Some Theoretical Insights into Wasserstein GANs

Gérard Biau, Maxime Sangnier, Ugo Tanielian

Some Theoretical Insights into Wasserstein GANs

Gérard Biau gerard.biau@sorbonne-universite.fr
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université
4 place Jussieu
75005 Paris, France

Maxime Sangnier maxime.sangnier@sorbonne-universite.fr
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université
4 place Jussieu
75005 Paris, France

Ugo Tanielian u.tanielian@criteo.com
Laboratoire de Probabilités, Statistique et Modélisation & Criteo AI Lab
Criteo AI Lab
32 rue Blanche
75009 Paris, France

Editor:

Abstract

Generative Adversarial Networks (GANs) have been successful in producing outstanding results in areas as diverse as image, video, and text generation. Building on these successes, a large number of empirical studies have validated the benefits of the cousin approach called Wasserstein GANs (WGANs), which brings stabilization in the training process. In the present paper, we add a new stone to the edifice by proposing some theoretical advances in the properties of WGANs. First, we properly define the architecture of WGANs in the context of integral probability metrics parameterized by neural networks and highlight some of their basic mathematical features. We stress in particular interesting optimization properties arising from the use of a parametric 1-Lipschitz discriminator. Then, in a statistically-driven approach, we study the convergence of empirical WGANs as the sample size tends to infinity, and clarify the adversarial effects of the generator and the discriminator by underlining some trade-off properties. These features are finally illustrated with experiments using both synthetic and real-world datasets.

Keywords: Generative Adversarial Networks, Wasserstein distances, deep learning theory, Lipschitz functions, trade-off properties

1. Introduction

Generative Adversarial Networks (GANs) is a generative framework proposed by Goodfellow et al. (2014), in which two models (a generator and a discriminator) act as adversaries in a zero-sum game. Leveraging the recent advances in deep learning, and specifically convolutional neural networks (LeCun et al., 1998), a large number of empirical studies have shown the impressive possibilities of GANs in the field of image generation (Radford

et al., 2015; Ledig et al., 2017; Karras et al., 2018; Brock et al., 2019). Lately, Karras et al. (2019) proposed an architecture able to generate hyper-realistic fake human faces that cannot be differentiated from real ones (see the website thispersondoesnotexist.com). The recent surge of interest in the domain also led to breakthroughs in video (Acharya et al., 2018), music (Mogren, 2016), and text generation (Yu et al., 2017; Fedus et al., 2018), among many other potential applications.

The aim of GANs is to generate data that look "similar" to samples collected from some unknown probability measure $\mu^\star$, defined on a Borel subset E of $\mathbb{R}^D$. In the targeted applications of GANs, E is typically a submanifold (possibly hard to describe) of a high-dimensional $\mathbb{R}^D$, which therefore prohibits the use of classical density estimation techniques. GANs approach the problem by making two models compete: the generator, which tries to imitate $\mu^\star$ using the collected data, vs. the discriminator, which learns to distinguish the outputs of the generator from the samples, thereby forcing the generator to improve its strategy.

Formally, the generator has the form of a parameterized class of Borel functions from $\mathbb{R}^d$ to E, say $\mathscr{G} = \{G_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^P$ is the set of parameters describing the model. Each function G_θ takes as input a d-dimensional random variable Z—it is typically uniform or Gaussian, with d usually small—and outputs the "fake" observation $G_\theta(Z)$ with distribution μ_θ. Thus, the collection of probability measures $\mathscr{P} = \{\mu_\theta : \theta \in \Theta\}$ is the natural class of distributions associated with the generator, and the objective of GANs is to find inside this class the distribution that generates the most realistic samples, closest to the ones collected from the unknown $\mu^\star$. On the other hand, the discriminator is described by a family of Borel functions from E to $[0,1]$, say $\mathscr{D} = \{D_\alpha : \alpha \in \Lambda\}$, $\Lambda \subseteq \mathbb{R}^Q$, where each D_α must be thought of as the probability that an observation comes from $\mu^\star$ (the higher $D(x)$, the higher the probability that x is drawn from $\mu^\star$).

In the original formulation of Goodfellow et al. (2014), GANs make $\mathscr{G}$ and $\mathscr{D}$ fight each other through the following objective:

$$\inf_{\theta\in\Theta} \sup_{\alpha\in\Lambda} \Big[\mathbb{E}\log(D_\alpha(X)) + \mathbb{E}\log(1 - D_\alpha(G_\theta(Z))) \Big], \tag{1}$$

where X is a random variable with distribution $\mu^\star$ and the symbol $\mathbb{E}$ denotes expectation. Since one does not have access to the true distribution, $\mu^\star$ is replaced in practice with the empirical measure μ_n based on independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_n$ distributed as X, and the practical objective becomes

$$\inf_{\theta\in\Theta} \sup_{\alpha\in\Lambda} \Big[\frac{1}{n}\sum_{i=1}^{n} \log(D_\alpha(X_i)) + \mathbb{E}\log(1 - D_\alpha(G_\theta(Z))) \Big]. \tag{2}$$

In the literature on GANs, both $\mathscr{G}$ and $\mathscr{D}$ take the form of neural networks (either feed-forward or convolutional, when dealing with image-related applications). This is also the case in the present paper, in which the generator and the discriminator will be parameterized by feed-forward neural networks with, respectively, rectifier (Glorot et al., 2011) and GroupSort (Chernodub and Nowicki, 2016) activation functions. We also note that from an optimization standpoint, the minimax optimum in (2) is found by using stochastic gradient descent alternatively on the generator's and the discriminator's parameters.

In the initial version (1), GANs were shown to reduce, under appropriate conditions, the Jensen-Shanon divergence between the true distribution and the class of parameterized distributions (Goodfellow et al., 2014). This characteristic was further explored by Biau et al. (2020), who stressed some theoretical guarantees regarding the approximation and statistical properties of problems (1) and (2). However, many empirical studies (e.g., Metz et al., 2016; Salimans et al., 2016) have described cases where the optimal generative distribution computed by solving (2) collapses to a few modes of the distribution $\mu^\star$. This phenomenon is known under the term of mode collapse and has been theoretically explained by Arjovsky and Bottou (2017). As a striking result, in cases where both $\mu^\star$ and μ_θ lie on disjoint supports, these authors proved the existence of a perfect discriminator with null gradient on both supports, which consequently does not convey meaningful information to the generator.

To cancel this drawback and stabilize training, Arjovsky et al. (2017) proposed a modification of criterion (1), with a framework called Wasserstein GANs (WGANs). In a nutshell, the objective of WGANs is to find, inside the class of parameterized distributions $\mathscr{P}$, the one that is the closest to the true $\mu^\star$ with respect to the Wasserstein distance (Villani, 2008). In its dual form, the Wasserstein distance can be considered as an integral probability metric (IPM, Mller, 1997) defined on the set of 1-Lipschitz functions. Therefore, the proposal of Arjovsky et al. (2017) is to replace the 1-Lipschitz functions with a discriminator parameterized by neural networks. To practically enforce this discriminator to be a subset of 1-Lipschitz functions, the authors use a weight clipping technique on the set of parameters. A decisive step has been taken by Gulrajani et al. (2017), who stressed the empirical advantage of the WGANs architecture by replacing the weight clipping with a gradient penalty. Since then, WGANs have been largely recognized and studied by the Machine Learning community (e.g., Roth et al., 2017; Petzka et al., 2018; Wei et al., 2018; Karras et al., 2019).

A natural question regards the theoretical ability of WGANs to learn $\mu^\star$, considering that one only has access to the parametric models of generative distributions and discriminative functions. Previous works in this direction are those of Liang (2018) and Zhang et al. (2018), who explore generalization properties of WGANs. In the present paper, we make one step further in the analysis of mathematical forces driving WGANs and contribute to the literature in the following ways:

(i) We properly define the architecture of WGANs parameterized by neural networks. Then, we highlight some properties of the IPM induced by the discriminator, and finally stress some basic mathematical features of the WGANs framework (Section 2).

(ii) We emphasize the impact of operating with a parametric discriminator contained in the set of 1-Lipschitz functions. We introduce in particular the notion of monotonous equivalence and discuss its meaning in the mechanism of WGANs. We also highlight the essential role played by piecewise linear functions (Section 3).

(iii) In a statistically-driven approach, we derive convergence rates for the IPM induced by the discriminator, between the target distribution $\mu^\star$ and the distribution output by the WGANs based on i.i.d. samples (Section 4).

(*iv*) Building upon the above, we clarify the adversarial effects of the generator and the discriminator by underlining some trade-off properties. These features are illustrated with experiments using both synthetic and real-world datasets (Section 5).

For the sake of clarity, proofs of the most technical results are gathered in the Appendix.

2. Wasserstein GANs

The present section is devoted to the presentation of the WGANs framework. After having given a first set of definitions and results, we stress the essential role played by IPMs and study some optimality properties of WGANs.

2.1 Notation and definitions

Throughout the paper, E is a Borel subset of $\mathbb{R}^D$, equipped with the Euclidean norm $\|\cdot\|$, on which $\mu^\star$ (the target probability measure) and the μ_θ's (the candidate probability measures) are defined. Depending on the practical context, E can be equal to $\mathbb{R}^D$, but it can also be a submanifold of it. We emphasize that there is no compactness assumption on E.

For $K \subseteq E$, we let $C(K)$ (respectively, $C_b(K)$) be the set of continuous (respectively, continuous bounded) functions from K to $\mathbb{R}$. We denote by Lip_1 the set of 1-Lipschitz real-valued functions on E, i.e.,

$$\mathrm{Lip}_1 = \big\{f : E \to \mathbb{R} : |f(x) - f(y)| \leqslant \|x - y\|,\ (x, y) \in E^2\big\}.$$

The notation $P(E)$ stands for the collection of Borel probability measures on E, and $P_1(E)$ for the subset of probability measures with finite first moment, i.e.,

$$P_1(E) = \Big\{\mu \in P(E) : \int_E \|x_0 - x\|\mu(\mathrm{dx}) < \infty\Big\},$$

where $x_0 \in E$ is arbitrary (this set does not depend on the choice of the point x_0). Until the end, it is assumed that $\mu^\star \in P_1(E)$. It is also assumed throughout that the random variable $Z \in \mathbb{R}^d$ is a sub-Gaussian random vector (Jin et al., 2019), i.e., Z is integrable and there exists $\gamma > 0$ such that

$$\forall v \in \mathbb{R}^d,\ \mathbb{E}e^{v\cdot(Z-\mathbb{E}Z)} \leqslant e^{\frac{\gamma^2\|v\|^2}{2}},$$

where $\cdot$ denotes the dot product in $\mathbb{R}^d$ and $\|\cdot\|$ the Euclidean norm. The sub-Gaussian property is a constraint on the tail of the probability distribution. As an example, Gaussian random variables on the real line are sub-Gaussian and so are bounded random vectors. We note that Z has finite moments of all nonnegative orders (Jin et al., 2019, Lemma 2). Assuming that Z is sub-Gaussian is a mild requirement since, in practice, its distribution is most of the time uniform or Gaussian.

As highlighted earlier, both the generator and the discriminator are assumed to be parameterized by feed-forward neural networks, that is,

$$\mathscr{G} = \{G_\theta : \theta \in \Theta\} \qquad \text{and} \qquad \mathscr{D} = \{D_\alpha : \alpha \in \Lambda\}$$

with $\Theta \subseteq \mathbb{R}^P$, $\Lambda \subseteq \mathbb{R}^Q$, and, for all $z \in \mathbb{R}^d$,

$$G_\theta(z) = \underset{D\times u_{p-1}}{U_p}\ \sigma\Big(\underset{u_{p-1}\times u_{p-2}}{U_{p-1}}\ \cdots \sigma\Big(\underset{u_2\times u_1}{U_2}\ \sigma\Big(\underset{u_1\times d}{U_1}\ z + \underset{u_1\times 1}{b_1}\Big) + \underset{u_2\times 1}{b_2}\Big)\cdots + \underset{u_{p-1}\times 1}{b_{p-1}}\Big) + \underset{D\times 1}{b_p}\,, \quad (3)$$

for all $x \in E$,

$$D_\alpha(x) = \underset{1\times v_{q-1}}{V_q}\ \tilde{\sigma}\Big(\underset{v_{q-1}\times v_{q-2}}{V_{q-1}}\ \cdots \tilde{\sigma}\Big(\underset{v_2\times v_1}{V_2}\ \tilde{\sigma}\Big(\underset{v_1\times D}{V_1}\ x + \underset{v_1\times 1}{c_1}\Big) + \underset{v_2\times 1}{c_2}\Big) + \cdots + \underset{v_{q-1}\times 1}{c_{q-1}}\Big) + \underset{1\times 1}{c_q}\,, \quad (4)$$

where $p, q \geqslant 2$ and the characters below the matrices indicate their dimensions (lines $\times$ columns). Some comments on the notation are in order. Networks in $\mathscr{G}$ and $\mathscr{D}$ have, respectively, $(p-1)$ and $(q-1)$ hidden layers. Hidden layers from depth 1 to $(p-1)$ (for the generator) and from depth 1 to $(q-1)$ (for the discriminator) are assumed to be of respective even widths u_i, $i = 1, \ldots, p-1$, and v_i, $i = 1, \ldots, q-1$. The matrices U_i (respectively, V_i) are the matrices of weights between layer i and layer $(i+1)$ of the generator (respectively, the discriminator), and the b_i's (respectively, the c_i's) are the corresponding offset vectors (in column format). We let $\sigma(x) = \max(x, 0)$ be the rectifier activation function (applied componentwise) and

$$\tilde{\sigma}(x_1, x_2, \ldots, x_{2n-1}, x_{2n}) = (\max(x_1, x_2), \min(x_1, x_2), \ldots, \max(x_{2n-1}, x_{2n}), \min(x_{2n-1}, x_{2n}))$$

be the GroupSort activation function with a grouping size equal to 2 (applied on pairs of components, which makes sense in (4) since the widths of the hidden layers are assumed to be even). GroupSort has been introduced in Chernodub and Nowicki (2016) as a 1-Lipschitz activation function that preserves the gradient norm of the input. This activation can recover the rectifier, in the sense that $\tilde{\sigma}(x, 0) = (\sigma(x), -\sigma(-x))$, but the converse is not true. The presence of GroupSort is critical to guarantee approximation properties of Lipschitz neural networks (Anil et al., 2019), as we will see later.

Therefore, denoting by $\mathscr{M}_{(j,k)}$ the space of matrices with j rows and k columns, we have $U_1 \in \mathscr{M}_{(u_1,d)}$, $V_1 \in \mathscr{M}_{(v_1,D)}$, $b_1 \in \mathscr{M}_{(u_1,1)}$, $c_1 \in \mathscr{M}_{(v_1,1)}$, $U_p \in \mathscr{M}_{(D,u_{p-1})}$, $V_q \in \mathscr{M}_{(1,v_{q-1})}$, $b_p \in \mathscr{M}_{(D,1)}$, $c_q \in \mathscr{M}_{(1,1)}$. All the other matrices U_i, $i = 2, \ldots, p-1$, and V_i, $i = 2, \ldots, q-1$, belong to $\mathscr{M}_{(u_i,u_{i-1})}$ and $\mathscr{M}_{(v_i,v_{i-1})}$, and vectors b_i, $i = 2, \ldots, p-1$, and c_i, $i = 2, \ldots, q-1$, belong to $\mathscr{M}_{(u_i,1)}$ and $\mathscr{M}_{(v_i,1)}$. So, altogether, the vectors $\theta = (U_1, \ldots, U_p, b_1, \ldots, b_p)$ (respectively, the vectors $\alpha = (V_1, \ldots, V_q, c_1, \ldots, c_q)$) represent the parameter space Θ of the generator $\mathscr{G}$ (respectively, the parameter space Λ of the discriminator $\mathscr{D}$). We stress the fact that the outputs of networks in $\mathscr{D}$ are not restricted to $[0, 1]$ anymore, as is the case for the original GANs of Goodfellow et al. (2014). We also recall the notation $\mathscr{P} = \{\mu_\theta : \theta \in \Theta\}$, where, for each θ, μ_θ is the probability distribution of $G_\theta(Z)$. Since Z has finite first moment and each G_θ is piecewise linear, it is easy to see that $\mathscr{P} \subset P_1(E)$.

Throughout the manuscript, the notation $\|\cdot\|$ (respectively, $\|\cdot\|_\infty$) means the Euclidean (respectively, the supremum) norm on $\mathbb{R}^k$, with no reference to k as the context is clear. For $W = (w_{i,j})$ a matrix in $\mathscr{M}_{(k_1,k_2)}$, we let $\|W\|_2 = \sup_{\|x\|=1} \|Wx\|$ be the 2-norm of W. Similarly, the ∞-norm of W is $\|W\|_\infty = \sup_{\|x\|_\infty=1} \|Wx\|_\infty = \max_{i=1,\ldots,k_1} \sum_{j=1}^{k_2} |w_{i,j}|$. We will also use the $(2, \infty)$-norm of W, i.e., $\|W\|_{2,\infty} = \sup_{\|x\|=1} \|Wx\|_\infty$. We shall constantly need the following assumption:

Assumption 1 *For all* $\theta = (U_1, \dots, U_p, b_1, \dots, b_p) \in \Theta$,

$$\max(\|U_i\|_2, \|b_i\|_2 : i = 1, \dots, p) \leqslant K_1,$$

where $K_1 > 0$ *is a constant. Besides, for all* $\alpha = (V_1, \dots, V_q, c_1, \dots, c_q) \in \Lambda$,

$$\|V_1\|_{2,\infty} \leqslant 1, \ \max(\|V_2\|_\infty, \dots, \|V_q\|_\infty) \leqslant 1, \ \text{and} \ \max(\|c_i\|_\infty : i = 1, \dots, q) \leqslant K_2,$$

where $K_2 \geqslant 0$ *is a constant.*

This compactness requirement is classical when parameterizing WGANs (e.g., Arjovsky et al., 2017; Zhang et al., 2018; Anil et al., 2019). In practice, one can satisfy Assumption 1 by clipping the parameters of neural networks as proposed by Arjovsky et al. (2017). An alternative approach to enforce $\mathscr{D} \subseteq \mathrm{Lip}_1$ consists in penalizing the gradient of the discriminative functions, as proposed by Gulrajani et al. (2017), Kodali et al. (2017), Wei et al. (2018), and Zhou et al. (2019). This solution was empirically found to be more stable. The usefulness of Assumption 1 is captured by the following lemma.

Lemma 1 *Assume that Assumption 1 is satisfied. Then, for each* $\theta \in \Theta$, *the function* G_θ *is* K_1^p*-Lipschitz on* $\mathbb{R}^d$. *In addition,* $\mathscr{D} \subseteq \mathrm{Lip}_1$.

Recall (e.g., Dudley, 2004) that a sequence of probability measures (μ_k) on E is said to converge weakly to a probability measure μ on E if, for all $\varphi \in C_b(E)$,

$$\int_E \varphi \, \mathrm{d}\mu_k \underset{k\to\infty}{\to} \int_E \varphi \, \mathrm{d}\mu.$$

In addition, the sequence of probability measures (μ_k) in $P_1(E)$ is said to converge weakly in $P_1(E)$ to a probability measure μ in $P_1(E)$ if (i) (μ_k) converges weakly to μ and if (ii) $\int_E \|x_0 - x\| \mu_k(\mathrm{d}x) \to \int_E \|x_0 - x\| \mu(\mathrm{d}x)$, where $x_0 \in E$ is arbitrary (Villani, 2008, Definition 6.7). The next proposition offers a characterization of our collection of generative distributions $\mathscr{P}$ in terms of compactness with respect to the weak topology in $P_1(E)$. This result is interesting as it gives some insight into the class of probability measures generated by neural networks.

Proposition 2 *Assume that Assumption 1 is satisfied. Then the function* $\Theta \ni \theta \mapsto \mu_\theta$ *is continuous with respect to the weak topology in* $P_1(E)$, *and the set of generative distributions* $\mathscr{P}$ *is compact with respect to the weak topology in* $P_1(E)$.

2.2 The WGANs and T-WGANs problems

We are now in a position to formally define the WGANs problem. The Wasserstein distance (of order 1) between two probability measures μ and ν in $P_1(E)$ is defined by

$$W_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{E\times E} \|x - y\| \pi(\mathrm{d}x, \mathrm{d}y),$$

where $\Pi(\mu, \nu)$ denotes the collection of all joint probability measures on $E \times E$ with marginals μ and ν (e.g., Villani, 2008). It is a finite quantity. In the present article, we will use the

dual representation of $W_1(\mu,\nu)$, which comes from the duality theorem of Kantorovich and Rubinstein (1958):

$$W_1(\mu,\nu) = \sup_{f\in \mathrm{Lip}_1} |\mathbb{E}_\mu f - \mathbb{E}_\nu f|,$$

where, for a probability measure π, $\mathbb{E}_\pi f = \int_E f \mathrm{d}\pi$ (note that for $f \in \mathrm{Lip}_1$ and $\pi \in P_1(E)$, the function f is Lebesgue integrable with respect to π).

In this context, it is natural to define the theoretical-WGANs (T-WGANs) problem as minimizing over Θ the Wasserstein distance between $\mu^\star$ and the μ_θ's, i.e.,

$$\inf_{\theta\in\Theta} W_1(\mu^\star,\mu_\theta) = \inf_{\theta\in\Theta} \sup_{f\in \mathrm{Lip}_1} |\mathbb{E}_{\mu^\star} f - \mathbb{E}_{\mu_\theta} f|. \tag{5}$$

In practice, however, one does not have access to the class of 1-Lipschitz functions, which cannot be parameterized. Therefore, following Arjovsky et al. (2017), the class Lip_1 is restricted to the smaller but parametric set of discriminators $\mathscr{D} = \{D_\alpha : \alpha \in \Lambda\}$ (it is a subset of Lip_1, by Lemma 1), and this defines the actual WGANs problem:

$$\inf_{\theta\in\Theta} \sup_{\alpha\in\Lambda} |\mathbb{E}_{\mu^\star} D_\alpha - \mathbb{E}_{\mu_\theta} D_\alpha|. \tag{6}$$

Problem (6) is the Wasserstein counterpart of problem (1). Provided Assumption 1 is satisfied, $\mathscr{D} \subseteq \mathrm{Lip}_1$, and the IPM (Mller, 1997) $d_\mathscr{D}$ is defined for $(\mu,\nu) \in P_1(E)^2$ by

$$d_\mathscr{D}(\mu,\nu) = \sup_{f\in\mathscr{D}} |\mathbb{E}_\mu f - \mathbb{E}_\nu f|. \tag{7}$$

With this notation, $d_{\mathrm{Lip}_1} = W_1$ and problems (5) and (6) can be rewritten as the minimization over Θ of, respectively, $d_{\mathrm{Lip}_1}(\mu^\star,\mu_\theta)$ and $d_\mathscr{D}(\mu^\star,\mu_\theta)$. So,

$$\text{T-WGANs: } \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star,\mu_\theta) \quad \text{and} \quad \text{WGANs: } \inf_{\theta\in\Theta} d_\mathscr{D}(\mu^\star,\mu_\theta).$$

Similar objectives have been proposed in the literature, in particular neural net distances (Arora et al., 2017) and adversarial divergences (Liu et al., 2017). These two general approaches include f-GANs (Goodfellow et al., 2014; Nowozin et al., 2016), but also WGANs (Arjovsky et al., 2017), MMD-GANs (Li et al., 2017), and energy-based GANs (Zhao et al., 2017). Using the terminology of Arora et al. (2017), $d_\mathscr{D}$ is called a neural IPM. If the theoretical properties of the Wasserstein distance d_{Lip_1} have been largely studied (e.g., Villani, 2008), the story is different for neural IPMs. This is why our next subsection is devoted to the properties of $d_\mathscr{D}$.

2.3 Some properties of the neural IPM $d_\mathscr{D}$

The study of the neural IPM $d_\mathscr{D}$ is essential to assess the driving forces of WGANs architectures. Let us first recall that a mapping $\ell : P_1(E) \times P_1(E) \to [0,\infty)$ is a metric if it satisfies the following three requirements:

(i) $\ell(\mu,\nu) = 0 \iff \mu = \nu$ (discriminative property)

(ii) $\ell(\mu,\nu) = \ell(\nu,\mu)$ (symmetry)

(iii) $\ell(\mu,\nu) \leqslant \ell(\mu,\pi) + \ell(\pi,\nu)$ (triangle inequality).

If (i) is replaced by the weaker requirement $\ell(\mu,\mu) = 0$ for all $\mu \in P_1(E)$, then one speaks of a pseudometric. Furthermore, the (pseudo)metric ℓ is said to metrize weak convergence in $P_1(E)$ (Villani, 2008) if, for all sequences (μ_k) in $P_1(E)$ and all μ in $P_1(E)$, one has $\ell(\mu,\mu_k) \to 0 \iff \mu_k$ converges weakly to μ in $P_1(E)$ as $k \to \infty$. According to Villani (2008, Theorem 6.8), d_{Lip_1} is a metric that metrizes weak convergence in $P_1(E)$.

As far as $d_{\mathscr{D}}$ is concerned, it is clearly a pseudometric on $P_1(E)$ as soon as Assumption 1 is satisfied. Moreover, an elementary application of Dudley (2004, Lemma 9.3.2) shows that if $\mathrm{span}(\mathscr{D})$ (with $\mathrm{span}(\mathscr{D}) = \{\gamma_0 + \sum_{i=1}^n \gamma_i D_i : \gamma_i \in \mathbb{R}, D_i \in \mathscr{D}, n \in \mathbb{N}\}$) is dense in $C_b(E)$, then $d_{\mathscr{D}}$ is a metric on $P_1(E)$, which, in addition, metrizes weak convergence. As in Zhang et al. (2018), Dudley's result can be exploited in the case where the space E is compact to prove that, whenever $\mathscr{D}$ is of the form (4), $d_{\mathscr{D}}$ is a metric metrizing weak convergence. However, establishing the discriminative property of the pseudometric $d_{\mathscr{D}}$ turns out to be more challenging without an assumption of compactness on E, as is the case in the present study. Our result is encapsulated in the following proposition.

Proposition 3 *Assume that Assumption 1 is satisfied. Then there exists a discriminator of the form (4) (i.e., a depth q and widths $v_1,\dots,v_{q-1}$) such that $d_{\mathscr{D}}$ is a metric on $\mathscr{P} \cup \{\mu^\star\}$. In addition, $d_{\mathscr{D}}$ metrizes weak convergence in $\mathscr{P} \cup \{\mu^\star\}$.*

Standard universal approximation theorems (Cybenko, 1989; Hornik et al., 1989; Hornik, 1991) state the density of neural networks in the family of continuous functions defined on compact sets but do not guarantee that the approximator respects a Lipschitz constraint. The proof of Proposition 3 uses the fact that, under Assumption 1, neural networks of the form (4) are dense in the space of Lipschitz continuous functions on compact sets, as revealed by Anil et al. (2019).

We deduce from Proposition 3 that, under Assumption 1, provided enough capacity, the pseudometric $d_{\mathscr{D}}$ can be topologically equivalent to d_{Lip_1} on $\mathscr{P} \cup \{\mu^\star\}$, i.e., the convergent sequences in $(\mathscr{P} \cup \{\mu^\star\}, d_{\mathscr{D}})$ are the same as the convergent sequences in $(\mathscr{P} \cup \{\mu^\star\}, d_{\mathrm{Lip}_1})$ with the same limit—see O'Searcoid (2006, Corollary 13.1.3). We are now ready to discuss some optimality properties of the T-WGANs and WGANs problems, i.e., conditions under which the infimum in $\theta \in \Theta$ and the supremum in $\alpha \in \Lambda$ are reached.

2.4 Optimality properties

Recall that for T-WGANs, we minimize over Θ the distance

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) = \sup_{f \in \mathrm{Lip}_1} |\mathbb{E}_{\mu^\star} f - \mathbb{E}_{\mu_\theta} f|,$$

whereas for WGANs, we use

$$d_{\mathscr{D}}(\mu^\star, \mu_\theta) = \sup_{\alpha \in \Lambda} |\mathbb{E}_{\mu^\star} D_\alpha - \mathbb{E}_{\mu_\theta} D_\alpha|.$$

A first natural question is to know whether for a fixed generator parameter $\theta \in \Theta$, there exists a 1-Lipschitz function (respectively, a discriminative function) that achieves the supremum in $d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ (respectively, in $d_{\mathscr{D}}(\mu^\star, \mu_\theta)$) over all $f \in \mathrm{Lip}_1$ (respectively, all $\alpha \in \Lambda$).

For T-WGANs, Villani (2008, Theorem 5.9) guarantees that the maximum exists, i.e.,

$$\{f \in \mathrm{Lip}_1 : |\mathbb{E}_{\mu^\star} f - \mathbb{E}_{\mu_\theta} f| = d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)\} \neq \varnothing. \tag{8}$$

For WGANs, we have the following:

Lemma 4 *Assume that Assumption 1 is satisfied. Then, for all $\theta \in \Theta$,*

$$\{\alpha \in \Lambda : |\mathbb{E}_{\mu^\star} D_\alpha - \mathbb{E}_{\mu_\theta} D_\alpha| = d_{\mathscr{D}}(\mu^\star, \mu_\theta)\} \neq \varnothing.$$

Thus, provided Assumption 1 is verified, the supremum in α in the neural IPM $d_{\mathscr{D}}$ is always reached. A similar result is proved by Biau et al. (2020) in the case of standard GANs.

We now turn to analyzing the existence of the infimum in θ in the minimization over Θ of $d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ and $d_{\mathscr{D}}(\mu^\star, \mu_\theta)$. Since the optimization scheme is performed over the parameter set Θ, it is worth considering the following two functions:

$$\xi_{\mathrm{Lip}_1} : \Theta \to \mathbb{R} \qquad \theta \mapsto d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \qquad \text{and} \qquad \xi_{\mathscr{D}} : \Theta \to \mathbb{R} \qquad \theta \mapsto d_{\mathscr{D}}(\mu^\star, \mu_\theta).$$

Theorem 5 *Assume that Assumption 1 is satisfied. Then ξ_{Lip_1} and $\xi_{\mathscr{D}}$ are Lipschitz continuous on Θ, and the Lipschitz constant of $\xi_{\mathscr{D}}$ is independent of $\mathscr{D}$.*

Theorem 5 extends Arjovsky et al. (2017, Theorem 1), which states that $d_{\mathscr{D}}$ is locally Lipschitz continuous under the additional assumption that E is compact. In contrast, there is no compactness hypothesis in Theorem 5 and the Lipschitz property is global. The lipschitzness of the function $\xi_{\mathscr{D}}$ is an interesting property of WGANS, in line with many recent empirical works that have shown that gradient-based regularization techniques are efficient for stabilizing the training of GANs and preventing mode collapse (Kodali et al., 2017; Roth et al., 2017; Miyato et al., 2018; Petzka et al., 2018).

In the sequel, we let $\Theta^\star$ and $\bar{\Theta}$ be the sets of optimal parameters, defined by

$$\Theta^\star = \operatorname*{arg\,min}_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \quad \text{and} \quad \bar{\Theta} = \operatorname*{arg\,min}_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta).$$

An immediate but useful corollary of Theorem 5 is as follows:

Corollary 6 *Assume that Assumption 1 is satisfied. Then $\Theta^\star$ and $\bar{\Theta}$ are non empty.*

Thus, any $\theta^\star \in \Theta^\star$ (respectively, any $\bar{\theta} \in \bar{\Theta}$) is an optimal parameter for the T-WGANs (respectively, the WGANs) problem. Note however that, without further restrictive assumptions on the models, we cannot ensure that $\Theta^\star$ or $\bar{\Theta}$ are reduced to singletons.

3. Optimization properties

We are interested in this section in the error made when minimizing over Θ the pseudo-metric $d_{\mathscr{D}}(\mu^\star, \mu_\theta)$ (WGANs problem) instead of $d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ (T-WGANs problem). This optimization error is represented by the difference

$$\varepsilon_{\mathrm{optim}} = \sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) - \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta).$$

It is worth pointing out that we take the supremum over all $\bar{\theta} \in \bar{\Theta}$ since there is no guarantee that two distinct elements $\bar{\theta}_1$ and $\bar{\theta}_2$ of $\bar{\Theta}$ lead to the same distances $d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_1})$ and $d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_2})$. The quantity $\varepsilon_{\mathrm{optim}}$ captures the largest discrepancy between the scores achieved by distributions solving the WGANs problem and the scores of distributions solving the T-WGANs problem. We emphasize that the scores are quantified by the Wasserstein distance d_{Lip_1}, which is the natural metric associated with the problem. We note in particular that $\varepsilon_{\mathrm{optim}} \geqslant 0$. A natural question is whether we can upper bound the difference and obtain some control of $\varepsilon_{\mathrm{optim}}$.

3.1 Approximating d_{Lip_1} with $d_{\mathscr{D}}$

As a warm-up, we observe that in the simple but unrealistic case where $\mu^\star \in \mathscr{P}$, provided Assumption 1 is satisfied and the neural IPM $d_{\mathscr{D}}$ is a metric on $\mathscr{P}$ (see Proposition 3), then $\Theta^\star = \bar{\Theta}$ and $\varepsilon_{\mathrm{optim}} = 0$. However, in the high-dimensional context of WGANs, the parametric class of distributions $\mathscr{P}$ is likely to be "far" from the true distribution $\mu^\star$. This phenomenon is thoroughly discussed in Arjovsky and Bottou (2017, Lemma 2 and Lemma 3) and is often referred to as dimensional misspecification (Roth et al., 2017).

From now on, we place ourselves in the general setting where we have no information on whether the true distribution belongs to $\mathscr{P}$, and start with the following simple observation. Assume that Assumption 1 is satisfied. Then, clearly, since $\mathscr{D} \subseteq \mathrm{Lip}_1$,

$$\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta). \tag{9}$$

Inequality (9) is useful to upper bound $\varepsilon_{\mathrm{optim}}$. Indeed,

$$\begin{aligned}
0 \leqslant \varepsilon_{\mathrm{optim}} &= \sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) - \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\
&\leqslant \sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \\
&= \sup_{\bar{\theta}\in\bar{\Theta}} \big[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) - d_{\mathscr{D}}(\mu^\star, \mu_{\bar{\theta}})\big] \\
&\quad \big(\text{since } \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) = d_{\mathscr{D}}(\mu^\star, \mu_{\bar{\theta}}) \text{ for all } \bar{\theta} \in \bar{\Theta}\big) \\
&\leqslant T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}),
\end{aligned} \tag{10}$$

where, by definition,

$$T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) = \sup_{\theta\in\Theta}\big[d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta)\big] \tag{11}$$

is the maximum difference in distances on the set of candidate probability distributions in $\mathscr{P}$. Note, since Θ is compact (by Assumption 1) and ξ_{Lip_1} and $\xi_{\mathscr{D}}$ are Lipschitz continuous (by Theorem 5), that $T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}) < \infty$. Thus, the loss in performance when comparing T-WGANs and WGANs can be upper-bounded by the maximum difference over $\mathscr{P}$ between the Wasserstein distance and $d_{\mathscr{D}}$.

Observe that when the class of discriminative functions is increased (say $\mathscr{D} \subset \mathscr{D}'$) while keeping the generator fixed, then the bound (11) gets reduced since $d_{\mathscr{D}}(\mu^\star, \cdot) \leqslant d_{\mathscr{D}'}(\mu^\star, \cdot)$. Similarly, when increasing the class of generative distributions (say $\mathscr{P} \subset \mathscr{P}'$) with a fixed discriminator, then the bound gets bigger, i.e., $T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}) \leqslant T_{\mathscr{P}'}(\mathrm{Lip}_1, \mathscr{D})$. It is important to note that the conditions $\mathscr{D} \subset \mathscr{D}'$ and/or $\mathscr{P} \subset \mathscr{P}'$ are easily satisfied for classes of functions parameterized with neural networks using either rectifier or GroupSort activation functions, just by increasing the width and/or the depth of the networks.

Our next theorem states that, as long as the distributions of $\mathscr{P}$ are generated by neural networks with bounded parameters (Assumption 1), then one can control $T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D})$.

Theorem 7 *Assume that Assumption 1 is satisfied. Then, for all $\varepsilon > 0$, there exists a discriminator $\mathscr{D}$ of the form (4) such that*

$$0 \leqslant \varepsilon_{\mathrm{optim}} \leqslant T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}) \leqslant c\varepsilon,$$

where $c > 0$ is a constant independent from ε.

Theorem 7 is important because it shows that for any collection of generative distributions $\mathscr{P}$ and any approximation threshold ε, one can find a discriminator such that the loss in performance $\varepsilon_{\mathrm{optim}}$ is (at most) of the order of ε. In other words, there exists $\mathscr{D}$ of the form (4) such that $T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D})$ is arbitrarily small. We note however that Theorem 7 is an existence theorem that does not give any particular information on the depth and/or the width of the neural networks in $\mathscr{D}$. The key argument to prove Theorem 7 is Anil et al. (2019, Theorem 3), which states that the set of Lipschitz neural networks are dense in the set of Lipschitz continuous functions on a compact space.

3.2 Equivalence properties

The quantity $T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D})$ is of limited practical interest, as it involves a supremum over all $\theta \in \Theta$. Moreover, another caveat is that the definition of $\varepsilon_{\mathrm{optim}}$ assumes that one has access to $\bar{\Theta}$. Therefor, our next goal is to enrich Theorem 7 by taking into account the fact that numerical procedures do not reach $\bar{\theta} \in \bar{\Theta}$ but rather an ϵ-approximation of it.

One way to approach the problem is to look for another form of equivalence between d_{Lip_1} and $d_{\mathscr{D}}$. As one is optimizing $d_{\mathscr{D}}$ instead of d_{Lip_1}, we would ideally like that the two IPMs behave "similarly", in the sense that minimizing $d_{\mathscr{D}}$ leads to a solution that is still close to the true distribution with respect to d_{Lip_1}. Assuming that Assumption 1 is satisfied, we let, for any $\mu \in P_1(E)$ and $\varepsilon > 0$, $\mathscr{M}_\ell(\mu, \varepsilon)$ be the set of ϵ-solutions to the optimization problem of interest, that is the subset of Θ defined by

$$\mathscr{M}_\ell(\mu, \varepsilon) = \big\{\theta \in \Theta : \ell(\mu, \mu_\theta) - \inf_{\theta \in \Theta} \ell(\mu, \mu_\theta) \leqslant \varepsilon\big\},$$

with $\ell = d_{\mathrm{Lip}_1}$ or $\ell = d_{\mathscr{D}}$.

Definition 8 *Let $\varepsilon > 0$. We say that d_{Lip_1} can be ε-substituted by $d_{\mathscr{D}}$ if there exists $\delta > 0$ such that*

$$\mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta) \subseteq \mathscr{M}_{d_{\mathrm{Lip}_1}}(\mu^\star, \varepsilon).$$

In addition, if d_{Lip_1} can be ε-substituted by $d_{\mathscr{D}}$ for all $\varepsilon > 0$, we say that d_{Lip_1} can be fully substituted by $d_{\mathscr{D}}$.

The rationale behind this definition is that by minimizing the neural IPM $d_{\mathscr{D}}$ close to optimality, one can be guaranteed to be also close to optimality with respect to the Wasserstein distance d_{Lip_1}. In the sequel, given a metric d, the notation $d(x, F)$ denotes the distance of x to the set F, that is, $d(x, F) = \inf_{f \in F} d(x, f)$.

Proposition 9 *Assume that Assumption 1 is satisfied. Then, for all $\varepsilon > 0$, there exists $\delta > 0$ such that, for all $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$, one has $d(\theta, \bar{\Theta}) \leqslant \varepsilon$.*

Corollary 10 *Assume that Assumption 1 is satisfied and that $\Theta^\star = \bar{\Theta}$. Then d_{Lip_1} can be fully substituted by $d_{\mathscr{D}}$.*

Proof Let $\varepsilon > 0$. By Theorem 5, we know that the function $\Theta \ni \theta \mapsto d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ is Lipschitz continuous. Thus, there exists $\eta > 0$ such that, for all $(\theta, \theta') \in \Theta^2$ satisfying $\|\theta - \theta'\| \leqslant \eta$, one has $|d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta'})| \leqslant \varepsilon$. Besides, using Proposition 9, there exists $\delta > 0$ such that, for all $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$, one has $d(\theta, \bar{\Theta}) \leqslant \eta$.

Now, let $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$. Since $d(\theta, \bar{\Theta}) \leqslant \eta$ and $\bar{\Theta} = \Theta^\star$, we have $d(\theta, \Theta^\star) \leqslant \eta$. Consequently, $|d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)| \leqslant \varepsilon$, and so, $\theta \in \mathscr{M}_{d_{\mathrm{Lip}_1}}(\mu^\star, \varepsilon)$. ■

Corollary 10 is interesting insofar as when both $d_{\mathscr{D}}$ and d_{Lip_1} have the same minimizers over Θ, then minimizing one close to optimality is the same as minimizing the other. The requirement $\Theta^\star = \bar{\Theta}$ can be relaxed by leveraging what has been studied in the previous subsection about $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D})$.

Lemma 11 *Assume that Assumption 1 is satisfied, and let $\varepsilon > 0$. If $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) \leqslant \varepsilon$, then d_{Lip_1} can be $(\varepsilon + \delta)$-substituted by $d_{\mathscr{D}}$ for all $\delta > 0$.*

Proof Let $\varepsilon > 0$, $\delta > 0$, and $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$, i.e., $d_{\mathscr{D}}(\mu^\star, \mu_\theta) - \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant \delta$. We have

$$\begin{aligned} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) &\leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \\ &\qquad \text{(by inequality (9))} \\ &\leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta) + \delta \\ &\leqslant T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) + \delta \leqslant \varepsilon + \delta. \end{aligned}$$

■

Lemma 11 stresses the importance of $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D})$ in the performance of WGANs. Indeed, the smaller $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D})$, the closer we will be to optimality after training. Moving on, to derive sufficient conditions under which d_{Lip_1} can be substituted by $d_{\mathscr{D}}$ we introduce the following definition:

Definition 12 *We say that d_{Lip_1} is monotonously equivalent to $d_{\mathscr{D}}$ on $\mathscr{P}$ if there exists a continuously differentiable, strictly increasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ and $(a, b) \in (\mathbb{R}_+^\star)^2$ such that*

$$\forall \mu \in \mathscr{P},\ af(d_{\mathscr{D}}(\mu^\star, \mu)) \leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu) \leqslant bf(d_{\mathscr{D}}(\mu^\star, \mu)).$$

Here, it is assumed implicitly that $\mathscr{D} \subseteq \mathrm{Lip}_1$. At the end of the subsection, we stress, empirically, that Definition 12 is easy to check for simple classes of generators. A consequence of this definition is encapsulated in the following lemma.

Lemma 13 *Assume that Assumption 1 is satisfied, and that d_{Lip_1} and $d_{\mathscr{D}}$ are monotonously equivalent on $\mathscr{P}$ with $a = b$ (that is, $d_{\mathrm{Lip}_1} = f \circ d_{\mathscr{D}}$). Then $\Theta^\star = \bar{\Theta}$ and d_{Lip_1} can be fully substituted by $d_{\mathscr{D}}$.*

To complete Lemma 13, we now tackle the case $a < b$.

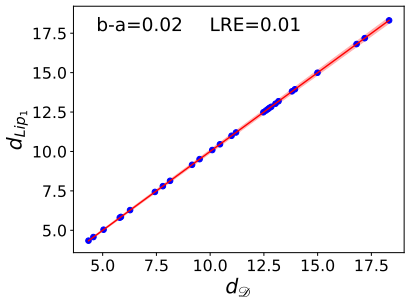
Proposition 14 *Assume that Assumption 1 is satisfied, and that d_{Lip_1} and $d_{\mathscr{D}}$ are monotonously equivalent on $\mathscr{P}$. Then, for any $\delta \in (0, 1)$, d_{Lip_1} can be ε-substituted by $d_{\mathscr{D}}$ with $\varepsilon = (b - a)f(\inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)) + O(\delta)$.*

Proposition 14 states that we can reach ϵ-minimizers of d_{Lip_1} by solving the WGANs problem up to a precision sufficiently small, for all ϵ larger than a bias induced by the model $\mathscr{P}$ and by the discrepancy between d_{Lip_1} and $d_{\mathscr{D}}$.

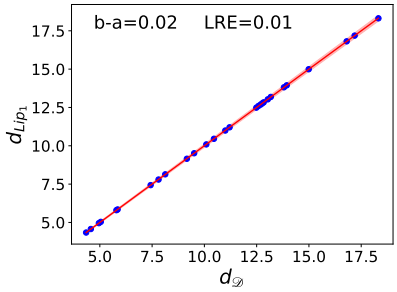
In order to validate Definition 12, we slightly depart from the WGANs setting and run a series of small experiments in the simplified setting where both $\mu^\star$ and $\mu \in \mathscr{P}$ are bivariate mixtures of independent Gaussian distributions with K components ($K = 1$, 2, 3, 25). We consider two classes of discriminators $\{\mathscr{D}_q : q = 2, 6\}$ of the form (4), with growing depth q (the width of the hidden layers is kept constant equal to 20). Our goal is to exemplify the relationship between the distances d_{Lip_1} and $d_{\mathscr{D}_q}$ by looking whether d_{Lip_1} is monotonously equivalent to $d_{\mathscr{D}_q}$.

First, for each K, we randomly draw 40 different pairs of distributions $(\mu^\star, \mu)$ among the set of mixtures of bivariate Gaussian densities with K components. Then, for each of these pairs, we compute an approximation of d_{Lip_1} by averaging the Wasserstein distance between finite samples of size 4096 over 20 runs. This operation is performed using the Python package by Flamary and Courty (2017). For each pair of distributions, we also calculate the corresponding IPMs $d_{\mathscr{D}_q}(\mu^\star, \mu)$. We finally compare d_{Lip_1} and $d_{\mathscr{D}_q}$ by approximating their relationship with a parabolic fit. Results are presented in Figure 1, which depicts in particular the best parabolic fit, and shows the corresponding Least Relative Error (LRE) together with the width $(b - a)$ from Definition 12. In order to enforce the discriminator to verify Assumption 1, we use the orthonormalization of Bjrck and Bowie (1971), as done for example in Anil et al. (2019).

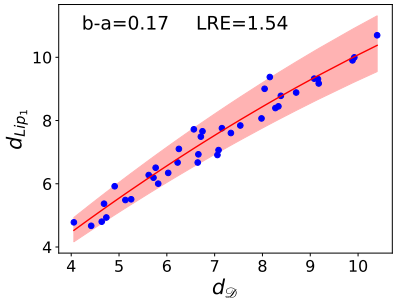
Interestingly, we see that when the class of discriminative functions gets larger (i.e., when q increases), then both metrics start to behave similarly (i.e., the range $(b - a)$ gets thinner), independently of K (Figure 1a to Figure 1f). This tends to confirm that d_{Lip_1} can be considered as monotonously equivalent to $d_{\mathscr{D}_q}$ for q large enough. On the other hand, for a fixed depth q, when allowing for more complex distributions, the width $(b - a)$ increases. This is particularly clear in Figure 1g and Figure 1h, which show the fits obtained when merging all pairs for $K = 1, 4, 9, 25$ (for both $\mu^\star$ and $\mathscr{P}$).



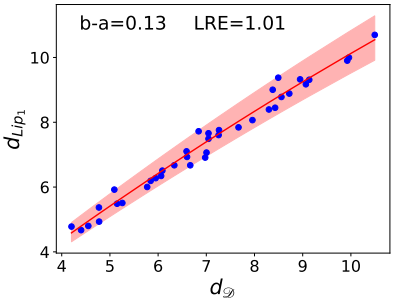
(a) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 2$, $K = 1$.



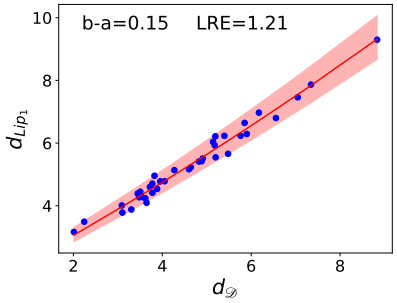
(b) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 5$, $K = 1$.



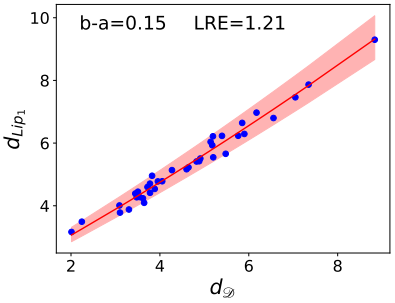
(c) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 2$, $K = 4$.



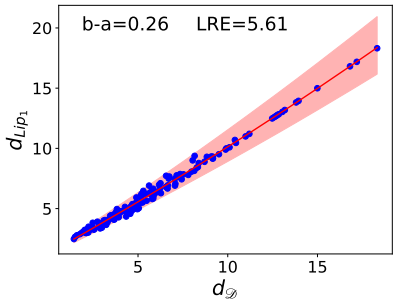
(d) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 5$, $K = 4$.



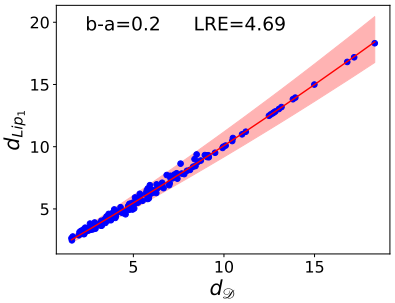
(e) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 2$, $K = 9$.



(f) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 5$, $K = 9$.



(g) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 2$, $K = 1, 4, 9, 25$.



(h) $d_{\mathscr{D}_q}$ vs. d_{Lip_1}, $q = 5$, $K = 1, 4, 9, 25$.

Figure 1: Scatter plots of 40 pairs of distances simultaneously measured with d_{Lip_1} and $d_{\mathscr{D}_q}$, for $q = 2, 5$ and $K = 1, 4, 9, 25$. The red curve is the optimal parabolic fitting and LRE refers to the Least Relative Error. The red zone is the envelope obtained by stretching the optimal curve from b to a.

These figures illustrate the fact that, for a fixed discriminator, the monotonous equivalence between d_{Lip_1} and $d_{\mathscr{D}}$ seems to be a more demanding assumption when the class of generative distributions becomes too large.

3.3 Motivating the use of deep GroupSort neural networks

The objective of this subsection is to provide some justification for the use of deep GroupSort neural networks in the field of WGANs. This short discussion is motivated by the observation of Anil et al. (2019, Theorem 1), who stress that norm-constrained ReLU neural networks are not well-suited for learning non-linear 1-Lipschitz functions.

The next lemma shows that networks of the form (4), which use GroupSort activations, can recover any 1-Lipschitz function belonging to the class AFF of real-valued affine functions on E.

Lemma 15 *Let $f : E \to \mathbb{R}$ be in $\mathrm{AFF} \cap \mathrm{Lip}_1$. Then there exists a neural network of the form (4) verifying Assumption 1, with $q = 2$ and $v_1 = 2$, that can represent f.*

Motivated by Lemma 15, we show that, in some specific cases, the Wasserstein distance d_{Lip_1} can be approached by only considering affine functions, thus motivating the use of neural networks of the form (4). Recall that the support S_μ of a probability measure μ is the smallest subset of μ-measure 1.

Lemma 16 *Let μ and ν be two probability measures in $P_1(E)$. Assume that S_μ and S_ν are one-dimensional disjoint intervals included in the same line. Then $d_{\mathrm{Lip}_1}(\mu, \nu) = d_{\mathrm{AFF}\cap\mathrm{Lip}_1}(\mu, \nu)$.*

Lemma 16 is interesting insofar as it describes a specific case where the discriminator can be restricted to affine functions while keeping the identity $d_{\mathrm{Lip}_1} = d_{\mathscr{D}}$. We consider in the next lemma a slightly more involved setting, where the two distributions μ and ν are multivariate Gaussian with the same covariance matrix.

Lemma 17 *Let $(m_1, m_2) \in (\mathbb{R}^D)^2$, and let $\Sigma \in \mathscr{M}_{(D,D)}$ be a positive semi-definite matrix. Assume that μ is Gaussian $\mathscr{N}(m_1, \Sigma)$ and that ν is Gaussian $\mathscr{N}(m_2, \Sigma)$. Then $d_{\mathrm{Lip}_1}(\mu, \nu) = d_{\mathrm{AFF}\cap\mathrm{Lip}_1}(\mu, \nu)$.*

Yet, assuming multivariate Gaussian distributions might be too restrictive. Therefore, we now assume that both distributions lay on disjoint compact supports sufficiently distant from one another. Recall that for a set $S \subseteq E$, the diameter of S is $\mathrm{diam}(S) = \sup_{(x,y)\in S^2} \|x - y\|$, and that the distance between two sets S and T is defined by $d(S, T) = \inf_{(x,y)\in S\times T} \|x - y\|$.

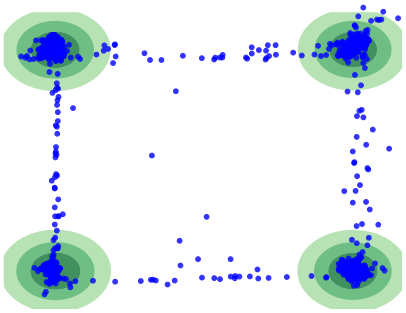
Proposition 18 *Let $\varepsilon > 0$, and let μ and ν be two probability measures in $P_1(E)$ with compact convex supports S_μ and S_ν. Assume that $\max(\mathrm{diam}(S_\mu), \mathrm{diam}(S_\nu)) \leqslant \varepsilon d(S_\mu, S_\nu)$. Then*

$$d_{\mathrm{AFF}\cap\mathrm{Lip}_1}(\mu, \nu) \leqslant d_{\mathrm{Lip}_1}(\mu, \nu) \leqslant (1 + 2\varepsilon) d_{\mathrm{AFF}\cap\mathrm{Lip}_1}(\mu, \nu).$$

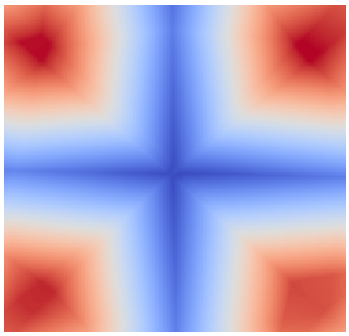
Observe that in the case where neither μ nor ν are Dirac measures, then the assumption of the lemma imposes that $S_\mu \cap S_\nu = \emptyset$. In the context of WGANs, it is highly likely that the generator badly approximates the true distribution $\mu^\star$ at the beginning of training. The

setting of Proposition 18 is thus interesting insofar as $\mu^\star$ and the generative distribution will most certainly verify the assumption on the diameters at this point in the learning process. However, in the common case where the true distribution lays on disconnected manifolds, the assumptions of the proposition are not valid anymore, and it would therefore be interesting to show a similar result using the broader set of piecewise linear functions on E.

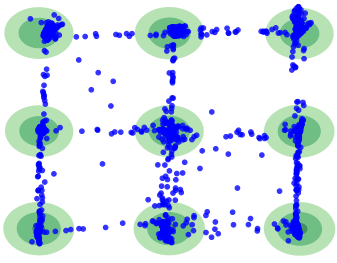
As an empirical illustration, consider the synthetic setting where one tries to approximate a bivariate mixture of independent Gaussian distributions with respectively 4 (Figure 2a) and 9 (Figure 2c) modes. As expected, the optimal discriminator takes the form of a piecewise linear function, as illustrated by Figure 2b and Figure 2d, which display heatmaps of the discriminator's output. Interestingly, we see that the number of linear regions increases with the number K of components of $\mu^\star$.



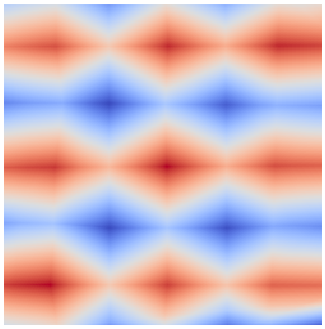
(a) True distribution $\mu^\star$ (mixture of $K = 4$ bivariate Gaussian densities, green circles) and 2000 data points sampled from the generator $\mu_{\bar{\theta}}$ (blue dots).



(b) Heatmap of the discriminator's output on a mixture of $K = 4$ bivariate Gaussian densities.



(c) True distribution $\mu^\star$ (mixture of $K = 9$ bivariate Gaussian densities, green circles) and 2000 data points sampled from the generator $\mu_{\bar{\theta}}$ (blue dots).



(d) Heatmap of the discriminator's output on a mixture of $K = 9$ bivariate Gaussian densities.

Figure 2: Illustration of the usefulness of GroupSort neural networks when dealing with the learning of mixtures of Gaussian distributions. In both cases, we have $p = q = 3$.

These empirical results stress that when $\mu^\star$ gets more complex, if the discriminator ought to correctly approximate the Wasserstein distance, then it should parameterize piecewise linear functions with growing numbers of regions. While we enlighten properties of Groupsort networks, many recent theoretical works have been studying the number of regions of deep ReLU neural networks (Pascanu et al., 2013; Montúfar et al., 2014; Arora et al., 2018; Serra et al., 2018). In particular, Montúfar et al. (2014, Theorem 5) states that the number of linear regions of deep models grows exponentially with the depth and polynomially with the width. This, along with our observations, is an interesting avenue to choose the architecture of the discriminator.

4. Asymptotic properties

In practice, one never has access to the distribution $\mu^\star$ but rather to a finite collection of i.i.d. observations $X_1, \dots, X_n$ distributed according to $\mu^\star$. Thus, for the remainder of the article, we let μ_n be the empirical measure based on $X_1, \dots, X_n$, that is, for any Borel subset A of E, $\mu_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A}$. With this notation, the empirical counterpart of the WGANs problem is naturally defined as minimizing over Θ the quantity $d_{\mathscr{D}}(\mu_n, \mu_\theta)$. Equivalently, we seek to solve the following optimization problem:

$$\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta) = \inf_{\theta\in\Theta} \sup_{\alpha\in\Lambda} \Big[\frac{1}{n}\sum_{i=1}^n D_\alpha(X_i) - \mathbb{E}D_\alpha(G_\theta(Z))\Big]. \tag{12}$$

Assuming that Assumption 1 is satisfied, we have, as in Corollary 6, that the infimum in (12) is reached. We therefore consider the set of empirical optimal parameters

$$\hat{\Theta}_n = \arg\min_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta),$$

and let $\hat{\theta}_n$ be a specific element of $\hat{\Theta}_n$ (note that the choice of $\hat{\theta}_n$ has no impact on the value of the minimum). We note that $\hat{\Theta}_n$ (respectively, $\hat{\theta}_n$) is the empirical counterpart of $\bar{\Theta}$ (respectively, $\bar{\theta}$). Section 3 was mainly devoted to the analysis of the difference $\varepsilon_{\text{optim}}$. In this section, we are willing to take into account the effect of having finite samples. Thus, in line with the above, we are now interested in the generalization properties of WGANs and look for upper-bounds on the quantity

$$0 \leqslant d_{\text{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} d_{\text{Lip}_1}(\mu^\star, \mu_\theta). \tag{13}$$

Arora et al. (2017, Theorem 3.1) states an asymptotic result showing that when provided enough samples, the neural IPM $d_{\mathscr{D}}$ generalizes well, in the sense that for any pair $(\mu, \nu) \in P_1(E)^2$, the difference $|d_{\mathscr{D}}(\mu, \nu) - d_{\mathscr{D}}(\mu_n, \nu_n)|$ can be arbitrarily small with high probability. However, this result does not give any information on the quantity of interest $d_{\text{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} d_{\text{Lip}_1}(\mu^\star, \mu_\theta)$. Closer to our current work, Zhang et al. (2018) provide bounds for $d_{\mathscr{D}}(\mu^\star, \mu_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)$, starting from the observation that

$$0 \leqslant d_{\mathscr{D}}(\mu^\star, \mu_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant 2d_{\mathscr{D}}(\mu^\star, \mu_n). \tag{14}$$

In the present article, we develop a complementary point of view and measure the generalization properties of WGANs on the basis of the Wasserstein distance d_{Lip_1}, as in equation

(13). Our approach is motivated by the fact that the neural IPM $d_{\mathscr{D}}$ is only used for easing the optimization process and, accordingly, that the performance should be assessed on the basis of the distance d_{Lip_1}, not $d_{\mathscr{D}}$.

Note that $\hat{\theta}_n$, which minimizes $d_{\mathscr{D}}(\mu_n, \mu_\theta)$ over Θ, may not be unique. Besides, there is no guarantee that two distinct elements $\theta_{n,1}$ and $\theta_{n,2}$ of $\hat{\Theta}_n$ lead to the same distance $d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_{n,1}})$ and $d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_{n,2}})$ (again, $\hat{\theta}_n$ is computed with $d_{\mathscr{D}}$, not with d_{Lip_1}). Therefore, in order to upper-bound the error in (13), we let, for each $\theta_n \in \hat{\Theta}_n$,

$$\bar{\theta}_n \in \operatorname*{arg\,min}_{\bar{\theta} \in \bar{\Theta}} \|\theta_n - \bar{\theta}\|.$$

The rationale behind the definition of $\bar{\theta}_n$ is that we expect it to behave "similarly" to θ_n. Following our objective, the error can be decomposed as follows:

$$\begin{aligned}
0 &\leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n}) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\
&\leqslant \sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\
&= \sup_{\theta_n \in \hat{\Theta}_n} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_n}) + d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_n}) \right] - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\
&\leqslant \sup_{\theta_n \in \hat{\Theta}_n} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_n}) \right] + \sup_{\bar{\theta} \in \bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\
&= \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}}, \qquad (15)
\end{aligned}$$

where we set $\varepsilon_{\mathrm{estim}} = \sup_{\theta_n \in \hat{\Theta}_n} [d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_n})]$. Notice that this supremum can be positive or negative. However, it can be shown to converge to 0 almost surely when $n \to \infty$.

Lemma 19 *Assume that Assumption 1 is satisfied. Then* $\lim_{n\to\infty} \varepsilon_{\mathrm{estim}} = 0$ *almost surely.*

Going further with the analysis of (13), the sum $\varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}}$ is bounded as follows:

$$\begin{aligned}
\varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} &\leqslant \sup_{\theta_n \in \hat{\Theta}_n} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_n}) \right] + T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) \\
&\qquad \text{(by inequality (10))} \\
&\leqslant \sup_{\theta_n \in \hat{\Theta}_n} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \right] + T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} \\
&\leqslant \sup_{\theta_n \in \hat{\Theta}_n} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - d_{\mathscr{D}}(\mu^\star, \mu_{\theta_n}) + d_{\mathscr{D}}(\mu^\star, \mu_{\theta_n}) - \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \right] + T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) \\
&\leqslant \sup_{\theta_n \in \hat{\Theta}_n} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - d_{\mathscr{D}}(\mu^\star, \mu_{\theta_n}) \right] + 2 d_{\mathscr{D}}(\mu^\star, \mu_n) + T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) \\
&\quad \text{(upon noting that inequality (14) is also valid for any } \theta_n \in \hat{\Theta}_n) \\
&\leqslant 2 T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) + 2 d_{\mathscr{D}}(\mu^\star, \mu_n). \qquad (16)
\end{aligned}$$

The above bound is a function of both the generator and the discriminator. The term $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D})$ is increasing when the capacity of the generator is increasing. The discriminator, however, plays a more ambivalent role, as already pointed out by Zhang et al. (2018). On the one hand, if the discriminator's capacity decreases, the gap between $d_{\mathscr{D}}$ and d_{Lip_1} gets bigger and $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D})$ increases. On the other hand, discriminators with bigger capacities ought to increase the contribution $d_{\mathscr{D}}(\mu^\star, \mu_n)$. In order to bound $d_{\mathscr{D}}(\mu^\star, \mu_n)$, Proposition 20 below extends Zhang et al. (2018, Theorem 3.1), in the sense that it does not require the set of discriminative functions nor the space E to be bounded. Recall that, for $\gamma > 0$, $\mu^\star$ is said to be γ sub-Gaussian (Jin et al., 2019) if

$$\forall v \in \mathbb{R}^d, \ \mathbb{E}e^{v\cdot(T-\mathbb{E}T)} \leqslant e^{\frac{\gamma^2\|v\|^2}{2}},$$

where T is a random vector with probability distribution $\mu^\star$ and $\cdot$ denotes the dot product in $\mathbb{R}^D$.

Proposition 20 *Assume that Assumption 1 is satisfied, let $\eta \in (0,1)$, and let $\mathscr{D}$ be a discriminator of the form (4).*

(i) If $\mu^\star$ has compact support with diameter B, then there exists a constant $c_1 > 0$ such that, with probability at least $1-\eta$,

$$d_{\mathscr{D}}(\mu^\star, \mu_n) \leqslant \frac{c_1}{\sqrt{n}} + B\sqrt{\frac{\log(1/\eta)}{2n}}.$$

(ii) More generally, if $\mu^\star$ is γ sub-Gaussian, then there exists a constant $c_2 > 0$ such that, with probability at least $1-\eta$,

$$d_{\mathscr{D}}(\mu^\star, \mu_n) \leqslant \frac{c_2}{\sqrt{n}} + 8\gamma\sqrt{eD}\sqrt{\frac{\log(1/\eta)}{n}}.$$

The result of Proposition 20 has to be compared with convergence rates of the Wasserstein distance. According to Fournier and Guillin (2015, Theorem 1), when the dimension D of E is such that $D > 2$, if $\mu^\star$ has a second-order moment, then there exists a constant c such that

$$0 \leqslant \mathbb{E}d_{\mathrm{Lip}_1}(\mu^\star, \mu_n) \leqslant \frac{c}{n^{1/D}}.$$

Thus, when the space E is of high dimension (e.g., in image generation tasks), under the conditions of Proposition 20, the pseudometric $d_{\mathscr{D}}$ provides much faster rates of convergence for the empirical measure. However, one has to keep in mind that both constants c_1 and c_2 grow in $O(qQ^{3/2}(D^{1/2}+q))$.

Our Theorem 7 states the existence of a discriminator such that $\varepsilon_{\mathrm{optim}}$ can be arbitrarily small. It is therefore reasonable, in view of inequality (16), to expect that the sum $\varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}}$ can also be arbitrarily small, at least in an asymptotic sense. This is encapsulated in Theorem 21 below.

Theorem 21 *Assume that Assumption 1 is satisfied, and let $\eta \in (0,1)$.*

(*i*) *If $\mu^\star$ has compact support with diameter B, then, for all $\varepsilon > 0$, there exists a discriminator $\mathscr{D}$ of the form (4) and a constant $c_1 > 0$ (independent of ε) such that, with probability at least $1-\eta$,*

$$0 \leqslant \varepsilon_{\text{estim}} + \varepsilon_{\text{optim}} \leqslant 2\varepsilon + \frac{2c_1}{\sqrt{n}} + 2B\sqrt{\frac{\log(1/\eta)}{2n}}.$$

(*ii*) *More generally, if $\mu^\star$ is γ sub-Gaussian, then, for all $\varepsilon > 0$, there exists a discriminator $\mathscr{D}$ of the form (4) and a constant $c_2 > 0$ (independent of ε) such that, with probability at least $1-\eta$,*

$$0 \leqslant \varepsilon_{\text{estim}} + \varepsilon_{\text{optim}} \leqslant 2\varepsilon + \frac{2c_2}{\sqrt{n}} + 16\gamma\sqrt{eD}\sqrt{\frac{\log(1/\eta)}{n}}.$$

Theorem 21 states that, asymptotically, the optimal parameters in $\hat{\Theta}_n$ behave properly. A caveat is that the definition of $\varepsilon_{\text{estim}}$ uses $\hat{\Theta}_n$. However, in practice, one never has access to $\hat{\theta}_n$, but rather to an approximation of this quantity obtained by gradient descent algorithms. Thus, in line with Definition 8, we introduce the concept of empirical substitution:

Definition 22 *Let $\varepsilon > 0$ and $\eta \in (0,1)$. We say that d_{Lip_1} can be empirically ε-substituted by $d_{\mathscr{D}}$ if there exists $\delta > 0$ such that, for all n large enough, with probability at least $1-\eta$,*

$$\mathscr{M}_{d_{\mathscr{D}}}(\mu_n, \delta) \subseteq \mathscr{M}_{d_{\text{Lip}_1}}(\mu^\star, \varepsilon). \tag{17}$$

The rationale behind this definition is that if (17) is satisfied, then by minimizing the IPM $d_{\mathscr{D}}$ close to optimality in (12), one can be guaranteed to be also close to optimality in (5) with high probability. We stress that Definition 22 is the empirical counterpart of Definition 8.

Proposition 23 *Assume that Assumption 1 is satisfied and that $\mu^\star$ is sub-Gaussian. Let $\varepsilon > 0$. If $T_{\mathscr{D}}(\text{Lip}_1, \mathscr{D}) \leqslant \varepsilon$, then d_{Lip_1} can be empirically $(\varepsilon+\delta)$-substituted by $d_{\mathscr{D}}$ for all $\delta > 0$.*

This proposition is the empirical counterpart of Lemma 11. It underlines the fact that by minimizing the pseudometric $d_{\mathscr{D}}$ between the empirical measure μ_n and the set of generative distributions $\mathscr{P}$ close to optimality, one can control the loss in performance under the metric d_{Lip_1}.

Let us finally mention that it is also possible to provide asymptotic results on the sequences of parameters $(\hat{\theta}_n)$, keeping in mind that $\hat{\Theta}_n$ and $\bar{\Theta}$ are not necessarily reduced to singletons.

Lemma 24 *Assume that Assumption 1 is satisfied. Let $(\hat{\theta}_n)$ be a sequence of optimal parameters that converges almost surely to $z \in \Theta$. Then $z \in \bar{\Theta}$ almost surely.*

Proof Let the sequence $(\hat{\theta}_n)$ converge almost surely to some $z \in \Theta$. By Theorem 5, the function $\Theta \ni \theta \mapsto d_{\mathscr{D}}(\mu^\star, \mu_\theta)$ is continuous, and therefore, almost surely, $\lim_{n\to\infty} d_{\mathscr{D}}(\mu^\star, \mu_{\hat{\theta}_n}) = d_{\mathscr{D}}(\mu^\star, \mu_z)$. Using inequality (14), we see that, almost surely,

$$\begin{aligned} 0 \leqslant d_{\mathscr{D}}(\mu^\star, \mu_z) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) &= \lim_{n\to\infty} d_{\mathscr{D}}(\mu^\star, \mu_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \\ &\leqslant \liminf_{n\to\infty} 2 d_{\mathscr{D}}(\mu^\star, \mu_n). \end{aligned}$$

Using Dudley (2004, Theorem 11.4.1) and the strong law of large numbers, we have that the sequence of empirical measures (μ_n) almost surely converges weakly to $\mu^\star$ in $P_1(E)$. Besides, since $d_{\mathscr{D}}$ metrizes weak convergence in $P_1(E)$ (by Proposition 3), we conclude that $z \in \bar{\Theta}$ almost surely. ■

5. Understanding the performance of WGANs

In order to better understand the overall performance of the WGANs architecture, it is instructive to decompose the final loss $d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n})$ as in (15):

$$\begin{aligned} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n}) &\leqslant \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} + \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\ &= \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} + \varepsilon_{\mathrm{approx}}, \end{aligned} \tag{18}$$

where

- (i) $\varepsilon_{\mathrm{estim}}$ matches up with the use of a data-dependent optimal parameter $\hat{\theta}_n$, based on the training set $X_1, \ldots, X_n$ drawn from $\mu^\star$;
- (ii) $\varepsilon_{\mathrm{optim}}$ corresponds to the loss in performance when using $d_{\mathscr{D}}$ as training loss instead of d_{Lip_1} (this term has been thoroughly studied in Section 3);
- (iii) and $\varepsilon_{\mathrm{approx}}$ stresses the capacity of the parametric family of generative distributions $\mathscr{P}$ to approach the unknown distribution $\mu^\star$.

Close to our work are the articles by Liang (2018), Singh et al. (2018), and Uppal et al. (2019), who study statistical properties of GANs. Liang (2018) and Singh et al. (2018) exhibit rates of convergence under an IPM-based loss for estimating densities that live in Sobolev spaces, while Uppal et al. (2019) explore the case of Besov spaces. Remarkably, Liang (2018) discusses bounds for the Kullback-Leibler divergence, the Hellinger divergence, and the Wasserstein distance between $\mu^\star$ and $\mu_{\hat{\theta}_n}$. These bounds are based on a different decomposition of the loss and offer a complementary point of view. We emphasize that, in the present article, no density assumption is made neither on the class of generative distributions $\mathscr{P}$ nor on the target distribution $\mu^\star$.

5.1 Synthetic experiments

Our goal in this subsection is to illustrate (18) by running a set of experiments on synthetic datasets. The true probability measure $\mu^\star$ is assumed to be a mixture of bivariate Gaussian distributions with either 1, 4, or 9 components. This simple setting allows us to control the complexity of $\mu^\star$, and, in turn, to better assess the impact of both the generator's and discriminator's capacities. We use growing classes of generators of the form (3), namely $\{\mathscr{G}_p : p = 2, 3, 5, 7\}$, and growing classes of discriminators of the form (4), namely $\{\mathscr{D}_q : q = 2, 3, 5, 7\}$. For both the generator and the discriminator, the width of the hidden layers is kept constant equal to 20.

Two metrics are computed to evaluate the behavior of the different generative models. First, we use the Wasserstein distance between the true distribution (either $\mu^\star$ or its

empirical version μ_n) and the generative distribution (either $\mu_{\bar{\theta}}$ or $\mu_{\hat{\theta}_n}$). This distance is calculated by using the Python package by Flamary and Courty (2017), via finite samples of size 4096 (average over 20 runs). Second, we use the recall metric (the higher, the better), proposed by Kynkäänniemi et al. (2019). Roughly, this metric measures "how much" of the true distribution (either $\mu^\star$ or μ_n) can be reconstructed by the generative distribution (either $\mu_{\bar{\theta}}$ or $\mu_{\hat{\theta}_n}$). At the implementation level, this score is based on k-nearest neighbor nonparametric density estimation. It is computed via finite samples of size 4096 (average over 20 runs).

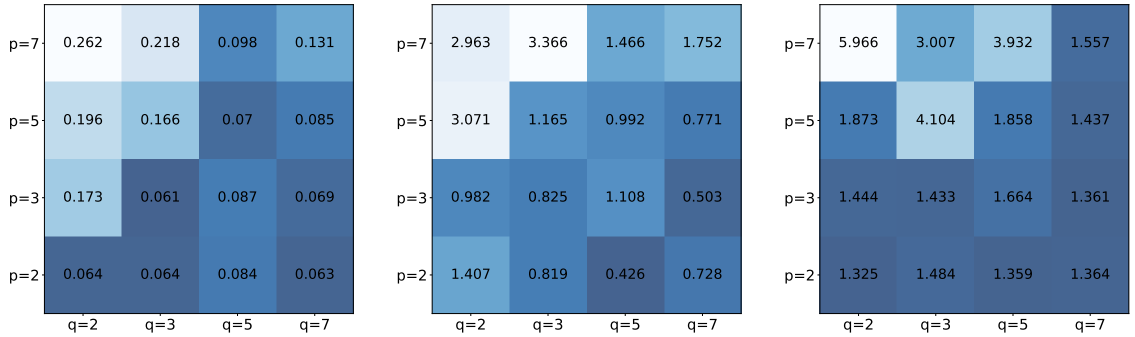
Our experiments were run in two different settings:

Asymptotic setting: in this first experiment, we assume that $\mu^\star$ is known from the experimenter (so, there is no dataset). At the end of the optimization scheme, we end up with one $\bar{\theta} \in \bar{\Theta}$. Thus, in this context, the performance of WGANs is captured by

$$\sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) = \varepsilon_{\mathrm{optim}} + \varepsilon_{\mathrm{approx}}.$$

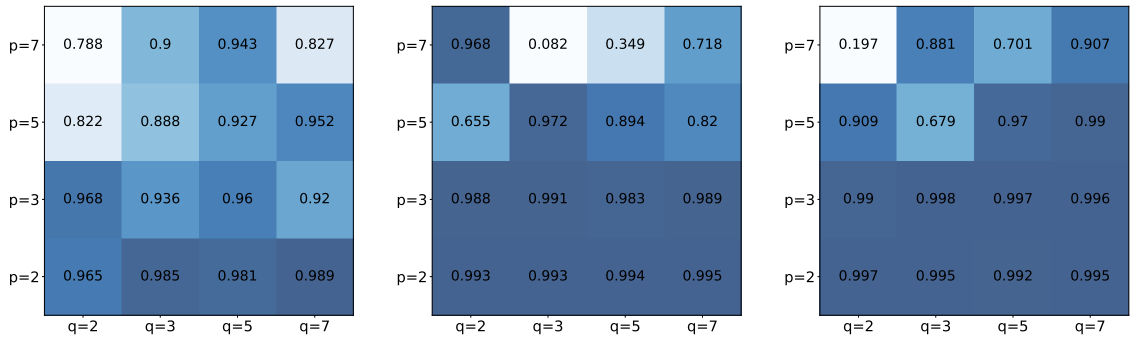
For a fixed discriminator, when increasing the generator's depth p, we expect $\varepsilon_{\mathrm{approx}}$ to decrease. Conversely, as discussed in Subsection 3.1, we anticipate an augmentation of $\varepsilon_{\mathrm{optim}}$, since the discriminator must now differentiate between larger classes of generative distributions. In this case, it is thus difficult to predict how $\sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}})$ behaves when p increases. On the contrary, in accordance with the results of Section 3, for a fixed p we expect the performance to increase with a growing q since, with larger discriminators, the pseudometric $d_{\mathscr{D}}$ is more likely to behave similarly to the Wasserstein distance d_{Lip_1}.

These intuitions are validated by Figure 3 and Figure 4 (the bluer, the better). The first one shows an approximation of $\sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}})$ computed over 5 different seeds as a function of p and q. The second one depicts the average recall of the estimator $\mu_{\bar{\theta}}$ with respect to $\mu^\star$, as a function of p and q, again computed over 5 different seeds. In both figures, we observe that for a fixed p, incrementing q leads to better results. On the opposite, for a fixed discriminator's depth q, increasing the depth p of the generator seems to deteriorate both scores (Wasserstein distance and recall). This consequently suggests that the term $\varepsilon_{\mathrm{optim}}$ dominates $\varepsilon_{\mathrm{approx}}$.



(a) $\sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}})$, $K = 1$. (b) $\sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}})$, $K = 9$. (c) $\sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}})$, $K = 25$.

Figure 3: Influence of the generator's depth p and the discriminator's depth q on the maximal Wasserstein distance $\sup_{\bar{\theta}\in\bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}})$.

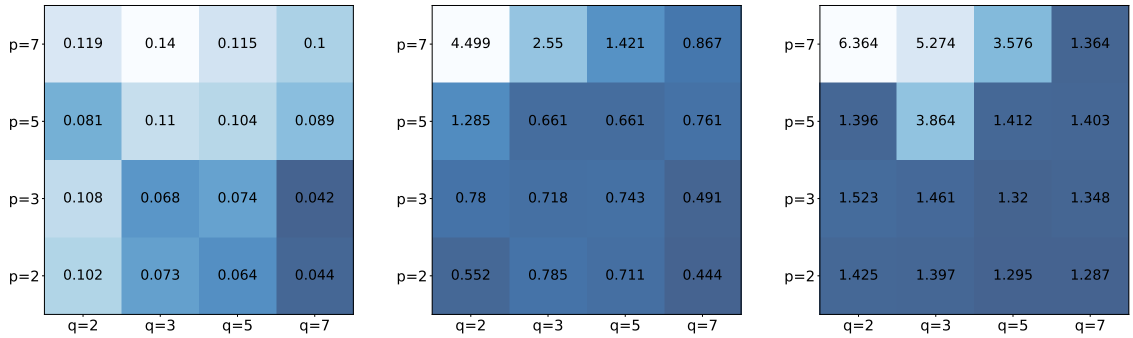


(a) Av. recall of $\mu_{\bar{\theta}}$ w.r.t. $\mu^\star$, $K = 1$. (b) Av. recall of $\mu_{\bar{\theta}}$ w.r.t. $\mu^\star$, $K = 9$. (c) Av. recall of $\mu_{\bar{\theta}}$ w.r.t. $\mu^\star$, $K = 25$.

Figure 4: Influence of the generator's depth p and the discriminator's depth q on the average recall of the estimators $\mu_{\bar{\theta}}$ w.r.t. $\mu^\star$.

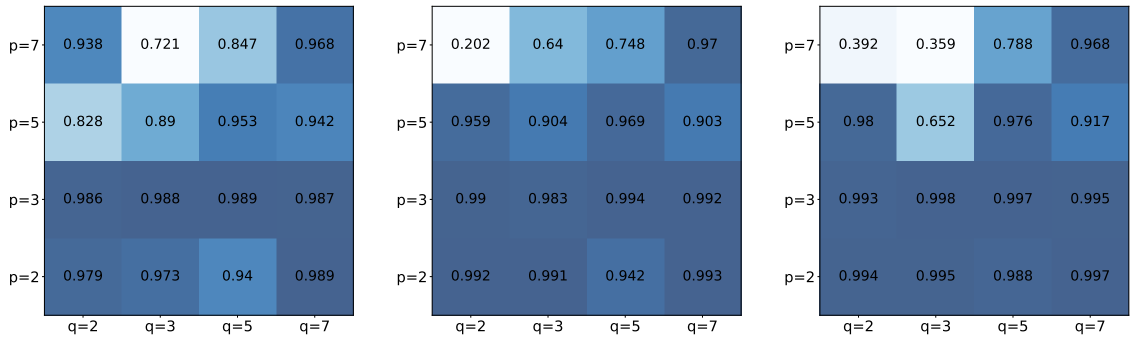
Finite-sample setting: in this second experiment, we consider the more realistic situation where we have at hand finite samples $X_1, \ldots, X_n$ drawn from $\mu^\star$ ($n = 5000$).

Recalling that $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) \leqslant \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} + \varepsilon_{\mathrm{approx}}$, we plot in Figure 5 the maximal Wasserstein distance $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n})$, and in Figure 6 the average recall of the estimators μ_{θ_n} with respect to $\mu^\star$, as a function of p and q. Anticipating the behavior of $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n})$ when increasing the depth q is now more involved. Indeed, according to inequality (16), which bounds $\varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}}$, a larger $\mathscr{D}$ will make $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D})$ smaller but will, on the opposite, increase $d_{\mathscr{D}}(\mu^\star, \mu_n)$. Figure 5 clearly shows that, for a fixed p, the maximal Wasserstein distance seems to be improved when q increases. This suggests that the term $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D})$ dominates $d_{\mathscr{D}}(\mu^\star, \mu_n)$. Similarly to the asymptotic setting, we also make the observation that bigger p require a higher depth q since larger class of generative distributions are more complex to discriminate.



(a) $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n})$, $K = 1$. (b) $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n})$, $K = 4$. (c) $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n})$, $K = 9$.

Figure 5: Influence of the generator's depth p and the discriminator's depth q on the maximal Wasserstein distance $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n})$, with $n = 5000$.



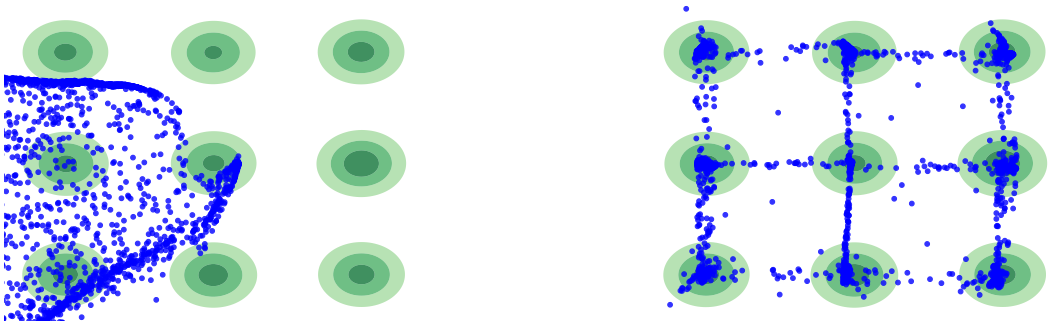
(a) Av. recall of $\mu_{\hat{\theta}_n}$ w.r.t. $\mu^\star$, $K = 1$.

(b) Av. recall of $\mu_{\hat{\theta}_n}$ w.r.t. $\mu^\star$, $K = 9$.

(c) Av. recall of $\mu_{\hat{\theta}_n}$ w.r.t. $\mu^\star$, $K = 25$.

Figure 6: Influence of the generator's depth p and the discriminator's depth q on the average recall of the estimators μ_{θ_n} w.r.t. $\mu^\star$, with $n = 5000$.

We end this subsection by pointing out a recurring observation across different experiments. In Figure 4 and Figure 6, we notice, as already stressed, that the average recall of the estimators is prone to decrease when the generator's depth p increases. On the opposite, the average recall increases when the discriminator's depth q increases. This is interesting because the recall metric is a good proxy for a stabilized training, insofar as a high recall means the absence of mode collapse. This is also confirmed in Figure 7, which compares two densities: in Figure 7a, the discriminator has a small capacity ($q = 3$) and the generator a large capacity ($p = 7$), whereas in Figure 7b, the discriminator has a large capacity ($q = 7$) and the generator a small capacity ($p = 3$). We observe that the first WGAN architecture behaves poorly compared to the second one. We therefore conclude that larger discriminators seem to bring some stability in the training of WGANS both in the asymptotic and finite sample regimes.



(a) $p = 7$ and $q = 3$.

(b) $p = 3$ and $q = 7$.

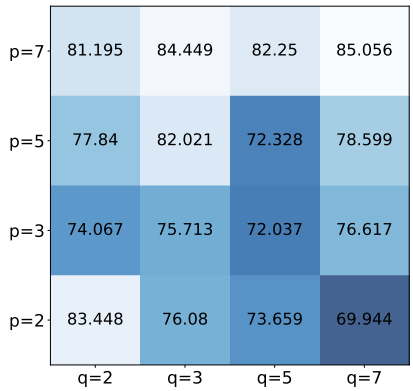
Figure 7: True distribution $\mu^\star$ (mixture of $K = 9$ bivariate Gaussian densities, green circles) and 2000 data points sampled from the generator $\mu_{\bar{\theta}}$ (blue dots).

5.2 Real-world experiments

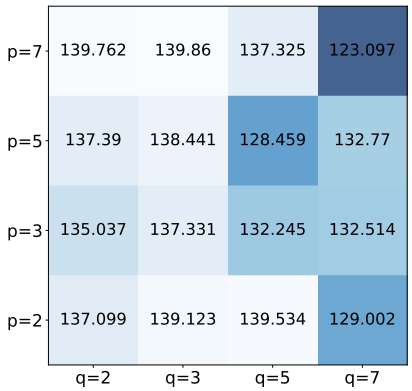
In this subsection, we further illustrate the impact of the generator's and the discriminator's capacities on two high-dimensional datasets, namely MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017). MNIST contains images in $\mathbb{R}^{28\times 28}$ with 10 classes representing the digits. Fashion-MNIST is a 10-class dataset of images in $\mathbb{R}^{28\times 28}$, with slightly more complex shapes than MNIST. Both datasets have a training set of 60,000 examples.

To measure the performance of WGANs when dealing with high-dimensional applications such as image generation, Brock et al. (2019) have advocated that embedding images into a feature space with a pre-trained convolutional classifier provides more meaningful information. Therefore, in order to assess the quality of the generator $\mu_{\hat{\theta}_n}$, we sample images both from the empirical measure μ_n and from the distribution $\mu_{\hat{\theta}_n}$. Then, instead of computing the Wasserstein (or recall) distance directly between these two samples, we use as a substitute their embeddings output by an external classifier and compute the Wasserstein (or recall) between the two new collections. Such a transformation is also done, for example, in Kynkäänniemi et al. (2019). Practically speaking, for any pair of images (a, b), this operation amounts to using the Euclidean distance $\|\phi(a)-\phi(b)\|$ in the Wasserstein and recall criteria, where ϕ is a pre-softmax layer of a supervised classifier, trained specifically on the datasets MNIST and Fashion-MNIST.

For these two datasets, as usual, we use generators of the form (3) and discriminators of the form (4), and plot the performance of $\mu_{\hat{\theta}_n}$ as a function of both p and q. The results of Figure 8 confirm the fact that the worst results are achieved for generators with a large depth p combined with discriminators with a small depth q. They also corroborate the previous observations that larger discriminators are preferred.



(a) $\sup_{\theta_n\in\hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu_n, \mu_{\theta_n})$, MNIST dataset.



(b) $\sup_{\theta_n\in\hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu_n, \mu_{\theta_n})$, FMNIST dataset.

Figure 8: Influence of the generator's depth p and the discriminator's depth q on the maximal Wasserstein distance $\sup_{\theta_n\in\hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu_n, \mu_{\theta_n})$ for the MNIST and F-MNIST datasets.

Acknowledgments

We thank Flavian Vasile (Criteo AI Lab) and Clément Calauzenes (Criteo AI Lab) for stimulating discussions and insightful suggestions.

References

D. Acharya, Z. Huang, D.P. Paudel, and L. Van Gool. Towards high resolution video generation with progressive growing of sliced Wasserstein GANs. *arXiv.1810.02419*, 2018.

C. Anil, J. Lucas, and R. Grosse. Sorting out Lipschitz function approximation. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 291–301. PMLR, 2019.

M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y.W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR, 2017.

R. Arora, A. Basu, P. Mianjy, and A. Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In D. Precup and Y.W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 224–232. PMLR, 2017.

G. Biau, B. Cadre, M. Sangnier, and U. Tanielian. Some theoretical properties of GANs. *The Annals of Statistics*, in press, 2020.

. Bjrck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8:358–364, 1971.

A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

A. Chernodub and D. Nowicki. Norm-preserving Orthogonal Permutation Linear Unit activation functions (OPLU). *arXiv.1604.02313*, 2016.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

R.M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2 edition, 2004.

W. Fedus, I. Goodfellow, and A.M. Dai. MaskGAN: Better text generation via filling in the ____. In *International Conference on Learning Representations*, 2018.

R. Flamary and N. Courty. POT: Python Optimal Transport library, 2017. URL https://github.com/rflamary/POT.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

C.R. Givens and R.M. Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31:231–240, 1984.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, and M. Dudk, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 315–323. PMLR, 2011.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and J. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A.C. Courville. Improved training of Wasserstein GANs. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc., 2017.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

C. Jin, P. Netrapalli, R. Ge, S.M. Kakade, and M. Jordan. A short note on concentration inequalities for random vectors with subGaussian norm. *arXiv.1902.03736*, 2019.

L.V. Kantorovich and G.S. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad University Mathematics*, 13:52–59, 1958.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

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

N. Kodali, J. Abernethy, J. Hays, and Z. Kira. On convergence and stability of GANs. *arXiv.1705.07215*, 2017.

A. Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In E.P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 28–36. PMLR, 2014.

T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3927–3936. Curran Associates, Inc., 2019.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. MMD GAN: Towards deeper understanding of moment matching network. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2203–2213. Curran Associates, Inc., 2017.

T. Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv.1811.03179*, 2018.

S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and convergence properties of generative adversarial learning. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5551–5559. Curran Associates, Inc., 2017.

C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series 141, pages 148–188. Cambridge University Press, Cambridge, 1989.

L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv.1611.02163*, 2016.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

O. Mogren. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. *arXiv.1611.09904*, 2016.

G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2924–2932. Curran Associates, Inc., 2014.

A. Mller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In D.D. Lee, M. Sugiyama, U. von Luxburg,

I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 271–279. Curran Associates, Inc., 2016.

M. O'Searcoid. *Metric Spaces*. Springer, Dublin, 2006.

R. Pascanu, G. Montúfar, and Y. Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. In *International Conference on Learning Representations*, 2013.

H. Petzka, A. Fischer, and D. Lukovnikov. On the regularization of Wasserstein GANs. In *International Conference on Learning Representations*, 2018.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. Stabilizing training of generative adversarial networks through regularization. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2018–2028. Curran Associates, Inc., 2017.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016.

T. Serra, C. Tjandraatmadja, and S. Ramalingam. Bounding and counting linear regions of deep neural networks. In *International Conference on Machine Learning*, pages 4565–4573, 2018.

S. Singh, A. Uppal, B. Li, C.-L. Li, M. Zaheer, and B. Poczos. Nonparametric density estimation under adversarial losses. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10225–10236. Curran Associates, Inc., 2018.

A. Uppal, S. Singh, and B. Poczos. Nonparametric density estimation and convergence rates for GANs under Besov IPM losses. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9089–9100. Curran Associates, Inc., 2019.

R. van Handel. *Probability in High Dimension*. APC 550 Lecture Notes, Princeton University, 2016.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018.

C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2008.

X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang. Improving the improved training of Wasserstein GANs: A consistency term and its dual effect. In *International Conference on Learning Representations*, 2018.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 28522858. AAAI Press, 2017.

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the discriminative-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Z. Zhou, J. Liang, Y. Song, L. Yu, H. Wang, W. Zhang, Y. Yu, and Z. Zhang. Lipschitz generative adversarial nets. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7584–7593. PMLR, 2019.

Appendix A.

A.1 Proof of Lemma 1

We know that for each $\theta \in \Theta$, G_θ is a feed-forward neural network of the form (3), which maps inputs $z \in \mathbb{R}^d$ into $E \subset \mathbb{R}^D$. In particular, for $z \in \mathbb{R}^d$, $G_\theta(z) = f_p \circ \cdots \circ f_1(z)$, where $f_i(x) = \sigma(U_i x + b_i)$ for $i = 1, \ldots, p-1$ (σ is applied componentwise), and $f_p(x) = U_p x + b_p$.

Recall that the notation $\|\cdot\|$ (respectively, $\|\cdot\|_\infty$) means the Euclidean (respectively, the supremum) norm, with no specific mention of the underlying space on which it acts. For $(z, z') \in (\mathbb{R}^d)^2$, we have

$$\begin{aligned}
\|f_1(z) - f_1(z')\| &\leqslant \|U_1 z + b_1 - U_1 z' - b_1\| \\
&\quad \text{(since } \sigma \text{ is 1-Lipschitz)} \\
&= \|U_1(z - z')\| \\
&\leqslant \|U_1\|_2 \, \|z - z'\| \\
&\leqslant K_1 \|z - z'\| \\
&\quad \text{(by Assumption 1).}
\end{aligned}$$

Repeating this for $i = 2, \ldots, p$, we thus have, for all $(z, z') \in (\mathbb{R}^d)^2$, $\|G_\theta(z) - G_\theta(z')\| \leqslant K_1^p \|z - z'\|$. We conclude that, for each $\theta \in \Theta$, the function G_θ is K_1^p-Lipschitz on $\mathbb{R}^d$.

Let us now prove that $\mathscr{D} \subseteq \mathrm{Lip}_1$. Fix $D_\alpha \in \mathscr{D}$, $\alpha \in \Lambda$. According to (4), we have, for $x \in E$, $D_\alpha(x) = f_q \circ \cdots \circ f_1(x)$, where $f_i(t) = \tilde{\sigma}(V_i t + c_i)$ for $i = 1, \ldots, q-1$ ($\tilde{\sigma}$ is applied on pairs of components), and $f_q(t) = V_q t + c_q$.

Consequently, for $(x, y) \in E^2$,

$$\begin{aligned}
\|f_1(x) - f_1(y)\|_\infty &\leqslant \|V_1 x - V_1 y\|_\infty \\
&\quad \text{(since } \tilde{\sigma} \text{ is 1-Lipschitz)} \\
&= \|V_1(x - y)\|_\infty \\
&\leqslant \|V_1\|_{2,\infty} \, \|x - y\| \\
&\leqslant \|x - y\| \\
&\quad \text{(by Assumption 1).}
\end{aligned}$$

Thus,

$$\begin{aligned}
\|f_2 \circ f_1(x) - f_2 \circ f_1(y)\|_\infty &\leqslant \|V_2 f_1(x) - V_2 f_1(y)\|_\infty \\
&\quad \text{(since } \tilde{\sigma} \text{ is 1-Lipschitz)} \\
&\leqslant \|V_2\|_\infty \, \|f_1(x) - f_1(y)\|_\infty \\
&\leqslant \|f_1(x) - f_1(y)\|_\infty \\
&\quad \text{(by Assumption 1)} \\
&\leqslant \|x - y\|.
\end{aligned}$$

Repeating this, we conclude that, for each $\alpha \in \Lambda$ and all $(x, y) \in E^2$, $|D_\alpha(x) - D_\alpha(y)| \leqslant \|x - y\|$, which is the desired result.

A.2 Proof of Proposition 2

We first prove that the function $\Theta \ni \theta \mapsto \mu_\theta$ is continuous with respect to the weak topology in $P_1(E)$. Let G_θ and $G_{\theta'}$ be two elements of $\mathscr{G}$, with $(\theta, \theta') \in \Theta^2$. Using (3), we write $G_\theta(z) = f_p \circ \cdots \circ f_1(z)$ (respectively, $G_{\theta'}(z) = f'_p \circ \cdots \circ f'_1(z)$), where $f_i(x) = \max(U_i x + b_i, 0)$ (respectively, $f'_i(x) = \max(U'_i x + b'_i, 0)$) for $i = 1, \ldots, p-1$, and $f_p(x) = U_p x + b_p$ (respectively, $f'_p(x) = U'_p x + b'_p$).

Clearly, for $z \in \mathbb{R}^d$,

$$\begin{aligned}\|f_1(z) - f'_1(z)\| &\leqslant \|U_1 z + b_1 - U'_1 z - b'_1\| \\ &\leqslant \|(U_1 - U'_1)z\| + \|b_1 - b'_1\| \\ &\leqslant \|U_1 - U'_1\|_2 \, \|z\| + \|b_1 - b'_1\| \\ &\leqslant (\|z\| + 1)\|\theta - \theta'\|.\end{aligned}$$

Similarly, for any $i \in \{2, \ldots, p\}$ and any $x \in \mathbb{R}^{u_i}$,

$$\|f_i(x) - f'_i(x)\| \leqslant (\|x\| + 1)\|\theta - \theta'\|.$$

Observe that

$$\begin{aligned}&\|G_\theta(z) - G_{\theta'}(z)\| \\ &\quad = \|f_p \circ \cdots \circ f_1(z) - f'_p \circ \cdots \circ f'_1(z)\| \\ &\quad \leqslant \|f_p \circ \cdots \circ f_1(z) - f_p \circ \cdots \circ f_2 \circ f'_1(z)\| + \cdots + \|f_p \circ f'_{p-1} \circ \cdots \circ f'_1(z) - f'_p \circ \cdots \circ f'_1(z)\|.\end{aligned}$$

As in the proof of Lemma 1, one shows that for any $i \in \{1, \ldots, p\}$, the function $f_p \circ \cdots \circ f_i$ is K_1^{p-i+1}-Lipschitz with respect to the Euclidean norm. Therefore,

$$\begin{aligned}&\|G_\theta(z) - G_{\theta'}(z)\| \\ &\quad \leqslant K_1^{p-1}\|f_1(z) - f'_1(z)\| + \cdots + K_1^0\|f_p \circ f'_{p-1} \circ \cdots \circ f'_1(z) - f'_p \circ \cdots \circ f'_1(z)\| \\ &\quad \leqslant K_1^{p-1}(\|z\| + 1)\|\theta - \theta'\| + \cdots + (\|f'_{p-1} \circ \cdots \circ f'_1(z)\| + 1)\|\theta - \theta'\| \\ &\quad \leqslant K_1^{p-1}(\|z\| + 1)\|\theta - \theta'\| + \cdots + (K_1^{p-1}\|z\| + \|f'_{p-1} \circ \cdots \circ f'_1(0)\| + 1)\|\theta - \theta'\|.\end{aligned}$$

Using the architecture of neural networks in (3), a quick check shows that, for each $i \in \{1, \ldots, p\}$,

$$\|f'_i \circ \cdots \circ f'_1(0)\| \leqslant \sum_{k=1}^{i} K_1^k.$$

We are led to

$$\|G_\theta(z) - G_{\theta'}(z)\| = (\ell_1\|z\| + \ell_2)\|\theta - \theta'\|, \tag{19}$$

where

$$\ell_1 = pK_1^{p-1} \quad \text{and} \quad \ell_2 = \sum_{i=1}^{p-1} K_1^{p-(i+1)} \sum_{k=1}^{i} K_1^k + \sum_{i=0}^{p-1} K_1^i.$$

Denoting by ν the probability distribution of the sub-Gaussian random variable Z, we note that $\int_{\mathbb{R}^d}(\ell_1\|z\| + \ell_2)\nu(\mathrm{d}z) < \infty$. Now, let (θ_k) be a sequence in Θ converging to $\theta \in \Theta$

with respect to the Euclidean norm. Clearly, for a given $z \in \mathbb{R}^d$, by continuity of the function $\theta \mapsto G_\theta(z)$, we have $\lim_{k\to\infty} G_{\theta_k}(z) = G_\theta(z)$ and, for any $\varphi \in C_b(E)$, $\lim_{k\to\infty} \varphi(G_{\theta_k}(z)) = \varphi(G_\theta(z))$. Thus, by the dominated convergence theorem,

$$\lim_{k\to\infty} \int_E \varphi(x)\mu_{\theta_k}(\mathrm{d}x) = \lim_{k\to\infty} \int_{\mathbb{R}^d} \varphi(G_{\theta_k}(z))\nu(\mathrm{d}z) = \int_{\mathbb{R}^d} \varphi\big(G_\theta(z)\big)\nu(\mathrm{d}z) = \int_E \varphi(x)\mu_\theta(\mathrm{d}x). \tag{20}$$

This shows that the sequence (μ_{θ_k}) converges weakly to μ_θ. Besides, for an arbitrary x_0 in E, we have

$$\begin{aligned}
&\limsup_{k\to\infty} \int_E \|x_0 - x\|\mu_{\theta_k}(\mathrm{d}x) \\
&\quad = \limsup_{k\to\infty} \int_{\mathbb{R}^d} \|x_0 - G_{\theta_k}(z)\|\nu(\mathrm{d}z) \\
&\quad \leqslant \limsup_{k\to\infty} \int_{\mathbb{R}^d} \big(\|G_{\theta_k}(z) - G_\theta(z)\| + \|G_\theta(z) - x_0\|\big)\nu(\mathrm{d}z) \\
&\quad \leqslant \limsup_{k\to\infty} \int_{\mathbb{R}^d} (\ell_1\|z\| + \ell_2)\|\theta_k - \theta\|\nu(\mathrm{d}z) + \int_{\mathbb{R}^d} \|G_\theta(z) - x_0\|\nu(\mathrm{d}z) \\
&\qquad \text{(by inequality (19)).}
\end{aligned}$$

Consequently,

$$\limsup_{k\to\infty} \int_E \|x_0 - x\|\mu_{\theta_k}(\mathrm{d}x) \leqslant \int_{\mathbb{R}^d} \|G_\theta(z) - x_0\|\nu(\mathrm{d}z) = \int_E \|x_0 - x\|\mu_\theta(\mathrm{d}x).$$

One proves with similar arguments that

$$\liminf_{k\to\infty} \int_E \|x_0 - x\|\mu_{\theta_k}(\mathrm{d}x) \geqslant \int_E \|x_0 - x\|\mu_\theta(\mathrm{d}x).$$

Therefore, putting all the pieces together, we conclude that

$$\lim_{k\to\infty} \int_E \|x_0 - x\|\mu_{\theta_k}(\mathrm{d}x) = \int_E \|x_0 - x\|\mu_\theta(\mathrm{d}x).$$

This, together with (20), shows that the sequence (μ_{θ_k}) converges weakly to μ_θ in $P_1(E)$, and, in turn, that the function $\Theta \ni \theta \mapsto \mu_\theta$ is continuous with respect to the weak topology in $P_1(E)$, as desired.

The second assertion of the proposition follows upon noting that $\mathscr{P}$ is the image of the compact set Θ by a continuous function.

A.3 Proof of Proposition 3

To show the first statement, we are to exhibit a specific discriminator, say $\mathscr{D}_{\max}$, such that, for all $(\mu, \nu) \in (\mathscr{P} \cup \{\mu^\star\})^2$, the identity $d_{\mathscr{D}_{\max}}(\mu, \nu) = 0$ implies $\mu = \nu$.

Let $\varepsilon > 0$. According to Proposition 2, under Assumption 1, $\mathscr{P}$ is a compact subset of $P_1(E)$ with respect to the weak topology in $P_1(E)$. Let $x_0 \in E$ be arbitrary. For any $\mu \in \mathscr{P}$ there exists a compact $K_\mu \subseteq E$ such that $\int_{K_\mu^\complement} \|x_0 - x\|\mu(\mathrm{d}x) \leqslant \varepsilon/4$. Also, for any such K_μ,

the function $P_1(E) \ni \rho \mapsto \int_{K_\mu^\complement} \|x_0 - x\| \rho(\mathrm{d}x)$ is continuous. Therefore, there exists an open set $U_\mu \subseteq P_1(E)$ containing μ such that, for any $\rho \in U_\mu$, $\int_{K_\mu^\complement} \|x_0 - x\| \rho(\mathrm{d}x) \leqslant \varepsilon/2$.

The collection of open sets $\{U_\mu : \mu \in \mathscr{P}\}$ forms an open cover of $\mathscr{P}$, from which we can extract, by compactness, a finite subcover $U_{\mu_1}, \ldots, U_{\mu_n}$. Letting $K_1 = \cup_{i=1}^n K_{\mu_i}$, we deduce that, for all $\mu \in \mathscr{P}$, $\int_{K_1^\complement} \|x_0 - x\| \mu(\mathrm{d}x) \leqslant \varepsilon/2$. We conclude that there exists a compact $K \subseteq E$ and $x_0 \in K$ such that, for any $\mu \in \mathscr{P} \cup \{\mu^\star\}$,

$$\int_{K^\complement} ||x_0 - x|| \mu(\mathrm{d}x) \leqslant \varepsilon/2.$$

By Arzelà-Ascoli theorem, it is easy to see that $\mathrm{Lip}_1(K)$, the set of 1-Lipschitz real-valued functions on K, is compact with respect to the uniform norm $\|\cdot\|_\infty$ on K. Let $\{f_1, \ldots, f_{\mathscr{N}_\varepsilon}\}$ denote an ε-covering of $\mathrm{Lip}_1(K)$. According to Anil et al. (2019, Theorem 3), for each $k = 1, \ldots, \mathscr{N}_\varepsilon$ there exists under Assumption 1 a discriminator $\mathscr{D}_k$ of the form (4) such that

$$\inf_{g \in \mathscr{D}_k} \|f_k - g\mathbb{1}_K\|_\infty \leqslant \varepsilon.$$

Since the discriminative classes of functions use GroupSort activations, one can find a neural network of the form (4) satisfying Assumption 1, say $\mathscr{D}_{\max}$, such that, for all $k \in \{1, \ldots, \mathscr{N}_\varepsilon\}$, $\mathscr{D}_k \subseteq \mathscr{D}_{\max}$. Consequently, for any $f \in \mathrm{Lip}_1(K)$, letting $k_0 \in \operatorname*{arg\,min}_{k \in \{1, \ldots, \mathscr{N}_\varepsilon\}} \|f - f_k\|_\infty$, we have

$$\inf_{g \in \mathscr{D}_{\max}} \|f - g\mathbb{1}_K\|_\infty \leqslant \|f - f_{k_0}\|_\infty + \inf_{g \in \mathscr{D}_{\max}} \|f_{k_0} - g\mathbb{1}_K\|_\infty \leqslant 2\varepsilon.$$

Now, let $(\mu, \nu) \in (\mathscr{P} \bigcup \{\mu^\star\})^2$ be such that $d_{\mathscr{D}_{\max}}(\mu, \nu) = 0$, i.e., $\sup_{f \in \mathscr{D}_{\max}} |\mathbb{E}_\mu f - \mathbb{E}_\nu f| = 0$. Let $f^\star$ be a function in Lip_1 such that $\mathbb{E}_\mu f^\star - \mathbb{E}_\nu f^\star = d_{\mathrm{Lip}_1}(\mu, \nu)$ (such a function exists according to (8)) and, without loss of generality, such that $f^\star(x_0) = 0$. Clearly,

$$\begin{aligned}
d_{\mathrm{Lip}_1}(\mu, \nu) &= \mathbb{E}_\mu f^\star - \mathbb{E}_\nu f^\star \\
&\leqslant \Big| \int_K f^\star \mathrm{d}\mu - \int_K f^\star \mathrm{d}\nu \Big| + \Big| \int_{K^\complement} f^\star \mathrm{d}\mu - \int_{K^\complement} f^\star \mathrm{d}\nu \Big| \\
&\leqslant \Big| \int_K f^\star \mathrm{d}\mu - \int_K f^\star \mathrm{d}\nu \Big| + \varepsilon.
\end{aligned}$$

Letting $g_{f^\star} \in \mathscr{D}_{\max}$ be such that

$$\|(f^\star - g_{f^\star})\mathbb{1}_K\|_\infty \leqslant \inf_{g \in \mathscr{D}_{\max}} \|(f^\star - g)\mathbb{1}_K\|_\infty + \varepsilon \leqslant 3\varepsilon,$$

we are thus led to

$$d_{\mathrm{Lip}_1}(\mu, \nu) \leqslant \Big| \int_K (f^\star - g_{f^\star}) \mathrm{d}\mu - \int_K (f^\star - g_{f^\star}) \mathrm{d}\nu + \int_K g_{f^\star} \mathrm{d}\mu - \int_K g_{f^\star} \mathrm{d}\nu \Big| + \varepsilon.$$

Observe, since $x_0 \in K$, that $|g_{f^\star}(x_0)| \leqslant 3\varepsilon$ and that, for any $x \in E$, $|g_{f^\star}(x)| \leqslant \|x_0 - x\| + 3\varepsilon$. Exploiting $\mathbb{E}_\mu g_{f^\star} - \mathbb{E}_\nu g_{f^\star} = 0$, we obtain

$$\begin{aligned}
d_{\mathrm{Lip}_1}(\mu, \nu) &\leqslant 7\varepsilon + \Big| \int_{K^\complement} g_{f^\star} \mathrm{d}\mu - \int_{K^\complement} g_{f^\star} \mathrm{d}\nu \Big| \\
&\leqslant 7\varepsilon + \int_{K^\complement} \|x_0 - x\| \mu(\mathrm{d}x) + \int_{K^\complement} \|x_0 - x\| \nu(\mathrm{d}x) + 6\varepsilon \\
&\leqslant 14\varepsilon.
\end{aligned}$$

Since ε is arbitrary and d_{Lip_1} is a metric on $P_1(E)$, we conclude that $\mu = \nu$, as desired.

To complete the proof, it remains to show that $d_{\mathscr{D}_{\max}}$ metrizes weak convergence in $\mathscr{P} \cup \{\mu^\star\}$. To this aim, we let (μ_k) be a sequence in $\mathscr{P} \cup \{\mu^\star\}$ and μ be a probability measure in $\mathscr{P} \cup \{\mu^\star\}$.

If (μ_k) converges weakly to μ in $P_1(E)$, then $d_{\mathrm{Lip}_1}(\mu, \mu_k) \to 0$ (Villani, 2008, Theorem 6.8), and, accordingly, $d_{\mathscr{D}_{\max}}(\mu, \mu_k) \to 0$.

Suppose, on the other hand, that $d_{\mathscr{D}_{\max}}(\mu, \mu_k) \to 0$, and fix $\varepsilon > 0$. There exists $M > 0$ such that, for all $k \geqslant M$, $d_{\mathscr{D}_{\max}}(\mu, \mu_k) \leqslant \varepsilon$. Using a similar reasoning as in the first part of the proof, it is easy to see that for any $k \geqslant M$, we have $d_{\mathrm{Lip}_1}(\mu, \mu_k) \leqslant 15\varepsilon$. Since the Wasserstein distance metrizes weak convergence in $P_1(E)$ and ε is arbitrary, we conclude that (μ_k) converges weakly to μ in $P_1(E)$.

A.4 Proof of Lemma 4

Using a similar reasoning as in the proof of Proposition 2, one easily checks that for all $(\alpha, \alpha') \in \Lambda^2$ and all $x \in E$,

$$\begin{aligned}|D_\alpha(x) - D_{\alpha'}(x)| &\leqslant Q^{1/2}\big(q\|x\| + K_2 \sum_{i=1}^{q-1} i + q\big)\|\alpha - \alpha'\| \\ &\leqslant Q^{1/2}\big(q\|x\| + \frac{q(q-1)K_2}{2} + q\big)\|\alpha - \alpha'\|,\end{aligned}$$

where q refers to the depth of the discriminator. Thus, since $\mathscr{D} \subset \mathrm{Lip}_1$ (by Lemma 1), we have, for all $\alpha \in \Lambda$, all $x \in E$, and any arbitrary $x_0 \in E$,

$$\begin{aligned}|D_\alpha(x)| &\leqslant |D_\alpha(x) - D_\alpha(x_0)| + |D_\alpha(x_0)| \\ &\leqslant \|x_0 - x\| + Q^{1/2}\big(q\|x_0\| + \frac{q(q-1)K_2}{2} + q\big)\|\alpha\| \\ &\quad \text{(upon noting that } D_0(x_0) = 0) \\ &\leqslant \|x_0 - x\| + Q^{1/2}\big(q\|x_0\| + \frac{q(q-1)K_2}{2} + q\big)Q^{1/2}\max(K_2, 1),\end{aligned}$$

where Q is the dimension of Λ. Thus, since $\mu^\star$ and the μ_θ's belong to $P_1(E)$ (by Lemma 1), we deduce that all $D_\alpha \in \mathscr{D}$ are dominated by a function independent of α and integrable with respect to $\mu^\star$ and μ_θ. In addition, for all $x \in E$, the function $\alpha \mapsto D_\alpha(x)$ is continuous on Λ. Therefore, by the dominated convergence theorem, the function $\Lambda \ni \alpha \mapsto |\mathbb{E}_{\mu^\star} D_\alpha - \mathbb{E}_{\mu_\theta} D_\alpha|$ is continuous. The conclusion follows from the compactness of the set Λ (Assumption 1).

A.5 Proof of Theorem 5

Let $(\theta, \theta') \in \Theta^2$, and let γ_Z be the joint distribution of the pair $(G_\theta(Z), G_{\theta'}(Z))$. We have

$$\begin{aligned}|\xi_{\mathrm{Lip}_1}(\theta) - \xi_{\mathrm{Lip}_1}(\theta')| &= |d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta'})| \\ &\leqslant d_{\mathrm{Lip}_1}(\mu_\theta, \mu_{\theta'}) \\ &= \inf_{\gamma \in \Pi(\mu_\theta, \mu_{\theta'})} \int_{E^2} \|x - y\| \gamma(\mathrm{d}x, \mathrm{d}y),\end{aligned}$$

where $\Pi(\mu_\theta, \mu_{\theta'})$ denotes the collection of all joint probability measures on $E \times E$ with marginals μ_θ and $\mu_{\theta'}$. Thus,

$$\begin{aligned}
|\xi_{\mathrm{Lip}_1}(\theta) - \xi_{\mathrm{Lip}_1}(\theta')| &\leqslant \int_{E^2} \|x - y\| \gamma_Z(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathbb{R}^d} \|G_\theta(z) - G_{\theta'}(z)\| \nu(\mathrm{d}z) \\
&\quad \text{(where } \nu \text{ is the distribution of } Z\text{)} \\
&\leqslant \|\theta - \theta'\| \int_{\mathbb{R}^d} (\ell_1 \|z\| + \ell_2) \nu(\mathrm{d}z) \\
&\quad \text{(by inequality (19))}.
\end{aligned}$$

This shows that the function $\theta \ni \Theta \mapsto \xi_{\mathrm{Lip}_1}(\theta)$ is L-Lipschitz, with $L = \int_{\mathbb{R}^d} (\ell_1 \|z\| + \ell_2) \nu(\mathrm{d}z)$. For the second statement of the theorem, just note that

$$\begin{aligned}
|\xi_{\mathscr{D}}(\theta) - \xi_{\mathscr{D}}(\theta')| &= |d_{\mathscr{D}}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_{\theta'})| \\
&\leqslant d_{\mathscr{D}}(\mu_\theta, \mu_{\theta'}) \\
&\leqslant d_{\mathrm{Lip}_1}(\mu_\theta, \mu_{\theta'}) \\
&\quad \text{(since } \mathscr{D} \subseteq \mathrm{Lip}_1\text{)} \\
&\leqslant L \|\theta - \theta'\|.
\end{aligned}$$

A.6 Proof of Theorem 7

The proof is divided into two parts. First, we show that under Assumption 1, for all $\varepsilon > 0$ and $\theta \in \Theta$, there exists a discriminator $\mathscr{D}$ (function of ε and θ) of the form (4) such that

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant 6\varepsilon.$$

Let $f^\star$ be a function in Lip_1 such that $\mathbb{E}_{\mu^\star} f^\star - \mathbb{E}_{\mu_\theta} f^\star = d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ (such a function exists according to (8)). We may write

$$\begin{aligned}
d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta) &= \mathbb{E}_{\mu^\star} f^\star - \mathbb{E}_{\mu_\theta} f^\star - \sup_{f \in \mathscr{D}} |\mathbb{E}_{\mu^\star} f - \mathbb{E}_{\mu_\theta} f| \\
&= \mathbb{E}_{\mu^\star} f^\star - \mathbb{E}_{\mu_\theta} f^\star - \sup_{f \in \mathscr{D}} (\mathbb{E}_{\mu^\star} f - \mathbb{E}_{\mu_\theta} f) \\
&= \inf_{f \in \mathscr{D}} (\mathbb{E}_{\mu^\star} f^\star - \mathbb{E}_{\mu_\theta} f^\star - \mathbb{E}_{\mu^\star} f + \mathbb{E}_{\mu_\theta} f) \\
&= \inf_{f \in \mathscr{D}} (\mathbb{E}_{\mu^\star}(f^\star - f) - \mathbb{E}_{\mu_\theta}(f^\star - f)) \\
&\leqslant \inf_{f \in \mathscr{D}} (\mathbb{E}_{\mu^\star} |f^\star - f| + \mathbb{E}_{\mu_\theta} |f^\star - f|). \qquad (21)
\end{aligned}$$

Next, for any $f \in \mathscr{D}$ and any compact $K \subseteq E$,

$$\begin{aligned}
\mathbb{E}_{\mu^\star} |f^\star - f| &= \mathbb{E}_{\mu^\star} |f^\star - f| \mathbb{1}_K + \mathbb{E}_{\mu^\star} |f^\star - f| \mathbb{1}_{K^\complement} \\
&\leqslant \|(f^\star - f) \mathbb{1}_K\|_\infty + \mathbb{E}_{\mu^\star} |f^\star| \mathbb{1}_{K^\complement} + \mathbb{E}_{\mu^\star} |f| \mathbb{1}_{K^\complement}.
\end{aligned}$$

Since $f^\star$ and f are integrable with respect to $\mu^\star$ and μ_θ, there exists a compact set K (function of ε and θ) such that

$$\max\left(\mathbb{E}_{\mu^\star}|f^\star|\mathbb{1}_{K^\complement},\ \mathbb{E}_{\mu^\star}|f|\mathbb{1}_{K^\complement},\ \mathbb{E}_{\mu_\theta}|f^\star|\mathbb{1}_{K^\complement},\ \mathbb{E}_{\mu_\theta}|f|\mathbb{1}_{K^\complement}\right) \leqslant \varepsilon.$$

Thus, for such a choice of K,

$$\mathbb{E}_{\mu^\star}|f^\star - f| \leqslant \|(f^\star - f)\mathbb{1}_K\|_\infty + 2\varepsilon.$$

Similarly,

$$\mathbb{E}_{\mu_\theta}|f^\star - f| \leqslant \|(f^\star - f)\mathbb{1}_K\|_\infty + 2\varepsilon.$$

Plugging the two inequalities above in (21), we obtain

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant 2\inf_{f\in\mathscr{D}} \|(f^\star - f)\mathbb{1}_K\|_\infty + 4\varepsilon.$$

According to Anil et al. (2019, Theorem 3), under Assumption 1, we can find a discriminator of the form (4) such that $\inf_{f\in\mathscr{D}} \|(f^\star - f)\mathbb{1}_K\|_\infty \leqslant \varepsilon$. We conclude that, for this choice of $\mathscr{D}$ (function of ε and θ),

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant 6\varepsilon, \tag{22}$$

as desired.

For the second part of the proof, we fix $\varepsilon > 0$ and let, for each $\theta \in \Theta$ and each discriminator of the form (4),

$$\hat{\xi}_{\mathscr{D}}(\theta) = d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta).$$

Arguing as in the proof of Theorem 5, we see that $\hat{\xi}_{\mathscr{D}}(\theta)$ is $2L$-Lipschitz in θ, where $L = \int_{\mathbb{R}^d}(\ell_1\|z\| + \ell_2)\nu(\mathrm{d}z)$ and ν is the probability distribution of Z.

Now, let $\{\theta_1, \ldots, \theta_{\mathscr{N}_\varepsilon}\}$ be an ε-covering of the compact set Θ, i.e., for each $\theta \in \Theta$, there exists $k \in \{1, \ldots, \mathscr{N}_\varepsilon\}$ such that $\|\theta - \theta_k\| \leqslant \varepsilon$. According to (22), for each such k, there exists a discriminator $\mathscr{D}_k$ such that $\hat{\xi}_{\mathscr{D}_k}(\theta_k) \leqslant 6\varepsilon$. Since the discriminative classes of functions use GroupSort activation functions, one can find a neural network of the form (4) satisfying Assumption 1, say $\mathscr{D}_{\max}$, such that, for all $k \in \{1, \ldots, \mathscr{N}_\varepsilon\}$, $\mathscr{D}_k \subseteq \mathscr{D}_{\max}$. Clearly, $\hat{\xi}_{\mathscr{D}_{\max}}(\theta)$ is $2L$-Lipschitz, and, for all $k \in \{1, \ldots, \mathscr{N}_\varepsilon\}$, $\hat{\xi}_{\mathscr{D}_{\max}}(\theta_k) \leqslant 6\varepsilon$. Hence, for all $\theta \in \Theta$, letting

$$\hat{k} \in \operatorname*{arg\,min}_{k\in\{1,\ldots,\mathscr{N}_\varepsilon\}} \|\theta - \theta_k\|,$$

we have

$$\hat{\xi}_{\mathscr{D}_{\max}}(\theta) \leqslant \left|\hat{\xi}_{\mathscr{D}_{\max}}(\theta) - \hat{\xi}_{\mathscr{D}_{\max}}(\theta_{\hat{k}})\right| + \hat{\xi}_{\mathscr{D}_{\max}}(\theta_{\hat{k}}) \leqslant (2L + 6)\varepsilon.$$

Therefore,

$$T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}_{\max}) = \sup_{\theta\in\Theta}\left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}_{\max}}(\mu^\star, \mu_\theta)\right] = \sup_{\theta\in\Theta} \hat{\xi}_{\mathscr{D}_{\max}}(\theta) \leqslant (2L + 6)\varepsilon.$$

We have just proved that, for all $\varepsilon > 0$, there exists a discriminator $\mathscr{D}_{\max}$ of the form (4) and a positive constant c (independent of ε) such that

$$T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}_{\max}) \leqslant c\varepsilon.$$

This is the desired result.

A.7 Proof of Proposition 9

Let us assume that the statement is not true. If so, there exists $\varepsilon > 0$ such that, for all $\delta > 0$, there exists $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$ satisfying $d(\theta, \bar{\Theta}) > \varepsilon$. Consider $\delta_n = 1/n$, and choose a sequence of parameters (θ_n) such that

$$\theta_n \in \mathscr{M}_{d_{\mathscr{D}}}\left(\mu^\star, \frac{1}{n}\right) \quad \text{and} \quad d(\theta_n, \bar{\Theta}) > \varepsilon.$$

Since Θ is compact by Assumption 1, we can find a subsequence (θ_{φ_n}) that converges to some $\theta_{\mathrm{acc}} \in \Theta$. Thus, for all $n \geqslant 1$, we have

$$d_{\mathscr{D}}(\mu^\star, \mu_{\theta_{\varphi_n}}) \leqslant \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) + \frac{1}{n},$$

and, by continuity of the function $\Theta \ni \theta \mapsto d_{\mathscr{D}}(\mu^\star, \mu_\theta)$ (Theorem 5),

$$d_{\mathscr{D}}(\mu^\star, \theta_{\mathrm{acc}}) \leqslant \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta).$$

We conclude that θ_{acc} belongs to $\bar{\Theta}$. This contradicts the fact that $d(\theta_{\mathrm{acc}}, \bar{\Theta}) \geqslant \varepsilon$.

A.8 Proof of Lemma 13

Since $a = b$, according to Definition 12, there exists a continuously differentiable, strictly increasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that, for all $\mu \in \mathscr{P}$,

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu) = f(d_{\mathscr{D}}(\mu^\star, \mu)).$$

For $(\theta, \theta') \in \Theta^2$ we have, as f is strictly increasing,

$$d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant d_{\mathscr{D}}(\mu^\star, \mu_{\theta'}) \iff f(d_{\mathscr{D}}(\mu^\star, \mu_\theta)) \leqslant f(d_{\mathscr{D}}(\mu^\star, \mu_{\theta'})).$$

Therefore,

$$d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant d_{\mathscr{D}}(\mu^\star, \mu_{\theta'}) \iff d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta'}).$$

This proves the first statement of the lemma.

Let us now show that d_{Lip_1} can be fully substituted by $d_{\mathscr{D}}$. Let $\varepsilon > 0$. Then, for $\delta > 0$ (function of ε, to be chosen later) and $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$, we have

$$\begin{aligned} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) &= f(d_{\mathscr{D}}(\mu^\star, \mu_\theta)) - \inf_{\theta \in \Theta} f(d_{\mathscr{D}}(\mu^\star, \mu_\theta)) \\ &= f(d_{\mathscr{D}}(\mu^\star, \mu_\theta)) - f(\inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)) \\ &\leqslant \sup_{\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)} \left| f(d_{\mathscr{D}}(\mu^\star, \mu_\theta)) - f(\inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)) \right|. \end{aligned}$$

According to Theorem 5, there exists a nonnegative constant c such that for any $\theta \in \Theta$, $d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant c$. Therefore, using the definition of $\mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$ and the fact that f is continuously differentiable, we are led to

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \leqslant \delta \sup_{x \in [0,c]} \left| \frac{\partial f(x)}{\partial x} \right|.$$

The conclusion follows by choosing δ such that $\delta \sup_{x \in [0,c]} |\frac{\partial f(x)}{\partial x}| \leqslant \varepsilon$.

A.9 Proof of Proposition 14

Let $\delta \in (0,1)$ and $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu^\star, \delta)$, i.e., $d_{\mathscr{D}}(\mu^\star, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \leqslant \delta$. As d_{Lip_1} is monotonously equivalent to $d_{\mathscr{D}}$, there exists a continuously differentiable, strictly increasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ and $(a,b) \in (\mathbb{R}_+^\star)^2$ such that

$$\forall \mu \in \mathscr{P},\ af(d_{\mathscr{D}}(\mu^\star, \mu)) \leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu) \leqslant bf(d_{\mathscr{D}}(\mu^\star, \mu)).$$

So,

$$\begin{aligned} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) &\leqslant bf(\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) + \delta) \\ &\leqslant bf(\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)) + O(\delta). \end{aligned}$$

Also,

$$\inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \geqslant af(\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)).$$

Therefore,

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \leqslant (b-a)f(\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)) + O(\delta).$$

A.10 Proof of Lemma 15

Let $f : \mathbb{R}^D \to \mathbb{R}$ be in $\mathrm{AFF} \cap \mathrm{Lip}_1$. It is of the form $f(x) = x \cdot u + b$, where $u = (u_1, \ldots, u_D)$, $b \in \mathbb{R}$, and $\|u\| \leqslant 1$. Our objective is to prove that there exists a discriminator of the form (4) with $q = 2$ and $v_1 = 2$ that contains the function f. To see this, define $V_1 \in \mathscr{M}_{(2,D)}$ and the offset vector $c_1 \in \mathscr{M}_{(2,1)}$ as

$$V_1 = \begin{bmatrix} u_1 & \cdots & u_D \\ u_1 & \cdots & u_D \end{bmatrix} \quad \text{and} \quad c_1 = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Letting $V_2 \in \mathscr{M}_{(1,2)}$, $c_2 \in \mathscr{M}_{(1,1)}$ be

$$V_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad c_2 = \begin{bmatrix} b \end{bmatrix},$$

we readily obtain $V_2\tilde{\sigma}(V_1 x + c_1) + c_2 = f(x)$. Besides, it is easy to verify that $\|V_1\|_{2,\infty} \leqslant 1$.

A.11 Proof of Lemma 16

Let μ and ν be two probability measures in $P_1(E)$ with supports S_μ and S_ν satisfying the conditions of the lemma. Let π be an optimal coupling between μ and ν, and let (X, Y) be a random pair with distribution π such that

$$d_{\mathrm{Lip}_1}(\mu, \nu) = \mathbb{E}\|X - Y\|.$$

Clearly, any function $f_0 \in \mathrm{Lip}_1$ satisfying $f_0(X) - f_0(Y) = \|X - Y\|$ almost surely will be such that

$$d_{\mathrm{Lip}_1}(\mu, \nu) = |\mathbb{E}_\mu f_0 - \mathbb{E}_\nu f_0|.$$

The proof will be achieved if we show that such a function f_0 exists and that it may be chosen linear. Since S_μ and S_ν are disjoint and convex, we can find a unit vector u of $\mathbb{R}^D$ included in the line containing both S_μ and S_ν such that $(x_0 - y_0) \cdot u > 0$, where (x_0, y_0) is an arbitrary pair of $S_\mu \times S_\nu$. Letting $f_0(x) = x \cdot u$ $(x \in E)$, we have, for all $(x, y) \in S_\mu \times S_\nu$, $f_0(x) - f_0(y) = (x - y) \cdot u = \|x - y\|$. Since f_0 is a linear and 1-Lipschitz function on E, this concludes the proof.

A.12 Proof of Lemma 17

For any pair of probability measures (μ, ν) on E with finite moment of order 2, we let $W_2(\mu, \nu)$ be the Wasserstein distance of order 2 between μ and ν. Recall (Villani, 2008, Definition 6.1) that

$$W_2(\mu, \nu) = \Big(\inf_{\pi \in \Pi(\mu,\nu)} \int_{E \times E} \|x - y\|^2 \pi(\mathrm{d}x, \mathrm{d}y) \Big)^{1/2},$$

where $\Pi(\mu, \nu)$ denotes the collection of all joint probability measures on $E \times E$ with marginals μ and ν. By Jensen's inequality,

$$d_{\mathrm{Lip}_1}(\mu, \nu) = W_1(\mu, \nu) \leqslant W_2(\mu, \nu).$$

Let $\Sigma \in \mathscr{M}_{(D,D)}$ be a positive semi-definite matrix, and let μ be Gaussian $\mathscr{N}(m_1, \Sigma)$ and ν be Gaussian $\mathscr{N}(m_2, \Sigma)$. Denoting by (X, Y) a random pair with marginal distributions μ and ν such that

$$\mathbb{E}\|X - Y\| = W_1(\mu, \nu),$$

we have

$$\|m_1 - m_2\| = \|\mathbb{E}(X - Y)\| \leqslant \mathbb{E}\|X - Y\| = W_1(\mu, \nu) \leqslant W_2(\mu, \nu) = \|m_1 - m_2\|,$$

where the last equality follows from Givens and Shortt (1984, Proposition 7). Thus, $d_{\mathrm{Lip}_1}(\mu, \nu) = \|m_1 - m_2\|$. The proof will be finished if we show that

$$d_{\mathrm{AFF} \cap \mathrm{Lip}_1}(\mu, \nu) \geqslant \|m_1 - m_2\|.$$

To see this, consider the linear and 1-Lipschitz function $f : E \ni x \mapsto x \cdot \frac{(m_1 - m_2)}{\|m_1 - m_2\|}$ (with the convention $0 \times \infty = 0$), and note that

$$\begin{aligned} d_{\mathrm{AFF} \cap \mathrm{Lip}_1}(\mu, \nu) &\geqslant \Big| \int_E x \cdot \frac{(m_1 - m_2)}{\|m_1 - m_2\|} \mu(\mathrm{d}x) - \int_E y \cdot \frac{(m_1 - m_2)}{\|m_1 - m_2\|} \nu(\mathrm{d}y) \Big| \\ &= \Big| \int_E x \cdot \frac{(m_1 - m_2)}{\|m_1 - m_2\|} \mu(\mathrm{d}x) - \int_E (x - m_1 + m_2) \cdot \frac{(m_1 - m_2)}{\|m_1 - m_2\|} \mu(\mathrm{d}x) \Big| \\ &= \|m_1 - m_2\|. \end{aligned}$$

A.13 Proof of Proposition 18

Let $\varepsilon > 0$, and let μ and ν be two probability measures in $P_1(E)$ with compact supports S_μ and S_ν such that $\max(\mathrm{diam}(S_\mu), \mathrm{diam}(S_\nu)) \leqslant \varepsilon d(S_\mu, S_\nu)$. Throughout the proof, it is

assumed that $d(S_\mu, S_\nu) > 0$, otherwise the result is immediate. Let π be an optimal coupling between μ and ν, and let (X, Y) be a random pair with distribution π such that

$$d_{\mathrm{Lip}_1}(\mu, \nu) = \mathbb{E}\|X - Y\|.$$

Any function $f_0 \in \mathrm{Lip}_1$ satisfying $\|X - Y\| \leqslant (1 + 2\varepsilon)(f_0(X) - f_0(Y))$ almost surely will be such that

$$d_{\mathrm{Lip}_1}(\mu, \nu) \leqslant (1 + 2\varepsilon)|\mathbb{E}_\mu f_0 - \mathbb{E}_\nu f_0|.$$

Thus, the proof will be completed if we show that such a function f_0 exists and that it may be chosen affine.

Since S_μ and S_ν are compact, there exists $(x^\star, y^\star) \in S_\mu \times S_\nu$ such that $\|x^\star - y^\star\| = d(S_\mu, S_\nu)$. By the hyperplane separation theorem, there exists a hyperplane $\mathscr{H}$ orthogonal to the unit vector $u = \frac{x^\star - y^\star}{\|x^\star - y^\star\|}$ such that $d(x^\star, \mathscr{H}) = d(y^\star, \mathscr{H}) = \frac{\|x^\star - y^\star\|}{2}$. For any $x \in E$, we denote by $p_{\mathscr{H}}(x)$ the projection of x onto $\mathscr{H}$. We thus have $d(x, \mathscr{H}) = \|x - p_{\mathscr{H}}(x)\|$, and $\frac{x^\star + y^\star}{2} = p_{\mathscr{H}}(\frac{x^\star + y^\star}{2}) = p_{\mathscr{H}}(x^\star) = p_{\mathscr{H}}(y^\star)$. In addition, by convexity of S_μ and S_ν, for any $x \in S_\mu$, $\|x - p_{\mathscr{H}}(x)\| \geqslant \|x^\star - p_{\mathscr{H}}(x^\star)\|$. Similarly, for any $y \in S_\nu$, $\|y - p_{\mathscr{H}}(y)\| \geqslant \|y^\star - p_{\mathscr{H}}(y^\star)\|$.

Let the affine function f_0 be defined for any $x \in E$ by

$$f_0(x) = (x - p_{\mathscr{H}}(x)) \cdot u.$$

Observe that $f_0(x) = f_0(x + \frac{x^\star + y^\star}{2})$. Clearly, for any $(x, y) \in E^2$, one has

$$\begin{aligned}
|f_0(x) - f_0(y)| &= \big|f_0\big(x - y + \frac{x^\star + y^\star}{2}\big)\big| \\
&= \big|\big(\big(x - y + \frac{x^\star + y^\star}{2}\big) - p_{\mathscr{H}}\big(x - y + \frac{x^\star + y^\star}{2}\big)\big).u\big| \\
&\leqslant \big\|\big(x - y + \frac{x^\star + y^\star}{2}\big) - p_{\mathscr{H}}\big(x - y + \frac{x^\star + y^\star}{2}\big)\big\| \\
&\leqslant \big\|x - y + \frac{x^\star + y^\star}{2} - \frac{x^\star + y^\star}{2}\big\| \\
&\quad \text{(since } \tfrac{x^\star + y^\star}{2} \in \mathscr{H}\text{)} \\
&= \|x - y\|.
\end{aligned}$$

Thus, f_0 belongs to Lip_1. Besides, for any $(x, y) \in S_\mu \times S_\nu$, we have

$$\begin{aligned}
\|x - y\| &\leqslant \|x - p_{\mathscr{H}}(x)\| + \|p_{\mathscr{H}}(x) - p_{\mathscr{H}}(y)\| + \|p_{\mathscr{H}}(y) - y\| \\
&\leqslant (x - p_{\mathscr{H}}(x)) \cdot u - (y - p_{\mathscr{H}}(y)) \cdot u + \big\|p_{\mathscr{H}}(x) - \frac{x^\star + y^\star}{2}\big\| + \big\|p_{\mathscr{H}}(y) - \frac{x^\star + y^\star}{2}\big\| \\
&= (x - p_{\mathscr{H}}(x)) \cdot u - (y - p_{\mathscr{H}}(y)) \cdot u + \|p_{\mathscr{H}}(x) - p_{\mathscr{H}}(x^\star)\| + \|p_{\mathscr{H}}(y) - p_{\mathscr{H}}(y^\star)\|.
\end{aligned}$$

Thus,

$$\begin{aligned}\|x-y\| &\leqslant (x-p_{\mathscr{H}}(x))\cdot u-(y-p_{\mathscr{H}}(y))\cdot u+2\max(\mathrm{diam}(S_\mu),\mathrm{diam}(S_\nu))\\ &\leqslant f_0(x)-f_0(y)+2\varepsilon d(S_\mu,S_\nu)\\ &= f_0(x)-f_0(y)+2\varepsilon(f_0(x^\star)-f_0(y^\star))\\ &= f_0(x)-f_0(y)+2\varepsilon(f_0(x^\star)-f_0(x)+f_0(x)-f_0(y)+f_0(y)-f_0(y^\star))\\ &\leqslant (1+2\varepsilon)(f_0(x)-f_0(y))\\ &\quad \text{(using the fact that } f_0(x^\star)-f_0(x)\leqslant 0 \text{ and } f_0(y^\star)-f_0(y)\geqslant 0).\end{aligned}$$

Since $f_0\in \mathrm{Lip}_1$, we conclude that, for any $(x,y)\in S_\mu\times S_\nu$,

$$|f_0(x)-f_0(y)|\leqslant \|x-y\|\leqslant (1+2\varepsilon)(f_0(x)-f_0(y)).$$

A.14 Proof of Lemma 19

Using Dudley (2004, Theorem 11.4.1) and the strong law of large numbers, the sequence of empirical measures (μ_n) almost surely converges weakly in $P_1(E)$ to $\mu^\star$. Thus, we have $\lim_{n\to\infty} d_{\mathrm{Lip}_1}(\mu^\star,\mu_n)=0$ almost surely, and so $\lim_{n\to\infty} d_{\mathscr{D}}(\mu^\star,\mu_n)=0$ almost surely. Hence, recalling inequality (14), we conclude that

$$\sup_{\theta_n\in\hat{\Theta}_n} d_{\mathscr{D}}(\mu^\star,\mu_{\theta_n})-\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star,\mu_\theta)\to 0 \quad \text{almost surely.} \tag{23}$$

Now, fix $\varepsilon>0$ and recall that, by our Theorem 5, the function $\Theta\ni\theta\mapsto d_{\mathrm{Lip}_1}(\mu^\star,\mu_\theta)$ is L-Lipschitz, for some $L>0$. According to (23) and Proposition 9, almost surely, there exists an integer $N>0$ such that, for all $n\geqslant N$, for all $\theta_n\in\hat{\Theta}_n$, the companion $\bar{\theta}_n\in\bar{\Theta}$ is such that $\|\theta_n-\bar{\theta}_n\|\leqslant\frac{\varepsilon}{L}$. We conclude by observing that $|\varepsilon_{\mathrm{estim}}|\leqslant \sup_{\theta_n\in\hat{\Theta}_n}|d_{\mathrm{Lip}_1}(\mu^\star,\mu_{\theta_n})-d_{\mathrm{Lip}_1}(\mu^\star,\mu_{\bar{\theta}_n})|\leqslant L\times\frac{\varepsilon}{L}$.

A.15 Proof of Proposition 20

Let μ_n be the empirical measure based on n i.i.d. samples $X_1,\dots,X_n$ distributed according to $\mu^\star$. Recall (equation (7)) that

$$d_{\mathscr{D}}(\mu^\star,\mu_n)=\sup_{\alpha\in\Lambda}|\mathbb{E}_{\mu^\star}D_\alpha-\mathbb{E}_{\mu_n}D_\alpha|=\sup_{\alpha\in\Lambda}\left|\mathbb{E}_{\mu^\star}D_\alpha-\frac{1}{n}\sum_{i=1}^n D_\alpha(X_i)\right|.$$

Let g be the real-valued function defined on E^n by

$$g(x_1,\dots,x_n)=\sup_{\alpha\in\Lambda}\left|\mathbb{E}_{\mu^\star}D_\alpha-\frac{1}{n}\sum_{i=1}^n D_\alpha(x_i)\right|.$$

Observe that, for $(x_1, \dots, x_n) \in E^n$ and $(x'_1, \dots, x'_n) \in E^n$,

$$\begin{aligned} |g(x_1, \dots, x_n) - g(x'_1, \dots, x'_n)| &\leqslant \sup_{\alpha \in \Lambda} \Big| \frac{1}{n} \sum_{i=1}^n D_\alpha(x_i) - \frac{1}{n} \sum_{i=1}^n D_\alpha(x'_i) \Big| \\ &\leqslant \frac{1}{n} \sup_{\alpha \in \Lambda} \sum_{i=1}^n |D_\alpha(x_i) - D_\alpha(x'_i)| \\ &\leqslant \frac{1}{n} \sum_{i=1}^n \|x_i - x'_i\|. \end{aligned} \tag{24}$$

We start by examining statement (i), where $\mu^\star$ has compact support with diameter B. In this case, letting X'_i be an independent copy of X_i, we have, almost surely,

$$|g(X_1, \dots, X_n) - g(X_1, \dots, X'_i, \dots, X_n)| \leqslant \frac{B}{n}.$$

An application of McDiarmid's inequality (McDiarmid, 1989) shows that for any $\eta \in (0, 1)$, with probability at least $1 - \eta$,

$$d_{\mathscr{D}}(\mu^\star, \mu_n) \leqslant \mathbb{E} d_{\mathscr{D}}(\mu^\star, \mu_n) + B \sqrt{\frac{\log(1/\eta)}{2n}}. \tag{25}$$

Next, for each $\alpha \in \Lambda$, let Y_α denote the random variable defined by

$$Y_\alpha = \mathbb{E}_{\mu^\star} D_\alpha - \frac{1}{n} \sum_{i=1}^n D_\alpha(X_i).$$

Using a similar reasoning as in the proof of Proposition 2, one shows that for any $(\alpha, \alpha') \in \Lambda^2$ and any $x \in E$,

$$|D_\alpha(x) - D_{\alpha'}(x)| \leqslant Q^{1/2} \big(q\|x\| + \frac{q(q-1)K_2}{2} + q \big) \|\alpha - \alpha'\|,$$

where we recall that q is the depth of the discriminator. Since $\mu^\star$ has compact support,

$$\ell = \int_E Q^{1/2} \big(q\|x\| + \frac{q(q-1)K_2}{2} + q \big) \mu^\star(\mathrm{d}x) < \infty.$$

Observe that

$$|Y_\alpha - Y_{\alpha'}| \leqslant \frac{1}{n} \|\alpha - \alpha'\| \; |\xi(n)|,$$

where

$$\xi_n = \sum_{i=1}^n Q^{1/2} \big(\ell + q\|X_i\| + \frac{q(q-1)K_2}{2} + q \big).$$

Thus, using Vershynin (2018, Proposition 2.5.2), there exists a positive constant $c = O(qQ^{1/2}(D^{1/2} + q))$ such that, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E} e^{\lambda (Y_\alpha - Y_{\alpha'})} \leqslant \mathbb{E} e^{\lambda \frac{1}{n} \|\alpha - \alpha'\| \; |\xi_n|} \leqslant e^{c^2 \frac{1}{n} \|\alpha - \alpha'\|^2 \lambda^2}.$$

We conclude that the process (Y_α) is sub-Gaussian (van Handel, 2016, Definition 5.20) for the distance $d(\alpha,\alpha') = \frac{c\|\alpha-\alpha'\|}{\sqrt{n}}$. Therefore, using van Handel (2016, Corollary 5.25), we have

$$\mathbb{E}d_{\mathscr{D}}(\mu^\star,\mu_n) = \mathbb{E}\sup_{\alpha\in\Lambda}\left|\mathbb{E}_{\mu^\star}D_\alpha - \frac{1}{n}\sum_{i=1}^n D_\alpha(X_i)\right| \leqslant \frac{12c}{\sqrt{n}}\int_0^\infty \sqrt{\log \mathscr{N}(\Lambda,\|\cdot\|,u)}\mathrm{d}u,$$

where $\mathscr{N}(\Lambda,\|\cdot\|,u)$ is the u-covering number of Λ for the norm $\|\cdot\|$. Since Λ is bounded, there exists $r>0$ such that $\mathscr{N}(\Lambda,\|\cdot\|,u)=1$ for $u \geqslant rQ^{1/2}$ and

$$\mathscr{N}(\Lambda,\|\cdot\|,u) \leq \left(\frac{rQ^{1/2}}{u}\right)^Q \quad \text{for } u < rQ^{1/2}.$$

Thus,

$$\mathbb{E}d_{\mathscr{D}}(\mu^\star,\mu_n) \leqslant \frac{c_1}{\sqrt{n}}$$

for some positive constant $c_1 = O(qQ^{3/2}(D^{1/2}+q))$. Combining this inequality with (25) shows the first statement of the lemma.

We now turn to the more general situation (statement (ii)) where $\mu^\star$ is γ sub-Gaussian. According to inequality (24), the function g is $\frac{1}{n}$-Lipschitz with respect to the 1-norm on E^n. Therefore, by combining Kontorovich (2014, Theorem 1) and Vershynin (2018, Proposition 2.5.2), we have that for any $\eta\in(0,1)$, with probability at least $1-\eta$,

$$d_{\mathscr{D}}(\mu^\star,\mu_n) \leqslant \mathbb{E}d_{\mathscr{D}}(\mu^\star,\mu_n) + 8\gamma\sqrt{eD}\sqrt{\frac{\log(1/\eta)}{n}}. \tag{26}$$

As in the first part of the proof, we let

$$Y_\alpha = \mathbb{E}_{\mu^\star}D_\alpha - \frac{1}{n}\sum_{i=1}^n D_\alpha(X_i),$$

and recall that for any $(\alpha,\alpha')\in\Lambda^2$ and any $x\in E$,

$$|D_\alpha(x) - D_{\alpha'}(x)| \leqslant Q^{1/2}\big(q\|x\| + \frac{q(q-1)K_2}{2} + q\big)\|\alpha-\alpha'\|.$$

Since $\mu^\star$ is sub-Gaussian, we have (see, e.g., Jin et al., 2019, Lemma 1),

$$\ell = \int_E Q^{1/2}\big(q\|x\| + \frac{q(q-1)K_2}{2} + q\big)\mu^\star(\mathrm{d}x) < \infty.$$

Thus,

$$|Y_\alpha - Y_{\alpha'}| \leqslant \frac{1}{n}\|\alpha-\alpha'\|\ |\xi(n)|,$$

where

$$\xi_n = \sum_{i=1}^n Q^{1/2}\big(\ell + q\|X_i\| + \frac{q(q-1)K_2}{2} + q\big).$$

According to Jin et al. (2019, Lemma 1), the real-valued random variable ξ_n is sub-Gaussian. We obtain that, for some positive constant $c_2 = O(qQ^{3/2}(D^{1/2}+q))$,

$$\mathbb{E}d_{\mathscr{D}}(\mu^\star,\mu_n) \leqslant \frac{c_2}{\sqrt{n}},$$

and the conclusion follows by combining this inequality with (26).

A.16 Proof of Theorem 21

Let $\varepsilon > 0$ and $\eta \in (0,1)$. According to Theorem 7, there exists a discriminator $\mathscr{D}$ of the form (4) (i.e., a collection of neural networks) such that

$$T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) \leqslant \varepsilon.$$

We only prove statement (i) since both proofs are similar. In this case, according to Proposition 20, there exists a constant $c_1 > 0$ such that, with probability at least $1-\eta$,

$$d_{\mathscr{D}}(\mu^\star, \mu_n) \leqslant \frac{c_1}{\sqrt{n}} + B\sqrt{\frac{\log(1/\eta)}{2n}}.$$

Therefore, using inequality (16), we have, with probability at least $1-\eta$,

$$0 \leqslant \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} \leqslant 2\varepsilon + \frac{2c_1}{\sqrt{n}} + 2B\sqrt{\frac{\log(1/\eta)}{2n}}.$$

A.17 Proof of Proposition 23

Observe that, for $\theta \in \Theta$,

$$\begin{aligned}
0 &\leqslant d_{\mathscr{D}}(\mu^\star, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \\
&= d_{\mathscr{D}}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu_n, \mu_\theta) + d_{\mathscr{D}}(\mu_n, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta) \\
&\quad + \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \\
&\leqslant d_{\mathscr{D}}(\mu^\star, \mu_n) + d_{\mathscr{D}}(\mu_n, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta) + d_{\mathscr{D}}(\mu^\star, \mu_n) \\
&= 2d_{\mathscr{D}}(\mu^\star, \mu_n) + d_{\mathscr{D}}(\mu_n, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta),
\end{aligned}$$

where we used respectively the triangle inequality and

$$|\inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta)| \leqslant \sup_{\theta\in\Theta} |d_{\mathscr{D}}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu_n, \mu_\theta)| \leqslant d_{\mathscr{D}}(\mu^\star, \mu_n).$$

Thus, assuming that $T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) \leqslant \varepsilon$, we have

$$\begin{aligned}
0 &\leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\
&\leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta) + d_{\mathscr{D}}(\mu^\star, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \\
&\leqslant T_{\mathscr{D}}(\mathrm{Lip}_1, \mathscr{D}) + d_{\mathscr{D}}(\mu^\star, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \\
&\leqslant \varepsilon + 2d_{\mathscr{D}}(\mu^\star, \mu_n) + d_{\mathscr{D}}(\mu_n, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta). \qquad (27)
\end{aligned}$$

Let $\delta > 0$ and $\theta \in \mathscr{M}_{d_{\mathscr{D}}}(\mu_n, \delta/2)$, that is,

$$d_{\mathscr{D}}(\mu_n, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta) \leqslant \delta/2.$$

For $\eta \in (0,1)$, we know from the second statement of Proposition 20 that there exists $N \in \mathbb{N}^\star$ such that, for all $n \geqslant N$, $2d_{\mathscr{D}}(\mu^\star, \mu_n) \leqslant \delta/2$ with probability at least $1-\eta$. Therefore, we conclude from (27) that for $n \geqslant N$, with probability at least $1-\eta$,

$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \leqslant \varepsilon + \delta.$$