Geometrical Insights for Implicit Generative Modeling

Leon Bottou [a,b], Martin Arjovsky [b], David Lopez-Paz [a], Maxime Oquab [a,c]

[a] Facebook AI Research, New York, Paris
[b] New York University, New York
[c] Inria, Paris.

August 21, 2019

Abstract

Learning algorithms for implicit generative models can optimize a variety of criteria that measure how the data distribution differs from the implicit model distribution, including the Wasserstein distance, the Energy distance, and the Maximum Mean Discrepancy criterion. A careful look at the geometries induced by these distances on the space of probability measures reveals interesting differences. In particular, we can establish surprising approximate global convergence guarantees for the 1-Wasserstein distance, even when the parametric generator has a nonconvex parametrization.

1 Introduction

Instead of representing the model distribution with a parametric density function, implicit generative models directly describe how to draw samples of the model distribution by first drawing a sample z from a fixed random generator and mapping into the data space with a parametrized generator function $G_\theta(z)$. The reparametrization trick [13, 42], Variational Auto-Encoders (VAEs) [27], and Generative Adversarial Networks (GANs) [20] are recent instances of this approach.

Many of these authors motivate implicit modeling with the computational advantage that results from the ability of using the efficient back-propagation algorithm to update the generator parameters. In contrast, our work targets another, more fundamental, advantage of implicit modeling.

Although unsupervised learning is often formalized as estimating the data distribution [24, §14.1], the practical goal of the learning process rarely consists in recovering actual probabilities. Instead, the probability models are often structured in a manner that is interpretable as a physical or causal model of the data. This is often achieved by defining an interpretable density $p(y)$ for well chosen latent variables y and letting the appearance model $p(x|y)$ take the slack. This approach is well illustrated by the *inverse graphics* approach to computer vision [31, 30, 43]. Implicit modeling makes this much simpler:

- The structure of the generator function $G_\theta(z)$ could be directly interpreted as a set of equations describing a physical or causal model of the data [28].
- There is no need to deal with latent variables, since all the variables of interest are explicitly computed by the generator function.
- Implicit modeling can easily represent simple phenomena involving a small set of observed or inferred variables. The corresponding model distribution cannot be represented with a density function because it is supported by a low-dimensional manifold. But nothing prevents an implicit model from generating such samples.

Unfortunately, we cannot fully realize these benefits using the popular Maximum Likelihood Estimation (MLE) approach, which asymptotically amounts to minimizing the Kullback-Leibler (KL) divergence $D_{KL}(Q, P_\theta)$ between the data distribution Q and the model distribution P_θ,

$$D_{KL}(Q, P_\theta) = \int \log\left(\frac{q(x)}{p_\theta(x)}\right) q(x) d\mu(x) \tag{1}$$

where p_θ and q are the density functions of P_θ and Q with respect to a common measure μ. This criterion is particularly convenient because it enjoys favorable statistical properties [14] and because its optimization can be written as an expectation with respect to the data distribution,

$$\underset{\theta}{\operatorname{argmin}}\, D_{KL}(Q, P_\theta) \;=\; \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{x\sim Q}[-\log(p_\theta(x))] \;\approx\; \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(x_i)\,.$$

which is readily amenable to computationally attractive stochastic optimization procedures [10]. First, this expression is ill-defined when the model distribution cannot be represented by a density. Second, if the likelihood $p_\theta(x_i)$ of a single example x_i is zero, the dataset likelihood is also zero, and there is nothing to maximize. The typical remedy is to add a noise term to the model distribution. Virtually all generative models described in the classical machine learning literature include such a noise component whose purpose is not to model anything useful, but merely to make MLE work.

Instead of using ad-hoc noise terms to coerce MLE into optimizing a different similarity criterion between the data distribution and the model distribution, we could as well explicitly optimize a different criterion. Therefore it is crucial to understand how the selection of a particular criterion will influence the learning process and its final result.

Section 2 reviews known results establishing how many interesting distribution comparison criteria can be expressed in adversarial form, and are amenable to tractable optimization algorithms. Section 3 reviews the statistical properties of two interesting families of distribution distances, namely the family of the Wasserstein distances and the family containing the Energy Distances and the Maximum Mean Discrepancies. Although the Wasserstein distances have far worse statistical properties, experimental evidence shows that it can deliver better performances in meaningful applicative setups. Section 4 reviews essential concepts about geodesic geometry in metric spaces. Section 5 shows how different probability distances induce different geodesic geometries in the space of probability measures. Section 6 leverages these geodesic structures to define various flavors of convexity for parametric families of generative models, which can be used to prove that a simple gradient descent algorithm will either reach or approach the global minimum regardless of the traditional nonconvexity of the parametrization of the model family. In particular, when one uses implicit generative models, minimizing the Wasserstein distance with a gradient descent algorithm offers much better guarantees than minimizing the Energy distance.

2 The adversarial formulation

The adversarial training framework popularized by the Generative Adversarial Networks (GANs) [20] can be used to minimize a great variety of probability comparison criteria. Although some of these criteria can also be optimized using simpler algorithms, adversarial training provides a common template that we can use to compare the criteria themselves.

This section presents the adversarial training framework and reviews the main categories of probability comparison criteria it supports, namely Integral Probability Metrics (IPM) (Section 2.4), f-divergences (Section 2.5), Wasserstein distances (WD) (Section 2.6), and Energy Distances (ED) or Maximum Mean Discrepancy distances (MMD) (Section 2.7).

2.1 Setup

Although it is intuitively useful to consider that the sample space $\mathcal{X}$ is some convex subset of $\mathbb{R}^d$, it is also useful to spell out more precisely which properties are essential to the development. In the following, we assume that $\mathcal{X}$ is a *Polish metric space*, that is, a complete and separable space whose topology is defined by a distance function

$$d: \quad \left\{ \begin{array}{ccc} \mathcal{X} \times \mathcal{X} & \to & \mathbb{R}_+ \cup \{+\infty\} \\ (x, y) & \mapsto & d(x, y) \end{array} \right.$$

satisfying the properties of a metric distance:

$$\forall x, y, z \in \mathcal{X} \quad \left\{ \begin{array}{rll} (\mathit{0}) & d(x, x) = 0 & \text{(zero)} \\ (i) & x \neq y \Rightarrow d(x, y) > 0 & \text{(separation)} \\ (ii) & d(x, y) = d(y, x) & \text{(symmetry)} \\ (iii) & d(x, y) \leq d(x, z) + d(z, y) & \text{(triangular inequality)} \end{array} \right. \tag{2}$$

Let $\mathfrak{U}$ be the Borel σ-algebra generated by all the open sets of $\mathcal{X}$. We use the notation $\mathcal{P}_\mathcal{X}$ for the set of probability measures μ defined on $(\mathcal{X}, \mathfrak{U})$, and the notation $\mathcal{P}_\mathcal{X}^p \subset \mathcal{P}_\mathcal{X}$ for those satisfying $\mathbb{E}_{x,y\sim\mu}[d(x, y)^p] < \infty$. This condition is equivalent to $\mathbb{E}_{x\sim\mu}[d(x, x_0)^p] < \infty$ for an arbitrary origin x_0 when d is finite, symmetric, and satisfies the triangular inequality.

We are interested in criteria to compare elements of $\mathcal{P}_\mathcal{X}$,

$$D: \quad \left\{ \begin{array}{ccc} \mathcal{P}_\mathcal{X} \times \mathcal{P}_\mathcal{X} & \to & \mathbb{R}_+ \cup \{+\infty\} \\ (Q, P) & \mapsto & D(Q, P) \end{array} \right. .$$

Although it is desirable that D also satisfies the properties of a distance (2), this is not always possible. In this contribution, we strive to only reserve the word *distance* for criteria that satisfy the properties (2) of a metric distance. We use the word *pseudodistance*[1] when a nonnegative criterion fails to satisfy the separation property (2.i). We use the word *divergence* for criteria that are not symmetric (2.ii) or fail to satisfy the triangular inequality (2.iii).

We generally assume in this contribution that the distance d defined on $\mathcal{X}$ is finite. However we allow probability comparison criteria to be infinite. When the distributions Q, P do not belong to the domain for which a particular criterion D is defined, we take that $D(Q,P)=0$ if $Q=P$ and $D(Q,P)=+\infty$ otherwise.

2.2 Implicit modeling

We are particularly interested in model distributions P_θ that are supported by a low-dimensional manifold in a large ambient sample space (recall Section 1). Since such distributions do not typically have a density function, we cannot represent the model family $\mathcal{F}$ using a parametric density function. Following the example of Variational Auto-Encoders (VAE) [27] and Generative Adversarial Networks (GAN) [20], we represent the model distributions by defining how to produce samples.

Let z be a random variable with known distribution μ_z defined on a suitable probability space $\mathcal{Z}$ and let G_θ be a measurable function, called the *generator*, parametrized by $\theta \in \mathbb{R}^d$,

$$G_\theta : \quad z \in \mathcal{Z} \;\mapsto\; G_\theta(z) \in \mathcal{X} \,.$$

The random variable $G_\theta(Z) \in \mathcal{X}$ follows the *push-forward* distribution[2]

$$G_\theta(z)\#\mu_Z(z) : \quad A \in \mathfrak{U} \;\mapsto \mu_z(G_\theta^{-1}(A)) \,.$$

By varying the parameter θ of the generator G_θ, we can change this push-forward distribution and hopefully make it close to the data distribution Q according to the criterion of interest.

This *implicit modeling approach* is useful in two ways. First, unlike densities, it can represent distributions confined to a low-dimensional manifold. Second, the ability to easily generate samples is frequently more useful than knowing the numerical value of the density function (for example in image superresolution or semantic segmentation when considering the conditional distribution of the output image given the input image). In general, it is computationally difficult to generate samples given an arbitrary high-dimensional density [37].

Learning algorithms for implicit models must therefore be formulated in terms of two sampling oracles. The first oracle returns training examples, that is, samples from the data distribution Q. The second oracle returns generated examples, that is, samples from the model distribution $P_\theta = G_\theta\#\mu_Z$. This is particularly easy when the comparison criterion $D(Q, P_\theta)$ can be expressed in terms of expectations with respect to the distributions Q or P_θ.

[1] Although failing to satisfy the separation property (2.i) can have serious practical consequences, recall that a pseudodistance always becomes a full fledged distance on the quotient space $\mathcal{X}/\mathcal{R}$ where $\mathcal{R}$ denotes the equivalence relation $x\mathcal{R}y \Leftrightarrow d(x,y)=0$. All the theory applies as long as one never distinguishes two points separated by a zero distance.

[2] We use the notation $f\#\mu$ or $f(x)\#\mu(x)$ to denote the probability distribution obtained by applying function f or expression $f(x)$ to samples of the distribution μ.

2.3 Adversarial training

We are more specifically interested in distribution comparison criteria that can be expressed in the form

$$D(Q,P) = \sup_{(f_Q,f_P)\in\mathcal{Q}} \mathbb{E}_Q[f_Q(x)] - \mathbb{E}_P[f_P(x)] \ . \tag{3}$$

The set $\mathcal{Q}$ defines which pairs (f_Q, f_P) of real-valued *critic* functions defined on $\mathcal{X}$ are considered in this maximization. As discussed in the following subsections, different choices of $\mathcal{Q}$ lead to a broad variety of criteria. This formulation is a mild generalization of the Integral Probability Metrics (IPMs) [36] for which both functions f_Q and f_P are constrained to be equal (Section 2.4).

Finding the optimal generator parameter θ^* then amounts to minimizing a cost function $C(\theta)$ which itself is a supremum,

$$\min_\theta \left\{ C(\theta) \stackrel{\Delta}{=} \max_{(f_Q,f_P)\in\mathcal{Q}} \mathbb{E}_{x\sim Q}[f_Q(x)] - \mathbb{E}_{z\sim\mu_z}[f_P(G_\theta(z))] \right\} \ . \tag{4}$$

Although it is sometimes possible to reformulate this cost function in a manner that does not involve a supremum (Section 2.7), many algorithms can be derived from the following variant of the envelope theorem [35].

Theorem 2.1. *Let C be the cost function defined in* (4) *and let θ_0 be a specific value of the generator parameter. Under the following assumptions,*

a. there is $(f_Q^*, f_P^*) \in \mathcal{Q}$ such that $C(\theta_0) = \mathbb{E}_Q\big[f_Q^*(x)\big] - \mathbb{E}_{\mu_z}[f_P^*(G_{\theta_0}(z))]$,
b. the function C is differentiable in θ_0,
c. the functions $h_z = \theta \mapsto f_P^*(G_\theta(z))$ are μ_z-almost surely differentiable in θ_0,
d. and there exists an open neighborhood $\mathcal{V}$ of θ_0 and a μ_z-integrable function $D(z)$ such that $\forall\theta\in\mathcal{V}$, $|h_z(\theta) - h_z(\theta_0)| \leq D(z)\|\theta - \theta_0\|$,

we have the equality $\operatorname{grad}_\theta C(\theta_0) = -\mathbb{E}_{z\sim\mu_z}[\operatorname{grad}_\theta h_z(\theta_0)]$.

This result means that we can compute the gradient of $C(\theta_0)$ without taking into account the way f_P^* changes with θ_0. The most important assumption here is the differentiability of the cost C. Without this assumption, we can only assert that $-\mathbb{E}_{z\sim\mu_z}[\operatorname{grad}_\theta h_z(\theta_0)]$ belongs to the "local" subgradient

$$\partial^{\text{loc}} C(\theta_0) \stackrel{\Delta}{=} \left\{ g\in\mathbb{R}^d \ : \ \forall\theta\in\mathbb{R}^d \ \ C(\theta) \geq C(\theta_0) + \langle g, \theta-\theta_0\rangle + o(\|\theta-\theta_0\|) \right\} \ .$$

Proof Let $\lambda \in \mathbb{R}_+$ and $u \in \mathbb{R}^d$ be an arbitrary unit vector. From (3),

$$\begin{aligned} C(\theta_0+\lambda u) &\geq \mathbb{E}_{z\sim Q}\big[f_Q^*(x)\big] - \mathbb{E}_{z\sim\mu_z}[f_P^*(G_{\theta_0+\lambda u}(z))] \\ C(\theta_0+\lambda u) - C(\theta_0) &\geq -\mathbb{E}_{z\sim\mu_z}[h_z(\theta_0+\lambda u) - h_z(\theta_0)] \ . \end{aligned}$$

Dividing this last inequality by λ, taking its limit when $\lambda \to 0$, recalling that the dominated convergence theorem and assumption (d) allow us to take the limit inside the expectation operator, and rearranging the result gives

$$Au \geq 0 \ \text{ with } \ A: u\in\mathbb{R}^d \mapsto \langle u, \operatorname{grad}_\theta C(\theta_0) + \mathbb{E}_{z\sim\mu_z}[\operatorname{grad}_\theta h_z(\theta_0)]\rangle \, .$$

Writing the same for unit vector $-u$ yields inequality $-Au \geq 0$. Therefore $Au = 0$. ∎

Thanks to this result, we can compute an unbiased[3] stochastic estimate $\hat{g}(\theta_t)$ of the gradient $\operatorname{grad}_\theta C(\theta_t)$ by first solving the maximization problem in (4), and then using the back-propagation algorithm to compute the average gradient on a minibatch $z_1 \dots z_k$ sampled from μ,

$$\hat{g}(\theta_t) = -\frac{1}{k}\sum_{i=1}^{k} \operatorname{grad}_\theta f_P^*(G_\theta(z_i)) \ .$$

Such an unbiased estimate can then be used to perform a stochastic gradient descent update iteration on the generator parameter

$$\theta_{t+1} = \theta_t - \eta_t\, \hat{g}(\theta_t) \ .$$

Although this algorithmic idea can be made to work relatively reliably [3, 22], serious conceptual and practical issues remain:

Remark 2.2. In order to obtain an unbiased gradient estimate $\hat{g}(\theta_t)$, we need to solve the maximization problem in (4) for the true distributions rather than for a particular subset of examples. On the one hand, we can use the standard machine learning toolbox to avoid overfitting the maximization problem. On the other hand, this toolbox essentially works by restricting the family $\mathcal{Q}$ in ways that can change the meaning of the comparison criteria itself [5, 34].

Remark 2.3. In practice, solving the maximization problem (4) during each iteration of the stochastic gradient algorithm is computationally too costly. Instead, practical algorithms interleave two kinds of stochastic iterations: gradient ascent steps on (f_Q, f_P), and gradient descent steps on θ, with a much smaller effective stepsize. Such algorithms belong to the general class of stochastic algorithms with two time scales [9, 29]. Their convergence properties form a delicate topic, clearly beyond the purpose of this contribution.

2.4 Integral probability metrics

Integral probability metrics (IPMs) [36] have the form

$$D(Q,P) = \left|\ \sup_{f\in\mathcal{Q}} \mathbb{E}_Q[f(X)] - \mathbb{E}_P[f(X)]\ \right| \ .$$

Note that the surrounding absolute value can be eliminated by requiring that $\mathcal{Q}$ also contains the opposite of every one of its functions.

$$\begin{aligned} D(Q,P) = \sup_{f\in\mathcal{Q}} \mathbb{E}_Q[f(X)] - \mathbb{E}_P[f(X)] \\ \text{where } \mathcal{Q} \text{ satisfies } \quad \forall f\in\mathcal{Q}\,,\ -f\in\mathcal{Q}\ . \end{aligned} \tag{5}$$

[3] Stochastic gradient descent often relies on unbiased gradient estimates (for a more general condition, see [10, Assumption 4.3]). This is not a given: estimating the Wasserstein distance (14) and its gradients on small minibatches gives severely biased estimates [7]. This is in fact very obvious for minibatches of size one. Theorem 2.1 therefore provides an imperfect but useful alternative.

Therefore an IPM is a special case of (3) where the critic functions f_Q and f_P are constrained to be identical, and where $\mathcal{Q}$ is again constrained to contain the opposite of every critic function. Whereas expression (3) does not guarantee that $D(Q,P)$ is finite and is a distance, an IPM is always a pseudodistance.

Proposition 2.4. *Any integral probability metric D,* (5) *is a pseudodistance.*

Proof To establish the triangular inequality (2.iii), we can write, for all $Q,P,R \in \mathcal{P}_\mathcal{X}$,

$$\begin{aligned} D(Q,P)+D(P,R) \quad &= \sup_{f_1,\, f_2 \in \mathcal{Q}} \mathbb{E}_Q[f_1(X)] - \mathbb{E}_P[f_1(X)] + \mathbb{E}_P[f_2(X)] - \mathbb{E}_R[f_2(X)] \\ &\geq \sup_{f_1=f_2 \in \mathcal{Q}} \mathbb{E}_Q[f_1(X)] - \mathbb{E}_P[f_1(X)] + \mathbb{E}_P[f_2(X)] - \mathbb{E}_R[f_2(X)] \\ &= \sup_{f \in \mathcal{Q}} \mathbb{E}_Q[f(X)] - \mathbb{E}_R[f(X)] \;=\; D(Q,R)\,. \end{aligned}$$

The other properties of a pseudodistance are trivial consequences of (5). ■

The most fundamental IPM is the Total Variation (TV) distance.

$$D_{TV}(Q,P) \;\stackrel{\Delta}{=}\; \sup_{A\in\mathfrak{A}} |P(A)-Q(A)| \;=\; \sup_{f\in C(\mathcal{X},[0,1])} \mathbb{E}_Q[f(x)] - \mathbb{E}_P[f(x)]\;, \tag{6}$$

where $C(\mathcal{X},[0,1])$ is the space of continuous functions from $\mathcal{X}$ to $[0,1]$.

2.5 *f*-Divergences

Many classical criteria belong to the family of *f*-divergences

$$D_f(Q,P) \;\stackrel{\Delta}{=}\; \int f\left(\frac{q(x)}{p(x)}\right) p(x)\, d\mu(x) \tag{7}$$

where p and q are respectively the densities of P and Q relative to measure μ and where f is a continuous convex function defined on R_+^* such that $f(1)=0$.

Expression (7) trivially satisfies (2.0). It is always nonnegative because we can pick a subderivative $u\in\partial f(1)$ and use the inequality $f(t) \geq u(t-1)$. This also shows that the separation property (2.i) is satisfied when this inequality is strict for all $t\neq 1$.

Proposition 2.5 ([38, 39] (informal))**.** *Usually,*[4]

$$D_f(Q,P) \;=\; \sup_{\substack{g \text{ bounded, measurable} \\ g(\mathcal{X})\subset \mathrm{dom}(f^*)}} \mathbb{E}_Q[g(x)] - \mathbb{E}_P[f^*(g(x))]\;.$$

where f^* denotes the convex conjugate of f.

Table 1 provides examples of *f*-divergences and provides both the function *f* and the corresponding conjugate function f^* that appears in the variational formulation. In particular, as argued in [39], this analysis clarifies the probability comparison criteria associated with the early GAN variants [20].

[4] The statement holds when there is an $M{>}0$ such that $\mu\{x : |f(q(x)/p(x))|{>}M\}{=}0$ Restricting μ to exclude such subsets and taking the limit $M\to\infty$ may not work because $\limsup \neq \sup\lim$ in general. Yet, in practice, the result can be verified by elementary calculus for the usual choices of *f*, such as those shown in Table 1.

Table 1: Various f-divergences and the corresponding f and f^*.

	$f(\mathbf{t})$	$\mathbf{dom}(f^*)$	$f^*(\mathbf{u})$
Total variation (6)	$\frac{1}{2}\lvert t-1\rvert$	$[-\frac{1}{2},\frac{1}{2}]$	u
Kullback-Leibler (1)	$t\log(t)$	$\mathbb{R}$	$\exp(u-1)$
Reverse Kullback-Leibler	$-\log(t)$	$\mathbb{R}_-$	$-1-\log(-u)$
GAN's Jensen Shannon [20]	$t\log(t)-(t+1)\log(t+1)$	$\mathbb{R}_-$	$-\log(1-\exp(u))$

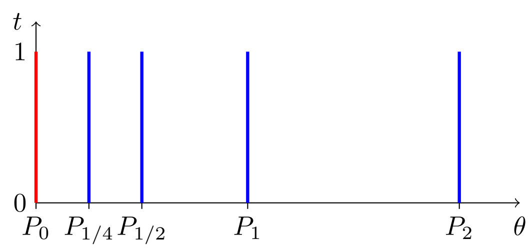


Figure 1: Let distribution P_θ be supported by the segment $\{(\theta,t)\ t\in[0,1]\}$ in $\mathbb{R}^2$. According to both the TV distance (6) and the f-divergences (7), the sequence of distributions $(P_{1/i})$ does not converge to P_0. However this sequence converges to P_0 according to either the Wasserstein distances (8) or the Energy distance (15).

Despite the elegance of this framework, these comparison criteria are not very attractive when the distributions are supported by low-dimensional manifolds that may not overlap. The following simple example shows how this can be a problem [3].

Example 2.6. Let U be the uniform distribution on the real segment $[0,1]$ and consider the distributions $P_\theta=(\theta,x)_\#U(x)$ defined on $\mathbb{R}^2$. Because P_0 and P_θ have disjoint support for $\theta\neq 0$, neither the total variation distance $D_{TV}(P_0,P_\theta)$ nor the f-divergence $D_f(P_0,P_\theta)$ depend on the exact value of θ. Therefore, according to the topologies induced by these criteria on $\mathcal{P}_\mathcal{X}$, the sequence of distributions $(P_{1/i})$ does not converge to P_0 (Figure 1).

The fundamental problem here is that neither the total variation distance (6) nor the f-divergences (7) depend on the distance $d(x,y)$ defined on the sample space $\mathcal{X}$. The minimization of such a criterion appears more effective for adjusting the probability values than for matching the distribution supports.

2.6 Wasserstein distance

For any $p\geq 1$, the p-Wasserstein distance (WD) is the p-th root of

$$\forall Q,P\in\mathcal{P}_\mathcal{X}^p \qquad W_p(Q,P)^p \;\stackrel{\Delta}{=}\; \inf_{\pi\in\Pi(Q,P)} \mathbb{E}_{(x,y)\sim\pi}[\,d(x,y)^p\,]\;, \tag{8}$$

where $\Pi(Q,P)$ represents the set of all measures π defined on $\mathcal{X}\times\mathcal{X}$ with marginals $x_\#\pi(x,y)$ and $y_\#\pi(x,y)$ respectively equal to Q and P. Intuitively, $d(x,y)^p$ represents the cost of transporting a grain of probability from point x to point y, and the joint distributions $\pi \in \Pi(Q,P)$ represent transport plans.

Since $d(x,y) \leq d(x,x_0) + d(x_0,y) \leq 2\max\{d(x,x_0), d(x_0,y)\}$,

$$\forall\, Q,P \in \mathcal{P}_{\mathcal{X}}^p \qquad W_p(Q,P)^p \leq \mathbb{E}_{\substack{x\sim Q\\ y\sim P}}[d(x,y)^p] < \infty\;. \tag{9}$$

Example 2.7. Let P_θ be defined as in Example 2.6. Since it is easy to see that the optimal transport plan from P_0 to P_θ is $\pi^* = ((0,t),(\theta,t))\#U(t)$, the Wassertein distance $W_p(P_0,P_\theta) = |\theta|$ converges to zero when θ tends to zero. Therefore, according to the topology induced by the Wasserstein distance on $\mathcal{P}_{\mathcal{X}}$, the sequence of distributions $(P_{1/i})$ converges to P_0 (Figure 1).

Thanks to the Kantorovich duality theory, the Wasserstein distance is easily expressed in the variational form (3). We summarize below the essential results useful for this work and we direct the reader to [55, Chapters 4 and 5] for a full exposition.

Theorem 2.8 ([55, Theorem 4.1])**.** *Let $\mathcal{X},\mathcal{Y}$ be two Polish metric spaces and $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+\cup\{+\infty\}$ be a nonnegative continuous cost function. Let $\Pi(Q,P)$ be the set of probablity measures on $\mathcal{X}\times\mathcal{Y}$ with marginals $Q\in\mathcal{P}_{\mathcal{X}}$ and $P\in\mathcal{P}_{\mathcal{Y}}$. There is a $\pi^*\in\Pi(Q,P)$ that minimizes $\mathbb{E}_{(x,y)\sim\pi}[c(x,y)]$ over all $\pi\in\Pi(Q,P)$.*

Definition 2.9. *Let $\mathcal{X},\mathcal{Y}$ be two Polish metric spaces and $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+\cup\{+\infty\}$ be a nonnegative continuous cost function. The pair of functions $f:\mathcal{X}\to\mathbb{R}$ and $g:\mathcal{Y}\to\mathbb{R}$ is c-conjugate when*

$$\forall x\in\mathcal{X}\;\, f(x) = \inf_{y\in\mathcal{Y}} g(y) + c(x,y) \quad \textit{and} \quad \forall y\in\mathcal{Y}\;\, g(y) = \sup_{x\in\mathcal{X}} f(x) - c(x,y)\;. \tag{10}$$

Theorem 2.10 (Kantorovich duality [55, Theorem 5.10])**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish metric spaces and $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+\cup\{+\infty\}$ be a nonnegative continuous cost function. For all $Q\in\mathcal{P}_{\mathcal{X}}$ and $P\in\mathcal{P}_{\mathcal{Y}}$, let $\Pi(Q,P)$ be the set of probability distributions defined on $\mathcal{X}\times\mathcal{Y}$ with marginal distributions Q and P. Let $\mathcal{Q}_c$ be the set of all pairs (f_Q,f_P) of respectively Q and P-integrable functions satisfying the property $\forall x\in\mathcal{X}\; y\in\mathcal{Y}$, $f_Q(x) - f_P(y) \leq c(x,y)$.*

$i)$ We have the duality

$$\min_{\pi\in\Pi(Q,P)} \mathbb{E}_{(x,y)\sim\pi}[c(x,y)] \quad = \tag{11}$$

$$\sup_{(f_Q,f_P)\in\mathcal{Q}_c} \mathbb{E}_{x\sim Q}[f_Q(x)] - \mathbb{E}_{y\sim P}[f_P(y)]\;. \tag{12}$$

$ii)$ Further assuming that $\mathbb{E}_{x\sim Q\, y\sim P}[c(x,y)] < \infty$,

- *a) Both* (11) *and* (12) *have solutions with finite cost.*
- *b) The solution (f_Q^*, fp^*) of* (12) *is a c-conjugate pair.*

Corollary 2.11 ([55, Particular case 5.16]). *Under the same conditions as Theorem 2.10.ii, when $\mathcal{X} = \mathcal{Y}$ and when the cost function c is a distance, that is, satisfies* (2)*, the dual optimization problem* (12) *can be rewritten as*

$$\max_{f \in \mathrm{Lip1}} \mathbb{E}_Q[f(x)] - \mathbb{E}_P[f(x)] \ ,$$

where Lip1 *is the set of real-valued* 1*-Lipschitz continuous functions on* $\mathcal{X}$.

Thanks to Theorem 2.10, we can write the p-th power of the p-Wasserstein distance in variational form

$$\forall Q, P \in \mathcal{P}^p_{\mathcal{X}} \qquad W_p(Q,P)^p = \sup_{(f_Q, f_P) \in \mathcal{Q}_c} \mathbb{E}_Q[f_Q(x)] - \mathbb{E}_P[f_P(x)] \ , \tag{13}$$

where $\mathcal{Q}_c$ is defined as in Theorem 2.10 for the cost $c(x,y) = d(x,y)^p$. Thanks to Corollary 2.11, we can also obtain a simplified expression in IPM form for the 1-Wasserstein distance.

$$\forall Q, P \in \mathcal{P}^1_{\mathcal{X}} \qquad W_1(Q,P) = \sup_{f \in \mathrm{Lip1}} \mathbb{E}_Q[f(x)] - \mathbb{E}_P[f(x)] \ . \tag{14}$$

Let us conclude this presentation of the Wassertein distance by mentioning that the definition (8) immediately implies several distance properties: zero when both distributions are equal (2.o), strictly positive when they are different (2.i), and symmetric (2.ii). Property 2.4 gives the triangular inequality (2.iii) for the case $p = 1$. In the general case, the triangular inequality can also be established using the Minkowsky inequality [55, Chapter 6].

2.7 Energy Distance and Maximum Mean Discrepancy

The Energy Distance (ED) [53] between the probability distributions Q and P defined on the Euclidean space $\mathbb{R}^d$ is the square root[5] of

$$\mathcal{E}(Q,P)^2 \ \stackrel{\Delta}{=}\ 2\mathbb{E}_{\substack{x\sim Q\\ y\sim P}}[\|x-y\|] - \mathbb{E}_{\substack{x\sim Q\\ x'\sim Q}}[\|x-x'\|] - \mathbb{E}_{\substack{y\sim P\\ y'\sim P}}[\|y-y'\|] \ , \tag{15}$$

where, as usual, $\|\cdot\|$ denotes the Euclidean distance.

Let $\hat{q}$ and $\hat{p}$ represent the characteristic functions of the distribution Q and P respectively. Thanks to a neat Fourier transform argument [53, 52],

$$\mathcal{E}(Q,P)^2 = \frac{1}{c_d} \int_{\mathbb{R}^d} \frac{|\hat{q}(t) - \hat{p}(t)|^2}{\|t\|^{d+1}} dt \quad \text{with} \ \ c_d = \frac{\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})} \ . \tag{16}$$

Since there is a one-to-one mapping between distributions and characteristic functions, this relation establishes an isomorphism between the space of probability distributions equipped with the ED distance and the space of the characteristic functions equipped with

[5] We take the square root because this is the quantity that behaves like a distance.

the weighted L_2 norm given in the right-hand side of (16). As a consequence, $\mathcal{E}(Q,P)$ satisfies the properties (2) of a distance.

Since the squared ED is expressed with a simple combination of expectations, it is easy to design a stochastic minimization algorithm that relies only on two oracles producing samples from each distribution [11, 7]. This makes the energy distance a computationally attractive criterion for training the implicit models discussed in Section 2.2.

Generalized ED It is therefore natural to ask whether we can meaningfully generalize (15) by replacing the Euclidean distance $\|x-y\|$ with a symmetric function $d(x,y)$.

$$\mathcal{E}_d(Q,P)^2 \;=\; 2\mathbb{E}_{\substack{x\sim Q\\ y\sim P}}[d(x,y)] - \mathbb{E}_{\substack{x\sim Q\\ x'\sim Q}}[d(x,x')] - \mathbb{E}_{\substack{y\sim P\\ y'\sim P}}[d(y,y')]\ . \tag{17}$$

The right-hand side of this expression is well defined when $Q,P\in\mathcal{P}^1_{\mathcal{X}}$. It is obviously symmetric (2.ii) and trivially zero (2.o) when both distributions are equal. The first part of the following theorem gives the necessary and sufficient conditions on $d(x,y)$ to ensure that the right-hand side of (17) is nonnegative and therefore can be the square of $\mathcal{E}_d(Q,P)\in\mathbb{R}_+$. We shall see later that the triangular inequality (2.iii) comes for free with this condition (Corollary 2.19). The second part of the theorem gives the necessary and sufficient condition for satisfying the separation property (2.i).

Theorem 2.12 ([57])**.** *The right-hand side of definition* (17) *is:*

$i)$ *nonnegative for all P,Q in $\mathcal{P}^1_{\mathcal{X}}$ if and only if the symmetric function d is a* negative definite kernel*, that is,*

$$\forall n\in\mathbb{N}\ \ \forall x_1\dots x_n\in\mathcal{X}\ \ \forall c_1\dots c_n\in\mathbb{R}$$
$$\sum_{i=1}^n c_i=0 \;\Longrightarrow\; \sum_{i=1}^n\sum_{j=1}^n d(x_i,x_j)c_ic_j\le 0\ . \tag{18}$$

$ii)$ *strictly positive for all $P\neq Q$ in $\mathcal{P}^1_{\mathcal{X}}$ if and only if the function d is a* strongly negative definite kernel*, that is, a negative definite kernel such that, for any probability measure $\mu\in\mathcal{P}^1_{\mathcal{X}}$ and any μ-integrable real-valued function h such that $\mathbb{E}_\mu[h(x)]=0$,*

$$\mathbb{E}_{\substack{x\sim\mu\\ y\sim\mu}}[\,d(x,y)h(x)h(y)\,]=0 \;\Longrightarrow\; h(x)=0\ \ \mu\text{-almost everywhere.}$$

Remark 2.13. The definition of a strongly negative kernel is best explained by considering how its meaning would change if we were only considering probability measures μ with finite support $\{x_1\dots x_n\}$. This amounts to requiring that (18) is an equality only if all the c_is are zero. However, this weaker property is not sufficient to ensure that the separation property (2.i) holds.

Remark 2.14. The relation (16) therefore means that the Euclidean distance on $\mathbb{R}^d$ is a strongly negative definite kernel. In fact, it can be shown that $d(x,y)=\|x-y\|^\beta$ is a strongly negative definite kernel for $0 < \beta < 2$ [52]. When $\beta = 2$, it is easy to see that $\mathcal{E}_d(Q, P)$ is simply the distance between the distribution means and therefore cannot satisfy the separation property $(2.i)$.

Proof of Theorem 2.12 Let $E(Q, P)$ be the right-hand side of (17) and let $S(\mu, h)$ be the quantity $\mathbb{E}_{x,y\sim\mu}[\, d(x, y)h(x)h(y)\,]$ that appears in clause (ii). Observe:

a) Let $Q, P \in \mathcal{P}^1_{\mathcal{X}}$ have respective density functions $q(x)$ and $p(x)$ with respect to measure $\mu = (Q + P)/2$. Function $h = q - p$ then satisfies $\mathbb{E}_\mu[h] = 0$, and

$$E(Q,P) = \mathbb{E}_{\substack{x\sim\mu \\ y\sim\mu}}\big[\, \big(q(x)p(y) + q(y)p(x) - q(x)q(y) - p(x)p(y)\big)\, d(x,y)\, \big] = -S(\mu,h)\ .$$

b) With $\mu \in \mathcal{P}^1_{\mathcal{X}}$, any h such that $\mu\{h{=}0\} < 1$ (ie., non-μ-almost-surely-zero) and $\mathbb{E}_\mu[h] = 0$ can be written as a difference of two nonnegative functions $h = \tilde{q} - \tilde{p}$ such that $\mathbb{E}_\mu[\tilde{q}] = \mathbb{E}_\mu[\tilde{p}] = \rho^{-1} > 0$. Then, $Q = \rho\, \tilde{q}\, \mu$ and $P = \rho\, \tilde{p}\, \mu$ belong to $\mathcal{P}^1_{\mathcal{X}}$, and

$$E(Q,P) = -\rho\, S(\mu,h)\ .$$

We can then prove the theorem:

$i)$ From these observations, if $E(Q, P) \geq 0$ for all P, Q, then $S(\mu, h) \leq 0$ for all μ and h such that $\mathbb{E}_\mu[h(x)] = 0$, implying (18). Conversely, assume there are $Q, P \in \mathcal{P}^1_{\mathcal{X}}$ such that $E(Q, P) < 0$. Using the weak law of large numbers [26] (see also Theorem 3.3 later in this document,) we can find finite support distributions Q_n, P_n such that $E(Q_n, P_n) < 0$. Proceeding as in observation (a) then contradicts (18) because $\mu = (Q_n + P_n)/2$ has also finite support.

$ii)$ By contraposition, suppose there is μ and h such that $\mu\{h{=}0\} < 1$, $\mathbb{E}_\mu[h(x)] = 0$, and $S(\mu, h) = 0$. Observation (b) gives $P \neq Q$ such that $E(Q, P) = 0$. Conversely, suppose $E(Q, P) = 0$. Observation (a) gives μ and $h = q - p$ such that $S(\mu, h) = 0$. Since h must be zero, $Q = P$. ■

Requiring that d be a negative definite kernel is a quite strong assumption. For instance, a classical result by Schoenberg [45] establishes that a squared distance is a negative definite kernel if and only if the whole metric space induced by this distance is isometric to a subset of a Hilbert space and therefore has a Euclidean geometry:

Theorem 2.15 (Schoenberg, [45])**.** *The metric space $(\mathcal{X}, d)$ is isometric to a subset of a Hilbert space if and only if d^2 is a negative definite kernel.*

Requiring d to be negative definite (not necessarily a squared distance anymore) has a similar impact on the geometry of the space $\mathcal{P}^1_{\mathcal{X}}$ equipped with the Energy Distance (Theorem 2.17). Let x_0 be an arbitrary origin point and define the symmetric *triangular gap* kernel K_d as

$$K_d(x,y) \ \stackrel{\Delta}{=}\ \tfrac{1}{2}\left(d(x,x_0) + d(y,x_0) - d(x,y)\right)\ . \tag{19}$$

Proposition 2.16. *The function d is a negative definite kernel if and only if K_d is a positive definite kernel, that is,*

$$\forall n \in \mathbb{N}\ \ \forall x_1 \ldots x_n \in \mathcal{X}\ \ \forall c_1 \ldots c_n \in \mathbb{R}\ \ \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_d(x_i, x_j) \geq 0\ .$$

Proof The proposition directly results from the identity

$$2\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_d(x_i,x_j) = -\sum_{i=0}^{n}\sum_{j=0}^{n} c_i c_j d(x_i,x_j)\;,$$

where x_0 is the chosen origin point and $c_0 = -\sum_{i=1}^{n} c_i$. ∎

Positive definite kernels in the machine learning literature have been extensively studied in the context of the so-called *kernel trick* [46]. In particular, it is well known that the theory of the Reproducing Kernel Hilbert Spaces (RKHS) [4, 1] establishes that there is a unique Hilbert space $\mathcal{H}$, called the RKHS, that contains all the functions

$$\Phi_x : y \in \mathcal{X} \mapsto K_d(x,y)$$

and satisfies the *reproducing property*

$$\forall x \in \mathcal{X}\;\; \forall f \in \mathcal{H} \quad \langle f, \Phi_x\rangle = f(x)\;. \tag{20}$$

We can then relate $\mathcal{E}_d(Q,P)$ to the RKHS norm.

Theorem 2.17 ([47] [40, Chapter 21])**.** *Let d be a negative definite kernel and let $\mathcal{H}$ be the RKHS associated with the corresponding positive definite triangular gap kernel* (19)*. We have then*

$$\forall Q,P \in \mathcal{P}^1_{\mathcal{X}} \quad \mathcal{E}_d(Q,P) \;=\; \|\, \mathbb{E}_{x\sim Q}[\Phi_x] - \mathbb{E}_{y\sim P}[\Phi_y]\;\|_{\mathcal{H}}\;.$$

Proof We can write directly

$$\begin{aligned}
\mathcal{E}_d(Q,P)^2 &= \mathop{\mathbb{E}}_{\substack{x,x'\sim Q\\ y,y'\sim P}}\big[d(x,y)+d(x',y')-d(x,x')-d(y,y')\big]\\
&= \mathop{\mathbb{E}}_{\substack{x,x'\sim Q\\ y,y'\sim P}}\big[K_d(x,x')+K_d(y,y')-K_d(x,y)-K_d(x',y')\big]\\
&= \langle \mathbb{E}_Q[\Phi_x],\mathbb{E}_Q[\Phi_x]\rangle + \langle \mathbb{E}_P[\Phi_y],\mathbb{E}_P[\Phi_y]\rangle - 2\,\langle \mathbb{E}_Q[\Phi_x],\mathbb{E}_P[\Phi_y]\rangle\\
&= \|\mathbb{E}_{x\sim Q}[\Phi_x] - \mathbb{E}_{y\sim P}[\Phi_y]\,\|^2_{\mathcal{H}}\;,
\end{aligned}$$

where the first equality results from (19) and where the second equality results from the identities $\langle \Phi_x,\Phi_y\rangle = K_d(x,y)$ and $\mathbb{E}_{x,y}[\langle\Phi_x,\Phi_y\rangle] = \langle \mathbb{E}_x[\Phi_x], \mathbb{E}_y[\Phi_y]\rangle$. ∎

Remark 2.18. In the context of this theorem, the relation (16) is simply an analytic expression of the RKHS norm associated with the triangular gap kernel of the Euclidean distance.

Corollary 2.19. *If d is a negative definite kernel, then $\mathcal{E}_d$ is a pseudodistance, that is, it satisfies all the properties* (2) *of a distance except maybe the separation property (2.i).*

Corollary 2.20. *The following three conditions are then equivalent:*

i) $\mathcal{E}_d$ satisfies all the properties (2) *of a distance.*

ii) d is a strongly negative definite kernel.

iii) the map $P \in \mathcal{P}^1_{\mathcal{X}} \mapsto \mathbb{E}_P[\Phi_x] \in \mathcal{H}$ is injective (characteristic kernel [21].)

Maximum Mean Discrepancy Following [21], we can then write $\mathcal{E}_d$ as an IPM:

$$\begin{aligned}
\mathcal{E}_d(Q,P) &= \|\mathbb{E}_Q[\Phi_x] - \mathbb{E}_P[\Phi_x]\,\|_{\mathcal{H}} \\
&= \sup_{\|f\|_{\mathcal{H}}\le 1} \langle f, \mathbb{E}_P[\Phi_x] - \mathbb{E}_Q[\Phi_x]\rangle \\
&= \sup_{\|f\|_{\mathcal{H}}\le 1} \mathbb{E}_P[\langle f,\Phi_x\rangle] - \mathbb{E}_Q[\langle f,\Phi_x\rangle] \\
&= \sup_{\|f\|_{\mathcal{H}}\le 1} \mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]\ . \qquad (21)
\end{aligned}$$

This last expression (21) is also called the Maximum Mean Discrepancy (MMD) associated with the positive definite kernel K_d [21]. Conversely, for any positive definite kernel K, the reader will easily prove that the symmetric function

$$d_K(x,y) = \|\Phi_x - \Phi_y\|^2_{\mathcal{H}} = K(x,x) + K(y,y) - 2K(x,y)\ ,$$

is a negative definite kernel, that $d_{K_d} = d$, and that

$$\|\,\mathbb{E}_Q[\Phi_x] - \mathbb{E}_P[\Phi_x]\,\|^2_{\mathcal{H}} \ = \ \mathcal{E}_{d_K}(Q,P)^2\ . \qquad (22)$$

Therefore the ED and MMD formulations are essentially equivalent [47]. Note however that the negative definite kernel d_K defined above may not satisfy the triangular inequality (its square root does.)

Remark 2.21. Because this equivalence was not immediately recognized, many important concepts have been rediscovered with subtle technical variations. For instance, the notion of characteristic kernel [21] depends subtly on the chosen domain for the map $P \mapsto \mathbb{E}_P[\Phi_x]$ that we want injective. Corollary 2.20 gives a simple necessary and sufficient condition when this domain is $\mathcal{P}^1_{\mathcal{X}}$ (with respect to the distance d). Choosing a different domain leads to complications [51].

3 Energy Distance vs. 1-Wasserstein Distance

The dual formulation of the 1-Wasserstein (14) and the MMD formulation of the Energy Distance (21) only differ by the use of a different family of critic functions: for all $Q, P \in \mathcal{P}^1_{\mathcal{X}}$,

$$\begin{aligned}
W_1(Q,P) &= \sup_{f\in\mathrm{Lip1}} \mathbb{E}_Q[f(x)] - \mathbb{E}_P[f(x)]\ , \\
\mathcal{E}_d(Q,P) &= \sup_{\|f\|_{\mathcal{H}}\le 1} \mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]\ .
\end{aligned}$$

At first sight, requiring that the functions f are 1-Lipschitz or are contained in the RKHS unit ball seem to be two slightly different ways to enforce a smoothness constraint. Nevertheless, a closer comparison reveals very important differences.

3.1 Three quantitative properties

Although both the WD [55, Theorem 6.9] and the ED/MMD [49, Theorem 3.2] metrize the weak convergence topology, they may be quantitatively very different and therefore hard to compare in practical situations. The following upper bound provides a clarification.

Proposition 3.1. *Let $\mathcal{X}$ be equipped with a distance d that is also a negative definite kernel. Let the 1-Wasserstein distance W_1 and the Energy Distance $\mathcal{E}_d$ be defined as in (8) and (17).*

$$\mathcal{E}_d(Q,P)^2 \leq 2W_1(Q,P)\ .$$

This inequality is tight. It is indeed easy to see that it becomes an equality when both P and Q are Dirac distributions.

The proof relies on an elementary geometrical lemma:

Lemma 3.2. *Let A, B, C, D be four points in $\mathcal{X}$ forming a quadrilateral. The perimeter length $d(A,B)+d(B,C)+d(C,D)+d(D,A)$ is longer than the diagonal lenghts $d(A,C)+d(B,D)$.*

Proof of the lemma Summing the following triangular inequalities yields the result.

$$\begin{aligned} d(A,C) &\leq d(A,B)+d(B,C) & \qquad d(A,C) &\leq d(C,D)+d(D,A) \\ d(B,D) &\leq d(B,C)+d(C,D) & \qquad d(B,D) &\leq d(D,A)+d(A,B) \end{aligned} \qquad \blacksquare$$

Proof of proposition 3.1 Let (x,y) and (x',y') be two independent samples of the optimal transport plan π with marginals Q and P. Since they are independent,

$$2\,\mathbb{E}_{\substack{x\sim Q\\ y\sim P}}[d(x,y)] = \mathbb{E}_{\substack{(x,y)\sim\pi\\ (x',y')\sim\pi}}\left[d(x,y')+d(x',y)\right]\ .$$

Applying the lemma and rearranging

$$\begin{aligned} 2\mathbb{E}_{\substack{x\sim Q\\ y\sim P}}[d(x,y)] &\leq \mathbb{E}_{\substack{(x,y)\sim\pi\\ (x',y')\sim\pi}}\left[d(x,y)+d(y,y')+d(y',x')+d(x',x)\right] \\ &= W_1(Q,P) + \mathbb{E}_{\substack{y\sim P\\ y'\sim P}}\left[d(y,y')\right] + W_1(Q,P) + \mathbb{E}_{\substack{x\sim Q\\ x'\sim Q}}\left[d(x,x')\right]\ . \end{aligned}$$

Moving the remaining expectations to the left-hand side gives the result. ∎

In contrast, the following results not only show that $\mathcal{E}_d$ can be very significantly smaller than the 1-Wasserstein distance, but also show that this happens in the particularly important situation where one approximates a distribution with a finite sample.

Theorem 3.3. *Let $Q, P \in \mathcal{P}^1_{\mathcal{X}}$ be two probability distributions on $\mathcal{X}$. Let $x_1 \dots x_n$ be n independent Q-distributed random variables, and let $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the corresponding empirical probability distribution. Let $\mathcal{E}_d$ be defined as in (17) with a kernel satisfying $d(x,x)=0$ for all x in $\mathcal{X}$. Then,*

$$\mathbb{E}_{x_1\dots x_n\sim Q}\left[\mathcal{E}_d(Q_n,P)^2\right] = \mathcal{E}_d(Q,P)^2 + \tfrac{1}{n}\,\mathbb{E}_{x,x'\sim Q}[d(x,x')]\ ,$$

and

$$\mathbb{E}_{x_1\dots x_n\sim Q}\left[\mathcal{E}_d(Q_n,Q)^2\right] = \tfrac{1}{n}\,\mathbb{E}_{x,x'\sim Q}[d(x,x')] = \mathcal{O}(n^{-1})\ .$$

Therefore the effect of replacing Q by its empirical approximation disappears quickly, like $\mathcal{O}(1/n)$, when n grows. This result is not very surprising when one notices that $\mathcal{E}_d(Q_n, P)$ is a V-statistic [56, 48]. However it gives a precise equality with a particularly direct proof.

Proof Using the following equalities in the definition (17) gives the first result.

$$\mathbb{E}_{x_1 \ldots x_n \sim Q}\left[\mathbb{E}_{\substack{x\sim Q_n \\ y\sim P}}[d(x,y)] \right] = \mathbb{E}_{x_1 \ldots x_n \sim Q}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{y\sim P}[d(x_i,y)] \right]$$
$$= \frac{1}{n}\sum_i \mathbb{E}_{\substack{x\sim Q \\ y\sim P}}[d(x,y)] = \mathbb{E}_{\substack{x\sim Q \\ y\sim P}}[d(x,y)] \ .$$
$$\mathbb{E}_{x_1 \ldots x_n \sim Q}\left[\mathbb{E}_{\substack{x\sim Q_n \\ x'\sim Q_n}}[d(x,x')] \right] = \mathbb{E}_{x_1 \ldots x_n \sim Q}\left[\frac{1}{n^2}\sum_{i\neq j} d(x_i,x_j) \right]$$
$$= \frac{1}{n^2}\sum_{i\neq j} \mathbb{E}_{\substack{x\sim Q \\ y\sim Q}}[d(x,y)] = \left(1-\frac{1}{n}\right) \mathbb{E}_{\substack{x\sim Q \\ y\sim Q}}[d(x,y)] \ .$$

Taking $Q = P$ then gives the second result. ∎

Comparable results for the 1-Wasserstein distance describe a convergence speed that quickly becomes considerably slower with the dimension $d > 2$ of the sample space $\mathcal{X}$ [50, 16, 18].

Theorem 3.4 ([18])**.** *Let $\mathcal{X}$ be $\mathbb{R}^d$, $d > 2$, equipped with the usual Euclidean distance. Let $Q \in \mathcal{P}^2_{R^d}$ and let Q_n be defined as in Theorem 3.3. Then,*

$$\mathbb{E}_{x_1 \ldots x_n \sim Q}[\, W_1(Q_n, Q)\,] = \mathcal{O}(n^{-1/d}) \ .$$

The following example, inspired by [5], illustrates this slow rate and its consequences.

Example 3.5. Let Q be a uniform distribution supported by the unit sphere in $\mathbb{R}^d$ equipped with the Euclidean distance. Let $x_1 \ldots x_n$ be n points sampled independently from this distribution and let Q_n be the corresponding empirical distribution. Let x be an additional point sampled from Q. It is well known[6] that $\min_i \|x - x_i\|$ remains arbitrarily close to $\sqrt{2}$, say, greater than 1.2, with arbitrarily high probability when $d \gg \log(n)$. Therefore,

$$W_1(Q_n, Q) \geq 1.2 \quad \text{when } n \ll \exp(d).$$

In contrast, observe

$$W_1(Q, \delta_0) = W_1(Q_n, \delta_0) = 1 \ .$$

In other words, as long as $n \ll \exp(d)$, a Dirac distribution in zero is closer to the empirical distribution than the actual distribution [5].

Theorem 3.3 and Example 3.5 therefore show that $\mathcal{E}_d(Q_n, Q)$ can be much smaller than $W_1(Q_n, Q)$. They also reveal that the statistical properties of the 1-Wasserstein distance are very discouraging. Since the argument of Example 3.5 naturally extends to the p-Wasserstein distance for all $p \geq 1$, the problem seems shared by all Wasserstein distances.

[6] The curious reader can pick an expression of $F_d(t) = P\{\|x - x_i\| < t\}$ in [23], then derive an asymptotic bound for $P\{\min_i \|x - x_i\| < t\} = 1 - (1 - F_d(t))^n$.

Remark 3.6. In the more realistic case where the 1-Lipschitz critic is constrained to belong to a parametric family with sufficient regularity, the bound of theorem 3.4 can be improved to $\mathcal{O}(\sqrt{\log(n)/n})$ with a potentially large constant [5]. On the other hand, constraining the critic too severely might prevent it from distinguishing distributions that differ in meaningful ways.

3.2 WD and ED/MMD in practice

Why should we consider the Wasserstein Distance when the Energy Distance and Maximum Mean Discrepancy offer better statistical properties (Section 3.1) and more direct learning algorithms [17, 33, 11] ?

The most impressive achievement associated with the implicit modeling approach certainly is the generation of photo-realistic random images that resemble the images provided as training data [15, 41, 25]. In apparent contradiction with the statistical results of the previous section, and with a couple notable exceptions discussed later in this section, the visual quality of the images generated using models trained by directly minimizing the MMD [17] usually lags behind those obtained with the WD [3, 22, 25] and with the original Generative Adversarial Network formulation[7] [41].

Before discussing the two exceptions, it is worth recalling that the visual quality of the generated images is a peculiar way to benchmark generative models. This is an incomplete criterion because it does not ensure that the model generates images that cover all the space covered by the training data. This is an interesting criterion because common statistical metrics, such as estimates of the negative log-likelihood, are generally unable to indicate which models generate the better-looking images [54]. This is a finicky criterion because, despite efforts to quantify visual quality with well-defined scores [44], the evaluation of the image quality fundamentally remains a beauty contest. Figure 2 nevertheless shows a clear difference.

A few authors report good image generation results by using the ED/MMD criterion in a manner that substantially changes its properties:

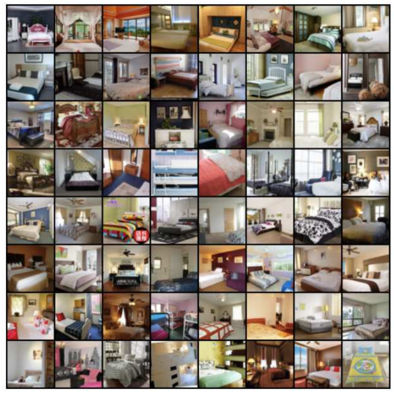
- The AE+GMMN approach [33] improves the pure MMD approach by training an implicit model that does not directly generate images but targets the compact representation computed by a pretrained auto-encoder network. This changes a high-dimensional image generation problem into a comparatively low-dimensional code generation problem with a good notion of distance. There is independent evidence that low-dimensional implicit models work relatively well with ED/MMD [11].

- The CramérGAN approach [7] minimizes the Energy Distance[8] computed on the representations produced by an adversarially trained 1-Lipschitz continuous *transformation layer* $T_\phi(x)$. The resulting optimization problem

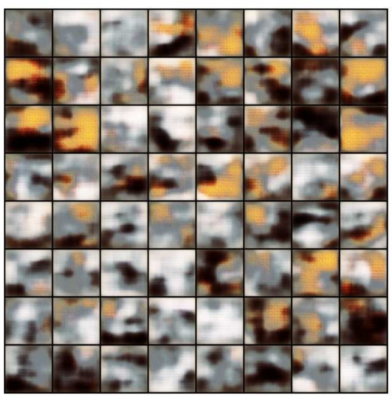
$$\min_{\theta} \left\{ \max_{T_\phi \in \mathrm{Lip1}} \mathcal{E}(T_\phi \# Q, T_\phi \# P_\theta) \right\} ,$$

[7]Note that it is then important to use the $\log(D)$ trick succinctly discussed in the original GAN paper [20].

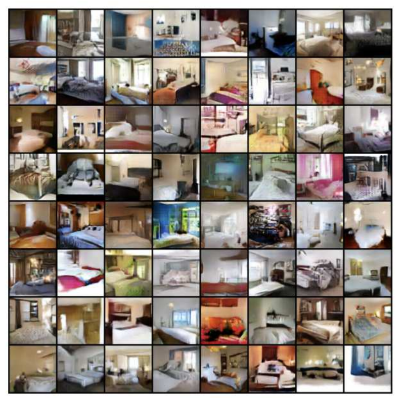
[8]See [53] for the relation between Energy Distance and Cramér distance.



A sample of 64 training examples



Generated by the ED trained model



Generated by the WD trained model

Figure 2: Comparing images generated by a same implicit model trained with different criteria. The top square shows a sample of 64 training examples represening bedroom pictures. The bottom left square shows the images generated by a model trained with ED using the algorithm of [11]. The bottom right square shows images generated by a model trained using the WGAN-GP approach [22].

can then be re-expressed using the IPM form of the energy distance

$$\min_{\theta} \left\{ D(Q,P) = \max_{T_\phi \in \mathrm{Lip1}} \max_{\|f\|_{\mathcal{H}} \le 1} \mathbb{E}_{x\sim Q}[f(T_\phi(x))] - \mathbb{E}_{x\sim P_\theta}[f(T_\phi(x))] \right\} .$$

The cost $D(Q,P)$ above is a new IPM that relies on critic functions of the form $f \circ T_\phi$, where f belongs to the RKHS unit ball, and T_ϕ is 1-Lipschitz continuous. Such hybrid critic functions still have smoothness properties comparable to that of the Lipschitz-continuous critics of the 1-Wasserstein distance. However, since these critic functions do not usually form a RKHS ball, the resulting IPM criterion no longer belongs to the ED/MMD family.

- The same hybrid approach gives comparable results in GMMN-C [32] where the authors replace autoencoder of GMMN+AE with an adversarially trained transformer layer.

On the positive side, such hybrid approaches may lead to more efficient training algorithms than those described in Section 2.3. The precise parametric structure of the transformation layer also provides the means to match what WGAN models achieve by selecting a precise parametric structure for the critic. Yet, in order to understand these subtle effects, it remains useful to clarify the similarities and differences between pure ED/MMD training and pure WD training.

4 Length spaces

This section gives a concise review of the elementary metric geometry concepts useful for the rest of our analysis. Readers can safely skip this section if they are already familiar with metric geometry textbooks such as [12].

Rectifiable curves A continuous mapping $\gamma : t \in [a,b] \subset \mathbb{R} \mapsto \gamma_t \in \mathcal{X}$ defines a curve connecting γ_a and γ_b. A curve is said to be *rectifiable* when its *length*

$$L(\gamma,a,b) \;\triangleq\; \sup_{n>1} \sup_{a=t_0<t_1<\cdots<t_n=b} \sum_{i=1}^{n} d(\gamma_{t_{i-1}},\gamma_t) \tag{23}$$

is finite. Intuitively, thanks to the triangular inequality, dividing the curve into n segments $[\gamma_{t-1},\gamma_t]$ and summing their sizes yields a quantity that is greater than $d(\gamma_a,\gamma_b)$ but smaller than the curvilinear length of the curve. By construction, $L(\gamma,a,b) \ge d(\gamma_a,\gamma_b)$ and $L(\gamma,a,c) = L(\gamma,a,b) + L(\gamma,b,c)$ for all $a \le b \le c$.

Constant speed curves Together with the continuity of γ, this additivity property implies that the function $t \in [a,b] \mapsto L(\gamma,a,t)$ is nondecreasing and continuous [12, Prop. 2.3.4]. Thanks to the intermediate value theorem, when a curve is rectifiable, for all $s \in [0,1]$, there is $t_s \in [a,b]$ such that $L(\gamma,a,t_s) = s\,L(\gamma,a,b)$. Therefore, we can construct a new curve $\bar\gamma : s \in [0,1] \mapsto \bar\gamma_s = \gamma_{t_s}$ that visits the same points in the same order as curve

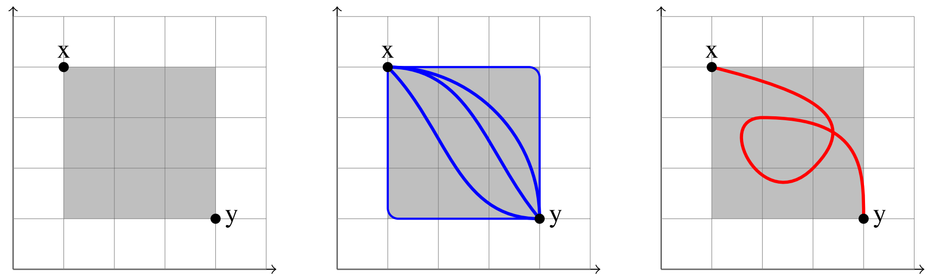


Figure 3: Consider $\mathbb{R}^2$ equipped with the L_1 distance. Left: all points z in the gray area are such that $d(x,z) + d(z,y) = d(x,y)$. Center: all minimal geodesics connecting x and y live in the gray area. Right: but not all curves that live in the gray area are minimal geodesics.

γ and satisfies the property $\forall s \in [0,1]$, $L(\bar{\gamma}, 0, s) = sL(\bar{\gamma}, 0, 1)$. Such a curve is called a *constant speed curve*.

Length spaces It is easy to check that the *distance induced by* d,

$$\hat{d} : (x,y) \in \mathcal{X}^2 \quad \mapsto \inf_{\substack{\gamma:[a,b]\to\mathcal{X} \\ \text{s.t. } \gamma_a = x \ \gamma_b = y}} L(\gamma, a, b) \ \in \mathbb{R}_+^* \cup \{\infty\}\,, \tag{24}$$

indeed satisfies all the properties (2) of a distance. It is also easy to check that the distance induced by $\hat{d}$ coincides with $\hat{d}$ [12, Prop. 2.3.12]. For this reason, a distance that satisfies $\hat{d} = d$ is called an *intrinsic distance*. A Polish metric space equipped with an intrinsic distance is called an *intrinsic Polish space*. A metric space $\mathcal{X}$ equipped with an intrinsic distance d is called a *length space*.

Minimal geodesics A curve $\gamma : [a,b] \to \mathcal{X}$ that achieves the infimum in (24) is called a *shortest path* or a *minimal geodesic* connecting γ_a and γ_b.

When the distance d is intrinsic, the length of a minimal geodesic γ satisfies the relation $L(\gamma, a, b) = \hat{d}(\gamma_a, \gamma_b) = d(\gamma_a, \gamma_b)$. When such a curve exists between any two points x, y such that $d(x,y) < \infty$, the distance d is called *strictly intrinsic*. A Polish space equipped with a strictly intrinsic distance is called a *strictly intrinsic Polish space.*

Conversely, a rectifiable curve $\gamma : [a,b] \to \mathcal{X}$ of length $d(\gamma_a, \gamma_b)$ is a minimal geodesic because no curve joining γ_a and γ_b can be shorter. If there is such a curve between any two points x, y such that $d(x,y) < \infty$, then d is a strictly intrinsic distance.

Characterizing minimal geodesics Let $\gamma : [a,b] \to \mathcal{X}$ be a minimal geodesic in a length space $(\mathcal{X}, d)$. Using the triangular inequality and (23),

$$\forall a \le t \le b \quad d(\gamma_a, \gamma_b) \le d(\gamma_a, \gamma_t) + d(\gamma_t, \gamma_b) \le L(\gamma, a, b) = d(\gamma_a, \gamma_b)\,. \tag{25}$$

This makes clear that every minimal geodesic in a length space is made of points γ_t for which the triangular inequality is an equality. However, as shown in Figure 3, this is not sufficient to ensure that a curve is a minimal geodesic. One has to consider two intermediate points:

Theorem 4.1. *Let $\gamma : [a,b] \to \mathcal{X}$ be a curve joining two points γ_a, γ_b such that $d(\gamma_a, \gamma_b) < \infty$. This curve is a minimal geodesic of length $d(\gamma_a, \gamma_b)$ if and only if $\forall\ a \le t \le t' \le b, \quad d(\gamma_a, \gamma_t) + d(\gamma_t, \gamma_{t'}) + d(\gamma_{t'}, \gamma_b) = d(\gamma_a, \gamma_b)\,.$*

Corollary 4.2. *Let $\gamma : [0,1] \to \mathcal{X}$ be a curve joining two points $\gamma_0, \gamma_1 \in \mathcal{X}$ such that $d(\gamma_0, \gamma_1) < \infty$. The following three assertions are equivalent:*
a) The curve γ is a constant speed minimal geodesic of length $d(\gamma_0, \gamma_1)$.
b) $\forall\ t, t' \in [0,1], \quad d(\gamma_t, \gamma_{t'}) = |t - t'|\, d(\gamma_0, \gamma_1)\,.$
c) $\forall\ t, t' \in [0,1], \quad d(\gamma_t, \gamma_{t'}) \le |t - t'|\, d(\gamma_0, \gamma_1)\,.$

Proof The necessity ($\Rightarrow$) is easily proven by rewriting (25) with two points t and t' instead of just one. The sufficiency ($\Leftarrow$) is proven by induction. Let

$$h_n = \sup_{a=t_0 \le t_1 \le \cdots \le t_n \le b} \ \sum_{i=1}^{n} d(\gamma_{t_{i-1}}, \gamma_{t_i}) \ + \ d(\gamma_{t_n}, \gamma_b)\,.$$

The hypothesis implies that $h_2 = d(\gamma_a, \gamma_b)$. We now assuming that the induction hypothesis $h_n = d(\gamma_a, \gamma_b)$ is true for some $n \ge 2$. For all partition $a = t_0 \le t_1 \ldots t_n \le b$, using twice the triangular inequality and the induction hypothesis,

$$d(\gamma_a, \gamma_b) \le d(\gamma_a, \gamma_{t_n}) + d(\gamma_{t_n}, \gamma_b) \le \sum_{i=1}^{n} d(\gamma_{t_{i-1}}, \gamma_{t_i}) + d(\gamma_{t_n}, \gamma_b) \le h_n = d(\gamma_a, \gamma_b)\,.$$

Therefore $\sum_{i=1}^{n} d(\gamma_{t_{i-1}}, \gamma_{t_i}) = d(\gamma_a, \gamma_{t_n})$. Then, for any $t_{n+1} \in [t_n, b]$,

$$\sum_{i=1}^{n+1} d(\gamma_{t_{i-1}}, \gamma_{t_i}) + d(\gamma_{t_{n+1}}, \gamma_b) = d(\gamma_a, \gamma_{t_n}) + d(\gamma_{t_n}, \gamma_{t_{n+1}}) + d(\gamma_{t_{n+1}}, \gamma_b) = d(\gamma_a, \gamma_b)\,.$$

Since this is true for all partitions, $h_{n+1} = d(\gamma_a, \gamma_b)$. We just have proved by induction that $h_n = d(\gamma_a, \gamma_b)$ for all n. Therefore $L(\gamma, a, b) = \sup_n h_n = d(\gamma_a, \gamma_b)$. ∎

5 Minimal geodesics in probability space

We now assume that $\mathcal{X}$ is a strictly intrinsic Polish space and we also assume that its distance d is never infinite. Therefore any pair of points in $\mathcal{X}$ is connected by at least one minimal geodesic. When the space $\mathcal{P}_\mathcal{X}$ of probability distributions is equipped with one of the probability distances discussed in section 2, it often becomes a length space itself and inherits some of the geometrical properties of $\mathcal{X}$. Since this process depends critically on how the probability distance compares different distributions, understanding the geodesic structure of $\mathcal{P}_\mathcal{X}$ reveals fundamental differences between probability distances.

This approach is in fact quite different from the celebrated work of Amari on *Information Geometry* [2]. We seek here to understand the geometry of the space of all probability measures equipped with different distances. Information Geometry characterizes the

Riemannian geometry of a parametric family of probability measures under the Kullback-Leibler distance. This difference is obviously related to the contrast between *relying on good distances* versus *relying on good model families* discussed in Section 1. Since we are particularly interested in relatively simple models that have a physical or causal interpretation but cannot truly represent the actual data distribution, we cannot restrict our geometrical insights to what happens within the model family.

5.1 Mixture geodesics

For any two distributions $P_0, P_1 \in \mathcal{P}_\mathcal{X}$, the mixture distributions

$$\forall t \in [0,1] \quad P_t = (1-t)P_0 + tP_1 \tag{26}$$

form a curve in the space of distributions $\mathcal{P}_\mathcal{X}$.

Theorem 5.1. *Let $\mathcal{P}_\mathcal{X}$ be equipped with a distance D that belongs to the IPM family* (5)*. Any mixture curve* (26) *joining two distributions $P_0, P_1 \in \mathcal{P}_\mathcal{X}$ such that $D(P_0, P_1) < \infty$ is a constant speed minimal geodesic, making D a strictly intrinsic distance.*

Proof The proof relies on Corollary 4.2: for all $t, t' \in [0,1]$,

$$\begin{aligned} D(P_t, P_{t'}) &= \sup_{f\in\mathcal{Q}} \left\{ \mathbb{E}_{(1-t)P_0+tP_1}[f(x)] - \mathbb{E}_{(1-t')P_0+t'P_1}[f(x)] \right\} \\ &= \sup_{f\in\mathcal{Q}} \left\{ -(t-t')\mathbb{E}_{P_0}[f(x)] + (t-t')\mathbb{E}_{P_1}[(] f(x)) \right\} \\ &= |t-t'| \sup_{f\in\mathcal{Q}} \left\{ \mathbb{E}_{P_0}[f(x)] - \mathbb{E}_{P_1}[f(x)] \right\} \,. \end{aligned}$$

where the last equality relies on the fact that if $f \in \mathcal{Q}$, then $-f \in \mathcal{Q}$. By Corollary 4.2, the mixture curve is a constant speed minimal geodesic. Since this is true for any $P_0, P_1 \in \mathcal{P}_\mathcal{X}$ such that $D(P_0, P_1) < \infty$, the distance D is strictly intrinsic. ∎

Remark 5.2. Although Theorem 5.1 makes $(\mathcal{P}_\mathcal{X}, D)$ a length space, it does not alone make it a strictly intrinsic Polish space. One also needs to establish the completeness and separability[9] properties of a Polish space. Fortunately, these properties are true for both $(\mathcal{P}^1_\mathcal{X}, W_1)$ and $(\mathcal{P}^1_\mathcal{X}, \mathcal{E}_d)$ when the ground space is Polish.[10]

Since both the 1-Wasserstein distance W_1 and the Energy Distance or MMD $\mathcal{E}_d$ belong to the IPM family, $\mathcal{P}_\mathcal{X}$ equipped with either distance is a strictly intrinsic Polish space. Any two probability measures are connected by at least one minimal geodesic, the mixture geodesic. We shall see later that the 1-Wasserstein distance admits many more minimal geodesics. However, in the case of ED/MMD distances, mixture geodesics are the only minimal geodesics.

[9] For instance the set of probability measures on $\mathbb{R}$ equipped with the total variation distance (6) is not separable because any dense subset needs one element in each of the disjoint balls $B_x = \{\, P{\in}\mathcal{P}_\mathbb{R} : D_{TV}(P, \delta_x){<}1/2 \,\}$.

[10] For the Wasserstein distance, see [55, Theorem 6.18]. For the Energy distance, both properties can be derived from Theorem 2.17 after recaling that $\Phi_\mathcal{X} \subset \mathcal{H}$ is both complete and separable because it is isometric to $\mathcal{X}$ which is Polish.

Theorem 5.3. *Let K be a characteristic kernel and let $\mathcal{P}_\mathcal{X}$ be equipped with the MMD distance $\mathcal{E}_{d_K}$. Then any two probability measures $P_0, P_1 \in \mathcal{P}_\mathcal{X}$ such that $\mathcal{E}_{d_K}(P_0, P_1) < \infty$ are joined by exactly one constant speed minimal geodesic, the mixture geodesic* (26).

Note that $\mathcal{E}_{d_K}$ is also the ED for the strongly negative definite kernel d_K.

Proof Theorem 5.1 already shows that any two measures $P_0, P_1 \in \mathcal{P}_\mathcal{X}$ are connected by the mixture geodesic P_t. We only need to show that it is unique. For any $t \in [0,1]$, the measure P_t belongs to the set

$$\left\{ P \in \mathcal{P}_\mathcal{X} : \mathcal{E}_{d_K}(P_0, P) = tD \text{ and } \mathcal{E}_{d_K}(P, P_1) = (1{-}t)D \right\} \subset \mathcal{P}_\mathcal{X} \tag{27}$$

where $D = \mathcal{E}_{d_K}(P_0, P_1)$. Thanks to Theorem 2.17, $\mathbb{E}_{P_t}[\Phi_x]$ must belong to the set

$$\left\{ \Psi \in \mathcal{H} : \|\mathbb{E}_{P_0}[\Phi_x] - \Psi\|_\mathcal{H} = tD \text{ and } \|\Psi - \mathbb{E}_{P_1}[\Phi_x]\,\|_\mathcal{H} = (1{-}t)D \right\} \subset \mathcal{H}\,. \tag{28}$$

with $D = \|\mathbb{E}_{P_0}[\Phi_x] - \mathbb{E}_{P_1}[\Phi_x]\,\|_\mathcal{H}$. Since there is only one point Ψ that satisfies these conditions in $\mathcal{H}$, and since Corollary 2.20 says that the map $P \mapsto \mathbb{E}_P[\Phi_x]$ is injective, there can only be one P satisfying (27) and this must be P_t. Therefore the mixture geodesic is the only one. ∎

5.2 Displacement geodesics

Displacement geodesics in the Euclidean case Let us first assume that $\mathcal{X}$ is a Euclidean space and $\mathcal{P}_\mathcal{X}$ is equipped with the p-Wasserstein distance W_p. Let $P_0, P_1 \in \mathcal{P}^p_\mathcal{X}$ be two distributions with optimal transport plan π. The *displacement curve* joining P_0 to P_1 is formed by the distributions

$$\forall t \in [0,1] \quad P_t = \big((1{-}t)x + ty\big)\#\pi(x,y)\,. \tag{29}$$

Intuitively, whenever the optimal transport plan specifies that a grain of probability mass must be transported from x to y in $\mathcal{X}$, we follow the shortest path connecting x and y, that is, in a Euclidean space, a straight line, but we drop the grain after performing a fraction t of the journey.

Proposition 5.4. *Let $\mathcal{X}$ be a Euclidean space and let $\mathcal{P}_\mathcal{X}$ be equipped with the p-Wasserstein distance* (8) *for some $p \geq 1$. Any displacement curve* (29) *joining two distributions P_0, P_1 such that $W_p(P_0, P_1) < \infty$ is a constant speed minimal geodesic, making W_p a strictly intrinsic distance.*

Proof Let π_{01} be the optimal transport plan between P_0 and P_1. For all $t, t' \in [0,1]$, define a tentative transport plan $\pi_{tt'}$ between P_t and $P_{t'}$ as

$$\pi_{tt'} = \big(\,(1{-}t)x + ty,\; (1{-}t')x + t'y\,\big)\#\pi_{01}(x,y) \;\in\; \Pi(P_t, P_{t'})\,.$$

Then

$$\begin{aligned} W_p(P_t, P_{t'})^p &\leq \mathbb{E}_{(x,y)\sim\pi_{tt'}}[\,\|x-y\|^p\,] \\ &= \mathbb{E}_{(x,y)\sim\pi}\big[\,\|(1{-}t)x + ty - (1{-}t')x - t'y\|^p\,\big] \\ &= |t-t'|^p\,\mathbb{E}_{(x,y)\sim\pi}[\,\|x-y\|\,] = |t-t'|^p\,W_p(P_0,P_1)^p\,. \end{aligned}$$

By Corollary 4.2, the displacement curve is a constant speed minimal geodesic. Since this is true for any $P_0, P_1 \in \mathcal{P}_\mathcal{X}$ such that $W_p(P_0, P_1) < \infty$, the distance W_p is strictly intrinsic. ∎

When $p > 1$, it is a well-known fact that the displacement geodesics are the only geodesics of $\mathcal{P}_\mathcal{X}$ equipped with the W_p distance.

Proposition 5.5. *The displacement geodesics* (29) *are the only constant speed minimal geodesics of $\mathcal{P}_\mathcal{X}$ equipped with the p-Wasserstein distance W_p with $p > 1$.*

This is a good opportunity to introduce a very useful lemma.

Lemma 5.6 (Gluing)**.** *Let $\mathcal{X}_i$, $i = 1, 2, 3$ be Polish metric spaces. Let probability measures $\mu_{12} \in \mathcal{P}_{\mathcal{X}_1\times\mathcal{X}_2}$ and $\mu_{23} \in \mathcal{P}_{\mathcal{X}_2\times\mathcal{X}_3}$ have the same marginal distribution μ_2 on $\mathcal{X}_2$. Then there exists $\mu \in \mathcal{P}_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3}$ such that $(x, y)\#\mu(x, y, z) = \mu_{12}$ and $(y, z)\#\mu(x, y, z) = \mu_{23}$.*

Proof notes for Lemma 5.6 At first sight, this is simply $P(x, y, z) = P(x|y)P(z|y)P(y)$ with $\mu_{12}{=}P(x, y)$, $\mu_{23}{=}P(y, z)$. Significant technical difficulties arise when $P(y) = 0$. This is where one needs the topological properties of a Polish space [8].

Proof of Proposition 5.5 Let $t \in [0, 1] \mapsto P_t$ be a constant speed minimal geodesic. Any point P_t must satisfy the equality

$$W_p(P_0, P_t) + W_p(P_t, P_1) = W_p(P_0, P_1)$$

Let π_0 and π_1 be the optimal transport plans associated with $W_p(P_0, P_t)$ and $W_p(P_t, P_1)$ and construct $\pi_3 \in \mathcal{P}_{\mathcal{X}^3}$ by gluing them. Then we must have

$$\begin{aligned}\left(\mathbb{E}_{(x,y,z)\sim\pi_3}[\,\|x-y\|^p\,]\right)^{1/p} &+ \left(\mathbb{E}_{(x,y,z)\sim\pi_3}[\,\|y-z\|^p\,]\right)^{1/p} \\ &= W_p(P_0, P_1) \;\leq\; \left(\mathbb{E}_{(x,y,z)\sim\pi_3}[\,\|x-z\|^p\,]\right)^{\frac{1}{p}} .\end{aligned}$$

Thanks to the properties of the Minkowski's inequality, this can only happen for $p > 1$ if there exists $\lambda \in [0, 1]$ such that, π_3-almost surely, $\|x{-}y\|{=}\lambda\|x{-}z\|$ and $\|y{-}z\|{=}(1{-}\lambda)\|x{-}z\|$. This constant can only be t because $W_P(P_0, P_t) = tW_p(P_0, P_1)$ on a constant speed minimal geodesic. Therefore $y = tx + (1{-}t)y$, π_3-almost surely. Therefore $P_t = y\#\pi(x, y, z)$ describes a displacement curve as defined in (29). ∎

Note however that the displacement geodesics are not the only minimal geodesics of the 1-Wasserstein distance W_1. Since W_1 is an IPM (14), we know that the mixture geodesics are also minimal geodesics (Theorem 5.1). There are in fact many more geodesics. Intuitively, whenever the optimal transport plan from P_0 to P_1 transports a grain of probability from x to y, we can drop the grain after a fraction t of the journey (displacement geodesics), we can randomly decide whether to transport the grain as planned (mixture geodesics), we can also smear the grain of probability along the shortest path connecting x to y, and we can do all of the above using different t in different parts of the space.

Displacement geodesics in the general case The rest of this section reformulates these results to the more general situation where $\mathcal{X}$ is a strictly intrinsic Polish space. Rather than following the random curve approach described in [55, Chapter 7], we chose a more elementary approach because we also want to characterize the many geodesics of W_1. Our definition is equivalent for $p{>}1$ and subtly weaker for $p{=}1$.

The main difficulties are that we may no longer have a single shortest path connecting two points $x, y \in \mathcal{X}$, and that we may not be able to use the push-forward formulation (29) because the function that returns the point located at position t along a constant speed minimal geodesic joining x to y may not satisfy the necessary measurability requirements.

Definition 5.7 (Displacement geodesic)**.** *Let $\mathcal{X}$ be a strictly intrinsic Polish metric space and let $\mathcal{P}^p_{\mathcal{X}}$ be equipped with the p-Wasserstein distance W_p. The curve $t \in [0,1] \mapsto P_t \in \mathcal{P}^p_{\mathcal{X}}$ is called a displacement geodesic if, for all $0{\le}t{\le}t'{\le}1$, there is a distribution $\pi_4 \in \mathcal{P}_{\mathcal{X}^4}$ such that*

$i)$ The four marginals of π_4 are respectively equal to P_0, P_t, $P_{t'}$, P_1.

$ii)$ The pairwise marginal $(x,z)\#\pi_4(x,u,v,z)$ is an optimal transport plan

$$W_p(P_0,P_1)^p = \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)^p] \ .$$

$iii)$ The following relations hold $\pi_4(x,u,v,z)$-almost surely:

$$d(x,u) = t\,d(x,z)\,, \quad d(u,v) = (t'-t)\,d(x,z)\,, \quad d(v,z) = (1{-}t')\,d(x,z)\ .$$

Proposition 5.8. *Definition 5.7 indeed implies that P_t is a constant speed minimal geodesic of length $W_p(P_0,P_1)$. Furthermore, for all $0 \le t \le t' \le 1$, all the pairwise marginals of π_4 are optimal transport plans between their marginals.*

Proof For all $0 \le t \le t' \le 1$, we have

$$\begin{aligned} W_p(P_t,P_{t'})^p &\le \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(u,v)^p] \\ &= (t{-}t')^p\,\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)^p] = (t{-}t')^p\,W_p(P_0,P_1)^p\ . \end{aligned}$$

By Corollary 4.2, the curve P_t is a constant speed minimal geodesic. We can then write

$$\begin{aligned} t'\,W_p(P_0,P_1) = W_p(P_0,P_{t'}) &\le \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,v)^p]\right)^{1/p} \\ &\le \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[(d(x,u)+d(u,v))^p]\right)^{1/p} \\ &\le \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,u)^p]\right)^{1/p} + \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(u,v)^p]\right)^{1/p} \\ &\le t\,\left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)^p]\right)^{1/p} + (t'{-}t)\,\left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)^p]\right)^{1/p} \\ &= t'\,W_p(P_0,P_1)\,, \end{aligned}$$

where the third inequality is Minkowski's inequality. Since both ends of this chain of inequalities are equal, these inequalities must be equalities, implying that $(x,v)\#\pi_4$ is an optimal transport plan between P_0 and $P_{t'}$. We can do likewise for all pairwise marginals of π_4. ■

The proposition above does not establish that a displacement geodesic always exists. As far as we know, this cannot be established without making an additional assumption such as the local compacity of the intrinsic Polish space $\mathcal{X}$. Since it is often easy to directly define a displacement geodesic as shown in (29), we omit the lengthy general proof.

Theorem 5.9. *Let $\mathcal{X}$ be a strictly intrinsic Polish metric space and let P_0, P_1 be two distributions of $\mathcal{P}^p_{\mathcal{X}}$ equipped with the p-Wasserstein with $p > 1$. The only constant speed minimal geodesics of length $W_p(P_0, P_1)$ joining P_0 and P_1 are the displacement geodesics.*

Proof Let P_t be a constant speed minimal geodesic of length $W_p(P_0, P_1)$. By Theorem 4.1, for all $0 \leq t \leq t' \leq 1$,

$$W_p(P_0, P_t) + W_p(P_t, P_{t'}) + W_p(P_{t'}, P_1) = W_p(P_0, P_1) \; .$$

Let π_4 be constructed by gluing optimal transport plans associated with the three distances appearing on the left hand side of the above equality. We can then write

$$\begin{aligned} W_p(P_0, P_1) &\leq \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)^p]\right)^{1/p} \\ &\leq \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[(d(x,u)+d(u,v)+d(v,z))^p]\right)^{1/p} \\ &\leq \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,u)^p]\right)^{1/p} + \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(u,v)^p]\right)^{1/p} \\ &\qquad\qquad + \left(\mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(v,z)^p]\right)^{1/p} \\ &= W_p(P_0, P_t) + W_p(P_t, P_{t'}) + W_p(P_{t'}, P_1) = W_p(P_0, P_1) \; . \end{aligned}$$

Since this chain of inequalities has the same value in both ends, all these inequalities must be equalities. The first one means that π_4 is an optimal transport plan for $W_p(P_0, P_1)$. The second one means that $(d(x,u) + d(u,v) + d(v,z) = d(x,z)$, π_4-almost surely. When $p > 1$, the third one, Minkowski's inequality can only be an inequality if there are scalars $\lambda_1{+}\lambda_2{+}\lambda_3{=}1$ such that, π_4-almost surely, $d(x,u) = \lambda_1 d(x,z)$, $d(x,u) = \lambda_2 d(x,z)$, and $d(v,z) = \lambda_3 d(x,z)$. Since P_t must satisfy Corollary 4.2, these scalars can only be $\lambda_1 = t$, $\lambda_2 = t'{-}t$, and $\lambda_3 = 1{-}t'$. ■

Minimal geodesics for the 1-Wasserstein distance We can characterize the many minimal geodesics of the 1-Wasserstein distance using a comparable strategy.

Theorem 5.10. *Let $\mathcal{X}$ be a strictly intrinsic Polish space and let $\mathcal{P}^1_{\mathcal{X}}$ be equipped with the distance W_1. A curve $t \in [a,b] \mapsto P_t \in \mathcal{P}^1_{\mathcal{X}}$ joining P_a and P_b is a minimal geodesic of length $W_1(P_a, P_b)$ if and only if, for all $a \leq t \leq t' \leq b$, there is a distribution $\pi_4 \in \mathcal{P}_{\mathcal{X}^4}$ such that*

$i)$ *The four marginals of π_4 are respectively equal to P_a, P_t, $P_{t'}$, P_b.*

$ii)$ *The pairwise marginal $(x,z)\#\pi_4(x,u,v,z)$ is an optimal transport plan*

$$W_p(P_a, P_b) = \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)] \; .$$

$iii)$ *The following relation holds $\pi_4(x,u,v,z)$-almost surely:*

$$d(x,u) + d(u,v) + d(v,z) = d(x,z) \; .$$

It is interesting to compare this condition to Theorem 4.1. Instead of telling us that two successive triangular inequalities in the probability space $(\mathcal{P}^1_{\mathcal{X}}, W_1)$ must be an equality, this result tells us that the same holds almost-surely in the sample space $(\mathcal{X}, d)$. In particular, this means that x, u, v, and z must be aligned along a geodesic of $\mathcal{X}$. In the case of a mixture geodesic, u and v coincide with x or z. In the case of a displacement geodesic, u

and v must be located at precise positions along a constant speed geodesic joining x to z. But there are many other ways to fulfil these conditions.

Proof When P_t is a minimal geodesic, Theorem 4.1 states

$$\forall a \leq t \leq t' \leq b \quad W_1(P_a, P_t) + W_1(P_t, P_{t'}) + W_1(P_{t'}, P_b) = W_1(P_a, P_b) \ .$$

Let π_4 be constructed by gluing optimal transport plans associated with the three distances appearing on the left hand side of the above equality. We can then write

$$\begin{aligned} \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)] &\leq \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,u) + d(u,v) + d(v,z)] \\ &= W_1(P_a, P_b) \leq \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)] \ . \end{aligned}$$

Since this chain of equalities has the same value on both ends, all these inequalities must be equalities. The first one means that $d(x,u) + d(u,v) + d(v,z) = d(x,z)$, π_4-almost surely. The second one means that $(x,z)\#\pi_4$ is an optimal transport plan.

Conversely, assume P_t satisfies the conditions listed in the proposition. We can then write, for all $a \leq t \leq t' \leq b$,

$$\begin{aligned} W_1(P_a, P_b) &= \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,z)] \\ &= \mathbb{E}_{(x,u,v,z)\sim\pi_4}[d(x,u) + d(u,v) + d(v,z)] \\ &= W_1(P_a, P_t) + W_1(P_t, P_{t'}) + W_1(P_{t'}, P_b) \ , \end{aligned}$$

and we conclude using Theorem 4.1. ∎

6 Unsupervised learning and geodesic structures

We have seen in the previous section that the geometry of the space $\mathcal{P}_\mathcal{X}$ of probability distributions changes considerably with our choice of a probability distance. Critical aspects of these possible geometries can be understood from the characterization of the shortest paths between any two distributions:

- With the Energy Distance $\mathcal{E}_d$ or the Maximum Mean Discrepancy $\mathcal{E}_{d_K}$, the sole shortest path is the mixture geodesic (Theorem 5.3.)
- With the p-Wasserstein distance W_p, for $p > 1$, the sole shortest paths are displacement geodesics (Theorem 5.9.)
- With the 1-Wasserstein distance W_1, there are many shortest paths, including the mixture geodesic, all the displacement geodesics, and all kinds of hybrid curves (Theorem 5.10.)

The purpose of this section is to investigate the consequences of these geometrical differences on unsupervised learning problems. In the following discussion, $Q \in \mathcal{P}_\mathcal{X}$ represents the data distribution which is only known through the training examples, and $\mathcal{F} \subset \mathcal{P}_\mathcal{X}$ represent the family of parametric models $P_\theta \in \mathcal{P}_\mathcal{X}$ considered by our learning algorithm.

Minimal geodesics in length spaces can sometimes be compared to line segments in Euclidean spaces because both represent shortest paths between two points. This association provides the means to extend the familiar Euclidean notion of convexity to length spaces. This section investigates the geometry of implicit modeling learning problems through the lens of this generalized notion of convexity.

6.1 Convexity *à-la-carte*

We now assume that $\mathcal{P}_\mathcal{X}$ is a strictly intrinsic Polish space equipped with a distance D. Let $\mathcal{C}$ be a family of smooth constant speed curves in $\mathcal{P}_\mathcal{X}$. Although these curves need not be minimal geodesics, the focus of this section is limited to three families of curves defined in Section 5:

- the family $\mathcal{C}_g(D)$ of all minimal geodesics in $(\mathcal{P}_\mathcal{X}, D)$,
- the family $\mathcal{C}_d(W_p)$ of the displacement geodesics in $(\mathcal{P}^p_\mathcal{X}, W_p)$,
- the family $\mathcal{C}_m$ of the mixture curves in $\mathcal{P}_\mathcal{X}$.

Definition 6.1. *Let $\mathcal{P}_\mathcal{X}$ be a strictly intrinsic Polish space. A closed subset $\mathcal{F} \subset \mathcal{P}_\mathcal{X}$ is called convex with respect to the family of curves $\mathcal{C}$ when, for all $P_0, P_1 \in \mathcal{P}_\mathcal{X}$, the set $\mathcal{C}$ contains a curve $t \in [0,1] \mapsto P_t \in \mathcal{X}$ connecting P_0 and P_1 whose graph is contained in $\mathcal{F}$, that is, $P_t \in \mathcal{F}$ for all $t \in [0,1]$.*

Definition 6.2. *Let $\mathcal{P}_\mathcal{X}$ be a strictly intrinsic Polish space. A real-valued function f defined on $\mathcal{P}_\mathcal{X}$ is called convex with respect to the family of constant speed curves $\mathcal{C}$ when, for every curve $t \in [0,1] \mapsto P_t \in \mathcal{P}_\mathcal{X}$ in $\mathcal{C}$, the function $t \in [0,1] \mapsto f(P_t) \in \mathbb{R}$ is convex.*

For brevity we also say that $\mathcal{F}$ or f is *geodesically convex* when $\mathcal{C} = \mathcal{C}_g(D)$, *mixture convex* when $\mathcal{C} = \mathcal{C}_m$, and *displacement convex* when $\mathcal{C} = \mathcal{C}_d(W_p)$.

Theorem 6.3 (Convex optimization *à-la-carte*)**.** *Let $\mathcal{P}_\mathcal{X}$ be a strictly intrinsic Polish space equipped with a distance D. Let the closed subset $\mathcal{F} \subset \mathcal{P}_\mathcal{X}$ and the cost function $f : \mathcal{P}_\mathcal{X} \mapsto \mathbb{R}$ be both convex with respect to a same family $\mathcal{C}$ of constant speed curves. Then, for all $M \geq \min_\mathcal{F}(f)$,*

$i)$ the level set $L(f, \mathcal{F}, M) = \{P \in \mathcal{F} : f(P) \leq M\}$ is connected,

$ii)$ for all $P_0 \in \mathcal{F}$ such that $f(P_0) > M$ and all $\epsilon > 0$, there exists $P \in \mathcal{F}$ such that $D(P, P_0) = \mathcal{O}(\epsilon)$ and $f(P) \leq f(P_0) - \epsilon(f(P_0) - M)$.

This result essentially means that it is possible to optimize the cost function f over $\mathcal{F}$ with a descent algorithm. Result (i) means that all minima are global minima, and result (ii) means that any neighborhood of a suboptimal distribution P_0 contains a distribution P with a sufficiently smaller cost to ensure that the descent will continue.

Proof (i): Let $P_0, P_1 \in L(f, \mathcal{F}, M)$. Since they both belong to $\mathcal{F}$, $\mathcal{C}$ contains a curve $t \in [0,1] \mapsto P_t \in \mathcal{F}$ joining P_0 and P_1. For all $t \in [0,1]$, we know that $P_t \in \mathcal{F}$ and, since $t \mapsto f(P_t)$ is a convex function, we can write $f(P_t) \leq (1-t)f(P_0) + tf(P_1) \leq M$. Therefore $P_t \in L(f, \mathcal{F}, M)$. Since this holds for all P_0, P_1, $L(f, \mathcal{F}, M)$ is connected.

(ii): Let $P_1 \in L(f, \mathcal{F}, M)$. Since $\mathcal{F}$ is convex with respect to $\mathcal{C}$, $\mathcal{C}$ contains a constant speed curve $t \in [0,1] \mapsto P_t \in \mathcal{F}$ joining P_0 and P_1. Since this is a constant speed curve, $d(P_0, P_\epsilon) \leq \epsilon D(P_0, P_1)$, and since $t \mapsto f(P_t)$ is convex, $f(P_\epsilon) \leq (1-\epsilon)f(P_0) + \epsilon f(P_1)$, implies $f(P_\epsilon \leq f(P_0) - \epsilon(f(P_0) - M)$. ∎

One particularly striking aspect of this result is that it does not depend on the parametrization of the family $\mathcal{F}$. Whether the cost function $C(\theta) = f(G_\theta \# \mu_z)$ is convex or not is irrelevant: as long as the family $\mathcal{F}$ and the cost function f are convex with respect to a well-chosen set of curves, the level sets of the cost function $C(\theta)$ will be connected, and there will be a nonincreasing path connecting any starting point θ_0 to a global optimum θ^*.

It is therefore important to understand how the definition of $\mathcal{C}$ makes it easy or hard to ensure that both the model family $\mathcal{F}$ and the training criterion f are convex with respect to $\mathcal{C}$.

6.2 The convexity of implicit model families

We are particularly interested in the case of implicit models (Section 2.2) in which the distributions P_θ are expressed by pushing the samples $z \in \mathcal{Z}$ of a known source distribution $\mu_z \in \mathcal{P}_\mathcal{Z}$ through a parametrized generator function $G_\theta(z) \in \mathcal{X}$. This push-forward operation defines a deterministic coupling between the distributions μ_z and P_θ because the function G_θ maps every source sample $z \in \mathcal{Z}$ to a single point $G_\theta(z)$ in $\mathcal{X}$. In contrast, a stochastic coupling distribution $\pi_\theta \in \Pi(\mu_z, P_\theta) \subset \mathcal{P}_{\mathcal{Z}\times\mathcal{X}}$ would be allowed to distribute a source sample $z \in \mathcal{Z}$ to several locations in $\mathcal{X}$, according to the conditional distribution $\pi_\theta(x|z)$.

The deterministic nature of this coupling makes it very hard to achieve mixture convexity using smooth generator functions G_θ.

Example 6.4. Let the distributions $P_0, P_1 \in \mathcal{F}$ associated with parameters θ_0 and θ_1 have disjoint supports separated by a distance greater than $D{>}0$. Is there a continuous path $t \in [0,1] \mapsto \theta_t$ in parameter space such that $G_{\theta_t} \# \mu_z$ is the mixture $P_t = (1{-}t)P_0 + P_1$?

If we assume there is such a path, we can write

$$\mu_z\{G_{\theta_0}(z) \in \text{supp}(P_0)\} = 1$$

and, for any $\epsilon{>}0$,

$$\mu_z\{G_{\theta_\epsilon}(z) \in \text{supp}(P_1)\} = \epsilon > 0 \ .$$

Therefore, for all $\epsilon{>}0$, there exists $z{\in}\mathcal{Z}$ such that $d(G_{\theta_0}(z), G_{\theta_\epsilon}(z)){\geq}D$. Clearly such a generator function is not compatible with the smoothness requirements of an efficient learning algorithm.

In contrast, keeping the source sample z constant, a small change of the parameter θ causes a small displacement of the generated sample $G_\theta(z)$ in the space $\mathcal{X}$. Therefore we can expect that such an implicit model family has a particular affinity for displacement geodesics.

It is difficult to fully assess the consequences of the quasi-impossibility to achieve mixture convexity with implicit models. For instance, although the Energy Distance $\mathcal{E}_d(Q, P_\theta)$ is a mixture convex function (see Proposition 6.6 in the next section), we cannot expect that a family $\mathcal{F}$ of implicit models will be mixture convex.

Example 6.5. Let μ_z be the uniform distribution on $\{-1, +1\}$. Let the parameter θ be constrained to the square $[-1, 1]^2 \subset \mathbb{R}^2$ and let the generator function be

$$G_\theta : z \in \{-1, 1\} \mapsto G_\theta(z) = z\theta \ .$$

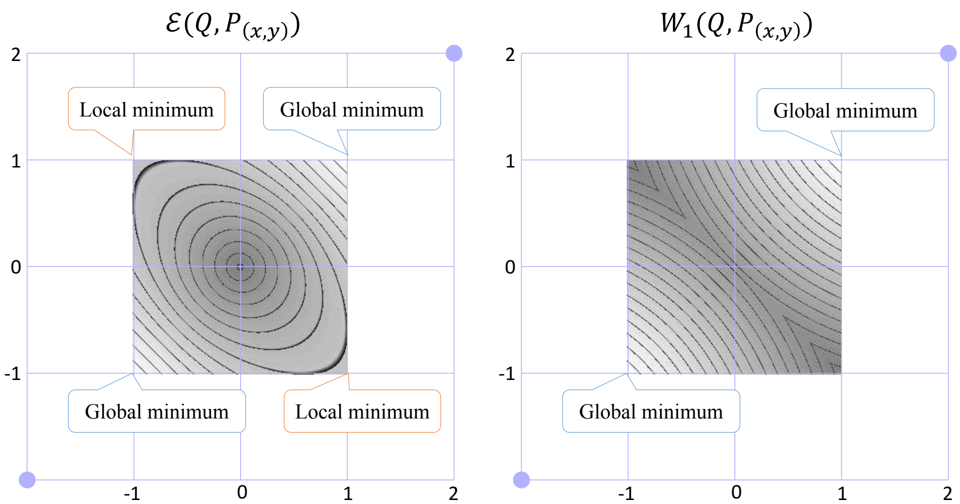


Figure 4: Level sets for the problems described in Example 6.5.

The corresponding model family is

$$\mathcal{F} = \left\{ P_\theta = \tfrac{1}{2}(\delta_\theta + \delta_{-\theta}) : \theta \in [-1,1] \times [-1,1] \right\} .$$

It is easy to see that this model family is displacement convex but not mixture convex. Figure 4 shows the level sets for both criteria $\mathcal{E}(Q, P_\theta)$ and $W_1(Q, P_\theta)$ for the target distribution $Q = P_{(2,2)} \notin \mathcal{F}$. Both criteria have the same global minima in $(1,1)$ and $(-1,-1)$. However the energy distance has spurious local minima in $(-1,1)$ and $(1,-1)$ with a relatively high value of the cost function.

Constructing such an example in $\mathbb{R}^2$ is nontrivial. Whether such situations arise commonly in higher dimension is not known. However we empirically observe that the optimization of a MMD criterion on high-dimensional image data often stops with unsatisfactory results (Section 3.2).

6.3 The convexity of distances

Let $Q \in \mathcal{P}_\mathcal{X}$ be the target distribution for our learning algorithm. This could be the true data distribution or the empirical training set distribution. The learning algorithm minimizes the cost function

$$\min_{P_\theta \in \mathcal{F}} C(\theta) \;\stackrel{\Delta}{=}\; D(Q, P_\theta)\,. \tag{30}$$

The cost function itself is therefore a distance. Since such a distance function is always convex in a Euclidean space, we can ask whether a distance in a strictly intrinsic Polish space is geodesically convex. This is not always the case. Figure 5 gives a simple counterexample in $\mathbb{R}^2$ equipped with the L_1 distance.

Yet we can give a positive answer for the mixture convexity of IPM distances.

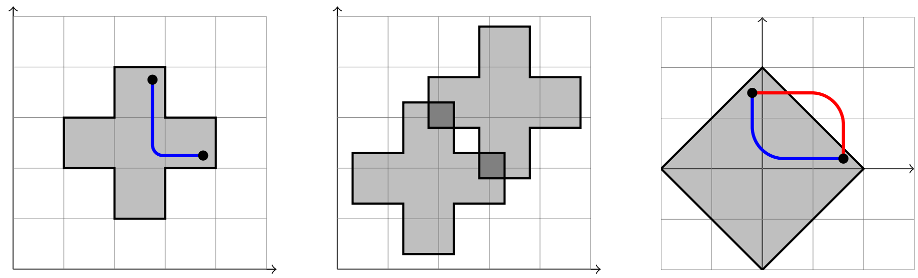


Figure 5: Geodesic convexity often differs from Euclidean convexity in important ways. There are many different minimal geodesics connecting any two points in $\mathbb{R}^2$ equipped with the L_1 distance (see also Figure 3). The cross-shaped subset of $\mathbb{R}_2$ shown in the left plot is geodesically convex. The center plot shows that the intersection of two geodesically convex sets is not necessarily convex or even connected. The right plot shows that two points located inside the unit ball can be connected by a minimal geodesic that does not stay in the unit ball. This means that the L_1 distance itself is not convex because its restriction to that minimal geodesic is not convex.

Proposition 6.6. *Let $\mathcal{P}_\mathcal{X}$ be equipped with a distance D that belongs to the IPM family* (5). *Then D is mixture convex.*

Proof Let $t \in [0,1] \mapsto P_t = (1{-}t)P_0 + tP_1$ be a mixture curve. Theorem 5.1 tells us that such mixtures are minimal geodesics. For any target distribution Q we can write

$$\begin{aligned} D(Q, P_t) &= \sup_{f\in\mathcal{Q}} \left\{ \mathbb{E}_Q[f(x)] - \mathbb{E}_{P_t}[f(x)] \right\} \\ &= \sup_{f\in\mathcal{Q}} \left\{ (1{-}t) \left(\mathbb{E}_Q[f(x)] - \mathbb{E}_{P_0}[f(x)]\right) + t \left(\mathbb{E}_Q[f(x)] - \mathbb{E}_{P_1}[f(x)]\right) \right\} \\ &\leq (1{-}t)\, D(Q, P_0) + t\, D(Q, P_1)\ . \end{aligned}$$

The same holds for any segment $t \in [t_1, t_2] \subset [0,1]$ because such segments are also mixture curves up to an affine reparametrization. Therefore $t \mapsto D(Q, P_t)$ is convex. ■

Therefore, when D is an IPM distance, and when $\mathcal{F}$ is a mixture convex family of generative models, Theorem 6.3 tells us that a simple descent algorithm can find the global minimum of (30). As discussed in Example 6.4, it is very hard to achieve mixture convexity with a family of implicit models. But this could be achieved with nonparametric techniques.

However the same does not hold for displacement convexity. For instance, the Wasserstein distance is not displacement convex, even when the sample space distance d is geodesically convex, and even when the sample space is Euclidean.

Example 6.7. Let $\mathcal{X}$ be $\mathbb{R}^2$ equipped with the Euclidean distance. Let Q be the uniform distribution on the unit circle, and let $P_{\ell,\theta}$ be the uniform distribution on a line segment of length 2ℓ centered on the origin (Figure 6, left). The distance $W_1(Q, P_{\ell,\theta})$ is independent on θ and decreases when $\ell \in [0,1]$ increases (Figure 6, center). Consider a displacement geodesic $t \in [0,1] \mapsto P_t$ where $P_0 = P_{\ell,\theta_0}$ and $P_1 = P_{\ell,\theta_1}$ for $0{<}\theta_0{<}\theta_1{<}\pi/2$. Since the

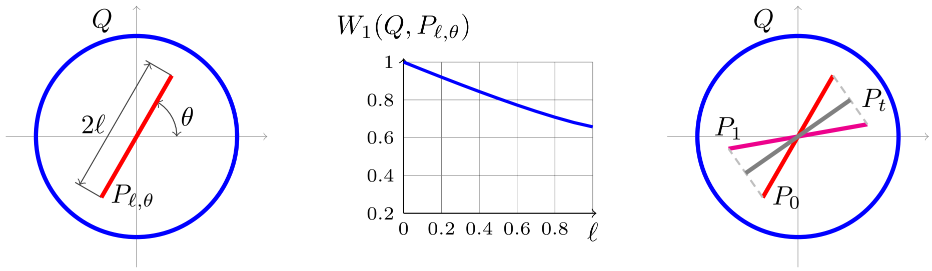


Figure 6: Example 6.7 considers a target distribution Q that is uniform on the $\mathbb{R}^2$ unit circle and a displacement geodesic between two line segments centered on the origin and with identical length (right plot.)

space $\mathbb{R}_2$ is Euclidean, displacements occur along straight lines. Therefore the distribution P_t for $0<t<1$ is uniform on a slightly shorter line segment (Figure 6, right), implying

$$W_1(Q, P_t) \ > \ W_1(Q, P_0) = W_1(Q, P_1) \ .$$

Therefore the distance function $P \mapsto W_1(Q, P)$ is not displacement convex.

Although this negative result prevents us from invoking Theorem 6.3 for the minimization of the Wasserstein distance, observe that the convexity violation in Example 6.7 is rather small. Convexity violation examples are in fact rather difficult to construct. The following section shows that we can still obtain interesting guarantees by bounding the size of the convexity violation.

6.4 Almost-convexity

We consider in this section that the distance d is geodesically convex in $\mathcal{X}$: for any point $x \in \mathcal{X}$ and any constant speed geodesic $t \in [0,1] \mapsto \gamma_t \in \mathcal{X}$,

$$d(x, \gamma_t) \leq (1-t)\, d(x, \gamma_0) + t\, d(x, \gamma_1) \ .$$

This requirement is of course verified when $\mathcal{X}$ is an Euclidean space. This is also trivially true when $\mathcal{X}$ is a Riemannian or Alexandrov space with nonpositive curvature [12].

The following result bounds the convexity violation:

Proposition 6.8. *Let $\mathcal{X}$ be a strictly intrinsic Polish space equipped with a geodesically convex distance d and let $\mathcal{P}^1_{\mathcal{X}}$ be equipped with the 1-Wasserstein distance W_1. For all $Q \in \mathcal{P}_{\mathcal{X}}$ and all displacement geodesics $t \in [0,1] \mapsto P_t$,*

$$\forall t \in [0,1] \quad W_1(Q, P_t) \leq (1-t)\, W_1(Q, P_0) + t\, W_1(Q, P_1) + 2t(1-t) K(Q, P_0, P_1)$$

with $K(Q, P_0, P_1) \leq 2 \min_{u_0 \in \mathcal{X}} \mathbb{E}_{u \sim Q}[d(u, u_0)]$.

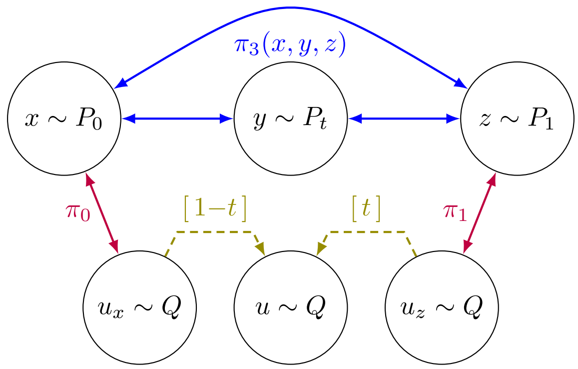


Figure 7: The construction of $\pi \in \mathcal{P}_{\mathcal{X}^6}$ in the proof of Proposition 6.8.

Proof The proof starts with the construction of a distribution $\pi \in \mathcal{P}_{\mathcal{X}^6}$ illustrated in Figure 7. Thanks to Proposition 5.8 we can construct a distribution $\pi_3(x, y, z) \in \mathcal{P}_{\mathcal{X}^3}$ whose marginals are respectively P_0, P_t, P_1, whose pairwise marginals are optimal transport plans, and such that, π_3-almost surely,

$$d(x, y) = t\, d(x, z) \quad d(y, z) = (1-t)\, d(x, z)\,.$$

We then construct distribution $\pi_5(x, y, z, u_x, u_z)$ by gluing π_3 with the optimal transport plans π_0 between P_0 and Q and π_1 between P_1 and Q. Finally $\pi(x, y, z, u_x, u_z, u)$ is constructed by letting u be equal to u_x with probability $1-t$ and equal to u_z with probability t. The last three marginals of π are all equal to Q.

Thanks to the convexity of d in $\mathcal{X}$, the following inequalities hold π-almost surely:

$$\begin{aligned}
d(u_x, y) &\leq (1-t)\, d(u_x, x) + t\, d(u_x, z) \\
&\leq (1-t)\, d(u_x, x) + t\, d(u_z, z) + t\, d(u_x, u_z) \\
d(u_z, y) &\leq (1-t)\, d(u_z, x) + t\, d(u_z, z) \\
&\leq (1-t)\, d(u_x, x) + t\, d(u_z, z) + (1-t)\, d(u_x, u_z)\ .
\end{aligned}$$

Therefore

$$\begin{aligned}
W_1(Q, P_t) &\leq \mathbb{E}_\pi[d(u, y)] \\
&= \mathbb{E}_\pi[(1-t) d(u_x, y) + t d(u_z, y)] \\
&\leq \mathbb{E}_\pi[(1-t)\, d(u_x, x) + t\, d(u_z, z) + 2t(1-t)\, d(u_x, u_z)] \\
&= (1-t)\, W_1(Q, P_0) + t\, W_1(Q, P_1) + 2t(1-t)\, \mathbb{E}_\pi[d(u_x, u_z)]\ .
\end{aligned}$$

For any $u_0 \in \mathcal{X}$, the constant K in the last term can then be coarsely bounded with

$$\begin{aligned}
K(Q, P_0, P_1) &= \mathbb{E}_\pi[d(u_x, u_z)] \\
&\leq \mathbb{E}_\pi[d(u_x, u_0)] + \mathbb{E}_\pi[d(u_0, u_z)] = 2\mathbb{E}_{u\sim Q}[d(u, u_0)]\ .
\end{aligned}$$

Taking the minimum over u_0 gives the final result. ∎

When the optimal transport plan from P_0 to P_1 specifies that a grain of probability must be transported from x to z, its optimal coupling counterpart in Q moves from u_x

to u_z. Therefore the quantity $K(Q, P_0, P_1)$ quantifies how much the transport plan from P_t to Q changes when P_t moves along the geodesic. This idea could be used to define a Lipschitz-like property such as

$$\forall P_0, P_1 \in \mathcal{F}_L \subset \mathcal{F} \qquad K(Q, P_0, P_1) \leq L W_1(P_0, P_1)\ .$$

Clearly such a property does not hold when the transport plan changes very suddenly. This only happens in the vicinity of distributions P_t that can be coupled with Q using multiple transport plans.

Unfortunately we have not found an elegant way to leverage this idea into a global description of the cost landscape. Proposition 6.8 merely bounds $K(Q, P_0, P_1)$ by the expected diameter of the distribution Q. We can nevertheless use this bound to describe some level sets of $W_1(Q, P_\theta)$

Theorem 6.9. *Let $\mathcal{X}$ be a strictly intrinsic Polish space equipped with a geodesically convex distance d and let $\mathcal{P}^1_{\mathcal{X}}$ be equipped with the 1-Wasserstein distance W_1. Let $\mathcal{F} \subset \mathcal{P}^1_{\mathcal{X}}$ be displacement convex and let $Q \in \mathcal{P}^1_{\mathcal{X}}$ have expected diameter*

$$D = 2 \min_{u_0 \in \mathcal{X}} \mathbb{E}_{u\sim Q}[d(u, u_0)]\ .$$

Then the level set $L(Q, \mathcal{F}, M) = \{P_\theta \in \mathcal{F} : W_1(Q, P_\theta) \leq M\}$ is connected if

$$M > \inf_{P_\theta \in \mathcal{F}} W_1(Q, P_\theta) + 2D\ .$$

Proof Choose $P_1 \in \mathcal{F}$ such that $W_1(Q, P_1) < M - 2D$. For any $P_0, P_0' \in L(Q, \mathcal{F}, M)$, let $t \in [0, 1] \mapsto P_t \in \mathcal{F}$ be a displacement geodesic joining P_0 and P_1 without leaving $\mathcal{F}$. Thanks to Proposition 6.8,

$$W_1(Q, P_t) \leq (1{-}t)\, M + t\, (M - 2D) + 2t(1 - t)D = M - 2t^2 D \leq M\ .$$

Therefore this displacement geodesic is contained in $L(Q, \mathcal{F}, M)$ and joins P_0 to P_1. We can similarly construct a second displacement geodesic that joins P_0' to P_1 without leaving $L(Q, \mathcal{F}, M)$. Therefore there is a continuous path connecting P_0 to P_0' without leaving $L(Q, \mathcal{F}, M)$. ∎

This result means that optimizing the Wasserstein distance with a descent algorithm will not stop before finding a generative model $P \in \mathcal{F}$ whose distance $W_1(Q, P)$ to the target distribution is within $2D$ of the global minimum. Beyond that point, the algorithm could meet local minima and stop progressing. Because we use a rather coarse bound on the constant $K(Q, P_0, P_1)$, we believe that it is possible to give much better suboptimality guarantee in particular cases.

Note that this result does not depend on the parametrization of G_θ and therefore applies to the level sets of potentially very nonconvex neural network parametrizations. Previous results on the connexity of such level sets [6, 19] are very tied to a specific parametric form. The fact that we can give such a result in an abstract setup is rather surprising. We hope that further developments will clarify how much our approach can help these efforts.

Finally, comparing this result with Example 3.5 also reveals a fascinating possibility: a simple descent algorithm might in fact be unable to find that the Dirac distribution at the center of the sphere is a global minimum. Therefore the effective statistical performance of the learning process may be subtantially better than what Theorem 3.4 suggests. Further research is necessary to check whether such a phenomenon occurs in practice.

7 Conclusion

This work illustrates how the geometrical study of probability distances provides useful —but still incomplete— insights on the practical performance of implicit modeling approaches using different distances. In addition, using a technique that differs substantially from previous works, we also obtain surprising global optimization results that remain valid when the parametrization is nonconvex.

Acknowledgments

We would like to thank Joan Bruna, Marco Cuturi, Arthur Gretton, Yann Ollivier, and Arthur Szlam for stimulating discussions and also for pointing out numerous related works.

References

[1] M. A. Aizerman, É. M. Braverman, and L. I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Society, 2007.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34nd International Conference on Machine Learning, ICML 2017, Sydney, Australia, 7-9 August, 2017*, 2017.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017.

[6] Antonio Auffinger and Gérard Ben Arous. Complexity of random smooth functions of many variables. *Annals of Probability*, 41(6):4214–4247, 2013.

[7] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

[8] Patrizia Berti, Luca Pratelli, Pietro Rigo, et al. Gluing lemmas and skorohod representations. *Electronic Communications in Probability*, 20, 2015.

[9] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[10] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *CoRR*, abs/1606.04838, 2016.

[11] Diane Bouchacourt, Pawan K Mudigonda, and Sebastian Nowozin. DISCO nets: DISsimilarity COefficients Networks. In *Advances in Neural Information Processing Systems 29*, pages 352–360, 2016.

[12] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*. volume 33 of AMS Graduate Studies in Mathematics. American Mathematical Society, 2001.

[13] Edward Challis and David Barber. Affine independent variational inference. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2186–2194. Curran Associates, Inc., 2012.

[14] Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[15] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1486–1494. Curran Associates, Inc., 2015.

[16] Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: approximation by empirical measures. *Annales de l'I.H.P. Probabilités et statistiques*, 49(4):1183–1203, 2013.

[17] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015*, pages 258–267, 2015.

[18] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015.

[19] C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

[20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[21] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

[23] J. M. Hammersley. The distribution of distance in a hypersphere. *The Annals of Mathematical Statistics*, 21(3):447–452, 1950.

[24] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Second Edition*. Springer Series in Statistics. Springer Verlag, New York, 2009.

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[26] Aleksandr Y. Khinchin. Sur la loi des grandes nombres. *Comptes Rendus de l'Académie des Sciences*, 1929.

[27] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[28] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

[29] Vijay R Konda and John N Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *Annals of applied probability*, pages 796–819, 2004.

[30] Tejas D Kulkarni, Pushmeet Kohli, Joshua B Tenenbaum, and Vikash Mansinghka. Picture: A probabilistic programming language for scene perception. In *Proceedings of the IEEE conference on Computer Vision And Pattern Recognition, CVPR 2015*, pages 4390–4399, 2015.

[31] Mun Wai Lee and Ramakant Nevatia. Dynamic human pose estimation using Markov Chain Monte Carlo approach. In *7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION 2005)*, pages 168–175, 2005.

[32] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *arXiv preprint arXiv:1705.08584*, 2017.

[33] Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1718–1727, 2015.

[34] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. *arXiv preprint arXiv:1705.08991*, 2017. to appear in NIPS 2017.

[35] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

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

[37] Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, Apr 2001.

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

[39] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29*, pages 271–279. 2016.

[40] Svetlozar T Rachev, Lev Klebanov, Stoyan V Stoyanov, and Frank Fabozzi. *The methods of distances in the theory of probability and statistics*. Springer, 2013.

[41] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[42] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, pages 1278–1286, 2014.

[43] Lukasz Romaszko, Christopher KI Williams, Pol Moreno, and Pushmeet Kohli. Vision-as-inverse-graphics: Obtaining a rich 3d explanation of a scene from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 851–859, 2017.

[44] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems 29*, pages 2234–2242, 2016.

[45] Isaac J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44:522–536, 1938.

[46] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[47] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[48] Robert J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, New York; Chichester, 1980.

[49] Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 08 2016.

[50] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[51] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[52] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[53] J. Gábor Székely. E-statistics: The energy of statistical samples. Technical Report 02-16, Bowling Green State University, Department of Mathematics and Statistics, 2002.

[54] Lucas Theis, Aaron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016.

[55] Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[56] Richard von Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

[57] A. A. Zinger, Ashot V. Kakosyan, and Lev B. Klebanov. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics*, 4(59):914–920, 1992. Translated from *Problemy Ustoichivosti Stokhasticheskikh Modelei-Trudi seminara*, 1989, pp 47-55.