SLICED-WASSERSTEIN FLOWS : NONPARAMETRIC GENERATIVE MODELING VIA OPTIMAL TRANSPORT AND DIFFUSION

MVA - COMPUTATIONAL OPTIMAL TRANSPORT : PROJECT REPORT

Raphaël Barboni
Département Mathématiques et Applications (DMA)
ENS Paris
raphael.barboni@ens.fr

January 8, 2021

ABSTRACT

The studied article [11] proposed an innovative numerical energy minimization method to solve the problem of generative modeling : starting from a referenced distribution, gradient flow of the distribution with respect to the Wasserstein topology is computed to fit the observed data. A regularization term is added so as to allow for generalization. The method strongly relies on the form of the energy functional, referred to as *Sliced-Wasserstein distance* which is the integral over the unit sphere of the optimal transport distance between the projected distributions. Taking advantage of the existence of a simple analytical solution for the problem of 1D optimal transport, both the energy and its gradient are easily computed. Taking the point of view of diffusion processes, the resulting flow is interpreted as the solution of a McKean-Vlasov equation which is discretized in a particle system.

It must be stressed out that the main advantage of the method is that it is fully non-parametric, computing gradient descent in the infinite dimensional space of probability measures whereas neural networks only consider a finite-dimensional space of parametrized measures. In addition to being elegantly interpretable we can thus expect the method to have a better adaptivity than 'black-box' generative networks.

As a part of this project mathematical foundations of the Sliced-Wasserstein flow are studied under the general approach of gradient flow. Numerical properties are compared to other standards methods used in the machine learning literature to address the problem of generative modeling.

1 Introduction

Generative modeling can be defined in a very general manner as the problem of automatically creating synthetic data that are credible. In practice, the notion of credibility is defined by data that are known to be real and assumed to be representative of the distribution p_{data} of the objects we want to generate. Generative modeling then reduces to the problem of sampling randomly from this distribution. As there is sometimes only few real data because they are expensive or hard to recover, for example in medical imaging, the field of possible industrial applications is consequent and synthetic data of good quality can be generated massively to train other network architectures.
In an attempt to motivate the research work that was pursued in [11] we try here to give an overview of the topic of generative modeling and of the methods and ideas that are commonly encountered to tackle this problem.

1.1 Generative Adversarial Networks

The concept of GANs, first introduced by [9], relies on a two part architecture, a *generator* and a *discriminator*, that are competing against each other : the generator's entry is a random variable z following the reference probability law p_z and maps it as $G(\theta_g, z)$, where θ_g is the set of weights of the generator. On the other side the discriminator D takes data into entry and return a score that we can interpret as the credibility of the data, or the likelihood that this data is real.

The score functional is parametrized by θ_d, the set of weights of the discriminator. G and D then play the following two player min-max game :

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}}\left[\log(D(x))\right] + \mathbb{E}_{z\sim p_z}\left[\log(1-D(G(z)))\right]$$

Numerically, an alternate minimization is performed using gradient descent or ascent on the parameters with back-propagation. When the network is trained, it suffices to remove the discriminator part to get a generator that create realistic synthetic data.
Theoretically, according to [9], assuming that G and D both have enough capacity and that the discriminator is allowed to reach its optimum given G at each step, the training should end up on the saddle-point solution of the problem. Denoting by $\#$ the push-forward action over probability distribution, we have :

$$G\#p_z \xrightarrow{training} p_{data}$$

However, this kind of result can't apply here because the finite-dimensional parametrization limits the representativity of G and D. In fact, even if several numerical evidence of GANs performances has been given, it is well-known that such architectures are very hard to train and very sensitive to the problem structure. A major issue is to control that the generator effectively generates new data and doesn't simply copy the objective data, a behavior referred to as *mode collapse*.

1.2 Wasserstein and Sliced-Wassrstein GAN

Recently, there has been a raising interest in the study of mathematical foundations for the theory of generative networks such as GANs and VAEs. In that matter, the theory of Optimal Transport is a particularly popular topic.

The idea proposed in [3] is to replace the discriminator part by an objective function that the generator should aim at minimizing during training. Intuitively, this objective function should measure some kind of distance between the generated data distribution $G\#p_z$ and p_{data}. Considering such a distance D, the problem of training a generative network becomes :

$$\min_{theta_g} D\left(G(\theta_g)\#p_z, p_{data}\right)$$

Different kind of distances or divergences over the space of probability distributions such as *Total-Variation (TV) distance, Jensen-Shannon (JS) divergence, Kullback-Leibler (KL) divergence* are considered in [3]. Arguing of the good mathematical properties of the *Wasserstein distance*, the authors then showed strong evidences of its good performances when used in the context of generative modeling : synthesized data are of the same quality or better than for classical GANs and the trained architecture is more robust.

More recently, many other works have investigated and confirmed the important role played by OT in the theory of GANs, outlined for example in [7]. The Sinkhorn algorithm presented in [4] provides a fast computation for entropic regularization of the Wasserstein distance. Whereas other losses such as MMD were before more appreciated because of the computational burden of OT, [8] showed how this method could be adapted in a machine learning framework. However, Sinkhorn doesn't allow to compute the OT cost between two atomless distributions and pure applications to the training of GANs or VAEs still suffer from the lack of reprensentativity induced by the parametric model. Other solutions has been studied that would alleviate these issues, for example stochastic gradient descent of MMD losses using RKHS theory as presented in [6].

More recently, other researches such as [5] have investigated the use of the Sliced-Wasserstein distance to compute the distance between probability distributions in a generative setting, benefiting from the simple analytical expression of the Wasserstein distance in 1D to perform a fast, efficient and easily adaptable training. However, the method is still applied in a parametric setting. This is the properties of the Sliced-Wasserstein distance that we will study in the following, in a non-parametric approach to generative modeling.

Mathematical foundations of the notion of gradient flow in the space of probability distributions will first be set in section 2 and 3. Then, details of the numerical implementations and the obtained numerical results will be presented in section 4. At the end of the report, open questions concerning possible future works are discussed.

2 Mathematical background : Optimal transport and (Sliced-)Wasserstein distance

Based on the extensive study presented in [15] or [14], we give an overview of the mathematical tools that are needed in order to justify the numerical methods that we will present after.
We will work on probability distributions, defined as positive measures on two complete and separable metric-spaces $\mathcal{X}$

and $\mathcal{Y}$. We could in fact restrict ourselves to $\mathcal{X}$ and $\mathcal{Y}$ being the same compact set $\Omega \subset \mathbb{R}^d$ for a given dimension d. We only consider distributions with finite second order moment, belonging to the space :

$$\mathcal{P}_2(\mathcal{X}) = \{ \int_{\mathcal{X}} |x|^2 d\mu(x) < \infty \text{ , } \mu \text{ probability distribution} \}$$

2.1 Wasserstein distances

Given an initial measure μ and an objective measure ν optimal transport focus on the problem of finding a transport plan T minimizing the transportation cost $\int_{\mathcal{X}} |x - T(x)|^2 d\mu(x)$ under the constrain $T\#\mu = \nu$, where $\#$ is the push-forward action. Relaxed version of this problem has been proposed by Kantorovitch :

$$(\mathbf{MK}) \ : \ \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} |x-y|^2 d\pi(x,y) \tag{1}$$

where $\Pi(\mu,\nu)$ is te set of joint probability distribution over $\mathcal{X}\times\mathcal{Y}$ with respective marginals μ and ν. The 2-Wasserstein distance between μ and ν is then defined as :

$$W_2^2(\mu,\nu) = \{ \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} |x-y|^2 d\pi(x,y) \}^{1/2} \tag{2}$$

We state here the fundamental properties of the Wasserstein distance :

Property 1. The Wasserstein distance W_2 defines a distance on $\mathcal{P}_2(\mathcal{X})$, i.e. it is *positive*, *definite* and satisfies the *triangular inequality*. Furthermore, the Wasserstein distance metricises the topology of weak convergence or convergence in law over $\mathcal{P}_2(\mathcal{X})$, i.e. for every μ and every sequence (μ_n) in $\mathcal{P}_2(\mathcal{X})$:

$$\mu_n \rightharpoonup \mu \iff \forall f \in \mathcal{C}_b(\mathcal{X}), \int f d\mu_n \xrightarrow{n\to+\infty} \int f d\mu \iff W_2(\mu_n,\mu) \xrightarrow{n\to+\infty} 0$$

Remark 2. Pointed out in [3], this property of metricising the weak-convergence topology is appealing in a machine learning framework as other distances or divergences such as Kullback-Leibler, or Total Variation are in a certain sense too "strong" and can't handle distributions with non-overlapping support : μ and ν can be very distant for these distances and very close in W_2 distance. In particular, this can append when the distributions are degenerate and supported on a low-dimensional sub-manifold of $\mathcal{X}$ which is often the case in generative modeling. GANs and VAEs in fact exploit this idea by trying to represent complex distributions with only a few entry or encoding parameters.

Optimal Transport in 1D We focus on the case $\mathcal{X} = \mathbb{R}$ where we can provide simple analytical solution to the problem of optimal transportation under mild assumptions :

Theorem 3 (1D monotone rearrangement)**.** *Let μ, ν be two probability distribution over $\mathbb{R}$ and let F, G be their two respective cumulative distribution function. Let's furthermore assume that μ is atomless, i.e. admits a density with respect to the Lebesgue measure. Then the map $T = G^{-1} \circ F$ is the unique solution of the (Monge) OT problem, i.e. $(Id, T)\#\mu$ solves (1) and we have an analytical form for the Wasserstein distance :*

$$W_2^2(\mu,\nu) = \int_{\mathbb{R}} |x - G^{-1} \circ F(x)|^2 d\mu(x) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt$$

Where F^{-1} and $G^{-}1$ stand for the generalize inverse of the non-decreasing, right-continuous functions F and G, also called quantile functions.
The function ψ on $\mathbb{R}$ such that $\psi' = Id - G^{-1} \circ F$ is called the Kantorovitch potential between the measures μ and ν.

In other words, the OT maps each quantile of the first distribution to the same quantile of the second distribution, a formulation that allows for direct numerical applications. Indeed, even if it does not fit with the assumptions made in theorem 3, we can consider two distributions defined over the same number of atoms and uniformly distributed over these atoms. Then the same kind of result apply :

Proposition 4. Let $(x_i)_{1\le i\le n}$ and $(y_j)_{1\le j\le n}$ be two families of points on $\mathbb{R}$ and let us define the atomic distributions $\mu = \frac{1}{n}\sum \delta_{x_i}$ and $\nu = \frac{1}{n}\sum \delta_{y_i}$. We note σ_μ and σ_ν the permutations sorting the families $(x_i)_{1\le i\le n}$ and $(y_j)_{1\le j\le n}$ respectively in ascending order. Then the OT cost between the two distributions is given by :

$$W_2^2(\mu,\nu) = \frac{1}{n}\sum_{i=1}^{n} |x_{\sigma_\mu(i)} - y_{\sigma_\nu(i)}|^2$$

defined by the monotone rearrangement T such that for all i, $T(x_{\sigma_\mu(i)}) = y_{\sigma_\nu(i)}$

Numerically, computation of the OT cost is straightforward using `argsort` function.

2.2 Sliced-Wasserstein distance

This last result motivated the idea of introducing the Sliced-Wasserstein distance which is based on the computation of Wasserstein distances between probability distributions projected on straight lines with different directions. We rely here on the extensive mathematical study that was performed in [12].

Definition 5 (Radon transform)**.** Let μ be a probability distribution over $\mathbb{R}^d$, and θ belonging to the centered unit sphere $\mathbb{S}^{d-1}$. We define the Radon transform $\theta\#\mu$ of μ in the direction θ, the probability distribution defined by the push-forward action against μ of the scalar product with θ. If we take $f \in \mathcal{C}_c(\mathbb{R})$ as test function we have :

$$\int_{\mathbb{R}} f(t)d\theta\#\mu(t) = \int_{\mathbb{R}^d} f(x.\theta)d\mu(t)$$

Interesting property is that Radon transform actually results in "slicing" in the Fourier mode : for every $\theta \in \mathbb{S}^{d-1}$ and $s \in \mathbb{R}$, if $\mathcal{F}$ denote the Fourier transform then $\mathcal{F}(\theta\#\mu)(s) = \mathcal{F}(\mu)(s\theta)$. In particular the Radon transform is injective. As pointed out in a remark in [14], Radon transform and the sliced-Fourier property has already proven useful in lots of applications such as medical imaging and computer tomography. It allows here to define a numerically appealing kind of distance.

Definition 6 (Sliced-Wasserstein distance)**.** Let μ, ν be two probability distribution in $\mathcal{P}_2(\mathbb{R}^d)$. Then for any $\theta \in \mathbb{S}^{d-1}$ the Radon transform $\theta\#\mu$ and $\theta\#\nu$ belong to $\mathcal{P}_2(\mathbb{R})$. In particular we can define the 2-Sliced-Wasserstein distance between μ and ν by :

$$SW_2^2(\mu,\nu) = \int_{\mathbb{S}^{d-1}} W_2^2(\theta\#\mu, \theta\#\nu)d\theta \tag{3}$$

where $d\theta$ is the uniform probability distribution over $\mathbb{S}^{d-1}$.

Positivity and triangle inequality directly follows from the fact that W_2 is already a distance, and the fact that SW_2 is definite comes from the above remark about injectivity of the Radon transform. In fact, it is provable that the this distance is equivalent to the classic Wasserstein distance, in particular they define the same topology on $\mathcal{P}_2(\Omega)$ when Ω is a compact subset of $\mathbb{R}^d$.

Theorem 7 (Equivalence of SW_2 and W_2)**.** *Let $R > 0$. There exist a constant $C_d > 0$ such that for every $\mu, \nu \in \mathcal{P}_2(B(0,R))$ we have :*

$$\frac{1}{C_d} SW_2^2(\mu,\nu) \leq W_2(\mu,\nu)^2 \leq C_d R^{1/(d-1)} SW_2(\mu,\nu)^{1/(d+1)}$$

Although equivalent to the W_2, SW_2 has the advantage to be numerically very simple to compute. Indeed, in the case of atomic distribution with uniform weights, considering a large number of directions $(\theta_j)_{1\leq j\leq N_\theta}$ sampled uniformly on $\mathbb{S}^{d-1}$ we can approximate $SW_2^2(\mu,\nu)$ by :

$$SW_2^2(\mu,\nu) = \frac{1}{N_\theta}\frac{1}{n}\sum_{j=1}^{N_\theta}\sum_{i=1}^{n}\langle\theta, x_{\sigma_{\theta_j\#\mu}(i)} - y_{\sigma_{\theta_j\#\nu}(i)}\rangle^2$$

Remark 8. Integration of the 1D transport maps $T_\theta = G_\theta^{-1} \circ F_\theta$ also allows to recover a general measure transport as $S : x \mapsto x + \int_{\mathbb{S}^{d-1}} (T_\theta(x) - x)\, d\theta$. Although the map S does not transport μ onto ν in general, we can expect it to be a good approximation of the exact optimal transport when both measures are close. This give rise to the idea of considering an iterative process $\mu_{n+1} = S_n\#\mu_n$ where S_n is the "optimal SW map" between measures μ_n and ν. This is the intuition behind the method originally proposed by [13] and used in [11].

Remark 9. We have considered here Radon transform that are the push-forward image of a measure by the scalar product. In terms of density, this boils done to the integration of the density on a $(d-1)$-dimensional hyper-plane. If μ admits density ρ, then $\theta\#\mu$ admits a density $\theta\#\rho$ with respect to the Lebesgue measure given by :

$$\forall z \in \mathbb{R},\ \theta\#\rho(z) = \int_{H_{d-1}(\theta)} \rho(z\theta + x)dx$$

where $H_{d-1}(\theta)$ stands for the orthogonal of θ. This consideration allows to generalize to other transformation given by integration of the density on various parametrized $(d-1)$-dimensional sub-manifolds. This is the idea exploited in [10] to define generalized form of SW distances and to study their numerical property, thereby demonstrating their superior performances over "classical" SW distance in various generative modeling applications.

2.3 Maximum Mean Discrepancy (MMD) loss

Their link with *Reproducing Kernel Hilbert Space (RKHS)* theory made MMD losses widely used in machine learning applications because of their nice computational property for appropriate choices of what is called a *reproducing kernel*. What use here the definitions proposed in [2].
Given a characteristic kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ we denote by $\mathcal{H}$ its corresponding RKHS endowed with the scalar product $\langle ., . \rangle_{\mathcal{H}}$. We recall that $\mathcal{H}$ is a space of functions over $\mathcal{X}$ with the fundamental property that for every $u \in \mathcal{H}$ and every $x \in \mathcal{X}$

$$\begin{aligned} &k(x, .) \in \mathcal{H} \\ &u(x) = \langle u, k(x, .) \rangle_{\mathcal{H}} \end{aligned}$$

A classical choice for k is generally the gaussian kernel $k(x, y) = e^{-|x-y|^2}$

Given two probability measures $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$, we can define the MMD loss between μ and ν :

Definition 10 (MMD loss)**.** Let μ, ν belong to $\mathcal{P}_2(\mathcal{X})$. $MMD(\mu, \nu)$ is the Hilbert norm of the of the unnormalized *witness function* $f_{\mu,\nu}$ which is defined as the difference between the mean embedding of the measures :

$$MMD_k(\mu, \nu) = \|f_{\mu,\nu}\|_{\mathcal{H}}, \quad f_{\mu,\nu}(z) = \int_{\mathcal{X}} k(z, .)d\mu - \int_{\mathcal{X}} k(z, .)d\nu, \ \forall z \in \mathcal{X}$$

or equivalently through the reproducing property of k :

$$MMD_k^2(\mu, \nu) = \int_{\mathcal{X}\times\mathcal{X}} k(x, x')d\mu(x)d\mu(x') + \int_{\mathcal{X}\times\mathcal{X}} k(y, y')d\nu(y)d\nu(y') - 2\int_{\mathcal{X}\times\mathcal{X}} k(x, y)d\mu(x)d\nu(y) \tag{4}$$

We note that which should not expect this loss to be a distance over $\mathcal{P}_2(\mathcal{X})$ in general, as it may not be definite or may not satisfy the triangular inequality.

2.4 Sinkhorn divergence

Sinkhorn divergence between two probability distributions μ and ν is defines through a regularization of the optimal transport cost. Given a regularization parameter ϵ, let us define the entropic regularization of the OT problem by :

$$(\mathcal{P}_\epsilon) : \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} |x - y|^2 d\pi(x, y) + \epsilon KL(\pi \| \mu \otimes \nu)$$

Definition 11 (Sinkhorn divergence)**.** Given the optimal coupling π_ϵ for the strictly convex problem $(\mathcal{P}_\epsilon)$, the associated regularized Wasserstein distance is defined by :

$$W_\epsilon^2(\mu, \nu) = \int_{\mathcal{X}\times\mathcal{X}} c d\pi_\epsilon$$

The Sinkhorn divergence is then defined as the symmetrization of the squared W_ϵ :

$$S_\epsilon(\mu, \nu) = W_\epsilon^2(\mu, \nu) - \frac{1}{2}W_\epsilon^2(\mu, \mu) - \frac{1}{2}W_\epsilon^2(\nu, \nu) \tag{5}$$

A fundamental property shown in [8] is that the Sinkhorn divergence can be viewed as an interpolation of between the classic OT cots W_2 and a MMD loss.

Theorem 12. *Let μ, ν belong to $\mathcal{P}_2(\mathcal{X})$ and let $MMD_{-|.|^2}$ be the MMD loss defined by equation (4) with kernel $k : (x, x') \mapsto -|x - x'|^2$. Then we have the following limiting behavior :*

1. $S_\epsilon(\mu, \nu) \xrightarrow{\epsilon\to 0} W_\epsilon(\mu, \nu)$
2. $S_\epsilon(\mu, \nu) \xrightarrow{\epsilon\to 0} MMD_{-|.|^2}(\mu, \nu)$

3 About gradient flow in the space of probability distributions

For the sake of the self-containment of this report and because these considerations build the foundations of the numerical methods that we will present later, we make here a note about continuity equations, based on the theory presented in [1].

3.1 Continuity equation

The following fundamental theorem gives a characterization of absolutely continuous curves in the Wasserstein space $\mathcal{P}_2(\mathcal{X})$ and make a link between the theory of OT and the theory of transport ODEs.

Theorem 13 (Absolutely continuous curves and the continuity equation)**.** *Let I be an open interval in $\mathbb{R}$ and $\mu_t : I \to \mathcal{P}_2(\mathcal{X})$ be an absolutely continuous curve for the metric W_2. Then there exist a vector field $v : (x,t) \mapsto v_t(x)$ such that $v_t \in L^2(\mu_t), \forall t \in I$ and the continuity equation*

$$\partial_t \mu_t + div(v_t \mu_t) = 0 \text{ in } \mathcal{X} \times I \tag{6}$$

holds in a weak sense, i.e. for every test function $\varphi \in \mathcal{D}(\mathcal{X} \times I)$:

$$\int_I \int_{\mathcal{X}} \left(\partial_x \varphi(x,t) + v_t(x).\nabla_x \varphi(x,t)\right) d\mu_t(x) dt = 0$$

Remark 14 (Link with the theory of stochastic processes)**.** Equation (6) is non-linear in general and is known in the theory of diffusion processes to belong to the class of McKean-Vlasov equations. Under few assumptions on the vector field v_t one can show that there exist an unique stochastic process $(X_t)_{t \geq 0}$ satisfying the Stochastic Differential Equation :

$$dX_t = v_t(X_t)dt, \quad X_0 \sim \mu_0 \tag{7}$$

Then at each time t, the law μ_t of X_t can be proved to satisfy equation (6).

The best example of application of this theorem is to geodesic curves of $\mathcal{P}_2(\mathcal{X})$. Considering two absolutely continuous measures μ, ν in $\mathcal{P}_2(\mathcal{X})$, we know thanks to Brenier theorem that there exist an (unique) optimal map T which is solution of the Monge problem. Then the curve defined for $t \in [0,1]$ by $\mu_t = ((1-t)Id + tT)\#\mu$ is a constant speed geodesic between μ and ν and satisfy the continuity equation :

$$\partial_t \mu_t + div(v_t \mu_t) = 0 \text{ in } \mathcal{X} \times I \tag{8}$$

where v_t is the vector field defined on $\mathcal{X}$ by $v_t((1-t)x + tT(x)) = T(x) - x$.

Inspired by the method proposed in [11] to discretize the continuity equation derived by the gradient descent of the SW_2 functional, one could try to directly discretize equation 8 in order obtain the same kind of non-parametric generative model. Unfortunately, this is impossible in practice as the computation of the vector field v_t precisely rely... on the knowledge of the optimal transport map T, which then become a "Catch 22" problem.

3.2 Gradient flow

Generalizing the notion of gradient flow in vector spaces, we take advantage of the metric structure of $\mathcal{P}_2$ endowed with W_2 to study the notion of gradient flow of functionals over probability distributions. We first define the notion of *first variation of a functional* based on [14] :

Definition 15 (First variation)**.** Let $F : \mathcal{P}_2(\mathcal{X}) \to \mathbb{R}$ be a functional. For every measure $\mu \in \mathcal{P}_2(\mathcal{X})$ we call first variation of F at μ, noted as $\frac{\delta F}{\delta \mu}(\mu)$, if it exists, any measurable function such that :

$$\lim_{\epsilon \to 0} \frac{F(\mu + \epsilon\chi) - F(\mu)}{\epsilon} = \int_{\mathcal{X}} \frac{\delta F}{\delta \mu}(\mu) d\chi$$

for every perturbation $\chi = \tilde{\mu} - \mu$ with $\tilde{\mu} \in \mathcal{P}_2(\mathcal{X})$

The gradient flow of F is defined under the form of a continuity equation :

Definition 16 (Gradient flow)**.** Let $F : \mathcal{P}_2(\mathcal{X}) \to \mathbb{R}$ be a lower semi-continuous functional. We will define as the gradient flow of μ of the functional F any solution of the continuity equation :

$$\partial_t \mu_t - div(\mu_t \nabla \frac{\delta F}{\delta \mu_t}(\mu_t)) = 0 \tag{9}$$

with initial data $\mu_0 = \mu$.

This definition of the gradient flow is in fact obtained when taking the limit $\tau \to 0$ in the piece-wise constant interpolation of the iterative process :

$$\mu_{k+1}^{\tau} = argmin_{\mu}\ F(\mu) + \frac{1}{2\tau} W_2(\mu, \mu_k^{\tau})$$

called *Minimizing Movement Scheme* of the functional F. Thus equation (9) can be interpreted as the continuous displacement of μ_t in the direction of steepest variation of F. Furthermore, decay of F along the flow is analytically justified by the formal computation :

$$\begin{aligned}\frac{dF}{dt}(\mu_t) &= -\int_{\mathcal{X}} \frac{\delta F}{\delta \mu_t}(\mu_t) \mathrm{d}div(\mu_t \nabla \frac{\delta F}{\delta \mu_t}(\mu_t)) \\ &= -\int_{\mathcal{X}} \left|\frac{\delta F}{\delta \mu_t}(\mu_t)\right|^2 d\mu_t\end{aligned}$$

However, one can not expect that the gradient flow μ_t will converge towards a global minimum of F in general.

Theorem 7 implies that for a given reference measure ν, the functional $F : \mu \mapsto SW_2^2(\mu, \nu)$ is continuous. The following proposition, proved in [12], gives a formula for its first variation :

Property 17. Let μ, ν belong to $\mathcal{P}_2(\mathcal{X})$ and let us assume that ν is absolutely continuous. Then for each $\theta \in \mathbb{S}^{d-1}$ there is a Kantorovitch potential ψ_θ between $\theta\#\mu$ and $\theta\#nu$ and if $\chi = \tilde{\mu} - \mu$ with $\tilde{\mu} \in \mathcal{P}_2(\mathcal{X})$ then :

$$\lim_{\epsilon \to 0^+} \frac{SW_2^2(\mu + \epsilon\chi, \nu) - SW_2^2(\mu, \nu)}{\epsilon} = \int_{\mathbb{S}^{d-1}} \int_{\mathcal{X}} \psi_\theta(x.\theta) d\chi(x) d\theta \tag{10}$$

In other word, the first variation of SW_2^2 is given by the average of the Kantorovitch potentials in each direction.

In the case of the squared MMD losses, as pointed out in [2], the computation of the first variation is directly given by the evaluation of the witness function $f_{\mu,\nu}$:

$$\lim_{\epsilon \to 0} \frac{MMD^2(\mu + \epsilon\chi, \nu) - MMD^2(\mu, \nu)}{2\epsilon} = \int_{\mathcal{X}} f_{\mu,\nu} d\chi \tag{11}$$

4 Numerical implementation

We focus on the numerical resolution of the non-parametric optimization problem :

$$\min_{\mu \in \mathcal{P}_2\mathcal{X}} F(\mu) \tag{12}$$

where F is a functional, assumed to be at least l.s.c., measuring a "notion of distance" between μ and an objective measure ν. Here, for computational facility and because it fits within the context of generative modeling, we will assume ν to be the atomic measure :

$$\nu = \sum_{j=1}^{M} a_j \delta_{Y^j}$$

. Where a_j are positive weights summing to one and Y^j can be interpreted as observed data. Note however that the presented method can be easily adapted to the case where ν admits a density.

In order to solve problem (12) we aim at simulating the McKean-Vlasov equation (9) starting from a reference measure μ_0, which is a problem we know to be equivalent to the simulation of the law of a process X_t satisfying equation 7 with vector field $v_t = -\nabla \frac{\delta F}{\delta \mu}(\mu_t)$:

$$dX_t = v_t(X_t)dt = -\nabla \frac{\delta F}{\delta \mu}(\mu_t)(X_t)dt, \quad X_0 \sim \mu_0 \tag{13}$$

In fact one only want to draw samples from μ_t at each time t, knowing the displacement vector field $v_t = -\nabla \frac{\delta F}{\delta \mu}(\mu_t)$, that can be done efficiently by drawing x from μ_0 and resolving the transport equation :

$$x_t' = v_t(x_t),\ x_0 = x$$

For a given time discretization step $\tau > 0$, one can for example use the explicit Euler scheme :

$$x_{n+1} = x_n + \tau v_{n\tau}(x_n),\ x_0 = x$$

However the dependence of the vector field with respect to the global law μ_t of X_t prevents us from doing so since the law of μ_t can't be estimated a priori. The solution proposed by [11] is to draw the flow of a whole system of N particles, approximating at each time step the real solution $\mu_{n\tau}$ by the empirical distribution $\hat{\mu}_n = \frac{1}{N}\sum_i \delta_{x_i}$.
The obtained procedure, called Euler-Maruyama discretization scheme, is detailed in algorithm 4 and reads as follows :

Algorithm 1 Gradient flow discretization

procedure $\min_\mu F(\mu)$
  Choose a reference measure μ, time step $\tau > 0$ and integer K and calculate the gradient flow μ_t at every time $k\tau$ with $\mu_0 = \mu$.
  $X_0^i \sim_{iid} \mu_0$, $\nu = \sum_j a_j \delta_{Y^j}$, `drift` $\in \{True, False\}$, λ
  for $k < K$ **do**
    $\hat{\mu}_k \leftarrow \frac{1}{N}\sum_i \delta_{X_k^i}$
    Compute for each i : $\hat{v}_k(X_k^i) = -\nabla \frac{\delta F}{\delta \mu}(\hat{\mu}_k)(X_k^i)$ (this step depends on the flow)
    Update for each i : $X_{k+1}^i = X_k^i + \hat{v}_k(X_k^i)$
    if `drift` **then**
      For each i : $X_{k+1}^i += \sqrt{2\lambda\tau} Z_k^i$ with $(Z_k^i)_{k,i} \sim_{iid} \mathcal{N}(0, I_d)$
  Output: $(X_K^i)_i$

1. For $i \in \{1, ..., N\}$, draw $X_0^i \sim_{iid} \mu_0$
2. For $k \in \{0, ..., K-1\}$, $X_{k+1}^i = X_k^i + \hat{v}_k(X_k^i) = X_k^i - \tau \nabla \frac{\delta F}{\delta \mu}(\hat{\mu}_k)(X_k^i)$

with $\hat{\mu}_k = \frac{1}{N}\sum_i \delta_{X_k^i}$ the empirical distribution at step k and $\hat{v}_k$ its associated displacement vector field.

See the subsections below for the details regarding the computation of the displacement vector field for each of the considered flows.

4.1 Computing the SW Flow

For the particular case of the Sliced-Wasserstein flow, F is given by the entropic regularization of the squared SW_2 distance between μ and the objective distribution ν, i.e. :

$$F_\lambda(\mu) = \frac{1}{2} SW_2^2(\mu, \nu) + \lambda H(\mu)$$

where H is given for a distribution μ by $H(\mu) = \int_{\mathcal{X}} \rho(x)\log(\rho(x))dx$ if μ admits a density ρ and $H(\mu) = +\infty$ otherwise. Following equation (10) the gradient of the first variation of F_λ is given by :

$$\nabla \frac{\delta F_\lambda}{\delta \mu}(\mu)(x) = \int_{\mathbb{S}^{d-1}} \psi'_\theta(x.\theta)\theta d\theta + \frac{\nabla \rho}{\rho}$$

where ψ_θ is the Kantorovitch potential between $\theta\#\mu$ and $\theta\#\nu$ and ρ is the density of μ. The continuity equation associated to the gradient flow of F_λ then reads :

$$\partial_t \mu_t = -div(v_t \mu_t) + \lambda \Delta \mu_t, \quad \text{with vector field} \quad v_t(x) = -\int_{\mathbb{S}^{d-1}} \psi'_\theta(x.\theta)\theta d\theta \tag{14}$$

Even if it makes assumptions that take us outside of the context of our numerical application, the authors of [11] justified the existence of a solution for this flow :

Theorem 18. *Let ν be a probability measure on $\overline{B(0,1)}$ with strictly positive smooth density. Choose a regularization constant $\lambda > 0$ and a radius $r > \sqrt{d}$ with d the data dimension. Assume that $\mu_0 \in \mathcal{P}_2(\overline{B(0,r)})$ with density ρ_0 with respect to the Lebesgue measure. Then there exist a solution $(\mu_t)_{t\geq 0}$ for the flow (14) with initial data μ_0. Furthermore, μ_t has density ρ_t for every $t \geq 0$.*

Numerically, [11] proposes a an approximation of the integral defining v_t via a Monte-Carlo estimate involving to draw randomly N_θ direction θ from $\mathbb{S}^{d-1}$. However, it works as well and as it induces important computational simplification, we preferred a stochastic approach. At each time step k, the displacement vector field is estimated by :

$$\hat{v}_k(x) = -\psi'_{k,\theta_k}(x.\theta_k)\theta_k, \quad \text{with} \quad \theta_k \sim_{iid} \mathcal{U}(\mathbb{S}^{d-1})$$

The Kantorovitch potentials are computed using sorting and cumulative distribution. At step k and for a projection direction θ, we note σ_k^θ and γ^θ the permutation sorting respectively $(\langle \theta, X_k^i \rangle)_i$ and $(\langle \theta, Y^j \rangle)_j$ in ascending order. The quantile function QF is defined for each index i as the index j such that : $\sum_{l=1}^{j} a_{\gamma^\theta(l)} \leq i/N < \sum_{l=1}^{j+1} a_{\gamma^\theta(l)}$. We have :

$$\psi'_{k,\theta}(x) = x - F^{-1}_{\theta\#\nu} \circ F_{\theta\#\hat{\mu}_k}(x)$$

$$\text{i.e.} \quad \forall k, i \quad \psi'_{k,\theta}(X_k^{\sigma_k^\theta(i)}) = X_k^{\sigma_k^\theta(i)} - Y^{\gamma^\theta(QF(i))}$$

Therefore, at each time step k, particles updates reads :

$$\forall i, \quad X_{k+1}^i = X_k^i - \tau \hat{v}_k(X_k^i) + \sqrt{2\lambda\tau} Z_k^i, \quad \text{with} \quad Z_k^i \sim_{iid} \mathcal{N}(0,1)$$

Remark 19. The brownian drift Z_k^i added at each step is induced by the entropic regularization of SW_2^2 and the laplacian term in the gradient flow equation. Brownian motion is indeed famously known in diffusion processes to generate solution of the heat equation. In our case, entropic regularization is introduced in order to generate sufficiently expressive results and avoid mode collapse.

Remark 20. The choice of performing stochastic optimization was also motivated by the following result from [13]:

Theorem 21 (Pitié-Kokaram-Dahyot)**.** *We assume μ is absolutely continuous and ν is a gaussian measure. At each step k we let $B_k = (e_k^1, ..., e_k^d)$ be an orthonormal basis of $\mathbb{R}^d$ used to compute the updates displacement vector field :*

$$v_k(x) = -\frac{1}{d} \sum_i \psi'_{k,e_k^i}(x.e_k^i) e_k^i$$

Then if the basis $(B_k)_{k\geq 0}$ are independent uniform random variables over the set of all orthonormal basis, $\mu_k \xrightarrow{k\to+\infty} \nu$ almost surely.

Note that in [11] the authors choose to draw first a set of directions (θ_i), supposed to be representative of the whole sphere $\mathbb{S}^{d-1}$, and then compute the flow using the same set of directions at each step. In a generative setting this approach allows to easily draw the flow of a supplementary particle as one just need to store one set of directions instead of a set of directions for each step. However according to this theorem, this approach may lose in precision.

4.2 Computing the MMD flow

In the case of an MMD flow, F is a squared MMD loss induced by a given kernel k as in (4) :

$$F(\mu) = \frac{1}{2} MMD_k^2(\mu, \nu)$$

Recalling equation (11), the gradient of the first variation is given by :

$$\nabla \frac{\delta F}{\delta \mu}(\mu)(z) = \int_{\mathcal{X}} \nabla_2 k(x,z) d\mu(x) - \nabla_2 k(u,z) d\nu(y)$$

Thus, the displacement vector field is numerically estimated at each step k by :

$$\hat{v}_k(z) = -\frac{1}{N} \sum_i \nabla_2 k(X_k^i, z) + \sum_j a_j \nabla_2 k(Y^j, z)$$

and particles updates reads :

$$\forall i, \quad X_{k+1}^i = X_k^i + \tau \hat{v}_k(X_k^i)$$

Remark 22 (Boosting the convergence rate)**.** In all the examples showed after we used the gaussian kernel $k(x,y) = e^{-|x-y|^2/2\sigma^2}$. Other kernels have been tried out, but none shown better results. Note that we set σ according to the data such that it scales with the maximum of the encountered norms during computation. This allows to boost the convergence which tends to be slowed by the fast decay of the gaussian kernel. However, this was not enough and we had to scale the step size by approximately 100 in order to obtain convergence rates that are comparable to the ones of the SW_2^2 flow.

Remark 23. As the displacement vector field is expressed as a mean, we have tried some stochastic optimization approach in order to gain on computational time. However, none of these attempts ever generated a good fitting of the objective data.

4.3 Computing the Sinkhorn divergence flow

For the Sinkhorn flow, F is set to be the Sinkhorn divergence of equation (5) for a given value of ϵ:

$$\begin{aligned} F(\hat{\mu}_k) &= \mathcal{S}_\epsilon(\hat{\mu}_{k\tau}, \nu) \\ &= \sum_{i,j} |X_k^i - Y^j|^2 P_{ij} + \epsilon KL(P \| (1/N) a^T) - \frac{1}{2} \left[\sum_{i,j} |X_k^i - X_k^j|^2 Q_{ij} + \epsilon KL(Q \| (1/N)(1/N)^T) \right] \end{aligned}$$

where P and Q are the two joint distributions solutions of the optimization problems associated respectively to $W_\epsilon(\hat{\mu}_k, \nu)$ and $W_\epsilon(\hat{\mu}_k, \hat{\mu}_k)$. They are obtained by Sinkhorn iterations.

The gradient of F is given in each index i by :

$$\begin{aligned}\nabla_i F(\hat{\mu}_k) &= 2\sum_j P_{ij}(X_k^i - Y^j) - \sum_j Q_{ij}(X_k^i - X_k^j) \\ &= \frac{1}{N}X_k^i - 2\sum_j P_{ij}Y^j + \sum_j Q_{ij}X_k^j\end{aligned}$$

And the particles updates at each time k reads :

$$\forall i, \quad X_{k+1}^i = X_k^i - \tau\nabla_i F(\hat{\mu}_k)$$

Remark 24. The level of entropic regularization ϵ is constant in our examples and set to 0.1. This allows for regularization and accelerate the convergence of the Sinkhorn iterations : only few (~ 10 or 20) are necessary at each time step in order to obtain a good gradient.

4.4 Numerical results

We show here some numerical examples in order to illustrate and compare the performance of each of the aforementioned flows. Used parameters are gathered in table 1.

	General parameters
Dimension	$d = 2$
Time step	$\tau = 0.01$
Number of iterations	$K = 1000$
Drift level	$\lambda = 0.001$
	Sinkhorn-flow specific parameters
Sinkhorn regularization level	$\epsilon = 0.1$
Sinkhorn iterations at each step	10 iterations
Gradient boost	$\tau' = 200\tau$
Drift reduction	$\lambda' = \lambda/10$
	MMD-flow specific parameters
Kernel	$k(x, y) = e^{-\lvert x-y\rvert^2/2\sigma^2}$
Space normalization	$\sigma = \max_{i,j}\{\lvert X_0^i - X_0^j\rvert, \lvert X_0^i - Y^j\rvert, \lvert Y^i - Y^j\rvert\}$
Gradient boost	$\tau' = 300\tau$
Drift reduction	$\lambda' = \lambda/100$

Table 1: Numerical parameters used in the experiments

In order to illustrate the geometric properties of the flow, we set the initial measure μ to be uniform on the circle $\mathbb{S}^1$ and the objective measure ν to be the continuous displacement of the same measure (no re-sampling) into a triangle, which is the gradient of a convex function. Results of the computation and associated losses are presented in figure 1. In order to fully compare the quantitative performances of the different flows, both scaled losses (i.e. such that $loss(0) = 1$) associated to each distance and reference OT losses are computed in log scale.

One can see that the best fit is performed by the SW-flow. In fact, equivalent fit would be realised by the Sinkhorn-flow after a few more iteration but the MMD-flow doesn't perform as well and generated figure is blurry. Every flow manage to perform significant decrease of both its associated loss and the OT loss. However, if the losses decrease fast along the first iterations for the MMD flow, convergence is slower at the end such that, in OT distance, the generated figure is a bit more far away than the Sinkhorn's one. The figure generated by the SW-flow is by far the closest to the objective.

In particular, ones sees that computation took much less time for the SW-flow, thanks to the adopted stochastic optimization method. On the other hand Sinkhorn- and MMD-flow took much more time to compute. This fact is explain for the Sinkhorn-flow that one has to compute several costly Sinkhorn iterations at each step. Although we only performed few (10 iterations) we see that it hinders performances.

We performed the same tests adding a drift to the flow as was originally proposed in [11]. The drift is designed to add in generalization and its level is managed by the constant λ that we set to a low level : $\lambda = 0.001$. However, we noticed that Sinkhorn-flow and MMD-flow responded particularly bad to the addition of this drift. For the Sinkhorn-flow this

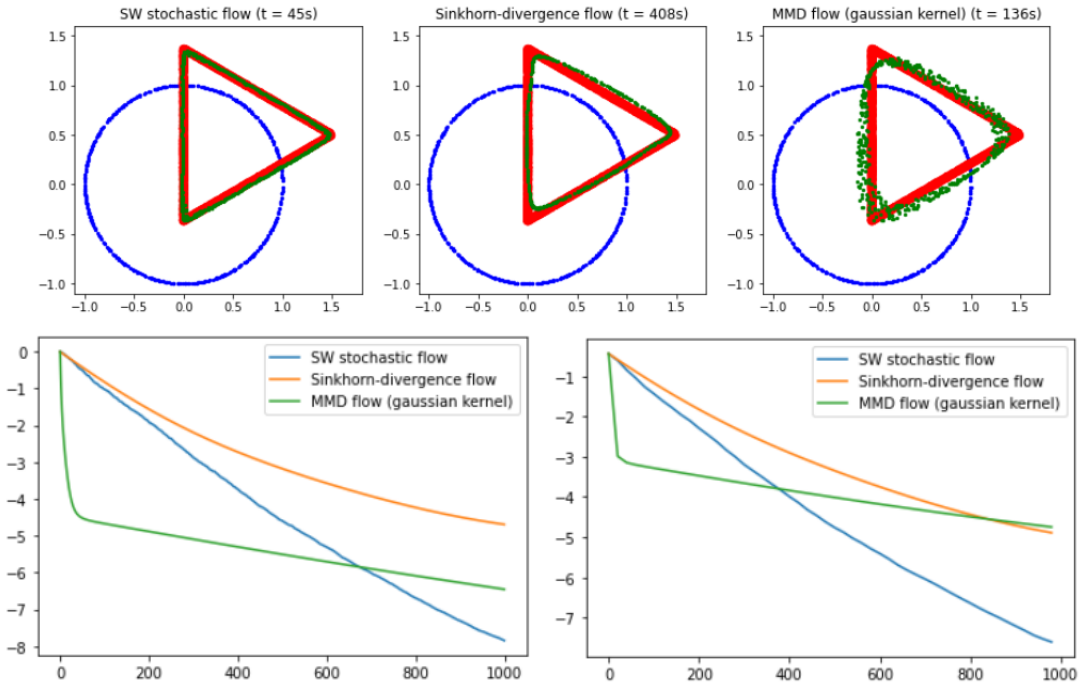

Figure 1: Fitted data and losses after a flow of $N = 500$ particles towards objective measure $\nu = \nabla\varphi\#\mu$, with times step $\tau = 0.01$ and number of step $K = 1000$, without drift.
Initial, objective and fitted data are respectively in blue, red and green and t stand for the computational time in the upper figure. Lower figures show the evolution along the flow of respectively the log scaled respective losses (right) and the log OT losses (left).

might be explained by the fact that it already performs generalization via the Sinkhorn iteration. For these flows, noise level was down-scaled to a factor 100 in order to obtain the results showed in figure 2.

One observes the same kind of behavior as without drift, excepted the fact that generated figures are more blurry therefore showing that generalization is indeed performed. Note however that these good results come from that fact the noise level λ was adapted for each flow, becoming very low for Sink- and MMD-flow $\lambda \sim 1e^{-5}$. Without that, results wouldn't have been good.

In many applications, such as medical imaging, the smoothness of the flow is a very important criteria. In particular tearing of the data should not appear. In order to observe the smoothness of the flows we computed the evolution of the generated figures in figure 3. One can see than no tearing appear and that every flow respect the geometry of the problem.

Remark 25. Concerning the performances of the MMD flow it is possible that the choice of a gaussian kernel if not fully adapted to the problem, and that the choice of an appropriate kernel would significantly boost the performances.

5 Open questions and sketch of ideas

As pointed out in section 3, it is not to be expected that the gradient flow of a functional F will converge towards a global minimum of this functional or even a local minimum. Some ideas to study the convergence of the flow of SW_2 are already given in [12] :

Property 26. Let $\mu \in \mathcal{P}_2(\mathcal{X})$ be a an absolutely continuous measure with strictly positive density. Then $\mu = \nu$ is and only if :

$$\int_{\mathbb{S}^{d-1}} \psi'_\theta(x.\theta)\theta d\theta = 0, \text{ for } \mu\text{-a.e. x}$$

where ψ_θ is the Kantorovitch potential between $\theta\#\mu$ and $\theta\#\nu$.

This in particular implies that if ν is absolutely continuous, then it is the only stationary point of the flow.

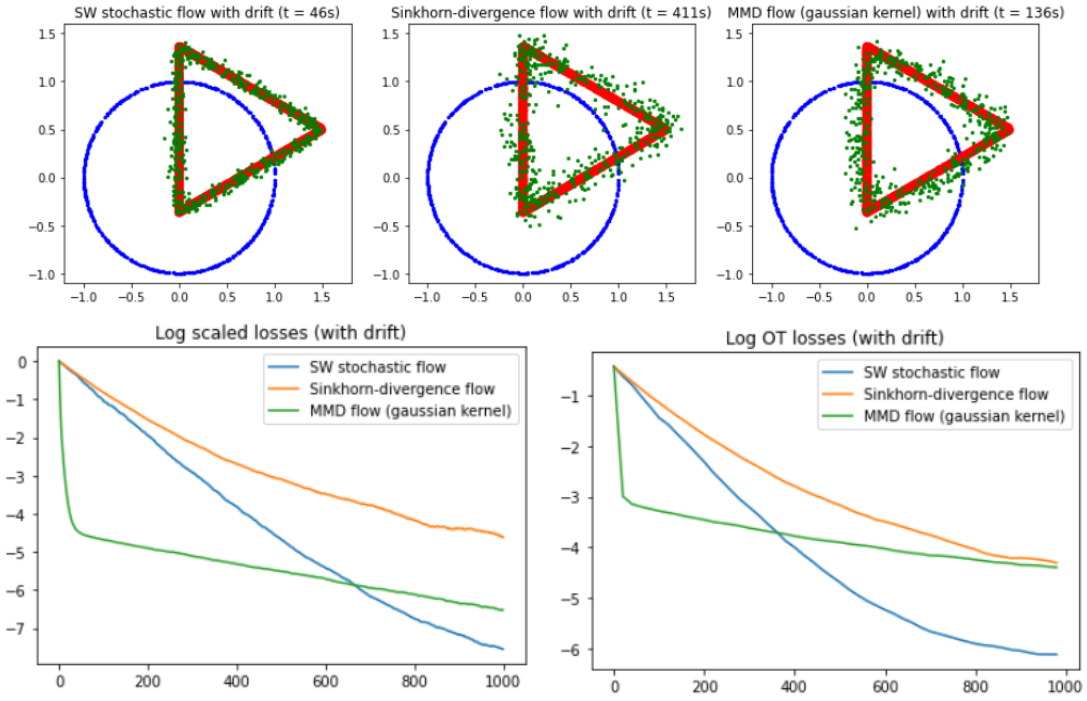

Figure 2: Fitted data and losses after a flow of $N = 500$ particles towards objective measure $\nu = \nabla\varphi\#\mu$, with times step $\tau = 0.01$ and number of step $K = 1000$, with drift level $\lambda = 0.001$ adapted for each flow.
Initial, objective and fitted data are respectively in blue, red and green and t stand for the computational time in the upper figure. Lower figures show the evolution along the flow of respectively the log scaled respective losses (right) and the log OT losses (left).

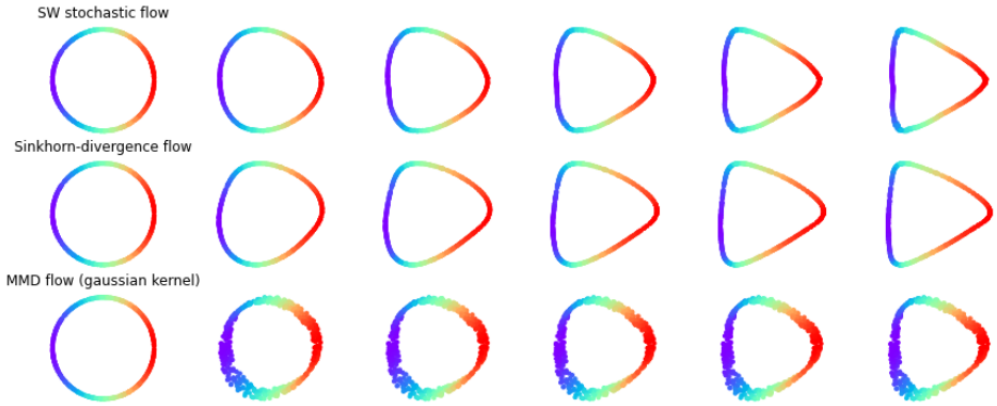

Figure 3: Transforming data along a flow of $N = 500$ particles towards objective measure $\nu = \nabla\varphi\#\mu$, with times step $\tau = 0.01$ and number of step $K = 1000$, without drift.
Figures correspond in each column respectively from left to right to iteration $\{0, 100, 200, 300, 400, 500\}$. For each particle its color is set at the beginning, doesn't change along iterations and is the same for each flow

However, this considerations are outside of the context of our numerical method where ν is by definition an atomic measure. In fact, one big obstacle to prove the convergence of the gradient flow is the lack of convexity of the functional F : it is known that W_2^2 is not geodesically convex and MMD losses are not convex in general. The question of wether SW_2^2 is convex is still open : on one hand, the Sliced-Wasserstein distance is based on the classic Wasserstein so it seems improbable that it will be convex. On the other hand SW_2 only relies on the 1D W_2 and averaging over the sphere could act as a convex regularization, inducing convexity where there is not for W_2^2. We give an example to

motivate this idea, based on an example proposed in [1]. We take three atomic probability distributions over $\mathbb{R}^2$

$$\nu = \frac{1}{2}(\delta_{(0,0)} + \delta_{(2,0)}),\ \mu_1 = \frac{1}{2}(\delta_{(0,0)} + \delta_{(-1,2)}),\ \mu_2 = \frac{1}{2}(\delta_{(0,0)} + \delta_{(-1,-2)})$$

Then the constant speed geodesic between μ_1 and μ_2 is given for $t \in [0, 1]$ by

$$\mu_t = \frac{1}{2}(\delta_{(-t,-2t)} + \delta_{(t-1,2-2t)})$$

and $W_2^2(\mu_t, \nu)$ posses a concave cusp at $t = 1/2$. However, calculation shows that $SW_2^2(\mu_t, \nu)$ is fully convex between on $[0, 1]$. Evolution profile of both functional are drawn in figure 4.

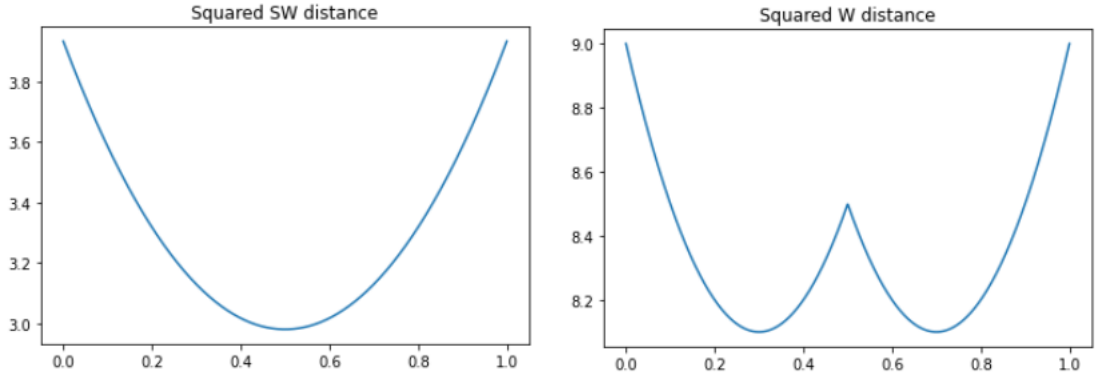

Figure 4: Evolution of the squared Sliced-Wasserstein (right) and Wasserstein (left) distance along the flow μ_t

One sees in this example, that concavity disappears when observing $W_2^2(\theta\#\mu_t, \theta\#\nu)$, the squared Wasserstein distance of the projected measures, for some direction θ. This suggests that in order to further investigate the convexity of SW_2^2 one should first study the case convexity of W_2^2 in 1D

6 Conclusion

This report's goal was to present the numerical method proposed in [11] to solve the problem of generative modeling in a fully non-parametric framework. We approached this problem under the general angle of gradient flow in the space of probability distribution whose mathematical foundations were set in section 2 and 3 and proposed different numerical implementation of this flow, one corresponding to the Sliced-Wasserstein flow of [11] and other corresponding to the Sinkhorn-flow and the MMD-flow with gaussian kernel. Detailed description of the implementations and presentation of the numerical results were performed in section 4.

Through these experimental results one could notice the superior performances of the SW-flow whose quality of the generated data is equivalent or better than the one of both other flows, visually and in term of OT distance. Furthermore, SW-flow benefits from the very low computational time that is induced by the choice of a stochastic optimization scheme whereas Sinkhorn-flow and MMD-flow suffer from large computational time. For the first one it is induced by the obligation to perform several Sinkhorn iterations at each time step. More generally, the experiments carried out the remarkable convergence properties and robustness of the SW-flow.

However, many theoretical questions are still to be answered to rigorously justify its use. In particular, one has no guarantee that the generated data will converge towards the objective distribution along the flow in a general setting. A first step to answer this question would be to study the geodesic convexity of the SW_2^2 functional.

References

[1] L. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient flows in metric spaces and in the space of probability measures*, Lectures in mathematics ETH Zürich, Birkhäuser, Basel, 2. ed ed., 2008. OCLC: 254181287.

[2] M. ARBEL, A. KORBA, A. SALIM, AND A. GRETTON, *Maximum Mean Discrepancy Gradient Flow*, arXiv:1906.04370 [cs, stat], (2019).

[3] M. ARJOVSKY, S. CHINTALA, AND L. BOTTOU, *Wasserstein generative adversarial networks*, in Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y. W. Teh, eds., vol. 70 of Proceedings of Machine Learning Research, International Convention Centre, Sydney, Australia, 06–11 Aug 2017, PMLR, pp. 214–223.

[4] M. CUTURI, *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances*, arXiv:1306.0895 [stat], (2013). arXiv: 1306.0895.

[5] I. DESHPANDE, Z. ZHANG, AND A. SCHWING, *Generative Modeling using the Sliced Wasserstein Distance*, arXiv:1803.11188 [cs], (2018). arXiv: 1803.11188.

[6] A. GENEVAY, M. CUTURI, G. PEYRÉ, AND F. BACH, *Stochastic Optimization for Large-scale Optimal Transport*, arXiv:1605.08527 [cs, math], (2016).

[7] A. GENEVAY, G. PEYRÉ, AND M. CUTURI, *GAN and VAE from an Optimal Transport Point of View*, arXiv:1706.01807 [stat], (2017).

[8] ——, *Learning Generative Models with Sinkhorn Divergences*, arXiv:1706.00292 [stat], (2017).

[9] I. J. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial networks*, 2014.

[10] S. KOLOURI, K. NADJAHI, U. SIMSEKLI, R. BADEAU, AND G. K. ROHDE, *Generalized sliced wasserstein distances*, CoRR, abs/1902.00434 (2019).

[11] A. LIUTKUS, U. ŞIMŞEKLI, S. MAJEWSKI, A. DURMUS, AND F.-R. STÖTER, *Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions*, arXiv:1806.08141 [cs, stat], (2019).

[12] B. NICOLAS, *Unidimensional and evolution methods for optimal transportation*, 2013.

[13] F. PITIÉ, A. C. KOKARAM, AND R. DAHYOT, *Automated colour grading using colour distribution transfer*, Computer Vision and Image Understanding, 107 (2007), pp. 123–137.

[14] F. SANTAMBROGIO, *Optimal Transport for Applied Mathematicians*, vol. 87 of Progress in Nonlinear Differential Equations and Their Applications, Springer International Publishing, Cham, 2015.

[15] C. VILLANI, *Optimal Transport*, vol. 338 of Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.