Introduction to Optimal Transport Theory

Filippo Santambrogio*

Grenoble, June 15th 2009

These very short lecture notes do not want to be an exhaustive presentation of the topic, but only a short list of results, concepts and ideas which are useful when dealing for the first time with the theory of Optimal Transport. I'm sure most of them will appear in the lectures by my colleagues, and I apologize for any possible superposition. Due to the limited time I had to write them, I'm sure there will be plenty of errors and you are very welcome to point them out.

The main references for the whole topic are the two books on the subject by C. Villani ([15, 16]). For what concerns curves in the space of probability measures, the best specifically focused reference is [2]. Moreover, I'm also very indebted to the approach that L. Ambrosio used in a course at SNS Pisa in 2001/02 and I want to cite as another possible reference [1].

The motivation for the whole subject is the following problem proposed by Monge in 1781 ([14]): given two densities of mass f, $g \geq 0$ on $\mathbb{R}^d$, with $\int f = \int g = 1$, find a map $T : \mathbb{R}^d \to \mathbb{R}^d$ pushing the first one onto the other, i.e. such that

$$\int_A g(x)dx = \int_{T^{-1}(A)} f(y)dy \quad \text{for any Borel subset } A \subset \mathbb{R}^d \tag{0.1}$$

and minimizing the quantity

$$\int_{\mathbb{R}^d} |T(x) - x| f(x)dx$$

among all the maps satisfying this condition. In the following, we will always define, similarly to (0.1), the image measure of a measure μ on X through a measurable map $T : X \to Y$, which is the measure denoted by $T_\#\mu$ on Y and caracterized by

$$T_\#\mu(A) = \mu(T^{-1}(A)) \quad \text{for every measurable set } A,$$
$$\text{or } \int_Y \phi \, d\,(T_\#\mu) = \int_X \phi \circ T \, d\mu \quad \text{for every measurable function } \phi.$$

The problem of Monge has stayed with no solution (does a minimizer exist? how to characterize it?...) for centuries. Only with the work by Kantorovich it has been inserted into a suitable framework which gave the possibility to approach it and, later, to find that actually solutions exist

* CEREMADE, UMR CNRS 7534, Université Paris-Dauphine, Pl. de Lattre de Tassigny, 75775 Paris Cedex 16, FRANCE filippo@ceremade.dauphine.fr , http://www.ceremade.dauphine.fr/∼filippo.

and to study them. The problem has been widely generalized, with very general cost functions $c(x,y)$ instead of the euclidean distance $|x-y|$ and more general measures and spaces. For simplicity, here we will not try to present a very wide theory on generic metric spaces, manifolds and so on, but we will deal only with the euclidean case.

1 Primal and dual problems

In what follows we will suppose Ω to be a (very often compact) domain of $\mathbb{R}^d$ and the cost function $c : \Omega \times \Omega \to [0, +\infty[$ will be supposed continuous and symmetric (i.e. $c(x,y) = c(y,x)$).

1.1 Kantorovitch and Monge problems

The generalization that appears as natural from the work of Kantorovich ([12]) of the problem raised by Monge is the following:

Problem 1. Given two probability measures μ and ν on Ω and a cost function $c : \Omega \times \Omega \to [0, +\infty]$ we consider the problem

$$(K) \quad \min \left\{ \int_{\Omega\times\Omega} c \, d\gamma \, | \gamma \in \Pi(\mu,\nu) \right\}, \tag{1.1}$$

where $\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(\Omega\times\Omega) : \, (p^+)_\#\gamma = \mu, \, (p^-)_\#\gamma = \nu, \}$ and p^+ and p^- are the two projections of $\Omega \times \Omega$ onto Ω. The minimizers for this problem are called *optimal transport plans* between μ and ν. Should γ be of the form $(id \times T)_\#\mu$ for a measurable map $T : \Omega \to \Omega$, the map T would be called *optimal transport map* from μ to ν.

Remark 1. It can be easily checked that if $(id \times T)_\#\mu$ belongs to $\Pi(\mu,\nu)$ then T pushes μ onto ν (i.e. $\nu(A) = \mu(T^{-1}(A))$ for any Borel set A) and the functional takes the form $\int c(x, T(x))\mu(dx)$, thus generalizing Monge's problem.

This generalized problem by Kantorovich is much easier to handle than the original one by Monge, for instance because in the Monge case we would need existence of at least a map T satisfying the constraints. This is not the case in the case $\mu = \delta_0$ if ν is not a single Dirac mass. On the contrary, there always exist transport plan in $\Pi(\mu,\nu)$ (for instance $\mu \otimes \nu \in \Pi(\mu,\nu)$). Moreover, one can state that (K) is the relaxation of the original problem by Monge: if one considers the problem in the same setting, where the competitors are transport plans, but sets the functional at $+\infty$ on all the plans that are not of the form $(id \times T)_\#\mu$, then one has a functional on $\Pi(\mu,\nu)$ whose relaxation is the functional in (K) (see [3]).

Anyway, it is important to notice that an easy use of the Direct Method of Calculus of Variations (i.e. taking a minimizing sequence, saying that is compact in some topology - here it is the weak convergence of probability measures - finding a limit, and proving semicontinuity (or continuity) of the functional we minimize so that the limit is a minimizer) proves that a minimum do exist.

For instance, if one is interested in the problem of Monge, the question becomes "does this minimum come from a transport map T?".

1.2 Duality

Since the problem (K) is a linear optimization under linear constraints, an important tool will be duality theory, which is typically used for convex problems. We will find a dual problem (D) for (K) and exploit the relations between dual and primal.

The first we will do is finding a formal dual problem, by means of an inf-sup exchange.

First express the constraint $\gamma \in \Pi(\mu,\nu)$ in the following way : notice that, if γ is a non-negative measure on $\Omega \times \Omega$, then we have

$$\sup_{\phi,\psi} \int \phi d\mu + \int \psi d\nu - \int (\phi(x)+\psi(y))\, d\gamma = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu,\nu) \\ +\infty & \text{otherwise} \end{cases}.$$

Hence, one can remove the constraints on γ if he adds the previous sup, since if they are satisfied nothing has been added and if they are not one gets $+\infty$ and this will be avoided by the minimization. Hence we may look at the problem we get and interchange the inf in γ and the sup in ϕ, ψ:

$$\min_{\gamma} \int c\, d\gamma + \sup_{\phi,\psi} \int \phi d\mu + \int \psi d\nu - \int (\phi(x)+\psi(y)) d\gamma = \\ \sup_{\phi,\psi} \int \phi d\mu + \int \psi d\nu + \inf_{\gamma} \int (c(x,y) - (\phi(x)+\psi(y)))\, d\gamma.$$

Obviously it is not always possible to exchange inf and sup, and the main tool to do it is a theorem by Rockafellar requiring concavity in one variable, convexity in the other one, and some compactness assumption. We will not investigate anymore whether in this case these assumptions are satisfied or not. But the result is true.

Afterwards, one can re-write the inf in γ as a constraint on ϕ and ψ, since one has

$$\inf_{\gamma \geq 0} \int (c(x,y) - (\phi(x)+\psi(y)))\, d\gamma = \begin{cases} 0 & \text{if } \phi(x)+\psi(y) \leq c(x,y) \text{ for all } (x,y) \in \Omega \times \Omega \\ -\infty & \text{otherwise} \end{cases}.$$

This leads to the following dual optimization problem.

Problem 2. Given the two probabilities μ and ν on Ω and the cost function $c : \Omega \times \Omega \to [0, +\infty]$ we consider the problem

$$(D) \quad \max \left\{ \int_\Omega \phi d\mu + \int_\Omega \psi d\nu \,\middle|\, \phi \in L^1(\mu), \psi \in L^1(\nu) \,:\, \phi(x)+\psi(y) \leq c(x,y) \text{ for all } (x,y) \in \Omega \times \Omega) \right\}, \tag{1.2}$$

This problem does not admit a straightforward existence result, since the class of admissible functions lacks compactness. Yet, we can better understand this problem and find existence once we have introduced the the notion of $c-$transform (a kind of generalization of the well-known Legendre transform).

Definition 1. Given a function $\chi : \Omega \to \overline{\mathbb{R}}$ we define its $c-$*transform* (or $c-$conjugate function) by

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

Moreover, we say that a function ψ is $c-$*concave* if there exists χ such that $\psi = \chi^c$ and we denote by $\Psi_c(\Omega)$ the set of $c-$concave functions.

It is quite easy to realize that, given a pair (ϕ,ψ) in the maximization problem (D), one can always replace it with (ϕ,ϕ^c), and then with (ϕ^{cc},ϕ^c). Actually one could go on but it is possible to prove that $\phi^{ccc} = \phi^c$ for any function ϕ. this is the same as saying that $\psi^{cc} = \psi$ for any $c-$concave function ψ, and this prefectly recalls what happens for the Legendre transform of convex funtions (which corresponds to the particular case $c(x,y) = x \cdot y$).

A consequence of these considerations is the following well-known result

Proposition 1.1. *We have*

$$\min(K) = \max_{\psi\in\Psi_c(\Omega)} \int_\Omega \psi\, d\mu + \int_\Omega \psi^c\, d\nu, \tag{1.3}$$

where the max on the right hand side is realized. In particular the minimum value of (K) *is a convex function of* (μ,ν)*, as it is a supremum of linear functionals.*

Definition 2. The functions ψ realizing the maximum in (1.3) are called *Kantorovich potentials* for the transport from μ to ν. This is in fact a small abuse, because usually this term is used only in the case $c(x,y) = |x-y|$, but it is usually understood in the general case as well.

Notice that any $c-$concave function shares the same modulus of continuity of the cost c. This is the reason why one can prove existence for (D) (which is the same of the right hand side problem in the previous proposition), by applying Ascoli-Arzelà's Theorem.

In, particular, in the case $c(x,y) = |x-y|^p$, if Ω is bounded with diameter D, any $\psi \in \Psi_c(\Omega)$ is $pD^{p-1}-$Lipschitz continuous. The case where c is a power of the distance is in fact of particular interest and two values of the exponent p are remarkable: the cases $p = 1$ and $p = 2$. In these two cases we provide characterizations for the set of $c-$concave functions. Let us denote by $\Psi_{(p)}(\Omega)$ the set of $c-$concave functions with respect the cost $c(x,y) = |x-y|^p/p$. It is not difficult to check that

$$\psi \in \Psi_{(1)}(\Omega) \Longleftrightarrow \psi \text{ is a 1-Lipschitz function;}$$

$$\psi \in \Psi_{(2)}(\Omega) \Longrightarrow\ x \mapsto \frac{x^2}{2} - \psi(x) \text{ is a convex function; if } \Omega = \mathbb{R}^d \text{ this is an equivalence.}$$

1.3 The case $c(x,y) = |x-y|$

The case $c(x,y) = |x-y|$ shows a lot of interesting features, even if from the point of the existence of an optimal map T it is one of the most difficult. A first interesting property is the following:

Proposition 1.2. *For any* $1-$*Lipschitz function* ψ *we have* $\psi^c = -\psi$*. In particular, Formula 1.3 may be re-written as*

$$\min(K) = \max(D) = \sup_{\psi \in \mathrm{Lip}_1} \int_\Omega \psi \, d(\mu - \nu).$$

Tha fact that $\psi^c = -\psi$ is easily proven if one considers that $\psi^c(y) = \inf_x |x-y| - \psi(x) \leq -\psi(x)$ (taking $x = y$, but also $\psi^c(y) = \inf_x |x-y| - \psi(x) \geq \inf_x |x-y| - |x-y| + \psi(y) = \psi(y)$ (making use of the Lipschitz behaviour of ψ).

Another peculiar feature of this case is the following:

Proposition 1.3. *Consider the problem*

$$(B) \quad \min \left\{ M(\lambda) \, : \lambda \in \mathcal{M}^d(\Omega); \nabla \cdot \lambda = \mu - \nu \right\}, \tag{1.4}$$

where $M(\lambda)$ *denotes the mass of the vector measure* λ *and the divergence condition is to be read in the weak sense, with Neumann boundary conditions, i.e.* $-\int \nabla\phi \cdot d\lambda = \int \phi d(\mu - \nu)$ *for any* $\phi \in C^1(\Omega)$*. If* Ω *is convex then it holds*

$$\min(K) = \min(B).$$

This proposition links the Monge-Kantorovich problem to a minimal flow problem which has been first proposed by Beckmann in [5], under the name of *continuous transportation model*, without knowing this link, as Kantorovich's theory was being developed independently almost in the same years. In Section 2.1 we will see some details more on this model and on the possibility of generalizing it to the case of distances $c(x, y)$ coming from Riemannian metrics. In particular, in the case of a nonconvex Ω, (B) would be equivalent to a Monge-Kantorovich problem where c is the geodesic distance on Ω.

To have an idea of why these equivalences between (B) and (K) hold true, one can look at the following considerations.

First, a formal computation. We take the problem (B) and re-write the constraint on λ by means of the quantity

$$\sup_\phi \int -\nabla\phi \cdot \, d\lambda + \int \phi d(\mu - \nu) = \begin{cases} 0 & \text{if } \nabla \cdot \lambda = \mu - \nu \\ +\infty & \text{otherwise} \end{cases}.$$

Hence one can write (B) as

$$\min_\lambda M(\lambda) + \sup_\phi \int -\nabla\phi \cdot \, d\lambda + \int \phi d(\mu - \nu) = \sup_\phi \int \phi d(\mu - \nu) + \inf_\lambda M(\lambda) - \int \nabla\phi \cdot \, d\lambda,$$

where inf and sup have been exchanged formally as in the previous computations. After that one notices that

$$\inf_\lambda M(\lambda) - \int \nabla\phi \cdot \, d\lambda = \inf_\lambda \int d|\lambda| \left(1 - \nabla\phi \cdot \frac{d\lambda}{d|\lambda|}\right) = \begin{cases} 0 & \text{if } |\nabla\phi| \leq 1 \\ -\infty & \text{otherwise} \end{cases}$$

and this leads to the dual formulation for (B) which gives

$$\sup_{\phi\,:\,|\nabla\phi|\le 1}\int_\Omega \phi\, d(\mu-\nu).$$

Since this problem is exactly the same as (D) (a consequence of the fact that Lip_1 functions are exactly those functions whose gradient is smaller than 1), this gives the equivalence between (B) and (K).

Most of the considerations above, especially those on the problem (B) do not hold for costs other than the distance $|x-y|$. The only possible generalizations I know concern either a cost c which comes from a Riemannian distance $k(x)$ (i.e. $c(x,y)=\inf\{\int_0^1 k(\sigma(t))|\sigma'(t)|dt\,:\,\sigma(0)=x,\sigma(1)=y\}$, which gives a problem (B) with $\int k(x)d|\lambda|$ instead of $M(\lambda)$) or the fact that $p-$homogeneous costs may become $1-$homogeneous through the introduction of time as an extra variable (see [11]). We will see something more on the problem (B) in the lectures on models and applications.

1.4 $c(x,y)=h(x-y)$ with h strictly convex and the existence of an optimal T

We summarize here some useful results for the case where the cost c is of the form $c(x,y)=h(x-y)$, for a strictly convex function h.

The main tool is the duality result. If we have equality between the minimum of (K) and the maximum of (D) and both extremal values are realized, one can consider an optimal transport plan γ and a Kantorovitch potential ψ and write

$$\psi(x)+\psi^c(y)\le c(x,y) \text{ on } \Omega\times\Omega \text{ and } \psi(x)+\psi^c(y)=c(x,y) \text{ on } \operatorname{spt}\gamma.$$

The equality on $\operatorname{spt}\gamma$ is a consequence of the inequality which is valid everywhere and of

$$\int c\,d\gamma=\int \psi\,d\mu+\int \psi^c\,d\nu=\int(\psi(x)+\psi^c(y))d\gamma,$$

which implies equality $\gamma-$a.e. These functions being continuous, the equality passes to the support of the measure.

Once we have that, let us fix a point $(x_0,y_0)\in\operatorname{spt}\gamma$. One may deduce from the previous computations that

$$x\mapsto \psi(x)-h(x-y_0) \quad \text{ is minimal at } x=x_0$$

and, if $\psi$is differentiable at x_0, one gets $\nabla\psi(x_0)\in\partial h(x_0-y_0)$. For a stricltly convex function h one may inverse the relation passing to ∇h^* thus getting

$$x_0-y_0=\nabla h^*(x_0)=(\partial h)^{-1}(x_0).$$

This solves several questions concerning the transport problem with this cost, provided ψ is differentiable a.e. with respect to μ.Actually, one may use the previous computation to deduce that one point y_0 only may be coupled to x_0 in the support of γ (i.e. γ is of the form $(id\times T)_\#\mu$) and also to get uniqueness of the optimal trasport plan and of the grdient of the Kantorovitch potential.

We may summarize everything in the following theorem:

Theorem 1.4. *Given μ and ν probability measures on a domain $\Omega \subset \mathbb{R}^d$ there exists an optimal transport plan π. It is unique and of the form $(id \times T)_{\#}\mu$, provided μ is absolutely continuous. Moreover there exists also at least a Kantorovich potential ψ, and the gradient $\nabla\psi$ is uniquely determined $\mu-$a.e. (in particular ψ is unique up to additive constants, provided the density of μ is positive a.e. on Ω). The optimal transport map T and the potential ψ are linked by $T(x) = x - (\nabla h^*)(\nabla\psi(x))$. Moreover we have $\psi(x) + \psi^c(T(x)) = c(x, T(x))$ for $\mu-$a.e. x. Conversely, every map T which is of the form $T(x) = x - (\nabla h^*)(\nabla\psi(x))$ for a function $\psi \in \Psi_c(\Omega)$ is an optimal transport plan from μ to $T_{\#}\mu$.*

Remark 2. Actually, the existence of an optimal transport map is true under weaker assumptions: we can replace the condition of being absolutely continuous with the condition $\mu(A) = 0$ for any $A \subset \mathbb{R}^d$ such that $\mathcal{H}^{d-1}(A) < +\infty$ or with any condition which ensures that the non-differentiability set of ψ is negligible. In the theorem we used the Lipschitz behavior of $\psi \in \Psi_c$ and applied Rademacher Theorem, but $c-$concave functions are often more regular than only Lipschitz.

Remark 3. In Theorem 1.4 only the part concerning the optimal map T is not symmetric in μ and ν: hence the uniqueness of the Kantorovich potential is true even if it ν (and not μ) has positive density a.e. (since one can retrieve ψ from ψ^c and viceversa).

Remark 4. Theorem 1.4 may be particularized to the quadratic case $c(x, y) = |x - y|^2/2$, thus getting the existence of an optimal transport map

$$T(x) = x - \nabla\psi(x) = \nabla\left(\frac{x^2}{2} - \psi(x)\right) = \nabla\phi(x)$$

for a convex function ϕ. By using the converse implication (sufficient optimality conditions), this also proves the existence and the uniqueness of a gradient of a convex function transporting μ onto ν. This well known fact has been investigated first by Brenier (see [6]) and is often known as Brenier's Theorem.

Not only, a specific approach for the case $|x - y|^2$, based on the fact that we can withdraw the parts of the cost depending on x or y only and maximize $\int x \cdot y d\gamma$, gives the same result in a easier way: we actually get $\phi(x_0) + \phi^*(y_0) = x_0 \cdot y_0$ for a convx function ϕ and its Legendre transform ϕ^* and we deduce $y_0 \in \partial\phi(x_0)$.

All the costs of the form $c(x, y) = |x - y|^p$ with $p > 1$ fall under Theorem 1.4.

We finish the part dedicated to positive results by noticing that the same method may not be used if h is only convex, or at least does not give results as strong as what it does if h is strictly convex. Yet, there is anyway something which is known for the case $c(x, y) = |x - y|$. The results are a bit weaker (and much harder) and are summarized below (this is the classical Monge case and we refer to [3], even if several different proofs have been provided by different methods). Notice that a lot of literature is currently being dedicated to the case of other norms than the euclidean one and other distance functions.

Theorem 1.5. *Given μ and ν probability measures on a domain $\Omega \subset \mathbb{R}^d$ there exists at least an optimal transport plan π. Moreover, one of such plans is of the form $(id \times T)_{\#}\mu$ provided μ is*

absolutely continuous. There exists a Kantorovich potential ψ, and its gradient is unique $\mu-$a.e.and we have $\psi(x) - \psi(T(x)) = |x - T(x)|$ for $\mu-$a.e. x, for any choice of optimal T and ψ.

Here the absolute continuity assumption is essential to have existence of an optimal transport map, in the sense that in general it cannot be replaced by weaker assumptions as in the strictly convex case.

Morevoer, we can provide a counter-exemple showing that in general it is necessary that μ does not give mass to "small" sets.

Example 1. Set

$$\mu = \mathcal{H}^1 \llcorner A \quad \text{and} \quad \nu = \frac{\mathcal{H}^1 \llcorner B + \mathcal{H}^1 \llcorner C}{2}$$

where A, B and C are three vertical parallel segments in $\mathbb{R}^2$ whose vertexes lie on the two line $y = 0$ and $y = 1$ and the abscissas are 0, -1 and 1, respectively. It is clear that no transport plan may realize a cost better than $1/2$ since, horizontally, every point needs to displaced of a distance $1/2$. Moreover, one can get a sequence of maps $T_n : A \to B \cup C$ by dividing A into $2n$ equal segments $(A_i)_{i=1,\dots,2n}$ and B and C into n segments each, $(B_i)_{i=1,\dots,n}$ and $(C_i)_{i=1,\dots,n}$ (all ordered upwards). Then define T_n as a piecewise affine map which sends A_{2i-1} onto B_i and A_{21} onto C_i. In this way the cost of the map T_n is less than $1/2 + 1/n$, which implies that the infimum of the Kantorovitch problem is $1/2$, as well as the infimum on trasnport maps only. Yet, no map T may obtain a cost $1/2$, as this would imply that any point is sent horizontally and but this cannot respect the push-forward constraint. On the other hand, the transport plan associated to T_n weakly converge to the transport plan $1/2T^+_\# \mu + 1/2T^-_\# \mu$, where $T^\pm(x) = x \pm e$ and $e = (1, 0)$. This transport plan turns out to be the only optimal transport plan and has a Kantorovich cost of $1/2$.

Notice that the same construction provides also an example of the relaxation procedure leading from Monge to Kantorovich.

2 Wasserstein distances and spaces

Starting from the values of the problem (K) in (1.1) we can define a set of distances over $\mathcal{P}(\Omega)$. For any $p \geq 1$ we can define

$$W_p(\mu, \nu) = \big(\min(K) \text{ with } c(x, y) = |x - y|^p\big)^{1/p}.$$

We recall that, by Duality Formula, we have

$$\frac{1}{p} W_p^p(\mu, \nu) = \sup_{\psi \in \Psi_p(\Omega)} \int_\Omega \psi \, d\nu + \int_\Omega \psi^c \, d\mu. \tag{2.1}$$

Theorem 2.1. *If Ω is compact, for any $p \geq 1$ the function W_p is in fact a distance over $\mathcal{P}(\Omega)$ and the convergence with respect to this distance is equivalent to the weak convergence of probability measures. In particular any functional $\mu \mapsto W_p(\mu, \nu)$ is continuous with respect to weak topology.*

To prove that he convergence according to W_p is equivalent to weak convergence one first establish this result for $p = 1$, through the use of the duality with the functions in Lip_1. Then it is possible to use the inequalities between the distances W_p to extend the result to a general p.

The case of a noncompact Ω is a little more difficult. First, the distance must be defined only on a subset of the whole space of probability measures, to avoid infinite values. We will use the space of probabilities with finite $p-$th momentum:

$$\mathcal{W}_p(\Omega) = \{\mu \in \mathcal{P}(\Omega) \,:\, m_p(\mu) := \int_\Omega |x|^p \mu(dx) < +\infty\}.$$

Theorem 2.2. *For any $p \geq 1$ the function W_p is a distance over $\mathcal{W}_p(\Omega)$ and, given a measure μ and a sequence $(\mu_n)_n$ in $\mathcal{W}_p(\Omega)$, the following are equivalent:*

- *$\mu_n \to \mu$ according to W_p;*
- *$\mu_n \rightharpoonup \mu$ and $m_p(\mu_n) \to m_p(\mu)$;*
- *$\int_\Omega \phi \, d\mu_n \to \int_\Omega \phi \, d\mu$ for any $\phi \in C^0(\Omega)$ whose growth is at most of order p (i.e. there exist constants A and B depending on ϕ such that $|\phi(x)| \leq A + B|x|^p$ for any x).*

Notice that, as a consequence of Hölder (or Jensen) inequalities, the Wasserstein distances are always ordered, i.e. $W_{p_1} \leq W_{p_2}$ if $p_1 \leq p_2$. Reversed inequalities are possible only if Ω is bounded, and in this case we have, if set $D = \mathrm{diam}(\Omega)$, for $p_1 \leq p_2$,

$$W_{p_1} \leq W_{p_2} \leq D^{1-p_1/p_2} W_{p_1}^{p_1/p_2}.$$

From the monotone behavior of Wasserstein distances with respect to p it is natural to introduce the following distance W_∞: set $\mathcal{W}_\infty(\Omega) = \{\mu \in \mathcal{P}(\Omega) \,:\, \mathrm{spt}(\mu) \text{ is bounded } \}$ (obviously if Ω itself is bounded one has $\mathcal{W}_\infty(\Omega) = \mathcal{P}(\Omega)$) and then

$$W_\infty(\mu, \nu) = \inf \left\{ \gamma - esssup_{x,y \in \Omega \times \Omega} |x - y| \,:\, \gamma \in \Pi(\mu, \nu) \right\}.$$

It is easy to check that $W_p \nearrow W_\infty$ and it is interesting to study the metric space $\mathcal{W}_\infty(\Omega)$. Curiously enough, this supremal problem in optimal transport theory, even if quite natural, has not deserved much attention, up to the very recent paper [10].

The W_∞ convergence is stronger than any W_p convergence and hence also than the weak convergence of probability measures. The converse is not true and W_∞ converging turns out to be actually rare: consequently there is a great lack of compactness in $\mathcal{W}_\infty$. For instance it is not difficult to check that, if we set $\mu_t = t\delta_{x_0} + (1-t)\delta_{x_1}$, where $x_0 \neq x_1 \in \Omega$, we have $W_\infty(\mu_t, \mu_s) = |x_0 - x_1|$ if $t \neq s$. This implies that the balls $B(\mu_t, |x_0 - x_1|/2)$ are infinitely many disjoint balls in $\mathcal{W}_\infty$ and prevents compactness.

The following statement summarizes the compactness properties of the spaces W_p for $1 \leq p \leq \infty$ and its proof is a direct application of the considerations above and of Theorem 2.2.

Proposition 2.3. *For $1 \leq p < \infty$ the space $W_p(\Omega)$ is compact if and only if Ω itself is compact. Moreover, for an unbounded Ω the space $W_p(\Omega)$ is not even locally compact. The space $W_\infty(\Omega)$ is neither compact nor locally compact for any choice of Ω with $\#\Omega > 1$.*

3 Geodesics, continuity equation and displacement convexity

3.1 Metric derivatives in Wasserstein spaces

We are concerned in this sections with several properties linked to the curves in the Wasserstein space W_p. For this subject the main reference is [**?**]. Before giving the main result we are interested in, we recall the definition of metric derivative, which is a concept that may be useful when studying curves which are valued in generic metric spaces.

Definition 3. Given a metric space (X, d) and a curve $\gamma : [0, 1] \to X$ we define *metric derivative* of the curve γ at time t the quantity

$$|\gamma'|(t) = \lim_{s \to t} \frac{d(\gamma(s), \gamma(t))}{|s - t|}, \tag{3.1}$$

provided the limit exists.

As a consequence of Rademacher Theorem it can be seen (see [4]) that for any Lipschitz curve the metric derivative exists at almost every point $t \in [0, 1]$. We will be concerned quite often with metric derivatives of curves which are valued in the space $\mathcal{W}_p(\Omega)$.

Definition 4. If we are given a Lipschitz curve $\mu : [0, 1] \to W_p(\Omega)$, we define velocity field of the curve any vector field $v : [0, 1] \times \Omega \to \mathbb{R}^d$ such that for a.e. $t \in [0, 1]$ the vector field $v_t = v(t, \cdot)$ belongs to $[L^p(\mu_t)]^d$ and the *continuity equation*

$$\frac{d}{dt}\mu_t + \nabla \cdot (v \cdot \mu_t) = 0$$

is satisfied in the sense of distributions: this means that for all $\phi \in C^1_c(\Omega)$ and any $t_1 < t_2 \in [0, 1]$ it holds

$$\int_\Omega \phi \, d\mu_{t_2} - \int_\Omega \phi \, d\mu_{t_1} = \int_{t_1}^{t_2} ds \int_\Omega \nabla\phi \cdot v_s \, d\mu_s,$$

or, equivalently, in differential form:

$$\frac{\partial}{\partial t} \int_\Omega \phi \, d\mu_t = \int_\Omega \nabla\phi \cdot v_t \, d\mu_t \qquad \text{for a.e. } t \in [0, 1].$$

We say that v is the *tangent* field to the curve μ_t if, for a.e. t, v_t has minimal $[L^p(\mu_t)]^d$ norm for any t among all the velocity fields (actually this is not the true definition of a tangent vector field, since this would involve the definition of a tangent space for the "manifold" W_p, but it is in this case the same).

The following proposition is concerned with the existence of tangent fields and comes from Theorem 8.3.1 and Proposition 8.4.5 in [2].

Theorem 3.1. *If $p > 1$ and $\mu = (\mu_t)_t$ is a curve in $\mathrm{Lip}([0,1]; W_p(\Omega))$ then there exist unique a vector field v characterized by*

$$\frac{\partial}{\partial t}\mu + \nabla \cdot (v \cdot \mu) = 0, \tag{3.2}$$

$$||v_t||_{L^p(\mu_t)} \leq |\mu'|(t) \text{ for a.e. } t, \tag{3.3}$$

where the continuity equation is satisfied in the sense of distributions as previously explained. Moreover, if (3.2) holds for a family of vector fields $(v_t)_t$ with $||v_t||_{L^p(\mu_t)} \leq C$ then $\mu \in \mathrm{Lip}([0,1]; \mathcal{W}_p(\Omega))$ and $|\mu'|(t) \leq ||v_t||_{L^p(\mu_t)}$ for a.e. t.

To have an idea of the meaning of the previous theorem and of the relationship between curves of measures and the continuity equation some considerations could be useful.

Actually, at least when the vector fields v_t are regular enough, the solution of the continuity equation $\partial\mu/\partial t + \nabla \cdot (v \cdot \mu) = 0$, are obtained by taking the images of the initial measure μ_0 through the maps $\sigma(t, \cdot)$ obtained by taking the solution of

$$\begin{cases} \sigma'(t,x) = v_t(\sigma(t,x)), \\ \sigma(0,x) = x. \end{cases}$$

This explains why the vector field v_t is considered the velocity field of the curve μ_t: if every particle follows at each time t the vector field v_t, then the position of all the particles at time t reconstructs exactly the measure μ_t that appears in the continuity equation together with v_t !

Think for a while to the case of two time steps only: there are two measures μ_t and μ_{t+h} and there are several ways for moving the particles so as to reconstruct the latter from the former. It is exactly as when we look for a transport. One of these trasnports is optimal in the sense that it minimizes $\int |T(x) - x|^p \mu_t(dx)$ and the value of this integral equals $W_p^p(\mu_t, \mu_{t+h})$. If we call $v_t(x)$ the "discrete velocity of the particle located at x at time t, i.e. $v_t(x) = (T(x) - x)/h$, one has $||v_t||_{L^p(\mu_t)} = \frac{1}{h} W_p(\mu_t, \mu_{t+h})$. The result of the previous theorem may be easily guessed as obtainable as a limit as $h \to 0$.

3.2 Geodesics and geodesic convexity

Once we know about curves in their generality, it is interesting to think about geodesics. The following result is a characterization of geodesics in $W_p(\Omega)$ when Ω is a convex domain in $\mathbb{R}^d$. This procedure is also known as *McCann's linear interpolation*.

Theorem 3.2. *All the spaces $\mathcal{W}_p(\Omega)$ are length spaces and if μ and ν belong to $\mathcal{W}_p(\Omega)$, any geodesic curve linking them, when parametrized by arc-length, is of the form*

$$\mu^\gamma(s) = (p_s)_\# \gamma$$

where $p_s : \Omega \times \Omega \to \Omega$ is given by $p_s(x,y) = x + s(y - x)$ and γ is an optimal transport plan from μ to ν for the cost $c_p(x,y) = |x - y|^p$. In the case $p > 1$ and μ, ν absolutely continuous, if T is the corresponding optimal transport map such that $\gamma = (id \times T)_\# \mu$,then the curve has the form

$$\mu^\gamma(s) = [(1 - s)id + sT]_\# \mu.$$

Conversely, any curve of this form, for a transport plan γ or a transport map T, is an arc-length geodesic.

By means of this characterization of geodesics we can also define the useful concept of displacement convexity introduced by McCann in [13].

Definition 5. Given a functional $F : \mathcal{W}_p(\Omega) \cap L^1 \to [0, +\infty]$, we say that it is *displacement convex* if all the maps $t \mapsto F(\mu^\gamma(t))$ are convex on $[0, 1]$ for every choice of μ and ν in $\mathcal{W}_p(\Omega)$ and γ optimal transport plan from μ to ν with respect to $c(x, y) = |x - y|^p$.

The following well-known result provides a wide set of displacement convex functionals. In the case $p = 2$ this result is due to McCann ([13]), while the generalization to any p can be found in [2].

Theorem 3.3. *Consider the following functionals on the space $\mathcal{W}_p(\Omega)$, where Ω is any convex subset of $\mathbb{R}^N$:*

$$\begin{aligned} J^1(\mu) &= \begin{cases} \int_\Omega f(u(x))\,dx & \text{if } \mu = u \cdot \mathcal{L}^d \\ +\infty & \text{if } \mu \text{ is not absolutely continuous;} \end{cases} \\ J^2(\mu) &= \int_\Omega V(x)\,\mu(dx); \\ J^3(\mu) &= \int_\Omega \int_\Omega w(x-y)\mu(dx)\mu(dy). \end{aligned}$$

Suppose $f : [0, +\infty] \to [0, +\infty]$ is a convex and superlinear lower semicontinuous function with $f(0) = 0$, $V : \Omega \to [0, +\infty]$ and $w : \mathbb{R}^d \to [0, +\infty]$ are convex functions. Then the functionals J^2 and J^3 are displacement convex in $\mathcal{W}_p(\Omega)$ and the functional J^1 is displacement convex provided the map

$$r \mapsto r^d f(r^{-d})$$

is convex and non-increasing on $]0, +\infty[$.

4 Monge-Ampère equation and regularity

The final issue that I'll approach in these lecture notes will be concerned with some regularity properties of T and ψ (the optimal transport map and the Kantorovich potential, respectively) and their relations with the densities of μ and ν. We will consider only the quadratic case $c(x, y) = |x-y|^2/2$, because it is the one where more results have been proven. Very recent results for generic costs have been developed by Ma, Trudinger, Wang, Loeper, Figalli... They require some very rigid assumptions on the costs, so that, surprinsingly enough, the quadratic cost is one of the few power that satisfies the suitable hypotheses.

It is easy, just by a change-of-variables formula, to transform, in the case of regular maps and densities, the equality $\nu = T_\# \mu$ into the PDE $v(T(x)) = u(x)/|JT|(x)$, where u and v are the

densities of μ and ν and J denotes the determinant of the Jacobian matrix. Recalling that we may write $T = \nabla\phi$ with ϕ convex (Remark 4), we get the Monge-Ampère equation

$$M\phi = \frac{u}{v(\nabla\phi)}, \tag{4.1}$$

where M denotes the determinant of the Hessian

$$M\phi = \det H\phi = \det\left[\frac{\partial^2\phi}{\partial x_i\,\partial x_j}\right]_{i,j}.$$

This equation up to now is satisfied by $\phi = \frac{x^2}{2} - \psi$ in a formal way only. We define various notions of solutions for (4.1):

- we say that ϕ satisfies (4.1) in the Brenier sense if $(\nabla\phi)_{\#}u \cdot \mathcal{L}^d = v \cdot \mathcal{L}^d$ (and this is actually the sense to be given to this equation);
- we say that ϕ satisfies (4.1) in the Alexandroff sense if $H\phi$, which is always a positive measure for ϕ convex, is absolutely continuous and its density satisfies (4.1) a.e.;
- we say that ϕ satisfies (4.1) in the viscosity sense if it satisfies the usual comparison properties required by viscosity theory but restricting the comparisons to regular convex test functions (since M is in fact monotone just when restricted to positively definite matrices);
- we say that ϕ satisfies (4.1) in the classical sense if it is of class C^2 and the equation holds pointwise.

Notice that any notion except the first may be also applied to the more general equation $M\phi = f$, while the first one just applies to this specific transportation case. The results we want to use are well summarized in Theorem 50 of [15]:

Theorem 4.1. *If u and v are $C^{0,\alpha}(\Omega)$ and are both bounded from above and from below on the whole Ω by positive constants and Ω is a convex open set, then the unique Brenier solution ϕ of (4.1) belongs to $C^{2,\alpha}(\Omega) \cap C^{1,\alpha}(\overline{\Omega})$ and ϕ satisfies the equation in the classical sense (hence also in the Alexandroff and viscosity senses).*

Even if this precise statement is taken from [15], we just detail a possible bibliographical path to arrive at this result. It is not easy to deal with Brenier solutions, so the idea is to consider viscosity solutions, for which it is in general easy to prove existence by Perron's method. Then prove some regularity result on viscosity solutions, up to getting a classical solution. Then, once we have a classical convex solution to Monge-Ampère equation, this will be a Brenier solution too. Since this is unique (up to additive constants) we have got a regularity statement for Brenier solutions. We can find results on viscosity solutions in [7], [9] and [8]. In [7] some conditions to ensure strict convexity of the solution of $M\phi = f$ when f is bounded from above and below are given. In [9] for the same equation it is proved $C^{1,\alpha}$ regularity provided we have strict convexity. In this way the term $u/v(\nabla\phi)$ becomes a $C^{0,\alpha}$ function and in [8] it is proved $C^{2,\alpha}$ regularity for solutions of $M\phi = f$ with $f \in C^{0,\alpha}$.

References

[1] L. Ambrosio, Lecture Notes on Optimal Transport Problems, *Mathematical Aspects of Evolving Interfaces,* Springer Verlag, Berlin, Lecture Notes in Mathematics (1812), 1–52, 2003.

[2] L. Ambrosio, N. Gigli and G. Savaré, *Gradient flows in metric spaces and in the spaces of probability measures.* Lectures in Mathematics, ETH Zurich, Birkhäuser, 2005.

[3] L. Ambrosio and A. Pratelli. Existence and stability results in the L^1 theory of optimal transportation, in *Optimal transportation and applications*, Lecture Notes in Mathematics (CIME Series, Martina Franca, 2001) 1813, L.A. Caffarelli and S. Salsa Eds., 123-160, 2003.

[4] L. Ambrosio and P. Tilli, *Topics on analysis in metric spaces.* Oxford Lecture Series in Mathematics and its Applications (25). Oxford University Press, Oxford, 2004.

[5] M. Beckmann, A continuous model of transportation, *Econometrica* (20), 643–660, 1952.

[6] Y. Brenier, Décomposition polaire et réarrangement monotone des champs de vecteurs. (French) *C. R. Acad. Sci. Paris Sér. I Math.* (305), no. 19, 805–808, 1987.

[7] L. Caffarelli, A localization property of viscosity solutions to the Monge-Ampère equation and their strict convexity. *Ann. of Math.* (131), no. 1, 129–134, 1990.

[8] L. Caffarelli, Interior $W^{2,p}$ estimates for solutions of the Monge-Ampère equation. *Ann. of Math.* (131), no. 1, 135–150, 1990.

[9] L. Caffarelli, Some regularity properties of solutions of Monge Ampère equation. *Comm. Pure Appl. Math.* (44), no. 8-9, 965–969, 1991.

[10] T. Champion, L. De Pacale, P. Juutinen, The $\infty-$Wasserstein distance: local solutions and existence of optimal transport maps, *SIAM J. Math. An.* (40), no. 1,1–20, 2008.

[11] C. Jimenez, Optimisation de Problèmes de Transport, PhD thesis of Université du Sud-Toulon-Var, 2005.

[12] L. Kantorovich, On the transfer of masses. *Dokl. Acad. Nauk. USSR*, (37), 7–8, 1942.

[13] R. J. McCann, A convexity principle for interacting gases. *Adv. Math.* (128), no. 1, 153–159, 1997.

[14] G. Monge, Mémoire sur la théorie des déblais et de remblais, *Histoire de l'Académie Royale des Sciences de Paris*, 666–704, 1781.

[15] C. Villani. *Topics in Optimal Transportation.* Graduate Studies in Mathematics, AMS, 2003.

[16] C. Villani, *Optimal transport: Old and New*, Springer Verlag (Grundlehren der mathematischen Wissenschaften), 2008