Lecture 1
Optimal Transport: the Theoretical Foundations

Luca Nenna

August 31, 2020

Some motivations for studying optimal transport.

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

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

  (synthetic notion of Ricci curvature for metric spaces), machine learning, inverse problems, etc.
- Quantum physics: electronic configuration in molecules and atoms.
- Economics : μ is the distribution of men and ν the distribution of women : how can we match men and women such that everyone has an happy marriage?
- Imaging, Game theory, Mean Field Games, Fluid Dynamics, Cosmology : **Optimal Transport is everywhere!**

References.

Introduction to optimal transport, with applications to PDE and/or calculus of variations can be found in books by Villani [8] and Santambrogio [7]. Villani's second book [9] concentrates on the application of optimal transport to geometric questions (e.g. synthetic definition of Ricci curvature). We also mention Gigli, Ambrosio and Savaré [2] for the study of gradient flows with respect to the Monge-Kantorovich/Wasserstein metric. On the Economics side we refer the interested reader to [3] and for the applications in data sciences we suggest [5].

1 The problems of Monge and Kantorovich

Let us start by giving some notations/remarks/definitions useful for the all the lecture.
Discrete measures: A discrete measure with weights $\mathbf{a}$ and locations $x_1, \cdots, x_n \in X \subset$

$\mathbb{R}^n$ reads

$$\mu = \sum_{I=1}^{n} \mathbf{a}_i \delta_{x_i},$$

Where δ_{x_i} is the Dirac at position x_i. Such a measure describes a probability measure if, additionally, $\mathbf{a} \in \Sigma_n := \{\mathbf{a} \in \mathbb{R}_+^n \mid \sum_{i=1}^n \mathbf{a}_i = 1\}$ and a more generally positive measure if all the elements of the vector $\mathbf{a}$ are nonnegative.

General measures: Let X a compact subset of $\mathbb{R}^n$ we denote by $\mathcal{P}(X)$ the set of probability measures on X, by $\mathcal{M}_+(X)$ the set of positive measures on X, that is $\mu(X) \geqslant 0$ and by $\mathcal{M}(X)$ the set of init measures on X.

Relative densities: a measure μ which is a weighting of another reference one $\mathrm{d}x$ is said to have a density, which is denoted $\mathrm{d}\mu = \overline{\mu}\mathrm{d}x$ (in the following we always assume that $\mathrm{d}x$ is the Lebesgue measure) that is

$$\forall f \in \mathcal{C}(X), \int_X f(x)\mathrm{d}\mu(x) = \int_X f(x)\overline{\mu}\mathrm{d}x.$$

Definition 1.1 (Push-forward)**.** Given $X, Y \subset \mathbb{R}^n$, for $T : X \to Y$, the push-forward measure $\nu = T_\sharp\mu \in \mathcal{M}(Y)$ of some $\mu \in \mathcal{M}(X)$ satisfies

$$\forall f \in \mathcal{C}(Y), \int_Y f(y)\mathrm{d}\nu(y) = \int_X f(T(x))\mathrm{d}\mu(x)$$

. Note that $T_\sharp$ preserves positivity and total mass, that if $\mu \in \mathcal{P}(X)$ then $T_\sharp\mu \in \mathcal{P}(Y)$.

Example 1.2. If μ is a discrete measure then

$$T_\sharp\mu := \sum_i \mathbf{a}_i \delta_{T(x_i)}.$$

Example 1.3 (Push-forward for densities)**.** Explicitly doing the change of variable $y = T(x)$ for measures with densities $\overline{\mu}, \overline{\nu}$ (assuming T is a $\mathcal{C}^1$ diffeomorphism), one has for all $f \in \mathcal{C}(Y)$

$$\int_Y f(y)\overline{\nu}(y)\mathrm{d}y = \int_X f(T(x))\overline{\nu}(T(x))\det(\mathrm{D}T(x))\mathrm{d}x = \int_X f(T(x))\overline{\mu}(x)\mathrm{d}x.$$

Hence,

$$\overline{\mu}(x) = \overline{\nu}(T(x))\det(\mathrm{D}T(x)).$$

1.1 The matching problem

Definition 1.4 (Matching problem)**.** Given a cost matrix $C \in \mathbb{R}^n \times \mathbb{R}^m$, assuming $n = m$, the optimal assignment problem seeks for a bijection σ in the set of permutations of n elements $\mathfrak{S}_n$ solving

$$\min_{\sigma \in \mathfrak{S}_n} \frac{1}{n} \sum_{i=1}^{n} C_{i,\sigma(i)}. \tag{1.1}$$

One can naively evaluate the cost function above by using all permutations in the set $\mathfrak{S}_n$. However, that set has size $n!$, which is gigantic even for small n!!!. In general σ is not unique.

Let us consider now a cost of the form $C_{ij} = h(x_i - y_j)$ where $h : \mathbb{R} \to \mathbb{R}_+$ is strictly convex, one has that an optimal σ must satisfy the following inequality: given $(x_i, y_{\sigma(i)})$ and $(x_j, y_{\sigma(j)})$ then

$$h(x_i - y_{\sigma(i)}) + h(x_j - y_{\sigma(j)}) \leqslant h(x_i - y_{\sigma(j)}) + h(x_j - y_{\sigma(i)}),$$

Otherwise it would be more efficient to move mass from x_i to $y_{\sigma(j)}$ and x_j to $y_{\sigma(i)}$. The above inequality and the strict convexity of h imply that the optimal σ defines an increasing map, that is

$$\forall (i,j) (x_i - x_j)(y_{\sigma(i)} - y_{\sigma(j)}) \geqslant 0.$$

Thus, the algorithm to compute an optimal transport, i.e. the optimal permutation σ, is to sort the points: find some pair of permutations σ_X, σ_Y such that

$$x_{\sigma_X(1)} \leqslant y_{\sigma_X(2)} \leqslant \cdots \text{ and } y_{\sigma_Y(1)} \leqslant y_{\sigma_Y(2)} \leqslant \cdots$$

and then an optimal matching is to send $x_{\sigma_X(k)}$ to $y_{\sigma_Y(k)}$, that is the optimal permutation is given by $\sigma = \sigma_Y^{-1} \circ \sigma_X$.

1.2 Monge problem

Definition 1.5 (Monge problem)**.** Consider $X, Y \subseteq \mathbb{R}^n$, 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

$$\text{(MP)} := \inf \left\{ \int_X c(x, T(x)) \mathrm{d}\mu(x) \mid T : X \to Y \text{ and } T_\# \mu = \nu \right\} \tag{1.2}$$

This problem exhibits several difficulties, one of which is that both the constraint $(T_\# \mu = \nu)$ and the functional are non-convex. For empirical measure with the same number $n = m$ of points, one retrieves the optimal matching problem.

Example 1.6. There might exist no transport map between μ and ν. For instance, consider $\mu = \delta_x$ for some $x \in X$. Then, $\mathrm{T}_\# \mu = \delta_{T(x)}$. In particular, if ν is not a single Dirac then there exists no transport map between μ and ν.

In the special case in which $c(x,y) = d^p(x,y)$ where d is a distance, we denote

$$\mathcal{W}_p^p(\mu, \nu) := \inf \left\{ \int_X d^p(x, T(x)) \mathrm{d}\mu(x) \mid T : X \to Y \text{ and } T_\# \mu = \nu \right\},$$

If the constraint set is empty, then we set $\mathcal{W}_p^p = +\infty$. In particular $\mathcal{W}_p^p$ defines a distance between probability measures!

Proposition 1.7. *$\mathcal{W}_p^p$ is a distance.*

Proof. If $\mathcal{W}_p^p(\mu, \nu) = 0$ then the optimal map is the identity Id which means that $\mu = \nu$. We have now to prove the triangle inequality

$$\mathcal{W}_p^p(\mu, \nu) \leqslant \mathcal{W}_p^p(\mu, \eta) + \mathcal{W}_p^p(\eta, \nu).$$

If $\mathcal{W}_p^p(\mu, \nu) = +\infty$, then either $\mathcal{W}_p^p(\mu, \eta) = +\infty$ or $\mathcal{W}_p^p(\eta, \nu) = +\infty$. Indeed, consider two maps S, T such that $S_\sharp \mu = \eta$ and $T_\sharp \eta = \nu$ then $(T \circ S)_\sharp \mu = \nu$ and we have $\mathcal{W}_p^p(\mu, \nu) \leqslant \int_X d^p(x, T \circ S(x)) \mathrm{d}\mu(x) < +\infty$. So consider $\mathcal{W}_p^p(\mu, \nu) < +\infty$ and restrict our attention to the case in which $\mathcal{W}_p^p(\mu,\eta) < +\infty$ and $\mathcal{W}_p^p(\eta,\nu) < +\infty$, otherwise the inequality is trivial. For any $\varepsilon > 0$, we consider $\varepsilon-$minimizers S and T such that

$$\Big(\int_X d^p(x,S(x))\mathrm{d}\mu(x)\Big)^{1/p} \leqslant \mathcal{W}_p(\mu,\eta)+\varepsilon \text{ and } \Big(\int_X d^p(x,T(x))\mathrm{d}\eta(x)\Big)^{1/p} \leqslant \mathcal{W}_p(\eta,\nu)+\varepsilon.$$

Take the map $T\circ S$, then we have

$$W_p(\mu,\nu) \leqslant \Big(\int_X d^p(x,T\circ S(x))\mathrm{d}\mu(x)\Big)^{1/p} \leqslant \Big(\int_X (d(x,S(x))+d(S(x),T\circ S(x)))^p\mathrm{d}\mu(x)\Big)^{1/p},$$

And be using the Minkowski inequality we obtain

$$W_p(\mu,\nu) \leqslant \Big(\int_X d^p(x,S(x))\mathrm{d}\mu(x)\Big)^{1/p} + \Big(\int_X d^p(S(x),T\circ S(x))\mathrm{d}\mu(x)\Big)^{1/p},$$

Thus

$$W_p(\mu,\nu) \leqslant W_p(\mu,\eta) + W_p(\eta,\nu) + 2\varepsilon,$$

and by letting $\varepsilon \to 0$ we have the desired inequality. □

We consider now the 1−dimensional case: for a measure μ on $\mathbb{R}$ we define the cumulative function

$$\forall x \in \mathbb{R},\ F_\mu(x) := \int_{-\infty}^{x} \mathrm{d}\mu(x),$$

Which is a function $F_\mu : \mathbb{R} \to [0,1]$ and its pseudo-inverse $F_\mu^{-1} : [0,1] \to \mathbb{R}\cup\{-\infty\}$

$$\forall s \in [0,1],\ F_\mu^{-1} = \min_x\{x \in \mathbb{R} \mid F_\mu(x) \geqslant s\}.$$

If μ has a density one can prove that for a strictly convex h the optimal transport map is given by $T = F_\nu^{-1}\circ F_\mu$. Notice that if $c(x,y) = d^p(x,y)$ with $p \geqslant 1$ on has

$$\mathcal{W}_p^p(\mu,\nu) = \int_X |x - F_\nu^{-1}\circ F_\mu(x)^p\mathrm{d}\mu(x) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)|^p\mathrm{d}s = ||F_\mu^{-1} - F_\nu^{-1}||_{L^p([0,1])}$$

. This formula shows that through the map $\mu \mapsto F_\mu^{-1}$, the Wasserstein distance is isometric to a linear space equipped with the L^p norm!

1.3 Kantorovich problem

Definition 1.8 (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, that is

$$\forall (f,g) \in \mathcal{C}(X)\times\mathcal{C}(y),\ \int_{X\times Y} f(x)\mathrm{d}\gamma(x,y) = \int_X f(x)\mathrm{d}\mu(x) \text{ and } \int_{X\times Y} g(y)\mathrm{d}\gamma(x,y) = \int_Y g(y)\mathrm{d}\nu(y).$$

Definition 1.9 (Transport plan)**.** A transport plan between two probabily measures μ,ν on 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 1.10 (Tensor product)**.** Note that the set of transport plans $\Pi(\mu, \nu)$ is never empty, as it contains the measure $\mu \otimes \nu$.

Example 1.11 (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 1.12 (Kantorovich problem)**.** Consider two compact subset of $\mathbb{R}^n$ 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

$$\text{(KP)} := \inf \left\{ \int_{X \times Y} c(x, y) \mathrm{d}\gamma(x, y) \mid \gamma \in \Pi(\mu, \nu) \right\} \tag{1.3}$$

Remark 1.13. 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.

Theorem 1.14 (Existence)**.** *Let X, Y be two compact subspaces, and $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ be a continuous cost function. Then Kantorovich's problem admits a minimizer.*

The main question is to establish the equality between the infimum in Monge problem and the minimum in Kantorovich problem. This part is taken from Santambrogio [7].

Theorem 1.15. *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,*

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

2 The dual problem

We now focus on duality theory without enter into details. 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

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

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)

$$\inf_{\gamma \in \mathcal{M}_+(X \times Y)} \int_{X \times Y} c \mathrm{d}\gamma + \Psi$$

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 2.1 (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

$$\text{(DP)} := \sup \left\{ \int_X \varphi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu \mid \varphi \in \mathcal{C}_b(X),\ \psi \in \mathcal{C}_b(Y),\ \varphi \oplus \psi \leqslant c \right\} \tag{2.4}$$

Remark 2.2. 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 $\inf$ (KP) $\geqslant \sup$ (DP). When $\sup$ (DP) $= \inf$ (KP), one talks of *strong duality.* Note that this is independent of whether the infimum and the supremum are attained.

Remark 2.3. 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.

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] Alfred Galichon, *Optimal transport methods in economics*, Princeton University Press, 2018.

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

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

[6] 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.

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

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

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