Lecture 1
Monge and Kantorovich problems: from primal to dual

Luca Nenna

February 5, 2020

These notes are based on the ones by Quentin Mérigot

Some motivations for studying optimal transport.

- Variational principles for (real) Monge-Ampère equations occuring in geometry (e.g. Gaussian curvature prescription) or optics.
- Wasserstein/Monge-Kantorovich distance between probability measures μ, ν on e.g. $\mathbb{R}^d$: how much kinetic energy does one require to move a distribution of mass described by μ to ν ?
  $\longrightarrow$ interpretation of some parabolic PDEs as Wasserstein gradient flows, construction of (weak) solutions, numerics, e.g.

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho v) = 0 \\ v = -\nabla \log \rho \end{cases} \quad \text{or} \quad \begin{cases} \partial_t \rho + \operatorname{div}(\rho v) = 0 \\ v = -\nabla p - \nabla V \\ p(1-\rho) = 0 \\ p \geqslant 0, \rho \leqslant 1 \end{cases}$$

  $\longrightarrow$ interesting geometry on $\mathcal{P}(X)$, with an embedding $X \hookrightarrow \mathcal{P}(X)$. Applications in geometry (synthetic notion of Ricci curvature for metric spaces), machine learning, inverse problems, etc.
- Quantum physics: electronic configuration in molecules and atoms.

References.

Introduction to optimal transport, with applications to PDE and/or calculus of variations can be found in books by Villani [6] and Santambrogio [5]. Villani's second book [7] concentrates on the application of optimal transport to geometric questions (e.g. synthetic definition of Ricci curvature), but its first chapters might be useful. We also mention Gigli, Ambrosio and Savaré [2] for the study of gradient flows with respect to the Monge-Kantorovich/Wasserstein metric.

Notation.

In the following, we assume that X is a *complete and separable metric space*. We denote $\mathcal{C}(X)$ the space of continuous functions, $\mathcal{C}_0(X)$ the space of continuous function vanishing

at infinity $\mathcal{C}_b(X)$ the space of bounded continuous functions. We denote $\mathcal{M}(X)$ the space of Borel regular measures on X with finite total mass and

$$\mathcal{M}^+(X) := \{\mu \in \mathcal{M}(X) \mid \mu \geqslant 0\}$$

$$\mathcal{P}(X) := \{\mu \in \mathcal{M}^+(X) \mid \mu(X) = 1\}$$

Some reminders.

Definition 0.1 (Lower semi-continuous function)**.** On a metric space Ω, a function $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ is said to be lower semi-continuous (l.s.c.) is for every sequence $x_n \to x$ we have $f(x) \leqslant \liminf_n f(x_n)$.

Definition 0.2. A metric space Ω is said to be compact if from any sequence x_n, we can extract a converging subsequence $x_{n_k} \to x \in \Omega$.

Theorem 0.3 (Weierstrass)**.** *If $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ is l.s.c. and Ω is compact, then there exists $x^\star \in \Omega$ such that $f(x^\star) = \min\{f(x) \mid x \in \Omega\}$.*

Definition 0.4 (weak and weak$-\star$ convergence)**.** A sequence x_n in a Banach space $\mathcal{X}$ is said to be weakly converging to x and we write $x_n \rightharpoonup x$, if for every $\eta \in \mathcal{X}'$ ($\mathcal{X}'$ is the topological dual of $\mathcal{X}$ and $\langle \cdot, \cdot \rangle$ is the duality product) we have $\langle \eta, x_n \rangle \to \langle \eta, x \rangle$. A sequence $\eta_n \in \mathcal{X}'$ is said to be weakly-$\star$ converging to $\eta \in \mathcal{X}'$, and we write $\eta_n \overset{\star}{\rightharpoonup} \eta$, if for every $x \in \mathcal{X}$ we have $\langle \eta_n, x \rangle \to \langle \eta, x \rangle$.

Theorem 0.5 (Banach-Alaoglu)**.** *If $\mathcal{X}'$ is separable and φ_n is a bounded sequence in $\mathcal{X}'$, then there exists a subsequence φ_{n_k} weakly-$\star$ converging to some $\varphi \in \mathcal{X}'$*

Theorem 0.6 (Riesz)**.** *Let X be a compact metric space and $\mathcal{X} = \mathcal{C}(X)$ then every element of $\mathcal{X}$ is represented in a unique way as an element of $\mathcal{M}^+(X)$, that is for every $\eta \in \mathcal{X}$ there exists a unique $\lambda \in \mathcal{M}^+(X)$ such that $\langle \eta, \varphi \rangle = \int_X \varphi \mathrm{d}\lambda$ for every $\varphi \in \mathcal{X}$.*

Definition 0.7 (Narrow convergence)**.** A sequence of finite measures $(\mu_n)_{n \geqslant 1}$ on X narrowly converges to $\mu \in \mathcal{M}(X)$ if

$$\forall \varphi \in \mathcal{C}_b(X), \quad \lim_{n \to \infty} \int_X \varphi \mathrm{d}\mu_n = \int_X \varphi \mathrm{d}\mu.$$

With a slightly abuse of notation we will denote it by $\mu_n \rightharpoonup \mu$.

Remark 0.8. Since we will mostly work on compact set X, then $\mathcal{C}(X) = \mathcal{C}_0(X) = \mathcal{C}_b(X)$. This means that the narrow convergence of measures, that is the notion of convergence in duality with $\mathcal{C}_b(X)$, corresponds to the weak-$\star$ convergence (the convergence in duality with $\mathcal{C}_0(X)$).

1 The problems of Monge and Kantorovich

1.1 Monge problem

Definition 1.1 (Push-forward and transport map)**.** Let X, Y be metric spaces, $\mu \in \mathcal{M}(X)$ and $T : X \to Y$ be a measurable map. The *push-forward* of μ by T is the measure $T_\# \mu$ on Y defined by

$$\forall B \subseteq Y, \quad \mathrm{T}_\# \mu(B) = \mu(T^{-1}(B)).$$

or equivalently if the following change-of-variable formula holds for all measurable and bounded $\varphi : Y \to \mathbb{R}$:

$$\int_Y \varphi(y)\mathrm{d}T_{\#}\mu(y) = \int_X \varphi(T(x))\mathrm{d}\mu(x).$$

A measurable map $T : X \to Y$ such that $T_{\#}\mu = \nu$ is also called a *transport map* between μ and ν.

Example 1.2. If $Y = \{y_1, \dots, y_n\}$, then $T_{\#}\mu = \sum_{1 \leqslant i \leqslant n} \mu(T^{-1}(\{y_i\}))\delta_{y_i}$.

Example 1.3. Assume that T is a $\mathcal{C}^1$ diffeomorphism between open sets X, Y of $\mathbb{R}^d$, and assume also that the probability measures μ, ν have continuous densities ρ, σ with respect to the Lebesgue measure. Then,

$$\int_Y \varphi(y)\sigma(y)\mathrm{d}y = \int_X \varphi(T(x))\sigma(T(x))\det(\mathrm{D}T(x))\mathrm{d}x.$$

Hence, T is a transport map between μ and ν iff

$$\forall \varphi \in \mathcal{C}_b(X), \int_X \varphi(T(x))\sigma(T(x))\det(\mathrm{D}T(x))\mathrm{d}x = \int_X \varphi(T(x))\rho(x)\mathrm{d}x$$

Hence, T is a transport map iff the non-linear Jacobian equation holds

$$\rho(x) = \sigma(T(x))\det(\mathrm{D}T(x)).$$

Definition 1.4 (Monge problem)**.** Consider two metric spaces X, Y, two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a *cost function* $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$. *Monge's problem* is the following optimization problem

$$(\mathrm{MP}) := \inf \left\{ \int_X c(x, T(x))\mathrm{d}\mu(x) \mid T : X \to Y \text{ and } T_{\#}\mu = \nu \right\} \tag{1.1}$$

This problem exhibits several difficulties, one of which is that both the constraint ($T_{\#}\mu = \nu$) and the functional are non-convex.

Example 1.5. There might exist no transport map between μ and ν. For instance, consider $\mu = \delta_x$ for some $x \in X$. Then, $\mathrm{T}_{\#}\mu(B) = \mu(T^{-1}(B)) = \delta_{T(x)}$. In particular, if $\mathrm{card}(\mathrm{spt}(\nu)) > 1$ (see Def. 1.15), there exists no transport map between μ and ν.

Example 1.6. The infimum might not be attained even if μ is atomless (i.e. for every point $x \in X$, $\mu(\{x\}) = 0$). Consider for instance $\mu = \frac{1}{2}\,\lambda|_{\{\pm 1\}\times[-1,1]}$ on $\mathbb{R}^2$ and $\nu = \lambda|_{\{0\}\times[-1,1]}$, where λ is the Lebesgue measure. One solution is to allow mass to split, leading to Kantorovich's relaxation of Monge's problem.

1.2 Kantorovich problem

Definition 1.7 (Marginals)**.** The *marginals* of a measure γ on a product space $X \times Y$ are the measures $\pi_{X\#}\gamma$ and $\pi_{Y\#}\gamma$, where $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ are their projection maps.

Definition 1.8 (Transport plan)**.** A transport plan between two probabily measures μ, ν on two metric spaces X and Y is a probability measure γ on the product space $X \times Y$ whose marginals are μ and ν. The space of transport plans is denoted $\Pi(\mu, \nu)$, i.e.

$$\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(X \times Y) \mid \pi_{X\#}\gamma = \mu, \quad \pi_{Y\#}\gamma = \nu\}.$$

Note that $\Pi(\mu, \nu)$ is a convex set.

Example 1.9 (Tensor product)**.** Note that the set of transport plans $\Pi(\mu,\nu)$ is never empty, as it contains the measure $\mu\otimes\nu$.

Example 1.10 (Transport plan associated to a map)**.** Let T be a transport map between μ and ν, and define $\gamma_T = (id,T)_{\#}\mu$. Then, γ_T is a transport plan between μ and ν.

Definition 1.11 (Kantorovich problem)**.** Consider two metric spaces X,Y, two probability measures $\mu\in\mathcal{P}(X)$, $\nu\in\mathcal{P}(Y)$ and a *cost function* $c: X\times Y\to\mathbb{R}\cup\{+\infty\}$. *Kantorovich's problem* is the following optimization problem

$$(\mathrm{KP}) := \inf\left\{\int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y)\mid \gamma\in\Pi(\mu,\nu)\right\} \tag{1.2}$$

Remark 1.12. The infimum in Kantorovich problem is less than the infimum in Monge problem. Indeed, consider a transport map satisfying $T_{\#}\mu=\nu$ and the associated transport plan γ_T. Then, by the change of variable one has

$$\int_{X\times Y} c(x,y)\mathrm{d}(id,T)_{\#}\mu(x,y) = \int_X c(x,T(x))\mathrm{d}\mu,$$

thus proving the claim.

Example 1.13 (Finite support)**.** Assume that $X=Y=\{1,\dots,N\}$ and that μ,ν are the uniform probability measures over X and Y. Then, Monge's problem can be rewritten as a minimization problem over bijections between X and Y:

$$\min\{\frac{1}{N}\sum_{1\leqslant i\leqslant N} c(i,\sigma(i))\mid\sigma\in\mathfrak{S}_N\}.$$

In Kantorovich's relaxation, the set of transport plans $\Pi(\mu,\nu)$ agrees with the set of bi-stochastic matrices :

$$\gamma\in\Pi(\mu,\nu)\Longleftrightarrow\gamma\geqslant 0,\sum_i\gamma(i,j)=1/N=\sum_j\gamma(i,j).$$

By Birkhoff's theorem, any extremal bi-stochastic matrix is induced by a permutation. This shows that, in this case, the solution to Monge's and Kantorovich's problems agree.

Remark 1.14. Proposition 1.16 shows that a transport plan concentrated on the graph of a function $T: X\to Y$ is actually induced by a transport map. One can prove that transport plans concentrated on graphs are extremal points in the convex set $\Pi(\mu,\nu)$, but the converse does not hold in general (the counter-examples are quite tricky to construct, see [1]). This means that one cannot resort to a simple argument such as Birkhoff's theorem to show that solutions to Kantorovich's problem (transport plans) are induced by transport maps.

Definition 1.15 (Support)**.** Let Ω be a separable metric space. The *support* of a non-negative measure μ is the smallest closed set on which μ is concentrated

$$\mathrm{spt}(\mu) := \bigcap\{A\subseteq\Omega\mid A\text{ closed and }\mu(X\setminus A)=0\}.$$

A point x belongs to $\mathrm{spt}(\mu)$ iff for every $r>0$ one has $\mu(\mathrm{B}(x,r))>0$.

Proposition 1.16. *Let $\gamma\in\Pi(\mu,\nu)$ and $T: X\to Y$ measurable be such that $\gamma(\{(x,y)\in X\times Y\mid T(x)\neq y\})=0$. Then, $\gamma=\gamma_T$.*

Proof. By definition of γ_T one has $\gamma_T(A \times B) = \mu(T^{-1}(B) \cap A)$ for all Borel sets $A \subseteq X$ and $B \subseteq Y$. On the other hand,

$$\begin{aligned}\gamma(A \times B) &= \gamma(\{(x,y) \mid x \in A, \text{ and } y \in B\}) \\ &= \gamma(\{(x,y) \mid x \in A, y \in B \text{ and } y = T(x)\}) \\ &= \gamma(\{(x,y) \mid x \in A \cap T^{-1}(B), y = T(x)\} \\ &= \mu(A \cap T^{-1}(B)),\end{aligned}$$

thus proving the claim. □

2 Existence of solutions to Kantorovich's problem

The proof of existence relies on the direct method in the calculus of variations, i.e. the fact that the minimized functional is lower semi-continuous and the set over which it is minimized is compact.

Theorem 2.1. *Let X, Y be two compact spaces, and $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous cost function, which is bounded from below. Then Kantorovich's problem admits a minimizer.*

Lemma 2.2. *Let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous function, which is also bounded from below. Define $\mathcal{F} : \mathcal{P}(X) \to \mathbb{R} \cup \{+\infty\}$ through $\mathcal{F}(\mu) = \int_X f \mathrm{d}\mu$. Then, $\mathcal{F}$ is lower-semicontinuous for the narrow convergence, i.e.*

$$\forall \mu_n \rightharpoonup \mu, \ \liminf_{n\to\infty} \mathcal{F}(\mu_n) \geqslant \mathcal{F}(\mu).$$

Proof. **Step 1.** We show that there exists a family of bounded and continuous functions f^k such that $k \mapsto f^k$ is pointwise increasing and $f = \sup_k f^k$. We assume that there exists x_0 such that $f(x_0) < +\infty$ (if not, there is nothing to prove). Define $g^k(x) = \inf_{y\in X} f(y) + k\mathrm{d}(x,y) \leqslant f(x_0) + k\mathrm{d}(x, x_0)$. The function g^k is k-Lipschitz as a minimum of k-Lipschitz functions, and one obviously has $g^k \leqslant g^\ell \leqslant f$ for $k \leqslant \ell$. Let us prove that $\sup_k g^k(x) = f(x)$ for any $x \in X$. Given x, and for every k, there exists a point x_k such that

$$f(x_k) + k\mathrm{d}(x, x_k) \leqslant g^k(x) + 1/k \leqslant f(x) + 1/k. \tag{2.3}$$

Using that $f \geqslant M > -\infty$ we get

$$\mathrm{d}(x, x_k) \leqslant \frac{1}{k}(f(x) + 1/k - f(x_k)) \leqslant \frac{1}{k}(f(x) + 1/k - M),$$

so that $x_k \to x$. Then, taking the limit in (2.3) and using the lower semicontinuity of f leads to $f(x) \leqslant \liminf_{k\to\infty} f(x_k) \leqslant \sup_{k\to\infty} g^k(x)$. Finally, set $f^k(x) = \min(g^k(x), k)$. Then f^k is k-Lipschitz, bounded by k and one has $\sup_k f^k = f$.

Step 2. Let $\mathcal{F}^k(\mu) = \int f^k \mathrm{d}\mu$. Since f^k is continuous and bounded, the linear form $\mathcal{F}^k$ is narrowly continuous. Thus, $\mathcal{F} = \sup_k \mathcal{F}^k$ is lower semi-continuous as a maximum of lower semi-continuous functions. □

Proof of Theorem 2.1. Define $\mathcal{F}(\gamma) := \int c \mathrm{d}\gamma$, then by Lemma 2.2 $\mathcal{F}$ is l.s.c. for the narrow convergence. We just need to show that the set $\Pi(\mu, \nu)$ is compact for narrow topology. Take a sequence $\gamma_n \in \Pi(\mu, \nu)$, since they are probability measures then they are bounded in the dual of $\mathcal{C}(X \times Y)$. Moreover, usual weak-$\star$ compactness of $\mathcal{P}(X \times Y)$ guarantees the existence of a converging subsequence $\gamma_{n_k} \rightharpoonup \gamma \in \mathcal{P}(X \times Y)$. We need to check that

$\gamma \in \Pi(\mu,\nu)$. Fix $\varphi \in \mathcal{C}(X)$, then $\int \varphi(x)\mathrm{d}\gamma_{n_k} = \int \varphi \mathrm{d}\mu$ and by passing to the limit we have $\int \varphi(x)\mathrm{d}\gamma = \int \varphi \mathrm{d}\mu$. This shows that $\pi_{X\#}\gamma = \mu$. The same may be done for π_Y which concludes the proof.

□

3 Kantorovich as a relaxation of Monge

The question that we consider here is the equality between the infimum in Monge problem and the minimum in Kantorovich problem. This part is taken from Santambrogio [5].

Theorem 3.1. *Let $X = Y$ be a compact subset of $\mathbb{R}^d$, $c \in \mathcal{C}(X \times Y)$ and $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$. Assume that μ is atomless. Then,*

$$\inf (\mathrm{MP}) = \min (\mathrm{KP}).$$

This theorem was first proved on $\mathbb{R}$ by Gangbo [3]. The proof presented here is taken from Santambrogio's book [5]. The next two counter examples are due to an article of Pratelli [4], where he also proves an extension of this theorem.

Example 3.2. Take the same measures on $\mathbb{R}^2$ as in example 1.6, but take the discontinuous (but lsc) cost $c(x,y) = 1$ if $\|x-y\| \leqslant 1$ and 2 if not. Then, the value of the infimum in Monge's problem is 2, while the minimum in Kantorovich's problem is 1.

Proof. Take any transport map T between μ and ν. It suffices to show that $\mu(\{x \mid \|T(x)-x\| = 1\}) = 0$, or equivalently that $\mu(E_\pm) = 0$ where $E_\pm = \{x \mid T(x) = x \pm (1,0)\}$. But, by definition of the measures, $\nu(T(E_+)) = 2\mu(E_+)$, which contradicts the property $T_\#\mu = \nu$ unless $\mu(E_+) = 0$. □

Example 3.3. Consider $\mu_i = \frac{1}{2}(\delta_{x_i} + \alpha\, \lambda|_{B(y_i,1)})$ with $\alpha = \frac{1}{\lambda(B(y_i,1))}$ on $\mathbb{R}^2$ with $c(x,y) = \|x-y\|$. Then, any transport map must transport the Dirac to the Dirac and the ball to the ball, so that its cost is $\|x_1-x_2\| + \|y_1-y_2\|$. On the other hand, a transport plan can transport δ_{x_1} to $\alpha\, \lambda|_{B(y_2,1)}$ with cost $\leqslant \|x_1-y_2\|+1$. The total cost of this transport plan is $2 + \|x_1-y_2\| + \|x_2-y_1\|$, which can be (much) lower that $\|x_1-x_2\| + \|y_1-y_2\|$ for suitable positions for these points.

We quote the following lemma without proof, see Corollary 1.28 in [5].

Lemma 3.4. *If $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and μ has no atoms, then $\exists T : \mathbb{R}^d \to \mathbb{R}^d$ measurable such that $T_\#\mu = \nu$.*

Lemma 3.5. *Let K be a compact metric space. For any $\varepsilon > 0$ there exists a (measurable) partition $K_1, \ldots, K_N$ of K such that for every i, $\mathrm{diam}(K_i) \leqslant \varepsilon$.*

Proof. By compactness, there exists N points $x_1, \ldots, x_N$ such that $K \subseteq \bigcup_i \mathrm{B}(x_i, \varepsilon)$. The partition $K_1, \ldots, K_N$ of K defined recursively by $K_i = \{x \in K \setminus K_1 \cup \ldots K_{i-1} \mid \forall j, d(x,x_i) \leqslant d(x,x_j)\}$ satisfies $K_i \subseteq \mathrm{B}(x_i, \varepsilon)$. □

Proof of Theorem 3.1. Using the continuity of the functional $\gamma \mapsto \int c\mathrm{d}\gamma$ (which uses the continuity of the cost), the statement will follow if we are able to prove that any transport plan $\gamma \in \Pi(\mu,\nu)$, there exists a sequence of transport maps $T^N : \mathbb{R}^d \to \mathbb{R}^d$ such that $T^N_\#\mu = \nu$ and γ_{T^N} narrowly converges to γ.

By Lemma 3.5, for any $\varepsilon > 0$ there exist a measurable partition $K_1, \ldots, K_N$ of X such that $\mathrm{diam}(K_i) \leqslant \varepsilon$. Define $\gamma_i := \gamma|_{K_i \times \mathbb{R}^d}$. Now, let $\mu_i := \pi_{X\#}\gamma_i$ and $\nu_i := \pi_{Y\#}\gamma_i$.

Since $\mu_i \leqslant \mu$, the measure μ_i has no atoms, so that by the previous Lemma, there exists a transport plan $S_i : K_i \to \mathbb{R}^d$ with $S_{i\#}\mu_i = \nu_i$. Then by gluing the transports S_i we get a transport T^N sending μ onto ν (here we use that $\mu = \sum_i \mu_i$ and $\nu = \sum_i \nu_i$) as the measures μ_i are concentrated on disjoint sets.

Since $\gamma_{S_i}, \gamma_i \in \mathcal{P}(K_i \times Y)$ both have marginals μ_i and ν_i, one has

$$\gamma_{S_i}(K_i \times K_j) = \nu_i(K_j) = \gamma_i(K_i \times K_j).$$

To prove narrow convergence, we consider a test function $\varphi \in \mathcal{C}_b(X \times Y)$. By compactness of $X \times Y$, this function has a uniform continuity modulus ω_φ with respect to the Euclidean norm on $\mathbb{R}^d \times \mathbb{R}^d$. Moreover,

$$\begin{aligned}\int_{X\times Y} \varphi \mathrm{d}(\gamma - \gamma_{T^N}) &= \sum_{ij} \int_{K_i\times K_j} \varphi \mathrm{d}(\gamma_i - \gamma_{S_i}) \\ &\leqslant \sum_{ij} \gamma_i(K_i \times K_j) \max_{K_i\times K_j} \varphi - \gamma_{S_i}(K_i \times K_j) \min_{K_i\times K_j} \varphi \\ &\leqslant \sum_{ij} \gamma_i(K_i \times K_j)\omega_\varphi(\mathrm{diam}(K_i \times K_j)) = \mathrm{O}(\omega_\varphi(2\varepsilon)).\end{aligned}$$

Since this holds for any function φ, one sees that γ_{T^N} converges to γ narrowly. In particular, if γ is the minimizer in Kantorovich's problem, then γ_{T^N} is a minimizing sequence. Then, $T^N_\# \mu = \nu$ and

$$\lim_{N\to\infty} \int_{X\times Y} c(x, T^N(x)) \mathrm{d}\mu(x) = \int_{X\times Y} c \mathrm{d}\gamma,$$

thus proving the statement. □

4 The dual problem

We now focus on duality theory. We firstly find a formal dual problem by exchanging $\inf - \sup$. Let write down the constraint $\gamma \in \Pi(\mu, \nu)$ as follows: if $\gamma \in \mathcal{M}^+(X \times Y)$ (we remind that X, Y are compact spaces) we have

$$\sup_{\varphi,\psi} \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu - \int_{X\times Y} (\varphi(x) + \psi(y)) \mathrm{d}\gamma = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu, \nu), \\ +\infty & \text{otherwise,} \end{cases}$$

where the supremum is taken on $\mathcal{C}_b(X) \times \mathcal{C}_b(Y)$. Thus we can now remove the constraint on γ in (KP)

$$\inf_{\gamma\in\mathcal{M}^+(X\times Y)} \int_{X\times Y} c \mathrm{d}\gamma + \sup_{\varphi,\psi} \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu - \int_{X\times Y} (\varphi(x) + \psi(y)) \mathrm{d}\gamma$$

and by interchanging sup and inf we get

$$\sup_{\varphi,\psi} \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu + \inf_{\gamma\in\mathcal{M}^+(X\times Y)} \int_{X\times Y} (c(x, y) - \varphi(x) - \psi(y)) \mathrm{d}\gamma.$$

One can now rewrite the inf in γ as constraint on φ and ψ as

$$\inf_{\gamma\in\mathcal{M}^+(X\times Y)} \int_{X\times Y} (c - \varphi \oplus \psi) \mathrm{d}\gamma = \begin{cases} 0 & \text{if } \varphi \oplus \psi \leqslant c \text{ on } X \times Y \\ -\infty & \text{otherwise} \end{cases},$$

where $\varphi \oplus \psi(x, y) := \varphi(x) + \psi(y)$.

Definition 4.1 (Dual problem)**.** Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a *cost function* $c \in \mathcal{C}(X \times Y)$. The *dual problem* is the following optimization problem

$$(\mathrm{DP}) := \sup \left\{ \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu \mid \varphi \in \mathcal{C}_b(X),\ \psi \in \mathcal{C}_b(Y),\ \varphi \oplus \psi \leqslant c \right\} \tag{4.4}$$

Remark 4.2. One trivially has the weak duality inequality (KP) $\geqslant$ (DP). Indeed, denoting

$$L(\gamma, \varphi, \psi) = \int_{X \times Y} (c - \varphi \oplus \psi) \mathrm{d}\gamma) + \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu,$$

one has for any $(\varphi, \psi, \gamma) \in \mathcal{C}_b(X) \times \mathcal{C}_b(Y) \times \mathcal{M}^+(X \times Y)$,

$$\inf_{\tilde{\gamma} \geqslant 0} L(\tilde{\gamma}, \varphi, \psi) \leqslant L(\gamma, \varphi, \psi) \leqslant \sup_{\tilde{\varphi}, \tilde{\psi}} L(\gamma, \tilde{\varphi}, \tilde{\psi})$$

Taking the supremum with respect to (φ, ψ) on the left and the infimum with respect to γ on the right gives $\inf(\mathrm{KP}) \geqslant \sup(\mathrm{DP})$. When $\sup(\mathrm{DP}) = \inf(\mathrm{KP})$, one talks of *strong duality*. Note that this is independent of whether the infimum and the supremum are attained.

Remark 4.3. As often, the Lagrange multipliers (or Kantorovich potentials) φ, ψ have an economic interpretation as prices. For instance, imagine that μ is the distribution of sand available at quarries, and ν describes the amount of sand required by construction work. Then, (KP) can be interpreted as finding the cheapest way of transporting the sand from μ to ν for a construction company. Imagine that this company wants to externalize the transport, by paying a loading coast $\varphi(x)$ at a point x (in a quarry) and an unloading coast $\psi(y)$ at a point y (at a construction place). Then, the constraint $\varphi(x) + \psi(y) \leqslant c(x, y)$ translates the fact that the construction company would not externalize if its cost is higher than the cost of transporting the sand by itself. Then, Kantorovich's dual problem (DP) describes the problem of a transporting company: maximizing its revenue $\int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu$ under the constraint $\varphi \oplus \psi \leqslant c$ imposed by the construction company. The economic interpretation of the strong duality (KP) = (DP) is that in this setting, externalization has exactly the same cost as doing the transport by oneself.

We now focus on the existence of a pair (ψ, ψ) which solves (DP) and postpone the proof of the strong duality to the next lecture.

Definition 4.4 (c-transform and $\bar{c}$-transform)**.** Given a function $f : x \to \overline{\mathbb{R}}$, we define its c-transform $f^c : Y \to \overline{\mathbb{R}}$ by

$$f^c(y) = \inf_{x \in X} c(x, y) - f(x).$$

We also define the $\bar{c}$-transform of $g : Y \to \overline{\mathbb{R}}$ by

$$g^{\bar{c}}(x) = \inf_{y \in Y} c(x, y) - g(y).$$

We also say that a function ψ on Y is $\bar{c}$-concave if there exists f such that $\psi = f^c$. Notice now that if c is continuous on a compact set, and hence uniformly continuous, then there exists am increasing function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ with $\omega(0) = 0$ such that

$$|c(x, y) - c(x', y')| \leqslant \omega(d_X(x, x') + d_Y(y, y')).$$

If we consider f^c we have that $f^c(y) = \inf_x \tilde{f}_x(y)$ with $\tilde{f}_x(y) = c(x,y) - f(x)$, and the functions $\tilde{f}_x$ satisfy $|\tilde{f}_x(y) - \tilde{f}_x(y)| \leqslant \omega(d_Y(y,y'))$. This implies that f^c actually share the same continuity modulus of c. It snow quite easy to see that given an admissible pair (ψ,ψ) in (DP), one cas always replace it with (φ,φ^c) and then $(\varphi^{c\bar{c}},\varphi^c)$ and the constrained are preserved and the integrals increased. The underlying idea of these transformations is actually to improve a maximizing sequence to get a uniform bound on its continuity.

Theorem 4.5. *Suppose that X and Y are compact and $c \in \mathcal{C}(X \times Y)$. Then there exists a pair $(\varphi^{c\bar{c}},\varphi^c)$ which solves* (DP).

Proof. Let us first denote by $\mathcal{J}(\varphi,\psi)$ the following functional

$$J(\varphi,\psi) = \int_X \varphi, \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu,$$

then it is clear that for every constant λ we have $\mathcal{J}(\varphi-\lambda,\psi+\lambda) = \mathcal{J}(\varphi,\psi)$. Given now a maximising sequence (φ_n,ψ_n) we can improve it by means of the c- and $\bar{c}$-transform obtaining a new one $(\varphi_n^{c\bar{c}},\varphi_n^c)$. Notice that by the consideration above the sequences $\varphi_n^{c\bar{c}}$ and φ_n^c are uniformly equicontinuous. Since φ_n^c is continuous on a compact set we can always subtract its minimum and assume that $\min_Y \varphi_n^c = 0$. This implies that the sequence φ_n^c is also equibounded as $0 \leqslant \varphi_n^c \leqslant \omega(\mathrm{diam}(Y))$. We also deduce uniform bounds on $\varphi_n^{c\bar{c}}$ as $\varphi_n^{c\bar{c}} = \inf_Y c(x,y) - \varphi^c(y)$. This let us apply Ascoli-Arzela's theorem and extract two uniformly converging subsequences $\varphi_{n_k}^{c\bar{c}} \to \overline{\varphi}$ and $\varphi_{n_k}^c \to \overline{\psi}$ where the pair $(\overline{\varphi},\overline{\psi})$ satisfies the inequality constraint. Moreover, since $(\varphi_n^{c\bar{c}},\varphi_n^c)$ is a maximising sequence we get that the pair $(\overline{\varphi},\overline{\psi})$ is optimal. now one can apply again the c- and $\bar{c}$-transforms obtaining an optimal pair of the form $(\overline{\varphi}^{c\bar{c}},\overline{\varphi}^c)$. □

References

[1] Najma Ahmad, Hwa Kil Kim, and Robert J McCann, *Extremal doubly stochastic measures and optimal transportation*, arXiv preprint arXiv:1004.4147 (2010).

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

[3] Wilfrid Gangbo, *The monge mass transfer problem and its applications*, Contemporary Mathematics **226** (1999), 79–104.

[4] Aldo Pratelli, *On the equality between monge's infimum and kantorovich's minimum in optimal mass transportation*, Annales de l'Institut Henri Poincare (B) Probability and Statistics **43** (2007), no. 1, 1–13.

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

[6] Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.

[7] ———, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

Lecture 2: Optimality Conditions and Consequences

Lénaïc Chizat

February 26, 2020

The material of today's lecture comes from [2, 3] and – for the most part of it – the lecture notes of Q. Mérigot.

Announcements. Register on the course webpage[1]. Bring your own laptops next week, with a running version of Python 3 and Jupyter notebooks.

1 Introduction

Let X and Y be compact metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \to \mathbb{R}$ a continuous cost function. In Lecture 1, we have defined the Kantorovich problem

$$\mathcal{T}_c(\mu,\nu) := \inf_{\gamma} \Big\{ \int_{X\times Y} c(x,y) \mathrm{d}\gamma(x,y) \mid \gamma \in \Pi(\mu,\nu) \Big\}. \tag{KP}$$

where $\Pi(\mu,\nu) := \{\gamma \in \mathcal{M}_+(X\times Y) \mid (\pi_X)_\#\gamma = \mu \text{ and } (\pi_Y)_\#\gamma = \nu\}$ is the set of *transport plans* between μ and ν. Rewriting the marginal constraints leads to the problem

$$\inf_{\gamma \geqslant 0} \sup_{\varphi,\psi} \Big\{ \int_X \varphi(x) \mathrm{d}\mu(x) + \int_Y \psi(y) \mathrm{d}\nu(y) + \int_{X\times Y} \Big(c(x,y) - \varphi(x) - \psi(y) \Big) \mathrm{d}\gamma(x,y) \Big\}.$$

After *formally* inverting the inf-sup, and minimizing over γ, we get the *dual* problem

$$\mathcal{T}_c^{dual}(\mu,\nu) := \sup_{\varphi,\psi} \Big\{ \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu \mid \varphi(x) + \psi(y) \leqslant c(x,y),\ \forall (x,y) \in X \times Y \Big\}. \tag{DP}$$

Let us recall some results from Lecture 1:

- There exists minimizers to (KP) in $\mathcal{P}(X \times Y)$.
- There exists maximizers to (DP) in $\mathcal{C}(X) \times \mathcal{C}(Y)$.
- It holds $\mathcal{T}_c^{dual}(\mu,\nu) \leqslant \mathcal{T}_c(\mu,\nu)$.
- We also recall the definition of c-transforms for $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$:

$$\varphi^c(y) = \inf_{x\in X} c(x,y) - \varphi(x) \qquad\qquad \psi^{\bar{c}}(x) = \inf_{y\in Y} c(x,y) - \psi(y).$$

  It always holds $\varphi^{c\bar{c}} \geqslant \varphi$. If $\varphi(x) = \psi^{\bar{c}}(y)$ for some ψ, then φ is said *c-concave* and it holds $\varphi^{c\bar{c}} = \varphi$ (exercise, or see [2, Prop. 1.3.4]).

Today, we will show *strong duality*, derive primal-dual optimality conditions and explore their consequences. We assume that X and Y are compact for the sake of simplicity, but most statement have their counterpart in non-compact spaces.

[1] http://lchizat.github.io/ot2020orsay.html

2 Strong duality

2.1 The case of discrete optimal transport

We start with the case of finite discrete probability measures, which is important because:

- It often comes up in applications (e.g. optimal matching in economy).
- Numerical methods for the continuous case often resort to discretization.
- It is a convenient way to study the general case, through density arguments.

Proposition 2.1 (Duality, discrete case)**.** *If μ and ν are finitely supported, then $\mathcal{T}_c^{dual}(\mu,\nu) = \mathcal{T}_c(\mu,\nu)$.*

Proof. Let us write $\mu = \sum_{i=1}^m \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$ where all μ_i and ν_j are strictly positive. Consider the linear program

$$\mathcal{T}_c^{lp}(\mu,\nu) := \min\Big\{ \sum_{i,j} c(x_i,y_j)\gamma_{i,j} \mid \gamma_{i,j} \geqslant 0,\ \sum_j \gamma_{i,j} = \mu_i,\ \sum_i \gamma_{i,j} = \nu_j \Big\}.$$

which admits a solution that we denote γ. By linear programming duality (which is standard in the finite dimensional case, see e.g. [1]), we have strong duality

$$\mathcal{T}_c^{lp}(\mu,\nu) = \max\Big\{ \sum_i \varphi_i\mu_i + \sum_j \psi_j\nu_j \mid \varphi_i + \psi_j \leqslant c(x_i,y_j) \Big\}$$

and at optimality $\gamma_{i,j}(c_{i,j} - \varphi_i - \psi_j) = 0$ (the complementary slackness in Karush-Kuhn-Tucker theorem). Let us now build a c-concave function φ such that $\varphi(x) \oplus \varphi^c(y) = c(x,y)$ on the set $\{(x_i,y_j) \mid \gamma_{i,j} > 0\}$. For this purpose, we introduce

$$\psi(y) = \begin{cases} \psi_i & \text{if } y = y_i, \\ +\infty & \text{otherwise,} \end{cases}$$

and let $\varphi = \psi^{\overline{c}}$. For $i_0 \in [n]$, there exists $j_0 \in [n]$ such that $\gamma_{i_0,j_0} > 0$ and thus, by complementary slackness, $\varphi_{i_0} + \psi_{j_0} = c(x_{i_0},y_{j_0})$ and thus

$$\varphi(x_{i_0}) = \inf_{y\in Y} \Big(c(x_{i_0},y) - \psi(y) \Big) = \min_{j\in[n]} \Big(c(x_{i_0},y_j) - \psi_j \Big) = c(x_{i_0},y_{j_0}) - \psi_{j_0} = \varphi_{i_0}.$$

Similarly, one can show that $\varphi^c(y_j) = \psi_j$ for all $j \in [n]$. Finally, we define $\gamma = \sum_{i,j} \gamma_{i,j}\delta_{(x_i,y_j)} \in \Pi(\mu,\nu)$. We conclude with Lemma 2.2. □

Lemma 2.2 (Duality criterion)**.** *Let $\gamma \in \Pi(\mu,\nu)$ and (φ,ψ) satisfying $\varphi(x) + \psi(y) \leqslant c(x,y)$. It $\varphi(x) + \psi(y) = c(x,y)$ for γ-almost every (x,y) then $\mathcal{T}_c^{dual}(\mu,\nu) = \mathcal{T}_c(\mu,\nu)$ and γ and (φ,ψ) are optimal for the primal and dual problem respectively.*

Proof. Observe that

$$\mathcal{T}_c(\mu,\nu) \leqslant \int c\mathrm{d}\gamma = \int \Big(\varphi(x) + \psi(y)\Big)\mathrm{d}\gamma(x,y) = \int \varphi\mathrm{d}\mu + \int \psi\mathrm{d}\nu \leqslant \mathcal{T}_c^{dual}(\mu,\nu)$$

Since we know that $\mathcal{T}_c^{dual}(\mu,\nu) \leqslant \mathcal{T}_c(\mu,\nu)$ this is sufficient to conclude. □

2.2 Density of discrete measures

In order the prove the general case, we will use the density of discrete measures for the weak topology and a stability property of optimal dual and primal solutions.

Lemma 2.3 (Density of discrete measures)**.** *Let X be a compact space and $\mu \in \mathcal{P}(X)$. Then, there exists a sequence of finitely supported probability measures weakly converging to μ.*

Proof. By compactness, for any $\epsilon > 0$, there exists N points $x_1, \dots, x_n$ such that $X \subset \bigcup_i B(x_i, \epsilon)$. We introduce the partition $K_1, \dots, K_n$ of X defined recursively by $K_i = B(x_i, \epsilon) \setminus K_1 \cup \dots K_{i-1}$ and

$$\mu_\epsilon := \sum_{i=1}^{n} \mu(K_i)\delta_{x_i}.$$

To prove weak convergence of μ_ϵ to μ as $\epsilon \to 0$, take $\varphi \in \mathcal{C}(X)$. By compactness of X, φ admits a modulus of continuity ω, i.e. an increasing function satisfying $\lim_{t\to 0} \omega(t) = 0$ and $|\varphi(x) - \varphi(y)| \leqslant w(\mathrm{dist}(x, y))$. Using that $\mathrm{diam}(K_i) \leqslant \epsilon$, we get

$$\Big| \int \varphi \mathrm{d}\mu - \int \varphi \mathrm{d}\mu_\epsilon \Big| \leqslant \sum_{i=1}^{n} \int_{K_i} |\varphi(x) - \varphi(x_i)| \mathrm{d}\mu(x) \leqslant \omega(\epsilon).$$

We deduce that μ_ϵ weakly converges to μ (remember that for measures on a compact space, tight, weak and weak* topologies are the same). □

Note that we even have weak density in $\mathcal{P}(X)$ of empirical measures, that is measures of the form $\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ for $n \in \mathbb{N}^*$ and $x_i \in X$. Indeed, take $x_1, \dots, x_n$ independent random variables with distribution μ. Then, by the uniform law of large numbers (a.k.a. Varadarajan's theorem) states that $\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ weakly converges to μ with probability 1.

2.3 Strong duality for the general case

Theorem 2.4 (Duality, general case)**.** *Let X, Y be compact metric spaces and $c \in \mathcal{C}(X \times Y)$. Then $\mathcal{T}_c(\mu, \nu) = \mathcal{T}_c^{dual}(\mu, \nu)$.*

Proof. By Lemma 2.3, there exists a sequence $\mu_k \in \mathcal{P}(X)$ (resp. $\nu_k \in \mathcal{P}(Y)$) of finitely supported measures which converge weakly to μ (resp. ν). By Proposition 2.1 and its proof, there exists for all k, γ_k and (φ_k, φ_k^c) with φ_k c-concave which are optimal primal-dual solutions to $\mathcal{T}_c(\mu_k, \nu_k)$ and such that γ_k is supported on the set

$$S_k := \{(x, y) \in X \times Y \mid \varphi_k(x) + \varphi_k^c(y) = c(x, y)\}.$$

Adding a constant if necessary, we can also assume that $\varphi_k(x_0) = 0$ for some point $x_0 \in X$. As in the previous lecture, we see that $\{\varphi_k\}$ and $\{\varphi_k^c\}$ are uniformly continuous and bounded so by Ascoli-Arzelà theorem converge uniformly to some (φ, ψ) up to a subsequence. We easily have that φ is c-concave and $\psi = \varphi^c$.

By weak compactness of $\mathcal{P}(X \times Y)$, we can assume that the sequence γ_k weakly converges to $\gamma \in \Pi(\mu, \nu)$. Moreover, by Lemma 2.5, every pair $(x, y) \in \mathrm{spt}(\gamma)$ can be approximated by a sequence of pairs $(x_k, y_k) \in \mathrm{spt}(\gamma_k)$ with $\lim_{k\to\infty}(x_k, y_k) = (x, y)$. Since γ_k is supported on S_k one has $c(x_k, y_k) = \varphi_k(x_k) + \varphi_k^c(y_k)$, which gives at the limit $c(x, y) = \varphi(x) + \varphi^c(y)$. We conclude with Lemma 2.2. □

Lemma 2.5. *If μ_n converges weakly to μ, then for any point $x \in \mathrm{spt}(\mu)$ there exists a sequence $x_n \in \mathrm{spt}(\mu_n)$ converging to x.*

Proof. Consider $x \in \mathrm{spt}(\mu)$. For any $k \in \mathbb{N}$, consider the function $\varphi_k(z) = \max\{0, 1 - k\mathrm{dist}(x,z)\}$ which is continuous. Then

$$\lim_{n\to\infty} \int \varphi_k \mathrm{d}\mu_n = \varphi_k \mathrm{d}\mu > 0.$$

Thus, there exists n_k such that for any $n \geqslant n_k$, $\int \varphi_k \mathrm{d}\mu_n > 0$. This implies the existence of a sequence $(x_n^{(k)}) \in X$ such that $x_n^{(k)} \in \mathrm{spt}(\mu_n)$ and $\mathrm{dist}(x_n^{(k)}, x) \leqslant 1/k$ for $n \geqslant n_k$. By a diagonal argument, we build the sequence $x_n = x_n^{k_n}$ where $k_n = \max\{k \mid k = 0 \text{ or } n \geqslant n_k\}$. Since by construction $k_n \to \infty$, we have $x_n \to x$. □

3 Optimality conditions and stability

Let us write down three important properties that follow from our previous results. First, remark that the proof of Theorem 2.4 can be used to prove the following stability property (the modifications are left as an exercise).

Proposition 3.1 (Stability)**.** *Let X, Y be compact metric spaces. Consider $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ converging weakly to μ and ν respectively and $(c_k)_{k\in\mathbb{N}}$ in $\mathcal{C}(X \times Y)$ converging uniformly to c.*

- *If γ_k is a minimizer for $\mathcal{T}_{c_k}(\mu_k, \nu_k)$ then, up to subsequences, (γ_k) converges weakly to a minimizer for $\mathcal{T}_c(\mu, \nu)$.*
- *Let $(\varphi_k, \varphi_k^{c_k})$ be a maximizer for $\mathcal{T}_{c_k}^{dual}(\mu_k, \nu_k)$ and be such that φ_k is c_k-concave and $\varphi_k(x_0) = 0$. Then, up to subsequences, $(\varphi_k, \varphi_k^{c_k})$ converges uniformly to (φ, φ^c) a maximizer for $\mathcal{T}_c^{dual}(\mu, \nu)$ with φ c-concave satisfying $\varphi(x_0) = 0$.*

Let us emphasize on the optimality conditions, which are just a continuous version of complementary slackness.

Proposition 3.2 (Optimality conditions)**.** *For $\gamma \in \Pi(\mu, \nu)$ and $(\varphi, \psi) \in \mathcal{C}(X) \times \mathcal{C}(Y)$ satisfying $\varphi \oplus \psi \leqslant c$, the following are equivalent:*

(i) $\varphi(x) + \psi(y) = c(x, y)$ holds γ-almost everywhere.

(ii) γ is a minimizer of (KP)*, (φ, ψ) is a maximizer of* (DP)*.*

Proof. The proof of $(i) \Rightarrow (ii)$, is given by Lemma 2.2. To show $(ii) \Rightarrow (i)$, notice that Theorem 2.4 and (ii) imply

$$0 = \int c(x,y)\mathrm{d}\gamma(x,y) - \int \varphi(x) + \psi(y)\mathrm{d}\gamma(x,y) = \int \Big(c(x,y) - \varphi(x) - \psi(y)\Big)\mathrm{d}\gamma(x,y).$$

Since the last integrand is nonnegative, it must vanish γ-almost everywhere. □

Another useful notion attached to optimal transport solutions is that of cyclical monotonicity.

Definition 3.3 (Cyclical monotonicity)**.** A set $S \subset X \times Y$ is said c-cyclically monotone if for any $n \in \mathbb{N}^*$ and $(x_i, y_i)_{i=1}^n \in S^n$, it holds

$$\sum_{i=1}^n c(x_i, y_i) \leqslant \sum_{i=1}^n c(x_i, y_{i+1})$$

with the convention $y_{n+1} = y_1$.

Proposition 3.4. *Let X, Y be compact metric spaces, $c \in \mathcal{C}(X \times Y)$ and $\gamma \in \Pi(\mu, \nu)$ an optimal transport plan between $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then* $\mathrm{spt}(\gamma)$ *is c-cyclically monotone.*

This result is rather direct in the discrete case and can also be proved without duality in the general case but our duality results lead to a straighforward proof.

Proof. Let $(x_i, y_i)_{i=1}^n$ be n points in $\mathrm{spt}(\gamma)$. By Proposition 3.2, we know that there exists (φ, ψ) such that $\varphi(x_i) + \psi(y_j) \leqslant c(x_i, y_j)$ for all i, j and such that $\varphi(x_i) + \psi(y_i) = c(x_i, y_i)$ for all i. Thus

$$\sum c(x_i, y_{i+1}) - \sum c(x_i, y_i) \geqslant \sum_i \Big(\varphi(x_i) + \psi(y_{i+1})\Big) - \sum_i \Big(\varphi(x_i) + \psi(y_i)\Big) = 0.$$

□

Remark 3.5. A stronger property holds: any c-cyclically monotonous set is contained in a set of the form $\{(x, y) \in X \times Y \,;\, \varphi(x) + \varphi^c(y) = c(x, y)\}$ for some c-concave function φ. This implies that any $\gamma \in \Pi(\mu, \nu)$ such that $\mathrm{spt}(\gamma)$ is c-cyclically monotone is optimal.

4 Applications

Let us exploit the optimality conditions and duality results to describe the behavior of optimal transport in specific situations.

4.1 Optimal transport on the real line

Theorem 4.1 (Optimality of the monotone transport plan)**.** *Let μ, ν be two probability measures on $\mathbb{R}$, and $c(x, y) := h(x - y)$ where h is strictly convex. Then, there exists a unique $\gamma \in \Gamma(\mu, \nu)$ satisfying the two following statements, which are equivalent*

(i) γ is optimal for the Kantorovich problem;

(ii) $\mathrm{spt}(\gamma)$ *is* monotone *in the sense*

$$\forall (x, y), (x', y') \in \mathrm{spt}(\gamma), (x' - x) \cdot (y' - y) \geqslant 0.$$

Proof. We first prove that there exists at most one transport plan satisfying (ii). Recall that a probability measure on $\mathbb{R}^2$ is uniquely defined from the values $\gamma((-\infty, a] \times (-\infty, b])$ for any $a, b \in \mathbb{R}$. This follows from the fact that such sets generate the Borel σ-algebra. Consider $A = (-\infty, a] \times (b, +\infty)$ and $B = (a, +\infty) \times (-\infty, b]$. Then, by monotonicity of $\mathrm{spt}(\gamma)$ one cannot have $\gamma(A) > 0$ and $\gamma(B) > 0$ at the same time. Hence,

$$\begin{aligned}\gamma(]-\infty, a] \times]-\infty, b]) &= \min(\gamma((]-\infty, a] \times]-\infty, b]) \cup A), \gamma((]-\infty, a] \times]-\infty, b]) \cup B)) \\ &= \min(\mu(]-\infty, a]), \nu(]-\infty, b])).\end{aligned}$$

This shows that $\gamma(]-\infty, a] \times (]-\infty, b])$ is uniquely defined from μ, ν, so that γ is unique.

Now by Proposition 3.4, we know that for an optimal transport plan γ and $(x_i, y_i)_{i=1}^2 \in \mathrm{spt}(\gamma)^2$, it holds

$$c(x_0, y_0) + c(x_1, y_1) \leqslant c(x_0, y_1) + c(x_1, y_0).$$

We conclude with $c(x, y) = |x - y|^2$, the case of a general strictly convex function can be found in Chapter 2 of [2]. Expanding the squares and simplifying, the above inequality can be rewritten as

$$-x_0 y_0 - x_1 y_1 \leqslant -x_0 y_1 - x_1 y_0,$$

giving exactly $(x_0 - x_1)(y_0 - y_1) \geqslant 0$ as desired. □

While in this proof cyclical monotonicity of order 2 was enough to conclude, we warn the reader that this is in general not the case in higher dimension.

Remark 4.2 (Book-shifting)**.** If $c(x,y) = |x-y|$ with the Euclidean norm, the solution to the optimal transport problem might be non-unique. Take for instance $\mu = \lambda|_{[0,1]}$ and $\nu = \lambda|_{[\varepsilon,1+\varepsilon]}$ for some $\varepsilon > 0$. Then, the maps $T : x \mapsto x+\varepsilon$ and $T'(x) = x$ if $x \in [\varepsilon, 1]$ and $T'(x) = x+1$ if $x \in [0,\varepsilon]$ are both optimal with the same cost. (NB: proving the optimality of a transport map is in general a difficult matter, to which Kantorovich duality provides an answer.)

It turns out that the unique monotone transport map can be built using *quantile* functions. Given $\mu \in \mathcal{P}(\mathbb{R})$, define its cumulative distribution function $F_\mu : \mathbb{R} \to [0,1]$ and its quantile function $Q_\mu : [0,1] \to \mathbb{R}$ by:

$$F_\mu(x) = \mu((-\infty, x]) \quad \text{and} \quad Q_\mu(t) = \inf\{x \in \mathbb{R} \mid F_\mu(x) \geqslant t\}.$$

As a simple consequence of these definitions, we have

$$Q_\mu(t) \leqslant x \Leftrightarrow F_\mu(x) \geqslant t \quad \text{and} \quad Q_\mu(t) > x \Leftrightarrow F_\mu(x) < t. \tag{4.1}$$

Proposition 4.3 (Characterization of the monotone transport plan)**.** *The unique monotone transport plan in $\Pi(\mu,\nu)$ is given by $\gamma_Q = (Q_\mu, Q_\nu)_\# \lambda$. In particular, for $c(x,y) = h(y-x)$ with h strictly convex, we have the following explicit optimal transport cost*

$$\mathcal{T}_c(\mu,\nu) = \int_0^1 h(Q_\nu(t) - Q_\mu(t))\mathrm{d}t$$

Proof. First, let us prove that Q_μ is a transport map between the Lebesgue measure on $[0,1]$ (denoted λ) and μ. Using Eq. (4.1), we write

$$(Q_\mu)_\# \lambda|_{[0,1]}(]-\infty, a]) = \lambda\Big(\{t \in [0,1] \mid Q_\mu(t) \leqslant a\}\Big) = \lambda\Big(\{t \in [0,1] \mid F_\mu(a) \geqslant t\}\Big) = F_\mu(a),$$

which proves that $(Q_\mu)_\# \lambda|_{[0,1]} = \mu$ using the characterization of a measure through its CFD. It directly follows that $\gamma_Q \in \Pi(\mu,\nu)$. Then, let us compute

$$\begin{aligned}
\gamma_Q(]-\infty, a] \times]-\infty, b]) &= \lambda\Big(\{t \in [0,1] \mid Q_\mu(t) \leqslant x,\ Q_\nu(t) \leqslant x\}\Big) \\
&= \lambda\Big(\{t \in [0,1] \mid F_\mu(x) \geqslant t,\ F_\nu(x) \geqslant t\}\Big) \\
&= \min\{F_\mu(x), F_\nu(x)\}
\end{aligned}$$

and we recover the characterization of the monotone transport plan in the proof of Theorem 4.1. □

4.2 Duality formula for the distance cost

The dual problem takes a particularly simple form when the cost is of the form $c(x,y) = \mathrm{dist}(x,y)$.

Proposition 4.4 (Kantorovich-Rubinstein)**.** *Let (X, dist) be a compact metric space and $\mu, \nu \in \mathcal{P}(X)$. Then*

$$\mathcal{T}_{\mathrm{dist}}(\mu,\nu) = \max_{\varphi : X \to \mathbb{R}} \Big\{ \int \varphi \,\mathrm{d}(\mu - \nu) \mid \varphi \text{ is 1-Lipschitz} \Big\}.$$

Proof. Note that $\psi^{\bar{c}}(x) = \inf_y \mathrm{dist}(x,y) - \psi(y)$ is 1-Lipschitz as a infimum of 1-Lipschitz functions, and the same holds for $\psi^{\bar{c}c}$. Moreover, if ψ is 1-Lipschitz, then $\mathrm{dist}(x,y) - \psi(y) \geqslant -\psi(x)$, so that

$$\psi^{\bar{c}}(x) = \inf_y \mathrm{dist}(x,y) - \psi(y) = -\psi(x).$$

Thus, $\varphi = -\psi$ and any 1-Lipschitz function is c-concave. Thus

$$\mathcal{T}_{\mathrm{dist}}(\mu,\nu) = \sup_{\psi: Y\to\mathbb{R}} \int \psi^{\bar{c}} \mathrm{d}\mu + \int \psi^{\bar{c}c} \mathrm{d}\nu = \sup_{\varphi \text{ 1-Lip}} \int \varphi \mathrm{d}\mu + \int \varphi^c \mathrm{d}\nu = \sup_{\varphi \text{ 1-Lip}} \int \varphi \mathrm{d}(\mu - \nu).$$

□

4.3 Optimal transport map for twisted costs

We recall the following characterization of solutions to Monge's problem from Lecture 1.

Lemma 4.5. *Let $\gamma \in \Pi(\mu,\nu)$ and $T : X \to Y$ measurable be such that $\gamma(\{(x,y) \in X \times Y \mid T(x) \neq y\}) = 0$. Then, $\gamma = \gamma_T := (\mathrm{id}, T)_\# \mu$.*

If γ is a minimizer for (KP) and (φ, φ^c) is a maximizer for (DP), we know that $\varphi \oplus \varphi^c = c$ γ-almost everywhere. To build a solution to Monge's problem, it is therefore sufficient to show that the set $\{\varphi \oplus \varphi^c = c\}$ is contained in the graph of a function. This will be possible for the following class of costs:

Definition 4.6 (Twisted cost)**.** A cost function $c \in \mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}^d)$ is said to satisfy the *twist condition* if

$$\forall x_0 \in \mathbb{R}^d, \quad \text{the map } y \mapsto \nabla_x c(x_0, y) \in \mathbb{R}^d \text{ is injective}$$

where $\nabla_x c(x_0, y)$ denotes the gradient of $x \mapsto c(\cdot, y)$ at $x = x_0$. Given $x, v \in \mathbb{R}^d$, we denote $y_c(x_0, v)$ the unique point such that $\nabla_x c(x_0, y_c(x_0, v)) = v$.

Theorem 4.7. *Let $c \in \mathrm{C}^1(\mathbb{R}^d \times \mathbb{R}^d)$ be a twisted cost, let $X, Y \subset \mathbb{R}^d$ be compact subsets and $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Assume that μ is absolutely continuous with respect to the Lebesgue measure. Then, there exists a c-concave function φ that is differentiable almost everywhere such that $\nu = T_\# \mu$ where $T(x) = y_c(x, \nabla\varphi(x))$. Moreover, the only optimal transport plan between μ and ν is γ_T.*

Proof. Enlarging X if necessary, we may assume that $\mathrm{spt}(\mu)$ is contained in the interior of X. First note that by compactness of $X \times Y$ and since c is $\mathcal{C}^1$, the cost c is Lipschitz continuous on $X \times Y$. Take (φ, φ^c) a maximizing pair for (DP) with φ c-concave. Since $\varphi(x) = \min_{y\in Y} c(x,y) + \varphi^c(y)$ we see that φ is Lipschitz. By Rademacher theorem, φ is thus differentiable Lebesgue almost everywhere and, since μ is assumed absolutely continuous, it is differentiable on a set $B \subset \mathrm{spt}(\mu)$ with $\mu(B) = 1$.

Consider an optimal transport plan $\gamma \in \Pi(\mu,\nu)$. For every pair of points $(x_0, y_0) \in \mathrm{spt}(\gamma) \cap (B \times Y)$, we have

$$\varphi^c(y_0) \leqslant c(x, y_0) - \varphi(x), \ \forall x \in X$$

with equality at $x = x_0$, so that x_0 minimizes the function $x \mapsto c(x, y_0) - \varphi(x)$. Since $x_0 \in \mathrm{spt}(\mu)$ and x_0 belongs to the interior of X, one necessarily has $\nabla\varphi(x_0) = \nabla_x c(x_0, y_0)$. Then, by the twist condition, one necessarily has $y_0 = y_c(x_0, \nabla\varphi(x_0))$. This shows that any optimal transport plan γ is supported on the graph of the map $T : x \in B \mapsto y_c(x_0, \nabla\varphi(x_0))$, and $\gamma = \gamma_T$ by Lemma 4.5. □

4.4 Square-norm cost and link with convexity

When the cost is given by $c(x,y) := \frac{1}{2}\|y-x\|_2^2$ there is a connection between c-concavity and the usual notion of convexity.

Proposition 4.8. *Given a function $\xi : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$, let us define $u_\xi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ through $u_\xi(x) = \frac{1}{2}\|x\|_2^2 - \xi(x)$. Then for $c(x,y) = \frac{1}{2}\|y-x\|_2^2$, we have $u_{\xi^c} = (u_\xi)^*$. In particular, a function ξ is c-concave iff $x \mapsto \frac{1}{2}\|x\|_2^2 - \xi(x)$ is convex and lower-semicontinuous.*

Proof. Observe that

$$u_{\xi^c}(x) = \frac{1}{2}\|x\|_2^2 - \xi^c(x) = \sup_y \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|x-y\|_2^2 + \xi(x) = \sup_y \langle x, y\rangle - \Big(\frac{1}{2}\|y\|_2^2 - \xi(y)\Big).$$

This proves the first part of the statement. The second part follows from the fact that convex l.s.c. functions are characterized by the fact that they are sup of affine functions. □

Theorem 4.9. *Let $c(x,y) = \frac{1}{2}\|y-x\|^2$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be compactly supported. If μ is absolutely continuous then there exists a unique optimal transport plan between μ and ν which is of the form $(\mathrm{id} \times \nabla\tilde{\varphi})_{\#}\mu$ for some convex function $\tilde{\varphi} : \mathbb{R}^d \to \mathbb{R}$.*

Proof. Consider two compact subsets $X, Y \subset \mathbb{R}^d$ that contain $\mathrm{spt}(\mu)$ and $\mathrm{spt}(\nu)$) in their respective interior. Then apply of Theorem 4.7. It holds $\nabla_x c(x_0, y) = x_0 - y$, which is injective for all x_0, thus $y_x(x_0, v) = x_0 - v$ and the optimal transport map is $T(x) = x - \nabla\varphi(x)$ for some c-concave φ. Finally, extend φ by $-\infty$ outside of X and define $\tilde{\varphi}(x) = \frac{1}{2}\|x\|^2 - \varphi(x)$ which is convex and l.s.c. by Proposition 4.8, with gradient $\nabla\tilde{\varphi}(x) = x - \nabla\varphi(x)$. □

References

[1] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

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

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

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
$$\mathcal{T}_c(\mu, \nu) = \sup_{\varphi \ 1-\mathrm{Lip}} \int \varphi \mathrm{d}(\mu - \nu).$$
  - 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 $p \geqslant 1$,

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

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. □

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 $x_0 \in X$,

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

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

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

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

□

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. □

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*

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

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. □

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

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

By minimizing this functional over all interpolations between μ and ν, we recover the optimal transport with cost $\|y - x\|^p$. 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

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

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

Lecture 4 : Entropic Optimal Transport and Numerics

Luca Nenna

March 11, 2020

Introduction: Discret Optimal Transport

We now consider the optimal transport problems between probability measures on two finite sets X and Y with, for simplicity, both cardinality N and we set

$$\mu = \sum_{x\in X} \mu_x \delta_x \qquad \nu = \sum_{y\in Y} \nu_y \delta_y.$$

Definition 0.1 (Discrete OT)**.** The discrete Optimal transport problem between two given measures μ and ν and a given cost function $c : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ is the following minimization problem

$$\inf \left\{ \sum_{x\in X} \sum_{y\in Y} \gamma_{xy} c(x,y) \mid \gamma \in \Pi(\mu,\nu) \right\}, \tag{0.1}$$

where the set of admissible couplings is now define as

$$\Pi(\mu,\nu) := \{\gamma \in X \times Y \mid \gamma_{xy} \geqslant 0,\ \sum_{y\in Y} \gamma_{xy} = \mu_x\ \forall x \in X,\ \sum_{x\in X} \gamma_{xy} = \nu_y\ \forall y \in Y\}.$$

Unfortunately, this linear programming problem has complexity $O(N^3)$ which actually means that it is infeasible for large N. A way to overcome this difficulty is by means of the **Entropic Regularization** which provides an approximation of Optimal Transport with lower computational complexity and easy implementation.

References: Entropic regularisation of Optimal Transport is a very active research field. We refer the interested reader to [1, 3, 6, 7, 4] and the citations therein. We also remark that these notes are inspired by the graduate classe on Numerical Optimal Transport given by F.-X. Vialard [8].

1 The Entropic Optimal Transport

1.1 The discrete case

We start from the primal formulation of the optimal transport problem, but instead of imposing the constraints $\gamma_{xy} \geqslant 0$, we add a term $\mathrm{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$, involving the (opposite of the) entropy

$$e(r) = \begin{cases} r(\log r - 1) & \text{if } r > 0 \\ 0 & \text{if } r = 0 \\ +\infty & \text{if } r < 0 \end{cases}$$

More precisely, given a parameter $\varepsilon > 0$ we consider

$$P_\varepsilon = \inf \left\{ \sum_{x,y} \gamma_{xy} c(x,y) + \varepsilon \operatorname{Ent}(\gamma) \mid \gamma \in X \times Y,\ \sum_{y \in Y} \gamma_{xy} = \mu_x,\ \sum_{x \in X} \gamma_{xy} = \nu_y \right\} \tag{1.2}$$

Theorem 1.1. *The problem P_ε has a unique solution $\gamma^\star$, which belongs to $\Pi(\mu,\nu)$. Moreover, if $\min(\min_{x\in X} \mu_x, \min_{y\in Y} \nu_y) > 0$ then*

$$\gamma_{x,y} > 0 \ \forall (x,y) \in X \times Y.$$

Before introducing the duality, it is important to state the following convergence result in ε.

Theorem 1.2 (Convergence in ε)**.** *The unique solution γ_ε to (1.2) converges to the optimal solution with minimal entropy within the set of all optimal solutions of the Optimal Transport problem, that is*

$$\gamma_\varepsilon \xrightarrow[\varepsilon \to 0]{} \operatorname{argmin} \left\{ \operatorname{Ent}(\gamma) \mid \gamma \in \Pi(\mu,\nu),\ \sum_{x,y} \gamma_{x,y} c(x,y) = \mathcal{MK}_c(\mu,\nu) \right\}. \tag{1.3}$$

Proof. Consider a sequence $(\varepsilon_k)_k$ such that $\varepsilon_k \to 0$ and $\varepsilon_k > 0$ and denote γ_k the solution to(1.2) with $\varepsilon = \varepsilon_k$. Since $\Pi(\mu,\nu)$ is bounded and close we can extract a converging subsequence $\gamma_k \to \gamma^\star \in \Pi(\mu,\nu)$. Take now any optimal γ for the unregularized problem then by optimality of γ_k and γ one has

$$0 \leqslant \langle \gamma_k | c \rangle - \langle \gamma | c \rangle \leqslant \varepsilon_k (\operatorname{Ent}(\gamma) - \operatorname{Ent}(\gamma_k)), \tag{1.4}$$

where $\langle \gamma | c \rangle := \sum_{x,y} \gamma_{x,y} c(x,y)$. Since Ent is continuous, by taking the limit $k \to +\infty$ in (1.4) we get $\langle \gamma^\star | c \rangle = \langle \gamma | c \rangle$. Furthermore, dividing by ε_k and taking the limit we obtain that $\operatorname{Ent}(\gamma) \geqslant \operatorname{Ent}(\gamma^\star)$ showing that $\gamma^\star$ is a solution to the minimization problem in (1.3). By strict convexity of Ent the optimization problem (1.3) has a unique solution and the whole sequence is converging to $\gamma^\star$. □

We want now to derive formally the dual problem. For this purpose we introduce the Lagrangian associated to (1.2)

$$\begin{aligned} \mathcal{L}(\gamma,\varphi,\psi) := \sum_{x,y} \gamma_{xy} c(x,y) + \varepsilon e(\gamma_{xy}) + \sum_{x \in X} \varphi(x) \left(\mu_x - \sum_{y \in Y} \gamma_{xy} \right) \\ + \sum_{y \in Y} \psi(y) \left(\nu_y - \sum_{y \in Y} \gamma_{xy} \right), \end{aligned} \tag{1.5}$$

where $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ are the Lagrange multipliers. Then,

$$P_\varepsilon = \inf_\gamma \sup_{\varphi,\psi} \mathcal{L}(\gamma,\varphi,\psi),$$

and the dual problem is obtained by interchanging the infimum and the supremum :

$$D_\varepsilon = \sup_{\varphi,\psi} \min_{\gamma} \sum_{x,y} \gamma_{xy}(c(x,y) - \psi(y) - \varphi(x) + \varepsilon(\log(\gamma_{xy}) - 1)) + \sum_{x\in X} \varphi(x)\mu_x + \sum_{y\in Y} \psi(y)\nu_y. \tag{1.6}$$

Taking the derivative with respect to γ_{xy}, we find that for a given φ, ψ, the optimal γ must satisfy:

$$\begin{aligned} c(x,y) - \psi(y) - \varphi(x) + \varepsilon \log(\gamma_{xy}) &= 0 \\ \text{i.e. } \gamma_{xy} &= \exp(\frac{1}{\varepsilon}(\varphi(x) + \psi(y) - c(x,y)) \end{aligned} \tag{1.7}$$

Putting these values in the definition of D_ε gives

$$\begin{aligned} D_\varepsilon &= \sup_{\varphi,\psi} \Phi_\varepsilon(\varphi,\psi) \text{ with} \\ \Phi_\varepsilon(\varphi,\psi) &:= \sum_{x\in X} \varphi(x)\mu_x + \sum_{y\in Y} \psi(y)\nu_y - \sum_{x,y} \varepsilon \exp(\frac{1}{\varepsilon}(\varphi(x) + \psi(y) - c(x,y))) \end{aligned} \tag{1.8}$$

Note that thanks to the relation (1.7), one can recover a solution to the primal problem from the dual one. This is true because, unlike the original linear programming formulation of the optimal transport problem, the regularized problem (1.2) is smooth and strictly convex. The following duality result holds

Theorem 1.3 (Strong duality)**.** *Strong duality holds and the maximum in the dual problem is attained, that is* $\exists \varphi, \psi$ *such that*

$$P_\varepsilon = D_\varepsilon = \Phi_\varepsilon(\varphi,\psi).$$

Corollary 1.4. *If* (φ,ψ) *is the solution to* (1.8)*, then the solution* $\gamma^\star$ *to* (1.2) *is given by*

$$\gamma_{x,y} = e^{\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}}$$

Notice now that the optimal coupling γ can be written as

$$\gamma_{x,y} = D_\varphi e^{\frac{-c(x,y)}{\varepsilon}} D_\psi,$$

where D_φ and D_ψ are the diagonal matrices associated to $e^{\varphi/\varepsilon}$ and $e^{\psi/\varepsilon}$, respectively. The problem is now similar to a matrix scaling problem

Definition 1.5 (Matrix scaling problem)**.** Let $K \in \mathbb{R}^{N\times N}$ be a matrix with positive coefficients. Find D_ψ and D_ψ positive diagonal matrices in $K \in \mathbb{R}^{N\times N}$ such that $D_\varphi K D_\psi$ is doubly stochastic, that is sum along each row and each column is equal to 1.

Remark 1.6. Uniqueness fails since if (D_φ, D_ψ) is a solution then so is $(cD_\varphi, \frac{1}{c}D_\psi)$ for every $c \in \mathbb{R}_+$.

Algorithm 1 Sinkhorn-Knopp algorithm for the matrix scaling problem

function SINKHORN-KNOPP(K)
  $D_\varphi^0 \leftarrow \mathbf{1}_N,\ D_\psi^0 \leftarrow \mathbf{1}_N$
  for $0 \leqslant k < k_{\max}$ **do**
    $D_\varphi^{k+1} \leftarrow \mathbf{1}_N ./ (K D_\psi^k)$
    $D_\psi^{k+1} \leftarrow \mathbf{1}_N ./ (K^T D_\varphi^{k+1})$
  end for
end function

Algorithm 2 Sinkhorn-Knopp algorithm for the regularised optimal transport problem

function SINKHORN-KNOPP(K_ε, μ, ν)
  $D_\varphi^0 \leftarrow \mathbf{1}_X,\ D_\psi^0 \leftarrow \mathbf{1}_Y$
  for $0 \leqslant k < k_{\max}$ **do**
    $D_\varphi^{k+1} \leftarrow \mu ./ (K D_\psi^k)$
    $D_\psi^{k+1} \leftarrow \nu ./ (K^T D_\varphi^{k+1})$
  end for
end function

The matrix scaling problem can be easily solved by using an iterative algorithm, known as Sinkhorn-Knopp algorithm, which simply alternates updating D_φ and D_ψ in order to match the marginal constraints (a vector $\mathbf{1}_N$ of ones in this simple case).

where ./ stand for the element-wise division. Denoting by $(K_\varepsilon)_{x,y} = e^{\frac{-c(x,y)}{\varepsilon}}$ the algorithm takes the form 2 for the regularized optimal transport problem.

Notice that one can recast the regularized OT in the framework of bistochastic matrix scaling by replacing the kernel $e^{\frac{-c(x,y)}{\varepsilon}}$ with $(K_\varepsilon)_{x,y} = \mathrm{diag}(\mu) e^{\frac{-c(x,y)}{\varepsilon}} \mathrm{diag}(\nu)$, where $\mathrm{diag}(\mu)$ ($\mathrm{diag}(\nu)$) denotes the diagonal matrix with the vector μ (ν) as main diagonal. In this case the problem (1.2) can be re-written as

$$P_\varepsilon(\mu,\nu) = \inf\left\{ \sum_{x,y} \gamma_{xy} c(x,y) + \varepsilon \mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma \in X\times Y,\ \sum_{y\in Y}\gamma_{xy} = \mu_x,\ \sum_{x\in X}\gamma_{xy} = \nu_y \right\}, \tag{1.9}$$

where $\mathcal{H}(\rho|\mu) := \sum_x \rho_x(\log(\frac{\rho_x}{\mu_x}) - 1)$ is the relative entropy or the Kullback-Leibler divergence.

Good to know: one can easily recast the regularized OT in the continuous framework as follows

$$\mathcal{P}_\varepsilon(\mu,\nu) = \inf\left\{ \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma\in\Pi(\mu,\nu) \right\}, \tag{1.10}$$

where

$$\mathcal{H}(\rho|\pi) = \begin{cases} \displaystyle\int_{X\times Y} \Big(\log\Big(\frac{\mathrm{d}\rho(x,y)}{\mathrm{d}\pi(x,y)}\Big) - 1\Big)\mathrm{d}\rho(x,y), & \text{if } \rho \ll \pi \\ +\infty, & \text{otherwise,} \end{cases}$$

and the marginals μ, ν are probability measures on the compact metric spaces X and Y, respectively. This problem is often referred to as the *static Schrödinger problem* [6] since it was initially considered by Schrödinger in statistical physics. Once again,

under mild assumptions on the cost functions, one can prove that the regularized problem converges to original one as $\varepsilon \to 0$; see [2, 5].

1.2 The convergence of Sinkhorn in the continuous setting

As presented in Lecture 1, the existence of Kantorovich potentials for the standard Optimal Transport problem can be proven by standard compactness arguments. By using similar arguments we show existence for the regularized dual problem (and convergence of Sinkhorn at the same time) in the continuous framework. We firstly recall that a coordinate ascent algorithm on a fucntion of two variables $f(x, y)$ can written as

$$\begin{aligned} y_{k+1} &= \operatorname{argmax}_y f(x_k, y), \\ x_{k+1} &= \operatorname{argmax}_x f(x, y_{k+1}). \end{aligned}$$

The Sinkhorn algorithm is actually a coordinate ascent algorithm: the main idea is indeed to maximize $\Phi_\varepsilon(\varphi, \psi)$ by maximizing alternatively in φ and ψ. From now on we assume for simplicity that $X = Y$.

Proposition 1.7. *The dual problem to (1.10) reads as*

$$D_\varepsilon = \sup\{\Phi_\varepsilon(\varphi, \psi) \mid \varphi, \psi \in \mathcal{C}_0(X)\}, \tag{1.11}$$

where

$$\begin{aligned} \Phi_\varepsilon(\varphi, \psi) := &\int_X \varphi(x) \mathrm{d}\mu(x) + \int_Y \psi(y) \mathrm{d}\nu(y) \\ &- \varepsilon \int_{X\times Y} \exp(\frac{1}{\varepsilon}(\varphi(x) + \psi(y) - c(x, y))) \mathrm{d}\mu \otimes \mathrm{d}\nu(x, y). \end{aligned}$$

It is strictly convex w.r.t. each argument φ and ψ and strictly convex w.r.t. $\varphi(x) + \psi(y)$. It is also Fréchet differentiable for the $(\mathcal{C}_0, \|\cdot\|_\infty)$ topology. Furthermore, if a maximizer exists it is unique up to a constant, that is $\Phi_\varepsilon(\varphi, \psi) = \Phi_\varepsilon(\varphi + C, \psi - C)$ for every $C \in \mathbb{R}$.

Proof. We leave the proof as an exercice. □

Proposition 1.8. *The maximization of $\Phi_\varepsilon(\varphi, \psi)$ w.r.t. each variable can be made explicit, and the Sinkhorn algorithm can be defined as*

$$\varphi_{k+1}(x) = -\varepsilon \log\left(\int_X \exp(\frac{1}{\varepsilon}(\psi_k(y) - c(x, y)) \mathrm{d}\nu(y)\right) := S_\nu(\psi_k), \tag{1.12}$$

$$\psi_{k+1}(y) = -\varepsilon \log\left(\int_X \exp(\frac{1}{\varepsilon}(\varphi_{k+1}(x) - c(x, y)) \mathrm{d}\mu(x)\right) := S_\mu(\varphi_{k+1}). \tag{1.13}$$

Moreover, the following properties hold

(i) $\Phi_\varepsilon(\varphi_k, \psi_k) \leqslant \Phi_\varepsilon(\varphi_{k+1}, \psi_k) \leqslant \Phi_\varepsilon(\varphi_{k+1}, \psi_{k+1})$;

(ii) The continuity modulus of $\varphi_{k+1}, \psi_{k+1}$ is bounded by that of $c(x, y)$;

(iii) If $\psi_k - C$ ($\varphi_{k+1} - C$) is bounded by M on the support of ν (μ), then so us φ_{k+1} (ψ_{k+1}).

Proof. (1.12) and (1.13) follow by writing the first-order necessary condition which gives us

$$1 - \exp\left(\frac{\varphi(x)}{\varepsilon}\right)\int_Y \exp(-\frac{1}{\varepsilon}(\psi(y) - c(x,y)))\mathrm{d}\nu(y) = 0,\ x - a.e.$$

implying the desired formula (and by symmetry, the same result on S_μ holds). Therefore, $S_\nu(\psi)$ is the unique maximizer of $\varphi \mapsto \Phi_\varepsilon(\varphi, \psi)$.
By definition of ascent on each coordinate, (i) is obtained directly. For (ii), remark that the derivative of $\log(\sum_i \exp(x_i))$ w.r.t. x_j is $\frac{\exp(x_j)}{sum_i \exp(x_i)}$ bounded from 1. Then, $x \mapsto \log \int_X \exp(\frac{c(x,y)-\psi(y)}{\varepsilon})\mathrm{d}\nu(y)$ is $L-$Lipschitz where L is the Lip. constant of c and the modulus of continuity of φ_{k+1} and ψ_{k+1} is bounded by that of c. The last point is just a bound on the iterates. □

Proposition 1.9. *The sequence (φ_k, ψ_k) defined by (1.12) and (1.13) converges in $(\mathcal{C}_0, \|\cdot\|_\infty)$ to the unique (up to a constant) couple of potentials (φ, ψ) which maximizes Φ_ε.*

Proof. Shifting the potentials by an additive constant, one can replace the optimization set by the couples (φ, ψ) which have uniformly bounded modulus of continuity and such that $\varphi(x_0) = 0$ for a given $x_0 \in X$. Recall that by proposition 1.7 the maximum of Φ is achieved at some couple (φ^*, ψ^*) which is unique up to a constant. Then, by prop. 1.8 (φ_k, ψ_k) are uniformly bounded and have uniformly modulus of continuity and one can extract a converging subsequence to $(\overline{\varphi}, \overline{\psi})$. By continuity of Φ and the monotonicity of the sequence, $\Phi_\varepsilon(\overline{\varphi}, S_\mu(\overline{\varphi})) \leqslant \Phi_\varepsilon(S_\nu \circ S_\mu(\overline{\varphi}), S_\mu(\overline{\varphi})) = \Phi_\varepsilon(\overline{\varphi}, S_\mu(\overline{\varphi}))$. Therefore, the maximizer coordinatewise being unique, one has

$$S_\nu(\overline{\psi}) = \overline{\varphi}, \tag{1.14}$$
$$S_\mu(\overline{\varphi}) = \overline{\psi}. \tag{1.15}$$

These show that $(\overline{\varphi}, \overline{\psi})$ is a critical point for Φ_ε, thus being a maximizer. □

The proof of convergence relies on some important properties of the log−sum−exp (LSE) function $\log \int \exp$ which we summarise in the next Lemma. Before that let define the pseudo-norm $\|\cdot\|_{\circ,\infty}$ of uniform convergence as

$$\|f\|_{\circ,\infty} := \frac{1}{2}(\sup f - \inf f) = \inf_{a\in\mathbb{R}} \|f + a\|_\infty .$$

Lemma 1.10. *The LSE function is convex and*

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \|\varphi_1 - \varphi_2\|_{\circ,\infty} . \tag{1.16}$$

Proof. Convexity is easily verified. We can get the 1−Lipschitz property as follows

$$\begin{aligned}
|S_\mu(\varphi_1)(x) - S_\mu(\varphi_2)(x)| &= \left|\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} S_\mu(\varphi_2 + t(\varphi_1 - \varphi_2))\mathrm{d}t\right| \\
&\leqslant \int_0^1 \left|\int_X (\varphi_1 - \varphi_2)\frac{\exp(\frac{1}{\varepsilon}(\varphi_2 + t(\varphi_1 - \varphi_2) - c))}{\int_X \exp(\frac{1}{\varepsilon}(\varphi_2 + t(\varphi_1 - \varphi_2) - c))\mathrm{d}\mu}\mathrm{d}\mu\right| \\
&\leqslant \|\varphi_1 - \varphi_2\|_\infty .
\end{aligned}$$

Notice that the equality occurs if and only if $\varphi_1 - \varphi_2$ is constant $\mu-$a.e.. In particular we would have $\varphi_1 = \varphi_2 + a$ and $S_\mu(\varphi_1) = S_\mu(\varphi_2) + a$. Thus it is natural to consider the set of continuous functions up to an additive constant $\mathcal{C}(X)/\mathbb{R}$ endowed with the pseudo-norm introduced above. Then, since $S_\mu(\varphi_1 + a) = S_\mu(\varphi_1) + a$ we got the same inequality for the norm $\|\cdot\|_{\circ,\infty}$. □

Lemma 1.11. *Let $u, v \in \mathcal{C}(X)$ and $\mu \in \mathcal{P}(X)$ and denote ν_u and ν_v the Gibbs measures associated to u and v, that is $\mathrm{d}\nu_u = \frac{1}{Z_u} e^u \mathrm{d}\mu$ and $\mathrm{d}\nu_v = \frac{1}{Z_v} e^v \mathrm{d}\mu$, where Z_u and Z_v are the normalizing constants, then*

$$\|\nu_u - \nu_v\|_{L^1} \leqslant 2(1 - e^{-2\|u-v\|_{\circ,\infty}}).$$

Proof. Consider a bounded function g on X and define

$$\eta_g(t) := \int_X g \frac{e^{tv+(1-t)u}}{Z_{t,g}} \mathrm{d}\mu,$$

where $Z_{t,g} = \int_X e^{tv+(1-t)u} \mathrm{d}\mu$. Differentiating we get

$$\eta_g'(t) + \eta_{v-u}(t)\eta_g(t) = \eta_{(v-u)g}(t),$$

and

$$e^{\int_0^t \eta_{v-u}(s)\mathrm{d}s}\eta_g(t) - \eta_g(0) = \int_0^t \eta_{(v-u)g}(s) e^{\int_0^s \eta_{v-u}(r)\mathrm{d}r} \mathrm{d}s.$$

Observe that

$$\begin{aligned} |e^{\int_0^t \eta_{v-u}(s)\mathrm{d}s}\eta_g(t) - \eta_g(0)| &\leqslant \|g\|_\infty \int_0^t \eta_{(u-v)}(s) e^{\int_0^s \eta_{u-v}(r)\mathrm{d}r} \mathrm{d}s \\ &\leqslant \|g\|_\infty \left(e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} - 1\right). \end{aligned}$$

Interchanging the role of u and v we have two possible cases: $\eta_g(1) \geqslant \eta_g(0) \geqslant 0$ or $\eta_g(1) \geqslant 0 \geqslant \eta_g(0)$. In the first case one has

$$|e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s}(\eta_g(t) - \eta_g(0))| \leqslant |e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s}\eta_g(t) - \eta_g(0)| \leqslant \|g\|_\infty \left(e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} - 1\right).$$

In the second case there exists $t_0 \in [0,1]$ such that $\eta_g(t_0) = 0$ and we get

$$\begin{aligned} |\eta_g(1)| &\leqslant \|g\|_\infty \underbrace{\left(1 - e^{\int_{t_0}^1 \eta_{u-v}(s)\mathrm{d}s}\right)}_{:=a_1} \\ |\eta_g(0)| &\leqslant \|g\|_\infty \underbrace{\left(1 - e^{\int_0^{t_0} \eta_{u-v}(s)\mathrm{d}s}\right)}_{:=a_0}. \end{aligned}$$

Thus,

$$|\eta_g(1) - \eta_g(0)| \leqslant |\eta_g(1)| + |\eta_g(0)| \leqslant 2\|g\|_\infty \max(a_1, a_0)$$

By exploiting the fact that $\eta_{u-v}(t) \leqslant 2\|u - v\|_{\circ,\infty}$ we obtain in both cases that

$$\|\nu_u - \nu_v\| \leqslant 2(1 - e^{-2\|u-v\|_{\circ,\infty}})$$

□

Theorem 1.12. *(Convergence of Sinkhorn) The map $S = S_\nu \circ S_\mu$ is a contraction for $\|\cdot\|_{\circ,\infty}$. In particular the sequence (φ_k, ψ_k) defined by the Sinkhorn algorithm linearly converges to the unique (up to a constant) maximiser of the dual problem.*

Proof. We actually have to prove that

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}, \tag{1.17}$$

and analogously for S_ν, with $\kappa_\mu < 1$ ($\kappa_\nu < 1$). Once we have established that S_μ (S_ν) is a contraction then it easily follows that

$$\|S(\varphi_1) - S(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu\kappa_\nu \|\varphi_1 - \varphi_2\|_{\circ,\infty},$$

which would conclude the proof.
In order to prove (1.17) we start by giving an estimation of the oscillations of S_μ

$$\frac{1}{2}|S_\mu(\varphi_1)(y) - S_\mu(\varphi_2)(y) - S_\mu(\varphi_1)(x) + S_\mu(\varphi_2)(x)| \leqslant \frac{1}{2}\left|\int_0^1 \int_X (\varphi_1 - \varphi_2)(\mathrm{d}\eta_{t,y} - \mathrm{d}\eta_{t,x})t\right|,$$

where $\mathrm{d}\eta_{t,z} := \frac{1}{Z} e^{\frac{t(\varphi_1-\varphi_2)+\varphi_2-c(z,\cdot)}{\varepsilon}} \mathrm{d}\mu$ where Z is the normalising constant. Since $\mathrm{d}\eta t, z$ is a Gibbs measure we can apply the L^1 bound of lemma 1.11 to estimate $\|\eta_{t,y} - \eta_{t,x}\|_{L^1}$ and get

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}$$

with $\kappa_\mu = 2(1 - e^{-2\frac{\|c\|_{\circ,\infty}}{\varepsilon}})$. □

References

[1] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré, *Iterative bregman projections for regularized transportation problems*, SIAM Journal on Scientific Computing **37** (2015), no. 2, A1111–A1138.

[2] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer, *Convergence of entropic schemes for optimal transport and gradient flows*, SIAM Journal on Mathematical Analysis **49** (2017), no. 2, 1385–1418.

[3] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, *Scaling algorithms for unbalanced transport problems*, arXiv preprint arXiv:1607.05816 (2016).

[4] Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems, 2013, pp. 2292–2300.

[5] Christian Léonard, *From the schr\" odinger problem to the monge-kantorovich problem*, arXiv preprint arXiv:1011.2564 (2010).

[6] ———, *A survey of the schr\" odinger problem and some of its connections with optimal transport*, arXiv preprint arXiv:1308.0215 (2013).

[7] Gabriel Peyré, Marco Cuturi, et al., *Computational optimal transport*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.

[8] François-Xavier Vialard, *An elementary introduction to entropic regularization and proximal methods for numerical optimal transport*, (2019).