations (CWR). CWR such as BERT, ELMO (Peters et al., 2018), HILAMOD (Chapuis et al., 2020, 2021) or ROBERTA (Liu et al., 2019b) are popular in NLP (Wiston et al., 2018) as they achieve SOTA performance on many tasks. The two most popular metrics are MoverScore and BertScore.

E.2 Evaluation

For the task of evaluation of text generation, we assume that we have access to a dataset $\{T_{R_i}, \{T^j_{G_i}, h(T^j_{G_i})\}_{j=1}^{n_S}\}_{i=1}^{n_T}$ where $T^j_{G_i}$ represents the i-th generated text by the j-th natural generation system, and $h(T^j_{G_i})$ represents score assigned by the human annotator[1] to $T^j_{G_i}$, and T_{R_i} is the reference text. n_T is the number of available texts, and n_S is the number of different systems.

To assess the relevance of an evaluation metric $\mathfrak{M}$, the correlation with the human judgment is considered one of the most important criteria (Banerjee and Lavie, 2005; Koehn, 2009; Chatzikoumi, 2020). To measure this correlation, two evaluation strategies are commonly adopted and built on top of a classical correlation measure, denoted C, e.g. Kendall (τ; Kendall, 1938), Pearson (r; Leusch et al., 2003) or Spearman (ρ; Melamed et al., 2003).

- *The text level correlation* (C_{text}) measures the ability of the metric to distinguish between badly and well generated text. Formally, C_{text} is defined as follows:

$$
\begin{aligned}
C_{text} &= \frac{1}{N_T}\sum_{i=1}^{n_T} C(\mathbf{M}_i^{text}, \mathbf{H}_i^{text}), \\
\mathbf{M}_i^{text} &= \left[\mathfrak{M}(T_{R_i}, T^1_{C_i}), \cdots, \mathfrak{M}(T_{R_i}, T^{n_S}_{C_i})\right], \\
\mathbf{H}_i^{text} &= \left[h(T^1_{C_i}), \cdots, h(T^{n_S}_{C_i})\right].
\end{aligned}
\tag{12}
$$

- *The system level correlation* (C_{sys}) assesses the ability of a metric to distinguish between good and bad systems.

[1] In practice an averaged score is considered as each sentence is annotated by 3 different annotators. The considered datasets directly provide the aggregated score.

Formally, C_{sys} is defined as follows:

$$
\begin{aligned}
C_{sys} =& C(\mathbf{M}^{sys}, \mathbf{H}^{sys}), \\
\mathbf{M}^{sys} =& \left[\frac{1}{n_T} \sum_{i=1}^{n_T} \mathfrak{M}(T_{R_i}, T^1_{C_i}), \cdots, \frac{1}{n_T} \sum_{i=1}^{n_T} \mathfrak{M}(T_{R_i}, T^{n_S}_{C_i}) \right], \\
\mathbf{H}^{sys} =& \left[\frac{1}{n_T} \sum_{i=1}^{n_T} h(T^1_{C_i}), \cdots, \frac{1}{n_T} \sum_{i=1}^{n_T} h(T^{n_S}_{C_i}) \right],
\end{aligned}
\tag{13}
$$

We refer the reader to Bhandari et al. (2020) for further details on the evaluation of text generation.

E.3 Results on Data2text

In this section, we gather further details and results on data2text automatic evaluation.

E.3.1 Task description

In WebNLG 2020, the goal is to create new efficient generation algorithms that can verbalise knowledge-based fragments. These algorithms are called Knowledge Base Verbalizers (Gardent et al., 2017) and are used during the micro-planning phase of NLG systems (Ferreira et al., 2018). WebNLG has been gathered to be more representative of the progress of recent NLG systems than previously existing task-oriented dialogue datasets (see e.g. SFHOTEL (Wen et al., 2015) and BAGEL (Mairesse et al., 2010)). As previously mentioned for the data2text task we work on the WebNLG2020 challenge (Gardent et al., 2017; Perez-Beltrachini et al., 2016). Data and system performances can be found in `https://webnlg-challenge.loria.fr/`. The task consists in mapping RDF triples to natural language (RDF format is used for many application including FOAF (see `http://www.foaf-project.org/`). For WebNLG 2020, the triplets are extracted from DBpedia (Auer et al., 2007). Data have been made freely available from the authors at `https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release_v3.0`. To compose this dataset, 15 systems (both symbolic and neural-based) have been used. The final dataset is composed of over 3k samples of human annotations (see `https://webnlg-challenge.loria.fr/files/WebNLG-2020-Presentation.pdf` for more details).

Example: Given the following triplet `(John_Blaha birthDate 1942_08_26) (John_Blaha birthPlace San_Antonio) (John_Blaha job Pilot)` the ground-truth reference is `John Blaha, born in San Antonio on 1942-08-26, worked as a pilot.`

E.3.2 Results

We gather in Table 3 complete results on the WebNLG tasks including results on ROUGE-2. To compare $DR_{p,\varepsilon}$ (with $\varepsilon = 0.01$, $n_\alpha = 5$, $p = 2$) with the different metrics (i.e. Wasserstein, Sliced-Wasserstein, MMD), we work on Roberta-based model from the HuggingFace hub (Wolf et al., 2019) and extract representation from the 11th layer. From Table 3, we observe a similar behavior from BertScore and MoverScore. This similarity has also been reported in a different setting in the previous work of Zhao et al. (2019). Overall, we observe that $DR_{p,\varepsilon}$ is always among its group's top-scoring metrics and achieves the best overall results on several configurations. It is worth noticing that $DR_{p,\varepsilon}$ only relies on information available in the candidate and the reference text. In contrast, BertScore and MoverScore use IDF information computed on every dataset.

E.4 Results on summarization

In this section, we gather experimental details and results on the automatic evaluation of the text summarization task.

E.4.1 Task description

Text summarization has attracted much attention in recent years (Zhang et al., 2020). Two types of models exist: *extractive* and *abstractive*. In extractive summarization, the system copies chunks of informative fragments from the input texts, whereas, in abstractive summarization, the system generates novel words. In this section, we describe our experimental setting. We present the tasks and the baseline metrics used for the automatic evaluation of summarization. We work with

	Correctness			Data Coverage			Relevance		
	r	τ	ρ	r	τ	ρ	r	τ	ρ
$DR_{p,\varepsilon}$	**89.4**	**80.0**	**92.6**	**84.2**	58.3	72.3	**86.2**	62.7	72.9
Wasserstein	86.2	73.0	86.7	80.4	45.3	62.3	83.8	51.3	67.6
Sliced-Wasserstein	86.1	73.0	85.8	80.9	45.5	60.0	82.0	51.3	68.2
MMD	25.4	71.7	8.3	19.1	45.3	10.0	26.1	51.3	15.0
BertScore	85.5	73.3	83.4	74.7	53.3	68.2	83.3	65.0	**79.4**
MoverScore	84.1	73.3	84.1	78.7	53.3	66.2	82.1	65.0	77.4
BLEU	77.6	60.0	66.3	55.7	36.6	50.2	63.0	51.6	65.2
ROUGE-1	80.6	65.0	65.0	76.5	**60.3**	**76.3**	64.3	56.7	69.2
ROUGE-2	73.6	58.3	63.3	54.7	35.0	43.1	62.0	46.7	60.8
METEOR	86.5	70.0	66.3	77.3	46.6	50.2	82.1	58.6	65.2
TER	79.6	58.0	78.3	69.7	38.0	58.2	75.0	**77.6**	70.2

Table 3: WebNLG 2020 (full results): absolute correlation at the system level with three human judgment criteria. Best overall results are indicated in bold, best results in their group are underlined.

the dataset from Bhandari et al. (2020) for this task. This dataset has been introduced to solve several flaws (Rankel et al., 2013) present in existing summarization datasets such as TAC (Dang and Owczarzak, 2008; McNamee and Dang, 2009). The dataset has been annotated using the pyramid score (Nenkova et al., 2007; Nenkova and Passonneau, 2004) and automatically built from the CNN/Daily News (Bhandari et al., 2020). It gathers 11 490 summaries coming from 11 extractive systems (See et al., 2017; Chen and Bansal, 2018; Raffel et al., 2019; Gehrmann et al., 2018; Dong et al., 2019; Liu and Lapata, 2019; Lewis et al., 2019; Yoon et al., 2020) and 14 abstractive systems (Zhou et al., 2018; Narayan et al., 2018; Kedzie et al., 2018; Zhong et al., 2019; Liu and Lapata, 2019; Dong et al., 2019; Wang et al., 2020; Zhong et al., 2020).

Example: The goal is to assign a similarity score between a reference text: "*Manchester United take on Manchester City on Sunday. Match will begin at 4 pm local time at United's Old Trafford home. Police have no objections to kick-off being so late in the afternoon. Last late afternoon weekend kick-off in the Manchester derby saw 34 fans arrested at Wembley in 2011 fa cup semi-final*" and the text generated by a NLG system: "*Manchester Derby takes place at Old Trafford on Sunday afternoon police have no objections to the late afternoon kick-off both sides are challenging for a top-four spot in the Premier League the man in charge of patrolling the sell-out clash has no such fears*".

E.4.2 Results

We gather in Table 4, the results on the summarization task. We use a bert-based uncased model and rely on the representations extracted from the 9th layer (similarly to BertScore). For this experiment the following parameters are used: $\varepsilon = 0.01$, $n_\alpha = 5$, $p = 2$. For this task, we can reproduce results from Bhandari et al. (2020) where the different behavior regarding the extractive and the abstractive systems is also observed. In this experiment, we observe that $DR_{p,\varepsilon}$ can achieve stronger results than other metrics based on Wasserstein, Sliced-Wasserstein and MMD. We also observe that $DR_{p,\varepsilon}$ outperforms MoverScore and BertScore on extractive systems (on r and τ). We believe these results support our approach.

	Abstractive			Extractive		
	r	τ	ρ	r	τ	ρ
$DR_{p,\varepsilon}$	<u>72.1</u>	<u>72.1</u>	70.1	<u>91.5</u>	**91.5**	<u>69.2</u>
Wasserstein	71.0	70.4	<u>71.1</u>	74.2	74.2	40.0
Sliced-Wasserstein	70.1	68.7	71.0	72.4	73.9	<u>69.2</u>
MMD	68.2	67.5	67.9	75.6	75.6	56.1
BertScore	71.7	<u>71.9</u>	72.0	70.9	72.9	**73.8**
MoverScore	<u>72.4</u>	<u>71.9</u>	<u>73.0</u>	<u>76.1</u>	<u>76.1</u>	47.4
ROUGE-1	**73.5**	73.0	**74.4**	72.2	<u>74.0</u>	<u>69.1</u>
ROUGE-2	73.0	**73.5**	73.0	55.1	53.2	<u>69.1</u>
JS-2	68.9	6.8	69.8	**92.9**	5.5	19.0

Table 4: Summarization: absolute correlation coefficients (using Pearson (r), Spearman (ρ) and Kendall (τ) coefficient) between different metrics on text summarization. Best overall results are indicated in bold, best results in their group are underlined.