Optimal Transport Kernel Embedding for Sets

Ambroise Odonnat
Master MVA
ENS Paris-Saclay
ambroise.odonnat@ens-paris-saclay.fr

Abstract

In scientific fields such as bioinformatics and natural language processing, sequences of data can be represented by sets of local features. Those collections of objects are delicate to work with due to varying size, long-range dependencies and possibly few labeled data. Hence, we are reviewing the paper - written by Mialon et al. - presenting a trainable optimal transport kernel embedding for sets of features. This representation embeds and aggregates elements of a set following the optimal transport plan between the input set and a trainable reference. The authors show the effectiveness of their method on biological and natural language sequences. However, their paper lacks a study on the impact of the number of points in reference. In our work, we show that the error on the 2-Wasserstein distance approximation obtained via the proposed embedding decreases with the number of points in reference. We propose in the same fashion an approximation of the p-Wasserstein distance and demonstrate the effectiveness of our method as a proxy for the 1-Wasserstein distance-based loss function in Wasserstein Generative Adversarial Networks. Finally, we illustrate the benefits and robustness of the optimal transport kernel embedding on a logistic classification task with synthetic distributions. We provide an open-source implementation of our experiments at https://github.com/AmbroiseOdonnat/OTKE.

1 Introduction

Machine Learning on sets has attracted a lot of attention in the recent years with the development of bioinformatics and natural language processing (NLP). Most conventional models expect fixed dimensional data instances and fail to handle sets. Learning algorithms on sets also need to be permutation-invariant. That is to say they must produce identical representations for any permutation of the elements of the input set. See Appendix A.1 for a formal definition.

Various deep learning architectures specifically designed for sets have been proposed recently. In [33], Zaheer et al. provide a family of functions to which any permutation invariant objective function must belong and present necessary and sufficient conditions to obtain permutation invariant deep neural networks. In [23], Qi et al. propose an efficient permutation-invariant deep network for vision tasks such as object classification, part segmentation, or scene semantic parsing. More recent approaches [16, 28] have shown great results on NLP tasks. Some efforts have been made to apply attention, a mechanism for feature aggregation introduced in [30], to sequence modelling tasks in NLP [6] and bioinformatics [25].

Several kernel methods for sets have been studied with mostly two lines of research. The first one focuses on match kernels, which rely on the comparison of all pairs of features between two sets via a similarity function [29, 17]. The second one matches features using the Wasserstein distance. Few of those Wasserstein distance-based kernels are positive definite and fast to compute [24, 15] and most of them do not enable trainable representations. Trainable kernel embeddings have also been proposed for biological sequences by [2, 3].

Preprint. Under review.

In [18], Mialon et al. go beyond sequence modelling and address the problem of learning on sets of features with positional information such as biological sequences or sentences. Their work can be seen as an extension of earlier approaches, more adapted to long sequences in practise and using optimal transport plan to aggregate local features. Driven by the need to perform pooling operations on long sequences of varying size, with long-range dependencies and scarce training data, the authors propose a trainable embedding which can be applied directly to the data or used in combination with existing deep architectures. Their Optimal Transport Kernel Embedding (OTKE) is in the line of the kernel learning theory [27, 12] and the optimal transport (OT) theory [22]. The proposed method embeds feature vectors of a given set to a reproducing kernel Hilbert space (RKHS) before performing a weighted pooling operation. The weights are given by the optimal transport plan between the input set and a trainable reference. This embedding is related to the attention mechanism and can be seen as a surrogate of the Earth Mover's distance-based kernel embedding.

The scalability of their method is mostly due to the use of effective kernel approximation techniques [32] to obtain a finite-dimensional embedding and fast OT algorithms [4] to compute the transport plan. Mialon et al. provide the OTKE, a fixed-size kernel embedding whose parameters can be learned with or without supervision. The proposed method is effective and adaptive as the trainable reference can be optimized with respect to a given task. The authors demonstrate the benefits of their method on biological sequence classification tasks, with state-of-the-art results on protein fold recognition and detection of chromatin profiles. They also show promising performance on NLP tasks.

The ICLR review of the paper points out that a study on the impact of the number of reference points is missing in their work. Motivated by this, we study the quality of the 2-Wasserstein distance approximation proposed by the authors when the number of points in reference increases. We propose in the same fashion an approximation of the p-Wasserstein distance and study its effectiveness by approximating the 1-Wasserstein distance-based loss in Wasserstein GAN [1, 10, 7]. We also use the OTKE in a simple logistic classification task.

Contributions Our contributions are three folds. In this paper, (i) we evaluate the quality of the Wasserstein distance approximation obtained via the proposed embedding. We show on 1D and 2D Gaussian distributions that the approximation error decreases with the number of points in reference. (ii) We propose an approximation of the p-Wasserstein distance and show that our 1-Wasserstein distance approximation can be used as a loss function in Wasserstein GAN. (iii) Finally, we display the benefits and robustness of the optimal transport kernel embedding on a logistic classification task with 2D Gaussian distributions.

2 Background

In this section, we introduce the basics of optimal transport, kernel methods and Wasserstein GAN.

2.1 Optimal Transport

Optimal Transport [31, 22] has to do with the problem of comparing two probability distributions and aims at finding the lowest cost to transfer the mass from one probability to another. In this paper, we will focus on the optimal transport plan between $\mathbf{x}$ and $\mathbf{y}$ seen as weighted point clouds or discrete, sampled measures on a Euclidean feature space $\mathcal{X} \subset \mathbb{R}^d$. The notions we introduce are detailed in [22].

Notations We denote $\mathcal{M}(\mathcal{X})$ the set of Radon measures on the space $\mathcal{X}$. $\mathcal{M}_+(\mathcal{X})$ is the set of all positive measures on $\mathcal{X}$, while $\mathcal{M}_+^1(\mathcal{X})$ is the set of all probability measures on $\mathcal{X}$. We denote $\Sigma_n = \{\mathbf{a} \in \mathbb{R}_+^n | \sum_i \mathbf{a}_i = 1\}$ the probability simplex and δ_x the Dirac distribution at position x. We consider two discrete measures $\alpha = \sum_i \mathbf{a}_i \delta_{\mathbf{x}_i}, \beta = \sum_i \mathbf{b}_i \delta_{\mathbf{y}_i}$ with respective weights $\mathbf{a} \in \Sigma_n$, $\mathbf{b} \in \Sigma_m$ and locations $\mathbf{x}, \mathbf{y}$.

Definition 1. For a continuous map $T : \mathcal{X} \mapsto \mathcal{Y}$, the push-forward measure $\beta = T_{\#}\alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ satisfies for any measurable set $B \subset \mathcal{Y}$:

$$\beta(B) = \alpha(\{x \in \mathcal{X} : T(x) \in B\}) = \alpha(T^{-1}(B))$$

For discrete measures, the push-forward operation defined in Def. 1 consists in moving the positions of all the points in the support of the measure $T_{\#}\alpha \coloneqq \sum_i \mathbf{a}_i \delta_{T(\mathbf{x}_i)}$.

Monge Problem The classical problem of optimal transportation mass between positions $\mathbf{x}$ and $\mathbf{y}$ introduced by Monge in [19] is parameterized by a cost function $c(x, y)$ for points in $\mathcal{X} \times \mathcal{Y}$ and consists in solving the following equation:

$$\min_{\mathcal{T}} \left\{ \sum_i c(\mathbf{x}_i, \mathcal{T}(\mathbf{x}_i)) : T_{\#}\alpha = \beta \right\} \tag{1}$$

Kantorovich Formulation The main issue with the Monge formulation in Eq. 1 is that such maps may not exist between two discrete measures. It is typically the case when the target measure has more points than the source, i.e. $n < m$. Kantorovich proposed a relaxation of the Monge Problem in [13], allowing mass splitting from a source toward several targets. Let $\mathbf{C} \in \mathbb{R}_+^{n \times m}$ be the cost matrix representing the pairwise costs for aligning the elements of $\mathbf{x}$ and $\mathbf{y}$ i.e $\mathbf{C}_{ij} \coloneqq c(\mathbf{x}_i, \mathbf{y}_j)$. Instead of a map $\mathcal{T}$, we are looking for a coupling $\mathbf{P} \in \mathbb{R}_+^{n \times m}$ where $\mathbf{P}_{ij}$ is the amount of mass flowing from bin i to bin j. The space of admissible coupling is:

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \coloneqq \{\mathbf{P} \in \mathbb{R}_+^{n \times m} : \mathbf{P}\mathbb{1}_m = \mathbf{a} \text{ and } \mathbf{P}^{\mathrm{T}}\mathbb{1}_n = \mathbf{b}\},$$

and the Kantorovich optimal transport problem from $\mathbf{x}$ to $\mathbf{y}$ reads:

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \sum_{ij} \mathbf{C}_{ij}\mathbf{P}_{ij} \tag{2}$$

Entropic Regularization Numerical schemes exist to efficiently find approximate solution of the Kantorovich formulation. The main idea is to add an entropic regularization penalty on the minimization problem in Eq. 2.

Definition 2. The discrete entropy of a coupling matrix $\mathbf{P}$ is defined as follows:

$$\mathbf{H}(\mathbf{P}) \coloneqq -\sum_{ij} \mathbf{P}_{ij}(\log(\mathbf{P}_{ij}) - 1)$$

The Kantorovich relaxation with entropic regularization, where ε controls the sparsity of $\mathbf{P}$, writes:

$$\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \sum_{ij} \mathbf{C}_{ij}\mathbf{P}_{ij} - \varepsilon \mathbf{H}(\mathbf{P}) \tag{3}$$

This problem can be solved with a matrix scaling algorithm called the Sinkhorn algorithm [14]. Cuturi et al. showed in [4] that the Sinkhorn's algorithm enables a parallel and GPU friendly computation. See Appendix A.2 for more details on the Sinkhorn algorithm. The optimal coupling $\mathbf{P}$ is called the transport plan and gives information on how to transport mass of $\mathbf{x}$ on $\mathbf{y}$ with minimal cost. The method proposed in [18] uses the optimal transport plan to align features from a given set $\mathbf{x}$ to a trainable reference $\mathbf{z}$. In this case, the authors consider the mass to be evenly distributed between the point and the weights $\mathbf{a}$ and $\mathbf{b}$ are supposed uniform.

Wasserstein Distance Let $p \geq 1$. We consider the case when $c(x, y) = ||x - y||^p$ with ground metric $|| \cdot ||$, the euclidean distance on $\mathbb{R}^d$. All the definitions below can be easily adapted to another ground metric d on $\mathbb{R}^d$. Formally, the p-Wasserstein distance between two probability measures $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$ can be expressed as:

$$W_p(\alpha, \beta) \coloneqq \left(\min_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} ||x - y||^p \, d\pi(x, y) \right)^{\frac{1}{p}} \tag{4}$$

with $\mathcal{U}(\alpha, \beta) \coloneqq \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : P_{\mathcal{X}\#}\pi = \alpha \text{ and } P_{\mathcal{Y}\#}\pi = \beta\}$, where $P_{\mathcal{X}\#}$ and $P_{\mathcal{Y}\#}$ are the push-forward operators (Def. 1) of the projection $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$.

For discrete measures, the cost matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$ representing the pairwise costs of aligning $\mathbf{x}$ and $\mathbf{y}$ verifies $\mathbf{C}_{ij} = ||x_i - y_j||^p$ and the p-Wasserstein distance between two discrete measures α, β of weights $\mathbf{a} \in \Sigma_n, \mathbf{b} \in \Sigma_m$ and locations $\mathbf{x}, \mathbf{y}$ writes:

$$W_p(\alpha, \beta) \coloneqq (\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \sum_{ij} \mathbf{C}_{ij} \mathbf{P}_{ij})^{\frac{1}{p}} = (\min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle)^{\frac{1}{p}} \tag{5}$$

Remark 1. When $p = 1$ in Eq. 4 (respectively Eq. 5 in the discrete case), we retrieve the Earth Mover's distance [26].

Remark 2. Monge and Kantorovich problems with $c(x, y) = ||x - y||^2$ are respectively equivalent to Monge and Kantorovich problems with $c(x, y) = -\langle x, y \rangle$. This equivalence is still true when $|| \cdot ||$ is replaced by another ground metric d on $\mathbb{R}^d$. In particular, for a positive definite kernel κ and an induced distance d_κ, one can show in the same fashion that Monge and Kantorovich problems with $c(x, y) = d_\kappa(x, y)^2$ are respectively equivalent to Monge and Kantorovich problems with $c(x, y) = -\kappa(x, y)$. The same result stands for discrete measure with cost matrices $\mathbf{C}$.

Proof. For $\pi \in \mathcal{U}(\alpha, \beta)$,

$$\begin{aligned}\int_{\mathcal{X} \times \mathcal{Y}} ||x - y||^2 \, d\pi(x, y) &= \int_{\mathcal{X} \times \mathcal{Y}} ||x||^2 \, d\pi(x, y) + \int_{\mathcal{X} \times \mathcal{Y}} ||y||^2 \, d\pi(x, y) + 2 \int_{\mathcal{X} \times \mathcal{Y}} -\langle x, y \rangle \, d\pi(x, y) \\ &= \underbrace{\int_{\mathcal{X}} ||x||^2 \, d\alpha(x)}_{\text{cst}} + \underbrace{\int_{\mathcal{Y}} ||y||^2 \, d\beta(y)}_{\text{cst}} + 2 \int_{\mathcal{X} \times \mathcal{Y}} -\langle x, y \rangle \, d\pi(x, y)\end{aligned}$$

□

The embedding proposed in [18] can be seen as a surrogate of the Earth Mover's Distance-based kernel embedding [26] and is related to the 2-Wasserstein distance. In this paper, we will study the quality of this surrogate when the number of points in reference increases.

2.2 Kernel Methods

Kernel methods [32] consists in mapping data $x \in \mathcal{X}$ to a high dimensional reproducing kernel Hilbert space $\mathcal{H}$ (RKHS) with scalar product $\langle \cdot, \cdot \rangle$ through a kernel mapping $\varphi : \mathcal{X} \mapsto \mathcal{H}$ associated to a positive definite kernel K verifying $K(x, y) = \langle \varphi(x), \varphi(y) \rangle$. Even though the obtained embedding $\varphi(x)$ might be infinite-dimensional, Williams & Seeger developed in [32] kernel approximation techniques to obtain finite-dimensional embeddings $\psi(x) \in \mathbb{R}^k$ such that $K(x, y) \approx \langle \psi(x), \psi(y) \rangle$. The Optimal Transport Kernel Embedding [18] we study in this paper is in line with the kernel learning theory and relies on kernel approximations methods.

2.3 Wasserstein GAN

The Wasserstein GAN proposed by Arjovsky et al. in [1] is a variant of the generative adversarial networks (GAN) proposed by Goodfellow et al. [11]. As explained by the authors, the original GAN is made of two networks, the generator G that captures the data distribution and the discriminator D that estimates the probability that a sample came from the training data rather than G.

Instead of using a discriminator to predict the probability that a generated image is fake or real, WGAN relies on a critic that measures the realness or fakeness of a generated image. The purpose of the critic is to minimize the 1-Wasserstein distance, also called the Earth Mover's Distance (Rem. 1), between the distribution of generated images and the true data distribution μ_d.

We denote the generator G_θ, the true data distribution μ_d and we consider the case where images are generated from a random distribution μ_n (typically a uniform or a Gaussian measure in a low-dimensional space Z). If we consider the generated samples $x_g = G_\theta(z)$, with $z \sim \mu_n$, the measure μ_θ from which x_g is drawn is $G_{\theta\#}\mu_n$ [10, 22]. Hence, the WGAN proposed by Arjovsky et al. aims at solving is the following optimization problem:

$$\min_{\theta} \quad W_1(\mu_d, G_{\theta\#}\mu_n) \tag{6}$$

In practice, the full distribution μ_d is only accessible via samples and the Wasserstein distance is intractable for large datasets. Arjovsky et al. proposed in [1] an approximation of the dual potential of 1-Wasserstein using a neural network to solve the optimization problem. This implementation enables to "improve the stability of learning, get rid of problems like mode collapse, and provide meaningful learning curves useful for debugging and hyperparameter searches" [1]. In particular, the training is more stable than the Vanilla GAN, especially when the generator learning is done in high dimension. The pseudo-code of the WGAN can be seen in Alg. 1.

In this paper, we will study the effectiveness of the WGAN model when the Wasserstein distance-based loss is replaced by a distance obtained via the embedding proposed in [18]. We will rely on the optimization of the expectation of the Wasserstein distance over minibatches proposed and studied in [10, 7].

Algorithm 1: WGAN proposed in [1] with: $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$

Require: α, the learning rate. c, the clipping parameter. m, the batch size.
n_{critic}, the number of iterations of the critic per generator iteration.
Requires: w_0, initial critic parameters. θ_0, initial generator's parameters.
while θ *has not converge* **do**
  for $t = 0, ..., n_{critic}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch from the prior samples.
    $\mathbf{g}_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)} - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch from the prior samples.
  $\mathbf{g}_\theta \leftarrow -\nabla_\theta \left[\frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end

3 Optimal Transport Kernel Embedding

3.1 Preliminaries

We consider sets of elements of $\mathbb{R}^d$ living in

$$\mathcal{X} \coloneqq \{\mathbf{x} = \{x_1, .., x_n\} | x_1, .., x_n \in \mathbb{R}^d \text{ for some } n \geq 1\}.$$

Several data structures can be represented by elements of $\mathcal{X}$: documents as sets of word embeddings, graphs as sets of nodes embeddings, images as sets of local features and sequences are sets of k-mers. The size n of a set $x \in \mathcal{X}$ can vary which is the reason why conventional machine learning architectures cannot handle sets. The Optimal Transport Kernel Embedding (OTKE) proposed in [18] can take as input a sequence of any type and provide a fixed-size embedding.

3.2 Proposed Embedding

Let us consider an input set $\mathbf{x} \in \mathcal{X}$ and a reference set $\mathbf{z} \in \mathcal{X}$ with p elements. The OTKE proposed in [18] is a pooling mechanism based on the transport plan between $\mathbf{x}$ and $\mathbf{z}$ seen as weighed point clouds or discrete, sampled measures. The weights of the discrete measure associated to the reference $\mathbf{z}$ are supposed uniform and in practise the same assumption applies to the weights of the input set $\mathbf{x}$. The proposed method described in Figure 1 is based on three steps: (i) embed $\mathbf{x}$ and $\mathbf{z}$ to a reproducible kernel Hilbert space (RKHS) $\mathcal{H}$; (ii) align elements of $\mathbf{x}$ and $\mathbf{z}$ via optimal transport; (iii) weighted linear pooling of elements of $\mathbf{x}$ into p bins, producing an embedding $\Phi_{\mathbf{z}}(\mathbf{x}) \in \mathcal{H}^p$. The formal definition of the OTKE proposed in [18] is the following:

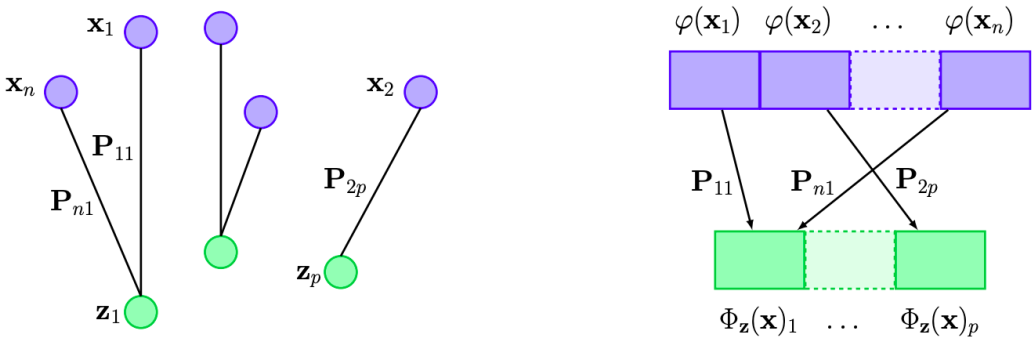


Figure 1: Proposed method to obtain the OTKE [18]

Definition 3. Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathcal{X}$ be an input set of vectors and $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_p\} \in \mathcal{X}$ be a reference set with p elements. Let κ be a positive definite kernel, *e.g.* a linear or Gaussian kernel, with RKHS $\mathcal{H}$ and associated embedding $\varphi : \mathcal{X} \mapsto \mathcal{H}$. $\kappa \in \mathbb{R}^{n \times p}$ carries the comparison between $\mathbf{x}$ and $\mathbf{z}$, before alignment. We denote $\mathbf{P}(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^{n \times p}$ the optimal transport plan between $\mathbf{x}$ and $\mathbf{z}$, solution of Eq. 3 when $\mathbf{C} = -\kappa$. We write $\varphi(\mathbf{x}) = [\varphi(\mathbf{x}_1), ..., \varphi(\mathbf{x}_n)]^{\mathrm{T}} \in \mathbb{R}^n$. The OTKE $\Phi_{\mathbf{z}}(\mathbf{x})$ is defined as follows:

$$\Phi_{\mathbf{z}}(\mathbf{x}) \coloneqq \sqrt{p} \left(\sum_{i=1}^{n} \mathbf{P}(\mathbf{x}, \mathbf{z})_{i1} \varphi(\mathbf{x}_i), ..., \sum_{i=1}^{n} \mathbf{P}(\mathbf{x}, \mathbf{z})_{ip} \varphi(\mathbf{x}_i) \right) = \sqrt{p} \times \mathbf{P}(\mathbf{x}, \mathbf{z})^{\mathrm{T}} \varphi(\mathbf{x}) \in \mathcal{H}^p \quad (7)$$

The obtained embedding is possibly infinite-dimensional. As we mentioned earlier, several methods exist to approximate kernel embeddings in finite dimension [32]. The reader can find more details on the process to obtain finite-dimensional OTKE in the section 3.3 of [18]. We also let the reader find details on the supervised and unsupervised learning of the reference $\mathbf{z}$ in [18].

3.3 Kernel Interpretation

Let us consider two sets $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ of size n and m and a reference $\mathbf{z} \in \mathcal{X}$ of size p. We introduce $\mathbf{P}_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \coloneqq p \times \mathbf{P}(\mathbf{x}, \mathbf{z}) \mathbf{P}(\mathbf{y}, \mathbf{z})^{\mathrm{T}} \in \mathbb{R}^{n \times m}$. The OTKE can be naturally associated to the positive definite kernel $K_{\mathbf{z}}$ with RKHS $\mathcal{H}^p$ defined as:

$$K_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \coloneqq \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{P}_{\mathbf{z}}(\mathbf{x}, \mathbf{y})_{ij} \kappa(\mathbf{x}_i, \mathbf{y}_j) = \langle \Phi_{\mathbf{z}}(\mathbf{x}), \Phi_{\mathbf{z}}(\mathbf{y}) \rangle \quad (8)$$

Thanks to the gluing lemma [31, 22], $\mathbf{P}_{\mathbf{z}}(\mathbf{x}, \mathbf{y})$ is a valid transport plan and roughly approximates $\mathbf{P}(\mathbf{x}, \mathbf{y})$. Using this approximation property, $K_{\mathbf{z}}$ can be seen as a surrogate of the Earth Mover's Distance-based kernel K_{OT} introduced in [26] and defined as:

$$K_{\mathrm{OT}}(\mathbf{x}, \mathbf{y}) \coloneqq \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{P}(\mathbf{x}, \mathbf{y})_{ij} \kappa(\mathbf{x}_i, \mathbf{y}_j) \quad (9)$$

Peyré and Cuturi showed in [22] that K_{OT} is not positive definite in general. Moreover, if we denote n_{sets} the number of sets to process and cost_K the number of transport plans to compute to obtain a given kernel K, we have $\mathrm{cost}_{K_{\mathrm{OT}}} = \mathcal{O}(n_{\mathrm{sets}}^2)$ as one has to compute all the transport plan $\mathbf{P}(\mathbf{x}, \mathbf{y})$ for all pairs $(\mathbf{x}, \mathbf{y})$. If we take into account the symmetry of $\mathbf{P}(\mathbf{x}, \mathbf{y})$, the number of transport plans to compute is $\frac{n_{\mathrm{sets}}(n_{\mathrm{sets}}-1)}{2}$, which leads to a complexity in $\mathcal{O}(n_{\mathrm{sets}}^2)$. On the contrary, $\mathrm{cost}_{K_{\mathbf{z}}} = \mathcal{O}(n_{\mathrm{sets}})$ as one only has to compute $\mathbf{P}(\mathbf{x}, \mathbf{z})$ for each set $\mathbf{x}$ to have $\mathbf{P}_{\mathbf{z}}(\mathbf{x}, \mathbf{y})$ for all pairs $(\mathbf{x}, \mathbf{y})$. In the end, the OTKE provides a computationally efficient surrogate of K_{OT}.

4 Approximating Wasserstein Distance

When ε is equal to 0 in Eq. 3, the kernel K_{OT} is equivalent to the 2-Wasserstein distance with ground metric d_κ, the distance induced by κ. Indeed, in this situation, the associated optimal plan $\mathbf{P}$ verifies

the Kantorovich relaxation (Eq. 2) and we showed in Rem. 2 that the Kantorovich OT with cost d_κ^2 is equivalent to the Kantorovich OT with cost $-\kappa$. It implies the equivalence of the associated objective functions of the two problems, respectively the 2-Wasserstein distance and K_{OT}.

4.1 2-Wasserstein Distance

Motivated by this equivalence, the authors deduce a relation between $\mathbf{P}(\mathbf{x}, \mathbf{z})$ and $\mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{z})$ when ε is equal to 0. It writes in terms of 2-Wasserstein distance as follows:

Lemma 1. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ of size n, m, p, denoting $W_2^\mathbf{z}(\mathbf{x}, \mathbf{y}) = \sqrt{\langle \mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y}), d_\kappa(\mathbf{x}, \mathbf{y})^2 \rangle}$ with d_κ ground metric induced by κ:

$$|W_2(\mathbf{x}, \mathbf{y}) - W_2^\mathbf{z}(\mathbf{x}, \mathbf{y})| \leq 2 \min\{W_2(\mathbf{x}, \mathbf{z}), W_2(\mathbf{y}, \mathbf{z})\} \tag{10}$$

Proof. [18], Appendix B.1

4.2 Extension to p-Wasserstein Distance

In this paper, we extend the work of Mialon et al. and propose $W_p^\mathbf{z}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y}), d_\kappa(\mathbf{x}, \mathbf{y})^p \rangle^{\frac{1}{p}}$ as an approximation of the p-Wasserstein distance $W_p(\mathbf{x}, \mathbf{y})$. We deduce an upper bound on the approximation error inspired from [18].

Proposition 1. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ of size n, m, q, the reference $\mathbf{z}$ having uniform weights, and denoting $W_p^\mathbf{z}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y}), d_\kappa(\mathbf{x}, \mathbf{y})^p \rangle^{\frac{1}{p}}$ with d_κ ground metric induced by κ, we have:

$$|W_p(\mathbf{x}, \mathbf{y}) - W_p^\mathbf{z}(\mathbf{x}, \mathbf{y})| \leq 2 \min\{W_p(\mathbf{x}, \mathbf{z}), W_p(\mathbf{y}, \mathbf{z})\} \tag{11}$$

Proof. To avoid confusion with p associated to the p-Wasserstein distance, we denote q the size of the reference. This proof is inspired from the proof of Lemma 3.1 in [18]. We use the fact that the weights $\mathbf{c}$ of the reference $\mathbf{z}$ are uniform. As $\mathbf{P}(\mathbf{y}, \mathbf{z})$ is solution of Eq. 3, we have $\mathbf{P}(\mathbf{y}, \mathbf{z})^\mathrm{T} \mathbb{1}_m = \mathbf{c}$ which implies $\sum_{j=1}^m q\mathbf{P}(\mathbf{y}, \mathbf{z})_{jk} = 1$ for all $k \in [\![1, q]\!]$ $(\star)$.
Thus,

$$\begin{aligned}
W_p(\mathbf{x}, \mathbf{z})^p &= \sum_{i=1}^n \sum_{k=1}^q \mathbf{P}(\mathbf{x}, \mathbf{z})_{ik} d_\kappa(\mathbf{x}_i, \mathbf{z}_k)^p \\
&= \sum_{i=1}^n \sum_{k=1}^q \sum_{j=1}^m q\mathbf{P}(\mathbf{y}, \mathbf{z})_{jk} \mathbf{P}(\mathbf{x}, \mathbf{z})_{ik} d_\kappa(\mathbf{x}_i, \mathbf{z}_k)^p \text{ using } (\star) \\
&= \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^q q\mathbf{P}(\mathbf{x}, \mathbf{z})_{ik} \mathbf{P}(\mathbf{y}, \mathbf{z})_{jk} d_\kappa(\mathbf{x}_i, \mathbf{z}_k)^p \\
&= ||u||_p^p
\end{aligned}$$

where $u \in \mathbb{R}^{n \times m \times q}$ has entries $(q\mathbf{P}(\mathbf{y}, \mathbf{z})_{jk} \mathbf{P}(\mathbf{x}, \mathbf{z})_{ik})^{\frac{1}{p}} d_\kappa(\mathbf{x}_i, \mathbf{z}_k)$ for $i \in [\![1, n]\!], j \in [\![1, m]\!], k \in [\![1, q]\!]$. Using the definition of $W_p^\mathbf{z}(\mathbf{x}, \mathbf{y})$ and $\mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y})$, we have:

$$\begin{aligned}
W_p^\mathbf{z}(\mathbf{x}, \mathbf{y})^p &= \sum_{i=1}^n \sum_{j=1}^m \mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y})_{ij} d_\kappa(\mathbf{x}_i, \mathbf{y}_j)^p \\
&= \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^q q\mathbf{P}(\mathbf{x}, \mathbf{z})_{ik} \mathbf{P}(\mathbf{y}, \mathbf{z})_{jk} d_\kappa(\mathbf{x}_i, \mathbf{y}_j)^p \\
&= ||v||_p^p
\end{aligned}$$

where $v \in \mathbb{R}^{n \times m \times q}$ has entries $(q\mathbf{P}(\mathbf{x}, \mathbf{z})_{ik} \mathbf{P}(\mathbf{y}, \mathbf{z})_{jk})^{\frac{1}{p}} d_\kappa(\mathbf{x}_i, \mathbf{y}_j))$ for $i \in [\![1, n]\!], j \in [\![1, m]\!], k \in [\![1, q]\!]$. We recall the reverse triangular identity $|||u||_p - ||v||_p| \leq ||u - v||_p$. This inequality also

stands for a metric d on $\mathbb{R}^d$, in particular, $|d(x,z) - d(x,y)| \leq d(z,y)$. See Appendix B.1 for the associated proofs.

$$
\begin{aligned}
|W_p(\mathbf{x},\mathbf{z}) - W_p^{\mathbf{z}}(\mathbf{x},\mathbf{y})| &= |||u||_p - ||v||_p| \\
&\leq ||u - v||_p \\
&= \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} |\left(q\mathbf{P}(\mathbf{x},\mathbf{z})_{ik}\mathbf{P}(\mathbf{y},\mathbf{z})_{jk}\right)^{\frac{1}{p}} \left(d_\kappa(\mathbf{x}_i,\mathbf{z}_k) - d_\kappa(\mathbf{x}_i,\mathbf{y}_j)\right)|^p \right)^{\frac{1}{p}} \\
&= \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} q\mathbf{P}(\mathbf{x},\mathbf{z})_{ik}\mathbf{P}(\mathbf{y},\mathbf{z})_{jk}|d_\kappa(\mathbf{x}_i,\mathbf{z}_k) - d_\kappa(\mathbf{x}_i,\mathbf{y}_j)|^p \right)^{\frac{1}{p}} \\
&\leq \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} q\mathbf{P}(\mathbf{x},\mathbf{z})_{ik}\mathbf{P}(\mathbf{y},\mathbf{z})_{jk} d_\kappa(\mathbf{z}_k,\mathbf{y}_j)^p \right)^{\frac{1}{p}} \\
&\leq \left(\sum_{j=1}^{m}\sum_{k=1}^{q} \mathbf{P}(\mathbf{y},\mathbf{z})_{jk} d_\kappa(\mathbf{y}_j,\mathbf{z}_k)^p \sum_{i=1}^{n} q\mathbf{P}(\mathbf{x},\mathbf{z})_{ik} \right)^{\frac{1}{p}} \text{ using the symmetry of } d_\kappa \\
&= \left(\sum_{j=1}^{m}\sum_{k=1}^{q} \mathbf{P}(\mathbf{y},\mathbf{z})_{jk} d_\kappa(\mathbf{y}_j,\mathbf{z}_k)^p \right)^{\frac{1}{p}} \text{ using } (\star) \text{ on } \mathbf{P}(\mathbf{x},\mathbf{z}) \\
&= W_p(\mathbf{y},\mathbf{z})
\end{aligned}
$$

Finally, we have:

$$
\begin{aligned}
|W_p(\mathbf{x},\mathbf{y}) - W_p^{\mathbf{z}}(\mathbf{x},\mathbf{y})| &\leq |W_p(\mathbf{x},\mathbf{y}) - W_p(\mathbf{x},\mathbf{z})| + |W_p(\mathbf{x},\mathbf{z}) - W_p^{\mathbf{z}}(\mathbf{x},\mathbf{y})| \\
&\leq W_p(\mathbf{y},\mathbf{z}) + W_p(\mathbf{y},\mathbf{z}) \\
&= 2W_p(\mathbf{y},\mathbf{z})
\end{aligned}
$$

As **x** and **y** plays similar roles, we obtain:

$$|W_p(\mathbf{x},\mathbf{y}) - W_p^{\mathbf{z}}(\mathbf{x},\mathbf{y})| \leq 2\min\{W_p(\mathbf{x},\mathbf{z}), W_p(\mathbf{y},\mathbf{z})\}$$

□

Corollary 1. Using this result, we deduce an upper bound on the deviation term between W_p and $W_p^{\mathbf{z}}$ for N samples $(\mathbf{x}^1, ..., \mathbf{x}^N)$ of elements of $\mathcal{X}$. It writes:

$$\mathcal{E}^p \coloneqq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} |W_p(\mathbf{x}^i,\mathbf{x}^j) - W_p^{\mathbf{z}}(\mathbf{x}^i,\mathbf{x}^j)|^p \leq 2^p \cdot \frac{1}{N}\sum_{i=1}^{N} W_p(\mathbf{x}^i,\mathbf{z})^p \tag{12}$$

Proof.

$$
\begin{aligned}
\mathcal{E}^p &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} |W_p(\mathbf{x}^i,\mathbf{x}^j) - W_p^{\mathbf{z}}(\mathbf{x}^i,\mathbf{x}^j)|^p \\
&\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} 2\min\{W_p(\mathbf{x}^i,\mathbf{z}), W_p(\mathbf{x}^j,\mathbf{z})\}^p \text{ using Proposition 1} \\
&\leq \frac{2^p}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} W_p(\mathbf{x}^i,\mathbf{z})^p \text{ as } \min\{W_p(\mathbf{x}^i,\mathbf{z}), W_p(\mathbf{x}^j,\mathbf{z})\} \leq W_p(\mathbf{x}^i,\mathbf{z}) \\
&= 2^p \cdot \frac{1}{N}\sum_{i=1}^{N} W_p(\mathbf{x}^i,\mathbf{z})^p
\end{aligned}
$$

□

5 Numerical Experiments

5.1 Impact of the number of points in reference

We study the quality of the 2-Wasserstein distance approximation when the number of points in reference increases. We conduct experiments on 1D and 2D distributions using the POT library [9] to compute 2-Wasserstein distances and optimal transport plans. We show that the approximation error when the number of points in reference increases.

1D Gaussian distributions We consider two discrete measures $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}, \beta = \sum_{i=1}^{m} \mathbf{b}_i \delta_{\mathbf{y}_i}$ with $n = 100$ and $m = 80$. Locations $\mathbf{x}$ and $\mathbf{y}$ are graduations of $[\![1, 100]\!]$ and $[\![1, 80]\!]$, while weights $\mathbf{a}$ and $\mathbf{b}$ are drawn from $\mathcal{N}(30, 10)$ and $\mathcal{N}(15, 5)$ respectively. The reference z are of size $p \in [\![1, 100]\!]$ with locations drawn from a normal distribution and uniform weights. Those measures are plotted in Figure 2.

Analysis We observe in Figure 3a the comparison between the optimal coupling $\mathbf{P}(\mathbf{x}, \mathbf{y})$ (left) and the approximation $\mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y})$ (right) when $p = 50$. We can see in Figure 3b that the approximation error, the blue curve plotted in semi-log scale, decreases when the size of the reference support increases. The upper bound on the right side of Eq. 10 is plotted in red.

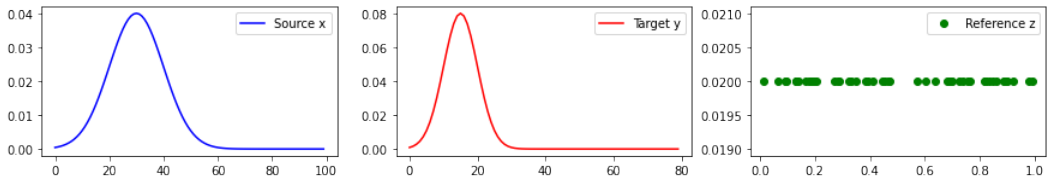
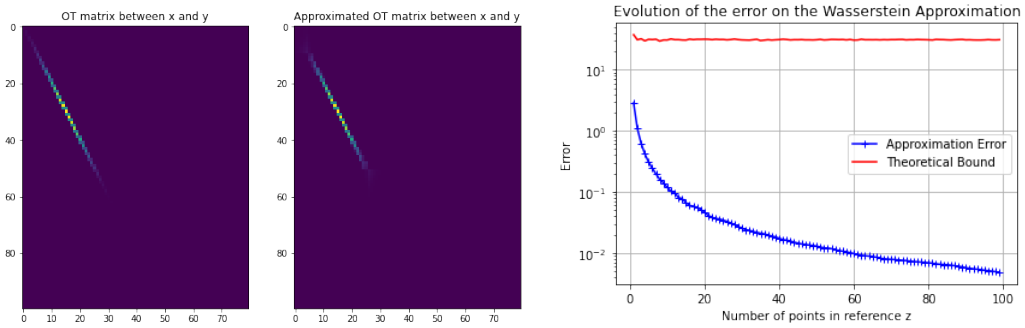


Figure 2: 1D Gaussians $\mathbf{x}$, $\mathbf{y}$, 50 random points as reference $\mathbf{z}$



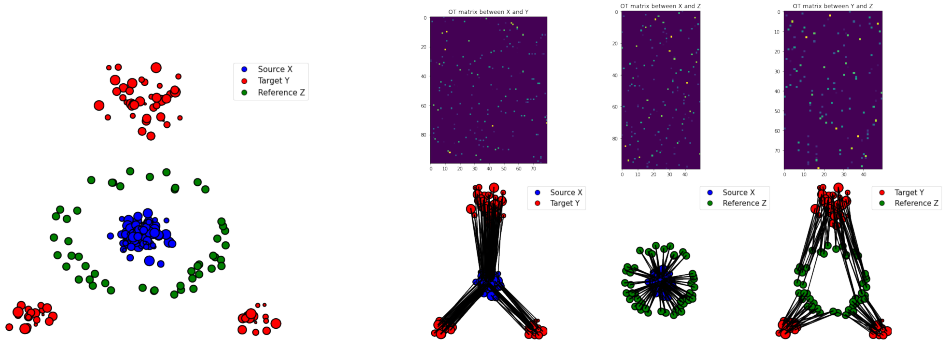
(a) $\mathbf{P}(\mathbf{x}, \mathbf{y})$ (left). $\mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y})$ (right)

(b) Evolution of $|W_2(\mathbf{x}, \mathbf{y}) - W_2^\mathbf{z}(\mathbf{x}, \mathbf{y})|$ (Eq. 10)

Figure 3: Study of the approximation for 1D discrete measures

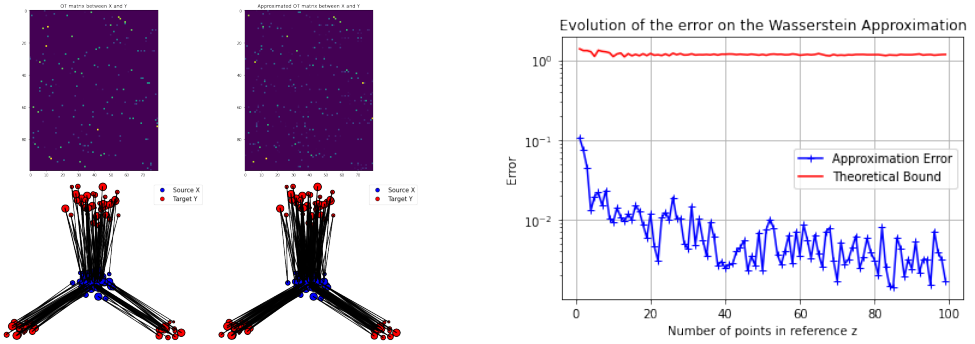
2D Gaussian distributions We consider two discrete measures $\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i}, \beta = \sum_{i=1}^{m} \mathbf{b}_i \delta_{\mathbf{y}_i}$ with $n = 100$ and $m = 80$. Locations $\mathbf{x}$ and $\mathbf{y}$ are random points of the plan, while weights $\mathbf{a}$ and $\mathbf{b}$ are drawn from $\mathcal{U}(100)$ and $\mathcal{U}(80)$ respectively. The reference z are of size $p \in [\![1, 100]\!]$ with locations randomly drawn in an annulus and uniform weights. Those measures are plotted in Figure 4a, while the transport plans between $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are plotted in Figure 4b.

Analysis We observe in Figure 5a the comparison between the optimal coupling $\mathbf{P}(\mathbf{x}, \mathbf{y})$ (left) and the approximation $\mathbf{P}_\mathbf{z}(\mathbf{x}, \mathbf{y})$ (right). We can see in Figure 5b that the approximation error, the blue curve plotted in semi-log scale, decreases when the size of the reference support increases. The upper bound on the right side of Eq. 10 is plotted in red.



(a) 100 random points for $\mathbf{x}$, 80 for $\mathbf{y}$, 50 for reference $\mathbf{z}$ (b) Plot of the transport plans between $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$

Figure 4: Data and transport plans between 2D Gaussians and a random reference $\mathbf{z}$



(a) $\mathbf{P}(\mathbf{x}, \mathbf{y})$ (left). $\mathbf{P}_{\mathbf{z}}(\mathbf{x}, \mathbf{y})$ (right) (b) Evolution of $|W_2(\mathbf{x}, \mathbf{y}) - W_2^{\mathbf{z}}(\mathbf{x}, \mathbf{y})|$ (Eq. 10)

Figure 5: Study of the approximation for 2D discrete measures

5.2 Wasserstein GAN

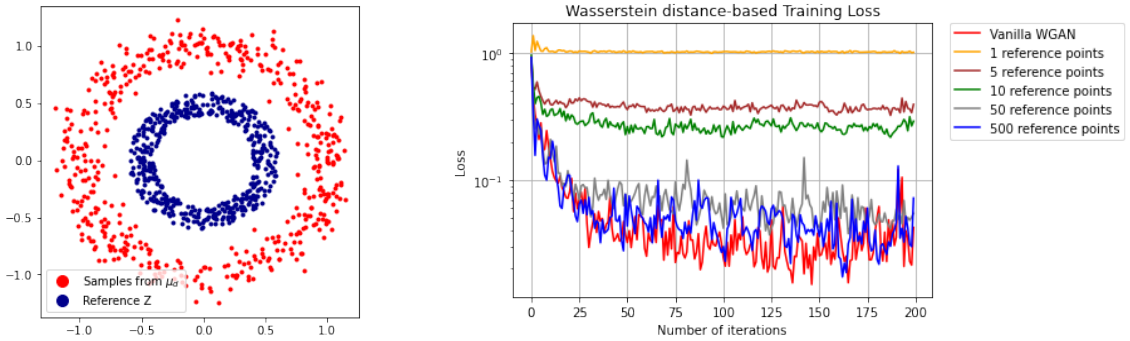
We study the quality of our 1-Wasserstein distance approximation when the number of points in reference increases. In particular, we study the capacity of this approximation to be used as a proxy for the 1-Wasserstein distance-based loss in Wasserstein GAN. The WGAN optimization problem is recalled in Eq. 6. We rely on PyTorch [20] and on the implementation of the vanilla WGAN with minibatches provided by Rémi Flamary in the POT library [8]. We show that the proposed 1-Wasserstein distance approximation can be used as a good proxy for the loss function in Wasserstein GAN.

Method In this experiment, we train a generator G_θ to generate realistic data from a Gaussian distribution μ_n, indistinguishable from the true data distribution μ_d. In minibatches WGAN, we sample batches $\mathbf{x}_d$ from the true distribution μ_d and batches $\mathbf{x}_n$ from a Gaussian distribution μ_n. Generated samples $\mathbf{x}_g = G_\theta(\mathbf{x}_n)$ are thus drawn from $G_{\theta\#}\mu_n$. The loss function is the 1-Wasserstein distance between μ_d and $G_{\theta\#}\mu_n$. These measures can be seen as sampled, discrete measures with locations $\mathbf{x}_d, \mathbf{x}_g$ and uniform weights on $[\![1, \text{batch size}]\!]$. Let us consider a fixed reference $\mathbf{z}$ of size p with uniform weights and random locations. We propose to compare the vanilla WGAN that minimizes the loss $W_1(\mathbf{x}_d, \mathbf{x}_g)$ to our WGAN_Approx that minimizes $W_1^{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{P}_{\mathbf{z}}(\mathbf{x}, \mathbf{y}), d_\kappa(\mathbf{x}, \mathbf{y}) \rangle$. We rely on the POT library [9] to compute W_1 and the optimal couplings.

Implementation details We consider 2D distributions and a batch size of 500. In particular, μ_d has a support of size 500, locations randomly sampled inside an annulus and uniform weights. Locations of the reference $\mathbf{z}$ of size $p \in [\![1, 500]\!]$ are also sampled inside an annulus. μ_n has a support of size 500 with uniform weights and locations drawn from a normal distribution. We train a simple 3-layer MLP as generator G_θ on 200 iterations. A sample from true data x_d and a reference z can be seen in Figure 6a. Examples of generative processes are shown in Appendix C, where we see the evolution

of the generated data across the iterations. The Figure 10 corresponds to the vanilla WGAN and the Figure 11 to our proposed approximation with 500 reference points.

Analysis We compare the training process with different number of points in reference. We can see in Figure 6b that the training loss trend decreases when the number of reference points increases and almost reach the training loss of the vanilla WGAN. We also compare in terms of quality of generated data the impact of the number of reference points. The Figure 7 clearly shows a drastic improvement of the generative model when the number of reference points is higher. It should be noted that only 50 reference points are sufficient to generate realistic samples of size 500. As the training losses show, the best generative model is obtained with 500 reference points.



(a) True distribution μ_d (red) and reference $\mathbf{z}$ (blue)

(b) Evolution of the training loss

Figure 6: Data and training losses

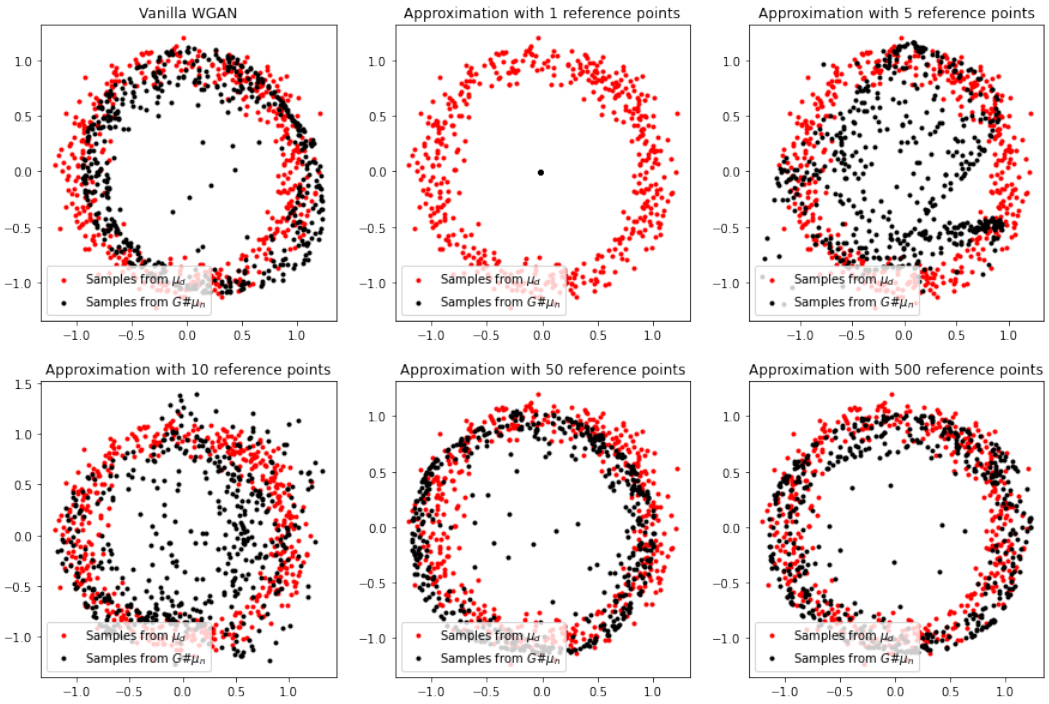


Figure 7: Evolution of generated data with the number of points in reference

5.3 Logistic Classification on 2D Gaussians

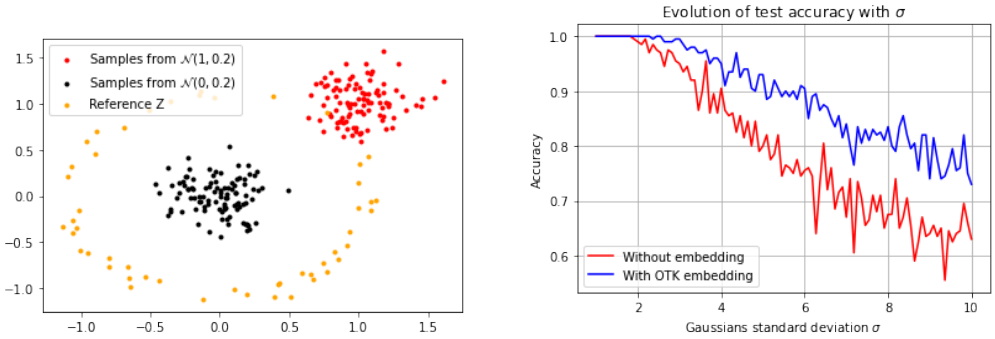
We study the benefits of the OTKE embedding proposed in [18] on a logistic binary classification task with a synthetic dataset. We rely on the scikit-learn library [21] to classify and plot T-SNE

embeddings of test data. We show that embedding the data with OTKE before performing the classification enables better results and first and foremost more robustness.

Method We run a logistic binary classification on 2D Gaussian distributions centered in 0 or 1 with standard deviation σ. The labels are the mean of the Gaussians. The classifier is fitted on training data and predictions are done on test data. We propose to compare the classification accuracy and class separation when we apply classification on raw data and classification on data embedded with the OTKE. We study the effect of the increase of σ in terms of accuracy and class separation with those two methods. In theory, when σ increases, the Gaussian distributions are wider and thus less easily separable.

Implementation details For a fixed value of σ, we build a balanced dataset of 400 Gaussian distributions with 100 data points. 200 of them have label 0 while the others have label 1. In practice, a training sample x lives in $\mathbb{R}^{100\times 2}$. We split the dataset equally in a training set with 200 distributions and a test set with 200 distributions. The split being randomly uniform, the balance between classes is maintained in each set. The OTKE embedding is done using the implementation of Mialon et al. in https://github.com/claying/OTK. We consider a reference $\mathbf{z}$ with 50 points, with uniform weights and locations randomly drawn inside an annulus. Two training samples, one with label 0 (black), one with label 1 (red) and a reference z are plotted in Figure 8a. Raw or embedded data are first flatten and then fed to the logistic classifier. We denote n_{data} the number of data in input. Without embedding, the model takes as input a matrix of $\mathbb{R}^{n_{\text{data}}\times 200}$. With embedding, the model takes as input a matrix of $\mathbb{R}^{n_{\text{data}}\times 100}$ for embedded data are in $\mathbb{R}^{50\times 2}$. The reference $\mathbf{z}$ is the same for all the experiment. The value of ϵ in the Sinkhorn algorithm used to compute optimal transport plans in the OTKE is fixed at σ. This trick enables a stable learning process. σ takes 100 values linearly spaced in $[\![1, 10]\!]$.

Analysis We plot the evolution of the accuracy on test set when the value of σ increases in Figure 8b. We can see that for all values, the embedding provides a better accuracy and that the accuracy gap grows with σ. Using the OTKE enables more robustness to the increase of σ. This property of the OTKE can be visualized in Figure 9. We can see that the separation between classes 0 and 1 is maintained for embedded test data while the class separation collapses as soon ass $\sigma > 3$. This experiment illustrates the benefits of the OTKE embedding in terms of classification and class separation.



(a) 2 training data (black, red) and reference $\mathbf{z}$ (orange) (b) Evolution of accuracy on test when σ increases

Figure 8: Data and accuracy on test set

6 Conclusion and Future Work

The Optimal Transport Kernel embedding (OTKE) proposed by Mialon et al. in [18] is deeply grounded in optimal transport theory [31, 22] and kernel methods [12, 27]. The authors provide an adaptive embedding for sets and achieve state-of-the-art results in protein fold recognition and chromatin profiles detection, as well as promising results in NLP tasks. The OTKE is scalable thanks to fast Sinkhorn algorithm [4] and kernel approximations methods [32]. Moreover, the OTKE can be learned in a supervised and unsupervised fashion which enables adaptability to the problem at hand.

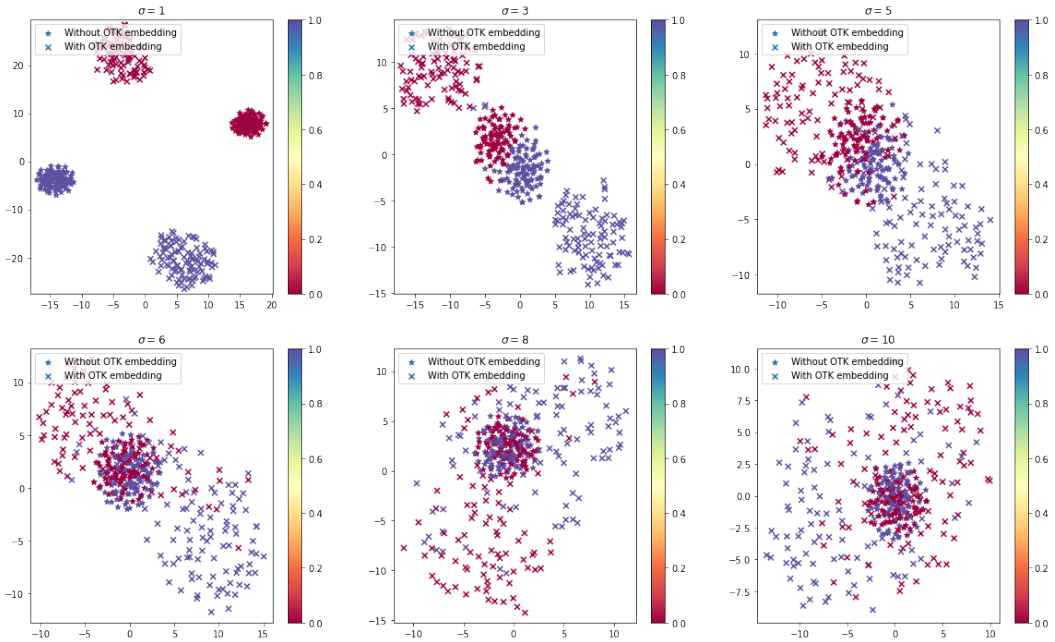


Figure 9: T-SNE Embeddings of test data when σ increases

However, the work of Mialon et al. lacks a study on the impact of number of points in reference. In this paper, we show on simple tasks that the error on the 2-Wasserstein distance approximation provided by OTKE decreases with the number of points in reference. Moreover, we propose an approximation of the p-Wasserstein distance via the OTKE embedding. In the same fashion as [18], we deduce a bound on the error approximation and show that our method can be used as a proxy for the 1-Wasserstein distance-based loss in Wasserstein GAN [1, 10, 7]. Finally, we illustrate the benefits of the OTKE on a logistic classification task. The embedding brings more robustness and better class separation.

A possible line of work would be to study the impact of the reference support size on biological and NLP sequences classification tasks. Another interesting idea would be to investigate the quality of our proposed p-Wasserstein distance approximation, for instance in Wasserstein barycenters problems [5], and the benefits of this approximation in terms of computational efficiency.

7 Connection with the course

The work of Mialon et al. is heavily connected to optimal transport notions introduced in [22]. In particular, it relies on the OT problems (Monge and Kantorovich formulations), on the notion of Wasserstein distance and on the Sinkhorn algorithm. In our work, we also introduced the Wasserstein GAN, linked to minimum Kantorovich estimators mentioned in section 9 of [22].

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, 2017.

[2] Dexiong Chen, Laurent Jacob, and Julien Mairal. Biological sequence modeling with convolutional kernel networks. *Bioinformatics*, 35(18):3294–3302, 02 2019.

[3] Dexiong Chen, Laurent Jacob, and Julien Mairal. Recurrent Kernel Networks. 2019.

[4] Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[5] Marco Cuturi and Arnaud Doucet. Fast Computation of Wasserstein Barycenters. 2013.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

[7] Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein : asymptotic and gradient properties, 2019.

[8] Rémi Flamary. Wasserstein 2 Minibatch GAN with PyTorch. `https://pythonot.github.io/auto_examples/backends/plot_wass2_gan_torch.html#sphx-glr-auto-examples-backends-plot-wass2-gan-torch-py`, 2021.

[9] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[10] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 09–11 Apr 2018.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, 2014.

[12] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008.

[13] L. V. Kantorovich. On the Translocation of Masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, 2006.

[14] Paul Knopp and Richard Sinkhorn. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343 – 348, 1967.

[15] Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein Kernels for Probability Distributions, 2015.

[16] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks, 2018.

[17] Siwei Lyu. Mercer kernels for object recognition with local features. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 223–229 vol. 2, 2005.

[18] Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention, 2020.

[19] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. 2018.

[23] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, 2016.

[24] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein, and Michael M. Bronstein, editors, *Scale Space and Variational Methods in Computer Vision*, pages 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[25] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, 2020.

[26] Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40:99–121, 11 2000.

[27] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[28] Konstantinos Skianis, Giannis Nikolentzos, Stratis Limnios, and Michalis Vazirgiannis. Rep the Set: Neural Networks for Learning Set Representations, 2019.

[29] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To Aggregate or Not to aggregate: Selective Match Kernels for Image Search. In *2013 IEEE International Conference on Computer Vision*, pages 1401–1408, 2013.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017.

[31] Cédric Villani. Optimal transport: Old and new. 2008.

[32] Christopher Williams and Matthias Seeger. Using the Nyström Method to Speed Up Kernel Machines. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[33] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep Sets, 2017.

A Additional Background

This section provides some mathematical background to notions mentioned in the core of the article.

A.1 Permutation Invariance

We consider the case where an input is typically a set $X = \{x_1, .., x_n\},$ with $x_m \in \mathfrak{X}$ and the input domain is the power set $\mathcal{X} = 2^{\mathfrak{X}}$. As stated in [33], learning on sets requires the response of a function to be indifferent from the ordering of the input set.

Definition 4. A function $f : 2^{\mathfrak{X}} \mapsto \mathcal{Y}$ acting on sets must be permutation invariant to the order of objects in the set, i.e. for any permutation $\pi : f(x_1, ..., x_M) = f(x_{\pi(1)}, ..., x_{\pi(M)})$.

A.2 Sinkhorn algorithm: regularized OT as matrix scaling

The Sinkhorn algorithm, introduced by [14], provides a solution to the Kantorovich relaxation with entropic regularization (Eq. 3). Based on the simple structure of the optimal coupling, Cuturi described in [4] a GPU friendly implementation of the sinkhorn. Formally, we denote $\mathbf{a}, \mathbf{x}$ and $\mathbf{b}, \mathbf{y}$ weighs and locations of two discrete measures with support of size n and m respectively. $\mathbf{C}$ is the cost matrix. Then, $\mathbf{P}(\mathbf{x}, \mathbf{y}) = \text{Sinkhorn}(\mathbf{C}, \varepsilon)$ is the optimal transport plan of Eq. 3. The Sinkhorn algorithm is initialized with a arbitrary positive vector $\mathbf{v}^{(0)} = \mathbb{1}_m$ and the iteration l relies on the following update rule:

$$\mathbf{u}^{(l+1)} \coloneqq \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(l)}} \text{ and } \mathbf{v}^{(l+1)} \coloneqq \frac{\mathbf{b}}{\mathbf{K}\mathbf{u}^{(l+1)}} \tag{13}$$

where $\mathbf{K} \coloneqq [e^{-\frac{\mathbf{C}_{ij}}{\varepsilon}}]_{ij}$ is the Gibbs kernel associated to $\mathbf{C}$.

B Proofs

B.1 Reverse triangular inequality

With the p-norm $||\cdot||_p$ Let $u, v \in \mathbb{R}^d$. Using the triangular inequality, we have: $||u||_p = ||u - v + v||_p \leq ||u - v||_p + ||v||_p$, i.e $||u||_p - ||v||_p \leq ||u - v||_p$. As u and v play symmetric roles, we also have: $||v||_p - ||u||_p \leq ||v - u||_p$. It leads $|||u||_p - ||v||_p| \leq ||u - v||_p$.

With a metric d Let $x, y, z \in \mathbb{R}^d$. $d(x, y) \leq d(x, z) + d(z, y)$, i.e. $d(x, y) - d(x, z) \leq d(z, y)$. As y and z play similar roles, we also have $d(x, z) - d(x, y) \leq d(y, z)$. It leads to $|d(x, z) - d(x, y)| \leq d(z, y)$.

C Additional figures

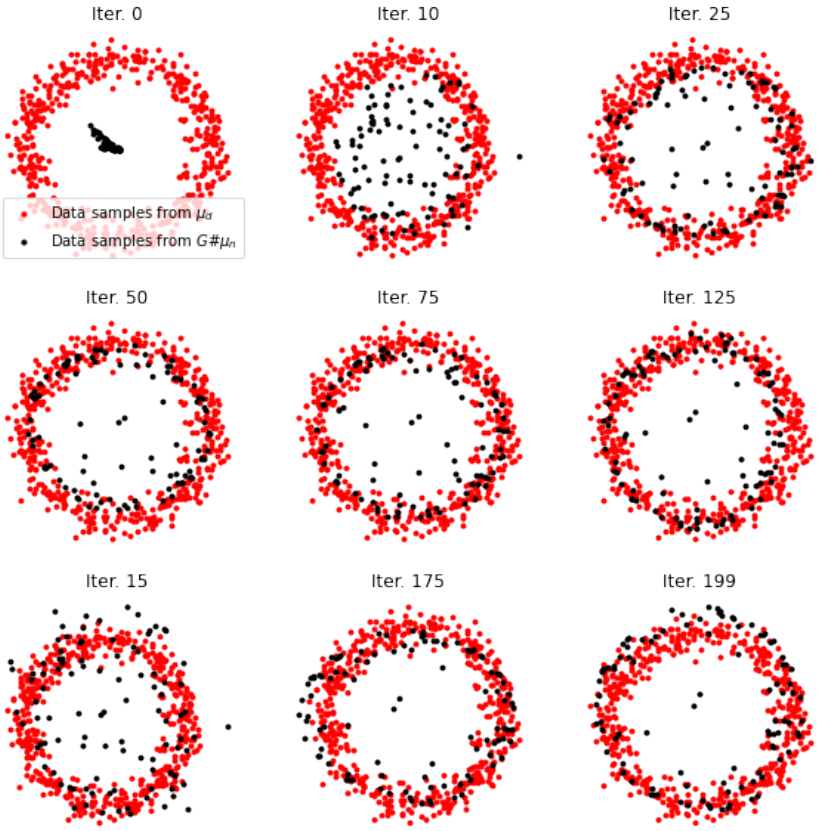


Figure 10: Evolution of the generative process of a vanilla WGAN across iterations

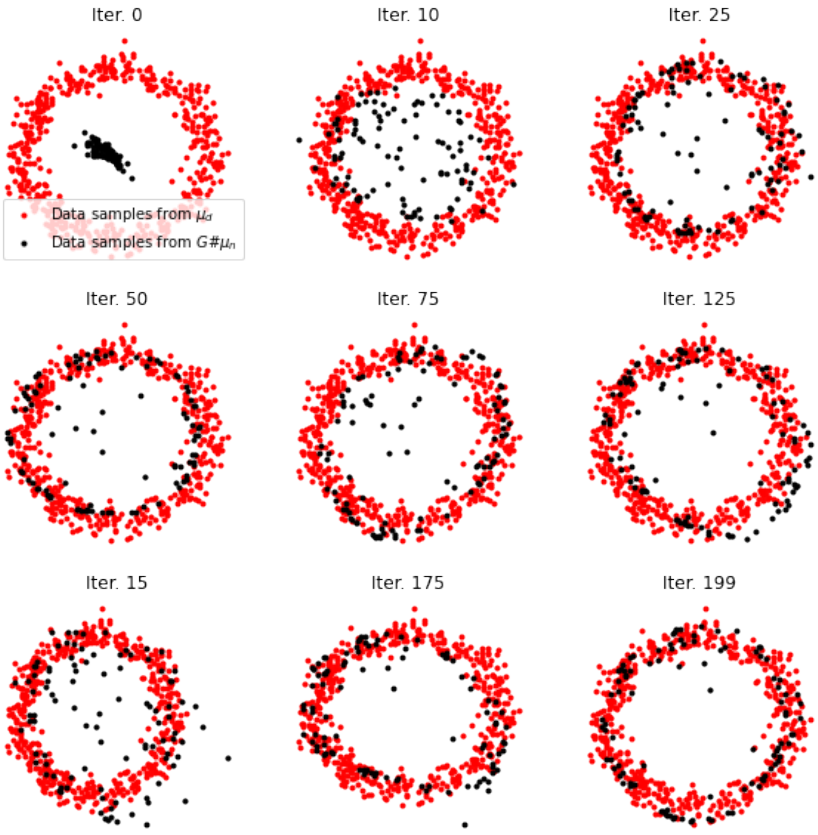


Figure 11: Evolution of the generative process of our WGAN_Approx across iterations