PART ONE

Short Courses

1

Introduction to optimal transport theory

FILIPPO SANTAMBROGIO

Abstract

These notes constitute a sort of crash course in optimal transport theory. The different features of the problem of Monge–Kantorovitch are treated, starting from convex duality issues. The main properties of space of probability measures endowed with the distances W_p induced by optimal transport are detailed. The key tools connecting optimal transport and partial differential equations are provided.

Contents

AMS Subject Classification (2010): 00-02, 49J45, 49Q20, 35J60, 49M29, 90C46, 54E35
Keywords: Monge problem, linear programming, Kantorovich potential, existence, Wasserstein distances, transport equation, Monge–Ampère, regularity

1.1 Introduction

These very short lecture notes are not intended 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. Several of these ideas have been used, and explained in greater detail, during the other classes of the Summer School "Optimal Transportation: Theory and Applications" which were the occasion for the redaction of these notes. The style that was chosen when preparing them, in view of their use during the Summer School, was highly informal, and this revised version will respect the same style.

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 am also very indebted to the approach that L. Ambrosio used in a course at SNS Pisa in 2001–02 and I want to cite this 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

$$\int_A g(x)dx = \int_{T^{-1}(A)} f(y)dy \quad \text{for any Borel subset } A \subset \mathbb{R}^d \tag{1.1}$$

and minimizing the quantity

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

among all the maps satisfying this condition. This means that we have a collection of particles, distributed with density f on $\mathbb{R}^d$, that have to be moved, so that they arrange according to a new distribution, whose density is prescribed and is g. The movement has to be chosen so as to minimize the average displacement. The map T describes the movement (that we must choose in an optimal way), and $T(x)$ represents the destination of the particle originally located at x. The constraint on T precisely accounts for the fact that we need to reconstruct the density g. In the following, we will always define, similarly to (1.1), the image measure of a measure μ on X (measures will indeed replace the densities f and g in the most general formulation of the problem) through a measurable map $T : X \to Y$: it is the measure denoted by $T_\#\mu$ on Y and characterized

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? . . .) until the progress made in the 1940s. Indeed, only with the work by Kantorovich in 1942 has it been inserted into a suitable framework which gave the possibility to approach it and, later, to find that solutions actually exist 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.

1.2 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. $c(x, y) = c(y, x)$).

1.2.1 Kantorovich 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

$$(K) \quad \min \left\{ \int_{\Omega\times\Omega} c \; d\gamma \,|\gamma \in \Pi(\mu, \nu) \right\}, \tag{1.2}$$

where $\Pi(\mu, \nu)$ is the set of the so-called *transport plans*, i.e. $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\Omega \times \Omega) : (p^+)_{\#}\gamma = \mu, \; (p^-)_{\#}\gamma = \nu\}$, where p^+ and p^- are the two projections of $\Omega \times \Omega$ onto Ω. These probability measures over $\Omega \times \Omega$ are an alternative way to describe the displacement of the particles of μ: instead of saying, for each x, which is the destination $T(x)$ of the particle originally located at x, we say for each pair (x, y) how many particles go from x to y. It is clear that this description allows for more general movements, since from a single point x particles can a priori move to different destinations y. If multiple

destinations really occur, then this movement cannot be described through a map T. Notice that the constraints on $(p^{\pm})_{\#}\gamma$ exactly mean that we restrict our attention to the movements that really take particles distributed according to the distribution μ and move them onto the distribution ν.

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 $T : \Omega \to \Omega$ (i.e. when no splitting of the mass occurs), the map T would be called an *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 proposed by Monge, for instance, in the Monge case we would need existence of at least a map T satisfying the constraints. This is not verified when $\mu = \delta_0$, if ν is not a single Dirac mass. On the contrary, there always exists a 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 it 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 does exist.

As a consequence, if one is interested in the problem of Monge, the question may become "Does this minimum come from a transport map T?" Actually, if the answer to this question is yes, then it is evident that the problem of Monge has a solution, which also solves a wider problem, that of minimizing among transport plans. In some cases, proving that the optimal transport plan comes from a transport map (or proving that there exists at least one optimal plan coming from a map) is equivalent to proving that the problem of Monge has a solution, since very often the infimum among transport plans and among transport maps is the same. Yet, in the presence of atoms, this is not always the case, but we will not insist any more on this degenerate case.

1.2.2 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 thing we will do is find 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 one 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:

$$\begin{aligned} &\min_{\gamma} \int c \, d\gamma + \sup_{\phi,\psi} \left(\int \phi \, d\mu + \int \psi \, d\nu - \int (\phi(x) + \psi(y)) \, d\gamma \right) \\ &\quad = \sup_{\phi,\psi} \int \phi \, d\mu + \int \psi \, d\nu + \inf_{\gamma} \int (c(x, y) - (\phi(x) + \psi(y))) \, d\gamma. \end{aligned}$$

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 rewrite the inf in γ as a constraint on ϕ and ψ, since one has

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

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

$$(D) \quad \max \left\{ \int_\Omega \phi \, d\mu + \int_\Omega \psi \, d\nu \;\middle|\; \phi \in L^1(\mu), \psi \in L^1(\nu) \,:\, \phi(x)+\psi(y) \leq c(x, y) \text{ for all } (x, y) \in \Omega\times\Omega \right\}. \quad (1.3)$$

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 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), and the constraints are preserved and the integrals increased. 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 perfectly 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*

$$\min(K) = \max_{\psi\in\Psi_c(\Omega)} \int_\Omega \psi \, d\mu + \int_\Omega \psi^c \, d\nu, \quad (1.4)$$

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 (1.4) 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 Proposition 1.1), 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. Notice that the case where c is a power of the distance is actually 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 to the cost $c(x, y) = |x - y|^p/p$. It is not difficult to check that

$$\psi \in \Psi_{(1)}(\Omega) \Longleftrightarrow \psi \text{ is a 1-Lipschitz function;}$$
$$\psi \in \Psi_{(2)}(\Omega) \Longrightarrow \ x \mapsto \frac{x^2}{2} - \psi(x) \text{ is a convex function;}$$
$$\text{if } \Omega = \mathbb{R}^d \text{ this is an equivalence.}$$

1.2.3 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, (1.4) may be rewritten as*

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

The key point of Proposition 1.2 is proving $\psi^c = -\psi$. This is easy 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 behavior of ψ).

Another peculiar feature of this case is the following:

Proposition 1.3. *Consider the problem*

$$(B) \quad \min\left\{M(\lambda) \;\middle|\; \lambda \in \mathcal{M}^d(\Omega);\ \nabla \cdot \lambda = \mu - \nu\right\}, \tag{1.5}$$

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* $\phi \in C^1(\overline{\Omega})$*. If* Ω *is convex then it holds*

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

This proposition links the Monge–Kantorovich problem to a minimal flow problem which was first proposed by Beckmann [5], under the name of *continuous transportation model*. He did not know this link, as Kantorovich's theory was being developed independently almost in the same years. In Section 2.1 we will see some more details 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 rewrite 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

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

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