CPSC 661: Sampling Algorithms in Machine Learning (March 22, 2021)

Lecture (14)

Lecturer: Andre Wibisono *Scribe: (Muhammed Yavuz Nuzumlalı)*

1 Introduction

Last time we looked at a very brief overview of optimization on manifold and we defined manifold, tangent space, metric on manifold and so on. Then we looked at continuous time dynamics and gradient flow for minimizing a function on the manifold, along with studying discrete settings via gradient descent and proximal gradient algorithms. Lastly we talked about properties of functions such as strong convexity and gradient domination, and how they imply exponential convergence rage (with smoothness in discrete case).

In this lecture, we will look at what's called the Wasserstein metric on the space of probability distributions and how they arise from optimal transport, and this will be a concrete example of a manifold that we will be working with

Optimal transport and Wasserstein metric are very wide and deep subjects, so we will only look at the very surface level overview of it. For further information, please refer to references [1], [2] and [3].

2 Space of Distributions

Let $\mathcal{P}(\mathcal{X})$ denote the space of probability distributions over a given space $\mathcal{X}$. For example, when is a discrete finite set where $\mathcal{X} = \{1, ..., n\}$, the space of probability distributions becomes the probability simplex:

$$\mathcal{P}(\mathcal{X}) = \Delta_{n-1} = \{p \in \mathbb{R}^n \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1\}$$

Notice here is that even when the space $\mathcal{X}$ is a discrete, the space of distributions $\mathcal{P}(\mathcal{X})$ is continuous already. This provides a nice structure such as ability to differentiate and ease of analysis. This can also apply when underlying space $\mathcal{X}$ is a graph or matroid where it is discrete but has more structure, like some notion or connectivity or distance.

In this lecture, we will generally be interested in the setting where space $\mathcal{X}$ is Euclidean space. However this also works for a manifold, or for a general metric space on which you have a notion of a distance metric.

Why is this space of probability distributions $\mathcal{P}(\mathcal{X})$ interesting? Actually $\mathcal{P}(\mathcal{X})$ can be thought as a generalization of the original space $\mathcal{X}$ such that it provides means to model uncertainty and

randomness in model and data for some problems. It's a generalization because any analysis in $\mathcal{X}$ can be done on a $\mathcal{P}(\mathcal{X})$ where is a one-to-one embedding from $\mathcal{X}$ to $\mathcal{P}(\mathcal{X})$ just by sending every point $x_i \in \mathcal{X}$ to just a point mass $\delta_i \in \mathcal{P}(\mathcal{X})$.

Additionally, if there is some dynamics on underlying space $\mathcal{X}$, we can construct an equivalent dynamics on $\mathcal{P}(\mathcal{X})$. Here we list some simple examples:

Example 1. *Any map $T : \mathcal{X} \to \mathcal{X}$ induces a pushforward map $T_\# : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ such that if random variable X has distribution ρ, then the new random variable Y by applying map T to X will have distribution ν that is application of pushforward map $T_\#$ to ρ. Concretely, if $X \sim \rho$, then $Y = T(X) \sim T_\#(\rho) = \nu$ where $\nu(A) = \rho(T^{-1}(A))$.*

Example 2. *Any algorithm $x_{k+1} = T(x_k)$ on $\mathcal{X}$ induces algorithm $\rho_{k+1} = T_\#(\rho_k)$ on $\mathcal{P}(\mathcal{X})$. Concretely, if $x_k \sim \rho_k$ and $x_{k+1} = T(x_k)$, then $x_{k+1} \sim \rho_{k+1} = T_\#\rho_k$*

Example 3. *Any stochastic dynamics on $\mathcal{X}$ with map $T : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ induces a deterministic dynamics on $\mathcal{P}(\mathcal{X})$.*

2.1 Geometry on space of distributions

Now, let's talk about why it's interesting to work on $\mathcal{P}(\mathcal{X})$ rather than $\mathcal{X}$ by remembering some distances of divergences we have seen so far such as:

$$\text{total variation distance } TV(\rho, \nu) = \frac{1}{2}\int_{\mathcal{X}} |\rho(x) - \nu(x)|dx$$

$$\text{chi-square divergence } \chi^2_\nu(\rho) = \int_{\mathcal{X}} \nu(x)\left(\frac{\rho(x)}{\nu(x)} - 1\right)^2 dx$$

$$\text{KL divergence } H_\nu(\rho) = \int_X \rho(x) log \frac{\rho(x)}{\nu(x)} dx$$

A common theme in these formulations is that they operate on any set $\mathcal{X}$ without considering any structure that may exist in the space. What we want is to build a better notion of distance such that if there is some geometry on $\mathcal{X}$, then it will get reflected on the geometry of $\mathcal{P}(\mathcal{X})$. While we look into the case of Euclidean space $\mathcal{X} = \mathbb{R}^n$ in this lecture, we will see later that a lot of the finding will also apply for manifolds.

Now that we have built the motivation, we can start talking about a specific choice of metric, that is the Wasserstein metric. In the following, we will discuss that the Wasserstein metric has a strong connection to optimal transport, of which we will provide a general definition below. Additionally, we will denote that there is a natural choice of metric, W_2, which is induced by the quadratic distance cost function. Moreover, we will show that W_2 has really nice properties and it has connections to convex duality. Lastly, we will discuss in the next lecture that W_2 metric has a really nice property such that if we have some smooth structure on $\mathcal{X}$, that becomes a smooth structure on $\mathcal{P}(\mathcal{X})$, so it formally becomes a manifold.

But before that, let's introduce some fundamentals to help understand the Wasserstein metric, starting with the notion of optimal transport.

3 Optimal transport

First let's make some definitions.

Definition 1 (Cost function)**.** *Let $\mathcal{X}$ be a space. Cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as*

$$c(x, y) = \textit{cost of transporting } x \textit{ to } y.$$

While cost function can be anything that follows $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, popular choices are distances or functions of distances. For $\mathcal{X} = \mathbb{R}^n$, we list some examples:

- $c(x, y) = \|x - y\|_2$
- $c(x, y) = \|x - y\|_2^2$
- $c(x, y) = \|x - y\|_2^p$
- $c(x, y) = \phi(\|x - y\|_2)$ where $\phi : [0, \infty) \to \mathbb{R}$ is convex, increasing
- $c(x, y) = \begin{cases} 1 \text{ if } x \neq y \\ 0 \text{ else} \end{cases}$

Definition 2 (Transport cost)**.** *Let $\mathcal{X}$ be a space, and $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a cost function. The transport cost between probability distributions $\rho, \nu \in \mathcal{P}(\mathcal{X})$ is*

$$\begin{aligned} \mathcal{T}_c(\rho, \nu) &= \inf_{\pi \in \Pi(\rho, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y) \\ &= \inf_{\pi \in \Pi(\rho, \nu)} \mathbb{E}_\pi[c(X, Y)] \end{aligned}$$

where infimum is over couplings π: joint distribution $(X, Y) \in \mathcal{X} \times \mathcal{X} \sim \pi$ with correct marginals $X \sim \rho$ and $Y \sim \nu$.

Above statement basically means that if you have a mass ρ and you want to transfer the mass into form ν, you would want to do this in such a way that you minimize the expected cost of transferring any random point X in ρ to another random point Y in ν. Intuitively, you are minimizing the expected maximum travel you have to make to move a point X in ρ to other point Y in ν. The coupling that gives you this minimum expected cost is called optimal transport cost.

Looking from another perspective, above formulation is actually an optimization problem over conditional distributions

$$P = (P_x \in \mathcal{P}(\mathcal{X}) : x \in \mathcal{X}) \text{ such that} \tag{1}$$

$$\int_{\mathcal{X}} P_x(y) d\rho(x) = \nu(y) \tag{2}$$

To see that, we can factorize the joint distribution π as:

$$\pi(x, y) = \rho(x)P(y|x)$$

Here, the free variable to choose is $P(y|x)$, and the optimization problem finds the best $P(y|x)$ values for each $x \in \mathcal{X}$ such that the constraint (2) holds, meaning the markov chain defined by P_x has to pushforward the distribution ρ to ν.

Example 4 (Earth mover's distance)**.** *Let $\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \|x - y\|_2$ (Euclidean distance). The transport cost is called Earth mover's distance (EMD)(also Wasserstein-1 distance), which is mathematically defined as:*

$$\mathcal{T}_c(\rho, \nu) = \inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}_\pi[\|X - Y\|_2] = W_1(\rho, \nu)$$

The original formulation of EMD dates back to 1781, where Monge formulated the transportation and allocation of resources as the minimization of total travel of the mass. After a long period of time in 1938, Kantorovich has derived the dual formulation of this problem, leading to the establishment of Linear Programming paradigm, and a Nobel prize in 1975. In the 90's, EMD distances started to be used in Computer Vision domain to measure the similarity between two images, or two histograms in general, leading to efficient approximation algorithms. More recently, the application of optimal transport and EMD distances to machine learning problems has increased a lot, with the advent of more computationally efficient algorithms and hardware to make it possible to calculate distances on large amount of data.

Example 5 (Total Variation distance)**.** *Let $\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \begin{cases} 1 \text{ if } x \neq y \\ 0 \text{ else} \end{cases}$ (Hamming distance). Then the transport cost is Total Variation distance which is defined as:*

$$\mathcal{T}_c(\rho, \nu) = \inf_{(X,Y)\sim\pi} \mathbb{P}(X \neq Y) = TV(\rho, \nu)$$

4 Wasserstein distance

Provided some classical examples of transport cost, let's now define the Wasserstein distance.

Definition 3 (Wasserstein distance)**.** *Let $\mathcal{X} = \mathbb{R}^n,\ c(x,y) = \|x-y\|_2^p$ for some $p \geq 1$.*
Let subset $\mathcal{P}_p(\mathbb{R}^n) = \{\rho \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_p[\|X\|_2^p] < \infty\}$, meaning $\mathcal{P}_p(\mathbb{R}^n)$ is the space of all distributions ρ that has finite width moment. Wasserstein distance is defined as

$$\begin{aligned} W_p(\rho,\nu) &= (\mathcal{T}_p(\rho,\nu))^{1/p} \\ &= (\inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}_\pi[\|X-Y\|_2^p])^{1/p} \end{aligned}$$

where $T_p(\rho,\nu)$ is transport cost.

A nice property of Wasserstein distance is that it satisfies Hölder's inequality such that if $p \leq q$, then $W_p \leq W_q$. Now let's state the more important theorem:

Theorem 1. *W_p defines a distance metric on subspace $\mathcal{P}_p(\mathbb{R}^n)$. This means that W_p is always non-negative, it's zero iff ρ and ν are the same, it's symmetric, and it holds triangular inequality.*

Among the family of Wasserstein distances, there is a specific distance, Wasserstein-2 (W_2) distance, with nice properties.

Definition 4 (Wasserstein-2 distance)**.** *Let $\mathcal{X} = \mathbb{R}^n,\ c(x,y) = \|x-y\|_2^2$. Wasserstein-2 distance W_2 is defined as:*

$$W_2(\rho,\nu) = \sqrt{\mathcal{T}_2(\rho,\nu)} = \inf_{\pi \in \Pi(\rho,\nu)} \sqrt{\mathbb{E}_\pi[\|X-Y\|^2]}$$

One nice property of W_2 is that it makes the embedding $\mathcal{X} \hookrightarrow \mathcal{P}(\mathcal{X})$ is an isometry. This basically means that the W_2 distance between point masses δ_x and δ_y is equal to Euclidean distance between x and y:

$$W_2(\delta_x, \delta_y) = \|x-y\|_2$$

This is easy to see as when there is only one point mass in each distribution to compare, there is only one coupling that couples δ_x with δ_y, resulting in the Euclidean distance between x and y. This property holds true for W_p distances and also when underlying space $\mathcal{X}$ is a manifold. In contrast, Total Variation distance for instance returns 1 for any two point masses that are not equal, losing the closeness information in the original space. On the other hand, chi-square or KL divergences return infinity as they require densities, and can't operate on point masses.

Another nice property of W_2 distance is that if you compare a density to a point mass δ_{x_0}, distance becomes the expected square distance to x_0:

$$W_2(\rho, \delta_{x_0})^2 = \mathbb{E}_\rho[\|X-x_0\|^2]$$

In particular, if we choose x_0 as the mean of the density, W_2 distance becomes the variance of the random variable $X \sim \rho$:

$$\mu = \mathbb{E}_\rho[X] \Rightarrow W_2(\rho, \delta_\mu)^2 = \mathbb{E}_\rho[\|X-\mu\|^2] = Var_\rho(X)$$

Hence we see that W_2 distance can also encode some statistical properties of the original space $\mathcal{X}$.

5 Optimal transport map

5.1 Optimal coupling

We previously denoted that the W_2 distance minimizes the transport cost over all possible couplings. While this is nice, it's still unclear how hard this minimization problem can be. For instance, one may wonder if there is always a solution for any possible pair ρ, ν and if it is unique. For instance, Figure 1 presents some possible couplings for different scenarios such as comparing discrete-discrete, discrete-continuous, or continuous-continuous distributions, that might be the optimal transport map or not.

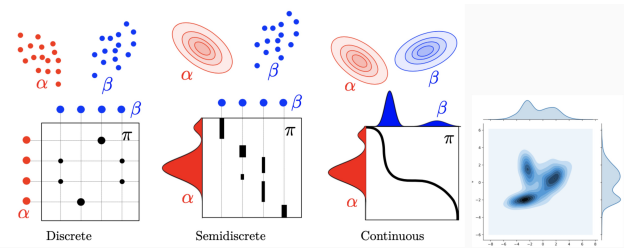

Figure 1: 1-D Wasserstein plots.

Luckily, we will now state a theorem that guarantees existence and uniqueness of this infimum problem in W_2 distance calculation.

Theorem 2 (Brenier's Theorem). *Assume $\rho \ll dx$ on $\mathbb{R}^n$, meaning ρ has a density w.r.t. the Lebesgue measure dx. Then there is a unique optimal coupling π, it's deterministic, and induced by the gradient of a convex function $\phi : \mathbb{R}^n \to \mathbb{R}$. Formally:*

$$d\pi(x, y) = d\rho(x)\delta_{\nabla\phi(x)(y)}$$

Proof for the theorem can be found in [2]. At a very high level, proof follows from the following Kantorovich dual formulation of the problem:

$$W_2(\rho, \nu)^2 = \sup\left\{\mathbb{E}_\rho[\varphi(X)] + \mathbb{E}_\nu[\psi(Y)] : \varphi(x) + \psi(y) \leq \|x - y\|^2\right\}$$

5.2 Optimal transport map

Given Brenier's theorem on optimal coupling, we can now define optimal transport map.

Definition 5 (Optimal transport map). *Let $\rho \ll dx$ with absolutely continuous density. For any ν, there is a unique convex $\phi : \mathbb{R}^n \to \mathbb{R}$ such that:*

1. *The gradient $\nabla\phi : \mathbb{R}^n \to \mathbb{R}^n$ pushes ρ forward to ν:*

$$X \sim \rho \Rightarrow Y = \nabla\phi(X) \sim \nu$$

*2. $\nabla\phi$ is the **optimal transport map** (or Brenier's map) from ρ to ν:*

$$W_2(\rho,\nu)^2 = \mathbb{E}_\rho[\|X - \nabla\phi(X)\|^2]$$

With this definition, we can write W_2 distance directly by getting rid of the infimum objective. However, finding this $\nabla\phi$ function is still non-trivial.

Corollary 1. *If $\nu \ll dx$, the optimal transport map from ν to ρ is*

$$\nabla\phi^* = (\nabla\phi)^{-1}$$

where $\phi^* : \mathbb{R}^n \to \mathbb{R}$ is the convex dual:

$$\phi^*(y) = \sup_{x\in\mathbb{R}^n} \langle x, y\rangle - \phi(x)$$

Here we list some examples where $\nabla\phi$ can be represented in explicit form, therefore Wasserstein distance can be found without solving an optimization problem.

Example 6 (Point mass)**.** *Let $\rho \ll dx$ and $\nu = \delta_a$. Optimal transport map is*

$$\nabla\phi(x) = a$$

where $\phi(x) = x^T a$. Then Wasserstein distance becomes

$$W_2(\rho,\nu)^2 = \mathbb{E}_\rho[\|X - a\|_2^2]$$

In this example, a density is mapped to a point mass at point a, therefore the optimal transport map need to assign any point on ρ to point a in ν. This results in a constant valued gradient function.

Example 7 (Standard normal)**.** *Let $\rho = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(\mu, \Sigma)$. Recall that if we have random variable $X \sim \mathcal{N}(0, I)$, then we can draw any random variable $Y \sim \mathcal{N}(\mu, \Sigma)$ by applying the following transform to X:*

$$Y = \mu + \Sigma^{\frac{1}{2}} X$$

So we can write optimal transport map as

$$\nabla\phi(x) = \mu + \Sigma^{\frac{1}{2}} x$$

where $\phi(x) = x^T\mu + \frac{1}{2}x^T\Sigma^{\frac{1}{2}}x$. Consequently Wasserstein distance becomes

$$\begin{aligned} W_2(\rho,\nu)^2 &= \mathbb{E}[\|\mu + (\Sigma^{\frac{1}{2}} - I)X\|_2^2] = \|\mu\|_2^2 + \|\Sigma^{\frac{1}{2}} - I\|_{\text{HS}}^2 \\ &= \mathbb{E}[\|X - \nabla\phi(X)\|^2] \end{aligned}$$

Example 8 (Gaussian)**.** *Let $\rho = \mathcal{N}(\mu_0, \Sigma_0)$ and $\nu = \mathcal{N}(\mu_1, \Sigma_1)$. Then optimal transport map is written as*

$$\nabla\phi(x) = \Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}(x-\mu_0)+\mu_1.$$

Above complex operation basically first subtracts the mean of the source distribution, then cancels out the source covariance and integrates the new covariance back, and then adds the target mean again.

Applying above gradient map, Wasserstein distance can be written as

$$W_2(\rho,\nu)^2 = \|\mu_0=\mu_1\|_2^2 + Tr(\Sigma_0+\Sigma_1-2(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}).$$

As a special case, if $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$ (covariances commute), optimal transport map and Wasserstein distance simplifies as

$$\nabla\phi(x) = \Sigma_0^{-\frac{1}{2}}\Sigma_1^{\frac{1}{2}}(x-\mu_0)+\mu_1$$
$$W_2(\rho,\nu)^2 = \|\mu_0-\mu_1\|_2^2 + \|\Sigma_0^{\frac{1}{2}}-\Sigma_1^{\frac{1}{2}}\|_{\text{HS}}^2$$

Example 9 (1-dimension)**.** *Let ρ, ν be probability distributions on $\mathbb{R}$ with cdf $F, G : \mathbb{R} \to [0,1]$ such that*

$$F(x) = \mathbb{P}_\rho(X \le x)$$
$$G(y) = \mathbb{P}_\nu(Y \le y)$$

Then optimal transport map from ρ to ν is defined as

$$T = G^{-1}\circ F$$

where $\circ$ is composition operator.

This optimal transport map actually corresponds to Inverse Transform approach to sample from another distribution using uniform distribution. Why is this map the optimal transport map? It's because it's increasing, and in 1-D, any increasing map is a derivative of a convex function.

Plugging in the optimal transport map, Wasserstein distance can be written as

$$W_2(\rho,\nu)^2 = \mathcal{T}_2(\rho,\nu) = \int_0^1 |F^{-1}(t)-G^{-1}(t)|^2 dt$$

For comparison, we can also list how transport cost changes according to the chosen cost function:

$$c(x,y) = |x-y| \Rightarrow \mathcal{T}_1(\rho,\nu) = \int_0^1 |F^{-1}(t)-G^{-1}(t)|^2 dt = \int_{\mathbb{R}} |F(x)-G(x)|dx$$
$$c(x,y) = c(|x-y|) \Rightarrow \mathcal{T}_c(\rho,\nu) = \int_0^1 c(|F^{-1}(t)-G^{-1}(t)|)dt \text{ (c is convex, non-negative)}$$

6 Bounds

As we discussed until now, Wasserstein distance has a very nice structure and properties, but the calculation of the distance is still very hard unless you know the explicit formulation of the optimal transport map, which only rarely happens. Otherwise, we have to solve the infimum problem. That's why finding bounds for the Wasserstein distance is important. Let's briefly state some trivial upper and lower bounds.

Corollary 2 (Upper bound)**.** *Let ρ and ν be probability distributions on $\mathbb{R}^n$ with mean $\mu_1, \mu_2 \in \mathbb{R}^n$ and covariance $\Sigma_1, \Sigma_2 \succ 0$.*

- *Any coupling $(X, Y) \sim \pi$ gives upper bound $W_2(\rho, \nu)^2 \leq \mathbb{E}_\pi[\|X - Y\|^2]$*
- *Trivial coupling: $X \sim \rho, Y \sim \nu$ independent gives upper bound:*

$$\begin{aligned} W_2(\rho, \nu)^2 &\leq \|\mathbb{E}_\rho[X] - \mathbb{E}_\nu[Y]\|_2^2 + \mathrm{Var}_\rho(X) + \mathrm{Var}_\nu(Y) \\ &= \|\mu_1 - \mu_2\|_2^2 + Tr(\Sigma_1 + \Sigma_2) \end{aligned}$$

Corollary 3 (Lower bound)**.** *Let ρ and ν be probability distributions on $\mathbb{R}^n$ with mean $\mu_1, \mu_2 \in \mathbb{R}^n$ and covariance $\Sigma_1, \Sigma_2 \succ 0$.*

- *Lower bound on Wasserstein distance:*

$$W_2(\rho, \nu)^2 \geq \|\mu_1 - \mu_2\|_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}})^{\frac{1}{2}})$$

- *If $\Sigma_1 \Sigma_2 = \Sigma_2 \Sigma_1$, then*

$$W_2(\rho, \nu)^2 \geq = \|\mu_1 - \mu_2\|_2^2 + \|\Sigma_1^{\frac{1}{2}} - \Sigma_2^{\frac{1}{2}}\|_{\mathrm{HS}}^2$$

7 Geodesic

Definition 6 (Displacement interpolation)**.** *Let $\rho, \nu \ll dx$ be probability distributions on $\mathbb{R}^n$. Let $T = \nabla\phi : \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν. For $0 \leq t \leq 1$, define $T_t : \mathbb{R}^n \to \mathbb{R}^n$ by*

$$T_t(x) = (1 - t)x + tT(x)$$

which is the linear interpolation between $T_0(x) = x$ and $T_1(x) = T(x)$ in $\mathbb{R}^n$. If we lift up this interpolation structure to the space of distributions $\mathcal{P}(\mathbb{R}^n)$, we get the displacement interpolation $\rho_t \in \mathcal{P}(\mathbb{R}^n)$ that is defined as

$$\rho_t = (T_t)_\# \rho.$$

This means that displacement interpolation ρ_t at time t can be found by applying pushforward of the transport map T_t on ρ. More concretely, pushforward is applied as follows:

$$X \sim \rho \Rightarrow X_t = (1-t)X + tT(X) \sim \rho_t$$

This interpolation interpolates between $\rho_0 = \rho$ and $\rho_1 = \nu$ in $\mathcal{P}(\mathbb{R}^n)$.

Theorem 3 (W_2 geodesic)**.** *Let $T = \nabla\phi : \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν. Displacement interpolation $\rho_t = (T_t)_{\#}\rho$ that is induced by linear interpolation*

$$T_t(x) = (1-t)x + tT(x)$$

is the geodesic for Wasserstein-2 W_2 distance.

Proof. We know from Brenier's theorem that $T(x) = \nabla\phi(x)$ for some convex ϕ. We can write $T_t(x)$ as

$$\begin{aligned} T_t(x) &= (1-t)x - tT(x) \\ &= (1-t)x - t\nabla\phi(x) \\ &= \nabla\left(\frac{(1-t)}{2}\|x\|^2 + t\phi(x)\right) \\ &= \nabla\phi_t(x) \end{aligned}$$

It's easy to see that $\phi_t(x)$ is a convex function as it's a convex combination between two convex functions $\phi(x)$ and $\|x\|^2$. Now that we have $T_t(x) = \nabla\phi_t(x)$ and $\phi_t(x)$ is convex, using Brenier's theorem we know that T_t is the optimal transport map from ρ to ρ_t.

Now let's compute W_2 distance between ρ and ρ_t:

$$\begin{aligned} W_2(\rho,\rho_t)^2 &= \mathbb{E}[\|X - T_t(X)\|^2] \\ &= \mathbb{E}[\|X - ((1-t)X + tT(X))\|^2] \\ &= \mathbb{E}[\|t(X - T(X))\|^2] \\ &= t^2\mathbb{E}[\|X - T(X)\|^2] \\ &= t^2\mathbb{E}[\|X - \nabla\phi(X)\|^2] \\ &= t^2 W_2(\rho,\nu)^2 \\ W_2(\rho,\rho_t) &= tW_2(\rho,\mu). \end{aligned}$$

Above equation is the defining property of the geodesic, indicating that the interpolation happens with constant velocity on the curve as value t. With that, we conclude the proof. □

The above theorem provides a very nice and systematic way to generate geodesics for Wasserstein distance. Once you know the optimal trnasport map, you can generate the geodesic on Wasserstein distance by forming linear interpolation in space X, and lift it up to get displacement interpolation in the space of distributions $\mathcal{P}(\mathbb{R}^n)$. This also holds for manifold space $\mathcal{X}$.

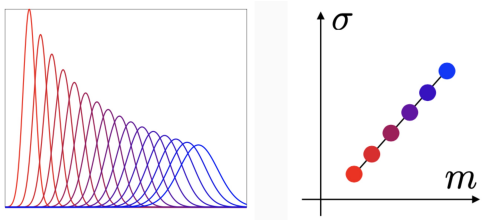

Figure 2: Geodesic behavior for Gaussian distributions. Geodesic transition happens by simply shifting the mean during movement. In addition, modes of the distributions μ and σ interpolate linearly.

Example 10 (Gaussian)**.** *Let $\rho = \mathcal{N}(\mu_0, \Sigma_0)$ and $\nu = \mathcal{N}(\mu_1, \Sigma_1)$. We saw that optimal transport map is given by*

$$\nabla\phi(x) = A(x - \mu_0) + \mu_1$$

where $A = \Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})\frac{1}{2}\Sigma_0^{-\frac{1}{2}}$. Then the displacement interpolation becomes $\rho_t = \mathcal{N}(\mu_t, \Sigma_t)$ where

$$\begin{aligned}\mu_t &= (1-t)\mu_0 + t\mu_1 \\ \Sigma_t &= ((1-t)I + tA)\Sigma_0((1-t)I + tA)\end{aligned}$$

For the special case of $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0 \Rightarrow \Sigma_t^{\frac{1}{2}} = (1-t)\Sigma_0^{\frac{1}{2}} + t\Sigma_1^{\frac{1}{2}}$ This property tells us that displacement interpolation between Gaussians stays Gaussian (Figure 2). In comparison, we know that a linear interpolation gives mixture of Gaussians in between:

$$\begin{aligned}\rho_t &= (1-t)\rho + t\nu \\ &= (1-t)\mathcal{N}(\mu_1, \Sigma_1) + t\mathcal{N}(\mu_2, \Sigma_2)\end{aligned}$$

Example 11 (1-dimension)**.** *Let ρ, ν be probability distributions on $\mathbb{R}$ with cdf $F_0, F_1 : \mathbb{R} \to [0, 1]$. Displacement interpolation ρ_t has cdf F_t (Figure 3) where*

$$F_t^{-1} = (1-t)F_0^{-1} + tF_1^{-1}$$

Definition 7 (Wasserstein barycenter)**.** *The average of distributions $(\rho_i)_{i=1}^m$ with weights $\Sigma_{i=1}^m\lambda_i = 1$ that is the solution of the following problem (Figure 4:*

$$\rho^* = \arg\min_{\rho=\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^m \lambda_i W_2(\rho, \rho_i)^2$$

In conclusion, in this lecture we define optimal transport map, Wasserstein distance, and geodesic on Wasserstein distance. Next time, we will start looking into manifold structure of the Wasserstein distance, and define notions of gradient and gradient flow on the Wasserstein manifold domain, which enables doing optimization.

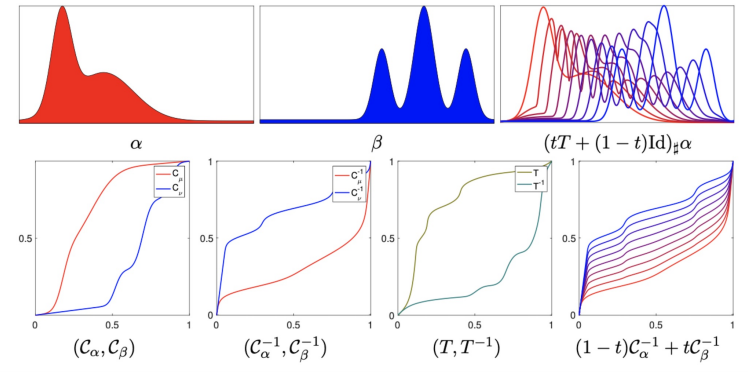

Figure 3: Geodesic behavior for 1-D distributions. We see similar smooth shifting behaviors compared to Gaussian geodesic.

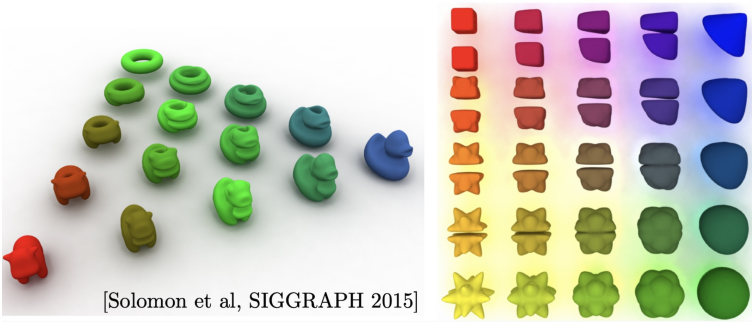

Figure 4: Wasserstein barycenter interpolation between images and shapes.

References

[1] G. Peyré and Marco Cuturi. "Computational Optimal Transport". In: *Found. Trends Mach. Learn.* 11 (2019), pp. 355–607.

[2] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media, 2008.

[3] Cédric Villani. *Topics in optimal transportation*. 58. American Mathematical Soc., 2003.