Lecture 4: Wasserstein space

I. Reminders

- Ingredients: \( X, Y \) compact metric spaces,
  - \( c \in C(X \times Y) \) cost function
  - \( \mu \in \mathcal{P}(X) \) and \( \nu \in \mathcal{P}(Y) \) marginals

- Primal/Kantorovich problem:
  \[
  T_c(\mu, \nu) := \min_{\pi \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) \, d\pi(x, y) \\
  \text{subject to transport plans}
  \]

- Dual problem:
  \[
  T_c(\mu, \nu) = \max_{\Psi \in \mathcal{P}(X)} \int_X \Psi \, d\mu + \int_Y \Psi \, d\nu \\
  \text{subject to } \Psi(x) + \Psi(y) \leq c(x, y) \forall (x, y) \in X \times Y
  \]

- At optimality, it holds \( \Psi(x) + \Psi(y) = c(x, y) \) for \( \lambda \)-a.e. \((x, y)\)

- A few special cases:
  - for \( X = Y \subset \mathbb{R} \) and \( c(x, y) = h(|x-y|) \), \( h \) strictly convex
    then optimal transport plan \( \pi = \text{(unique) monotone plan} \)
    
    \rightarrow subject of the first "practical session" (see the website)
  - for \( X = Y \) and \( c(x, y) = \text{dist}(x, y) \), we have

    \[T_c(\mu, \nu) = \sup_{\pi \in \mathcal{P}(X)} \int_X \Psi \, d\pi - \int_Y \Psi \, d\nu\]

    (Kantorovich-Rubinstein)

  - for \( X = Y \subset \mathbb{R}^d \) and \( c(x, y) = \frac{1}{2} \| y - x \|^2 \), if \( \nu \) is absolutely continuous, there exists a unique transport plan, it is of the form \( \pi = (\text{id}, \mathcal{D}^{\psi})_\# \mu \)
    for some \( \psi \in C(\mathbb{R}^d) \) convex.
II Wasserstein space

II. 1 Definition and first properties

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

\[
W_p(\mu, \nu) := \left( \min_{\gamma \in \Pi(\mu, \nu)} \int \text{dist}(x, y)^p \, d\gamma(x, y) \right)^{1/p} = \left( \inf_{\delta \in \mathcal{M}_p(\mu, \nu)} \int \text{dist}(x, y)^p \, d\delta(x, y) \right)^{1/p}
\]

Property: the map \(X \to \mathcal{P}_p(X)\) is an isometry (i.e. \(W_p(\delta x, \delta y) = \text{dist}(x, y)\)).

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

Proof. - Symmetry is obvious, non-negativity too.

- If \(W_p(\mu, \nu) = 0\) then there exists \(\gamma \in \Pi(\mu, \nu)\) such that \(\int \text{dist}^p \, d\gamma = 0\), so \(\gamma\) is concentrated on the diagonal, thus \(\gamma = (\text{id}, \text{id}) \# \mu\), in other words \(\nu = \text{id} \# \mu = \mu\).

- For the triangle inequality, we rely on:

Lemma (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 \(\gamma_i \in \Pi(\mu_i, \nu_i)\). Then, there exists \(\gamma \in \mathcal{P}(X_1 \times \ldots \times X_N)\) such that for all \(i \in \{1, \ldots, N-1\}\), \((\Pi_i, \text{id}) \# \gamma = \gamma_i\) where \(\Pi_i, \text{id}: X_1 \times \ldots \times X_N \to X_i \times X_{i+1}\) is the projection. (proof see ref.)
be optimal in the definition of $W_p$. There exists $\sigma \in \Sigma(X)$ such that $(\Pi_{x,y})_\# \sigma = \chi_1$ and $(\Pi_{x,y})_\# \sigma = \chi_2$. A fortiori $(\Pi_{x,y})_\# \sigma \in \Pi(\mu, \nu)$. We have

$$W_p(\mu, \nu) \leq \left( \int \text{dist}^p(x, y) \, d\left( (\Pi_{x,y})_\# \sigma \right)(x, y) \right)^\frac{1}{p} \nu \left( \int \text{dist}^p(x, y) \, d\sigma(x, y) \right)^\frac{1}{p}$$

$$= \left( \int \text{dist}^p(x_1, x_3) \, d\sigma(x_1, x_2, x_3) \right)^\frac{1}{p} \nu \left( \int \text{dist}^p(x_1, x_3) \, d\sigma(x_1, x_2, x_3) \right)^\frac{1}{p}$$

$$\leq \left( \int \text{dist}^p(x_1, x_2) \, d\sigma(x_1, x_2, x_3) + \int \text{dist}^p(x_2, x_3) \, d\sigma(x_1, x_2, x_3) \right)^\frac{1}{p} \nu \left( \int \text{dist}^p(x_1, x_3) \, d\sigma(x_1, x_2, x_3) \right)^\frac{1}{p}$$

$$= \left( \int \text{dist}^p(x_1, x_2) \, d\left( (\Pi_{x,y})_\# \sigma \right)(x_1, x_2) \right)^\frac{1}{p} + \ldots$$

$$= W_p(\mu, \nu) + W_p(\nu, \mu)$$

This concludes the proof.

**Exercise:** Prove the triangle inequality assuming the existence of transport maps.

**Remark (non-compact case).** When $X$ is complete and separable metric space then $\Sigma(X)$ as the set of probability measures $\mu \in \Sigma(X)$ such that for some $x_0 \in X$ (and thus any $x \in X$), it holds

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

**Exercise:** show that $W_p$ is finite on this set.

## II.2 Comparisons

- By Jensen's inequality, for any $\gamma \in \Pi(\mu, \nu)$, $p < q$, it holds

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

$$=) \ W_p(\mu, \nu) \leq W_1(\mu, \nu)$$

In particular \( W_p(\mu, \nu) \leq W_1(\mu, \nu) \quad \forall \ p \geq 1 \quad \text{(in full generality)} \).

- For \( X \) compact and thus bounded, we have \( \forall \ Y \in \mathcal{T}(\mu, \nu) : \)

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

  $$=) \quad W_p(\mu, \nu) \leq \text{diam}(X)^{p-1} W_1(\mu, \nu)^{1/p}$$

### II.3 Topological property

**Theorem:** Assume that \( X \) is compact. For \( p \in [1, \infty) \), we have \( \mu_n \rightarrow \mu \) weakly if and only if \( W_p(\mu_n, \mu) \rightarrow 0 \).

**Proof:** Thanks to the comparison inequality, we only need to prove it for \( W_1 \).

- Let \( \mu_n \) be a sequence such that \( W_1(\mu_n, \mu) \rightarrow 0 \).
  - By the Kantorovich-Rubinstein formula, \( \forall \, \psi \in \text{Lip}(X) \), \( \int \psi d(\mu_n - \mu) \rightarrow 0 \).
  - By linearity, \( \forall \, \psi \in \text{Lip}(X) \), \( \int \psi d(\mu_n - \mu) \rightarrow 0 \).
  - By density, \( \forall \, \psi \in \mathcal{C}(X) \), \( \int \psi d(\mu_n - \mu) \rightarrow 0 \).
  - So \( W_1(\mu_n, \mu) \rightarrow 0 \) implies that \( \mu_n \rightarrow \mu \).

- Now assume that \( \mu_n \rightarrow \mu \). Let us fix a subsequence \( (\mu_{n_k}) \) that satisfies \( \lim_k W_1(\mu_{n_k}, \mu) = \lim \sup_n W_1(\mu_n, \mu) \).
  - For \( k \), pick a function \( \psi_{n_k} \in \text{Lip}(X) \) such that
    $$W_1(\mu_{n_k}, \mu) = \int \psi_{n_k} d(\mu_{n_k} - \mu)$$

Assuming that all \( (\psi_{n_k}) \) vanish at the same point, the sequence \( (\mu_{n_k}) \) converges weakly to \( \mu \).
is equi-bounded & equi-continuous. So we can extract a subsequence 
that converges uniformly to \( \Psi \in L^1(\mathbb{P}) \) (by Arzela–Ascoli). Then 
up to taking a subsequence, we have

\[
W_1(\rho_{n}, \rho) = \int \Psi_n d(\rho_{n} - \rho) \rightarrow \int \Psi d(\rho - \rho) = 0
\]

This shows that \( \limsup_n W_1(\rho_{n}, \rho) \leq 0 \) Thus \( W_1(\rho_{n}, \rho) \rightarrow 0 \) \hfill \( \blacksquare \)

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

\[
\int \operatorname{dist}(x, y)^p d\rho_n(y) \rightarrow \int \operatorname{dist}(x_0, y)^p d\rho(y)
\]

III. Geodesics in Wasserstein space

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

Proposition. Let \( \rho_0, \rho_1 \in \mathcal{P}(\mathbb{X}) \) with \( \mathbb{X} \subset \mathbb{R}^d \) compact and convex. 
Let \( \Pi \in \Pi(\rho_0, \rho_1) \) be an optimal transport plan in \( W_1(\rho_0, \rho_1) \). Define 

\[
\mu_t = (\Pi t) \# \Psi \text{ where } \Pi_t(x, y) = (1 - t)x + ty.
\]

Then the curve \( (\mu_t) \) is a constant speed geodesic between \( \rho_0 \) and \( \rho_1 \). 

\[
\int_{\mathbb{R}^d} d\mu_t
\]
Remarks: if there exists an optimal transport map \( T \) between \( \mu_0 \) and \( \mu_1 \), then the geodesic in the proportion is \( \mu_t = ((1-t)\text{id} + tT)\#\mu_0 \).

In fact all the geodesics one of the form given in the proportion.

Proof: First note that if \( 0 < s < t < 1 \),
\[
W_p(\mu_0, \mu_t) \leq W_p(\mu_0, \mu_s) + W_p(\mu_s, \mu_t) + W_p(\mu_t, \mu_1)
\]
so it is enough to prove \( W_p(\mu_s, \mu_t) \leq |t-s| W_p(\mu_0, \mu_1) \) to get equality.

Let \( \gamma \in \Pi(\mu_0, \mu_1) \) an optimal transport plan.

Take \( \gamma_{st} = (\mu_s, \mu_t) \# \gamma \in \Pi(\mu_s, \mu_t) \).

It holds:
\[
W_p(\mu_s, \mu_t)^p = \int \|x-y\|^p d\gamma_{st}(x,y)
\]
\[
= \int \| \pi_s(x,y) - \pi_t(x,y) \|^p d\gamma(x,y)
\]
\[
= \int \| (1-s)x + sy - ((1-t)x + ty) \|^p d\gamma(x,y)
\]
\[
= \int \| (t-s)(x-y) \|^p d\gamma(x,y) = |t-s|^p W_p(\mu_0, \mu_1)^p
\]
so \( W_p(\mu_s, \mu_t) \leq |t-s| W_p(\mu_0, \mu_1) \) and thus \( (\mu_t) \) is a constant speed geodesic.

Corollary: The space \( (\mathbb{P}_p(X), W_p) \) with \( X \subset \mathbb{R}^d \) compact and convex is a geodesic space, meaning that any \( \mu_0, \mu_1 \in \mathbb{P}_p(X) \) can be joined by (at least one) constant speed geodesic.
$W_p(\mu_0, \mu_1) \leq W_p(\mu_0, \mu_3) + W_p(\mu_3, \mu_4) + W_p(\mu_4, \mu_1)$

$\leq |s| W_p(\mu_0, \mu_1) + |s-1| W_p(\mu_0, \mu_1) + |t-1| W_p(\mu_0, \mu_1)$

$= W_p(\mu_0, \mu_1)$

**Barycenters in $\mathcal{P}_2(X)$.**

- The notion of geodesic allows to define barycenter between two probability distributions. Can this be generalized to more than 2 distributions?

- In $\mathbb{R}^d$, the barycenter of $x_1, ..., x_n$ with weights $\lambda_1, ..., \lambda_n > 0$ is the unique point $y$ that minimizes $\sum_{i=1}^n \lambda_i \| y - x_i \|_2^2$.

- This motivate to define $W_2$-barycenters between $\mu_1, ..., \mu_n \in \mathcal{P}_2(X)$ with weights $\lambda_1, ..., \lambda_n > 0$ as any measure that solves

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

Remark: when $\mu_i = Sx_i$, we recover the usual notion of barycenter on $\mathbb{R}^d$. 

$\mu_i \in \mathcal{S}(\mathbb{R}^2)$
Theorem: Let $\sigma, \rho, \rho', \in \mathcal{P}(X)$. Assume that there exists a unique pair $(\Psi_0, \Psi_0)$ of Kantorovich potentials between $\sigma$ and $\rho$ which are $c$-conjugate to each other and satisfy $\Psi_0(x_0) = 0$ for some $x_0 \in X$. Then,

$$\frac{d}{dt} T_c(\sigma, \rho + t (\rho' - \rho_0)) \bigg|_{t=0} = \int \Psi_0 \, d(\rho' - \rho_0) \quad \blacklozenge$$

NB. Taking $c = \text{dist}^\rho$, this allows to differentiate $t \mapsto W_p^\rho(\sigma, \rho)$. 

Proof: Denote $\rho_t = (1-t)\rho_0 + t \rho_1 = \rho_0 + t(\rho' - \rho_0)$. By Kantorovich duality, we have:

$$T_c(\sigma, \rho_t) \geq \int \Psi_0 \, d\sigma + \int \Psi_0 \, d\rho_t$$

So,

$$T_c(\sigma, \rho_t) - T_c(\sigma, \rho_0) \geq \int \Psi_0 \, d\sigma + \int \Psi_0 \, d\rho_0 + t \int \Psi_0 \, d(\rho' - \rho_0)$$

$$\geq \int \Psi_0 \, d(\rho' - \rho_0)$$

Thus,

$$\frac{1}{t} (T_c(\sigma, \rho_t) - T_c(\sigma, \rho_0)) \geq \int \Psi_0 \, d(\rho' - \rho_0)$$

To show the converse inequality, let $(\Psi_t, \Psi_t)$ be $c$-conjugate Kantorovich potentials between $\sigma$ and $\rho_t$ satisfying $\Psi_t(x_0) = 0$, we have,

$$\frac{1}{t} (T_c(\sigma, \rho_0) - T_c(\sigma, \rho_t)) \geq \int \Psi_t \, d(\rho' - \rho_0)$$

By uniqueness of $(\Psi_0, \Psi_0)$, we get that $(\Psi_t, \Psi_t)$ converges uniformly to $(\Psi_0, \Psi_0)$ as $t \to 0$. This concludes the proof. \blacklozenge
Remark: \( T_c(\mu, \nu) = \sup_{(\zeta, \zeta') \in \mathcal{E}} \int \zeta d\mu + \int \zeta' d\nu \)

The assumption of uniqueness can be guaranteed in particular in the following setting.

**Proposition:** If \( X \subset \mathbb{R}^d \) is the closure of a bounded and connected open set, \( x_0 \in X \), \((\mu, \nu) \in \mathcal{M}(X)\) such that \( \mu \) absolutely continuous and \( \text{spt}(\mu) = X \). Then there exists a unique pair \((\Psi, \Psi')\) of Kantorovich potentials optimal for \( c(x, y) = \frac{1}{2} \| x - y \|^2 \), \( c \)-conjugate to each other and satisfying \( \Psi(x_0) = 0 \).

**Proof:** Since \( c \) is Lipschitz on \( X \), \( \Psi \) and \( \Psi' \) are Lipschitz and therefore differentiable almost everywhere. Take \((x_0, y_0) \in \text{spt}(\mu)\) where \( Y \in \mathcal{T}(\mu, \nu) \) is an optimal Kantorovich plan, such that \( \Psi \) is differentiable at \( x_0 \in X \). As shown in Lecture 2, for any optimal pair \((\Psi, \Psi')\) we have

\[
y_0 = x_0 - D\Psi(x_0)
\]

so if \((\Psi', \Psi'')\) is another optimal pair, we should have \( D\Psi = D\Psi' \) \( \mu \)-almost everywhere. Since \( \text{spt}(\mu) = X \), and \( X \) is the closure of a connected open set, \( \Psi = \Psi' + C \) for a constant \( C \), which is \( C = 0 \) if \( \Psi(x_0) = \Psi'(x_0) = 0 \).
Dynamic formulation of optimal transport

Discussion at a informal level (see references for proofs).

When $X \subset \mathbb{R}^d$, interpret $\mu$ and $\nu \in \mathcal{P}(X)$ as distributions of particles at $t = 0$ and at $t = 1$. We call $(\rho_t)_{t \in [0,1]}$, the distribution of particles that evolve in time.

Assume that there is a velocity field $V_t : \mathbb{R}^d \to \mathbb{R}^d$ which move the particles around. The relation between $\rho_t$ and $V_t$ is given by

$$\partial_t \rho_t + \nabla \cdot (\rho_t V_t) = 0 \quad \text{(Continuity equation)}$$

(understood in the distributional sense)

Let us denote by $\mathcal{CE}(\mu, \nu) = \{ (\rho, v) \text{ solves the continuity equation } \}
\begin{cases} 
\rho_0 = \mu, \, \rho_1 = \nu \\
\rho \to \rho_t \text{ is weakly continuous}
\end{cases}$

Consider the functional

$$A_p(\rho, v) = \int_0^1 \int_X \| V_t(x) \|^p \, d\rho_t(x) \quad \text{"kinetic energy"}$$

Theorem (Benamou-Brenier formula). Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be compactly supported. For $p \geq 1$, it holds:

$$W_p(\mu, \nu) = \inf \left\{ \int_{\mathbb{R}^2} (\rho, v) ; (\rho, v) \in \mathcal{CE}(\mu, \nu) \right\}$$
Remark (Riemannian interpretation) - In case $p=2$, we can understand the Benamou-Brenier formula as Riemannian formula for $W_2$: the tangent space at $p \in \mathcal{S}(\mathcal{L}^2)$ are measures of the form $Sp = -\nabla \cdot (\nu p)$ with $\nu \in L^2(p, \mathbb{R}^d)$, the metric is given by:

$$\|Sp\|^2 = \inf_{\nu \in L^2(p, \mathbb{R}^d)} \left\{ \int \|\nu(x)\|^2 \, dp(x) \mid Sp = -\nabla \cdot (\nu p) \right\}$$

⇒ Next week you will see "Wasserstein gradient flows".