Optimal transport

III Wasserstein Space

Seongho, Joo

Seoul National University

Reminders

- Let X, Y be compact metric spaces, $c \in C(X \times Y)$ the cost function $(\mu, \nu) \in \mathcal{P}(X) \times \mathcal{P}(Y)$ the marginals. We call the following results:
  - minimizer/maximizers exist for both problems and, for the dual, can be chosen as (φ, φ^c) with φ c-concave.
  - at optimality, it holds $\varphi(x) + \psi(y) = c(x, y)$ for γ-almost every (x, y).
  - we have the following special cases:
    - for $X = Y \subset \mathsf{r}$ and $c(x, y) = h(y - x)$ with h strictly convex, the (unique) optimal transport plan, which can be characterized with the quantile functions of μ and ν.
    - for $X = Y$ and $c(x, y) = \mathsf{dist}(x, y)$, we have the Kantorovich-Rubinstein formula

$$T_c(\mu, \nu) = \sup_{\varphi \in 1\text{-Lip}} \int \varphi \, \mathrm{d}\mu - \nu$$

    - for $X = Y \subset \mathsf{r}^d$ and $c(x, y) = \frac{1}{2}|y - x|^2$, and when μ is absolutely continuous, there exists a unique optimal transport plan. It is of the from $\gamma = (\mathsf{id}, \nabla\tilde{\varphi})_{\#}\mu$ for some $\tilde{\varphi} \in C(\mathsf{r}^d)$ convex.

Wasserstein space

Definition 1 (Wassetstein space)

Let (X, dist) be a compact metric space. For $p \geq 1$, we denote by $\mathcal{P}_p(X)$ the set of probability measures on X endowed with the p-Wasserstein distance, defined as

$$W_p(\mu,\nu) := \left(\min_{\gamma \in \Pi(\mu,\nu)} \int \mathsf{dist}(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p} = \mathcal{T}_{\mathsf{dist}^p}(\mu,\nu)^{\frac{1}{p}}$$

- This distance is a natural way to build a distance on $cP(X)$ from a distance on X.

In particular, the map $\delta : X \to \mathcal{P}_p(X)$ mapping a point $x \in X$ to the Dirac mass δ_x is an isometry.

Wasserstein space

Definition 1 (Wassetstein space)

Let (X, dist) be a compact metric space. For $p \geq 1$, we denote by $\mathcal{P}_p(X)$ the set of probability measures on X endowed with the p-Wasserstein distance, defined as

$$W_p(\mu, \nu) := \left(\min_{\gamma \in \Pi(\mu,\nu)} \int \mathsf{dist}(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p} = \mathcal{T}_{\mathsf{dist}^p}(\mu, \nu)^{\frac{1}{p}}$$

• This distance is a natural way to build a distance on $cP(X)$ from a distance on X. In particular, the map $\delta : X \to \mathcal{P}_p(X)$ mapping a point $x \in X$ to the Dirac mass δ_x is an isometry.

Proposition 1

W_p defines the axioms of a distance on $\mathcal{P}_p(X)$.

The symmetry of the Wasserstein distance is obvious. Moreover, $W_p(\mu, \nu) = 0$ implies that there exists a $\gamma \in \Pi(\mu, \nu)$ such that $\int \mathsf{dist}^p \, \mathrm{d}\gamma = 0$. This implies that γ is concentrated on the diagonal, so that $\gamma = (\mathsf{id}, \mathsf{id})_{\#}\mu$ is induced by the identity map.

Proposition 1 proof

To prove the triangle inequality we will use the gluing lemma below with $N = 3$.

Lemma 1 (Gluing)

Let $X_1, \ldots X_N$ be complete and separable metric spaces, and for any $1 \leq i \leq N-1$ consider a transport plan $\gamma_i \in \Pi(\mu_i, \mu_{i+1})$. Then, there exists $\gamma \in \mathcal{P}(X_1, \ldots, X_N)$ such that for all $i \in \{1, \ldots N-1\}$, $(\pi_{i,i+1})_{\#}\gamma = \gamma_i$, where $\pi_{i,i+1} : X_1 \times \cdots \times X_N \to X_i \times X_{i+1}$ is the projection.

Proposition 1 proof

To prove the triangle inequality we will use the gluing lemma below with $N = 3$.

Lemma 1 (Gluing)

Let $X_1, \dots X_N$ be complete and separable metric spaces, and for any $1 \leq i \leq N-1$ consider a transport plan $\gamma_i \in \Pi(\mu_i, \mu_{i+1})$. Then, there exists $\gamma \in \mathcal{P}(X_1, \dots, X_N)$ such that for all $i \in \{1, \dots N-1\}$, $(\pi_{i,i+1})_{\#}\gamma = \gamma_i$, where $\pi_{i,i+1} : X_1 \times \cdots \times X_N \to X_i \times X_{i+1}$ is the projection.

Let $\mu_i \in \mathcal{P}_p(X)$ for $i \in \{1, 2, 3\}$ and let $\gamma_1 \in \Pi(\mu_1, \mu_2)$ and $\gamma_2 \in \Pi(\mu_2, \mu_3)$ be optimal in the definition of W_p. Then, there exists $\sigma \in \mathcal{P}(X^3)$ such that $(\pi_{i,i+1})_{\#}\sigma = \gamma_i$ for $i \in \{1, 2\}$. A fortiori one has $(\pi_1)_{\#}\sigma = \mu_1$ and $(\pi_3)_{\#}\sigma = \mu_3$, so that $(\pi_{1,3})_{\#}\sigma \in \Pi(\mu_1, \mu_3)$. In particular,

$$
\begin{aligned}
W_p(\mu_1, \mu_3) &\leq \left(\int_{X^2} \mathsf{dist}(x, y)^p \, \mathrm{d}(\pi_{1,3})_{\#}\sigma(x, y) \right)^{1/p} \\
&= \left(\int_{X^3} \mathsf{dist}(x_1, x_3)^p \, \mathrm{d}\sigma(x_1, x_2, x_3) \right)^{1/p} \\
&\leq \left(\int_{X^3} (\mathsf{dist}(x_1, x_2) + \mathsf{dist}(x_2, x_3))^p \, \mathrm{d}\sigma(x_1, x_2, x_3) \right)^{1/p} \\
&\leq \left(\int_{X^3} \mathsf{dist}(x_1, x_2)^p \, \mathrm{d}\sigma(x_1, x_2, x_3) \right)^{1/p} + \left(\int_{X^3} \mathsf{dist}(x_2, x_3)^p \, \mathrm{d}\sigma(x_1, x_2, x_3) \right)^{1/
\end{aligned}
$$

Comparision between Wasserstein distances

Note that, due to Jensen's inequality, since all $\gamma \in \Pi(\mu, \nu)$ are probability measures, for $p \leq q$ we have $(\int \mathsf{dist}(x,y)^p \, \mathrm{d}\gamma)^{q/p} \leq \int \mathsf{dist}(x,y)^q \, \mathrm{d}\gamma$ and so

$$\left(\int \mathsf{dist}(x,y)^p \, \mathrm{d}\gamma \right)^{\frac{1}{p}} \leq \left(\int \mathsf{dist}(x,y)^p \, \mathrm{d}\gamma \right)^{\frac{1}{q}},$$

which implies $W_p(\mu,\nu) \leq W_q(\mu,\nu)$. In particular, $W_1(\mu,\nu) \leq W_p(\mu,\nu)$ for every $p \geq 1$. In particular, $W_1(\mu,\nu) \leq W_p(\mu,\nu)$. On the other hand, for compact (and thus bounded) X, an opposite inequality also holds, since

$$\left(\int \mathsf{dist}(x,y)^p \, \mathrm{d}\gamma \right)^{1/p} \leq \mathsf{diam}(X)^{\frac{p-1}{p}} \left(\int \mathsf{dist}(x,y) \, \mathrm{d}\gamma \right)^{\frac{1}{p}}$$

This implies that for all $p \geq 1$,

$$W_1(\mu,\nu) \leq W_p(\mu,\nu) \leq \mathsf{diam}(X)^{\frac{p-1}{p}} \left(W_1(\mu,\nu) \right)^{\frac{1}{p}}$$

Topologial properties

Theorem 1

Assume that X is compact. For $p \in [1, +\infty]$, we have $\mu_n \rightharpoonup \mu$ if and only if $W_p(\mu, \mu) \to 0$.

Proof.

Topologial properties

Theorem 1

Assume that X is compact. For $p \in [1, +\infty]$, we have $\mu_n \rightharpoonup \mu$ if and only if $W_p(\mu, \mu) \to 0$.

Proof. We only need to prove the result for W_1 thanks to the comparison inequalities between W_1 and W_p. Consider a sequence μ_n such that $W_1(\mu_n, \mu) \to 0$. Thanks to the duality formula, for every $\varphi \in \mathrm{Lip}_1(X)$, we have $\int \varphi(\mu_n - \mu) \to 0$. By linearity, the same is true for any Lipschitz function. By density, this holds for any function in $C(X)$. This shows that convergence in W_1 implies weak convergence.
To prove the opposite implication, consider a subsequence ν_{n_k} that satisfies $\lim_k W_1(\mu_{n_k}, \mu) = \limsup_n W_1(\mu_n, n)$. For every k pick a function $\varphi_{n_k} \in \mathrm{Lip}_1(X)$ such that $\int \varphi_{n_k}(\mu_{n_k} - \mu) = W_1(\mu_{n_k}, \mu)$. We may assume that the φ_{n_k} all vanish at the same point, and they are hence uniformly bounded and equi-continuous. By Ascoli-Arzelà theorem, we can extract a sub-sequence uniformly converging to a certain fucntion $\varphi \in \mathrm{Lip}_1(X)$. By replacing the original subsequence with this new one, we have now

$$W_1(\mu_{n_k}, \mu) = \int \varphi_{n_k} \, \mathrm{d}(\mu_{n_k} - \mu) \to \int \varphi \, \mathrm{d}(\mu - \mu) = 0$$

where the convergence of the integral is justified by the weak convergence $\mu_{n_k} \rightharpoonup \mu$ together with the strong convergence in $C(X)$ $\varphi_{n_k} \to \varphi$. This shows that $\limsup_n W_1(\mu_n, \mu) \leq 0$ and concludes the proof.

Geodesics in Wasserstein space

Definition 2

Let (X, dist) be a metric space. A constant speed geodesic between two points $x_0, x_1 \in X$ is a continuous curve $x : [0,1] \to X$ such that for every $s, t \in [0,1]$, $\text{dist}(x_s, x_t) = |s - t| \text{dist}(x_0, x_1)$

Proposition 2 (Geodesic between measures)

Let $\mu_0, \mu_1 \in \mathcal{P}_p(X)$ with $X \subset \mathrm{r}^d$ compact and convex. Let $\gamma \in \Pi(\mu_0, \mu_1)$ be an optimal transport plan. Define

$$\mu_t := (\pi_t)_\# \gamma \text{ where } \pi_t(x, y) = (1-t)x + ty$$

Then, the curve μ_t is a constant speed geodesic between μ_0 and μ_1.

Example 3.3 If there exists an optimal transport map T between μ_0 and μ_1, then the geodesic defined above is $\mu_t = ((1-t)\text{id} + tT)_\# \mu_0$.

Geodesics in Wasserstein space

Definition 2

Let (X, dist) be a metric space. A constant speed geodesic between two points $x_0, x_1 \in X$ is a continuous curve $x : [0,1] \to X$ such that for every $s, t \in [0,1]$, $\text{dist}(x_s, x_t) = |s-t|\text{dist}(x_0, x_1)$

Proposition 2 (Geodesic between measures)

Let $\mu_0, \mu_1 \in \mathcal{P}_p(X)$ with $X \subset \mathrm{r}^d$ compact and convex. Let $\gamma \in \Pi(\mu_0, \mu_1)$ be an optimal transport plan. Define

$$\mu_t := (\pi_t)_\# \gamma \text{ where } \pi_t(x,y) = (1-t)x + ty$$

Then, the curve μ_t is a constant speed geodesic between μ_0 and μ_1.

Example 3.3 If there exists an optimal transport map T between μ_0 and μ_1, then the geodesic defined above is $\mu_t = ((1-t)\text{id} + tT)_\# \mu_0$.

Corollary 1

The space $(\mathcal{P}_p(X), W_p)$ with $X \subset \mathrm{r}^d$ compact and convex is a geodesic space, meaning that any $\mu_0, \mu_1 \in \mathcal{P}_p(X)$ can be joined by (at least one) constant speed geodesic.

Geodesics in Wasserstein space

Prop 2 Proof.

Geodesics in Wasserstein space

Barycenters in $\mathcal{P}_2(X)$. The notion of geodesics allow to define the notion of a midpoint, or more generally barycenters, between two probability distributions. How to generalize the notion of "Wasserstein barycenters" to more than two probability distributions?

In r^d, the barycenters of $x_1, \ldots x_n$ with weights $\lambda_1, \ldots, \lambda_n > 0$ is the unique point y that minimizes $\sum_i \lambda_i \|y_i - x_i\|_2^2$. This motivates us to define *Wasserstein-2* barycenters between $\mu_1 \ldots \mu_n \in \mathcal{P}_2(X)$ with weights $\lambda_1, \ldots \lambda_n > 0$ as any measures that solves

$$\min_{\nu \in \mathcal{P}_2(X)} \left\{ \sum_{i=1}^{n} \lambda_i W_2^2(\mu_i, \nu) \right\}$$

Observe that when $\mu_1 = \delta_{x_i}$ we recover the usual notion of barycenters on r^d.

Differentiability of the Wasserstein distance

Theorem 2

Let $\sigma, \rho_0, \rho_1 \in \mathcal{P}(X)$. Assume that there exists unique Kantorovich potentials (φ_0, ψ_0) between σ and ρ_0 which are c-conjugate to each other and satisfy $\psi_0(x_0) = 0$ for some $x_0 \in X$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{T}_c(\sigma + \rho_0 + t(\rho_1 - \rho_0)|_{t=0} = \int \psi \,\mathrm{d}(\rho_1 - \rho_0)$$

Proof.

Differentiability of the Wasserstein distance

• The assumption on the uniqueness of the potentials can be guaranteed a priori in several settings. Let us give one example of sufficient conditions which corresponds to the distance W_2 (one could prove it for W_p, with $p > 1$ similarly).

Proposition 3 (Uniqueness of potentials)

If $X \subseteq \mathrm{r}^d$ is the closure of a bounded and connected open set, $x_0 \in X$, $(\mu, \nu) \in \mathcal{P}(X)$ are such that μ is absolutely continuous and spt$(\mu) = X$ then, there exists a unique pair of Kantorovich potentials (φ, ψ) optimal for $c(x, y) = \frac{1}{2} \|x - y\|^2$, c-conjugate to each other, and satisfying $\varphi(x_0) = 0$.

Proof.

Dynamic formulation of optimal transport

- When $X \subset \mathsf{r}^d$, we can interpret the marginals $\mu, \nu \in \mathcal{P}(X)$ as distributions of particles at times $t = 0$ and $t = 1$ respectively. Assume that for each time t, there is a velocity field $v_t : rr^d \to \mathsf{r}^d$ which moves particles around. The relation between the velocity field and the distribution is given by the continuity equation (satisfied in the sense of distributions)

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0.$$

- When v_t is regular enough (e.g. Lipschitz continuous in x, uniformly in t), then we can defines its flow $T : [0,1] \to X \to \mathsf{r}^d$ which is such that $T_t(x)$ gives the position at time t of a particle which is at x at time 0. It solves $T_0(X) = x$ and

$$\frac{\mathrm{d}}{\mathrm{d}t} T_t(x) = v_t(T_t(x)).$$

- The relation between the evolution of the distribution ρ_t- the *Eulerian* description- and the evolution of the flow T_t - the *Lagrangian* description - is simply $\rho_t = (T_t)_{\#}\mu$.

Dynamic formulation of optimal transport

- Let us denote $\mathsf{CE}(\mu, \nu)$ the sets of solutions (ρ, ν) to the continuity equation such that $t \mapsto \rho_t$ is weakly continuous and satisfies $\rho_0 = \mu$ and $\rho_1 = \nu$. Consider also the integrated (generalized) "kinetic energy" functional

$$A_p(\rho, v) := \int_0^1 \int_X \|v_t(x)\|^p \, \mathrm{d}\mu_t(x) \, \mathrm{d}t.$$

By minimizing the functional over all interpolation between μ and ν, we recover the optimal transport with cost $\|y - x\|^p$. This is called the Benamou-Brenier formunlation.

Theorem 3 (Dynamic formulation)

Let $\mu, \nu \in \mathcal{P}(\mathsf{r}^d)$ be compactly supported. For $p \geq 1$ it holds

$$W_p^p(\mu, \nu) = \int \{A_p(\rho, \nu) \,|\, (\rho, \nu) \in \mathsf{CE}(\mu, \nu)\}$$

Justifications for Theorem 3

- First argue that for $(\rho,\nu) \in \mathsf{CE}(\mu,\nu)$ it holds $A_p(\rho,\nu) \geq W_p^p(\mu,\nu)$. Assume (ρ,ν) is regular enough and consider the flow $T_t(x)$, that satisfies $\rho_t = (T_t)_\# \rho_0$. It holds

$$\begin{aligned} A(\rho,\nu) &= \int_0^1 \int_X \|v_t(T_t(x))\|^p \, \mathrm{d}\rho_0(x) \, \mathrm{d}t \\ &= \int_X \left(\int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t} T_t(x) \right\|^p \mathrm{d}t \right) \mathrm{d}\rho_0(x) \\ &\geq \int_X \|T_1(x) - T_0(X)\|^p \, \mathrm{d}\rho_0(x) \end{aligned}$$

by Jensen's inequality. Since $(T_1)_\# \rho_0 = \rho_1 = \nu$ and $\rho_0 = \mu$, the last quantity is larger than $W_p^p(\mu,\nu)$.

- Let us build an admissible $(\rho,\nu) \in \mathsf{CE}(\mu,\nu)$ such that $A(\rho,v) = W_p^p(\mu,\nu)$ using the geodesic between μ and ν. Assume that **there exists an optimal transport map** T between μ and ν, and set $\rho_t = (T_t)_\# \mu$ with $T_t(x) = (1-t)x + tT(x)$. Now define the velocity field

$$v_t = \left(\frac{\mathrm{d}}{\mathrm{d}t} \right) \circ T_t^{-1} = (T - \mathsf{id}) \circ T_t^{-1},$$

which, by construction, is that (ρ_t, v_t) satisfies the continuity equation in the weak sense. We have the desired equality:

$$A(\rho,v) = \int \|v_t(x)\|^p \, \mathrm{d}\rho_t(x) = \int |T(x) - x|^p \, \mathrm{d}\rho_0(x) = W_p^p(\mu,\nu).$$

Riemannian interpretation

- In the case $p = 2$, we can understand (at least as the formal level) the Benamou-Brenier formula as a Riemannaian formulation for w_2. In this interpretation, the tangent space at $\rho \in \mathcal{P}(X)$ are measures of form $\delta\rho = -\nabla \cdot (v\rho)$ with a velocity field $v \in L^2(\rho, \mathsf{r}^d)$ and the metric is given by

$$\|\delta\rho\|_p^2 = \int_{v \in L^2(\mathsf{r}^d, \rho)} \left\{ \int \|v(x)\|_2^2 \, \mathrm{d}\rho(x) \,|\, \delta\rho = -\nabla \cdot (v\rho) \right\}.$$