Lecture 4 : Wasserstein space

I Reminders

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

- Primal / Kantorovich problem :

$$T_c(\mu, \nu) := \min_{\gamma \in \Pi(\mu, \nu)} \int_{X \times Y} c(x,y)\, d\gamma(x,y)$$

($\Pi(\mu,\nu)$ ← set of transport plans)

- Dual problem :

$$T_c(\mu, \nu) = \max_{\substack{\varphi \in \mathcal{C}(X) \\ \psi \in \mathcal{C}(Y)}} \int_X \varphi \, d\mu + \int_Y \psi \, d\nu \quad \text{s.t. } \varphi(x) + \psi(y) \leq c(x,y) \ \forall (x,y) \in X \times Y$$

(strong duality)

- At optimality, it holds $\varphi(x) + \psi(y) = c(x,y)$ for γ-a.e. (x,y)

- A few special cases :

  • for $X = Y \subset \mathbb{R}$ and $c(x,y) = h(y-x)$, h strictly convex - then optimal transport plan γ = (unique) monotone plan
  → subject of the first "practical session" (see the web site).

  • for $X = Y$ and $c(x,y) = \text{dist}(x,y)$, we have

(Kantorovich - Rubinstein) $$T_c(\mu, \nu) = \sup_{\varphi \text{ 1-Lipschitz}} \int_X \varphi \, d(\mu - \nu)$$

  • for $X = Y \subset \mathbb{R}^d$ and $c(x,y) = \frac{1}{2}\|y-x\|_2^2$, if μ is absolutely continuous, there exists a unique transport plan, it is of the form $\gamma = (\text{id}, \nabla\tilde{\varphi})_\# \mu$ for some $\tilde{\varphi} \in \mathcal{C}(\mathbb{R}^d)$ convex.

II Wasserstein space

II.1 Definition and first properties

Def (Wasserstein space). Let (X, 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:

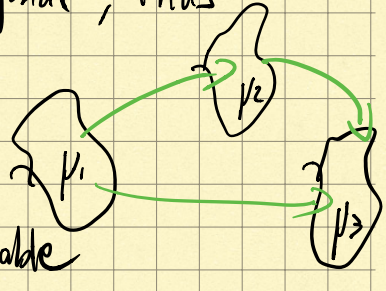
$$W_p(\mu,\nu) := \left(\min_{\gamma \in \Pi(\mu,\nu)} \int \text{dist}(x,y)^p \, d\gamma(x,y) \right)^{1/p}$$
$$= T_{\text{dist}^p}(\mu,\nu)^{1/p}$$

Property: the map $\begin{cases} X \to \mathcal{P}_p(X) \\ x \mapsto \delta_x \end{cases}$ is an isometry (i.e $W_p(\delta_x, \delta_y) = \text{dist}(x,y)$)

δ_x ← Dirac mass located at x

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

$\pi_{i,i+1} : X_1 \times \dots \times X_N \to X_i \times X_{i+1}$ is the projection. (proof see refs).

Let $\mu_i \in \mathcal{P}(X)$ for $i \in \{1,2,3\}$ 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. There exists $\sigma \in \mathcal{P}(X^3)$ such that $(\pi_{1,2})_\# \sigma = \gamma_1$ and $(\pi_{2,3})_\# \sigma = \gamma_2$. A fortiori $(\pi_{1,3})_\# \sigma \in \Pi(\mu_1, \mu_3)$.

We have

$$W_p(\mu_1,\mu_3) \leq \left(\int \mathrm{dist}(x,y)^p \, d[(\pi_{1,3})_\# \sigma](x,y) \right)^{1/p}$$

$$= \left(\int_{X^3} \mathrm{dist}(x_1,x_3)^p \, d\sigma(x_1,x_2,x_3) \right)^{1/p}$$

$$\leq \left(\int_{X^3} \left(\mathrm{dist}(x_1,x_2) + \mathrm{dist}(x_2,x_3) \right)^p d\sigma(x_1,x_2,x_3) \right)^{1/p}$$

Triangle inequality in $L^p(\sigma)$ / Minkowski inequality

$$\leq \left(\int_{X^3} \mathrm{dist}(x_1,x_2)^p \, d\sigma(x_1,x_2,x_3) \right)^{1/p} + \left(\int_{X^3} \mathrm{dist}(x_2,x_3)^p \, d\sigma(x_1,x_2,x_3) \right)^{1/p}$$

$$= \left(\int \mathrm{dist}(x_1,x_2)^p \, d\underbrace{[(\pi_{1,2})_\# \sigma]}_{=\gamma_1}(x_1,x_2) \right)^{1/p} + \dots$$

$$= W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3)$$

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 $\mathcal{P}_p(X)$ as the set of probability measures $\mu \in \mathcal{P}(X)$ such that for some $x_0 \in X$ (and thus any $x_0 \in X$), it holds

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

Exercice: show that W_p is finite on this set. ∎

II.2 Comparisons

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

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

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

$$\Rightarrow \quad W_p(\mu,\nu) \leq W_q(\mu,\nu)$$

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

- For X compact and thus bounded, we have for $\gamma \in \Pi(\mu,\nu)$:

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

$$\Rightarrow \quad \underline{W_p(\mu,\nu) \leq \mathrm{diam}(X)^{\frac{p-1}{p}} W_1(\mu,\nu)^{1/p}}$$

II.3 Topological property

Theorem: Assume that X is compact. For $p \in [1,+\infty[$ we have

$\mu_n \rightharpoonup \mu$ (weak convergence) if and only if $W_p(\mu_n,\mu) \to 0$.

(weak* = weak convergence)

Proof: Thanks to the comparison inequalities, we only need to prove it for W_1.

- Let μ_n be a sequence such that $W_1(\mu_n,\mu) \to 0$.

  By the Kantorovich-Rubinstein formula, $\forall \varphi \in 1\text{-Lip}(X)$ $\int \varphi d(\mu_n - \mu) \to 0$

  By linearity $\forall \varphi \in \mathrm{Lip}(X)$, $\int \varphi d(\mu_n - \mu) \to 0$

  By density $\forall \varphi \in \mathcal{C}(X)$, $\int \varphi d(\mu_n - \mu) \to 0$

  So $W_1(\mu_n,\mu) \to 0$ implies that $\mu_n \rightharpoonup \mu$.

- Now assume that $\mu_n \rightharpoonup \mu$. Let us fix a subsequence (μ_{n_k}) that satisfies $\lim_k W_1(\mu_{n_k},\mu) = \limsup_n W_1(\mu_n,\mu)$

  $\forall k$, pick a function $\varphi_{n_k} \in 1\text{-Lip}(X)$ such that

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

Assuming that all (φ_{n_k}) vanish at the same point, the sequence (φ_{n_k})

is equi-bounded & equi-continuous. So we can extract a subsequence that converges uniformly to $\varphi \in 1\text{-Lip}(X)$ (by Ascoli-Arzelà). Then up to taking a subsequence, we have:

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

($\varphi_{n_k} \to \varphi$, $\mu_{n_k} \to \mu$)

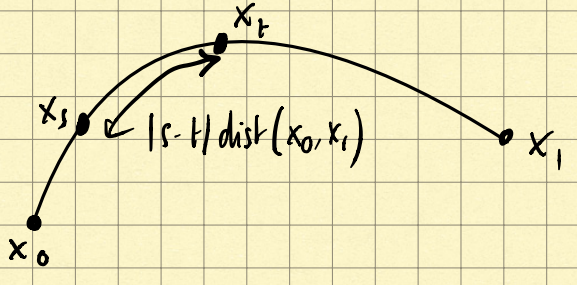
This shows that $\limsup_n W_1(\mu_n, \mu) \leq 0$ thus $W_1(\mu_n, \mu) \to 0$ ∎

Remark (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 $x_0 \in X$,

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

III Geodesics in Wasserstein space

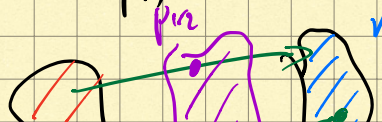
Definition. Let (X, dist) be a metric space. A constant speed geodesic between two points $x_0, x_1 \in X$ is a continuous curve $\gamma : [0,1] \to X$ such that for every $s, t \in [0,1]$, $\text{dist}(x_s, x_t) = |s-t| \, \text{dist}(x_0, x_1)$.

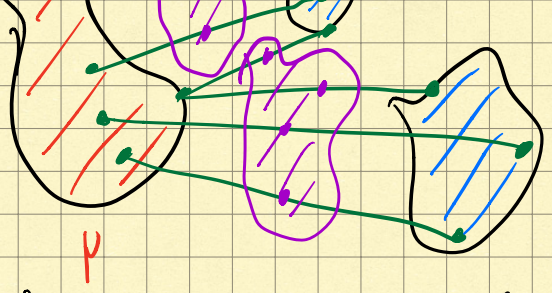


Proposition. Let $\mu_0, \mu_1 \in \mathcal{P}_p(X)$ with $X \subset \mathbb{R}^d$ compact and convex. Let $\gamma \in \Pi(\mu_0, \mu_1)$ be an optimal transport plan for W_p ($p \geq 1$). Define

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

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





Remarks: • if there exists an optimal transport map T between μ_0 and μ_1, then the geodesic in the proposition is $\mu_t = ((1-t)\mathrm{id} + tT)_\# \mu_0$

• in fact all the geodesics are of the form given in the Proposition.

Proof: First note that if $0 \leq s \leq t \leq 1$,

$$W_p(\mu_0, \mu_1) \leq W_p(\mu_0, \mu_s) + W_p(\mu_s, \mu_t) + W_p(\mu_t, \mu_1)$$ (triangle inequality)

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 $\gamma_{st} := (\pi_s, \pi_t)_\# \gamma \in \Pi(\mu_s, \mu_t)$.

It holds:

$$W_p(\mu_s, \mu_t)^p \leq \int \|x-y\|^p \, d\gamma_{st}(x,y)$$

$$= \int \|\pi_s(x,y) - \pi_t(x,y)\|^p \, d\gamma(x,y)$$

$$= \int \|(1-s)x + sy - ((1-t)x + ty)\|^p \, d\gamma(x,y)$$

$$= \int \|(t-s)(x-y)\|^p \, d\gamma(x,y) = |t-s|^p W_p(\mu_0, \mu_1)^p$$

so $W_p(\mu_s, \mu_t) \leq |t-s| W_p(\mu_0, \mu_1)$ and thus (μ_t) is a constant speed geodesic. ∎

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

$$W_p(\mu_0,\mu_1) \le W_p(\mu_0,\mu_s) + W_p(\mu_s,\mu_t) + W_p(\mu_t,\mu_1)$$
$$\le |s|\, W_p(\mu_0,\mu_1) + |s-t|\, W_p(\mu_0,\mu_1) + |t-1|\, W_p(\mu_0,\mu_1)$$
$$= W_p(\mu_0,\mu_1)$$

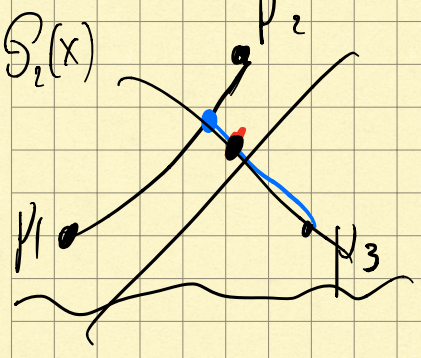
Barycenters in $\mathcal{P}_2(X)$.

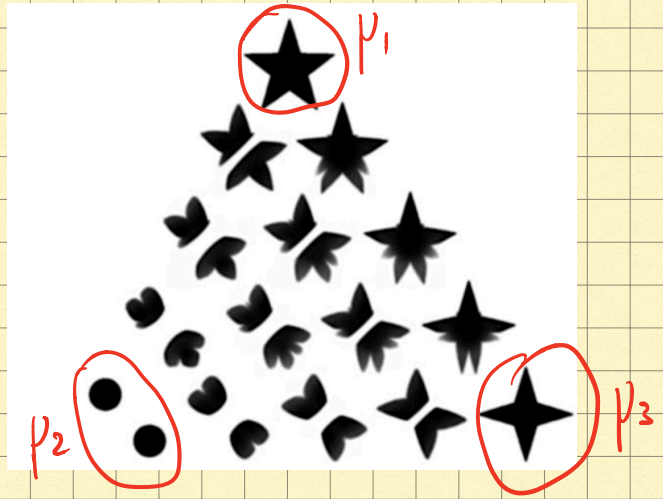
- The notion of geodesic allows to define barycenters between two probability distributions. Can this be generalized to more than 2 distributions?

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

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

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





$\mu_i \in \mathcal{P}(\mathbb{R}^2)$

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

V Differentiability of the Wasserstein distance

Theorem: Let $\sigma, \rho_0, \rho_1 \in \mathcal{P}(X)$ (X compact). Assume that there exists a unique pair (φ_0, ψ_0) of Kantorovich potentials 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,

$$\frac{d}{dt} T_c(\sigma, \rho_0 + t(\rho_1 - \rho_0))\Big|_{t=0} = \int \psi_0 \, d(\rho_1 - \rho_0) \quad \blacksquare$$

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

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

$$T_c(\sigma, \rho_t) \geq \int \varphi_0 \, d\sigma + \int \psi_0 \, d\rho_t$$

So $T_c(\sigma, \rho_t) - T_c(\sigma, \rho_0) \geq \int \varphi_0 \, d\sigma + \int \psi_0 \, d\rho_0 + t\int \psi_0 \, d(\rho_1 - \rho_0) - \int \varphi_0 \, d\sigma - \int \psi_0 \, d\rho_0$

$$\geq t \int \psi_0 \, d(\rho_1 - \rho_0)$$

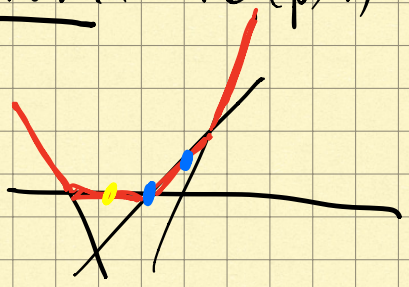
Thus $\frac{1}{t}\left(T_c(\sigma, \rho_t) - T_c(\sigma, \rho_0)\right) \geq \int \psi_0 \, d(\rho_1 - \rho_0)$

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

$$\frac{1}{t}\left(T_c(\sigma, \rho_0) - T_c(\sigma, \rho_t)\right) \geq \int \psi_t \, d(\rho_1 - \rho_0)$$

By uniqueness of (φ_0, ψ_0), we get that (φ_t, ψ_t) converges uniformly to (φ_0, ψ_0) as $t \to 0$. This concludes the proof $\blacksquare$

Remark: $T_c(\mu,\nu) = \sup_{(\varphi,\psi)\in\mathcal{C}} \int \varphi\, d\mu + \int \psi\, d\nu$



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, $x_0 \in X$, $(\mu,\nu) \in \mathcal{P}(X)$ such that μ absolutely continuous (with respect to Lebesgue) and $\mathrm{spt}(\mu) = X$. Then there exists a unique pair (φ,ψ) of Kantorovich potentials optimal for $c(x,y) = \frac{1}{2}\|x-y\|_2^2$, c-conjugate to each other and satisfying $\varphi(x_0) = 0$.

Proof: Since c is Lipschitz on X, φ and ψ are Lipschitz and therefore differentiable almost everywhere. Take $(x_0, y_0) \in \mathrm{spt}(\gamma)$ where $\gamma \in \Pi(\mu,\nu)$ is an optimal Kantorovich plan, such that φ is differentiable at $x_0 \in \mathring{X}$. As shown in Lecture 2, (OT for twisted costs) for any optimal pair (φ,ψ) we have

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

so if (φ',ψ') is another optimal pair, we should have $\nabla\varphi = \nabla\varphi'$ μ almost everywhere. Since $\mathrm{spt}(\mu) = X$, and X is the closure of a connected open set, $\varphi = \varphi' + C$ for a constant C, which is $C = 0$ if $\varphi(x_0) = \varphi'(x_0) = 0$.

V Dynamic formulation of optimal transport

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

When $X \subset \mathbb{R}^d$, interpret μ 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 moves the particles around. The relation between ρ_t and v_t is given by

$$\partial_t \rho_t + \underbrace{\nabla \cdot}_{\text{divergence}} (\rho_t v_t) = 0 \qquad \text{(Continuity equation)}$$

(understood in the distributional sense).

Let us denote by $CE(\mu,\nu) = \left\{ (\rho, v) \text{ solves the continuity equation} \right.$
$\rho_0 = \mu,\ \rho_1 = \nu$
$\left. t \to \rho_t \text{ is weakly continuous} \right\}$

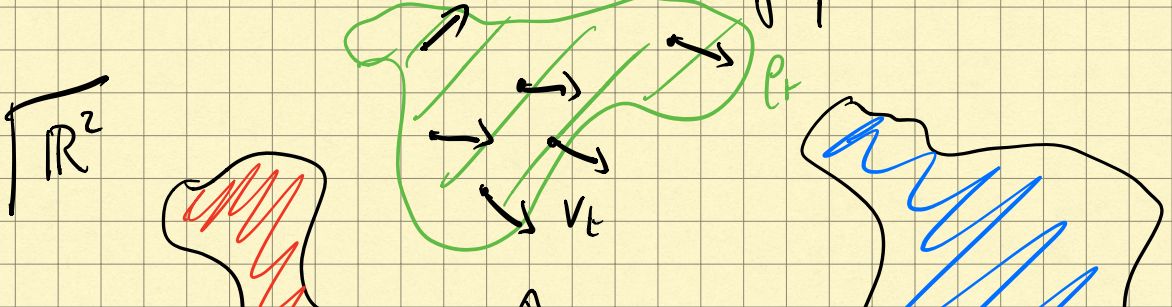
Consider the functional

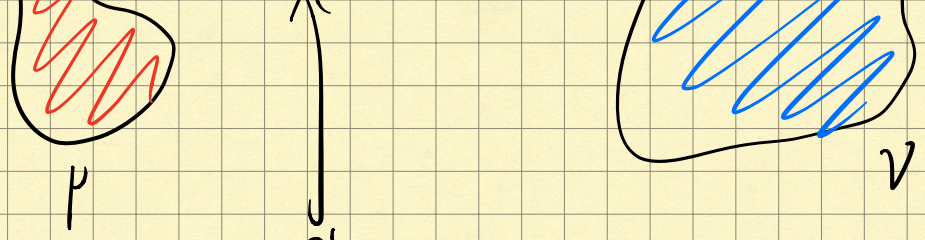
$$A_p(\rho, v) = \int_0^1 \int_X \|v_t(x)\|^p \, d\rho_t(x)$$

"kinetic energy integrated in time".

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

$$W_p^p(\mu,\nu) = \inf \left\{ A_p(\rho, v) \,;\, (\rho, v) \in CE(\mu,\nu) \right\}$$





$$\int_0^1 \int_X |v_t(x)|^p \, d\rho_t(x)$$

Remark (Riemannian interpretation) - In case $p=2$, we can understand the Benamou-Brenier formula as Riemannian formulation for W_2:

- the tangent space at $\rho \in \mathcal{P}_2(X)$ are measures of the form $\delta\rho = -\nabla\cdot(v\rho)$ with $v \in L^2(\rho, \mathbb{R}^d)$
- the metric is given by:

$$\|\delta\rho\|_\rho^2 = \inf_{v \in L^2(\rho,\mathbb{R}^d)} \left\{ \int \|v(x)\|_2^2 \, d\rho(x) \; ; \; \delta\rho = -\nabla\cdot(v\rho) \right\}$$

$\rightsquigarrow$ Next week you will see "Wasserstein gradient flows".
