Lecture 3: Wasserstein Space

Lénaïc Chizat

February 26, 2020

The material of today's lecture is adapted from Q. Mérigot's lecture notes and [3, 4].

1 Reminders

Let X, Y be compact metric spaces, $c \in \mathcal{C}(X \times Y)$ the cost function and $(\mu, \nu) \in \mathcal{P}(X) \times \mathcal{P}(Y)$ the marginals. In previous lectures, we have seen that the optimal transport problem can be formulated as an optimization over the space of transport plans $\Pi(\mu, \nu)$ — the primal or Kantorovich problem — and as an optimization over potential functions $\{(\varphi, \psi) \in \mathcal{C}(X) \times \mathcal{C}(Y) \mid \varphi \oplus \psi \leqslant c\}$ — the dual problem. We recall 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 $X = Y \subset \mathbb{R}$ and $c(x, y) = h(y - x)$ with h strictly convex, the optimal transport plan is the (unique) monotone plan, which can be characterized with the quantile functions of μ and ν.
  - for $X = Y$ and $c(x, y) = \mathrm{dist}(x, y)$, we have the Kantorovich-Rubinstein formula
  $$\mathcal{T}_c(\mu, \nu) = \sup_{\varphi \text{ 1-Lip}} \int \varphi \mathrm{d}(\mu - \nu).$$
  - for $X = Y \subset \mathbb{R}^d$ 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 form $\gamma = (\mathrm{id}, \nabla\tilde{\varphi})_{\#}\mu$ for some $\tilde{\varphi} \in \mathcal{C}(\mathbb{R}^d)$ convex.

2 Wasserstein space

2.1 Definition and elementary properties

Definition 2.1 (Wasserstein space)**.** Let (X, dist) be a compact metric space. For $p \geqslant 1$, we denote by $\mathcal{P}_p(X)$ the set of probability measures on X endowed with the p-Wasserstein distance, defined as

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

This distance is a natural way to build a distance on $\mathcal{P}(X)$ 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 2.2. *W_p satisfies the axioms of a distance on $\mathcal{P}_p(x)$.*

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

To prove the triangle inequality we will use the gluing lemma below (Lemma 2.3) with $N = 3$. 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 $(\pi_{13})_\#\sigma \in \Pi(\mu_1,\mu_3)$. In particular,

$$\begin{aligned}
W_p(\mu_1,\mu_3) &\leqslant \left(\int_{X^2} \mathrm{dist}(x,y)^p \mathrm{d}(\pi_{1,3})_\#\sigma(x,y)\right)^{1/p} \\
&= \left(\int_{X^3} \mathrm{dist}(x_1,x_3)^p \mathrm{d}\sigma(x_1,x_2,x_3)\right)^{1/p} \\
&\leqslant \left(\int_{X^3} (\mathrm{dist}(x_1,x_2) + \mathrm{dist}(x_2,x_3))^p \mathrm{d}\sigma(x_1,x_2,x_3)\right)^{1/p} \\
&\leqslant \left(\int_{X^3} \mathrm{dist}(x_1,x_2)^p \mathrm{d}\sigma(x_1,x_2,x_3)\right)^{1/p} + \left(\int_{X^3} \mathrm{dist}(x_2,x_3)^p \mathrm{d}\sigma(x_1,x_2,x_3)\right)^{1/p} \\
&= W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3),
\end{aligned}$$

where we used the Minkowski inequality in $L^p(\sigma)$ to get the second inequality, and the property $(\pi_{i,i+1})_\#\sigma = \gamma_i$ to get the last equality. □

Lemma 2.3 (Gluing)**.** *Let $X_1,\dots,X_N$ be complete and separable metric spaces, and for any $1 \leqslant i \leqslant N-1$ 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.*

Proof. See Lemma 5.3.2 and Remark 5.3.3 in [1]. □

Exercise 2.4. Prove the triangle inequality assuming the existence of optimal transport maps between μ_1,μ_2 and μ_2,μ_3.

Remark 2.5 (Non-compact case)**.** As usual, the compactness assumption is only here for clarity of presentation. In general, when X is a complete and separable metric space, the space $\mathcal{P}_p(X)$ is defined as the set of probability measures such that for some (and thus any) $x_0 \in X$ it holds

$$\int \mathrm{dist}(x_0,y)^p \mathrm{d}\mu(y) < \infty.$$

It can be shown that this set endowed with the distance W_p is also a complete and separable metric space. Exercice: show that the Wasserstein distance W_p is finite on this set.

2.2 Comparisons

Comparison between Wasserstein distances Note that, due to Jensen's inequality, since all $\gamma \in \Pi(\mu,\nu)$ are probability measures, for $p \leqslant q$ we have

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

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

$$\Big(\int \mathrm{dist}(x,y)^p \mathrm{d}\gamma\Big)^{\frac{1}{p}} \leqslant \mathrm{diam}(X)^{\frac{p-1}{p}} \Big(\int \mathrm{dist}(x,y) \mathrm{d}\gamma\Big)^{\frac{1}{p}}.$$

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

$$W_1(\mu,\nu) \leqslant W_p(\mu,\nu) \leqslant \mathrm{diam}(X)^{\frac{p-1}{p}} W_1(\mu,\nu)^{\frac{1}{p}}.$$

Comparison with L^p distances Take $X \subset \mathbb{R}$ with its usual distance. Consider the translation $t_x : y \in \mathbb{R} \mapsto x + y$. Then, since the map t_x is increasing, for any $\mu \in \mathcal{P}_p(X)$ one has $W_p(\mu, t_{x\#}\mu) = |x|$. However,

- if $\rho \in \mathcal{P}(X) \cap L^p(X)$ with $\mathrm{spt}(\rho) \subseteq [0,1]$, then for all $|x| \geqslant 1$ one has $\|\rho - t_x\rho\|_{\mathrm{L}^p(X)} = 2\,\|\rho\|_{L^p(X)}$ where $t_x\rho(y) = \rho(y-x)$. Unlike the $L^p(X)$ norm, the Wasserstein distance is geometry "aware".
- If $\rho \in \mathcal{P}(X) \cap L^2(X)$ is such that $\|t_x\rho - \rho\|_{L^2(X)} = \mathrm{O}(|x|)$, then ρ belongs to the Sobolev space H^1 (see Proposition VIII.3 in [2]). In other words, $\|t_x\rho - \rho\|_{\mathrm{L}^2(X)}$ can be much larger than $|x|$ unless ρ is very regular.

Because of these two examples, the Wasserstein distance is a very appealing notion of distance for data analysis (e.g. measuring the distance between signals). The flipside is that the definition of the Wasserstein requires the signal to belong to $\mathcal{P}(X)$ i.e. to be non-negative and with unit mass.

2.3 Topological properties

Theorem 2.6. *Assume that X is compact. For $p \in [1,+\infty[$, we have $\mu_n \rightharpoonup \mu$ if and only if $W_p(\mu_n,\mu) \to 0$.*

Proof. We only need to prove the result for W_1 thanks to the comparison inequalities between W_1 and W_p in previous section. Let us start from 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 $\mathcal{C}(X)$. This shows that convergence in W_1 implies weak convergence.

To prove the opposite implication, let us first fix a subsequence μ_{n_k} that satisfies $\lim_k W_1(\mu_{n_k},\mu) = \limsup_n W_1(\mu_n,\mu)$. For every k, pick a function $\varphi_{n_k} \in \mathrm{Lip}_1(X)$ such that $\int \varphi_{n_k}(\mu_{n_k} - \mu) = W_1(\mu_{n_k},\mu)$. Up to adding a constant, which does not affect the integral, we can 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 $\varphi \in \mathrm{Lip}_1(X)$. 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 $\mathcal{C}(X)$ $\varphi_{n_k} \to \varphi$. This shows that $\limsup_n W_1(\mu_n,\mu) \leqslant 0$ and concludes the proof. $\square$

Remark 2.7. In the non-compact case, it can be shown that convergence in $\mathcal{P}_p(X)$ is equivalent to tight convergence (in duality with continuous and bounded functions) and convergence of the p-th order moments i.e. for all $x_0 \in X$,

$$\int \operatorname{dist}(x_0, y)^p \mathrm{d}\mu_n(y) \to \int \operatorname{dist}(x_0, y)^p \mathrm{d}\mu(y).$$

3 Geodesics in Wasserstein space

Definition 3.1. 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 $s, t \in [0,1]$, $\operatorname{dist}(x_s, x_t) = |s-t| \operatorname{dist}(x_0, x_1)$.

Proposition 3.2. *Let $\mu_0, \mu_1 \in \mathcal{P}_p(X)$ with $X \subset \mathbb{R}^d$ 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 $\mu_t = ((1-t)\mathrm{id} + tT)_\# \mu_0$.

Remark 3.4. In fact, it can be shown that any geodesic between μ_0 and μ_1 can be constructed as in Proposition 3.2.

Proof. First note that if $0 \leqslant s \leqslant t \leqslant 1$,

$$W_p(\mu_0, \mu_1) \leqslant W_p(\mu_0, \mu_s) + W_p(\mu_s, \mu_t) + W_p(\mu_t, \mu_1),$$

so that it suffices to prove the inequality $W_p(\mu_s, \mu_t) \leqslant |t-s| \, W_p(\mu_0, \mu_1)$ for all $0 \leqslant s \leqslant t \leqslant 1$ to get equality. The inequality is easily checked by building an explicit transport plan using an optimal transport plan γ. Take $\gamma_{st} := (\pi_s, \pi_t)_\# \gamma \in \Pi(\mu_s, \mu_t)$, so that

$$\begin{aligned} W_p(\mu_s, \mu_t)^p &\leqslant \int \|x-y\|^p \, \mathrm{d}\gamma_{st}(x,y) = \int \|\pi_s(x,y) - \pi_t(x,y)\|^p \, \mathrm{d}\gamma(x,y) \\ &= \int \|(1-s)x + sy - ((1-t)x + ty)\|^p \, \mathrm{d}\gamma(x,y) \\ &= \int \|(t-s)(x-y)\|^p \, \mathrm{d}\gamma(x,y) = (t-s)^p W_p(\mu, \nu)^p \end{aligned}$$

$\square$

Corollary 3.5. *The space $(\mathcal{P}_p(X), W_p)$ with X 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.*

4 Differentiability of the Wasserstein distance

In this section, we will compute the differential of the Wasserstein distance under additive perturbations.

Theorem 4.1. *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 $\varphi_0(x_0) = 0$ for some $x_0 \in X$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{T}_c(\sigma, \rho_0 + t(\rho_1 - \rho_0)|_{t=0} = \int \psi_0 \mathrm{d}(\rho_1 - \rho_0).$$

Proof. Denote $\rho_t = (1-t)\rho_0 + t\rho_1 = \rho_0 + t(\rho_1 - \rho_0)$. By Kantorovich duality, we have

$$\mathcal{T}_c(\sigma, \rho_t) \geqslant \int \varphi_0 \mathrm{d}\sigma + \int \psi_0 \mathrm{d}\rho_t.$$

This immediately gives

$$\frac{1}{t}(\mathcal{T}_c(\sigma, \rho_t) - \mathcal{T}_c(\sigma, \rho_0)) \geqslant \int \psi_0 \mathrm{d}(\rho_1 - \rho_0).$$

To show the converse inequality, we let (φ_t, ψ_t) be c-conjugate Kantorovich potentials between σ and ρ_t satisfying $\psi_t(x_0) = 0$, giving

$$\frac{1}{t}(\mathcal{T}_c(\sigma, \rho_0) - \mathcal{T}_c(\sigma, \rho_t)) \geqslant \int \psi_t \mathrm{d}(\rho_1 - \rho_0).$$

Moreover, by uniqueness of (φ_0, ψ_0), we get that φ_t, ψ_t converges uniformly to (φ_0, ψ_0) as $t \to 0$, thus concluding the proof. $\square$

The assumption on the uniqueness of the potentials can be guaranteed a priori in the following setting, which corresponds to the distance W_2 (one could prove it for W_p, with $p > 1$ similarly).

Proposition 4.2 (Uniqueness of potentials)**.** *If $X \subseteq \mathbb{R}^d$ is the closure of a bounded and connected open set, $x_0 \in X$, $(\sigma, \rho) \in \mathcal{P}(X)$ satisfies*

$$\mathrm{spt}(\rho) = X \ \textit{or}\ \mathrm{spt}(\sigma) = X,$$

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. Assume that $\mathrm{spt}(\sigma) = X$. Since c is Lipschitz on the bounded set X, φ, ψ are Lipschitz and therefore differentiable almost everywhere. Take $(x_0, y_0) \in \mathrm{spt}(\gamma)$ where $\gamma \in \Pi(\sigma, \rho)$ is the optimal transport plan, such that φ is differentiable at $x_0 \in \mathring{X}$. As we have already shown, for any optimal pair (φ, ψ) we necessarily have

$$y_0 = x_0 - \nabla\varphi(x_0),$$

so that if (φ', ψ') is another optimal pair, we should have $\nabla\varphi = \nabla\varphi'$ σ-a.e. Since $\mathrm{spt}(\sigma) = X$ and since X is the closure of a connected open set, this implies $\varphi = \varphi' + C$ for a constant C as desired, and $C = 0$ since $\varphi(x_0) = \varphi'(x_0)$. Moreover, $\psi' = \varphi'^c = \varphi^c = \psi$, allowing to deal with the case where $\mathrm{spt}(\rho) = X$ by symmetry. $\square$

5 Dynamic formulation of optimal transport

We conclude this lecture with a discussion around a fluid dynamic interpretation of optimal transport. The material in this section is only treated at an informal level and we refer to [3] for a rigorous treatment.

When $X \subset \mathbb{R}^d$, 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 $v_t : \mathbb{R}^d \to \mathbb{R}^d$ 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 define its flow $T : [0,1] \times X \to \mathbb{R}^d$ 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

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

Let us denote $\mathrm{CE}(\mu, \nu)$ the set of solutions (ρ, v) 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

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

Among all interpolations between μ and ν, it turns out that optimal transport with cost $\|y - x\|^p$ is the one that minimizes A_p. This is called the Benamou-Brenier formulation.

Theorem 5.1 (Dynamic formulation)**.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be compactly supported. For $p \geqslant 1$ it holds*

$$W_p^p(\mu, \nu) = \inf \Big\{ A_p(\rho, v) \mid (\rho, v) \in \mathrm{CE}(\mu, \nu) \Big\}.$$

Let us give some informal arguments to understand this result.

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

$$\begin{aligned} A(\rho, v) &= \int_0^1 \int_{\mathbb{R}^d} \|v_t(T_t(x))\|^p \mathrm{d}\rho_0(x) \mathrm{d}t \\ &= \int_{\mathbb{R}^d} \Big(\int_0^1 \Big\| \frac{d}{dt} T_t(x) \Big\|^p \mathrm{d}t \Big) \mathrm{d}\rho_0(x) \\ &\geqslant \int_{\mathbb{R}^d} \|T_1(x) - T_0(x)\|^p \mathrm{d}\rho_0(x) \end{aligned}$$

  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 $(\rho, v) \in \mathrm{CE}(\mu, \nu)$ 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

$$v_t = \Big(\frac{d}{dt} T_t \Big) \circ T_t^{-1} = (T - \mathrm{id}) \circ T_t^{-1},$$

  which, by construction, is such 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 at the formal level) the Benamou-Brenier formula as a Riemannian formulation for W_2 (this point of view is due to Otto). In this interpretation, the tangent space at $\rho \in \mathcal{P}_2(X)$ are measures of the form $\delta\rho = -\nabla \cdot (v\rho)$ with a velocity field $v \in L^2(\rho, \mathbb{R}^d)$ and the metric is given by

$$\|\delta\rho\|_\rho^2 = \inf_{v \in L^2(\rho, \mathbb{R}^d)} \Big\{ \int \|v(x)\|_2^2 \mathrm{d}\rho(x) \mid \delta\rho = -\nabla \cdot (v\rho) \Big\}.$$

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[2] Haïm Brezis, *Analyse fonctionnelle*, Masson, Halsted Press, 1983.

[3] Filippo Santambrogio, *Optimal transport for applied mathematicians*, Springer, 2015.

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