Wasserstein Riemannian Frameworks

Octav Dragoi

May 1, 2020

Abstract

The point of this short writeup is to describe how to construct Riemannian-like objects, such as geodesics, tangent spaces, exponential and log maps, etc, over the space of point clouds endowed with the Wasserstein metric.

1 Introduction

1.1 Definitions

Let X be a Riemannian manifold, in our case we can take $X = \mathbb{R}^d$ for some d. In literature, the space $\mathscr{P}_2(X)$ is defined as the space of all Borel measures over X with finite second moments (we need this condition to ensure the existence of the Wasserstein distance).

There are two important ways in which these measures relate, through transport plans and transport maps. We shall define each one in turn. Let $\mu \in \mathscr{P}_2(X), \nu \in \mathscr{P}_2(Y)$.

- A **transport map** is simply a function $T : X \to Y$ for which $\nu = T_\# \mu$. The measure $T_\# \mu \in \mathscr{P}_2(Y)$ is called the *pushforward of* μ *through* T and is defined as:

  $$T_\# \mu(E) = \mu(T^{-1})(E), \; \forall E \subset Y \text{ Borel}$$

  Intuitively, this is a point to point mapping from X to Y that preserves the μ measure on the image set.

- A **transport plan** is a measure $\boldsymbol{\gamma} \in \mathscr{P}_2(X \times Y)$, for which the marginals agree with μ and ν, specifically:

  $$\boldsymbol{\gamma}(A \times Y) = \mu(A), \boldsymbol{\gamma}(X \times B) = \nu(B)$$

  Denote by $\textsc{Adm}(\mu, \nu)$ the set of all transport maps between μ and ν.

  Intuitively, this is a pairing that describes how much weight is transported from each point to possibly more points in the image set. This is a generalization of transport maps, since weight from one point can (and usually does) flow into multiple points. Formally, each transport map T can be expressed as a transport plan:

  $$\nu = T_\#(\mu) \implies \boldsymbol{\gamma} := (Id \times T)_\# \mu \in \textsc{Adm}(X \times Y)$$

1.2 Discussion on point clouds

It is important to see how we can apply this theory to our point cloud embeddings. The theory is defined on *continuous* probability distributions across X, whereas point clouds are inherently discrete. Therefore, we want to restrict ourselves to looking at **measures with finite support**.

Under this constraint, transport maps become quite restrictive. Basically, they only move points to other points, keeping the probability distribution fixed between them, at most merging two points together. We need to see whether, for our purposes, this kind of restriction is workable.

Transport plans, where one point can be mapped to a multitude of other ones, grant us much wider movement space, at the cost of more complex parametrization. We shall see exactly how this is reflected mathematically, in the next section.

2 The weak Riemannian structure of $(\mathscr{P}_2(X), W_2)$

The Wasserstein distance W_2 between μ and ν is defined as:

$$W_2(\mu,\nu) = \sqrt{\inf_{\boldsymbol{\gamma}\in \textsc{Adm}(\mu,\nu)} \int d^2(x,y)d\boldsymbol{\gamma}(x,y)}$$

If X is a Riemannian manifold, as it is in our case where $X = \mathbb{R}^d$, then $(\mathscr{P}_2(X), W_2)$ is also a Riemannian manifold. In [1], the authors describe the topology of $(\mathscr{P}_2(X), W_2)$ and the construction of a Riemannian-like framework by mostly looking at transport maps, not at transport plans.

As defined in [1], the set $\mathscr{P}_2(X)$ contains all distributions, but for our use case we would like to restrict this to probabilities with finite support, where each distribution is a sum of discrete weighted Dirac indicator distributions. Something like:

$$\mathscr{D}_2(X) = \left\{ X = \sum_{i=1}^{n} \lambda_i \delta_{x_i} : n \in \mathbb{N}, \lambda_i \in \mathbb{R}, x_i \in X \right\}$$

Details have to be ironed out, but from the subsequent constructions it looks likely that $(\mathscr{D}_2(X), W_2)$ also shares the same Riemannian-like structure.

2.1 Geodesics and tangent spaces induced by transport maps

If X is a geodesic space, then for $t \mapsto \gamma_t$ a constant speed geodesic between $x, y \in X$, then $t \mapsto \delta_{\gamma_t}$ is a constant speed geodesic over $\mathscr{P}_2(X)$. The natural way of taking geodesics therefore is by this so-called *displacement interpolation*, i.e. taking intermediary points along the geodesic curve between two points. This is good; this process should therefore generate new point clouds, rather than just shifting weight between already defined points.

To illustrate the previous point, let us consider a simple example, with two distributions $\mu = \delta_x, \nu = \delta_y$, for some $x, y \in X$. The geodesic $(\mu_t) \subset \mathscr{P}_2(X)$, for which $\mu_0 = \mu, \mu_1 = \nu$, is then defined accordingly as:

$$\mu_t = \delta_{tx+(1-t)y}$$

Notice that it is not $t\delta_x + (1-t)\delta_y$; this interpolation process actually yields new points from X, rather than just redistributing weight among existing points.

Also notice that, if μ and ν have finite support, then also μ_t will have finite support (i.e. both of them are point clouds in $(\mathscr{D}_2(X), W_2)$).

In generaal, an optimal map T naturally gives rise to (”induces”) a constant speed geodesic $(\mu_t) \subset \mathscr{P}_2(X)$ on $\mathscr{P}_2(X) = \mathscr{P}_2(\mathbb{R}^d)$:

$$\mu_t = ((1-t)Id + tT)_{\#}\mu_0$$

for which, if $\mu_0, \mu_1 \in \mathscr{D}_2$, then $(\mu_t) \subset \mathscr{D}_2$.

Based on this, one can define the ”instantaneous velocity” v_t along this curve:

$$v_t := (T - Id)((1-t)Id + tT)^{-1}$$

which satisfies the continuity equation:

$$\frac{d}{dt}\mu_t + \nabla \cdot (v_t \mu_t) = 0 \tag{1}$$

It is natural to think that integrating the instantaneous velocity along the geodesic curve should give us an indication of the curve's length. The **Benamou-Brenier** formula recovers the Wasserstein distance under this "dynamic" formulation:

$$W_2(\mu^0, \mu^1) = \inf \left\{ \int_0^1 \|v_t\|_{\mu_t} \, dt \right\}$$

where the infimum ranges over all solutions (μ_t, v_t) of (1) which link μ_0 and μ_1. If these vectors v_t are induced by an optimal transport plan, then the infimum is attained there and the integral is indeed equal to the Wasserstein distance.

These vectors v_t of minimal size describe the movement of curves around μ_t, so we want to define the tangent space as the collection of all such vectors, in a rigorous manner. Turns out, we can do this in two equivalent ways, actually.

- The fact that these vectors act only on smooth functions suggests we should consider all gradients of such functions, or, more rigorously:

$$\overline{\{\nabla\phi : \phi \in C_c^\infty(X)\}}^{L_2(\mu)} \tag{2}$$

- Wanting to consider only v_t of minimal norm, from the continuity eqauation 1, one can check that this is equivalent to having v_t in the set:

$$\left\{ v \in L^2(\mu_t) : \int \langle v, w \rangle d\mu_t = 0, \ \forall w \in L^2(\mu_t) \ s.t. \ \nabla \cdot (w\mu_t) = 0 \right\} \tag{3}$$

The expressions from (2) and (3) are the same! Therefore, the tangent space is defined as:

$$\begin{aligned} \mathrm{Tan}_\mu(\mathscr{P}_2(X)) &\coloneqq \overline{\{\nabla\phi : \phi \in C_c^\infty(X)\}}^{L_2(\mu)} \\ &= \left\{ v \in L^2(\mu) : \int \langle v, w \rangle d\mu = 0, \ \forall w \in L^2(\mu) \ s.t. \ \nabla \cdot (w\mu) = 0 \right\} \end{aligned}$$

2.2 Generalization to transport plans

[1] also discuss an alternative definition to the tangent space. Define:

$$\mathrm{Geod}_\mu \coloneqq \{\text{geodesics starting from}\mu\}/\approx$$

where $(\mu_t) \approx (\mu'_t)$ if they coincide on a neighborhood around μ.

The natural distance on this space is:

$$D((\mu_t), (\mu'_t)) = \overline{\lim}_{t\downarrow 0} \frac{W_2(\mu_t, \mu'_t)}{t}$$

Then define the *geometric tangent space* $\mathbf{Tan}_\mu(\mathscr{P}_2(\mathbb{R}^d))$ as the completion of Geod_μ with respect to D.

This description is purposefully vague; since we will likely only use the solutions over transport maps, where computations are usually more tractable and most of the literature is concentrated anyway, this part does not necessarily concern us directly.

2.3 Discussion

The bolded, larger "space of directions" $\mathbf{Tan}_\mu(\mathscr{P}_2(\mathbb{R}^d))$ is larger than the "space of gradients" $\mathrm{Tan}_\mu(\mathscr{P}_2(\mathbb{R}^d))$. For example, if $\mu = \delta_x$ for some $x \in \mathbb{R}^d$, then:

$$\mathrm{Tan}_\mu(\mathscr{P}_2(\mathbb{R}^d)) \sim \mathbb{R}^d$$

however:

$$(\mathbf{Tan}_\mu(\mathscr{P}_2(\mathbb{R}^d)), D) \sim (\mathscr{P}_2(\mathbb{R}^d), W_2)$$

This intuitively happens because when displacing a point with a tangent map, it inevitably gets sent to another point, so the set of all directions is isomorphic to $\mathbb{R}^d$. However, when displacing a point with a transport plan, we can essentially recover any distribution over our space, i.e. the set of all directions is $\mathscr{P}_2(\mathbb{R}^d)$.

We don't even need to go here, though, and this is the reason why that section only throws a furtive glance at general transportation plans. When solving Wasserstein transport plans for point cloud embeddings with an equal number of vertices, we actually retrieve transport maps. That's because the W_2 distance is convex in linear combinations of transport plans, since the L^2 distance is also convex and therefore the minimum will be attained in points on the border of the transportation polytope.

3 Principal Geodesic Analysis - Computing in Practice

The main problem with the aforementioned concepts of tangent spaces and exponential maps is that most of the computations are intractable for an infinite-dimensional space like $\mathscr{P}_2(X)$. In the paper [2], the authors describe a series of approximations for tangent vectors and interpolations in the W_2 space, together with algorithmic implementations for the case of $\mathscr{D}_2(X)$. The paper's goal is to eventually arrive at a PCA-like analysis within Wasserstein spaces, but we shall use it rather for its computational framework.

For computations of Wasserstein barycenters, the authors of [2] use other implementations of other algorithms, such as [3] or [4]. Most of the geometric constructions that they devise then, are actually used to compute the so-called "principal geodesics" and conduct some form of principal component analysis.

3.1 Theory

Let $\{\mu_i : 1 \leq i \leq n, n \in \mathbb{N}\} \subset \mathscr{D}_2(X)$ be a finite set of probability measures with finite support, and let $\bar{\mu}$ be its average within the Wasserstein space. A geodesic through $\bar{\mu}$ is parametrized by a vector $v \in L^2(\bar{\mu})$ on a neighborhood around $\bar{\mu}$ as:

$$g_t(v) \coloneqq (id + tv)_{\#}\bar{\mu}$$

In [5], Lemma 7.2.1. shows that *all* geodesics through $\bar{\mu}$ are parametrized in such a way, so this does not restrict the generality in any way. However, making sure that this is a geodesic is no trivial matter, and proves to be computationally too intensive. Therefore, the authors adopt a relaxed definition of a geodesic, built upon *two* vector fields arount $\bar{\mu}$.

Let $\sigma, \nu, \eta \in \mathscr{P}_2(X)$ and take optimal mappings $T^{(\sigma,\nu)}$ and $T^{(\sigma,\eta)}$ from $\sigma \to \nu$ and $\sigma \to \eta$, respectively. A *generalized geodesic* between ν and η with base σ is defined as:

$$g_t = \left((1-t)T^{(\sigma,\nu)} + tT^{(\sigma,\eta)}\right) \#\sigma,\ t \in [0,1]$$

In our case, for $\sigma = \bar{\mu}$ and v_1, v_2 two vector fields such that $id - v_1, id + v_2$ are optimal mappings, the generalized geodesic can be written as:

$$g_t(v_1, v_2) \coloneqq (id - v_1 + t(v_1 + v_2))\,\#\bar{\mu},\ t \in [0,1]$$

In Figure 1 the authors of [2] illustrate how these generalized geodesics look like, for $t = \frac{1}{3}$ and $t = \frac{2}{3}$. The linear behavior of this kind of interpolation makes it cheaper to compute than the purely theoretical method.

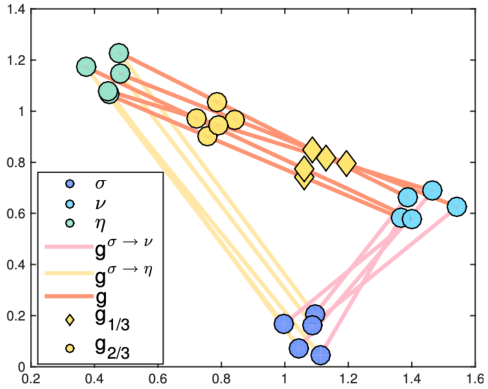

Figure 1: Generalized geodesic interpolation between two measures ν and η using a base measure σ, all within $\mathscr{P}_2(X)$

3.2 Computations in practice

The Wasserstein mean $\bar{\mu}$ of the input measures (μ_i) is given, and we denote $\bar{\mu} = \sum_{k=1}^{p} b_k \delta_{y_k}$, where $\sum b_i = 1$ and $Y = [y_1, \ldots, y_p] \in \mathbb{R}^(d \times p)$ is the matrix containing the support of $\bar{\mu}$. In [2], they treat this $\bar{\mu}$ as given, and they compute it using some other barycentric algorithm.

For each of the p points in $\bar{\mu}$, they consider *two* velocity vectors. These vectors are represented by $V_1 = [v_1^1, \ldots, v_p^1]$ and $V_2 = [v_1^2, \ldots, v_p^2]$ in $\mathbb{R}^{d \times p}$. Under the assumption that these velocity fields yield optimal mappings, the generalized geodesic is defined as:

$$g_t(V_1, V_2) = \sum_{k=1^p} b_k \delta_{z_k^t}$$

where the locations Z_t are computed as:

$$Z_t = [z_1^t, \ldots, z_p^t] \coloneqq Y - V_1 + t(V_1 + V_2)$$

3.3 Discussion

At this point, only the specific setup of the algorithm to construct these geodesics has been summarized here. To me it seems that this algorithm has a good potential to be useful, if we are interested in computing these tangent vectors V_1, V_2 and the principal geodesics. That's a big **if**!

Let us convene and decide what parts of these computations would be useful for us, and how can we adapt them to our use case scenario.

References

[1] Luigi Ambrosio and Nicola Gigli. *A User's Guide to Optimal Transport.* 2011.

[2] Vivien Seguy and Marco Cuturi. *Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric.* 2015.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. *Iterative bregman projections for regularized transportation problems.* 2015

[4] Marco Cuturi and Arnaud Doucet. *Fast Computation of Wasserstein Barycenters.* 2014.

[5] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures.* 2006.