Exploring interpolation between optimal transport and MMD using geodesic distances

Kerrian Le Caillec
MVA
Paris, France

Antoine Vialle
MVA
Paris, France

ABSTRACT

While empirical point-wise comparisons fail to capture the geometric complexity of a distribution, optimal transport as well as Maximum mean discrepancy (MMD) norms are standard ways to convey the geometry of the distribution into a scalar. The Sinkhorn divergence appears as an interpolation between these two methods and has been studied in many instances. Recent advances have extended its application beyond Euclidean spaces to Riemannian manifolds, enabling the study of optimal transport in geometrically complex settings. This work explores the adaptation of the Sinkhorn algorithm and gradient flow to Riemannian manifolds, where the underlying distance metric is derived from the manifold's intrinsic geometry. Key challenges include efficiently computing geodesic distances, defining cost matrices that respect the manifold's curvature, and ensuring numerical stability in the iterative updates. Numerical experiments demonstrate the algorithm robustness and accuracy in approximating transport maps, highlighting its potential for advancing manifold-based machine learning and geometric data analysis. All the numerical experiments are available at: https://github.com/kerrianlc/Sinkhorn.

KEYWORDS

Optimal transport - Sinkhorn's algorithm - Geodesic distance - Riemannian manifolds - Gradient Flow

1 INTRODUCTION

Probability-based predictions and comparisons present a challenge with far-reaching implications across various machine learning tasks, finding applications in fields such as medicine, natural language processing, and computer vision. Many models rely on directly or indirectly comparing the probabilistic outcome of a model to a ground-truth distribution. Penalizing traditional loss functions using entropic terms has shown interesting results in different learning approaches, Burgess et al. [5] proposes a penalized approach to variational autoencoder. Pereyra et al. [14] proposes a confidence penalty to avoid overconfident predictions in neural networks.

Approaches for probability distribution comparison have been developed, one of which is maximum mean discrepancy, that has shown results in the context of generative adversarial networks (GANs) as stated in [7].

Optimal transport is one of the most common way to compare two probability distributions. Those methods compute a plane to "transport" one distribution to the other [13]. The theory of optimal transport has been extended to computer vision problems [17]. Nonetheless, the difficulty in computing the solution to the optimal transport problem has led to approaches using regularization terms [3]. One of the most widely used regularized algorithms in modern computational optimal transport is the Sinkhorn algorithm [6], which provides an efficient way to approximate Wasserstein distances.

1.1 Related Works

Properties of Sinkhorn algorithm convergence have been extensively studied, offering a detailed understanding of its theoretical behavior and practical computational advantages [8],

In addition to its theoretical appeal, Sinkhorn divergence has been successfully employed as a loss function for training classification models [9], as well as generative models, demonstrating its versatility and efficacy in applications such as GANs and other generative frameworks [10], [18].

Furthermore, the sample complexity of Sinkhorn divergence has been shown to be comparable to that of Maximum Mean Discrepancy (MMD), which is favorable, especially in higher dimensions, as it significantly outperforms the complexity associated with Kantorovich's Optimal Transport (OT) problem [11]. This makes it an attractive choice for scalable applications where computational efficiency is critical. Furthermore, Sinkhorn algorithm approximates the optimal transport problem in almost linear time [2].

Methods to include geodesic into the Sinkhorn algorithm have been developed, notably in [12]. The paper uses a heat kernel approximation to estimate the geodesic on a Riemannian manifold, [1].

1.2 Our contribution

The standard Sinkhorn algorithm relies on Euclidean distance calculations, which are not suitable for data that resides on curved or non-Euclidean manifolds. In these cases, the ability to compute geodesic distances, which represent the shortest paths between points on the manifold, becomes crucial for more accurate and meaningful results.

This paper aims to address these challenges by introducing two key concepts:

Efficient Geodesic Distance Approximation:
We implemented a method for approximating geodesic distances on discrete manifolds. By utilizing a graph-based representation of the manifold and leveraging techniques such as the heat kernel approximation and Chebyshev polynomials, we efficiently compute geodesic distances in a computationally feasible manner.

Geodesic Manifold-Aware Vector Field:
We present a Vector Field which incorporates manifold structures. We will use it for two different algorithms :

- Vector Field Descent for creating smooth path between a target point and a source point.
- Sinkhorn algorithm where the Euclidean cost function has been modified such that its gradient were replaced by the vector field.

This approach enables optimization algorithms to respect the local geometry of the manifold while maintaining the simplicity and effectiveness of Euclidean-based loss functions.

2 MATHEMATICAL FOUNDATION

This section aims to define the theory behind probability comparisons using optimal transport.

2.1 Optimal transport problem

In this paper, we will focus on the Kantorovich's formulation of the optimal transport problem. Given a metric space X, and two probability Radon measures μ and ν both defined on X, we will refer the set of probability Radon measures on X as $\mathcal{M}_1^+(X)$.

Given a source distribution μ on X, a target distribution ν on X, and a cost function $C : X^2 \to \mathbb{R}$, the aim of the OT problem is to find the optimal measure $\pi^* \in \mathcal{M}_1^+\left(X^2\right)$ called the *transportation plan* on the space X^2 such that the marginal distribution on the first space X is equal to μ and the marginal distribution on the second space X is equal to ν. We denote $\Gamma(\mu,\nu) \subset \mathcal{M}_1^+\left(X^2\right)$ the set of all measures on X^2 that verify the previous properties. For two distributions $\mu, \nu \in \mathcal{M}_1^+(X)$ we define product measure as follows : $\forall A, B \in \mathcal{B}(X)$(the Borel set of X), $\mu \otimes \nu(A,B) = \mu(A)\nu(B)$. We denote as $C(X)$ the set of continuous test functions going from X to $\mathbb{R}$ and we will use the inner product notation $\langle f, \mu\rangle := \int_X f \mathrm{d}\mu$. The **optimal transport metric** can thus be expressed as:

$$OT_0 := \inf_{\pi\in\Gamma(\mu,\nu)} \int_{X^2} C(x,y)\mathrm{d}\pi(\mathrm{d}x,\mathrm{d}y),$$

where C is a cost function from X^2 to $\mathbb{R}$. The cost function is useful to "embed" the geometric of the two distributions μ and ν in the integral. One common example of a cost function C is $(x,y) \in X^2 \mapsto \|x-y\|_p^p$, for a given natural integer $p \geq 1$. The optimal transport metric is then referred as the p-**Wasserstein metric**.In order to minimize the previous integral, we want to find the transportation plan that gives the most "weight" to the points closest to the other distribution.

Evaluating OT_0 given μ, ν is actually more difficult and less time-efficient to do numerically than a regularized optimal transport problem in which an entropic term is added, according to Cuturi [6]. The problem is usually penalized using a Kullblack-Leibler (KL) divergence term, for any given measures p and q on X^2 :

$$D_{KL}(p\|q) = \int_{X^2} \log\left(\frac{\mathrm{d}p}{\mathrm{d}q}\right)\mathrm{d}p,$$

where $\frac{\mathrm{d}p}{\mathrm{d}q}$ represents the Radon-Nikodym derivative of p with respect to q.

For any $\varepsilon > 0$, we define the **entropic regularization** of the optimal transport distance as follows :

$$OT_\varepsilon := \inf_{\pi\in\Gamma(\mu,\nu)} \int_{X^2} C(x,y)\mathrm{d}\pi(\mathrm{d}x,\mathrm{d}y) + \varepsilon D_{KL}(\pi\|\mu\otimes\nu). \quad (1)$$

The KL divergence acts as a relative entropic term and it is positive or null by definition. Thus, the further away two probability measures are, the greater the divergence value is. Indeed, the KL divergence is minimal when the transportation plan $\pi = \mu\otimes\nu$, as $D_{KL}(\mu\otimes\nu\|\mu\otimes\nu) = 0$. The penalty makes sure that the found probability distribution π is point-wise similar to $\mu\otimes\nu$.

We now introduce the **Sinkhorn divergences** for any $\varepsilon > 0$:

$$S_\varepsilon(\mu,\nu) := OT_\varepsilon(\mu,\nu) - \frac{1}{2}OT_\varepsilon(\mu,\mu) - \frac{1}{2}OT_\varepsilon(\nu,\nu). \quad (2)$$

The previous quantity is actually an interpolation between the OT_0 distance and the **Maximum Mean Discrepancy norm**-induced distance defined as such:

$$\|\mu-\nu\|_k := \int_{X^2} k(x,y)\mathrm{d}(\mu-\nu)(x)\mathrm{d}(\mu-\nu)(y),$$

where k is a radius basis function kernel on X^2 and $\mu-\nu$ is a signed measure. Then, ε is the interpolation factor between the two divergences; in fact $S_\varepsilon(\mu,\nu) \xrightarrow[\varepsilon\to 0]{} OT_0(\mu,\nu)$ and when $\varepsilon \to +\infty$, we can easily show that $S_\varepsilon(\mu,\nu) \to \|\mu-\nu\|_{-C}$, using the fact that the optimal transport quantity is only defined when $D_{KL}(\mu\otimes\nu\|\mu\otimes\nu) = 0$ [16].

The main result provided by Feydy et al. [8] is the following:

Theorem 2.1. *Let X be a compact metric space, and a Lipschitz cost function C. For a given $\varepsilon > 0$, we have the following items for S_ε:*

Th.1 *S_ε is symmetric positive definite ,*
Th.2 *S_ε is convex in each of its input variables ,*
Th.3 *S_ε metricizes the convergence in law .*

We will summarize the proof of the theorem within the next subsections.

2.2 Dual problem

We can link the OT_0 distance to a dual linear optimization problem based on the convexity behaviour of OT_0 and the Fenchel's duality theorem :

$$OT_0(\mu,\nu) = \sup_{(f,g)\in C(X)} \langle f,\mu\rangle + \langle g,\nu\rangle + \min_{X^2}(C - f\oplus g). \quad (3)$$

The operator $\oplus$ denote the tensor sum between the functions f and g, such that $f\oplus g : (x,y) \in X^2 \mapsto f(x) + g(y)$.

Similarly, according to Peyré and Cuturi [15], the optimized quantity OT_ε can be written without the constraints and thus the problem takes the following form:

$$OT_\varepsilon(\mu,\nu) = \sup_{f,\, g\in C(X)} \langle f,\mu\rangle + \langle g,\nu\rangle - \varepsilon\left\langle \exp\left(-\frac{1}{\varepsilon}(f\oplus g - C)\right), \mu\otimes\nu\right\rangle, \tag{4}$$

The later term in the sum is often referred as the *softMin* of $f\oplus g - C$. Furthermore, a solution (f^*, g^*) of (4) exists and is unique up to a scalar addition; i.e. for all $K\in\mathbb{R}$, (f^*+K, g^*-K) is also solution of (4). The quantity OT_ε can be rewritten as:

$$OT_\varepsilon(\mu,\nu) = \langle f^*,\mu\rangle + \langle g^*,\nu\rangle. \tag{5}$$

Considering the strong solution of the primal-dual pairing, the optimal π^* solving (1) is given by

$$\pi^* = \exp\left(-\frac{1}{\varepsilon}(f\oplus g - C)\right)\cdot(\mu\otimes\nu). \tag{6}$$

Considering small increments in the space of probability Radon distributions ie $\nu_h := \nu_0 + h\delta\nu$, where $(\nu_0, \delta\nu)$ are two measures in $\mathcal{M}^+(X)$ and $h\in\mathbb{R}$.With this definition, when $h\to 0$, $\nu_h \rightharpoonup \nu$ (where $\rightharpoonup$ denotes the *weak convergence*). The ratio $\Delta_h = \frac{OT_\varepsilon(\mu_h,\nu_h) - OT_\varepsilon(\mu,\nu)}{h}$, can be lower-bounded knowing that $OT_\varepsilon(\mu_h,\nu_h)$ is the solution of the dual problem (4) for μ_h, ν_h. In fact, it is the supremum for all $f, g\in C(X)$, included for (f^*, g^*) solution of the dual problem for μ,ν. For the upper bound the idea is similar, The solution (f_h, g_h) of $OT_\varepsilon(\mu_h,\nu_h)$ needs to be considered. By combining the two bounds, and thanks to the uniform convergence of f_h, g_h towards f, g when $h\to 0$, we find that $\Delta_h \xrightarrow{h\to 0} \langle f-\varepsilon, \delta\mu\rangle + \langle g-\varepsilon,\delta\nu\rangle$. Using the translation invariance of the optimal transport solutions, ε can be removed. Thus, the gradient of OT_ε is defined as:

$$\nabla OT_\varepsilon = (f, g). \tag{7}$$

As $OT_\varepsilon(\mu,\nu)$ is a maximum of inner products, $OT_\varepsilon(\mu,\nu)$ is convex with respect to the marginal variables μ and ν. S_ε is thus convex with respect to the inputs as a sum of OT_ε functions and **Th.2** is verified.

We can use the convex inequality to find a lower bound of $OT_\varepsilon(\mu,\nu)$ given the gradient of OT_ε,

$$\begin{aligned} OT_\varepsilon(\mu,\mu) + \langle \nu-\mu, \nabla_2 OT_\varepsilon(\mu,\mu)\rangle &\le OT_\varepsilon(\mu,\nu),\\ OT_\varepsilon(\nu,\nu) + \langle \mu-\nu, \nabla_1 OT_\varepsilon(\nu,\nu)\rangle &\le OT_\varepsilon(\mu,\nu). \end{aligned} \tag{8}$$

To simplify the expression, let's define $F_\varepsilon : \mu \mapsto OT_\varepsilon(\mu,\mu)$. Given (7), F_ε is differentiable. We also define Bregman's divergence [4] for F_ε, and two distributions $\mu,\nu\in\mathcal{M}^+(X)$ as follows :

$$D_F(\mu,\nu) := F_\varepsilon(\mu) - F_\varepsilon(\nu) - \langle \mu-\nu, \nabla F_\varepsilon(\nu)\rangle. \tag{9}$$

We can establish the three point lemma [21] on this divergence and build its symmetric expression:

$$D_F(\mu,\nu) = \frac{1}{2}\langle \mu-\nu, \nabla F_\varepsilon(\mu) - \nabla F_\varepsilon(\nu)\rangle \ge 0. \tag{10}$$

We deduce from (8) that $S_\varepsilon \ge D_F \ge 0$ and thus $S_\varepsilon(\mu,\nu) = 0 \Rightarrow D_F(\mu,\nu) = 0$. **Th.1** is then verified.

The key arguments of the proof of **Th.3** are detailed in the appendix.

2.3 Sinkhorn's algorithm

In practice, the probability distribution μ and ν are defined by their empirical measures i.e. $\mu = \sum_{i=1}^{N} \mu_i \delta_{\mathbf{x}_i}$, where $x := (\mathbf{x}_i)_{0\le i\le N}$ is the family of atoms that corresponds to the points of $X = \mathbb{R}^d$, with d a natural integer, where the probability measure is defined. We can then map the distributions by their empirical density coefficients: $\mu \simeq (\mu_i)_{0\le i\le N}$, $\nu \simeq (\nu_j)_{0\le j\le M}$ and their respective observation points $x\in\mathbb{R}^{N\times d}$ and $y\in\mathbb{R}^{M\times d}$.

In that setting, the dual solution (f^*, g^*) belongs to $\mathbb{R}^N\times\mathbb{R}^M$, and the transportation plan π^* belongs to $\mathbb{R}^{N\times M}$. Knowing that $X = \mathbb{R}^d$ is metric and compact, we also introduce the positive universal kernel $k_\varepsilon(x,y) := \exp\left(-\frac{C(x,y)}{\varepsilon}\right)$, and its fitted kernel matrix $K = \left(k_\varepsilon\left(\mathbf{x}_i, \mathbf{y}_j\right)\right)_{\substack{1\le i\le N\\ 1\le j\le M}}$. The expression (6) can be rewritten as:

$$\pi^* = \mathrm{diag}(\exp(f^*/\varepsilon))\; K\; \mathrm{diag}(\exp(g^*/\varepsilon)). \tag{11}$$

We will use the notation $(u^*, v^*) = (\exp(f^*/\varepsilon), \exp(g^*/\varepsilon))$ in the next paragraph for simplicity.

We can easily build a numerical scheme to solve the problem. The scheme is usually referred as the **Sinkhorn's algorithm**. Throughout a sequence of iterations $t\in\{0,\dots,T\}$, referred as the *Sinkhorn iterations*, we build a sequence of transportation plans $(\pi_t)_{t\in\{0,\dots,T\}}$.
The iterations can be summed up as follows, for each $t\in\{1,\dots,T\}$, starting from $(u^{(0)}, v^{(0)}) = (\mathbb{1}_N, \mathbb{1}_M)$, following [15]:

$$u^{(t+1)} \leftarrow \frac{\mu}{Kv^{(t)}}, \qquad v^{(t+1)} \leftarrow \frac{\nu}{K^T u^{(t+1)}}. \tag{12}$$

The whole idea behind the algorithm is to update one of the vector given the past value of the other one under the boundary constraints (μ,ν) . Indeed, these iterations come from the fact that $u^*\odot(Kv^*) = \mu$ and $v^*\odot(K^T u^*) = \nu$, where $\odot$ denotes the Hadamard product. We note that this approach is GPU-friendly as all operations are tensor-based.

Furthermore, if we consider the symmetric case i.e. we want to compute the transportation plan for $OT_\varepsilon(\mu,\mu)$, the iterations become:

$$u^{(t+1)} \leftarrow \sqrt{u^{(t)}} \odot K\left(\mu\odot u^{(t)}\right). \tag{13}$$

We now have an efficient way to compute the Sinkhorn divergence numerically.

2.4 Vector Field Flow

A **vector field flow** describes the evolution of points in a space according to a given vector field $V : \mathbb{R}^d \to \mathbb{R}^d$. Specifically, the trajectory of a point : $t\in\mathbb{R}^+ \mapsto x(t)\in\mathbb{R}^d$ is governed by the ordinary differential equation:

$$x'(t) = V(x(t)), \qquad t\in\mathbb{R}^+,$$

where the vector field $V(x)$ determines the velocity at every point x. This framework is fundamental in dynamical systems and

physics, where the vector field represents the direction and magnitude of motion at each location in the space.

2.4.1 Gradient Flow. A particular case of vector field flow arises when the vector field is defined as the negative gradient of a scalar-valued differentiable function $F : \mathbb{R}^d \to \mathbb{R}$. In this case, the evolution equation becomes:

$$x'(t) = -\nabla F(x(t)), \qquad t \in \mathbb{R}^+,$$

and the trajectory x is called the **gradient flow** of F. Gradient flow provides a natural framework for minimizing F over time. Starting from an initial point $x(0) = x_0$, solving this differential equation yields under simple conditions a continuous trajectory that converges to a critical point of F, often a local minimum.

2.4.2 Gradient Flow for Distributions. Now, consider a scenario where we wish to minimize a loss function defined between two discrete probability distributions. Let the target distribution be

$$\beta = \frac{1}{n}\sum_{k=1}^{n} \delta_{y_k},$$

where $y_k \in \mathbb{R}^d$ are fixed points, and let the evolving distribution be

$$\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)},$$

where $x_i(t) \in \mathbb{R}^d$ are the time-dependent points forming the distribution. To align $\alpha(x)$ with β, we minimize a loss function $\mathcal{L}(\alpha(x), \beta)$ between the two distributions. The gradient flow for this loss is then given by:

$$x_i'(t) = -\nabla_i \mathcal{L}(\alpha(x(t)), \beta), \quad i = 1, \dots, n.$$

This equation governs the evolution of the point cloud $\{x_i(t)\}_{i=1}^n$, ensuring that the empirical distribution $\alpha(x)$ progressively aligns with the target β.

2.4.3 Scope of Study. In this work, we will study both vector field flow and gradient flow in the context of discrete distributions. Specifically, we focus on the dynamics of point cloud evolution under these frameworks and analyze how different vector fields and loss functions influence the alignment process. Our aim is to explore the interplay between these approaches and their effectiveness in transforming discrete measures in high-dimensional or geometrically complex settings.

3 LIMITS AND NEW DEVELOPMENTS

In this section, we aim to highlight some limitations of the approach proposed by Feydy et al. [8]. We want to extend Euclidean-based methods for both the Sinkhorn iterates and the gradient flow. The goal is to build a Sinkhorn divergence that takes as a cost function a shortest path length induced kernel based on a Riemannian manifold, as well as, continuous transformation from two point distributions following the manifold local geometry.

Furthermore, we will assume that we can sample an evenly distributed set of $N \in \mathbb{N}^*$ points that belongs to the manifold in order to numerically approximate its structure. We construct a graph $G = (V, E)$, with V the set of vertices and E a set of edges connecting the $k \in \mathbb{N}^*$ nearest neighbors, to approximate the underlying topological structure of the manifold.

3.1 Geodesic Sinkhorn algorithm

3.1.1 General case.
Let us assume the probability Radon distributions to be embedded in a more intricate space than the entire space $X = \mathbb{R}^d$. For $d = 2$, applying the gradient flow based on the Sinkhorn loss in the setting of Fig. 1 (obtained using [8] Python package[1]) yields a Euclidean transformation that may not be the most suited for the distribution. Whereas in Fig. 1b, if we consider the angular parametrization of points over the unit circle, we can use the distance over the angles $N_1(\theta_1, \theta_2) := |\theta_1 - \theta_2|$ as the cost function in the Sinkhorn divergence.

Given two probability Radon distributions μ and ν that lie on the same Riemannian manifold. The objective of this part is to consider the constrained construction of a transportation plan π found in (1), so that the algorithm gives weight vectors adapted to the geodesic distance on the manifold.
We assume that each chosen probability Radon measures μ and ν in this section are embedded on the same Riemannian manifold $\mathfrak{M}$ and let $d_{\mathfrak{M}}$ be the geodesic distance on the manifold defined as

$$d_{\mathfrak{M}} : (x, y) \mapsto \inf\left\{L(\gamma), \quad \gamma \in C^\infty([0,1], \mathfrak{M}), \gamma(0) = x,\ \gamma(1) = y\right\},$$

where $L : \gamma \mapsto \int_0^1 \|\dot{\gamma}(t)\| \mathrm{d}t$.

We then define the natural cost $C : (x, y) \mapsto d_{\mathfrak{M}}(x, y)^p$, with $p \geq 1$.

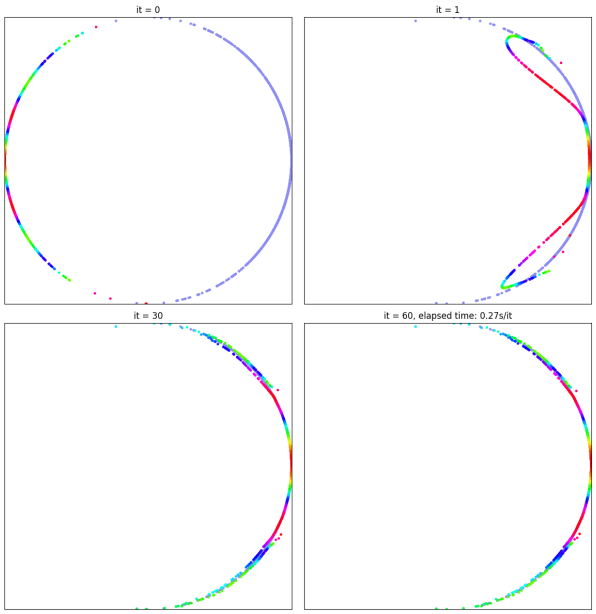
3.1.2 Approximating the geodesic distance.
Determining the geodesic distance is usually difficult in the continuous setting, thus, approximating the distance using point-wise estimations can be necessary. Several approaches exist, a notable one consists in approximating the geodesic distance using Varadhan's state heat approximation [20]:

$$d_{\mathfrak{M}}(x, y)^2 = \lim_{t \to 0^+} -4t \log(H_t(x, y)), \tag{14}$$

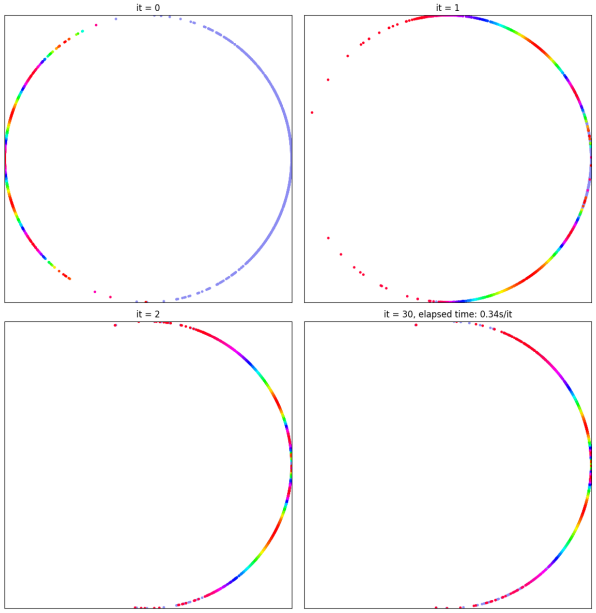
where H_t is called the *heat kernel*. The geodesic distance of a manifold can thus be recovered as a heat transfer problem when the time factor t tends to 0. The two main difficulties with this formula are determining the limit when $t \to 0^+$ as well as approximating the heat kernel. The next section will tackle approximation of the geodesic distance using the heat-kernel distance $d_{H_t}(x, y)^2 := -4t \log(H_t(x, y))$.

We will use a similar implementation as Huguet et al. [12]. Based on the graph G, we can then compute an affinity matrix $A = (A_{i,j})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}}$ based on the $\mathbb{Z}/2\mathbb{Z}$ adjacency matrix on G and the degree matrix D defined as $D_{i,i} := \sum_{j=1}^{n} A_{i,j}$, for all $i \in \{1 \dots, n\}$ and equal to 0 elsewhere. We then compute the Laplacian of the graph

[1] https://www.kernel-operations.io/geomloss/



(a) Gradient flow iterations over $X = \mathbb{R}^2$ for two distributions embedded in the unit circle.



(b) Gradient flow iterations for two distributions embedded on the unit circle using the geodesic $d_\circ(\theta_1, \theta_2) = |\theta_1 - \theta_2|$.

Figure 1: Manifold-agnostic approach vs. manifold-aware approach. The first approach takes an unexpected shape before converging towards the other distribution the second one keeps the geometric properties of the distribution i.e. for each iteration the distribution is on the unit circle.

defined as such $L := D - A$. Solving the heat equation on the graph (V, E), simplifies into solving the ODE:

$$\frac{\mathrm{d}f}{\mathrm{d}t} + Lf = 0 \qquad f(0) = f_0 \in \mathbb{R}. \tag{15}$$

The solution kernel to (15) is the heat kernel $H_t = e^{-tL}$ and thus the solution of the heat equation is $f(t) = H_t f_0$.

Calculating H_t can be computational-heavy as the eigendecomposition of e^{-tL} is in $O(n^3)$. Several methods to approximate the exponential of a matrix exist, notably, orthogonal polynomial approximations are good candidates for the task. Huguet et al. [12] provides a proof of convergence for an approximation using Chebyshev's polynomials $(T_p)_{p\in\mathbb{N}}$. We have the following series with $(b_p)_{p\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ [19]:

$$H_t = \frac{b_0}{2} I_n + \sum_{p=1}^{\infty} b_p T_p(t(I_n - L)). \tag{16}$$

The sum of Chebyshev polynomials in (16) can be truncated up to $P \in \mathbb{N}$ in order to be numerically approximated, we will denote this approached polynomial as $\hat{T}_P(t, L)$. With this in mind, the algorithm detailed in (12) can be re-expressed as such[2]:

$$u^{(t+1)} \leftarrow \frac{\mu}{\hat{T}_P(t, L)v^{(t)}}, \qquad v^{(t+1)} \leftarrow \frac{\nu}{\hat{T}_P(t, L)u^{(t+1)}}. \tag{17}$$

A point of attention with this method lies in the Sinkhorn's iterations, indeed the matrix $\hat{T}_P(t, L)$ must be of the same size as μ and ν. Thus, sparse distributions (i.e. the default distributions with added zeros) are usually required. On a more practical note, a substantial amount of points are required to sample a manifold structure while only a fraction of these points may be used by the mass distributions μ and ν.

3.2 Vector Field Geodesic Awareness

We introduce a novel manifold-aware vector field that extends the gradient field of the Euclidean squared norm distance to incorporate manifold-specific information.

Our goal is to create a manifold-aware transformation from a given starting point $x \in X$ to a specific goal $y \in X$.

Notations:

- The tilde operation $\tilde{x}$ normalizes a vector:

$$x \in \mathbb{R}^n \setminus \{0\} \mapsto \tilde{x} := \frac{x}{\|x\|}.$$

- $\pi(x)$ denotes an Euclidean projection of x onto the manifold M.

In order to do so, we decompose our vector field into three basic components:

- **Euclidean gradient:** The Euclidean distance negative gradient is defined as:

$$-\nabla_x d^{\mathrm{euc}}(x, y) = 2(y - x).$$

[2]Implementation of Huguet et al. [12] is available here: https://github.com/KrishnaswamyLab/GeoSinkhorn. Our modified implementation is also available at: https://github.com/kerrianlc/Sinkhorn.

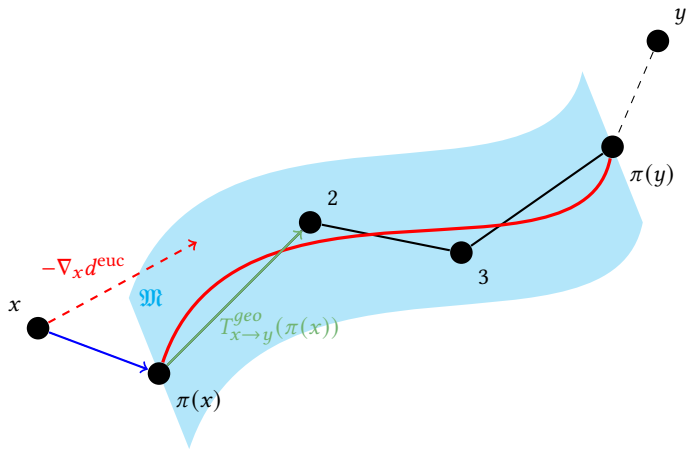


Figure 2: Intuition between the two path concepts. The dashed red arrow represents the Euclidean term, the blue arrow corresponds to the manifold projection term and the green arrow represent the tangent $T^{geo}_{x\to y}(\pi(x))$.

This term is useful to make sure that the trajectory converges towards y.

- **Projection onto the manifold:** We will make use of the projection direction on our manifold : $\pi(x) - x$. It makes our points move towards the manifold.

- **Geodesic-directed tangent:** The aim of this term is to use the structure of the manifold to guide our transformation. By computing the shortest path in G between $\pi(x)$ and $\pi(y)$, we will use the direction of the shortest path between $\pi(x)$ and $\pi(y)$ starting at $\pi(x)$. We will denote the tangent following the geodesic path between $\pi(x)$ and $\pi(y)$ as $z \mapsto T^{geo}_{x\to y}(z)$, the mapping is defined for all points of the path.
$T^{geo}_{x\to y}(\pi(x))$ aligns the movement from x to y with the movement of $\pi(x)$ to $\pi(y)$ on the manifold.

  One could draw a parallel between this discrete approach and the continuous one as mentioned in appendix 5.

A representation of the decomposition of the vector field is illustrated in Fig. 2.

The resulting vector field on $\mathbb{R}^n \times \mathbb{R}^n$ is defined as:

$$F(x,y) = \alpha\,\widetilde{2(y-x)} + \beta\,\widetilde{\pi(x)-x} + \gamma\,\widetilde{T^{geo}_{x\to y}(\pi(x))},$$

where the coefficients α, β and γ satisfy $\alpha + \beta + \gamma = 1$. These terms represent the contributions of the Euclidean gradient, the projection gradient, and the tangent gradient, respectively. It enables us to control the weight of each "force".

We can show the convergence of a numerical vector field scheme following the vector field flow.

Proposition 3.1. *Let $x, y \in \mathbb{R}^n$ and $\alpha > \frac{1}{2}$.*
Let's define the sequence x_n as follows :

$$\begin{cases} \forall n \in \mathbb{N}, \quad x_{n+1} = x_n + \eta_n F(x_n, y), \\ x_0 = x. \end{cases}$$

We define $D_n := \dfrac{\left(\alpha - \frac{1}{2}\right)^2}{\|y - x_n\|^2}$.

- *For $\delta \le \max_{k\in\mathbb{N}} D_k^2$, if for $n \in \mathbb{N}$,*
  $\eta_n \in \left[D_n - \sqrt{D_n - \delta}, D_n + \sqrt{D_n + \delta}\right]$,
  x_n converges towards y with at least a geometrical speed of $\sqrt{1-\delta}$.
- *if for $n \in \mathbb{N}$, $\eta_n \le 2D_n$,*
  then $\|y - x_n\|$ is a decreasing function.

(Proof in Appendix 5.)

Based on the proposition 3.1, with suitable hyperparameters our vector field flow method converges. However, in certain cases, it may be preferable to place greater emphasis on the coefficients associated with the structure of the Riemannian manifold.

4 EXPERIMENTS

4.1 Geodesic Sinkhorn toy example

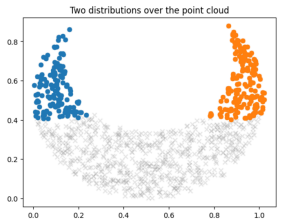
In order to evaluate the contribution of choosing a geodesic distance instead of the standard Euclidean distance on X, we implemented some toy examples. We first sample ($\ge$ 1000 points) a manifold and, on this sample, we define two smaller distributions. The first example (Fig. 3a) is defined by two colored distributions on the "moon crescent" manifold ends. The second example (Fig. 3b) is defined by a non-connected manifold with two distributions on the two connected components.

The results of the Sinkhorn's algorithm using the Wasserstein distance are plotted on Fig. 3c, Fig. 3d. We assigned to each point of the two distributions the coefficients of the converged vectors (u^*, v^*) given by the equation (11). We note that the coefficients given to the points are proportional to the Euclidean distance between the points of the two distributions, i.e. the closer the point is to the other distribution, the greater the vector coefficient is.

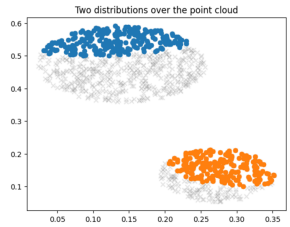
The results based on the heat kernel distance reflects the local Riemannian manifold structure. The most distant point using the geodesic distance are the ones on the tips the moon crescents, as observed in Fig. 3e, they are the points with the lowest attributed weight. For the second example (Fig. 3f), by fixing the number of neighbors k in the graph to be small, the distance between the two blobs cannot be computed, thus the coefficients have not converged. This is to be expected as we are not in the conditions required by the theorem 2.1. In fact, the cost function is not Lipschitz because the geodesic distance can be equal to ∞ for low values of k in a non connected setting.

4.2 Vector field flow experiments

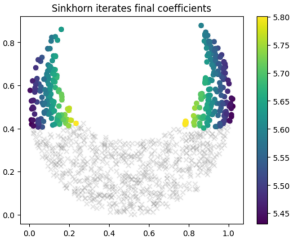
We conducted experiments to evaluate our manifold-aware vector field approach on the MNIST dataset. we aim to determine if our



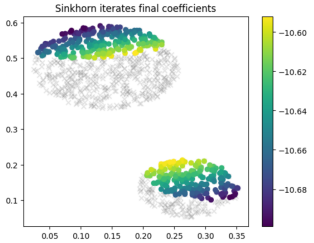
(a) Distributions over a moon crescent manifold respectively represented by the orange and blue point clouds.



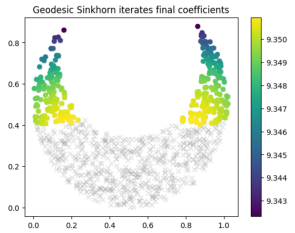
(b) Distributions over two Gaussian blobs.



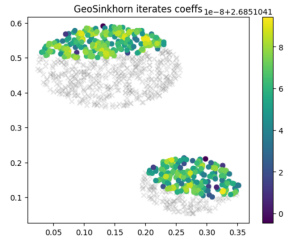
(c) Coefficients of the Sinkhorn iterates given on the two distributions using the Wasserstein distance.



(d) Coefficients of the Sinkhorn iterates given on the two distributions using the Wasserstein distance.

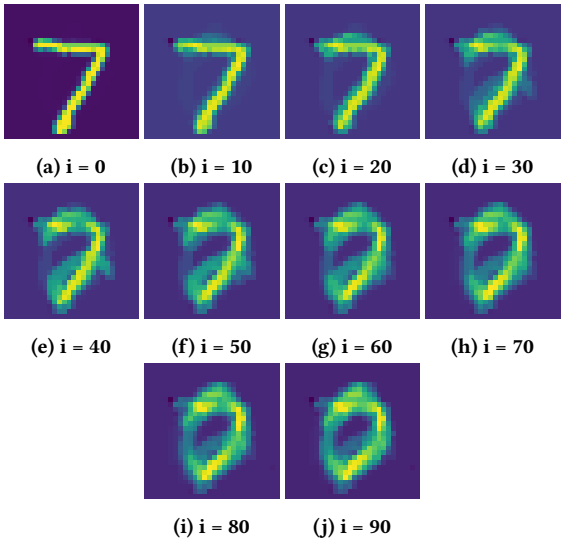


(e) Coefficients of the Sinkhorn iterates given on the two distributions using the geodesic distance on the manifold ($k = 350$).



(f) Coefficients of the Sinkhorn iterates given on the two distributions using the geodesic distance on the manifold ($k = 5$).

Figure 3: Toy example results for different manifold-based distributions.

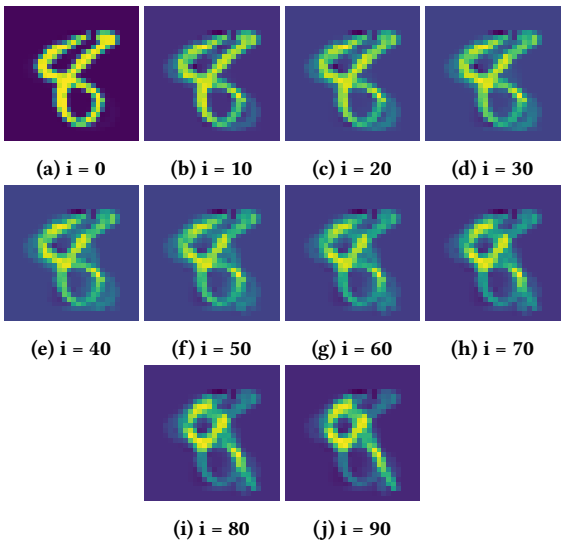


(a) i = 0 (b) i = 10 (c) i = 20 (d) i = 30

(e) i = 40 (f) i = 50 (g) i = 60 (h) i = 70

(i) i = 80 (j) i = 90

Figure 4: Transformation from digit '7' to '0' using our vector field with $k = 15$ for the graph kNN construction. Each subfigure illustrates the result of the gradient flow at a specific iteration, starting from the initial digit ('7') at iteration 0 and progressively moving toward the target digit ('0') by iteration 90. The intermediate steps demonstrate the smooth trajectory going from $7 \longrightarrow 2 \longrightarrow 0$.



(a) i = 0 (b) i = 10 (c) i = 20 (d) i = 30

(e) i = 40 (f) i = 50 (g) i = 60 (h) i = 70

(i) i = 80 (j) i = 90

Figure 5: Transformation from digit '8' to '9' using our vector field with $k = 15$ for the graph kNN construction. Subfigures show the outcomes of the gradient flow across iterations, beginning with the initial digit ('8') at iteration 0 and converging to the target digit ('9') by iteration 90. The experiment challenges the expected results as the intermediate steps appear to be an amalgamation of digits '8' and '9' rather than a clear digit.

method successfully produces the expected manifold aware transformation from a starting image to a target image both excluded from the sampled dataset. The vector field parameters were set to $\alpha = 0.4$, $\beta = 0.2$, and $\gamma = 0.4$, with a learning rate (lr) of 0.1. vector field vector were renormalized at each step to match the norm $\|y - x\|$ (with $\eta_n = \text{lr} \times \frac{\|y - x_n\|}{\|F(x_n, y)\|}$).

As shown in Fig. 4 and Fig. 5, the gradient flow exhibits smooth, progressive movement through the data's intrinsic manifold structure, transitioning between digits while adhering to the local geometry and geodesic paths.

4.3 Modified Sinkhorn loss

We explore the use of gradient flow on the Sinkhorn loss, both with the classical Euclidean distance and a modified version of the gradient. The classical Sinkhorn loss is optimized using the standard Euclidean gradient, which assumes a direct, straight-line path between points. However, to better account for the manifold structure of the data, we modified Euclidean loss gradient to incorporate our constructed vector field (Section 3.2). This modified gradient, termed the manifold-aware gradient, replaces the straight-line Euclidean path with trajectories that follow the underlying manifold of the data. By aligning the Sinkhorn loss optimization process with the data geometry, it leads to more natural and interpretable transformations. Fig. 6 demonstrates the stark contrast between the two methods.

5 CONCLUSION

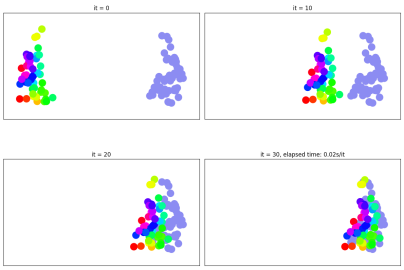
In this work, we introduced a novel method to incorporate the structure of a manifold into optimization and distance computation frameworks.

On the one hand, we focused on approximating geodesic distances on discrete manifolds using Varadhan's heat kernel formulation and graph-based methods. By leveraging k-nearest neighbor (k-NN) graphs, the Laplacian, and Chebyshev polynomial approximations, we efficiently computed heat kernels and geodesic distances while addressing computational challenges.

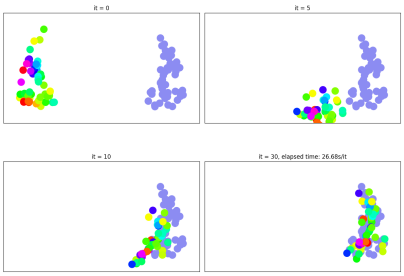
On the other hand, we proposed a manifold-aware vector field approach, combining Euclidean, distance-to-manifold, and tangent gradients into a unified framework. This allows optimization paths to follow the intrinsic manifold structure. The approach was validated on toy datasets and real-world examples such as MNIST, demonstrating that the vector field flow adheres to the manifold geometry.

Future research could extend this framework by comparing two parametrized density distributions instead of considering mass distributions. We could extend the experiments of the vector field flow to Riemannian manifold where geodesic distance and shortest path tangent can be computed, and thus removing the graph approximation.

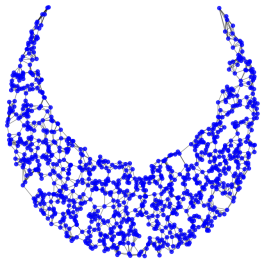
The appendix 5 provides a glimpse into a framework that unifies the discrete and continuous approaches. Studying its structure could entail interesting results regarding the convergence of similar schemes.



(a) Evolution of data points using Euclidean gradient flow.



(b) Evolution of data points using manifold-aware gradient flow. As expected, the point cloud goes through the moon crescent.



(c) The considered k-NN graph ($k = 7$) illustrating the dataset used for the manifold-aware gradient

Figure 6: Evolution of data points using the modified Sinkhorn loss with gradient flow. The subfigures (a) and (b) show the alignment of points using Euclidean and manifold-aware gradient flows, respectively, while (c) displays the k-NN graph emphasizing local geometric relationships.

REFERENCES

[1] One Point Isometric Matching with the Heat Kernel. 29. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2010.01764.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2010.01764.x.

[2] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/491442df5f88c6aa018e86dac21d3606-Abstract.html.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman Projections for Regularized Transportation Problems, December 2014. URL http://arxiv.org/abs/1412.5154. arXiv:1412.5154.

[4] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, January 1967. ISSN 0041-5553. doi: 10.1016/0041-5553(67)90040-7. URL https://www.sciencedirect.com/science/article/pii/0041555367900407.

[5] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018. URL https://arxiv.org/abs/1804.03599.

[6] 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. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[7] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via Maximum Mean Discrepancy optimization, May 2015. URL http://arxiv.org/abs/1505.03906. arXiv:1505.03906.

[8] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018. URL https://arxiv.org/abs/1810.08278.

[9] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a Wasserstein Loss, December 2015. URL http://arxiv.org/abs/1506.05439. arXiv:1506.05439.

[10] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences, 2017. URL https://arxiv.org/abs/1706.00292.

[11] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences, 2019. URL https://arxiv.org/abs/1810.02733.

[12] Guillaume Huguet, Alexander Tong, María Ramos Zapatero, Christopher J. Tape, Guy Wolf, and Smita Krishnaswamy. Geodesic sinkhorn for fast and accurate optimal transport on manifolds, 2023. URL https://arxiv.org/abs/2211.00805.

[13] L. Kantorovitch. On the Translocation of Masses. *Management Science*, 5(1):1–4, 1958. URL http://www.jstor.org/stable/2626967.

[14] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing Neural Networks by Penalizing Confident Output Distributions, January 2017. URL http://arxiv.org/abs/1701.06548. arXiv:1701.06548.

[15] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

[16] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[17] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000. ISSN 1573-1405. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054.

[18] Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, and Hamed Hassani. Sinkhorn natural gradient for generative models, 2020. URL https://arxiv.org/abs/2011.04162.

[19] David I Shuman, Pierre Vandergheynst, and Pascal Frossard. Chebyshev polynomial approximation for distributed signal processing. In *2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, pages 1–8, June 2011. doi: 10.1109/DCOSS.2011.5982158. URL https://ieeexplore.ieee.org/document/5982158/?arnumber=5982158. ISSN: 2325-2944.

[20] Sathamangalam R Srinivasa Varadhan. On the behavior of the fundamental solution of the heat equation with variable coefficients. *Communications on Pure and Applied Mathematics*, 20(2):431–455, 1967.

[21] Yi Zhou, Yingbin Liang, and Lixin Shen. A Simple Convergence Analysis of Bregman Proximal Gradient Algorithm, December 2017. URL http://arxiv.org/abs/1503.05601. arXiv:1503.05601.

APPENDIX

Proof of Th.3

- The weak convergence of a sequence $(\mu_n)_{n\in\mathbb{N}}$ towards $\mu \in \mathcal{M}_1^+$, implies that $S_\varepsilon(\mu_n, \mu) \to 0$ following the positive definite characteristic of S_ε.
- For a given sequence $(\mu_n)_{n\in\mathbb{N}^*}$ and μ in $\mathcal{M}_1^+(X)$, assuming that $S_\varepsilon(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$, we use a weak continuity argument to extract a sub-sequence $(\mu_{n_k})_{k\in\mathbb{N}^*}$ that weakly converges towards a distribution μ_∞, and by the positive definiteness of S_ε, we find that $\mu_\infty = \mu$, and we conclude that $(\mu_n)_{n\in\mathbb{N}}$ weakly converges towards μ.

Useful lemmas for Proposition 3.1

LEMMA 5.1. *for $\alpha > 0.5$, $(x, y) \in X^2$, $u \in \mathbb{R}^d$ such that $\|u\| \leq \frac{1}{2}$*

$$\langle y - x, \alpha \frac{(y-x)}{\|(y-x)\|} + u\rangle \geq (\alpha - \frac{1}{2})\|y - x\|$$

PROOF.
Using Cauchy-Schwarz inequality,

$$|\langle y - x, u\rangle| \leq \|y - x\|\|u\| \leq \frac{1}{2}\|y - x\|$$

$$\langle y - x, \frac{1}{2}\frac{(y-x)}{\|(y-x)\|} + u\rangle = \frac{1}{2}\|y - x\| + \langle y - x, u\rangle \geq 0$$

$$\langle y - x, F(x, y)\rangle = (\alpha - \frac{1}{2})\|y - x\| + \langle y - x, \frac{1}{2}\frac{(y-x)}{\|(y-x)\|} + \beta u + \gamma v\rangle \geq (\alpha - \frac{1}{2})\|y - x\|$$

To summarize :

$$\langle y - x, \alpha \frac{(y-x)}{\|(y-x)\|} + u\rangle \geq (\alpha - \frac{1}{2})\|y - x\|$$

□

LEMMA 5.2. *Let $x \neq y \in \mathbb{R}^d$, $u \in \mathbb{R}^d$ such that $\|u\| \leq \frac{1}{2}$.*

if we define $x' := x + \eta\alpha\frac{y-x}{\|y-x\|}$ and $D := \frac{(\alpha - \frac{1}{2})}{\|y-x\|}$

(1) for $\eta \leq D^2$:

$$\|y - x'\| \leq \|y - x\|$$

(2) for $\delta \leq D^2$, for $\eta \in [D - \sqrt{D^2 - \delta}, D + \sqrt{D^2 + \delta}]$:

$$\|y - x'\|^2 \leq (1 - \delta)\|y - x\|^2$$

The figure 7 represent the result in a graphical way.

PROOF. Let us develop $\|y - x'\|^2$:

$$\|y - x'\|^2 - 2\eta\langle y - x, \alpha\frac{y-x}{\|y-x\|} + u\rangle + \eta^2\|(\alpha\frac{y-x}{\|y-x\|} + u))\|^2$$

Let $\delta \geq 0$, we need to find a reformulation of our aimed inequality.

$$\|y-x'\|^2 \leq (1-\delta)\|y-x\|^2 \quad (*)$$
$$\iff -2\eta\langle y-x, (\alpha\frac{y-x}{\|y-x\|}+u)\rangle + \eta^2\|(\alpha\frac{y-x}{\|y-x\|}+u)\|^2 \leq -\delta\|y-x\|^2$$
$$\iff (\delta+\eta^2)\|y-x\|^2 \leq 2\eta\langle y-x, (\alpha\frac{y-x}{\|y-x\|}+u)\rangle$$

we know from (5) that : $2\eta\langle y-x, \alpha\frac{y-x}{\|y-x\|}+u\rangle \geq 2\eta(\alpha-\frac{1}{2})\|y-x\|$

Using the inequality above, The following inequality is sufficient to get $(*)$:

$$(\delta+\eta^2)\|y-x\|^2 \leq 2\eta(\alpha-\frac{1}{2})\|y-x\|$$

ie

$$(\delta+\eta^2)\|y-x\| \leq 2(1-\frac{1}{2})\eta$$

This is an polynomial inequality of degree 2 in η. Let's define

$$D := \frac{(\alpha-\frac{1}{2})}{\|y-x\|}$$

To solve this problem for η, we need to have a positive discriminant ie $(\alpha-\frac{1}{2})^2 \geq \delta\|y-x_n\|^2$ ie $\delta \leq D^2$.

The inequality holds for η in :
$I := [D-\sqrt{D^2-\delta}, D+\sqrt{D^2-\delta}]$
To conclude, For any $\delta \leq D^2$, for $\eta \in I$, we have :

$$\|y-x'\|^2 \leq (1-\delta)\|y-x\|^2$$

for $\delta = 0$, for $\eta \leq 2D$, we have :

$$\|y-x'\| \leq \|y-x\|$$

□

Proof of proposition 3.1

Proof. Let us recall the formula of our vector field:

$$F(x,y) = \alpha\frac{2(y-x)}{\|2(y-x)\|} + \beta\frac{\pi(x)-x}{\|\pi(x)-x\|} + \gamma\frac{T_y(x)-\pi(x)}{\|T_y(x)-\pi(x)\|}.$$

let $u := \beta\frac{\pi(x)-x}{\|\pi(x)-x\|} + \gamma\frac{T_y(x)-\pi(x)}{\|T_y(x)-\pi(x)\|}$
by triangular inequality : $\|u\| \leq \beta+\gamma \leq \frac{1}{2}$
We can apply lemma 5.2
with $D = \frac{\alpha-\frac{1}{2}}{\|y-x\|}$,
for $\eta \in [D-\sqrt{D^2-\delta}, D+\sqrt{D^2-\delta}]$

$$\|y-x-\eta F(x,y)\|^2 \leq (1-\delta)\|y-x\|^2$$

and if $\delta = 0$:

$$\|y-x-\eta F(x,y)\|^2 \leq \|y-x\|$$

We conclude by induction. □

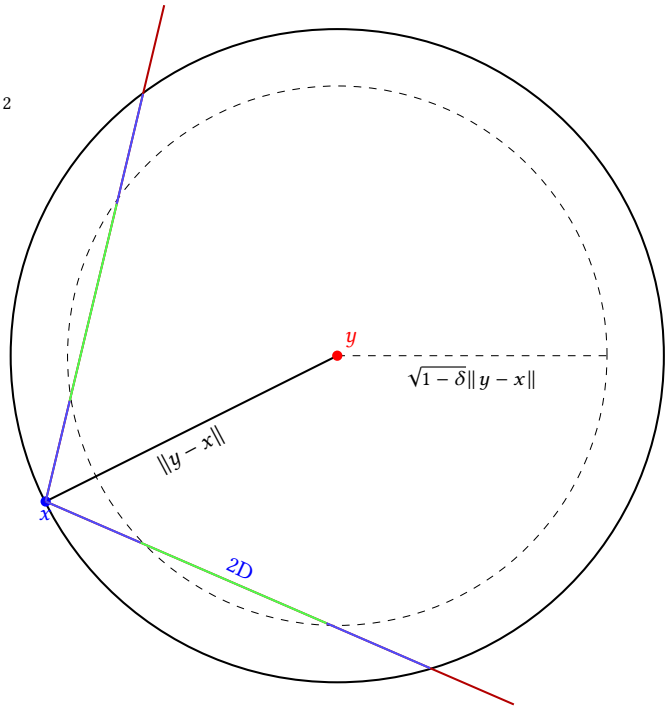


Figure 7: Visualisation of lemma 2. For a given δ, the green region are the range values of η such that we fall in the dashed circle. If $\eta > 2D$, we get too far from y and the distance between x and y increase.

Algebraic Structure

In order to build the concept of geodesic tangent, we define an algebraic structure for subspaces of $\mathbb{R}^n$ that captures the essence of shortest paths and tangents.

Notations:

- $\mathcal{F}(A,B)$: The set of all maps from A to B.
- $C(A,B)$: The set of all continuous maps between two topological sets A and B.

Let $(E, \leq, O)$ be a totally ordered, bounded set included in $\mathbb{R}$ equipped with the order topology O.

Define E^B as the collection of open intervals within E:

$$E^B = \{A \subset E \mid \exists x,y \in A, \forall a \in A, x \leq a \leq y\} = \{(x,y) \mid x,y \in E\}.$$

We consider M, a subspace of $\mathbb{R}^n$. We define a triple (Γ, I, D), where:

- $\Gamma \subset \mathcal{F}(E,M)$ is the set of paths we consider.
- $D : \Gamma \to C(E,\mathbb{R}^n)$ is a linear map (*differentiation operator*).
- $I : E^B \times C(E,\mathbb{R}^n) \to \mathbb{R}^n$ is a map (*integration operator*) with the following properties:
  - For any interval $(x,y) \in E^B$, $I((x,y),\cdot)$ is linear.
  - For any $x,y,z \in E$, $I((x,y),\cdot) + I((y,z),\cdot) = I((x,z),\cdot)$.
- A path $\gamma \in \Gamma$ satisfies:

$$\forall x \leq y \in E, \quad \gamma(x) + I((x,y), D(\gamma)) = \gamma(y).$$

Finally, let $L : \Gamma \to \mathbb{R}^+$ represent the length of a path.

We suppose that for any $u, v \in M$ there exists a shortest path γ_{geo} from u to v such that:

- $\gamma_{\text{geo}}(\min E) = u$, $\gamma_{\text{geo}}(\max E) = v$.
- $L(\gamma_{\text{geo}}) = \inf\{L(\gamma) \mid \gamma \in \Gamma, \gamma(\min E) = u, \gamma(\max E) = v\}$.

The tangent at u, $D(\gamma_{\text{geo}})(\min E)$, represents the initial tangent direction of the shortest path. This tangent direction is crucial for constructing our manifold-aware vector field.

To ground this abstraction, we consider two main examples:

- **Graph Setting:**
  - Let $E = P_f(\mathbb{N})$, and $G = (V, E_{\text{edge}})$ be a graph where $V \subset \mathbb{R}^n$ is the set of vertices and $E_{\text{edge}} \subset V \times V$ is the set of edges.
  - Define $\Gamma = \{\gamma \in \mathcal{F}(\{1, \ldots, k\}, M) \mid \exists k \in \mathbb{N}, \forall i < k, (\gamma(i), \gamma(i+1)) \in E_{\text{edge}}\}$.
  - $D(\gamma) = \gamma(i+1) - \gamma(i)$, and $I([l, m], f) = \sum_{k=l}^{m-1} f(k)$.
  - The length $L(\gamma) = \sum_{k=0}^{\infty} \|\gamma(k+1) - \gamma(k)\|$.

  In this setup, the shortest path γ_{geo} exists and can be computed using Dijkstra's algorithm for example and $D(\gamma_{\text{geo}})(\min_{k \in E} k) = \gamma_{\text{geo}}(1) - \gamma_{\text{geo}}(0)$, representing the tangent.

- **Riemannian Manifold Setting:**
  - Let $E = [0, 1]$, and M be a Riemannian manifold.
  - Γ is the set of C^∞ paths from $[0, 1]$ to M.
  - Define $L(\gamma) = \int_0^1 \|\dot{\gamma}(t)\| \, dt$, I and D as the standard integration and differentiation operator over a curve.

  Assuming M allows minimizing paths, the tangent $D(\gamma_{\text{geo}})(\min E)$ points in the direction of minimizing distance.