NWI-WM246 - OPTIMAL TRANSPORT LECTURE NOTES

RICCARDO CRISTOFERI

Contents

1. Introduction

1.1. **What these notes are.** These are the notes for the class *NWI-WM246 Optimal Transport* I first taught in Spring 2021. They are intended as a guide for the students who attended the class and they are specifically designed for them and for the specific path I chose to take on the subject. Of course, they do not aim neither to be a complete introduction on the subject, nor to substitute any book on Optimal Transport (see next section).

Being this an introductory class to the subject, I intend to present the main ideas and proofs. Therefore, sometimes results are not stated in the full generality they hold. The reason that I want students to focus on the ideas rather than on technical details needed to adapt the idea to the most general setting. Grasping the ideas will then allow the interested students to continue their path into digging more in the theory and applications of Optimal Transport.

The course is basically self-contained, in the sense that all the results needed are at least stated in the notes. Classical results in Analysis and in Measure Theory are not proven, but references are given for where to find the proofs. Some results are only stated, but the proofs were given as exercise to the students.

A final note. The style in which these notes are written is that of a dialogue with the reader. The reason being that they are designed for students (or whoever) that encounters Optimal Transport for the first time and wants to get familiar with the ideas and techniques of this field. They are intended to be read and thought over, not to be only quickly looked to search for a result. Finally, communicating by writing has its own advantages and drawbacks. If on the one hand it is possible to include more details and expand the material presented during the live classes (to be fair, livestramed, since you know, pandemic!), on the other hand communicating by writing makes harder to present comments and ideas that could be better conveyed by using a chalk, a blackboard, and the most important ingredient of all for facilitating understanding: human interaction!

1.2. **Guide to the literature.** There are several references for Optimal Transport available, each with a specific audience and goal in mind. I write here some that I know about (being this not a list of the best ones whatsoever - except for the AGS: that's the best if you are interested in gradiet flows!):

Theoretical aspects

- Ambrosio, *Lecture notes on Optimal Transport problems*, [1]
- Ambrosio, Gigli *A user's guide to optimal transport*, [2]
- Ambrosio, Gigli, Savarè, *Gradient flows in metric spaces and in the space of probability measures*, [3]
- Bourne *A brief introduction to optimal transport theory*, [4]
- Evans, *Partial Differential Equations and Monge-Kantorovich mass transport* [8]
- McCann, Guillen, *Five lectures on Optimal Transportation: geometry, regularity and applications*, [15]
- Santambrogio, *Optimal Transport for Applied Mathematicians*, [17]
- Villani, *Topics in Optimal Transportation*, [19]
- Villani, *Optimal Transport: Old and New*, [20]

Computational aspects

- Peyré, Cuturi, *Computational optimal transport*, [16]

Applications

- Buttazzo, Santambrogio, *A Mass transportation model for the optimal planning of an urban region*, [6]
- Carlier, *Optimal Transport and Economic applications*, [7]
- Galichon, *Optimal Transport Methods in Economics*, [11]
- Peyré, Cuturi, *Computational optimal transport*, [16]
- Santambrogio, *Optimal Transport for Applied Mathematicians*, [17]

1.3. **The problem of Monge.** The question thanks to which we are all here studying Optimal Transport dates back to 1781, when Gaspard Monge was interested in understanding how to move a collection of particle to one specific configuration to another by minimizing the average displacement of the particles. Written in (modern) mathematical terms the problem of Monge can be stated like this: consider two continuous densities $f, g : \mathbb{R}^N \to [0, \infty)$, where $N = 2, 3$, with

$$\int_{\mathbb{R}^N} f\,dx = \int_{\mathbb{R}^N} g\,dx.$$

Find a map $T : \mathbb{R}^N \to \mathbb{R}^N$ that minimizes

$$\int_{\mathbb{R}^N} |T(x) - x| f(x)\,dx \tag{1.1}$$

among those such that

$$\int_A g(y)\,dy = \int_{T^{-1}(A)} f(x)\,dx \tag{1.2}$$

for any Borel set $A \subset \mathbb{R}^N$. The reason why condition (1.2) is written in this way is in order to take into consideration the possibility that the map T is not injective (see Figure 1).

If f and g satisfies (1.2) we write $g = T_\# f$. The minimum problem we need to solve is thus

$$\min\left\{ \int_{\mathbb{R}^N} |T(x) - x| f(x)\,dx \;:\; T : \mathbb{R}^N \to \mathbb{R}^N \text{ with } g = T_\# f \right\}. \tag{1.3}$$

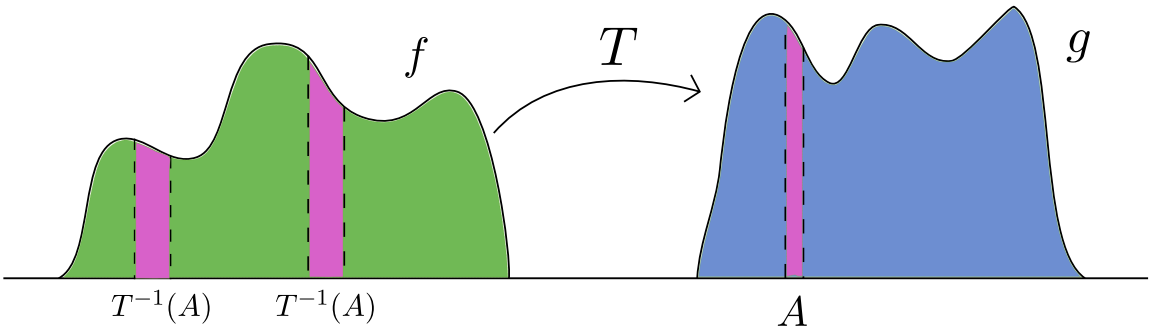

FIGURE 1. The Monge problem: the intial configuration on the left described by the map f has to be transformed through the map T into the final configuration on the right described by the map g. This is ensures if the pink area of g over a region A on the right is the same as the pink area of f over $T^{-1}(A)$.

Few comments are in order: for each $x \in \mathbb{R}^N$, the value $T(x)$ describes the position to which we move the mass at x, while the *energy functional* in (1.1) measures the average displacement of particles. Note that this is weighted by the initial density f, since *heuristically*, at the point $x \in \mathbb{R}^N$ there is $f(x)$ amount of particles. Finally, condition (1.2) ensures that the displacement T we use actually moves particles from the configuration given by f to that given by g.

This problem looks like an innocent one, and thus, despite many people studied it and investigated fine properties of solutions (even if they did not know they existed!), it might come as a surprise that it took more than 150 years to prove that, at least in certain cases, it has a solution.

The original problem of Monge can be generalized as follows. The quantity $|T(x)-x|$ quantifies the *cost* needed to move a unit of mass from position x to position $T(x)$. We can write it as

$$c(x, T(x)) = |T(x) - x|\,,$$

and write (1.3) as

$$\int_{\mathbb{R}^N} c(x, T(x)) f(x)\, dx\,.$$

With this writing it is not surprising to consider more general costs $c : \mathbb{R}^N \times \mathbb{R}^N \to [0, +\infty]$, where the interpretation of the quantity $c(x, y)$ is the cost to move a unitary mass from position $x \in \mathbb{R}^N$ to position $y \in \mathbb{R}^N$. Note that the function c needs not to be symmetric, namely $c(x, y) \neq c(y, x)$: if x is at the bottom of a mountain and y is at the top, you realize that it makes sense to consider also non-symmetric costs. Moreover, in this formulation, the cost does not care to the specifics of *how* we move the mass from x to y, and *in what order* we move each *particle* in order to transform the density f into the density g.

1.4. **Some examples.** In this section we present some examples that highlight interesting situations that can arise with the generalization of the Monge problem. Before going into the parade of the examples, let us present a test for optimality for the original problem of Monge, namely for the cost $c(x, y) = |x - y|$. In this case it holds that

$$\begin{aligned}\inf \left\{ \int_{\mathbb{R}^N} |T(x) - x| f(x)\, dx \,:\, T : \mathbb{R}^N \to \mathbb{R}^N \text{ with } g = T_{\#} f \right\} \\ \geq \sup \left\{ \int_{\mathbb{R}^N} u(x)\, (g(x) - f(x))\, dx \,:\, u \in \mathrm{Lip}_1(\mathbb{R}^N) \right\}, \end{aligned} \tag{1.4}$$

where $\mathrm{Lip}_1(\mathbb{R}^N)$ denotes the space of Lipschitz maps $u : \mathbb{R}^N \to \mathbb{R}$ with Lipschitz constants 1:

$$\sup_{x \neq y} \frac{|u(x) - u(y)|}{|x - y|} \leq 1\,.$$

The constrain (1.2) implies that

$$\int_{\mathbb{R}^N} u(x) g(x)\, dx = \int_{\mathbb{R}^N} u(T(x)) f(x)\, dx$$

and therefore

$$\int_{\mathbb{R}^N} u(x)\, (g(x) - f(x))\, dx = \int_{\mathbb{R}^N} [u(T(x)) - u(x)]\, f(x)\, dx \leq \int_{\mathbb{R}^N} |T(x) - x| f(x)\, dx\,,$$

where the last inequality follows by the 1-Lipschitzianty of u.

Example 1 - Book shifting. Consider the densities (see Figure 2)

$$f(x) := \begin{cases} 1 & \text{if } x \in [0, 2], \\ 0 & \text{else}\,, \end{cases} \qquad g(x) := \begin{cases} 1 & \text{if } x \in [1, 3], \\ 0 & \text{else}\,, \end{cases}$$

Among the several ways that there are in order to transform the green rectangle into the blue one, let us consider two:

(i) Shifting everything by 1 to the right:

$$T_1(t) := t + 1\,;$$

(ii) Shifting only the non-overlapping part by 2 to the right:

$$T_2(t) := \begin{cases} t + 2 & \text{if } t \in [0, 1], \\ t & \text{else}\,. \end{cases}$$

We also consider three costs, corresponding to three important families of costs that will be further analyzed during the course:

$$c_{\frac{1}{2}}(x, y) := |x - y|^{\frac{1}{2}}\,, \qquad c_1(x, y) := |x - y|\,, \qquad c_2(x, y) := |x - y|^2\,.$$

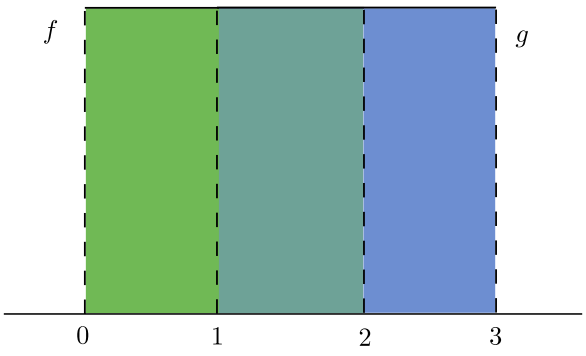

FIGURE 2. The book shifting: we want to transform the green rectangle into the blue one.

We want to compare the *energy* of T_1 of T_2 with respect to the three costs. We have that

$$\int_0^2 c_{\frac{1}{2}}(x, T_1(x))\,dx = 2\,, \qquad \int_0^2 c_{\frac{1}{2}}(x, T_2(x))\,dx = \sqrt{2}\,,$$
$$\int_0^2 c_1(x, T_1(x))\,dx = 2\,, \qquad \int_0^2 c_1(x, T_2(x))\,dx = 2\,,$$
$$\int_0^2 c_2(x, T_1(x))\,dx = 2\,, \qquad \int_0^2 c_2(x, T_2(x))\,dx = 4\,.$$

We see that there is a difference in behaviour in the three cases. In particular, moving the common mass (the part where the green and the blue rectangles intersect) is not convenient for $c_{\frac{1}{2}}$, not convenient for c_2, and does not matter for c_1. This follows from the fact that, if $x < z < y$,

$$|x-y|^{\frac{1}{2}} < |x-z|^{\frac{1}{2}} + |z-y|^{\frac{1}{2}},$$
$$|x-y| = |x-z| + |z-y|,$$
$$|x-y|^2 > |x-z|^2 + |z-y|^2.$$

Moreover, note that the optimality criterion (1.4) implies that both T_1 and T_2 are optimal for the cost c_1.

Example 2 - Concentration of mass. We introduce informally the notion of a mass concentrated in a point (this will be made rigorous later by using Measure Theory). Given a point $x \in \mathbb{R}$ and $m > 0$, we denote by $m\delta_x$ a mass m concentrated at the point x. Different masses m_i's concentrated at different points x_i's can be written by $\sum_{i=1}^k m_i \delta_{x_i}$.

Consider the density f as above and suppose that we want to concentrate all the mass at the extreme points of the interval $[0, 2]$ (see Figure 3). By using the language introduced above, we consider the 'density' g given by

$$g := \delta_0 + \delta_2.$$

In this case it is easy to see that the optimal map for all the three costs above is given by

$$T(x) := \begin{cases} 0 & \text{if } x \in [0,1], \\ 2 & \text{if } x \in (1,2]. \end{cases}$$

Example 3 - Non existence. Consider now the case where g is as above, f is $2\delta_1$, two unities of mass concentrated at the point $x = 1$ (see Figure 4).

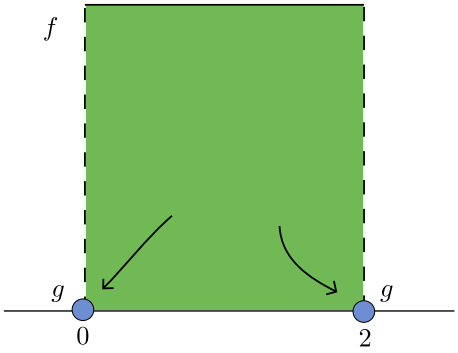

FIGURE 3. Mass concentration: we want to to concentrate the mass of the green rectangle in the two blue points.

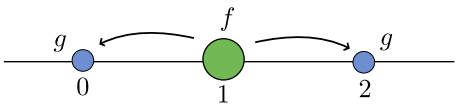

FIGURE 4. In this case we want to *split* the two unities of mass at the point $x = 1$ (green circle) and send each unity to one of the two blue points.

In this case *there is no map that satisfies the constrain!!!*. Indeed, it $T : \mathbb{R} \to \mathbb{R}$ is a map, then $T(1)$ will be just *one* of the two points $x = 0$ and $x = 2$ where g is concentrated. In this case we would like to *split* the mass ad move one unity of mass to the point $x = 0$ and the other unity of mass to the point $x = 2$. But maps *do not allow* for such a splitting!

Example 4 - All are the best. We now present a degenerate case, where all admissible maps are optimal. We go in the plane $\mathbb{R}^2$, and consider the following situation (see Figure 5): g is concentrated in two points $A := (-1, 0), B := (1, 0) \in \mathbb{R}^2$, namely $g = \delta_A + \delta_B$, and f is concentrated in the vertical segment

$$\{(x, y) \in \mathbb{R}^2 \, : \, x = 0, \, y \in [-1, 1]\}.$$

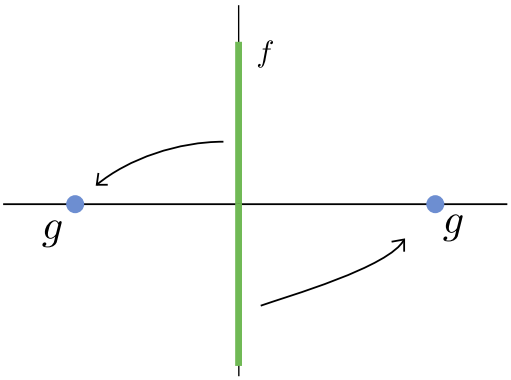

FIGURE 5. The degenerate example: we want to move the mass from the green line into the two blue points.

Then, *any* admissible map is optimal for all of the three costs above!

1.5. **Difficulties in solving the problem.** Why is the problem of Monge so difficult? Well, let's try to solve it!

Before tackling the Monge problem, let us take inspiration from a simpler case. Suppose you are given a *continuous* function $F : \mathbb{R} \to \mathbb{R}$, and a *compact* set $K \subset \mathbb{R}$. Consider the minimization problem

$$\min\{F(x) \ : \ x \in K\}\,.$$

The set K plays the role of a *constrain*. Solving this problem means to find a point $\bar{x} \in K$ such that

$$F(\bar{x}) = \inf\{F(x) \ : \ x \in K\}\,.$$

Note that the inf of a set is always defined. What you (should) do to solve the problem is the following:

(Step 1.) Consider a minimizing sequence $\{x_n\}_{n\in\mathbb{N}} \subset K$; namely a sequence such that

$$\lim_{n\to\infty} F(x_n) = \inf\{F(x) \ : \ x \in K\}\,;$$

(Step 2.) Since K is compact, there exists a subsequence $\{x_{n_i}\}_{i\in\mathbb{N}}$ and a point $\bar{x} \in K$ such that $x_{n_i} \to \bar{x}$ (with respect to the Euclidean topology);

(Step 3.) Since F is continuous

$$F(\bar{x}) = \lim_{i\to\infty} F(x_{n_i}) = \lim_{n\to\infty} F(x_n) = \inf\{F(x) \ : \ x \in K\}$$

and thus we solved the problem!

The above strategy is nothing but the proof of the so called *Weierstraß's Theorem*.

Let us try to use the same strategy also in our case. Let us denote by X the space of (Borel) maps from $\mathbb{R}^N$ to $\mathbb{R}^N$. The constrain will then be

$$K \coloneqq \{T \in X \ : \ (1.2) \text{ holds } \}\,.$$

The functional $F : X \to [0,\infty]$ will then be

$$F(T) \coloneqq \int_{\mathbb{R}^N} |T(x) - x| f(x)\, dx\,,$$

and the problems writes as

$$\min\{F(T) \ : \ T \in K\}\,.$$

To run the above scheme, we first have to answer the following questions:

(i) Is there a notion of convergence that makes K compact?
(ii) Can we ensure that a minimizing sequence is (pre)-compact in that topology?
(iii) Is the functional F continuous with respect to that topology?

First of all, we start by rewriting condition (1.2) as a pointwise condition rather than an integral one as follows: fix a point $\overline{y} \in \mathbb{R}^N$ and consider, for $\varepsilon > 0$, the ball $B(\overline{y}, \varepsilon)$ centered at $\overline{y}$ and with radius ε. Assume the map $T : \mathbb{R}^N \to \mathbb{R}^N$ to be of class C^1 and injective, and let $\overline{x} \coloneqq T^{-1}(\overline{y})$. We can change variable in the integral on the left-hand side of (1.2) and get

$$\int_{T^{-1}(B(\overline{y},\varepsilon))} g(T(x)) \det(DT(x))\, dx = \int_{T^{-1}(B(\overline{y},\varepsilon))} f(x)\, dx.$$

We now want to send $\varepsilon \to 0$ in the above equality. All the integrands are continuous, and thus they will shrink to the integrand computed at the limiting point of $T^{-1}(B(\overline{y},\varepsilon))$ that is, by the injectivity of the map T and of the definition of $\overline{x}$, the point $\overline{x}$ itself. Therefore we get that (by dropping the bar over x)

$$g\left(T(x)\right) \det(DT(x)) = f(x) \tag{1.5}$$

for all $x \in \mathbb{R}^N$. We can now rewrite the constrain (note: under the very strong assumptions that T is injective and C^1!, but let's forget about that) as

$$K \coloneqq \{T \in X \, : \, (1.5) \text{ holds } \}.$$

The problem is the following. Take a minimizing sequence $\{T_n\}_{n\in\mathbb{N}}$ and assume that

$$\inf\{F(T) \ : T \in K\} < \infty.$$

Thus, what we know is that

$$\sup_{n\in\mathbb{N}} F(T_n) < \infty.$$

Let us answer question (ii): by using the above bound, can we say that the sequence $\{T_n\}_{n\in\mathbb{N}}$ is (pre)-compact in some topology? Yes, we can! It is compact in the weak*-L^∞ topology. In particular, there exists a subsequence $\{T_{n_i}\}_{i\in\mathbb{N}}$ and $T \in X$ (note, not in K!) such that T_n *to*T with respect to the weak*-L^∞ topology. Good! Now let us go to question (iii): is the functional F continuous with respect to this topology? No, but it holds that

$$F(T) \leq \liminf_{i\to\infty} F(T_{n_i})$$

and this suffices for our purposes to find a minimum. Good! Just one question left to answer: is K compact with respect to the weak*-L^∞ topology? Unfortunately, the answer is no!!! The reason is that equation (1.5) is highly non-linear and the weak*-L^∞ topology behaves nicely with linear equations. Thus, despite we found a limiting map T such that

$$F(T) \leq \lim_{i\to\infty} F(T_{n_i}) = \inf\{F(T) \ : T \in K\},$$

we cannot conclude that T is the desired minimizer, since *we are not sure that it belongs to the class we are minimizing F on.*

So is the problem condemned not to have a solution? It turns out that we were asking, in general, the wrong question! The right framework that allows to get a solution to the (natural generalization) of the Monge problem is that of*measures*. These mathematical objects have two fundamental properties of interest for us:

- They allow a rigorous definition of splitting mass;
- They possess the good compactness properties that maps fail to have.

The right framework for the problem was discovered by Kantorovich in 1942 (see [14]) who proved existence of a generalized solution. Brenier in 1987 (see [5]) proved that in the case the costs is $c(x, y) = |x - y|^2$ then a generalized solution is a solution of the Monge problem (namely a map T). This result was extended to more general costs functionals of the form $c(x, y) = h(y - x)$ with h *strictly* convex by Gangbo and McCann in 1996 (see [12]). Finally, the linear cost initially considered by Monge was address by Sudakov in [18], and his proof was later fixed by Ambrosio in [1]. For more general costs it is difficult to tell whether or not there exists a generalized solution that is a map.

1.6. **Why it is interesting to study this problem?** After having read the previous pages, you might ask yourself: 'Well, nice mathematical problem. But why should I care about it?'. Granted that everybody as their own taste on what they like, let me answer the question by appealing to the importance of the problem.

From the mathematical point of view, the above problem has challenges generations of mathematicians that tried to prove the existence of a solution, and lately to investigate fine properties of the (generalized) solutions. The quest of solving the problem has led to the development of interesting mathematical results. In particular, it has been used in connection with functional and geometrical inequalities, nonlinear Partial Differential Equations, and Dynamical Systems.

From the interdisciplinary point of view (again from the point of view of research), Optimal Transport has been used to find a framework that allows to interpret several equations in Physics as gradient flows (namely evolutions that decreases 'as fast as possible' a certain potential). This has the advantage of providing more structure to the equations allows to have more tools to study stable equilibria, and to design numerical approximations. In particular, in recent years mathematicians have been able to provide a different point of view on the equations of General Relativity. This is ongoing research!, and we are all thrilled to see what this will lead to.

From the point of view of applications of the theory, Optimal Transport is widely used by practitioners in a wide range of problems:

- In vision theory: registration and segmentation of images, histograms balance;
- In economics: equilibria of supplies and demands, maximization of profit, social welfare, urban planning;
- In engineering: optimal shape, material design, aerodynamic resistance;
- In atmosphere and ocean dynamics: semigeostrophic equation;
- In biology: irrigation, leaf growth;
- In big data: clustering of the data set
- In machine learning: generative models, image recognition, neural networks.

I am positive that you are excited in taking this journey in Optimal Transport!

2. A quick tour in Measure Theory

In this section we are going to give an overview on the main ideas of Measure Theory that we will need to study the Optimal Transport problem. The basic results are only stated. For more on Measure Theory refer to classical references such as [9, 13, 10]. An important example of (outer) measure is the Lebesgue measure: we will introduce it and present its main properties.

2.1. Measures and outer measures.

The definition of a measure generalizes the heuristic idea we have of *natural* measures such as length, area, volume: a function that assigns a (positive) number to each set of a space and that is additive on disjoint sets.

We first introduce the family of objects a measure is defined on. For a set X we denote by 2^X the class of subsets of X.

Definition 2.1. Let X be a set. A family of sets $\mathcal{A} \subset 2^X$ is called an *algebra* if

(i) $\emptyset \in \mathcal{A}$;
(ii) if $E_1, E_2 \in \mathcal{A}$, then $E_1 \cap E_2 \in \mathcal{A}$;
(iii) if $E \in \mathcal{A}$, then $X \setminus E \in \mathcal{A}$.

Remark 2.2. It is easy to see that if $\mathcal{A}$ is an algebra on X, then

(i) $E_1, E_2, \dots, E_k \in \mathcal{A}$, then $E_1 \cap E_2 \cap \cdots \cap E_k \in \mathcal{A}$;
(ii) $E_1, E_2, \dots, E_k \in \mathcal{A}$, then $E_1 \cup E_2 \cup \cdots \cup E_k \in \mathcal{A}$.

Example 2.3. Let X be a set and let $\mathcal{A} \subset 2^X$ be the collection of sets that are finite or have finite complement. Then $\mathcal{A}$ is an algebra.

Definition 2.4. Let X be a set and $\mathcal{A}$ be an algebra on it. A function $\mu : \mathcal{A} \to [0, +\infty]$ is called a *positive finitely additive measure* on the algebra $\mathcal{A}$ if

(i) $\mu(\emptyset) = 0$;
(ii) for every $E_1, E_2 \in \mathcal{A}$ with $E_1 \cap E_2 = \emptyset$ it holds $\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2)$.

Remark 2.5. Property (i) is clearly essential to have a definition that makes sense. Moreover, the notion of measure does not require any structure on the space X where it is defined (namely it does not have to be a topological space, metric space, vector space).

It is easy to see that, for every $k \in \mathbb{N}$ and every pairwise disjoint sets $E_1, \dots, E_k \in \mathcal{A}$ (namely $E_i \cap E_j = \emptyset$ for any $i \neq j$), then

$$\mu(E_1 \cup \cdots \cup E_k) = \mu(E_1) + \cdots + \mu(E_k).$$

holds for finitely additive measures μ.

Example 2.6. Let X be a set and let $\mathcal{A} \subset 2^X$ be the collection of sets that are finite or have finite complement. Then $\mathcal{A}$ is an algebra. Define $\mu : \mathcal{A} \to [0, \infty]$ by

$$\mu(E) := \min\{\#E, \#(X \setminus E)\},$$

where with $\#A$ we denote the number of elements of the set A. Then μ is a finitely additive measure.

Property (ii) of Definition 2.4 holds also for finitely many sets. But want more! Namely we would like to say something about the measure of the union of *as many* sets as we can obtain by adding one set after the other. This *as many* turns out to be *countably many*. This is why we need to enlarge the family of objects where such a measure is defined.

Definition 2.7. A family of sets $\mathcal{A} \subset 2^X$ is called a *σ-algebra* if it is an algebra and

$$\{E_i\}_{i \in \mathbb{N}} \subset \mathcal{A} \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} E_i \in \mathcal{A}$$

holds. In this case we say that $(X, \mathcal{A})$ is a *measure space*.

Example 2.8. Note that the collection of sets $\mathcal{A}$ introduced in Example 2.3 is *not* a sigma algebra. An important example of σ-algebra is the following. Let $X = \mathbb{R}$ and consider the family $\mathcal{E}$ of subsets of $\mathbb{R}$ of the form $[a, \infty)$, for some $a \in \mathbb{R}$. What is the smallest algebra that contains $\mathcal{E}$? And the smallest σ-algebra that contains $\mathcal{E}$?

Definition 2.9. Let $(X, \mathcal{A})$ be a measure space. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a *positive countably additive measure* on the measure space $(X, \mathcal{A})$ if

(i) $\mu(\emptyset) = 0$;
(ii) for every countable collection $\{E_i\}_{i\in\mathbb{N}} \subset \mathcal{A}$ of pairwise disjoint sets it holds

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i).$$

In this case, with an abuse of terminology, we say that $(X, \mathcal{A}, \mu)$ is a *measure space*.

Remark 2.10. Note that countably additivity is the best we can hope for in order to get something that makes sense. Indeed, if we were to ask for

$$\mu\left(\bigcup_{i\in I} E_i\right) = \sum_{i\in I} \mu(E_i) \tag{2.1}$$

for *any* family of indexes I, even more than countable, we would have something nonsense. Indeed consider the measure μ to be the area (vaguely defined) and the unit cube $Q = (0, 1)^2$ in the plane. From (2.1) we would get

$$1 = \mu(Q) = \mu\left(\bigcup_{q\in Q} \{q\}\right) = \sum_{q\in Q} \mu(q) = 0,$$

where the last equality follows from the fact that point $q \in Q$ has zero area.

Example 2.11. Let X be a set and $\mathcal{A} = 2^X$. Define the *counting measure* $\mu : \mathcal{A} \to [0, \infty]$ by $\mu(E) \coloneqq \#E$. Then μ is a positive countably additive measure.

Unfortunately (or luckly, if you enjoy things like the Banach-Tarski paradox!), the notions of length, area, and volume properly defined (by using the notion of Lebesgue measure) are not additive measures in the sense of Definition 2.9. The reason being that there exists sets that make the additivity requirement to fail. These are usually pathological sets, and construct them require a bit of effort.

Definition 2.12. Let X be a set. We say that $\mu : 2^X \to [0, \infty]$ is an *outer measure* if

(i) $\mu(\emptyset) = 0$;
(ii) for every $E \subset X$ and every countable family $\{E_i\}_{i\in\mathbb{N}} \subset 2^X$ (not necessarily pairwise disjoint) with $E \subset \bigcup_{i=1}^{\infty} E_i$ it holds

$$\mu(E) \leq \sum_{i=1}^{\infty} \mu(E_i).$$

Remark 2.13. Countably additive measures are outer measures. Moreover, note that, *since* μ *takes values in* $[0, \infty]$, $\mu(A) \leq \mu(B)$ if $A \subset B$.

The most important example of out measure is the Lebesgue measure $\mathcal{L}^N$ in dimension N: this is the rigorous definition of *length* in one dimension, *area* in two dimensions, and *volume* in three dimensions. Here we introduce it by using the Hausdorff construction (see Proposition 2.25 for a similar idea). The main idea is the following: we know what number to assign as the length of a segment, the area of a square, the volume of a cube, and so on. The number we decide to assign to a general set is that resulting by the *best* covering of the set by segments, squares, cubes, and so on. We will see later how this is connected to the product of Lebesgue measure in one dimension.

Definition 2.14. For $x \in \mathbb{R}^N$ and $r > 0$, let

$$Q_r(x) \coloneqq \{y \in \mathbb{R}^N \,:\, |x_i - y_i| < r/2 \text{ for all } i = 1, \dots, N\}$$

denote the open cube centered at x with radius r. For $E \subset \mathbb{R}^N$ we define the *Lebesgue outer measure* of E by

$$\mathcal{L}^N(E) \coloneqq \inf \left\{ \sum_{i=1}^{\infty} r_i^N \,:\, E \subset \bigcup_{i=1}^{\infty} Q_{r_i}(x_i), x_i \in \mathbb{R}^N, r_i \geq 0 \right\}.$$

Only because we called it a measure does not make it a measure. The name is, nevertheless, appropriate.

Proposition 2.15. *The function $\mathcal{L}^N : 2^X \to [0, \infty]$ is an outer measure.*

Given an outer measure, there are two classes of sets that are interesting: negligible sets, and measurable sets.

Definition 2.16. Let $\mu : 2^X \to [0, \infty]$ be an outer measure. We say that a set $E \subset X$ is *μ-negligible* if $\mu(E) = 0$.

There is an easy way to determine whether a set is negligible.

Lemma 2.17. *Let $\mu : 2^X \to [0, \infty]$ be an outer measure and $E \subset X$. The following are equivalent*

(i) E is μ-negligible;
(ii) For every $\varepsilon > 0$ there exists a family $\{A_i\}_{i=1}^{\infty}$ with $E \subset \bigcup_{i=1}^{\infty} A_i$ such that

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) < \varepsilon.$$

Definition 2.18. Let $\mu : 2^X \to [0, \infty]$ be an outer measure. We say that a property holds for μ-almost every $x \in X$ (μ-a.e. in X) if it holds for all $x \in X \setminus N$, where $N \subset X$ is μ-negligible.

The second class of sets we introduce finds its reason to be considered when one tries to answer the following question: given an outer measure μ on X, is it possible to restrict it to a σ-algebra for which it turns out to be a measure, namely for which it is countably additive on pairwise disjoint families of that σ-algebra? The answer is yes! and the class of sets for which (countable) additivity holds can be easily characterized as follows.

Definition 2.19. Let μ be an outer measure on X. A set $E \subset X$ is called *μ-measurable* if

$$\mu(F) = \mu(E \setminus F) + \mu(E \cap F)$$

for each $F \subset X$.

Remark 2.20. The name *measurable* can be misleading: an outer measure is defined for *all* subsets of X. Those who behave in a good way are called μ-measurable.

Theorem 2.21 (Carathéodory)**.** *Let X be a set and let $\mu : 2^X \to [0, \infty]$ be an outer measure. Then the family of sets*

$$\mathcal{M}_\mu \coloneqq \{\, E \subset X \,:\, E \text{ is } \mu\text{-measurable} \,\}$$

is a σ-algebra and

$$\{E \in X : \mu(E) = 0\} \subset \mathcal{M}_\mu.$$

Moreover $\mu : \mathcal{M}_\mu \to [0, \infty]$ is a (countably additive) measure.

Remark 2.22. In particular, Theorem 2.21 states that $\mathcal{M}_\mu$ is the greatest σ-algebra where the restriction of μ is a measure. This allows, given on outer measure μ, to consider its *natural* σ-algebra $\mathcal{M}_\mu$.

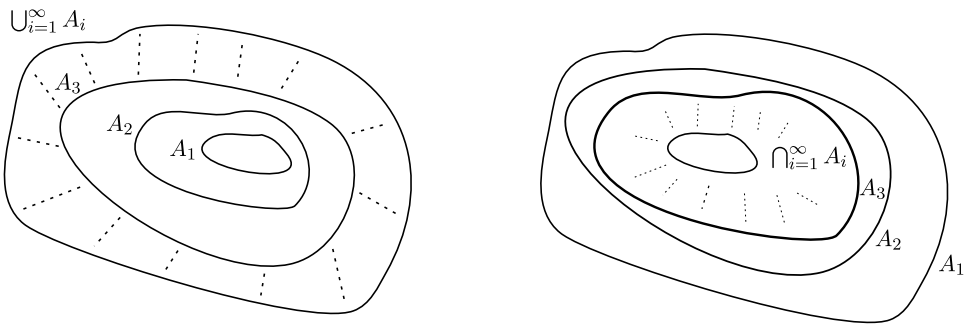

FIGURE 6. The limiting sets for increasing and decreasing sequences of sets.

Terminology. When we will say that (X, μ) is a *measure space* we will mean that μ is an outer measure on X and we consider its restriction to the σ-algebra of μ-measurable sets, where μ is a coutably additive measure.

On measurable sets it is possible to prove continuity properties for the outer measure μ (see Figure 6).

Lemma 2.23. *Let $\mu : 2^X \to [0, \infty]$ be an outer measure. Let $\{A_i\}_{i\in\mathbb{N}} \subset X$ be a sequence of μ-measurable sets. Then:*

(i) If $A_1 \subset A_2 \subset A_3 \subset \dots$, then

$$\lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right);$$

(ii) If $A_1 \supset A_2 \supset A_3 \supset \dots$ and $\mu(A_1) < \infty$, then

$$\lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcap_{i=1}^{\infty} A_i\right).$$

Remark 2.24. Note that in both cases on the right-hand side we have the measure of the *limiting* set. In the second case the assumption $\mu(A_1) < \infty$ is crucial. Consider the case $X = \mathbb{R}$, $\mu = \mathcal{L}^1$, and the decreasing sequence of sets $A_i := (i, \infty)$. Then $\mu(A_i) = \infty$ for each $i \in \mathbb{N}$, but since $\bigcap_{i\in\mathbb{N}} A_i = \emptyset$, we have that $\mu(\bigcap_{i\in\mathbb{N}} A_i) = 0$.

Theorem 2.21 allows to get a measure from an outer measure by restricting it to the family of measurable sets. Is it possible to do the opposite, namely to *extend* a measure on a σ-algebra to an outer measure?

Proposition 2.25. *Let $(X, \mathcal{A}, \mu)$ be a measure space. Define the function $\mu^* : 2^X \to [0, \infty]$ by*

$$\mu^*(E) := \inf\{\, \mu(F) \,:\, E \subset F,\, F \in \mathcal{A} \,\}.$$

for any $E \subset X$. Then μ^* is an outer measure and $\mu^*(E) = \mu(E)$ for every $E \in \mathcal{A}$.

Remark 2.26. Note that the idea of Proposition 2.25 is similar to the one we used to introduce the Lebesgue measure. In the latter case though, the family of cubes is *not* a σ-algebra (not even an algebra!). The construction used in Proposition 2.25 is known as Hausdorff construction.

2.2. **Regularity properties of measures.** A measure on a set X can be wild. If the set X has an additional structure, like of a topological or of a metric space, it is possible to consider classes of measures that behave in a good way.

Definition 2.27. Let $\mu : 2^X \to [0, \infty]$ be an outer measure. We say that μ is *regular* if for each set $E \subset X$ there exists a μ-measurable set $F \supset E$ such that $\mu(E) = \mu(F)$.

For regular measures the continuity properties that hold for measurable sets hold for any family of sets.

Lemma 2.28. *Let $\mu : 2^X \to [0,\infty]$ be an outer measure. Let $\{A_i\}_{i\in\mathbb{N}}$ be a sequence of sets (not necessarily μ-measurable). Then*

(i) If $A_1 \subset A_2 \subset A_3 \subset \dots$, then

$$\lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right);$$

(ii) If $A_1 \supset A_2 \supset A_3 \supset \dots$ and $\mu(A_1) < \infty$, then

$$\lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcap_{i=1}^{\infty} A_i\right).$$

We now consider the case where the set X has additional structure: topological space and metric space. In both cases there is a σ-algebra related to the topology and a σ-algebra related to the μ-measurable sets. It is interesting when the first one is contained in the latter.

Definition 2.29. Let (X,τ) be a topological space. The *Borel σ-algebra* is the smallest σ-algebra that contains the open sets. Elements of the Borel σ-algebra are called *Borel* sets.

Definition 2.30. Let (X,τ) be a topological space, and $\mu : 2^X \to [0,\infty]$ be an outer measure on X. We say that μ is:

(ii) *Borel* if every Borel set is μ-measurable;

(ii) *Borel regular* if it is a Borel outer measures, and for every $E \subset X$ there exists a Borel set $B \subset X$ with $E \subset B$ such that $\mu(E) = \mu(B)$;

There is a nice characterization of Borel measures in metric spaces.

Theorem 2.31 (Carathèodory criterion)**.** *Let (X,d) be a metric space and let $\mu : 2^X \to [0,\infty]$ be an outer measure. Then μ is Borel if and only if*

$$\mu(A\cup B) = \mu(A) + \mu(B)$$

for every $A, B \subset X$ with $\mathrm{d}(A,B) := \inf\{\,\mathrm{d}(x,y) \,:\, x\in A, y\in B\,\} > 0$

Remark 2.32. Note that the Carathèodory criterion requires to check the additivity of μ only on sets that are more than disjoint: they have to be *far apart* with respect to the distance d. This is usually simpler because for measure that are defined by using the Hausdorff construction we have space between the two sets A and B.

By using the Carathèodory criterion it is possible to see that the Lebesgue measure $\mathcal{L}^N$ is a Borel measure.

Borel regular measures allow to restrict to specific classes of sets, namely Borel sets, when computing the measure of a general set. In many applications it is sufficient to be able to *approximate* the measure of a general set by using classes of sets that enjoy desirable properties.

Definition 2.33. Let (X,τ) be a topological space and let $\mu : 2^X \to [0,\infty]$ be an outer measure on X. A set $E \subset X$ is said to be

(i) *Inner regular* if

$$\mu(E) = \sup\{\,\mu(K) \,:\, K\subset E, K \text{ compact}\,\}$$

(ii) *Outer regular* if

$$\mu(E) = \inf\{\,\mu(A) \,:\, E\subset A, A \text{ open}\,\}$$

A class of measures that plays a central role in Measure Theory is that of Radon measures.

Definition 2.34. Let (X,τ) be a topological space and let $\mu : 2^X \to [0,\infty]$ be an outer measure on X. We say that μ is a *Radon outer measure* if

(i) μ is a Borel regular outer measure;
(ii) $\mu(K) < \infty$ for every compact set $K \subset X$;
(iii) Every open set $A \subset X$ is inner regular;
(iv) Every set $E \subset X$ is outer regular.

Proposition 2.35. *The Lebesgue measure $\mathcal{L}^N$ is a Radon measure.*

Actually, the inner regularity holds for a larger class of sets. We first need to introduce a special class of sets that is needed when one tries to prove properties for sets with infinite measure.

Definition 2.36. A set $A \subset X$ is said to be *σ-finite* for the outer measure $\mu : 2^X \to [0,\infty]$ if there exists a family $\{F_i\}_{i\in\mathbb{N}}$ with $\mu(F_i) < \infty$ for all $i \in \mathbb{N}$, such that

$$E = \bigcup_{i=1}^{\infty} F_i \,.$$

We say that the outer measure μ is *σ-finite* if X is σ-finite.

Remark 2.37. The Lebesgue measure $\mathcal{L}^N$ is σ-finite. Indeed one can consider the sequence $F_i := B(0,i)$.

Lemma 2.38. *Let (X,τ) be a topological space and let $\mu : 2^X \to [0,\infty]$ be an outer Radon measure on X. Then every σ-finite μ-measurable set $E \subset X$ is inner regular.*

The relation between Borel regular and Radon outer measures is the following.

Proposition 2.39. *Let (X,τ) be a topological space and let $\mu : 2^X \to [0,\infty]$ be an outer measure on X. The followings hold:*

(i) If μ is a Radon outer measure, then it is a Borel regular;
(ii) Assume that X is a locally compact Hausdorff space that can be written as a countable union of compact sets. Moreover, assume that μ is a Borel outer measure that is finite on compact set. Then μ is a Radon outer measure.

2.3. **Integration and limiting theorems.** We now turn to the definition of integral by following the idea of Lebesgue's integration. In all of this section X will be a set and $\mu : 2^X \to [0,\infty]$ an outer measure. Moreover, we will denote by $\bar{\mathbb{R}}$ the extended reals $\mathbb{R} \cup \{\pm\infty\}$.

Definition 2.40. Let (Y,d) be a metric space. A function $f : X \to Y$ is said to be *μ-measurable* if $f^{-1}(A) \in \mathcal{M}_\mu$ for all open sets $A \subset Y$.

Remark 2.41. The definition of measurability is not that strange. Recall indeed that a function $g : Z \to Y$, where (Z,τ) is a topological space, is continuous if $g^{-1}(A) \in \tau$ for each open set $A \subset Y$.

The class of μ-measurable functions is closed under the most important operations among functions.

Proposition 2.42. *Let $\mu : 2^X \to \mathbb{R}$ be an outer measure. Then*

(i) If $f, g : X \to [0,\infty]$ are μ-measurable, then so are

$$f+g, \qquad \min\{f,g\}, \qquad \max\{f,g\}\,,$$

and also f/g, provided $g > 0$;
(ii) If $f_i : X \to \mathbb{R}$ are μ-measurable for each $i \in \mathbb{N}$, then also

$$\inf_{i\in\mathbb{N}} f_i, \qquad \sup_{i\in\mathbb{N}} f_i, \qquad \liminf_{i\to\infty} f_i, \qquad \limsup_{i\to\infty} f_i$$

are μ-measurable.

Remark 2.43. The above result is typically used to prove that a function is μ-measurable because it an be obtained as a combination of the above stated operations (sum, inf, sup,...) of functions that are more easily checked to be μ-measurable.

An important result that relates measurability and continuity is the following.

Theorem 2.44 (Lusin's Theorem). *Let $\mu : 2^X \to [0,\infty]$ be a Borel regular outer measure, and $f : X \to \mathbb{R}^M$ be a μ-measurable function. Fix a μ-measurable set $A \subset X$ with $\mu(A) < \infty$, and $\varepsilon > 0$. Then there exists a compact set $K \subset A$ with $\mu(A \setminus K) < \varepsilon$ and $f : K \to \mathbb{R}^M$ is continuous (in the relative topology of K).*

We now start introducing the objects needed to define the integral of a function. As for the Riemann integral, the Lebesgue integral of a function will be obtained as a limiting procedure by approximating the function with simple ones, for which we know what the integral should be. The difference between the two integrals is that the Lebesgue's integral allows for a larger class of simple functions.

Definition 2.45. Given a set $E \subset X$, we define the *characteristic function* of E as

$$\mathbb{1}_E(x) \coloneqq \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{if } x \notin E. \end{cases}$$

Remark 2.46. In some books, the characteristic function of a set E is denoted by χ_E.

Definition 2.47. We say that a μ-measurable function $f : X \to \mathbb{R}$ is *simple* if its image is finite, namely if it is possible to write

$$f(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i$$

for al $x \in X$, where $k \in \mathbb{N}$, $E_i \subset X$, and $y_1, \dots, y_k \in Y$.

Remark 2.48. Note that it is always possible to write a simple function $f : X \to \mathbb{R}$ as

$$f(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i$$

where the sets $E_1, \dots, E_k$ are pairwise disjoint.

We now define the notion of Lebesgue integral.

Definition 2.49 (integral of a positive simple function)**.** Let $f : X \to \mathbb{R}$ be a simple μ-measurable function

$$f(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i \,.$$

We define the *(Lebesgue) integral* of f with respect to μ by

$$\int_X f \, d\mu \coloneqq \sum_{i=1}^{k} y_i \mu(E_i) \,,$$

with the convention that if $y_i = 0$ and $\mu(E_i) = \infty$, then $y_i \mu(E_i) = 0$.

Remark 2.50. The μ-measurability is needed in order to have a well-defined object. Indeed, if

$$f(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i = \sum_{j=1}^{m} \mathbb{1}_{F_j}(x) z_j$$

we would like

$$\sum_{i=1}^{k} y_i \mu(E_i) = \sum_{i=j}^{m} z_j \mu(F_j) \,,$$

for *any* choice of the sets F_j's. This is precisely requiring that the sets E_i's are μ-measurable.

Definition 2.51. Given a function $f : X \to \mathbb{R} \cup \{\infty\}$ we define its *positive* and *negative* part by

$$f^+ := \max\{f, 0\}, \qquad f^- := \max\{-f, 0\},$$

respectively.

Remark 2.52. Note that $f^+, f^- \geq 0$, and that $f = f^+ - f^-$, $|f| = f^+ + f^-$.

Definition 2.53 (integral of a generic positive function)**.** Given a μ-measurable function $f : X \to [0, \infty]$, we define the *(Lebesgue) integral* of f with respect to μ by

$$\int_X f \, d\mu := \sup \left\{ \int_X g \, d\mu \, : \, g \text{ simple, } \mu\text{-measurable, } g \leq f \right\}.$$

Definition 2.54 (integral of a generic function)**.** Let $f : X \to \overline{\mathbb{R}}$ be a μ-measurable function. Assume that f is *μ-integralble*, namely that

$$\int_X f^+ \, d\mu < \infty, \qquad \text{or} \qquad \int_X f^- \, d\mu < \infty. \tag{2.2}$$

We define the *(Lebesgue) integral* of f with respect to μ by

$$\int_X f \, d\mu := \int_X f^+ \, d\mu - \int_X f^- \, d\mu.$$

Remark 2.55. Assumption (2.2) is in order to avoid $+\infty - \infty$ in the definition of the integral.

Definition 2.56. We say that a μ-integrable function $f : X \to \overline{\mathbb{R}}$ belongs to the space $L^1(X; \mu)$, if

$$\int_X |f| \, d\mu < \infty.$$

The Lebesgue integral satisfies some basic properties.

Lemma 2.57. *Let $f, g : X \to \overline{\mathbb{R}}$ be μ-integrable. Then*

$$\int_X (af + bg) \, d\mu = a \int_X f \, d\mu + b \int_X g \, d\mu.$$

for all $a, b \in \mathbb{R}$. Moreover, if $f \leq g$ μ-a.e., then

$$\int_X f \, d\mu \leq \int_X g \, d\mu.$$

Finally, if $f = g$ μ-a.e., then

$$\int_X f \, d\mu = \int_X g \, d\mu.$$

Discussion: The Lebesgue and the Riemann integral In the first analysis classes you have been introduced to the notion of Riemann integration. From the point of view of the construction, the two differ from the fact that Riemann integral requires to partition the *domain*, while the Lebesgue one requires to partition the *target space* (see Figure 7). In particular, the simple functions used to define the Riemann integral are a subset of the simple functions used to define the Lebesgue's one. This translates to the fact that Riemann integrability requires a strong regularity of the function.

Theorem 2.58 (Riemann-Lebesgue Theorem)**.** *A function $f : \mathbb{R}^N \to \mathbb{R}^M$ is Riemann integrable if and only if the set of discontinuities of f has Lebesgue measure zero.*

Lebesgue integral allows to consider also functions that are discontinuous everywhere. As an example, let us consider the characteristic function of the irrational numbers $\mathbb{1}_{\mathbb{R}\setminus\mathbb{Q}}$. This function is *not* Riemann integrable, but is it easily seen to be Lebesgue integrable, and

$$\int_{(a,b)} \mathbb{1}_{\mathbb{R}\setminus\mathbb{Q}} \, dx = b - a$$

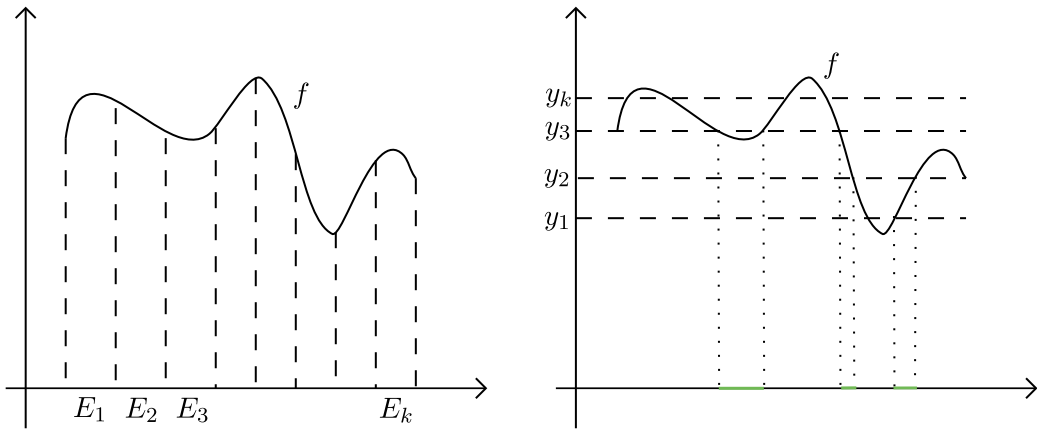

FIGURE 7. The paradigm shift from the Riemann integral (on the left) and the Lebesgue's one (on the right). For the Riemann integral we partition the domain in set $E_1, \dots, E_k$ and that determines what values to assign to each set in the partition. For the Lebesgue integral, instead, we partition the domain space $y_1, \dots, y_k$ and those value determines the sets on which we assign each of those value. For instance, in the figure on the right, the set E_3 (depicted in green) is the set where f is in between y_2 and y_3.

for every $a, b \in \mathbb{R}$ with $a < b$.

Lebesgue integration extends the notion of Riemann integration. Indeed, the two notion of integrations agree on the set of Riemann integrable functions.

Lemma 2.59. *If $f : \mathbb{R}^N \to \mathbb{R}^M$ is Riemann integrable the it is Lebesgue integrable and the two integrals coincide.*

Why is there the need for the notion of Lebesgue integration? Despite the notion of Riemann integral might seem more intuitive from the geometrical point of view, it is very limited from the point of view of applications, since important limiting theorem requires very restrictive assumptions.

Theorem 2.60 (Continuity of the Riemann integral)**.** *Let $f_n : [a, b] \to \mathbb{R}$ be a sequence of Riemann integrable functions. Assume that $f_n \to f$ uniformly. Then f is Riemann integrable and*

$$\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b \lim_{n\to\infty} f_n(x)\,dx = \int_a^b f(x)\,dx\,.$$

Uniform continuity is a very strong assumption and it is very often not satisfied in applications, where weaker information on the asymptotic behaviour of the sequence of functions are known. The following three results address important situations. They are not valid for the Riemann integral. We start by investigating the case of a general sequence of functions, for which the pointwise limit does not need to exist. Nevertheless, we have the following bounds for the sequence of integrals.

Theorem 2.61 (Fatou's lemma)**.** *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of μ-measurable functions. If $f_n \geq g$ for all $n \in \mathbb{N}$, where $g \in L^1(X;\mu)$, then*

$$\int_X \liminf_{n\to\infty} f_n\,d\mu \leq \liminf_{n\to\infty} \int_X f_n\,d\mu\,.$$

If $f_n \leq g$ for all $n \in \mathbb{N}$, where $g \in L^1(X;\mu)$, then

$$\limsup_{n\to\infty} \int_X f_n\,d\mu \leq \int_X \limsup_{n\to\infty} f_n\,d\mu.$$

Remark 2.62. It could be that the above inequalities are strict even if $\lim_{n\to\infty} f_n$ exists.

A special case is when the sequence $\{f_n\}_{n\in\mathbb{N}}$ is monotone, since in that case the pointwise limit $\lim_{n\to\infty} f_n(x)$ exists for all $x \in X$.

Theorem 2.63 (Lebesgue's Monotone convergence theorem)**.** *Let $\{f_n\}_{n\in\mathbb{N}}$ be an increasing sequence of μ-measurable functions such that $f_n \geq g$ with $g \in L^1(X,\mu)$. Then*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X \lim_{n\to\infty} f_n \, d\mu.$$

Finally, if the pointwise limit of a sequence of functions is known, but the sequence is not monotone, we wonder whether or not this translates into convergence of the integrals of that sequence. Next important result gives a positive answer under very mild assumptions.

Theorem 2.64 (Lebesgue's dominated convergence theorem)**.** *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of μ-measurable functions such that*

$$f_n(x) \to f(x)$$

for μ-a.e. $x \in X$. Assume that

$$|f_n| \leq g$$

where $g \in L^1(X;\mu)$ Then $f \in L^1(X;\mu)$ and

$$\lim_{n\to\infty} \int_X |f_n - f| \, d\mu = 0.$$

In particular,

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu.$$

2.4. **Product measure.** We now consider a construction that generalizes one of the natural ways in which the Lebesgue measure in higher dimensions can be defined (see Figure 8).

Definition 2.65 (Product measure)**.** Let (X,μ) and (Y,ν) be a measures spaces. We define the *product measure* $\mu\otimes\nu$ on $X\times Y$ by

$$\mu\otimes\nu(C) \coloneqq \inf\left\{ \sum_{i=1}^{\infty} \mu(A_i)\nu(B_i) \,:\, C \subset \bigcup_{i=1}^{\infty} (A_i\times B_i)\,, A_i \in \mathcal{M}_\mu,\, B_i \in \mathcal{M}_\nu \right\},$$

for every set $C \subset X\times Y$.

Remark 2.66. It is easy to see that $\mu\otimes\nu(A\times B) = \mu(A)\nu(B)$ for every $A \ni \mathcal{M}_\mu$ and $B \in \mathcal{M}_\nu$.

The idea of the above formula is the following: starting from the basic knowledge of the measure of product sets $A\times B$, we approximate each set $C \subset X\times Y$ with a countable union of product sets and we define its measure as the infimum over these approximations. This operation comes naturally if you think about what you do in order to compute the area of figures in the plane: the only *primitive* measure you know if the one of rectangles (objects of the for $[a,b]\times[c,d]$); then, the are of any region of the plane is obtained as limit of finer and finer approximations of it by means of rectangles. Think about how you learned in high-school on how to compute the area of a circle by using the so called *method of exhaustion*.

The Lebesgue measure introduced in Definition 2.14 can be also be seen as a product of measures.

Proposition 2.67. *For each $N \in \mathbb{N}$ it holds that $\mathcal{L}^N = \underbrace{\mathcal{L}^1\otimes\mathcal{L}^1\otimes\cdots\otimes\mathcal{L}^1}_{N \text{ times}}$.*

We want to write the integration with respect to the product measure $\mu\otimes\nu$ in terms of integration with respect to μ and ν. For

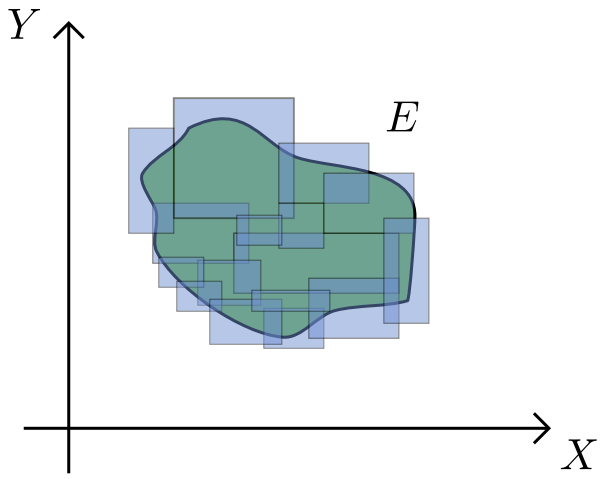

FIGURE 8. The product measure of a set $E \subset X \times Y$ is obtained by approximating E with product sets as efficiently as possible from the point of view of the product of the measures.

Theorem 2.68 (Tonelli's theorem)**.** *Let (X, μ) and (Y, ν) be two measure spaces that are σ-finite. Let $f : X \times Y \to [0, \infty]$ be a $\mu \otimes \nu$-measurable function. Then for μ-a.e. $x \in X$ the map $y \mapsto f(x, y)$ is ν-measurable, and the map*

$$x \mapsto \int_Y f(x, y)\, d\nu(y)$$

is μ-measurable. Moreover for ν-a.e. $y \in Y$ the map $x \mapsto f(x, y)$ is μ-measurable,

$$y \mapsto \int_X f(x, y)\, d\mu(y)$$

is ν-measurable. Finally

$$\int_{X\times Y} f\, d(\mu \otimes \nu) = \int_X \left[\int_Y f(x, y)\, d\nu(y) \right] d\mu(x) = \int_Y \left[\int_X f(x, y)\, d\mu(x) \right] d\nu(y).$$

For functions non-necessarily positive, we have the following result.

Theorem 2.69 (Fubini's theorem)**.** *Let (X, μ) and (Y, ν) be two outer measures. Let $f : X \times Y \to [-\infty, \infty]$ be a $\mu \otimes \nu$-integrable function. Then for μ-a.e. $x \in X$ the map $y \mapsto f(x, y)$ is ν-integrable, and the map*

$$x \mapsto \int_Y f(x, y)\, d\nu(y)$$

is μ-measurable. Moreover for ν-a.e. $y \in Y$ the map $x \mapsto f(x, y)$ is μ-integrable,

$$y \mapsto \int_X f(x, y)\, d\mu(y)$$

is ν-measurable. Finally

$$\int_{X\times Y} f\, d(\mu \otimes \nu) = \int_X \left[\int_Y f(x, y)\, d\nu(y) \right] d\mu(x) = \int_Y \left[\int_X f(x, y)\, d\mu(x) \right] d\nu(y).$$

Remark 2.70. A deep theorem in measure theory states that every reasonable measure can be written as a generalized product of measures, if we allow the measures in the product to depend on the point $x \in X$. This is the so called *disintegration theorem*.

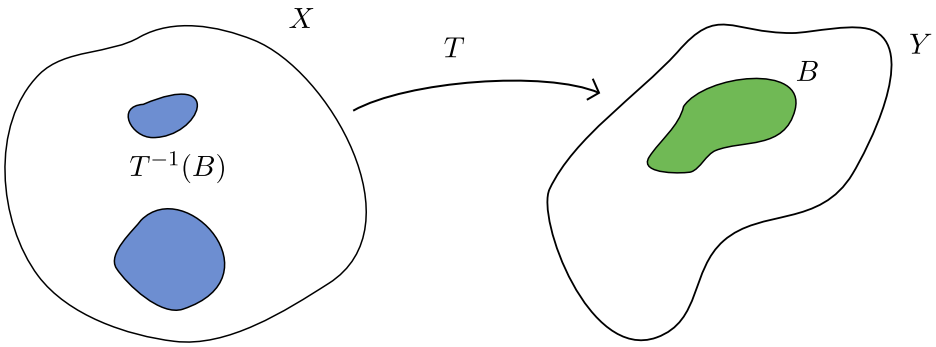

FIGURE 9. The push-forward measure $T_{\#}\mu$.

2.5. **The space of measures.**

2.5.1. *Operations with measures.* In this section we investigate ways to construct new measures by using existing ones. The easiest way is to exploit the fact that the spaces of (outer) measures over a set X is a vector space.

Lemma 2.71. *Let $\mu, \lambda : 2^X \to [0,\infty]$ be outer measures, and $a, b \geq 0$. Then the function $a\mu + b\lambda : 2^X \to [0,\infty]$ defined by*

$$(a\mu + b\lambda)(E) \coloneqq a\mu(E) + b\lambda(E)$$

for each $E \in 2^X$ is an outer measure on X. It is a measure on the family of sets that belong to both $\mathcal{M}_\mu$ and $\mathcal{M}_\nu$. Moreover, every regularity property shared by μ and λ is inherited by $a\mu + b\lambda$.

Next construction will be used repetitively during the course. It is a was to obtain a measure as a counter-image of another measure via a function (see Figure 9).

Definition 2.72 (Push-forward)**.** Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{M})$ be measure spaces. Let $T : X \to Y$ be a map such that $T^{-1}(E) \in \mathcal{A}$ for every $E \in \mathcal{M}$. Then we define the *push-forward* $T_{\#}\mu : \mathcal{M} \to [0,\infty]$ by

$$T_{\#}\mu(B) \coloneqq \mu(T^{-1}(B)),$$

for each $B \in Y$.

Lemma 2.73. *The function $T_{\#}\mu : \mathcal{M} \to [0,\infty]$ is a measure. Moreover, if $f : Y \to [0,\infty]$ is $T_{\#}\mu$-integrable, then $f \circ T$ is μ-integrable and*

$$\int_Y f \, dT_{\#}\mu = \int_X f \circ T \, d\mu \,.$$

It is useful to know how to recover the measures μ and ν (the *marginals*, in the language of probability) from the product measure $\mu \otimes \nu$.

Lemma 2.74. *Let μ and ν be measures on X and Y respectively, with $\mu(X) = \nu(Y) = 1$. Let $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ be the natural projects on X and Y respectively, namely $\pi_1(x,y) \coloneqq x$, $\pi_2(x,y) \coloneqq y$. Then*

$$(\pi_1)_{\#}(\mu \otimes \nu) = \mu \qquad\qquad (\pi_2)_{\#}(\mu \otimes \nu) = \nu \,.$$

Finally we see how to combine a measure and a function to define a new measure.

Definition 2.75. Let $\mu : 2^X \to [0,\infty]$ be an outer measure space and let $f : X \to [0,\infty]$ be μ-integrable. We define the function $f\mu : \mathcal{M}_\mu \to [0,\infty]$ by

$$(f\mu)(E) \coloneqq \int_E f \, d\mu$$

for $E \in \mathcal{M}_\mu$.

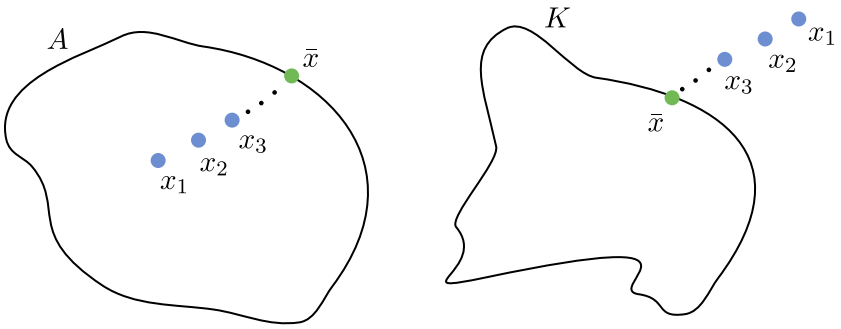

FIGURE 10. The two situations where the first attempt to define a notion of convergence for measures fails: on the left the case of an open set where we loose mass to the boundary in the limit; on the right the case of a compact set where we gain mass at the boundary in the limit.

Lemma 2.76. *The function $f\mu : \mathcal{M}_\mu \to [0,\infty]$ is a measure whose σ-algebra of measurable sets coincides with that of μ. Moreover, the functionals $f \mapsto f\mu$ and $\mu \mapsto f\mu$ are linear.*

2.5.2. *Convergence of measures.* We would like to talk about convergence of measures. A first attempt could be this: we say that a sequence of measures $\{\mu_n\}_{n\in\mathbb{N}}$ converges to the measure μ if $\mu_n(E) \to \mu(E)$ for all sets $E \subset X$. This is too optimistic! Let us indeed consider the following example: let $X = \mathbb{R}^N$ and let $\mu_n = \delta_{x_n}$ for some points $x_\in\mathbb{R}^N$. Assume that $x_n \to \bar{x}$. We *expect* that $\mu_n \to \mu$, where $\mu = \delta_{\bar{x}}$. Let us see if this is true by checking the definition we just introduced. We consider two cases where things go wrong (see Figure 10). Let $A \subset \mathbb{R}^N$ be an *open* set and assume that $x_n \in A$ for each $n \in \mathbb{N}$, but $\bar{x} \in \partial A$. In this case

$$0 = \mu(A) < \lim_{n\to\infty} \mu_n(A) = 1\,.$$

The problem is that on open sets we can *loose* mass in the limit because part of the set where μ_n is supported moves to the boundary of the open set and it is then not seen in the limit.

On the other hand, we can run into the opposite problem of *gaining* mass in the limit. Let $K \subset \mathbb{R}^N$ be n *compact* set and assume that $x_n \notin K$ for each $n \in \mathbb{N}$, but $\bar{x} \in \partial K$. In this case

$$1 = \mu(K) > \lim_{n\to\infty} \mu_n(K) = 0\,.$$

Motivated by the above heuristic, we introduce a natural notion of convergence of measures In the following, we denote by $\mathcal{M}(X)$ the set of Radon measures on the topological space X. In order to avoid technicalities, we will assume X to be a Borel subset of $\mathbb{R}^N$.

Definition 2.77. Let $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$, and $\mu \in \mathcal{M}(X)$. We say that the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ *converges locally weakly** to μ, and we write $\mu_n \overset{w*}{\rightharpoonup} \mu$ if

(i) for every open set $A \subset X$ it holds

$$\mu(A) \leq \liminf_{n\to\infty} \mu_n(A);$$

(ii) for each compact set $K \subset X$ it holds

$$\mu(K) \geq \limsup_{n\to\infty} \mu_n(K).$$

If in addition $\sup_{n\in\mathbb{N}} \mu_n(X) < \infty$, then we say that $\{\mu_n\}_{n\in\mathbb{N}}$ *converges weakly** to μ.

It is interesting to find conditions on a set that ensures that there is actual converges of the value of the measures. Moreover, it is useful to understand the behaviour of integrals of functions with respect to sequences that are (locally) weakly* convergent. For this reason we introduce a

special class of functions that will play a fundamental role in the following. The idea is that, if we want to talk about the asymptotic behaviour of

$$\int_X f \, d\mu_n$$

we need to make sure that the above quantity is well defined (namely f is μ_n-integrable for each $n \in \mathbb{N}$) and finite. This is the reason why we introduce the following class of functions.

Definition 2.78. Let (X, τ) be a topological space, and $f : X \to \overline{\mathbb{R}}$ be a function. We define the *support* of f by

Definition 2.79. We denote by $C_c(X)$ the space of continuous functions $f : X \to \mathbb{R}$ whose support

$$\operatorname{supp}(f) := \overline{\{x \in X \,:\, f(x) \neq 0\}}.$$

is compact. We endow $C_c(X)$ with the sup-norm

$$\|f\|_\infty := \sup_{x \in X} |f(x)|.$$

Remark 2.80. The sup-norm is the natural norm to endowed the space of continuous functions $C(X)$ with (not necessarily with compact support). Indeed $(C(X); \|\cdot\|_\infty)$ is a Banach space: every Cauchy sequence (with respect to the sup-norm) is convergent.

Note that a continuous function is μ-measurable for each Borel measure μ on X. Moreover, if $f \in C_c(X)$ we have that

$$\int_X f \, d\mu = \int_{\operatorname{supp}(f)} f \, d\mu \leq \|f\|_\infty \, \mu(\operatorname{supp}(f)) < \infty$$

since $\|f\|_\infty < \infty$ being the maximum of a continuous function on the compact set $\operatorname{supp}(f)$, and $\mu(\operatorname{supp}(f)) < \infty$ since μ is finite on compact sets.

We have the following equivalent definitions of weak* convergence for $X \subset \mathbb{R}^N$.

Lemma 2.81. *Let $\{\mu_n\}_{n \in \mathbb{N}} \subset \mathcal{M}(X)$, and $\mu \in \mathcal{M}(X)$. Then the followings are equivalent:*

(i) $\mu_n \overset{w*}{\rightharpoonup} \mu$;

(ii) For each bounded Borel set $B \subset X$ with $\mu(\partial B) = 0$, it holds

$$\mu(B) = \lim_{n \to \infty} \mu_n(B)\,.$$

(iii) It holds

$$\lim_{n \to \infty} \int_X f \, d\mu_n \to \int_X f \, d\mu$$

for each $f \in C_c(X)$.

Proof. Step 1: (i)⇒(ii). Since μ is a Radon measure, both ∂B (negligible set) and B (Borel set) are μ-measurable. Therefore

$$\begin{aligned}
\mu(B) &= \mu(B \setminus \partial B) && (\mu(\partial B) = 0)\\
&\leq \liminf_{n \to \infty} (\mu_n(B \setminus \partial B) && (B \setminus \partial B \text{ open})\\
&\leq \liminf_{n \to \infty} \mu_n(\bar{B}) && (B \setminus \partial B \subset \bar{B})\\
&\leq \limsup_{n \to \infty} \mu_n(\bar{B}) && (\liminf \leq \limsup)\\
&\leq \mu(\bar{B}) && (\bar{B} \text{ compact})\\
&= \mu(B) && (\mu(\partial B) = 0)\,.
\end{aligned}$$

Thus all inequalities above are actually equalities. This concludes the proof of this implication.

Step 2: (ii)⇒(iii). Let $f \in C_c(X)$. In order to use assumption (ii) we approximate the integral of f with respect to μ and to μ_n as follows. We assume $f \geq 0$. The general case is addressed by using the definition of the integral by using the positive and the negative part of a function. Fix $\varepsilon > 0$ and a radius $R > 0$ such that $\operatorname{supp}(f) \subset B_R$ (where B_R denotes the ball of radius R centered at the origin) and $\mu(\partial B_R) = 0$. Let $0 = y_0 < y_1 < \cdots < y_k = \|f\|_\infty + 1$ be such that $y_{i+1} - y_i < \varepsilon$ and that

$$\mu\left(\{f^{-1}(\{y_i\})\}\right) = 0\,,$$

for each $i = 1, \dots, k$. It is possible to choose $R > 0$ such that $\mu(\partial B_R) = 0$ and the y_i's satisfying the second assumption because μ is a Radon measure. Indeed, for the second case (the first being similar) for every $0 < a < b < \infty$ and every $j \in \mathbb{N}$ consider the set

$$M_j := \left\{ y \in [a,b] : \frac{1}{j+1} \leq \mu(f^{-1}(\{y\})) < \frac{1}{j} \right\}.$$

We have that the M_j's are μ-measurable, disjoint, and

$$f^{-1}([a,b]) = \bigcup_{j \in \mathbb{N}} M_j\,.$$

Therefore, by definition of the M_j's, we have that

$$\mu\left(f^{-1}([a,b])\right) = \sum_{j \in \mathbb{N}} \mu(M_j) \geq \sum_{j \in \mathbb{N}} \frac{1}{j+1} \#(M_j)\,,$$

where we recall that $\#(M_j)$ denotes the cardinality of M_j. Since μ is a Radon measure and $f^{-1}([a,b])$ is contained in a compact set (since $a > 0$), we get that $\#(M_j) = 0$ for all but finitely many indexes j's. In particular, this implies that $\mu\left(f^{-1}(\{t\})\right) = 0$ for all but countably many $t \in (0, \infty)$.

Let us continue with the proof. Define the Borel sets $B_i := f^{-1}\left((y_i, y_{i+1}]\right)$ for each $i = 0, \dots k-1$. We have that $y_i < f(x) \leq y_{i+1}$ for each $x \in B_i$ and that, by the choice of the y_i's $\mu(\partial B_i) = 0$ for each $i = 1, \dots, k$. Therefore, by (ii) we get

$$\lim_{n \to \infty} \mu_n(B_i) = \mu(B_i) \tag{2.3}$$

for all $i = 1, \dots, k-1$, and

$$\lim_{n \to \infty} \mu_n(B_R) = \mu(B_R)\,. \tag{2.4}$$

Thus

$$\sum_{i=1}^{k-1} y_i \mu_n(B_i) < \int_X f \, d\mu_n \sum_{i=1}^{k-1} y_{i+1} \mu_n(B_i) + y_1 \mu_n(B_R)$$

and

$$\sum_{i=1}^{k-1} y_i \mu(B_i) < \int_X f \, d\mu \sum_{i=1}^{k-1} y_{i+1} \mu_n(B_i) + y_1 \mu(B_R)$$

Therefore

$$\begin{aligned}\left| \int_X f \, d\mu_n - \int_X f \, d\mu \right| &\leq \sum_{i=1}^{k-1} |y_{i+1} - y_i| \, |\mu_n(B_i) - \mu(B_i)| + |y_1| \, [\mu_n(B_R) + \mu(B_R)] \\ &\leq \varepsilon \sum_{i=1}^{k-1} |\mu_n(B_i) - \mu(B_i)| + \varepsilon \, [\mu_n(B_R) + \mu(B_R)]\end{aligned}$$

By using (2.3) and (2.4) we get

$$\lim_{n \to \infty} \left| \int_X f \, d\mu_n - \int_X f \, d\mu \right| \leq 2\varepsilon \mu(B_R)\,.$$

Since $\varepsilon > 0$ is arbitrary, we conclude.

Step 3: (iii)⇒(i). Let $A \subset X$ be an open set. By inner regularity we have that

$$\mu(A) = \sup\{\mu(K) : K \subset A,\ K \text{ compact }\}. \tag{2.5}$$

Fix $K \subset A$ compact and let $f \in C_c(A)$ with $f(x) \in [0,1]$ for all $x \in A$ be such that $f \equiv 1$ on K. Then

$$\mu(K) \leq \int_X f\, d\mu = \lim_{n\to\infty} \int_X f\, d\mu_n \leq \mu_n(A),$$

where in the second step we used assumption (ii), while in the last the fact that f has compact support in A. Thus

$$\mu(K) \leq \liminf_{n\to\infty} \mu_n(A)$$

for each $K \subset A$ compact. This, together with (2.5) gives

$$\mu(A) \leq \liminf_{n\to\infty} \mu_n(A)$$

for each $A \subset X$ open. To prove that

$$\mu(K) \geq \limsup_{n\to\infty} \mu_n(K)$$

for all compact sets $K \subset X$ we reason in a similar way by using the outer regularity of μ

$$\mu(K) = \inf\{\mu(A) : A \subset K,\ A \text{ open }\}.$$

We fix an open set $A \supset K$ and a function $f \in C_c(A)$ with $f(x) \in [0,1]$ for all $x \in A$ be such that $f \equiv 1$ on K. This gives the desired inequality. □

In some applications to Optima Transport, it will be useful to understand the behaviour of integrals of larger classes of functions.

Definition 2.82. Let $f : X \to \mathbb{R}$. We say that f is

- *Lower semi-continuous* if
$$f(\bar{x}) \leq \liminf_{n\to\infty} f(x_n)$$
for each $\bar{x} \in X$ and each $\{x_n\}_{n\in\mathbb{N}} \subset X$ with $x_n \to \bar{x}$;
- *Upper semi-continuous* if
$$f(\bar{x}) \geq \liminf_{n\to\infty} f(x_n)$$
for each $\bar{x} \in X$ and each $\{x_n\}_{n\in\mathbb{N}} \subset X$ with $x_n \to \bar{x}$;
- *Borel* if $f^{-1}(B)$ is a Borel set for each $B \subset \mathbb{R}$ Borel.

Proposition 2.83. *Let $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ be such that $\mu_n \overset{w*}{\rightharpoonup} \mu$, where $\mu \in \mathcal{M}(X)$. Then*

$$\int_X f\, d\mu \leq \liminf_{n\to\infty} \int_X f\, d\mu_n,$$

for every lower semi-continuous function $f : X \to [0,\infty)$, and

$$\int_X f\, d\mu \geq \limsup_{n\to\infty} \int_X f\, d\mu_n,$$

for every upper semi-continuous function $f : X \to [0,\infty]$ with compact support. In particular,

$$\int_X f\, d\mu = \lim_{n\to\infty} \int_X f\, d\mu_n,$$

for every Borel function $f : X \to \mathbb{R}$ with compact support and whose set of discontinuities is μ-negligible.

A notion of convergence is of little use if it does not provide compactness (up to subsequences) of bounded sequences, namely for sequences $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ with

$$\sup_{n\to\infty} \mu_n(X) < \infty\,.$$

To prove that the weak* convergence actually gives compactness, we need to take a different point of view on measures, and see them as objects acting on functions via integration. This is a *dual* point of view of what we introduced in Definition (2.75). The following is a fundamental result connecting measure theory to functional analysis.

Theorem 2.84 (Riesz's representation theorem in $C_c(X)$)**.** *Let $L : C_c(X) \to [0,\infty]$ be a additive and locally bounded functional, namely*

$$L(f_1 + f_2) = L(f_1) + L(f_2)$$

for each $f_1, f_2 \in C_0(X)$, and

$$\sup\{L(f) \,:\, f \in C_c(K),\, \|f\|_\infty \leq 1\} < \infty\,.$$

for each compact set $K \subset X$. Then there exists a unique Radon measure μ on X such that

$$L(f) = \int_X f\, d\mu$$

for each $f \in C_c(X)$.

Remark 2.85. In other words, the Riesz representation theorem states that Radon measures on a topological space X are the (topological) dual of $C_c(X)$. in particular, the weak* convergence for measures we introduced is the weak* convergence in $(C_c(X))'$.

The desired compactness then follows from abstract results in functional analysis. Here, we state it as an independent result that can e proved *by hands*.

Theorem 2.86 (De La Vallée Poussin)**.** *Let $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ be a sequence of Radon measures such that*

$$\sup_{n\in\mathbb{N}} \mu_n(K) < \infty\,,$$

for each compact set $K \subset X$. Then there exists a subsequence $\{\mu_{n_j}\}_{j\in\mathbb{N}}$ and a Radon measure μ on X such that $\mu_{n_j} \overset{w^*}{\rightharpoonup} \mu$.

Proof. Step 1: on a compact set. Fix a compact set $K \subset X$. Let $\{f_i\}_{i\in\mathbb{N}}$ be a dense subset of $C_c(X)$ with respect the the sup-norm. By assumption

$$M := \sup_{n\in\mathbb{N}} \int_K f_i\, d\mu_n < \infty \tag{2.6}$$

for each $i \in \mathbb{N}$. We first construct the limiting measure μ and the subsequence and then we prove the convergence.

Step 1.1: construction of the subsequence. We proceed by induction. Fix $i = 1$. By (2.6) we can extract a subsequence $\{\mu_n^1\}_{n\in\mathbb{N}}$ of $\{\mu_n\}_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \int_K f_1\, d\mu_n^1 = a_1$$

for some $a_1 \in \mathbb{R}$. Let us now assume that we have extracted a subsequence $\{\mu_n^k\}_{n\in\mathbb{N}}$ and found $a_1, \dots, a_k \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \int_K f_i\, d\mu_n^k = a_i$$

for each $i = 1, \dots, k$. By using again (2.6) we can extract a subsequence $\{\mu_n^{k+1}\}_{n\in\mathbb{N}}$ of $\{\mu_n^k\}_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \int_K f_{k+1}\, d\mu_n^{k+1} = a_{k+1}$$

for some $a_{k+1} \in \mathbb{R}$. We thus found a subsequence $\{\mu_{n_j}\}_{j\in\mathbb{N}}$ and a numbers $\{a_i\}_{i\in\mathbb{N}}$ such that

$$\lim_{j\to\infty} \int_K f_i \, d\mu_{n_j} = a_i$$

for each $i \in \mathbb{N}$.

Step 1.2: construction of the measure μ. Define the functional $L : \{f_i\}_{i\in\mathbb{N}} \to \mathbb{R}$ by $L(f_i) \coloneqq a_i$. This is linear and bounded, since

$$L(f_i) \leq M \|f_i\|_\infty .$$

By density of $\{f_i\}_{i\in\mathbb{N}}$ it is possible to uniquely extend L to a bounded linear functional $\bar{L}$ on $C_c(K)$. By the Riesz Representation Theorem (Theorem 2.84) there exists a Radon measure μ on K such that

$$\bar{L}(f) = \int_K f \, d\mu$$

for each $f \in C_c(K)$. Note that μ is finite on K, being this compact and μ Radon.

Step 2: convergence. By construction of L we know that

$$\lim_{j\to\infty} \int_X f_i \, d\mu_{n_j} = \int_K f_i \, d\mu$$

for each $i \in \mathbb{N}$. The idea now is to use the density of $\{f_i\}_{i\in\mathbb{N}}$ to prove that the same holds for each $f \in C_c(K)$. Fix $f \in C_c(K)$ and $\varepsilon > 0$. Let $i \in \mathbb{N}$ be such that

$$\|f_i - f\|_\infty < \varepsilon , \tag{2.7}$$

and let $\bar{j} \in \mathbb{N}$ be such that

$$\left| \int_K f_i \, d\mu_{n_j} - \int_K f_i \, d\mu \right| < \varepsilon \tag{2.8}$$

for each $j \geq \bar{j}$. Then, for $j \geq \bar{j}$, by the triangle inequality we get

$$\begin{aligned} \left| \int_K f \, d\mu_{n_j} - \int_K f \, d\mu \right| \leq & \left| \int_K f_i \, d\mu_{n_j} - \int_K f \, d\mu_{n_j} \right| + \left| \int_K f_i \, d\mu_{n_j} - \int_K f_i \, d\mu \right| \\ & + \left| \int_K f_i \, d\mu - \int_K f \, d\mu \right| \\ \leq & \, \varepsilon(2M + 1) \end{aligned}$$

where in the last step we used (2.6), (2.7) and (2.8). Being $\varepsilon > 0$ arbitrary, we conclude.

Step 3: on X. The case where $\sup_{n\in\mathbb{N}} \mu_n(X) = \infty$ is handled as follows: consider an increasing sequence of positive numbers $\{R_i\}_{i\in\mathbb{N}}$ with $\lim_{i\to\infty} R_i = \infty$. On $\overline{B_{R_1}}$ the previous steps allow us to construct a subsequence $\{\mu_n^1\}_{n\in\mathbb{N}}$ and a Radon measure μ^1 on B_{R_1} such that μ_n^1 converges (locally) weakly* to μ^1 on B_{R_1}. We then consider the sequence $\{\mu_n^1\}_{n\in\mathbb{N}}$ on B_{R_2}. Again the previous steps allow us to extract a subsequence $\{\mu_n^2\}_{n\in\mathbb{N}}$ and a Radon measure μ^2 on B_{R_2} such that μ_n^2 converges (locally) weakly* to μ^2 on B_{R_2}. Note that this implies that $\mu^2 = \mu^1$ on B_{R_1}. We then continue to extract further subsequences and Radon measures μ^i that extends μ^{i-1} and we conclude. □

The local weak* convergence of measure allows to understand the behaviour of integrals of functions in $C_c(X)$ with respect to converging sequences of measures. It is useful to extend the class of functions for which we can say something about their integrals with respect to converging sequences of measures. We start with some definitions, the first of which is motivated by the fact that a Cauchy sequence in $C_c(X)$ does not necessarily converge to an element of $C_c(X)$.

Definition 2.87. Let (X, τ) be a topological space. We denote by $C_0(X)$ the closure of $C_c(X)$ with respect to the $\|\cdot\|_\infty$ norm. Namely, $f \in C_0(X)$ if and only if there exists $\{f_n\}_{n\in\mathbb{N}} \subset C_c(X)$ such that $\|f_n - f\|_\infty \to 0$ as $n \to \infty$.

Remark 2.88. It is easy to see that $C_0(X) \subset C(X)$ with $C_0(X) \neq C(X)$. Indeed, $C_0(X)$ is the set of functions that *vanish at infinity (or at the boundary of X)*: $f \in C_0(X)$ if and only if for each $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset X$ with $|f| \leq \varepsilon$ in $X \setminus K_\varepsilon$.

In particular, X is compact it holds that $C_c(X) = C_0(X)$ and it can be identified as the space of functions that are zero on ∂X.

Theorem 2.89 (Riesz's representation theorem in $C_0(X)$)**.** *Let $L : C_0(X) \to [0, \infty]$ be a additive and bounded functional, namely*

$$L(f_1 + f_2) = L(f_1) + L(f_2)$$

for each $f_1, f_2 \in C_0(X)$, and

$$\|L\| = \sup\{L(f) \,:\, f \in C_0(X),\, \|f\|_\infty \leq 1\} < \infty\,.$$

Then there exists a unique finite Radon measure μ on X such that

$$L(f) = \int_X f \, d\mu$$

for each $f \in C_0(X)$. In particular $\mu(X) = \|L\|$.

Remark 2.90. In other words, the Riesz representation theorem states that *finite* Radon measures on a topological space X are the (topological) dual of $C_0(X)$. Indeed, note that the Radon measure obtained in the Riesz's representation theorem in $C_c(X)$ can be such that $\mu(X) = \infty$. Nevertheless

$$\mu(K) = \sup\{L(f) \,:\, f \in C_c(K),\, \|f\|_\infty \leq 1\} < \infty\,,$$

for each compact set $K \subset X$. This is why it is a local version of Theorem 2.89.

The first part of the proof of the local version of the compactness theorem gives the global version.

Theorem 2.91 (De La Vallée Poussin)**.** *Let $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ be a sequence of Radon measures such that*

$$\sup_{n\in\mathbb{N}} \mu_n(X) < \infty\,.$$

Then there exists a subsequence $\{\mu_{n_j}\}_{j\in\mathbb{N}}$ and a finite Radon measure μ on X such that μ_{n_j} converges weakly to μ. In particular*

$$\lim_{j\to\infty} \int_X f \, d\mu_{n_j} = \int_X f \, d\mu$$

for each $f \in C_0(X)$.

So far we considered functions that vanish at the boundary of X (see Remark 2.88). We now consider the case of bounded continuous functions.

Definition 2.92. We say that a sequence $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ of *finite* Radon measures *converges tightly* to the Radon measure μ if

$$\lim_{n\to\infty} \int_X f \, d\mu_n = \int_X f \, d\mu$$

for each $f \in C_b(X)$, continuous bounded function $f : X \to \mathbb{R}$.

Remark 2.93. Since $C_c(X) \subset C_0(X) \subset C_b(X)$, the following holds:

$$\text{tight convergence} \;\Rightarrow\; \text{weak* convergence} \;\Rightarrow\; \text{local weak* convergence}\;.$$

Moreover, if X is compact, then $C_c(X) = C_0(X) = C_b(X)$ and all the three notions of converge for Radon measures are equivalent.

It is possible to infer the tight convergence from the weak* convergence if the sequence of measures *does not loose mass* at the boundary of X (that could be at infinity if X is unbounded).

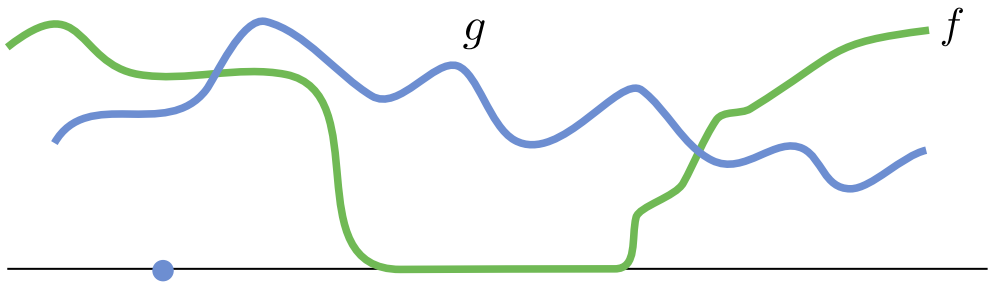

FIGURE 11. The relation between two measures $\mu = f\mathcal{L}^1$, depicted in green, and $\nu = \delta_{x_0} + g\mathcal{L}^1$, depicted in blue.

Proposition 2.94. *Let $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ with $\mu_n \overset{w*}{\rightharpoonup} \mu$, where $\mu \in \mathcal{M}(X)$, and such that*

$$\lim_{n\to\infty} \mu_n(X) = \mu(X)\,.$$

Then μ_n convergences to μ tightly.

Finally, we get compactness with respect to the tight convergence. This result is well known in probability.

Theorem 2.95 (Prohorov theorem)**.** *Let $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ with*

$$\sup_{n\in\mathbb{N}} \mu_n(X) < \infty$$

and such that, for each $\varepsilon > 0$, it is possible to find a compact set $K_\varepsilon \subset X$ such that

$$\sup_{n\in\mathbb{N}} \mu_n(X \setminus K_\varepsilon) < \varepsilon\,.$$

Then there exists a subsequence $\{\mu_{n_j}\}_{j\in\mathbb{N}}$ and a finite Radon $\mu \in \mathcal{M}(X)$ such that μ_{n_j} converges tightly to μ. Also the vice versa holds.

2.5.3. *Relation between measures*. Finally, we want to investigate the relation between two measures. In particular, we wonder, given a measure μ on X for which class of measure ν on X it holds $\nu = f\mu$, where $f \in L^1(X;\mu)$. Namely, when it is possible to write

$$\nu(E) = \int_E f d\mu \tag{2.9}$$

for every $E \in \mathcal{M}_\mu$. In the above formula, f plays the role of a *density* of the measure ν with respect to the measure μ. As an example, consider two measures $\mu = f\mathcal{L}^1$ and $\nu = \delta_{x_0} + g\mathcal{L}^1$ (see Figure 11). The Dirac delta present in ν cannot be written in terms of μ, since the former is concentrated on a set that is μ-negligible. Same consideration in the region where $f = 0$ but $g > 0$. On the other hand, where $f > 0$, it is possible to write the part $g\mathcal{L}^1$ of the measure ν in terms of μ as follows:

$$\nu(E) = \int_E \frac{g}{f}\, d\mu$$

if $E \subset \mathbb{R}$ with $(\{x_0\} \cup \{f = 0\}) \cap E = \emptyset$. In particular, wee see that if ν assigns a positive measure to μ-negligible sets, formula (2.9) cannot hold. This justifies the following definition.

Definition 2.96. Let $\mu, \nu : \mathcal{A} \to [0,\infty]$ be two measures on a space X, where $\mathcal{A}$ is a σ-algebra. We say that ν is *absolutely continuous* with respect to μ, and we write $\nu \ll \mu$, if

$$\mu(E) = 0 \quad \Rightarrow \quad \nu(E) = 0$$

for all $E \subset X$.

The amazing fact is that absolute continuity characterizes the couple of measures for which (2.9) holds.

Theorem 2.97 (Radon-Nikodym Theorem)**.** *Let $\mu, \nu : \mathcal{A} \to [0, \infty]$ be measures on a space X, where $\mathcal{A}$ is a σ-algebra. Assume that X is σ-finite with respect to μ. If $\nu \ll \mu$, then there exists a function $f \in L^1(X; \mu)$ such that*

$$\nu(E) = \int_E f \, d\mu$$

for each $E \in \mathcal{A}$. The function f is unique up to sets of measures zero.

The function f obtained in the above theorem is called the *Radon-Nikodym derivative* of ν with respect to μ, and it is denoted by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. The proof of the Radon-Nikodym theorem is not constructive. Nevertheless, for certain measures it is possible to obtain the Radon-Nikodym derivative more explicitly.

Theorem 2.98 (Lebesgue-Besikovitch differentiation Theorem)**.** *Assume that $X = \mathbb{R}^N$ and that μ, ν are Radon measures on $\mathbb{R}^N$, with $\nu \ll \mu$. Fix $C \subset \mathbb{R}^N$, a convex set containing the origin Then, for μ-a.e. $x \in \mathbb{R}^N$, the limit*

$$\lim_{r \to 0} \frac{\nu(x + rC)}{\mu(x + rC)} =: f(x)$$

exists, and $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

Remark 2.99. There is a notion of absolute continuity also for functions, and the Radon-Nikodym Theorem together with the Lebesgue-Besikovitch differentiation theorem give the validity of the Fundamental Theorem of Calculus for absolutely continuous functions.

Next question is if it is possible to say something in case ν is not absolutely continuous with respect to μ.

Definition 2.100. Given two measures $\mu, \nu : \mathcal{A} \to [0, \infty]$, we say that they are *mutually singular*, and we write $\mu \perp \nu$ if there exists $X_\mu, X_\nu \in \mathcal{A}$ with $X_\mu \cap X_\nu = \emptyset$ and $X = X_\nu \cup X_\nu$ such that

$$\mu(E) = \mu(E \cap X_\mu), \qquad \nu(E) = \nu(E \cap X_\nu).$$

for each $E \in \mathcal{A}$.

Remark 2.101. Mutually singular measures *do not see each other* because they are concentrate on different sets.

Theorem 2.102 (Lebesgue decomposition theorem)**.** *Let $\mu, \nu : \mathcal{A} \to [0, \infty]$ be two measures and assume μ to be σ-finite. Then there exist measures $\nu_{ac}, \nu_s : \mathcal{A} \to [0, \infty]$ such that*

$$\nu = \nu^{ac} + \nu^s$$

where $\nu^{ac} \ll \mu$. Moreover, if also ν is σ-finite, then $\nu^s \perp \mu$, and the measures ν^{ac}, ν^s of the decomposition are unique.

Remark 2.103. In the case where $X = \mathbb{R}^N$ and both measures are Radon and σ-finite, it is possible to characterize the set where ν^s is concentrated by the set of points $x \in \mathbb{R}^N$ for which

$$\lim_{r \to 0} \frac{\nu(x + rC)}{\mu(x + rC)} = \infty \, .$$

2.6. **This section in a nutshell.**

- The rigorous definition of length, area, volume, etc. leads to the definition of outer measures;
- Outer measures are *not* countably additive on an arbitrary choice of pairwise disjoint sets. Only for *measurable* set this holds;
- On measurable sets we have nice continuity properties with respect to increasing and decreasing sequences of sets;

- Radom measures are an important class of measures (good for regularity properties);
- The definition of Lebesgue integral has the advantage that the class of functions where it is defined on is larger than that where the Riemann integral is. Moreover, limiting theorems hold with weaker assumptions;
- Product measures are a natural way to define a measure on a product space. Integrals behave well with respect to the product measure (Tonelli's and Fubini's Theorems allow to integrate *one variable at the time*);
- The push-forward measure allows to transfer a measure on the target space of a function;
- Measures are not only functionals on sets, but functionals on functions. This duality is the Riesz Representation Theorem;
- It is possible to define several notions of convergence for sequences of measures, each treating a different class of functions: $C_c(X)$, $C_0(X)$, and bounded continuous functions. Each notion has the good taste to ensure compactness of bounded sequences of Radon measures.
- It is possible to write a measure as a sum of a density with respect to another measure and a singular part (Lebesgue decomposition Theorem and Radon-Nikodym Theorem). The density can be obtained by actual differentiation of measures (Lebesgue-Besikovitch Theorem);

3. The Kantorovich problem

The original Monge problem (see (1.3)) asks to find the *transport map* $T : \mathbb{R}^N \to \mathbb{R}^N$ that transform the density f into the density g in such a way to minimize the cost c. In mathematical terms it is written as

$$\min \left\{ \int_{\mathbb{R}^N} c(x, T(x)) f(x)\, dx \ : \ T : \mathbb{R}^N \to \mathbb{R}^N \text{ with } g = T_{\#} f \right\}.$$

We saw (see Example 3 in Introduction) that a problem with asking for a *map* is that splitting mass is not allowed, and this might be a problem. This is the reason why Kantorovich in [14] proposed to extend the Monge problem in such a way to allow for splitting of mass.

3.1. The Kantorovich formulation.

The idea of splitting mass can be made rigorous by using the notion of measure (see Figure 12). Moreover, it allows to consider more general initial and final measures (not necessarily generated by densities f and g).

Definition 3.1. We denote by $\mathcal{P}(X)$ the set of *probability measures* on X, namely Radon measures μ on X with $\mu(X) = 1$.

Definition 3.2. Let (X, μ), (Y, ν) be two measure spaces, with $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. We say that a measure $\gamma \in \mathcal{P}(X \times Y)$ is a *transport plan* between μ and ν if

(i) $(\pi_1)_{\#}\gamma = \mu$;
(ii) $(\pi_2)_{\#}\gamma = \nu$.

We denote by $\Pi(\mu, \nu)$ the set of transport plans between μ and ν.

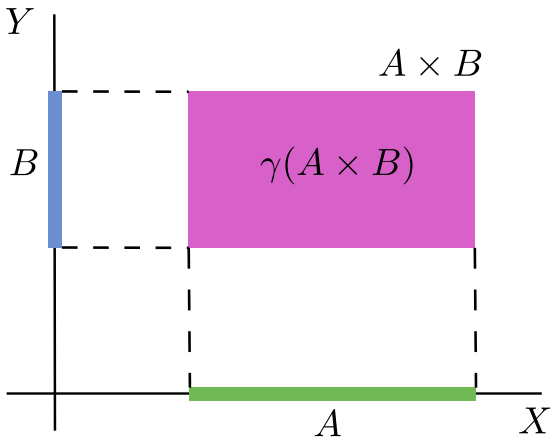

FIGURE 12. The notion of measures allow to make rigorous the idea of splitting mass: for each $A \subset X$ and $B \subset Y$ the quantity $\gamma(A \times B)$ represents the quantity of mass that has been moved from A to B.

Remark 3.3. In the language of probability, γ has μ and ν as first and second marginal. Note that $\Pi(\mu, \nu) \neq \emptyset$, since $\mu \otimes \nu \in \Pi(\mu, \nu)$.

Remark 3.4. The idea is that a transport plan γ determines, for each $A \subset X$, how the mass in A is spread out in the target space Y. Since all the initial mass μ at each point has to be moved somewhere, and all the final mass ν has to be matched, we have condition (i) and (ii) of the above definition respectively .

To get a better feeling on the notion of transport plan, think about how the transport plan $\mu \otimes \nu$ spreads the mass.

The notion of transport plan extends that of a transport maps.

Lemma 3.5. *Let (X,μ), (Y,ν) be two measure spaces, with $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Let $T : X \to Y$ be a Borel map such that $T_\#\mu = \nu$. Then the transport plan induced by T is*

$$\gamma_T := (\mathrm{Id}, T)_\#\mu\,.$$

Here $(\mathrm{Id}, T) : X \to X \times Y$ is defined by $(\mathrm{Id}, T)(x) := (x, T(x))$.

Proof. Let $A \subset X$ and $B \in Y$ Borel sets. The mass moved by the map T from A to B is

$$\mu(A \cap T^{-1}(B)) = \mu(\mathrm{Id}^{-1}(A) \cap T^{-1}(B)) = \mu\left((\mathrm{Id}, T)^{-1}(A, B)\right) = (\mathrm{Id}, T)_\#\mu(A, B)\,.$$

This is precisely the definition of the transport plan γ_T.

We now check that $\gamma_T \in \Pi(\mu,\nu)$. By definition, $\gamma_T \in \mathcal{P}(X \times Y)$. We need to prove that (i) and (ii) of Definition 3.2. Let us start with (i). Let $A \subset X$ be a μ-measurable set. Then

$$\begin{aligned}(\pi_1)_\#\gamma_T(A) &= \gamma_T\left((\pi_1)^{-1}(A)\right)\\ &= \gamma_T(A \times Y)\\ &= (\mathrm{Id}, T)_\#\mu(A \times Y)\\ &= \mu\left((\mathrm{Id}, T)^{-1}(A \times Y)\right)\\ &= \mu(A)\,.\end{aligned}$$

This proves (i). In order to prove (ii), let $B \subset Y$ be ν-measurable. Then

$$\begin{aligned}(\pi_2)_\#\gamma_T(B) &= \gamma_T\left((\pi_2)^{-1}(B)\right)\\ &= \gamma_T(X \times B)\\ &= (\mathrm{Id}, T)_\#\mu(X \times B)\\ &= \mu\left((\mathrm{Id}, T)^{-1}(X \times B)\right)\\ &= \mu\left(T\right)^{-1}(B))\\ &= \nu(B)\,,\end{aligned}$$

where in the last step we used the fact that $T_\#\mu = \nu$. □

We are now in position to state the Kantorovich problem Let (X,μ), (Y,ν) be two measure spaces, with $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Consider a continuous cost function $c : X \times Y \to [0,\infty)$. The *Kantorovich problem* writes as

$$\min\left\{ \int_{X\times Y} c(x,y)\, d\gamma \;:\; \gamma \in \Pi(\mu,\nu) \right\}. \tag{3.1}$$

Remark 3.6. The Kantorovich problem is a generalized version of the Monge problem. Indeed

$$\int_{X\times Y} c(x,y)\, d\gamma_T = \int_{\mathbb{R}^N} c(x, T(x))\, d\mu\,,$$

and in the case $\mu = f\mathcal{L}^N$ this writes as

$$\int_{X\times Y} c(x,y)\, d\gamma_T = \int_{\mathbb{R}^N} c(x, T(x)) f(x)\, dx\,.$$

Moreover, note that, since by Lemma 3.5 the set $\Pi(\mu,\nu)$ contains the set of transport maps, we have that

$$\inf\left\{ \int_{X\times Y} c(x,y)\, d\gamma \;:\; \gamma \in \Pi(\mu,\nu) \right\} \le \inf\left\{ \int_{\mathbb{R}^N} c(x, T(x)) f(x)\, dx \;:\; g = T_\# f \right\}.$$

Note that we write inf in place of min since we do not (yet) know whether or not the problems admit a solution. We have already seen a case where the strict inequality holds: $\mu = 2\delta_{x_0}$, $\nu = \delta_{y_1} + \delta_{y_2}$ with x_0, y_1, y_3 three distinct points. In this case there is *no* transport map.

3.2. **Existence for the Kantorovich problem.** The advantage of the Kantorovich formulation (namely allowing splitting of mass) is that existence follows easily. In order to focus on the main ideas, we consider the case where X and Y are *compact* spaces. Extensions will be treated later on.

Theorem 3.7. *Let (X,μ), (Y,ν) be two measure spaces, with $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, and X, Y be compact. Let $c : X \times Y \to [0,\infty)$ be continuous. Then the Kantorovich problem admits a solution.*

Proof. Step 1. Let $\{\gamma_n\}_{n\in\mathbb{N}} \subset \Pi(\mu,\nu)$ be a minimizing sequence for the Kantorovich problem. Namely

$$\lim_{n\to\infty} \int_{X\times Y} c(x,y)\, d\gamma_n = \inf\left\{ \int_{X\times Y} c(x,y)\, d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\}. \tag{3.2}$$

Since $\sup_{n\in\mathbb{N}} \gamma_n(X\times Y) = 1$, by Theorem 2.91 together with Remark 2.93 we have that there exists a subsequence $\{\gamma_{n_j}\}_{j\in\mathbb{N}}$ and $\gamma \in \mathcal{P}(X\times Y)$ such that γ_{n_j} converges to γ tightly. In particular

$$\lim_{n\to\infty} \int_{X\times Y} c(x,y)\, d\gamma_{n_j} = \lim_{n\to\infty} \int_{X\times Y} c(x,y)\, d\gamma\,,$$

and thus, by (3.2), γ achieves the infimum of the Kantorovich problem.

Step 2. In order to prove that γ is a solution to the Kantorovich problem, we need to prove that it is an admissible competitor. This is what was not possible to prove in the case of transport maps. We need to prove that $(\pi_1)_\#\gamma = \mu$ and $(\pi_2)_\#\gamma = \nu$. We will do that by using the dual nature of measures. Let $\varphi \in C(X)$. We have to prove that

$$\int_X \varphi\, d(\pi_1)_\#\gamma = \int_X \varphi\, d\mu\,. \tag{3.3}$$

Indeed, since by assumption $(\pi_1)_\#\gamma_{n_j} = \mu$ for each $j \in \mathbb{N}$, we have that

$$\int_X \varphi\, d\mu = \int_X \varphi\, d(\pi_1)_\#\gamma_{n_j} = \int_X \varphi\circ\pi_1\, d\gamma_{n_j}\,, \tag{3.4}$$

where in the last step we used Lemma 2.73 (φ satisfies the assumptions of the lemma because it is continuous on a compact set). Note that $\varphi\circ\pi_1 \in C(X\times Y)$. Therefore, since γ_{n_j} converges to γ tightly, we have that

$$\int_X \varphi\circ\pi_1\, d\gamma_{n_j} \to \int_X \varphi\circ\pi_1\, d\gamma \tag{3.5}$$

as $j\to\infty$. Thus, (3.4) together with (3.5) yield (3.3). In a similar way it is possible to prove that $(\pi_2)_\#\gamma = \nu$. This proves that $\gamma \in \Pi(\mu,\nu)$, and in turn that it is a minimizer to the Kantorovich problem. □

The compactness of the spaces X and Y was only in order to focus on the main ideas of the proof. This assumption can be relaxed (here we present it only in $\mathbb{R}^N$, but it actually holds for more general metric spaces), as well as the continuity of the cost function.

Theorem 3.8. *Let $X, Y \subset \mathbb{R}^N$ be Borel sets, and let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Let $c : X\times Y \to \mathbb{R}\cup\{\infty\}$ be lower semi-continuous and bounded from below. Then the Kantorovich problem admits a solution.*

We are now in the following situation: we started with the Monge problem, and we saw some difficulties in solving it. One being the necessity of spreading mass. This lead to the Kantorovich formulation that, by extending the class of objects that are admissible competitors, allows for existence of a (generalized) solution. We know nothing about uniqueness of the solution. We now have the following questions:

(i) Is there a solution to the Kantorovich problem that is also a solution to the Monge problem?

(ii) When is it true that

$$\inf\left\{\int_{X\times Y} c(x,y)\,d\gamma \,:\, \gamma\in\Pi(\mu,\nu)\right\} = \inf\left\{\int_{\mathbb{R}^N} c(x,T(x))f(x)\,dx \,:\, g=T_{\#}f\right\}?\,.$$

(iii) Is there another way to extend the class of admissible competitors in such a way that we have existence of a (generalized) solution, but the class of admissible objects is in between transport maps and transport plans?

Question (i) is about *regularity* of solutions: are there optimal transport plans that are induced by transport maps? Question (ii) relates to the density of transport maps in the space of transport plans in such a way to approximate the optimal costs. Finally, question (iii) is about the minimality of the extension we used.

In order to understand better the above questions, let us consider this simple example. Do not focus on the actual solution to the problem, but on the similarities with our case. Consider the minimization problem

$$\min\left\{\, F(x) \coloneqq |x^2-2| \,:\, x\in\mathbb{Q}\,\right\}.$$

Assume that we want to apply the Weirstrass Theorem: take a minimizing sequence $\{x_n\}_{n\in\mathbb{N}}\subset\mathbb{Q}$. Then we have that

$$\sup_{n\in\mathbb{N}} F(x_n) < \infty\,.$$

What we can conclude is that there exists a subsequence $\{x_{n_j}\}_{j\in\mathbb{N}}$ and $\bar{x}\in\mathbb{R}$ such that $x_{n_j}\to\bar{x}$. We are thus forced by the problem itself to extend the class of admissible competitors from $\mathbb{Q}$ to $\mathbb{R}$, and consider the following generalization of the problem:

$$\min\left\{\, \widetilde{F}(x) \coloneqq |x^2-2| \,:\, x\in\mathbb{R}\,\right\}.$$

I wrote $\widetilde{F}$ because, in principle (not in this case, of course), the functional F might make no sense on the new objects we consider (namely on $\mathbb{R}\setminus\mathbb{Q}$). For the generalized problem, we can prove existence of a solution. The three questions above still make sense in this case:

(i) Is there a solution to the minimization problem for $\widetilde{F}$ that is also a solution to the minimization problem for F?

(ii) Is it true that

$$\min\left\{\, \widetilde{F}(x) \coloneqq |x^2-2| \,:\, x\in\mathbb{R}\,\right\} = \min\left\{\, F(x) \coloneqq |x^2-2| \,:\, x\in\mathbb{Q}\,\right\}?\,.$$

(iii) Is there another way to extend the class of admissible competitors in such a way that we have existence of a (generalized) solution, but the class of admissible objects is in between transport maps and transport plans?

In this case we see better what we mean with those questions. In particular, we comment on the last one: we could have extended the competitors from $\mathbb{Q}$ to $\mathbb{C}$, but this would have been a too large extension that is not necessary to get existence of a (generalized) solution. In this regard, we say that $\mathbb{R}$ is the minimal extension of the class of admissible competitors $\mathbb{Q}$ for the initial problem that ensures existence of a solution.

In the next sections we will answer to all the above three questions.

4. Existence for the Monge problem

We now want to prove that, for important classes of costs and under some general assumptions on the measures μ and ν, the Monge problem admits a solution, and that in certain cases it is also unique. The idea is to prove *regularity* of a optimal transport plan for the Kantorovich problem. The path toward our goal passes through another problem, the so called *dual problem*, a common tool used in linear optimization to gain more information on the primal problem.

4.1. **The dual problem.** The main difficulty in solving the Monge problem was the lack of compactness. This was solved by using a weak formulation for which compactness holds and, in turn, existence of a generalized solution can be easily proved. It is nevertheless the reason at the core of the lack of compactness for the Monge problem (namely the absence of any derivative of the transport map in the problem) that allows to use a powerful tool in linear programming: the dual problem. This is a problem that is obtained as follows: the constrain in the minimization problem is included in the functional as a penalization term (a supremum). The problem then becomes as infimum of a supremum and we *formally* switch them in order to obtain a supremum of an infimum. This last infimum is then removed from the functional an inserted in the maximization problem as a constrain. Basically, we switch the role of competitors and of constraints. The advantage of this dual problem is that (usually) it has a solution even when the primal problem doesn't. Moreover it is used to get information on the primal problem and on its minimizers (if any).

Let us see the above idea in action in our specific case. In order to avoid technicalities and focus on the main ideas, in this section X and Y are compact metric spaces and $c : X \times Y \to [0, \infty)$ is a continuous function, hence *uniformly continuous*. In the Kantorovich problem

$$\min \left\{ \int_{X\times Y} c \, d\gamma \, : \, \gamma \in \Pi(\mu, \nu) \right\}$$

we want to write the constrain $\gamma \in \Pi(\mu, \nu)$ as a penalization term. For, we exploit the dual nature of measures seen as functionals on the space of continuous functions. Note that, since X is compact, we have $C_c(X) = C_0(X) = C_b(X) = C(X)$. We have that

$$(\pi_1)_\# \gamma = \mu \qquad \Leftrightarrow \qquad \int_X \varphi \, d(\pi_1)_\# \gamma = \int_X \varphi \, d\mu$$

for all $\varphi \in C(X)$, and

$$(\pi_2)_\# \gamma = \nu \qquad \Leftrightarrow \qquad \int_Y \psi \, d(\pi_2)_\# \gamma = \int_Y \psi \, d\nu$$

for all $\psi \in C(Y)$. Recalling that

$$\int_X \varphi \, d(\pi_1)_\# \gamma = \int_{X\times Y} \varphi(x) \, d\gamma \, , \qquad \int_Y \psi \, d(\pi_2)_\# \gamma = \int_{X\times Y} \psi(y) \, d\gamma \, ,$$

we can say that $\gamma \in \Pi(\mu, \nu)$ if and only if

$$\int_X \varphi \, d\mu + \int_Y \psi \, d\nu = \int_{X\times Y} [\, \varphi(x) + \psi(y) \,] \, d\gamma$$

for all $\varphi \in C(X)$ and $\psi \in C(Y)$. Denote by $\mathcal{M}(X \times Y)$ the space of finite Radon measures on $X \times Y$. From these considerations, it is easy to see that

$$\sup_{\varphi\in C(X), \psi \in C(Y)} \left\{ \int_X \varphi \, d\mu + \int_Y \psi \, d\nu - \int_{X\times Y} [\, \varphi(x) + \psi(y) \,] \, d\gamma \right\} = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu, \nu) \, , \\ +\infty & \text{if } \gamma \in \mathcal{M}(X \times Y) \setminus \Pi(\mu, \nu) \, . \end{cases}$$

Therefore, we get

$$\inf \left\{ \int_{X\times Y} c \, d\gamma \, : \, \gamma \in \Pi(\mu, \nu) \right\}$$

$$= \inf_{\gamma\in\mathcal{M}(X\times Y)} \left\{ \int_{X\times Y} c\,d\gamma + \sup_{\varphi\in C(X),\psi\in C(Y)} \left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu - \int_{X\times Y} [\,\varphi(x)+\psi(y)\,]\,d\gamma \right\} \right\}$$
$$= \inf_{\gamma\in\mathcal{M}(X\times Y)} \sup_{\varphi\in C(X),\psi\in C(Y)} \left\{ \int_{X\times Y} c\,d\gamma + \left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu - \int_{X\times Y} [\,\varphi(x)+\psi(y)\,]\,d\gamma \right\} \right\}$$

We now swap the operation of infimum and supremum *formally*, namely without any justification (yet), getting

$$\inf_{\gamma\in\mathcal{M}(X\times Y)} \sup_{\varphi\in C(X),\psi\in C(Y)} \left\{ \int_{X\times Y} c\,d\gamma + \left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu - \int_{X\times Y} [\,\varphi(x)+\psi(y)\,]\,d\gamma \right\} \right\}$$
$$= \sup_{\varphi\in C(X),\psi\in C(Y)} \inf_{\gamma\in\mathcal{M}(X\times Y)} \left\{ \int_{X\times Y} c\,d\gamma + \int_X \varphi\,d\mu + \int_Y \psi\,d\nu - \int_{X\times Y} [\,\varphi(x)+\psi(y)\,]\,d\gamma \right\}$$
$$= \sup_{\varphi\in C(X),\psi\in C(Y)} \left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu + \inf_{\gamma\in\mathcal{M}(X\times Y)} \left\{ \int_{X\times Y} [\,c-\varphi(x)-\psi(y)\,]\,d\gamma \right\} \right\}.$$

Now we note that

$$\inf_{\gamma\in\mathcal{M}(X\times Y)} \left\{ \int_{X\times Y} [\,c(x,y)-\varphi(x)-\psi(y)\,]\,d\gamma \right\} = \begin{cases} 0 & \text{if } \varphi(x)+\psi(y)\le c(x,y)\,, \\ -\infty & \text{else}\,. \end{cases}$$

Thus, this inner infimum acts as a penalization for the constraint $\varphi(x)+\psi(y)\le c$, since we want to maximize in φ and ψ. As a shorthand notation we will write $\varphi\oplus\psi : X\times Y\to\mathbb{R}$ for the function $\varphi\oplus\psi(x,y) \coloneqq \varphi(x)+\psi(y)$. Thus, we have obtained *formally* that

$$\inf\left\{ \int_{X\times Y} c\,d\gamma \,:\, \gamma\in\Pi(\mu,\nu) \right\} = \sup\left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu \,:\, \varphi\oplus\psi\le c \right\}.$$

This equality is called *duality formula*, and the maximization problem on the right-hand side is called the *dual problem* for the Kantorovich problem: note that the cost c is not in the functional to maximize, but on the constrain, while the variables φ and ψ that were used to describe the constraint in the primal problem (the Kantorovich problem) are now the free variables of the optimization problem.

Definition 4.1. A pair (φ,ψ) of functions that solve the dual problem is called a *pair of Kantorovich potentials*.

Remark 4.2. Dual problems are very useful in convex optimization: they (usually) admit a solution even if the original problem doesn't and the study of the relation between the primal and the dual problem gives insight on the former. We will see that this is the case also for the Kantorovich problem.

4.1.1. *Existence of a solution to the dual problem.* We now want to prove that the dual problem has a solution. The idea is, as usual, to start from a maximizing sequence $\{(\varphi_n,\psi_n)\}_{n\in\mathbb{N}}$ and to be able to extract a converging subsequence. Since φ_n and ψ_n are continuous functions, and no derivative is present in the functional to maximize (that usually ensures some compactness), the only (general) result at our disposal is the Ascoli-Arzelà Theorem. We recall it here by first fixing some terminology.

Definition 4.3. Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of continuous functions on a metric space (Z,d). We say that the sequence is

(i) *equi-continuous* if for each $\varepsilon>0$ there exists $\delta>0$ such that

$$|f_n(x)-f_n(y)|<\varepsilon$$

for all $x,y\in Z$ with $\mathrm{d}(x,y)<\delta$ and all $n\in\mathbb{N}$;

(ii) *equi-bounded* if there exists $C \in (0, \infty)$ such that

$$\sup_{x \in Z} |f_n(x)| \leq C$$

for all $n \in \mathbb{N}$.

Remark 4.4. The above conditions mean that the parameters for continuity and boundness do *not* depend on the index $n \in \mathbb{N}$, but they are uniform. Another way to express the equi-continuity is to say that all the functions in the sequence have the same *modulus of continuity*. A modulus of continuity for a function $f : Z \to \mathbb{R}$ is an increasing continuous function $\omega : (0, \infty) \to (0, \infty)$ with $\lim_{t \to 0^+} \omega(t) = 0$ such that

$$|f(x) - f(y)| \leq \omega\left(\mathrm{d}(x, y)\right)$$

for all $x, y \in Z$. Note that this is equivalent to the definition of continuity with ε and δ.

Theorem 4.5 (Ascoli-Arzelà Theorem)**.** *Let $\{f_n\}_{n \in \mathbb{N}}$ be a sequence of equi-continuous and equi-bounded functions on a compact metric space (Z, d). Then there exists a subsequence $\{f_{n_j}\}_{j \in \mathbb{N}}$ and a continuous function $f \in C(Z)$ such that f_{n_j} converges to f uniformly (namely in the sup-norm) as $j \to \infty$, that is*

$$\lim_{j \to \infty} \sup_{x \in Z} |f_n(x) - f(x)| = 0\,.$$

Also the vice-versa holds: if $\{f_n\}_{n \in \mathbb{N}}$ is a sequence such that $f_n \to f$ uniformly for some $f \in C(Z)$, then $\{f_n\}_{n \in \mathbb{N}}$ is equi-convinuous and equi-bounded.

We would like to apply the Ascoli-Arzelà compactness theorem to our case. The problem is that for a *generic* maximizing sequence $\{(\varphi_n, \psi_n)\}_{n \in \mathbb{N}}$ we cannot prove (a priori) that it is equi-continuous and equi-bounded. The idea is to exploit the constrain $\varphi \oplus \psi \leq c$ to modify the maximizing sequence into a new maximizing sequence for which equi-continuity and equi-boundness hold.

Since we want to maximize

$$\int_X \varphi \, d\mu + \int_Y \psi \, d\nu$$

and μ and ν are positive measures, we would like to have φ and ψ as large as possible, in a compatible way with the constrain. The constrain $\varphi \oplus \psi \leq c$ means, in particular, that for each $y \in Y$ it holds

$$\psi(y) \leq c(x, y) - \varphi(x)$$

for all $x \in X$. Thus, given φ, the largest ψ we can take is

$$\psi(y) \coloneqq \inf_{x \in X} \left[\, c(x, y) - \varphi(x) \,\right].$$

In a similar way, given ψ the largest φ we can take is

$$\varphi(y) \coloneqq \inf_{y \in Y} \left[\, c(x, y) - \psi(y) \,\right].$$

These considerations justifies the following definition. In the following we will use the notation $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$. Moreover, in some definitions and results, we will consider a costs $c : X \times Y \to \mathbb{R}$, since non-negativity is not required.

Definition 4.6. Let $c : X \times Y \to \mathbb{R}$ be continuous and let $\varphi : X \to \overline{\mathbb{R}}$. We define $\varphi^c : Y \to \overline{\mathbb{R}}$, the *c-transform* of φ, by

$$\varphi^c(y) \coloneqq \inf_{x \in X} \left[\, c(x, y) - \varphi(x) \,\right].$$

Similarly, given $\psi : Y \to \overline{\mathbb{R}}$, we define $\psi^c : X \to \overline{\mathbb{R}}$, the *c-transform* of ψ, by

$$\psi^c(x) \coloneqq \inf_{y \in Y} \left[\, c(x, y) - \psi(y) \,\right].$$

Finally, we say that a function $f : X \to \overline{\mathbb{R}}$ is *c-concave* if there exists a function $g : Y \to \overline{\mathbb{R}}$ such that $f = g^c$, and that a function $h : Y \to \overline{\mathbb{R}}$ is *c-concave* if there exists $\xi : X \to \overline{\mathbb{R}}$ such that $h = \xi^c$.

Remark 4.7. To be precise, the above definition is an abuse of language: it is correct when c is symmetric in x and y, while in general we should distinguish the c-transform fin X and Y. For our purposes though, this distinction would just be a matter of notation, and therefore we will omit it.

Thanks to the observations above, for each $\varphi \in C(X)$ and $\psi \in C(Y)$, we have that

$$\int_X \varphi \, d\mu + \int_Y \psi \, d\nu \leq \int_X \varphi \, d\mu + \int_Y \varphi^c \, d\nu \leq \int_X \varphi^{cc} \, d\mu + \int_Y \varphi^c \, d\nu \, .$$

Next proposition shows that it is enough to stop iterating taking c-transforms, namely that $f^{ccc} = f$ for any function.

Proposition 4.8. *Let $f : X \to \mathbb{R} \cup \{-\infty\}$. Then $f^{cc} \geq f$ and equality holds if and only if f is c-concave. In particular f^{cc} is the lowest c-concave function that is greater that f, and $f^{ccc} = f^c$.*

The advantage of the c-transform is that its modulus of continuity is the same of that of the cost c. For, we need the following technical result.

Lemma 4.9. *Let $\{f_\alpha\}_\alpha$ be a family of functions (not necessarily countable) on a metric space (Z, d) such that*

$$|f_\alpha(x) - f_\alpha(y)| \leq \omega\left(\mathrm{d}(x, y)\right)$$

for all $x \neq y \in Z$ and all indexes α, where $\omega : (0, \infty) \to (0, \infty)$ is a continuous function. Define

$$f(x) \coloneqq \inf_\alpha f_\alpha(x) \, .$$

Then

$$|f(x) - f(y)| \leq \omega\left(\mathrm{d}(x, y)\right)$$

for all $x \neq y \in Z$.

Lemma 4.10. *Assume that $\omega : (0, \infty) \to (0, \infty)$ is a modulus of continuity for the function $c : X \times Y \to \mathbb{R}$. Then, for each $\varphi : X \to \mathbb{R}$, the function ω is also a modulus of continuity for φ^c and φ^{cc}.*

Proof. We prove the result for φ^c. The same argument works for φ^{cc}. For each $x \in X$, consider the function $g_x : Y \to \mathbb{R}$ defined by

$$g_x(y) \coloneqq c(x, y) - \varphi(x) \, .$$

Then

$$|g_x(y) - g_x(y')| \leq \omega\left(d(y, y')\right)$$

for all $y, y' \in Y$. Therefore, Lemma 4.9 yields that

$$\varphi^c(y) = \inf_{x \in X} g_x(y)$$

satisfies

$$|\varphi^c(y) - \varphi^c(y')| \leq \omega\left(d(y, y')\right)$$

for all $y, y' \in Y$. □

We are now in position to prove the existence of a solution to the dual problem.

Theorem 4.11. *Let X and Y be compact metric spaces, and let $c : X \times Y \to [0, \infty)$ be a continuous cost. Then the dual problem admits a solution.*

Proof. Let $\{(\varphi_n, \psi_n)\}_{n \in \mathbb{N}}$ be a maximizing sequence for the dual problem. Without loss of generality, thanks to Proposition 4.8, we can assume the maximizing sequence to be of the form $\{\varphi_n, (\varphi_n)^c\}_{n \in \mathbb{N}}$ with φ_n c-concave for each $n \in \mathbb{N}$. Namely we have that

$$\int_X \varphi_n \, d\mu + \int_Y (\varphi_n)^c \, d\nu \to \sup \left\{ \int_X \varphi \, d\mu + \int_Y \psi \, d\nu \, : \, \varphi \oplus \psi \leq c \right\} \tag{4.1}$$

as $n \to \infty$. Let $\omega : (0,\infty) \to (0,\infty)$ be a modulus of continuity for c. In particular, from Lemma 4.10 we have that, for all $n \in \mathbb{N}$, φ_n and $(\varphi_n)^c$ have ω as a modulus of continuity.

In particular, the sequence $\{\varphi_n, (\varphi_n)^c\}_{n\in\mathbb{N}}$ is equi-continuous. In order to apply the Ascoli-Arzelà compactness Theorem, we need to prove equi-boudneness of the sequence. For, we note that, for each $t \in \mathbb{R}$ it holds

$$\int_X (\varphi_n - t)\, d\mu + \int_Y ((\varphi_n)^c + t)\, d\nu = \int_X \varphi_n\, d\mu + \int_Y (\varphi_n)^c\, d\nu$$

for all $n \in \mathbb{N}$. In particular, up to subtracting the minimum of φ_n from it, we can assume without loss of generality that

$$\min_X \varphi_n = 0$$

for all $n \in \mathbb{N}$. Thus, if $x_n \in X$ is a point with $\varphi_n(x_n) = 0$, we have, for all $x \in X$ that

$$|\varphi_n(x)| = |\varphi_n(x) - \varphi_n(x_n)| \leq \mathrm{d}(x, x_n) \leq \mathrm{d}\,(\mathrm{diam}(X))$$

where $\mathrm{diam}(X) \coloneqq \sup\{\mathrm{d}(x,x') \,:\, x, x' \in X\}$ is the diameter of the space X, which is finite since X is compact. In particular, $\{\varphi_n\}_{n\in\mathbb{N}}$ is equi-bounded. To prove that also $\{(\varphi_n)^c\}_{n\in\mathbb{N}}$ is equi-bounded we note that, for each $y \in Y$,

$$(\varphi_n)^c(y) = \inf_{x\in X} [\, c - \varphi_n(x)\,] \in [\min c, \max c + \mathrm{d}\,(\mathrm{diam}(X))]$$

since $\min_X \varphi_n = 0$. Thus, also the sequence $\{(\varphi_n)^c\}_{n\in\mathbb{N}}$ is equi-bounded. Thanks to the Ascoli-Arzelà Theorem we get the existence of a subsequence $\{\varphi_{n_j}\}_{j\in\mathbb{N}}$ and of a function $\varphi \in C(X)$ such that φ_{n_j} converges uniformly to φ_∞. By applying again the Ascoli-Arzelà Theorem to the sequence $\{(\varphi_{n_j})^c\}_{j\in\mathbb{N}}$ we get the existence of a subsequence $\{(\varphi_{n_{j_i}})^c\}_{i\in\mathbb{N}}$ and of a function $\psi_\infty \in C(Y)$ such that $(\varphi_{n_{j_i}})^c$ converges to ψ_∞ uniformly. Note that $\varphi_{n_{j_i}}$ keeps converging uniformly to φ_∞. The uniform convergence of the sequences implies that

$$\int_X \varphi_{n_{j_i}}\, d\mu + \int_Y (\varphi_{n_{j_i}})^c\, d\nu \to \int_X \varphi_\infty\, d\mu + \int_Y \psi_\infty\, d\nu\,.$$

as $i \to \infty$. Indeed, fix $\varepsilon > 0$ and let $i_0 \in \mathbb{N}$ be such that for all $i \geq i_0$ it holds

$$\sup_{x\in X} |\varphi_{n_{j_i}}(x) - \varphi_\infty(x)| < \varepsilon\,.$$

Then

$$\begin{aligned}
\left| \int_X \varphi_{n_{j_i}}\, d\mu - \int_X \varphi_\infty\, d\mu \right| &= \left| \int_X \left(\varphi_{n_{j_i}} - \varphi_\infty \right) d\mu \right| \\
&\leq \int_X \left| \varphi_{n_{j_i}} - \varphi_\infty \right| d\mu \\
&\leq \sup_{x\in X} |\varphi_{n_{j_i}}(x) - \varphi_\infty(x)| \mu(X) \\
&\leq \varepsilon \mu(X)
\end{aligned}$$

for $i \geq i_0$, where in the first inequality we used that, for any $f \in C(X)$, it holds that

$$\begin{aligned}
\left| \int_X f\, d\mu \right| &= \left| \int_X f^+\, d\mu - \int_X f^-\, d\mu \right| \\
&\leq \left| \int_X f^+\, d\mu \right| + \left| \int_X f^-\, d\mu \right| \\
&= \int_X f^+\, d\mu + \int_X f^-\, d\mu \\
&= \int_X |f|\, d\mu\,.
\end{aligned}$$

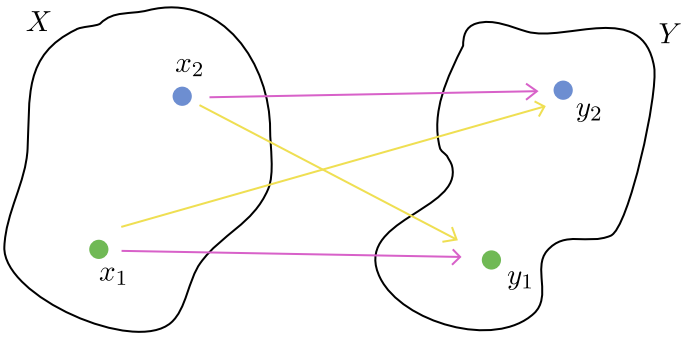

FIGURE 13. Since γ is optimal, moving mass from x_1 to y_1 and from x_2 to y_2 (pink arrows) is more convenient than moving mass from x_1 to y_2 and from x_2 to y_1 (yellow arrows).

Since $\mu(X) = 1 < \infty$ and ε is arbitrary, we conclude that

$$\lim_{i\to\infty} \int_X \varphi_{n_{j_i}}\, d\mu = \int_X \varphi_\infty\, d\mu\,.$$

A similar argument shows that the same holds for the integrals of the other sequence. Thus, thanks to (4.1) we have that

$$\int_X \varphi_\infty\, d\mu + \int_Y \psi_\infty\, d\nu = \sup\left\{ \int_X \varphi\, d\mu + \int_Y \psi\, d\nu \,:\, \varphi\oplus\psi \le c \right\}.$$

Finally, since for all $j \in \mathbb{N}$ it holds

$$\varphi_{n_j}(x) + (\varphi_{n_j})^c(y) \le c$$

for all $x \in X$ and $y \in Y$, and using the fact that $\varphi_{n_{j_i}}$ converges uniformly to φ_∞ and $(\varphi_{n_{j_i}})^c$ converges uniformly to ψ_∞, we get

$$\varphi_\infty(x) + \psi_\infty(y) \le c(x,y)$$

for all $x \in X$ and $y \in Y$. Thus, $(\varphi_\infty, \psi_\infty)$ is an admissible pair for the dual problem and therefore a Kantorovich potential.

Again by using the observations on the c-transform, we can assume φ_∞ to be c-concave, and $\psi_\infty = (\varphi_\infty)^c$. □

4.1.2. *Regularity of optimal transport plans.* Here we prove a regularity result for an optimal transport plan $\gamma \in \Pi(\mu,\nu)$. This property is based on this heuristic observation: assume that that an optimal transport plan $\gamma \in \Pi(\mu,\nu)$ moves some mass from a point x_1 to a point y_1 and from a point x_2 to a point y_2. Since γ is optimal, this means that this choice of moving mass is no worse than moving mass from x_1 to y_2 and from x_2 to y_1 (see Figure 13). Namely

$$c(x_1,y_1) + c(x_2,y_2) \le c(x_1,y_2) + c(x_2,y_1)\,.$$

This observation is at the core of what follows.

Definition 4.12. Let $c : X \times Y \to \mathbb{R}$ be a continuous function. A set $\Gamma \subset X \times Y$ is called *c-cyclically monotone* if

$$\sum_{i=1}^k c(x_i,y_i) \le \sum_{i=1}^k c(x_i, y_{\sigma(i)})$$

for any $k \in \mathbb{N}$, for any $(x_1,y_1),\dots,(x_k,y_k) \in \Gamma$, and for any permutation $\sigma : \{1,\dots,k\} \to \{1,\dots,k\}$.

We now introduce the set that is relevant for a transport plan.

Definition 4.13. Given a Radon measure λ on a separable metric space Z, we define its *support* $\mathrm{supp}(\gamma) \subset Z$ by

$$\mathrm{supp}\,(\gamma) \coloneqq \bigcap \{\, C \subset Z \,:\, Z \text{ closed }, \lambda(X \setminus Z) = 0 \,\}\,.$$

Remark 4.14. Note that the support of a measure is well defined because the separability of the space Z ensures that the above intersection can be taken countable. In particular, the support of a measure is a closed set, being a countable intersection of closed sets. It is the smallest closed set on which the measure is concentrated. Note that this might strictly contain the set $\{\lambda > 0\}$. As an example, consider the measure λ on $\mathbb{R}$ defined by

$$\lambda(E) \coloneqq \mathcal{L}^1\left(E \cap (0,1)\right)\,.$$

Then $\{\lambda > 0\} = (0,1)$, but $\mathrm{supp}(\lambda) = [0,1]$.

Theorem 4.15. *Let $\gamma \in \Pi(\mu,\nu)$ be an optimal transport plan for the continuous cost $c : X \times Y \to [0,\infty)$. Then the set* $\mathrm{supp}(\gamma)$ *is c-cyclically monotone.*

Proof. We argue by contradiction and assume that there exists $k \in \mathbb{N}$, points $(x_1,y_1),\dots,(x_k,y_k) \in \mathrm{supp}(\gamma)$ and a permutation $\sigma : \{1,\dots,k\} \to \{1,\dots,k\}$ such that

$$\sum_{i=1}^{k} c(x_i,y_i) > \sum_{i=1}^{k} c(x_i,y_{\sigma(i)})\,. \tag{4.2}$$

The heuristic idea is the following. For simplicity of exposition of this idea, we assume $k = 2$. We would like to do what is depicted in Figure 13, namely swap the mass from the points x_1 to y_2 and from x_2 to y_1 because the above inequality says that it is more convenient for the cost c. There are two technical points to pay attention to: the first one is that, in general, we cannot just move mass from one point to the other, because μ might not be some Dirac delta at the points x_1 and x_2. Therefore we will consider balls $B_r(x_1)$ and $B_r(y_1)$ and we would like to modify the transport plan γ by moving the mass from $B_r(x_1)$ to $B_r(y_2)$ instead that on $B_r(y_1)$. The other technical point to pay attention to (see Figure 14) is that the mass that γ moves from $B_r(x_1)$ to $B_r(y_1)$ might be different from the mass that γ moves from $B_r(x_2)$ to $B_r(y_2)$. Thus, we will just swap a bit of that mass.

We proceed as follows. Fix $\varepsilon > 0$ that will be chosen later. Thanks to the continuity of c, let $r > 0$ be such that, for all $i = 1,\dots,k$,

$$c(x,y) \geq c(x_i,y_i) - \varepsilon \tag{4.3}$$

for all $(x,y) \in B_r(x_i) \times B_r(y_i)$, and such that

$$c(x,y) \leq c(x_i,y_{\sigma(i)}) + \varepsilon \tag{4.4}$$

for all $(x,y) \in B_r(x_i) \times B_r(y_{\sigma(i)})$. Set $V_i \coloneqq B_r(x_i) \times B_r(y_i)$ for all $i = 1,\dots,k$. Up to further decrease the value of $r > 0$, we can also assume that $V_i \cap V_j = \emptyset$ if $i \neq j$.

Define the measure $\lambda_i \in \mathcal{P}(X \times Y)$ by

$$\lambda_i(E) \coloneqq \frac{\gamma(E \cap V_i)}{\gamma(V_i)}\,,$$

for each Borel set $E \subset X \times Y$. Note that $\gamma(V_i) > 0$ for each $i = 1 \dots,k$ since $(x_i,y_i) \in \mathrm{supp}(\gamma)$. In particular, $\gamma(V_i)$ is the mass that γ moves from $B_r(x_i)$ to $B_r(y_i)$. The reason why we divide by $\gamma(V_i)$ is because we want to work with probability measures in order to have a way to swap the mass. Moreover, let

$$\mu_i \coloneqq (\pi_1)_{\#}\lambda_i, \qquad\qquad \nu_i \coloneqq (\pi_2)_{\#}\lambda_i\,,$$

for each $i = 1,\dots,k$. We would like to move the mass μ_i to match $\nu_{\sigma(i)}$. There are several ways to do that. We choose to do it with the measure $\widetilde{\gamma}_i \coloneqq \mu_i \otimes \nu_{\sigma(i)}$. Now the issue is that these

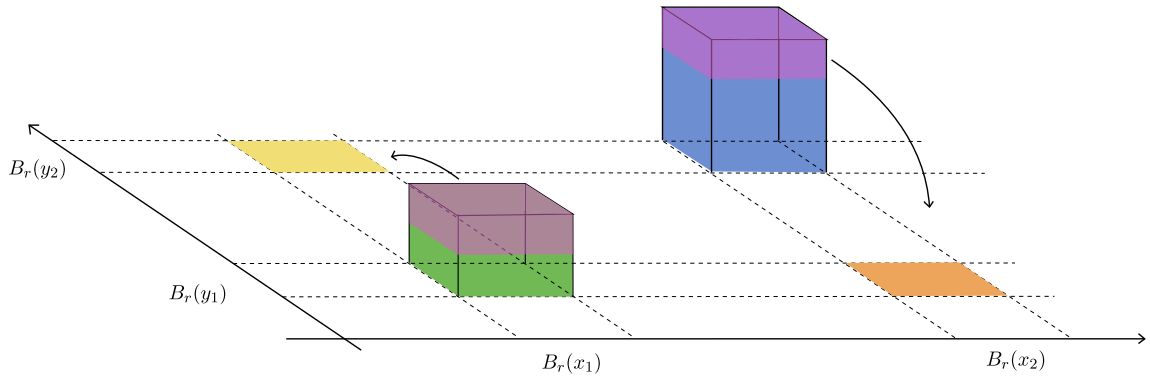

FIGURE 14. The idea of the proof. For simplicity, we assume $k = 2$. We would like to *move* the green and pink paralleleped to the yellow region and the blue and pink parallelepiped to the orange region. Since these two amounts of mass might be different, we just move the pink parallelepipeds.

masses could be different, and thus swapping the whole mass, that translates into considering the transport plan

$$\gamma - \sum_{i=1}^{k} \lambda_i + \sum_{i=1}^{k} \widetilde{\gamma}_i$$

might result in having the above modified transport plan negative in some regions, namely on those where the mass originally moved $\nu_i(X \times Y)$ is larger than the one that we would like to move, $\nu_{\sigma(i)}(X \times Y)$. To solve this problem we swap just a bit of that mass and this is enough to reach the desired contradiction. For $\delta > 0$, that will be chosen later, define

$$\widetilde{\gamma} := \gamma - \delta \sum_{i=1}^{k} \lambda_i + \delta \sum_{i=1}^{k} \widetilde{\gamma}_i \,.$$

We now prove that it is possible to choose $\delta > 0$ in such a way that $\widetilde{\gamma}$ is a non-negative measure. For each Borel set $E \subset X \times Y$, we have that

$$\begin{aligned}
\widetilde{\gamma}(E) &= \gamma(E) - \delta \sum_{i=1}^{k} \lambda_i(E) + \delta \sum_{i=1}^{k} \widetilde{\gamma}_i(E) \\
&\geq \gamma(E) - \delta \sum_{i=1}^{k} \frac{\gamma(E \cap V_i)}{\gamma(V_i)} \\
&\geq \gamma(E) - \frac{\delta}{\min_i \gamma(V_i)} \sum_{i=1}^{k} \gamma(E \cap V_i) \\
&\geq \left[1 - \frac{\delta}{\min_i \gamma(V_i)} \right] \gamma(E),
\end{aligned}$$

where in the last step we used the fact that the V_i's are disjoint. Therefore, by choosing (obviously we do not move more mass that the minimum mass that could be moved!)

$$\delta < \min_{i=1,\dots,k} \gamma(V_i),$$

we get that $\gamma(E) \geq 0$ for each Borel set $E \subset X \times Y$. Moreover, it is easy to see that $\widetilde{\gamma} \in \Pi(\mu, \nu)$. Indeed, by the linearity of the push-forward measure, we have that

$$(\pi_1)_\# \widetilde{\gamma} = (\pi_1)_\# \gamma - \delta \sum_{i=1}^k (\pi_1)_\# \lambda_i + \delta \sum_{i=1}^k (\pi_1)_\# \widetilde{\gamma}_i = \mu - \delta \sum_{i=1}^k \mu_i + \delta \sum_{i=1}^k \mu_i = \mu \,,$$

and

$$(\pi_2)_\# \widetilde{\gamma} = (\pi_2)_\# \gamma - \delta \sum_{i=1}^k (\pi_2)_\# \lambda_i + \delta \sum_{i=1}^k (\pi_2)_\# \widetilde{\gamma}_i = \nu - \delta \sum_{i=1}^k \nu_i + \delta \sum_{i=1}^k \nu_i = \nu \,.$$

Finally, we prove that it is possible to choose $\varepsilon > 0$ in such a way to reach the desired contradiction. Indeed, by using (4.3) and (4.4), we have

$$\begin{aligned} \int_{X\times Y} c \, d\gamma - \int_{X\times Y} c \, d\widetilde{\gamma} &= \delta \sum_{i=1}^k \int_{X\times Y} c \, d\lambda_i - \delta \sum_{i=1}^k \int_{X\times Y} c \, d\widetilde{\gamma}_i \\ &\geq \delta \left[\sum_{i=1}^k \big(c(x_i, y_i) + \varepsilon \big) - \sum_{i=1}^k \big(c(x_i, y_{\sigma(i)}) - \varepsilon \big) \right] \\ &= \delta \left[\sum_{i=1}^k c(x_i, y_i) - \sum_{i=1}^k c(x_i, y_{\sigma(i)}) - 2k\varepsilon \right] . \end{aligned}$$

Therefore, by choosing

$$0 < \varepsilon < \frac{1}{2k} \left[\sum_{i=1}^k c(x_i, y_i) - \sum_{i=1}^k c(x_i, y_{\sigma(i)}) \right] ,$$

which is possible by (4.2), we get

$$\int_{X\times Y} c \, d\gamma > \int_{X\times Y} c \, d\widetilde{\gamma}$$

that contradicts the minimality of the transport plan γ. □

Finally, we prove that a c-cyclically monotone set $\Gamma \subset X \times Y$ is contained in the sub-differential of a c-concave function.

Theorem 4.16. *Let $\Gamma \subset X \times Y$ be non-empty and c-cyclically monotone. Then there exists a c-concave function $\varphi : X \to \mathbb{R} \cup \{-\infty\}$, not identically $-\infty$, such that*

$$\Gamma \subset \{ \, (x, y) \in X \times Y \, : \, \varphi(x) + \varphi^c(y) = c(x, y) \, \} \, .$$

Proof. First of all, we would like to understand the idea behind the construction of the function φ that does the job. Since

$$\varphi(x) + \varphi^c(y) \leq c(x, y)$$

for all $x \in X$ and $y \in Y$, we have to find a c-concave function φ such that

$$\varphi(x) + \varphi^c(y) \geq c(x, y) \tag{4.5}$$

for all $(x, y) \in \Gamma$. Assume that such a function exists, and let's see if we can find conditions on what it has to be. Since adding constants to φ means subtracting the same quantity from φ^c, we can fix a point $(x_0, y_0) \in \Gamma$, and assume that $\varphi(x_0) = 0$. Fix $(x_n, y_n) \in \Gamma$. By definition of $\varphi^c(y_n)$ since X is compact and c is continuous (and, in turn also φ, being c-concave) there exists $x \in X$ such that

$$\varphi^c(y_n) = c(x, y_n) - \varphi(x) \, .$$

By using this equality in (4.5) (at the point (x_n, y_n)), we get

$$\varphi(x_n) + c(x, y_n) - \varphi(x) \geq c(x_n, y_n) \, ,$$

that we re-write as

$$\varphi(x) - \varphi(x_n) \leq c(x, y_n) - c(x_n, y_n) \, . \tag{4.6}$$

We would like to impose this condition for all $x \in X$ and $(x_n, y_n) \in \Gamma$. In this case, taking $(x_{n-1}, y_{n-1}) \in \Gamma$ we have

$$\varphi(x_n) - \varphi(x_{n-1}) \leq c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1})\,. \tag{4.7}$$

Thus, by summing (4.6) and (4.7) we get

$$\varphi(x) - \varphi(x_{n-1}) \leq c(x, y_n) - c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1})\,.$$

This can be done as many times as we want. In particular, we can choose any *path*

$$(x_n, y_n), (x_{n-1}, y_{n-1}), \dots, (x_0, y_0)$$

of points in Γ, to get

$$\begin{aligned}\varphi(x) \leq c(x, y_n) - c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + c(x_{n-1}, y_{n-2}) - c(x_{n-2}, y_{n-2})\\ + \dots + c(x_0, y_1) - c(x_0, y_0)\,,\end{aligned}$$

where on the left-hand side we used the fact that $\varphi(x_0) = 0$.

Therefore, it comes natural to define

$$\begin{aligned}\varphi(x) \coloneqq \inf\Big\{ c(x, y_n) - c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + c(x_{n-1}, y_{n-2}) - c(x_{n-2}, y_{n-2})\\ + \dots + c(x_0, y_1) - c(x_0, y_0) \,:\, (x_i, y_i) \in \Gamma \text{ for all } i \in \{1, \dots, n\},\, n \in \mathbb{N} \Big\}\,.\end{aligned}$$

Note that $\varphi(x_0) = 0$. Indeed, since Γ is c-cyclically monotone, by using the permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ defined by $\sigma(i) \coloneqq (i-1) \bmod n$, we obtain that $\varphi(x_0) \geq 0$. By choosing $n = 1$ and $(x_1, y_1) = (x_0, y_0)$ we get the desired conclusion. In particular, $\varphi \not\equiv -\infty$. Moreover, since Γ is not empty (and thus we are infimizing over a non-empty set), and c is never $+\infty$, we get $\varphi(x) < +\infty$ for all $x \in X$.

We now want to prove that φ is c-concave and that (4.5) holds. It is possible to prove that φ is c-concave, namely that there exists $\xi : Y \to \mathbb{R} \cup \{\infty\}$ such that $\varphi = \xi^c$, by explicitly defining ξ. Indeed, such a ξ would have to be such that

$$\varphi(x) = \inf_{y \in Y} [c(x, y) - \xi(y)]$$

for all $x \in X$. In particular, by using the definition of φ, we have the equation

$$\begin{aligned}&\inf\Big\{ c(x, y_n) - c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + c(x_{n-1}, y_{n-2}) - c(x_{n-2}, y_{n-2})\\ &\qquad + \dots + c(x_0, y_1) - c(x_0, y_0) \,:\, (x_i, y_i) \in \Gamma \text{ for all } i \in \{1, \dots, n\},\, n \in \mathbb{N} \Big\}\\ &= \inf_{y \in Y} [c - \xi(y)]\,.\end{aligned}$$

In particular, if on the right-hand side we consider points $y_n \in Y$ such that there exists $x_n \in X$ with $(x_n, y_n) \in \Gamma$, we get that ξ has to satisfy

$$\begin{aligned}-\xi(y_n) = \inf\Big\{ -c(x_n, y_n) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + c(x_{n-1}, y_{n-2}) - c(x_{n-2}, y_{n-2})\\ + \dots + c(x_0, y_1) - c(x_0, y_0) \,:\, (x_i, y_i) \in \Gamma \text{ for all } i \in \{1, \dots, n\},\, n \in \mathbb{N} \Big\}\,.\end{aligned}$$

We thus define

$$\begin{aligned}-\xi(y) \coloneqq \inf\Big\{ -c(x_n, y) + c(x_n, y_{n-1}) - c(x_{n-1}, y_{n-1}) + c(x_{n-1}, y_{n-2}) - c(x_{n-2}, y_{n-2})\\ + \dots + c(x_0, y_1) - c(x_0, y_0) \,:\, (x_i, y_i) \in \Gamma \text{ for all } i \in \{1, \dots, n-1\},\\ x_n \in \pi_1(\Gamma),\, n \in \mathbb{N},\, \Big\}\,.\end{aligned}$$

Note that, by using the fact that Γ is c-cyclically monotone, we get that $\xi(y) \geq 0$ for all $y \in \pi_2(\Gamma)$. Moreover, from the above considerations, we obtain that $\varphi = \xi^c$.

Finally, we prove (4.5). Since $\varphi = \xi^c$, then by Proposition 4.8, we get $\varphi^c \geq \xi$. Therefore, if we prove

$$\varphi(x) + \xi(y) \geq c(x,y)$$

for all $(x,y) \in \Gamma$, we are done. Fix $(\overline{x},\overline{y}) \in \Gamma$ and let $\varepsilon > 0$. Since $\varphi = \xi^c$, there exists $y \in Y$ such that

$$\varphi(\overline{x}) \geq c(\overline{x},y) - \xi(y) - \varepsilon\,. \tag{4.8}$$

Moreover, from the definition of ξ, we have that

$$-\xi(y) + \xi(\overline{y}) \leq c(\overline{x},\overline{y}) - c(\overline{x},y)\,. \tag{4.9}$$

Indeed, given any admissible path for the infimum defining $-\psi(y)$, we can construct an admissible path for the infimum defining $-\psi(\overline{y})$ by adding the points $(\overline{x},\overline{y})$ and $(\overline{x},y)$. By thus taking a minimizing path for $-\psi(y)$, we get the desired inequality. Thus, from (4.8) and (4.9) we obtain

$$\varphi(x) + \xi(y) \geq c(x,y) - \varepsilon\,.$$

Since $\varepsilon > 0$ is arbitrary, we conclude. □

Remark 4.17. In the proof above it might seem that the only property that we used was that

$$\sum_{i=0}^{n} c(x_i,y_i) \leq \sum_{i=1}^{n} c(x_i,y_{i-1})$$

where $y_{-1} = y_n$. Actually, just by renaming the points it is possible to obtain *any* permutation.

4.1.3. *The duality formula.* We are now in position to prove the duality formula.

Theorem 4.18. *Let $c : X \times Y \to [0,\infty)$ be a continuous cost, and X and Y are compact metric spaces. Then*

$$\min\left\{ \int_{X\times Y} c\,d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\} = \max\left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu \,:\, \varphi\oplus\psi \leq c \right\}.$$

Proof. We first prove the inequality $\geq$. Let (φ_0,φ_0^c) be a pair of Kantorovich potentials, that we know exists by Theorem 4.11. Then, for each $\gamma \in \Pi(\mu,\nu)$, we have that

$$\begin{aligned}\max\left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu \,:\, \varphi\oplus\psi \leq c \right\} &= \int_X \varphi\,d\mu + \int_Y \varphi^c\,d\nu \\ &= \int_{X\times Y} [\,\varphi_0(x) + \varphi_0^c(y)\,]\,d\gamma \\ &\leq \int_{X\times Y} c\,d\gamma\,,\end{aligned}$$

where in the previous to last step we used the constraint $\varphi_0 \oplus \varphi_0^c \leq c$. Thus

$$\max\left\{ \int_X \varphi\,d\mu + \int_Y \psi\,d\nu \,:\, \varphi\oplus\psi \leq c \right\} \leq \min\left\{ \int_{X\times Y} c\,d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\}.$$

To prove the opposite inequality, let $\gamma_0 \in \Pi(\mu,\nu)$ be an optimal transport plan, that we know exists by Theorem 3.7. By Theorem 4.15 we know that $\mathrm{supp}(\gamma_0)$ is c-cyclically monotone. Thus, Theorem 4.16 yields the existence of a c-concave function $\varphi : X \to \mathbb{R}\cup\{-\infty\}$ such that

$$\mathrm{supp}(\gamma_0) \subset \{\,(x,y) \in X\times Y \,:\, \varphi(x) + \varphi^c(y) = c\,\}\,.$$

Since φ is c-concave, it is continuous (see Lemma 4.10). The same holds for φ^c. Moreover, we have that the constrain $\varphi\oplus\varphi^c \leq c$ is satisfied. Therefore, the couple (φ,φ^c) is an admissible competitor for the dual problem. Thus

$$\begin{aligned}\min\left\{ \int_{X\times Y} c\,d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\} &= \int_{X\times Y} c\,d\gamma_0 \\ &= \int_{X\times Y} [\,\varphi(x) + \varphi^c(y)\,]\,d\gamma_0\end{aligned}$$

$$= \int_X \varphi \, d\mu + \int_Y \varphi^c \, d\nu$$
$$\leq \max \left\{ \int_X \varphi \, d\mu + \int_Y \psi \, d\nu \, : \, \varphi \oplus \psi \leq c \right\} .$$

This proves the opposite inequality and concludes the proof. □

4.2. **Strictly convex costs.** We now turn our attention to proving the existence of a solution to the Monge problem in the case $X, Y \subset \mathbb{R}^N$ are Borel sets and the cost function $c : X \times Y \to [0, \infty)$ is of the form $c = h(x - y)$, with h strictly convex.

We would like to be able to say the following: for any $x_0 \in \mathrm{supp}(\mu)$ we want to find a *unique* $y_0 \in \mathrm{supp}\,(\nu)$ such that $(x_0, y_0) \in \mathrm{supp}(\gamma)$. In this way, if everything is nice enough, we can construct a map $T : X \to Y$ such that $T(x_0) = y_0$ for any $x_0 \in \mathrm{supp}(\mu)$, where y_0 is the point found above. Thus, we can conclude that $\gamma = (\mathrm{Id}, T)_{\#}\mu$. The idea to do this is the following. Let $\gamma \in \Pi(\mu, \nu)$ be an optimal transport plan and let (φ, φ^c) be a pair of Kantorovich potentials for the dual problem. Then

$$\varphi(x) + \varphi^c(y) \leq c \, , \qquad \text{for } (x, y) \in X \times Y \, ,$$

because of the constrain $\varphi \oplus \varphi^c \leq c$, and

$$\varphi(x) + \varphi^c(y) = c \, , \qquad \text{for } (x, y) \in \mathrm{supp}(\gamma) \, ,$$

thanks to Theorem 3.7 and Theorem 4.15. Fix a point $(x_0, y_0) \in \mathrm{supp}(\gamma)$ and consider the function $\xi : X \to \mathbb{R} \cup \{-\infty\}$ defined by $\xi(x) := \varphi(x) - c(x, y_0)$. Then the point $x_0 \in X$ is a point of maximum for ξ. In particular, if $x \mapsto c(y_0, x)$ and φ are both differentiable at x_0, and $x_0 \notin \partial X$, then

$$0 = \nabla \xi(x_0) = \nabla \varphi(x_0) - \nabla_x c(x_0, y_0) \, .$$

We would like to use the above condition in order to obtain *the unique* point $y_0 \in \mathrm{supp}(\nu)$ by inverting $y_0 \to \nabla_x c(\cdot, y_0)$. The question is then: when is it possible to invert $\nabla_x c(\cdot, y_0)$? And when can we say that $x \mapsto c(x, y_0)$ and φ are both differentiable at x_0, for all points $x_0 \in \mathrm{supp}(\mu)$?

We start by recalling a fundamental result in Analysis.

Theorem 4.19 (Rademacher's Theorem)**.** *Let $h : \mathbb{R}^N \to \mathbb{R}$ be a locally Lipschitz map, namely for each compact set $K \subset \mathbb{R}^N$ there exists $C_K < \infty$ such that*

$$\sup_{x \neq y} \frac{|h(x) - h(y)|}{|x - y|} \leq C_K \, .$$

Then h is differentiable $\mathcal{L}^N$-almost everywhere. Namely if $D \subset \mathbb{R}^N$ is the set of points $x \in \mathbb{R}^N$ such that the limit

$$\nabla h(x) := \lim_{y \to x} \frac{h(y) - h(x)}{y - x}$$

exists, then $\mathcal{L}^N(\mathbb{R}^N \setminus D) = 0$. In particular, the function $\nabla h : \mathbb{R}^N \to \mathbb{R}^N$ is Lebesgue measurable.

We now introduce the class of functions that we will use.

Definition 4.20. A function $h : \mathbb{R}^N \to \mathbb{R}$ is said to be *convex* if

$$h(tx + (1 - t)y) \leq th(x) + (1 - t)h(y)$$

for any $x, y \in \mathbb{R}^N$ and $t \in [0, 1]$. The function h is said to be *strictly convex* if the above inequality is strict for any $x \neq y$ and $t \in (0, 1)$.

Strictly convex functions have the good taste to satisfy all the assumptions that we need.

Proposition 4.21. *Let $h : \mathbb{R}^N \to \mathbb{R}$ be a convex function. Then h is locally Lipschitz. Moreover, if h is strictly convex, then the map $x \mapsto \nabla h(x)$ is invertible. Moreover, the function $(\nabla h)^{-1} : \mathbb{R}^N \to \mathbb{R}^N$ is Lebesgue measurable.*

We are now in position to prove the existence of a solution to the Monge problem. In order to avoid the use of the subdifferential of a convex function, we assume h to be C^1, even if the result holds true even without that assumption.

Theorem 4.22. *Let $X, Y \subset \mathbb{R}^N$ be compact sets, and assume $\mathcal{L}^N(\partial X) = 0$ (in particular, this is satisfied if ∂X is regular). Assume that the probability measure μ is of the form $\mu = f\mathcal{L}^N$ for some non-negative function $f \in L^1(\mathbb{R}^N; \mathcal{L}^N)$ with $\int_X f\,dx = 1$. Finally, assume that the costs $c : X \times Y \to [0, \infty)$ is of the form*

$$c(x, y) = h(x - y)$$

where $h : \mathbb{R}^N \to [0, \infty)$ is a strictly convex function. Then the Monge problem admits a unique solution, and the solution $T : X \to Y$ satisfies

$$T(x) = x - \nabla h^{-1}\left(\nabla \varphi(x)\right)$$

for each $x \in X$ with $f(x) > 0$ and such that h is differentiable at x, where $\varphi \in C(X)$ is a (Lipschitz) Kantorovich potential for the dual problem.

Proof. Thanks to Theorem 4.11 there exists a pair of Kantorovich potentials (φ, φ^c) where φ is c-concave.

Step 1. First we want to make sure that we can differentiate φ for μ-a.e. in X. Note that, since $\mu = f\mathcal{L}^N$, if we prove differentiability $\mathcal{L}^N$-a.e. in X, this implies differentiability μ-a.e. in X. This is another point where we use the fact that μ is *not* concentrated on a set with zero Lebesgue measure, since that could be part of the set where φ is *not* differentiable (this assumption could be relaxed though).

By Proposition 4.21, the function c is locally Lipschitz (composition of locally Lipschitz functions) and thus, by Rademacher's Theorem (Theorem 4.19) it is differentiable $\mathcal{L}^N$-a.e. in X. This implies that the c-concave function φ is also locally Lipschitz, and thus differentiable $\mathcal{L}^N$-a.e. in X.

Step 2. Let $\gamma \in \Pi(\mu, \nu)$ be a solution to the Kantorovich problem. In particular, note that if $D \subset X$ is such that

$$\int_D f\,dx = 0$$

then $\gamma(D \times Y) = 0$. In particular, by setting $D \subset X$ to be the union of ∂X with the set of points where h or φ are not differentiable, we get that $\mathcal{L}^N(D) = 0$, and thus $\gamma(D \times Y) = 0$. Without loss of generality, by using the Borel regularity of μ, we can assume D to be a negligible Borel set.

Let $(x_0, y_0) \in \operatorname{supp}(\gamma) \setminus (D \times Y)$. Thanks to the observations above, we get that $x_0 \in X$ is a point of minimum of the function $x \mapsto \varphi(x) - h(x - y_0)$. Since $x_0 \in X \setminus D$, we get that

$$\nabla \varphi(x_0) = \nabla_x c(x_0, y_0) = \nabla h(x_0 - y_0)\,. \tag{4.10}$$

Using the strict convexity of h, Proposition 4.21 ensures that ∇h (which exists everywhere, since h is C^1) is invertible, and thus we obtain

$$y_0 = x_0 - \nabla h^{-1}\left(\nabla \varphi(x_0)\right)\,.$$

Fix a point $\overline{y} \in Y$ and define the function $T : X \to Y$ by

$$T(x) := \begin{cases} x - \nabla h^{-1}\left(\nabla \varphi(x)\right) & \text{if } x \in X \setminus D, \\ \\ \bar{y} & \text{otherwise}\,. \end{cases}$$

Since D is negligible and ∇h^{-1} is Lebesgue measurable, the function T is a Lebesgue measurable function. We then proved that $\operatorname{supp}(\gamma) \setminus (D \times Y)$ is contained in the graph of the function T. In particular, since $\gamma(D \times Y) = 0$, we get that the optimal transport plan γ is induced by the

transport map T, and thus that T is an optimal map for the Monge problem.

Step 3. We finally prove uniqueness. In step 2 we showed that any optimal transport plan is actually induced by a transport map. Let γ_1 and γ_2 be two optimal transport plans. Fix $x_0 \in X \setminus D$ such that there exist $y_1, y_2 \in Y$ with $(x_0, y_1) \in \mathrm{supp}(\gamma_1)$ and $(x_0, y_2) \in \mathrm{supp}(\gamma_2)$. Thanks to (4.10) we get

$$\nabla h(x_0 - y_1) = \nabla h(x_0 - y_2)\,,$$

and thus, by the injectivity of ∇h, that $y_1 = y_2$. This proves the uniqueness of the optimal transport plan and, in turn, of the optimal transport map. □

Remark 4.23. The class of costs $c := |x-y|^p$ with $p > 1$ satisfied the assumptions of the above theorem, and thus the related Monge problem admits a (unique) solution.

4.3. **Strictly concave costs.** You are very familiar with the following situation: if you take a public or private transport, the more you travel, the less you spend in proportion. In mathematical terms, the cost of the service is sub-additive: it costs more for two people to go from two cities that are 100km far away from each other, than for a single person to travel for 200km.

In this section we briefly address the case of existence of a solution to the Monge problem for strictly concave costs, that is for costs of the form

$$c(x, y) = l(|x - y|),$$

for $x, y \in \mathbb{R}^N$, where $l : [0, \infty) \to [0, \infty)$ is a *strictly concave and increasing* function with $l(0) = 0$. Notice that in this case this function is composed with $|x - y|$ and not with the more general $x - y$ as for the strictly convex case.

Remark 4.24. Under the assumptions above, the function $c : \mathbb{R}^N \times \mathbb{R}^N \to [0, \infty)$ is a *distance* on $\mathbb{R}^N$ that satisfies the strict triangle inequality. Namely, it holds

- $c(x, x) = 0$;
- $c < c(x, z) + c(z, y)$ for $x, y, z, \in \mathbb{R}^N$ three distinct points.

The idea is to use the same strategy as for strictly convex case. Namely noticing that, for a point (x_0, y_0) in the support of an optimal transport plan for the Kantorovich problem, if $x \mapsto c(y_0, x)$ and φ are both differentiable at x_0, and $x_0 \notin \partial X$, then

$$0 = \nabla \xi(x_0) = \nabla \varphi(x_0) - \nabla_x c(x_0, y_0) = \nabla \varphi(x_0) - l'(|x_0 - y_0|)\frac{x_0 - y_0}{|x_0 - y_0|}\,.$$

Note that it is possible to obtain y_0 by inverting the above relation as follows: since

$$l'(|x_0 - y_0|)\frac{x_0 - y_0}{|x_0 - y_0|} = \nabla \varphi(x_0)$$

and $l'(t) \neq 0$ for all $t \in (0, \infty)$ (if $t = 0$ then $y_0 = x_0$), we have that

$$y_0 = x_0 - \frac{1}{l'(|x_0 - y_0|)}\nabla \varphi(x_0)\,.$$

The main technical difficulty in order to apply rigorously the above strategy is the lack of differentiability of the costs at the origin. This reflects in a lack of differentiability of the Kantorovich potentials. This technical point is solved by using the notion of approximate gradient. One quick fix for this is to avoid this singularity to take place by assuming μ and ν with disjoint support (so that the cost l is never computed at the point $t = 0$). In order to consider a more general situation, we first establish an important property that distinguish the concave costs from the convex ones: common mass does not move!

Theorem 4.25. *Let $\gamma \in \Pi(\mu, \nu)$ be an optimal transport plan for the Kantorovich problem with cost*

$$c(x, y) = l(|x - y|)$$

with $l : [0,\infty) \to [0,\infty)$ strictly concave, increasing, and with $l(0) = 0$. Let γ_D denotes the restriction of γ to the diagonal $D := \{(x,x) : x \in \mathbb{R}^N\}$. Then

$$\gamma_D = (\mathrm{Id}, \mathrm{Id})_\#(\mu \wedge \nu)$$

where $\mu \wedge \nu := \mu - (\mu - \nu)_+$.

Remark 4.26. In the section on Measure Theory we did not treat signed measures, because they are not relevant for the content of the course. Other than in this specific point, where we need them to rigorously define the *common mass*! To quickly justify the definition, think about the case where $\mu = f\mathcal{L}^N$ and $\nu = g\mathcal{L}^N$. Then the shared mass between μ and ν is given by the measure that considers the *minimum* density at each point. Namely the quantity $\min\{f,g\}$. Since by notation, $f \wedge g := \min\{f,g\}$, we get that the common mass between μ and ν is $(f \wedge g)\mathcal{L}^N$. This justifies the notation $\mu \wedge \nu$. Moreover, noticing that

$$\min\{f,g\} = f - (f-g)_+$$

where $(x)_+ := x$ if $x \geq 0$ and 0 otherwise, it is easy to be persuaded that the same can be done with general measures.

Finally, we note that the condition *the common mass does not move* writes mathematically as $\gamma_D = (\mathrm{Id}, \mathrm{Id})_\#(\mu \wedge \nu)$.

Proof of Theorem 4.25. It is easy to see that $\gamma_D \leq (\mathrm{Id}, \mathrm{Id})_\#(\mu \wedge \nu)$. Namely that for each set $E \subset X \times Y$ it holds $\gamma_D(E) \leq (\mathrm{Id}, \mathrm{Id})_\#(\mu \wedge \nu)(E)$. Suppose by contradiction that

$$\gamma_D < (\mathrm{Id}, \mathrm{Id})_\#(\mu \wedge \nu).$$

This means that some common mass has been moved somewhere. Moreover, since we must have $(\pi_1)_\#\gamma = \mu$ and $(\pi_2)_\#\gamma = \nu$, the common mass that has been moved has to be replaced by mass moved there from somewhere else. In more mathematical terms, there exists a point $z \in X$ belonging to the support of the common mass $\mu \wedge \nu$ with the following property: there exists a point $x \neq z \in X$ where mass is moved from and sent to the point z, and a point $y \neq z \in X$ to which some mass of z is moved to (see Figure 15). Moreover note that $(x,z) \notin D$, and $(z,y) \notin D$.

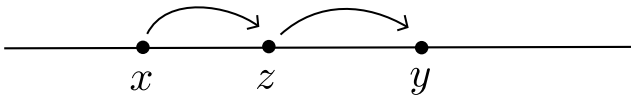

FIGURE 15. The point z to which some mass from x is moved, and from which some mass is moved to the point y.

Since γ is an optimal transport plan, we know that its support is c-cyclically monotone (see Theorem 4.15). Thus, we have that

$$c(x,z) + c(z,y) \leq c + c(z,z)$$

that writes as

$$l(|x-z|) + l(|z-y|) \leq l(|x-y|) + l(|z-z|) = l(|x-y|)$$

since we are assuming $l(0) = 0$. But the strict concavity of l implies that

$$l(|x-y|) < l(|x-z|) + l(|z-y|)$$

thus giving the desired contradiction. □

Remark 4.27. In particular, the above results implies that if in the region where there is common mass, μ is larger than the common mass, then necessarily the optimal transport plan cannot be induced by a map (see Figure 16). Indeed, if $A \subset X$ is contained in the support of the common mass, and for which $\mu(A) > \mu \wedge \nu(A)$, then on the one hand for each point $x \in A$ the

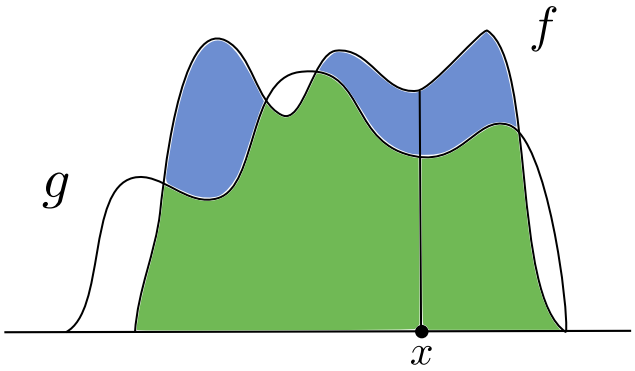

FIGURE 16. The common mass of the measure with density f and that with density g is depicted in green. In blue is the *excess* mass, namely that mass that forces the transport plan at the point x to leave the green mass where it is, but to move the blue one somewhere else. This prevents the transport plan to be induced by a transport map. This is why, by knowing what happens for the common mass, it is possible to remove it and to consider the optimal transport problem for the remaining mass.

common mass has to stay in place, and the *exceeding* mass $\mu - \mu \wedge \nu$ has to be moved somewhere. The idea is then to see what happens when the common mass is removed from μ, namely when we consider $\mu - \mu \wedge \nu$.

We can now state a result for strictly concave costs saying that the common mass does not move, and if we consider the problem of moving the remaining mass, that can be done with a map.

Theorem 4.28. *Let $c(x,y) = l(|x-y|)$ with $l : [0,\infty) \to [0,\infty)$ strictly concave, increasing, with $l(0) = 0$, and of class C^1 on $(0,\infty)$. Moreover, assume that $\mu - \mu \wedge \nu = f\mathcal{L}^N$ for some non-negative function $f \in L^1(\mathbb{R}^N; \mathcal{L}^N)$. Then there exists a unique optimal transport plan γ, and it ig given by*

$$\gamma = (\mathrm{Id}, \mathrm{Id})_\#(\mu \wedge \nu) + (\mathrm{Id}, T)_\#(\mu - \mu \wedge \nu)\,,$$

where $T : \mathbb{R}^N \to \mathbb{R}^N$ is a measurable map.

4.4. **Linear costs and more general costs.** Here we briefly address the case of the original problem of Monge, namely the linear cost $c = |x - y|$. Note that this is not included in any of the previous cases, due to the *strict* convexity or concavity. In particular, note that by using the strategy outlined at the beginning of the previous section we get

$$\nabla\varphi(x_0) = \frac{x_0 - y_0}{|x_0 - y_0|}\,.$$

Thus we know in what *direction* each y_0 has to be, but we do not know how far is from x_0. This lack of knowledge prevents us from concluding. In particular, we know that there are cases of optimal transport plans for the linear cost that are not induced by an optimal map. Nevertheless, it is possible to prove that there exists *an* optimal transport plan that is induced by an optimal transport map.

Theorem 4.29. *There exists a solution to the original Monge problem.*

The proof of this theorem was firstly given by Sudakok in [18] and subsequently fixed by Ambrosio in [**?**].

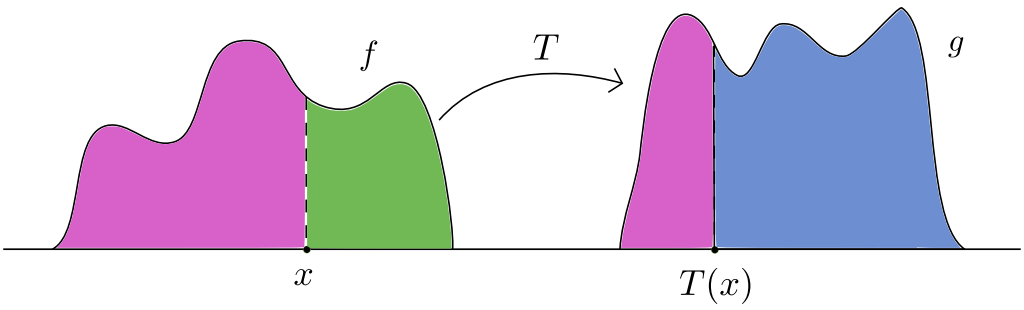

FIGURE 17. For a monotone map $T : \mathbb{R} \to \mathbb{R}$ to transform f into g, we want the two pink regions to have the same mass.

Finally, we mention that for general costs, it is difficult to state that the Monge problem has a solution. Notable cases for which it is possible are the following:

- convex cost (not strictly convex!);
- for crystalline norms;
- the relativistic cost $c := 1 - \sqrt{1 - |x - y|^2}$.

4.5. **The one dimensional case.** In one dimension it is possible to solve more explicitly the Monge problem. In particular, we will consider the case where the measures μ and ν are possibly supported in the *whole* space $\mathbb{R}$.

To understand the main idea of this section, let us consider a convex cost $c(x, y) = h(x - y)$, two measures $\mu = f\mathcal{L}^1$ and $\nu = g\mathcal{L}^1$. We would like to know if it is possible to transport μ into ν in a *ordered* way, namely with a monotone map $T : \mathbb{R} \to \mathbb{R}$. That is in such a way that $T(x) \leq T(y)$ if $x \leq y$. How to construct such a map? Well, if the point x is mapped to the point $T(x)$ it means that all the initial mass before the point x has to be moved to all the final mass before the point $T(x)$. Namely we want that (see Figure 17)

$$\int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{T(x)} g(t)\, dt\,. \tag{4.11}$$

Is it possible to invert the above relation in order to get $T(x)$? Yes! First, we need some language.

Definition 4.30. Given a probability measure $\lambda \in \mathcal{P}(\mathbb{R})$ we define its *cumulative distribution function* $F_\mu : \mathbb{R} \to [0, 1]$ by

$$F_\lambda(t) := \lambda\left((-\infty, t]\right)\,.$$

for all $t \in \mathbb{R}$.

Remark 4.31. It is easy to see that the cumulative distribution F_λ is right-continuous. Namely it holds

$$\lim_{y \to x^+} F_\lambda(y) = F_\lambda(x)$$

for all $x \in \mathbb{R}$. Moreover, F_λ is continuous at a point $x \in \mathbb{R}$ if and only if $\lambda(\{x\}) = 0$.

By using the notion of cumulative function, (4.11) writes as

$$F_\mu(x) = F_\nu(T(x))$$

and we would like to obtain $T(x)$ as

$$T(x) = F_\nu^{-1}\left(F_\mu(x)\right)\,.$$

The question is then: is it really possible to *invert* F_ν? The answer is *almost*! There are two problems with using the usual inverse of a function (see Figure 18): the first one is that F_ν could

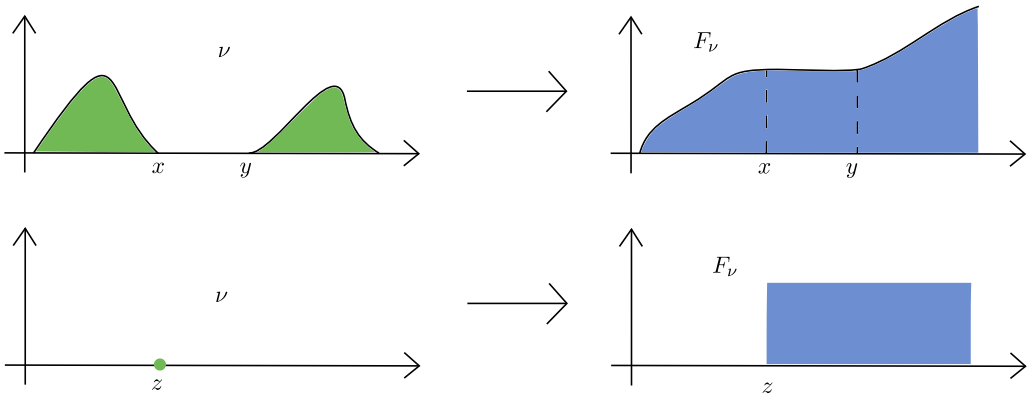

FIGURE 18. The two phenomena that can cause the cumulative distribution F_ν to be non-invertible. On the top two figures: when the support of ν is not connected this causes F_ν to be constant on the region between connected components of the support of ν. What value to choose for the counter-image? On the lower two figures: if ν is a Dirac delta at z, then F_ν will be discontinuous at z. How to define the inverse in between the two values F_ν jumps from?

be constant in some interval; the second being that F_ν could have jumps. The first situation happens if the support of ν is not connected, while the second situation happens if $\nu(x) > 0$ for some $x \in \mathbb{R}$. Nevertheless, it is possible to overcome these difficulties by using a generalized notion of inverse of a function.

Definition 4.32. Let $F : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function. We define $F^{[-1]} : \mathbb{R} \to \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$, the *pseudo-inverse* of F by

$$F^{[-1]}(t) := \inf\{s \in \mathbb{R} : F(s) \geq t\}$$

for all $t \in \mathbb{R}$.

Remark 4.33. We refer to Figure 19 for a graphic representation of the comments in here. Note that if F_μ is a constant t_0 in $[x, y]$, then $F^{[-1]}(t_0) = x$. Moreover if μ has a Dirac delta of mass $t_2 - t_1$ at the point x, then $F^{[-1]}$ will be constant between t_1 and t_2. Finally, note that $F^{[-1]}$ is non-decreasing.

Some basic properties of the pseudo-inverse are the followings.

Lemma 4.34. *Let $F : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function, and $F^{[-1]} : \mathbb{R} \to \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$ it pseudo-inverse. Then*

- $F^{[-1]}(t) = -\infty$ *if* $t < \inf_{\mathbb{R}} F$;
- $F^{[-1]}(t) = +\infty$ *if* $t > \sup_{\mathbb{R}} F$;
- $F^{[-1]}$ *is left-continuous;*
- $F^{[-1]}(t) \leq a$ *if and only if* $F(a) \geq t$;
- *If F is right-continuous, then the* inf *in the definition of $F^{[-1]}$ is a* min.

We now investigate what happens when we consider the push forward of μ and of the Lebesgue measure via F_μ and $F_\mu^{[-1]}$. First we need to introduce a particular class of measures.

Definition 4.35. A measure $\mu \in \mathcal{P}(\mathbb{R})$ is called *atomless* if for all points $x \in \mathbb{R}$ it holds $\mu(\{x\}) = 0$.

In the following $\mathcal{L}^1 \llcorner [0, 1]$ will denote the one dimensional Lebesgue measure restricted to the interval $[0, 1]$. We need to consider this restriction to avoid issues due to the fact that $F_\mu^{[-1]}(t) = -\infty$ if $t < 0$ and $F_\mu^{[-1]}(t) = +\infty$ if $t > 1$ (see Lemma 4.34).

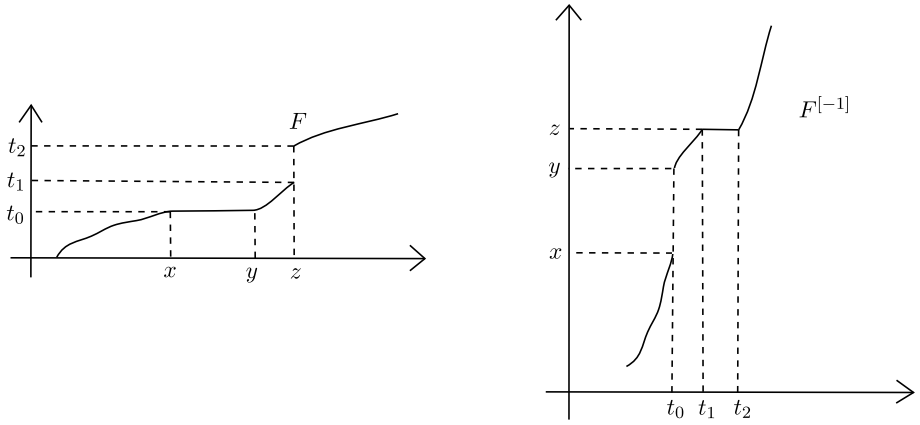

FIGURE 19. The idea behind the definition of the pseudo-inverse of a non-decreasing function. In this example we have that $F^{[-1]}(t_0) = x$, and that $F^{[-1]}(t) = z$ for all $t \in [t_1, t_2]$

Lemma 4.36. *Given a measure $\mu \in \mathcal{P}(\mathbb{R})$, let $F_\mu : \mathbb{R} \to [0,1]$ be its cumulative distribution, and let $F_\mu^{[-1]}$ be its pseudo-inverse. Then*

$$(F_\mu^{[-1]})_\# \left(\mathcal{L}^1 \llcorner [0,1]\right) = \mu.$$

Moreover, if μ is atomless, then

$$(F_\mu)_\# \mu = \mathcal{L}^1 \llcorner [0,1].$$

Proof. Let us prove the first equality. It suffices to prove that

$$\left((F_\mu^{[-1]})_\# \left(\mathcal{L}^1 \llcorner [0,1]\right)\right)((-\infty, a]) = \mu\left((-\infty, a]\right)$$

for all $a \in \mathbb{R}$. Fix $a \in \mathbb{R}$; then

$$\begin{aligned}\left((F_\mu^{[-1]})_\# \left(\mathcal{L}^1 \llcorner [0,1]\right)\right)((-\infty, a]) &= \mathcal{L}^1\left(\left\{x \in [0,1] \,:\, F_\mu^{[-1]}(x) \le a\right\}\right)\\ &= \mathcal{L}^1\left(\{x \in [0,1] \,:\, F_\mu(a) \ge x\}\right)\\ &= \mu\left((-\infty, a]\right).\end{aligned}$$

To prove the second equality, we first recall that μ atomless implies that F_μ is continuous (see Remark 4.31). In particular, for every $a \in (0,1)$

$$(F_\mu)^{-1}\left([0,a]\right) = (-\infty, x_a]$$

where $x_a \in \mathbb{R}$ is such that $F_\mu(x_a) = a$. The continuity ensures that the interval on the right-hand side is closed. Let $a \in (0,1)$; then

$$\begin{aligned}(F_\mu)_\# \mu\left((0,a]\right) &= \mu\left(\{x \in \mathbb{R} \,:\, F_\mu(x) \le a\}\right)\\ &= \mu\left((-\infty, x_a]\right)\\ &= F_\mu(x_a)\\ &= a\\ &= \mathcal{L}^1\left((0,a]\right).\end{aligned}$$

This concludes the proof. □

We now introduce the notion of monotonicity for a transport plan.

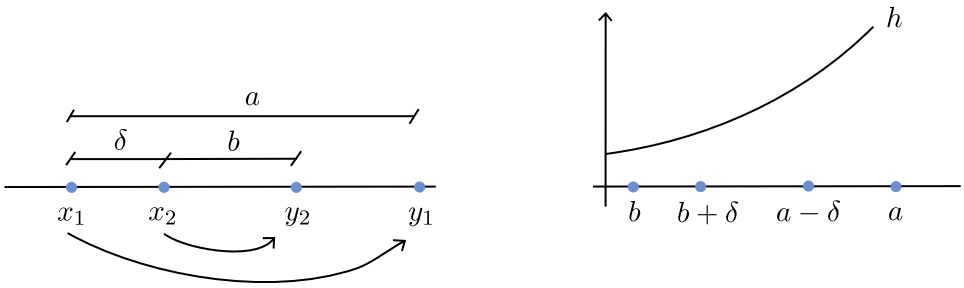

FIGURE 20. The idea behind step 1 of the proof of Theorem 4.40.

Definition 4.37. Given $\mu, \nu \in \mathcal{P}(\mathbb{R})$, we define $\gamma_{mon}(\mu,\nu)$ the *co-monotone* transport plan between them, by

$$\gamma_{mon}(\mu,\nu) := \left(F_\mu^{[-1]}, F_\nu^{[-1]}\right)_{\#} \left(\mathcal{L}^1 \llcorner [0,1]\right).$$

The following result, whose proof is left as an exercise, helps to understand how the co-monotone transport plan works.

Lemma 4.38. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Then $\gamma_{mon}(\mu,\nu) \in \Pi(\mu,\nu)$ and*

$$\gamma_{mon}(\mu,\nu)\left((-\infty,a] \times (-\infty,b]\right) = F_\mu(a) \wedge F_\nu(b)$$

for all $a, b \in \mathbb{R}$.

Remark 4.39. The importance of Lemma 4.38 lies in the fact that the knowledge of the value of a measure $\lambda \in \mathcal{M}(\mathbb{R}\times\mathbb{R})$ on sets of the form $(-\infty,a]\times(-\infty,b]$, for any $a, b \in \mathbb{R}$ is enough to know the value of λ on every rectangle $[x,y]\times[s,t]$ for every $x, y, s, t \in \mathbb{R}$ and, in turn, on every Borel subset of $X \times Y$ by using the regularity of the measure λ.

We are now in position to state the main result for convex costs in the one dimensional case. To stress the fact that we do not consider only costs depending on the *distance*, but also on the *direction*, we write $h(y-x)$ in place of the most common $h(x-y)$.

Theorem 4.40. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Consider a cost $c : \mathbb{R}\times\mathbb{R} \to [0,\infty)$ of the form*

$$c(x,y) = h(y-x)$$

where $h : \mathbb{R} \to [0,\infty)$ is a convex function. Assume that the infimum of the Kantorovich problem is finite. Then a solution to the Kantorovich problem is given by $\gamma_{mon}(\mu,\nu)$.

Moreover, if μ is atomless, then $\gamma_{mon}(\mu,\nu)$ is induced by the (only) non-decreasing transport map

$$F_\nu^{[-1]} \circ F_\mu.$$

Finally, if h is strictly convex then the solution in both of the above cases is unique.

Proof. We start by assuming h strictly convex. The convex case will be addressed by approximation.

Step 1. Let $\gamma \in \Pi(\mu,\nu)$ be a solution to the Kantorovich problem. We claim that

$$(x_2 - x_1)(y_2 - y_1) \geq 0 \tag{4.12}$$

for all $(x_1,y_1), (x_2,y_2) \in \mathrm{supp}(\gamma)$.

Assume $x_2 > x_1$ and, by contradiction, that $y_2 < y_1$. Then by Theorem 4.15 we know that its support is c-cyclically monotone. This means that for all $(x_1,y_1), (x_2,y_2) \in \mathrm{supp}(\gamma)$ it holds

$$h(y_1 - x_1) + h(y_2 - x_2) \leq h(y_1 - x_2) + h(y_2 - x_1). \tag{4.13}$$

Note that the powerfulness of the c-cyclically monotonicity is that it allows to work as we were in the discrete case (namely μ and ν were unitary Dirac deltas centred at those points). We

would like to say that it is more convenient to move mass from x_1 to y_2 and from x_2 to y_1, namely violating (4.13). Set (see Figure 20)

$$a \coloneqq y_1 - x_1, \qquad b \coloneqq y_2 - x_2, \qquad \delta \coloneqq x_2 - x_1.$$

Note that the assumption $x_2 - x_1 > 0$ implies $\delta > 0$, while $y_2 < y_1$ yields that $b < b + \delta < a$ and $b < a - \delta < a$. The right-hand side of (4.13) writes as

$$h(b + \delta) + h(a - \delta).$$

By writing

$$b + \delta = ta + (1 - t)b, \qquad a - \delta = (1 - t)a + tb$$

where $t = \frac{\delta}{a-b}$, the strict convexity of h gives

$$\begin{aligned} h(x_1 - y_2) + h(x_2 - y_1) &= h(b + \delta) + h(a - \delta) \\ &= h(ta + (1 - t)b) + h((1 - t)a + tb) \\ &< th(a) + (1 - t)h(b) + (1 - t)h(a) + th(b) \\ &= h(a) + h(b) \\ &= h(y_1 - x_1) + h(y_2 - x_2). \end{aligned}$$

This contradicts (4.13).

Step 2. We now show that $\gamma = \gamma_{mon}(\mu, \nu)$. For, let $a, b \in \mathbb{R}$. We will prove that

$$\gamma\left((-\infty, a] \times (-\infty, b]\right) = F_\mu(a) \wedge F_\nu(b).$$

Thanks to Lemma 4.38 this implies the desired result. Fix $a, b \in \mathbb{R}$. If

$$\inf\{y \in \mathbb{R} : (a, y) \in \operatorname{supp}(\gamma)\} > b \tag{4.14}$$

then, (4.12) yields that $\gamma\left([a, \infty) \times (-\infty, b]\right) = 0$. Therefore

$$\gamma\left((-\infty, a] \times (-\infty, b]\right) = \gamma\left(\mathbb{R} \times (-\infty, b]\right) = \nu((-\infty, b]) = F_\nu(b)$$

where in the second step we used that $(\pi_2)_\# \gamma = \nu$. Moreover, note that (4.14) together with (4.12) implies that $F_\nu(b) \leq F_\mu(a)$.

On the other hand, if

$$\inf\{y \in \mathbb{R} : (a, y) \in \operatorname{supp}(\gamma)\} \leq b$$

then (4.12) yields that $\gamma\left((-\infty, a) \times [b, \infty)\right) = 0$. Therefore

$$\gamma\left((-\infty, a] \times (-\infty, b]\right) = \gamma\left((-\infty, a] \times \mathbb{R}\right) = \mu((-\infty, a]) = F_\mu(b)$$

where in the second step we used that $(\pi_2)_\# \gamma = \nu$. Moreover, note that (4.14) together with (4.12) implies that $F_\mu(b) \leq F_\nu(a)$. This proves the desired result.

Step 3. In the case where μ is atomless, we would like to show that $\gamma_{mon}(\mu, \nu)$ is induced by the non-decreasing map

$$F_\nu^{[-1]} \circ F_\mu.$$

We will show that in two steps. First we show that $\gamma_{mon}(\mu, \nu)$ is induced by a non-decreasing map, and then that there exists a unique non-decreasing map pushing μ to ν.

Step 3a. For each $x \in \mathbb{R}$ let $I_x \subset \mathbb{R}$ be the smallest interval (here we do not specify whether it is open, closed, because it does not matter for our purposes) such that

$$\operatorname{supp}(\gamma) \cap (\{x\} \times \mathbb{R}) \subset I_x.$$

By using (4.12) we get that $x_1 < x_2$ implies

$$\sup\{y \in X_{x_1}\} \leq \inf\{y \in X_{x_2}\}. \tag{4.15}$$

We now want to prove that, for μ almost every point $x \in \mathbb{R}$ I_x is a singleton. We start by noticing that if I_x is not a singleton, then it contains at least one rational number in its interior. Call it q_x. By (4.15) we get that if $x_1 < x_2$ are such that I_{x_1} and I_{x_2} are not singletons, then

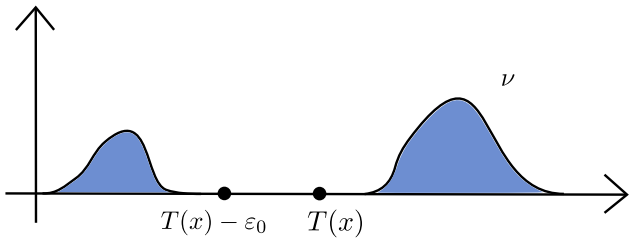

FIGURE 21. The point $T(x)$ is in an interval where F_ν is constant. This means that that interval is outside the support of ν.

$q_{x_1} < q_{x_2}$. This implies that there can be at most countably many $x \in \mathbb{R}$ such that I_x is not a singleton. Since μ is atomless, then this set of points is μ-negligible. Therefore, the map $T : \mathbb{R} \to \mathbb{R}$ given by $T(x) := I_x$ is well defined μ-almost everywhere. Note that by construction the map T is non-decreasing and $T_\#\mu = \nu$.

Step 3b. We now show that if $T : \mathbb{R} \to \mathbb{R}$ is a non-decreasing map such that $T_\#\mu = \nu$, then T coincides with $F_\nu^{[-1]} \circ F_\mu$ $\mu-$almost everywhere. By using the fact that T is non-decreasing, but not necessarily strictly increasing, we get

$$(-\infty, x] \subset T^{-1}\left((-\infty, T(x)]\right),$$

and, by using the fact that μ is a non-negative measure, in turn that

$$F_\mu(x) = \mu\left((-\infty, x]\right) \leq \mu\left(T^{-1}\left((-\infty, T(x)]\right)\right) = \nu\left((-\infty, T(x)]\right) = F_\nu(T(x)),$$

where in the previous to last step we used the fact that $T_\#\mu = \nu$. Thus

$$T(x) \geq F_\nu^{[-1]}\left(F_\mu(x)\right) \tag{4.16}$$

for all $x \in \mathbb{R}$. We now claim that the set of points $x \in \mathbb{R}$ for which $T(x) > F_\nu^{[-1]}\left(F_\mu(x)\right)$ is μ-negligible. Let $x \in \mathbb{R}$ be such that $T(x) > F_\nu^{[-1]}\left(F_\mu(x)\right)$ and take $\varepsilon_0 > 0$ such that $F_\nu\left(T(x) - \varepsilon_0\right) \geq F_\mu(x)$. Thus, for any $\varepsilon \in (0, \varepsilon_0)$ we get

$$F_\nu\left(T(x) - \varepsilon\right) \geq F_\mu(x) \tag{4.17}$$

for all $\varepsilon \in [0, \varepsilon_0)$. On the other hand, by using the fact that T is non-decreasing, we get

$$T^{-1}\left((-\infty, T(x) - \varepsilon)\right) \subset (-\infty, x).$$

Therefore, by using the fact that μ is atomless, we obtain

$$\begin{aligned} F_\mu(x) &= \mu\left((\infty, x]\right) \\ &= \mu\left((\infty, x)\right) \\ &\geq \mu\left(T^{-1}\left((-\infty, T(x) - \varepsilon)\right)\right) \\ &= \nu\left((-\infty, T(x) - \varepsilon)\right) \end{aligned}$$

for all $\varepsilon \in [0, \varepsilon_0)$. By using an approximation argument, this implies that

$$F_\mu(x) \geq F_\nu\left(T(x) - \varepsilon\right) \tag{4.18}$$

for all $\varepsilon \in [0, \varepsilon_0)$. Therefore, (4.17) and (4.18) yield

$$F_\nu\left(T(x) - \varepsilon\right) = F_\mu(x)$$

for all $\varepsilon \in [0, \varepsilon_0)$. This means that F_ν is constant in the interval $(T(x) - \varepsilon_0, T(x)]$ (see Figure 21). That is, the interval $(T(x) - \varepsilon_0, T(x)]$ is outside of the support of the measure ν. Note that there are at most countably many intervals where F_ν is constant, since each of these intervals must contain a rational number, and two different intervals do not overlap.

We now reason as follows (see Figure **??**): let $\{(a_i, b_i)\}_{i\in I}$ be the intervals where F_ν is constant. Here $I \subset \mathbb{N}$ could also be finite (or empty). For each $i \in I$ let $t_i \in [0,1]$ be such that $F_\nu(y) = t_i$ for all $y \in (a_i, b_i)$. From what we discovered above, we have that

$$\left\{ x \in \mathbb{R} \,:\, T(x) > F_\nu^{[-1]}\left(F_\mu(x)\right) \right\} \subset \bigcup_{i\in I} \{\, y \in \mathbb{R} \,:\, F_\mu(y) = t_i \,\}. \tag{4.19}$$

Now, since μ is atomless, Lemma 4.36 yields that

$$(F_\mu)_\# \mu = \mathcal{L}^1 \llcorner [0,1],$$

which implies that, for all $l \in [0,1)$

$$\mu\left(\{x \in \mathbb{R} \,:\, F_\mu(x) = l\}\right) = 0.$$

Thus, each of the sets on the union on the right-hand side of (4.19) is μ-negligible. Therefore, since the set of indexes I is at most countable, we get that

$$\mu\left(\left\{ x \in \mathbb{R} \,:\, T(x) > F_\nu^{[-1]}\left(F_\mu(x)\right) \right\}\right) = 0.$$

This, together with (4.16), implies that T coincides with $F_\nu^{[-1]} \circ F_\mu$ $\mu-$almost everywhere.

Step 4. The case where h is convex (but not strictly convex) is treated as follows. Assume h is not constant (otherwise the Kantorovich problem is trivial) and let $\gamma \in \Pi(\mu,\nu)$ be a solution to the Kantorovich problem (note that we do not know whether is it unique or not!). Let $\{h_k\}_{k\in\mathbb{N}}$ be a sequence of strictly convex functions with

$$h \leq h_k \leq \left(1 + \frac{1}{k}\right) h + \frac{1}{k}. \tag{4.20}$$

Then by steps 1 and 2 we know that $\gamma_{mon}(\mu,\nu)$ is the only solution to the Kantorovich problem with cost $c_k : \mathbb{R}\times\mathbb{R} \to [0,\infty)$ where $c_k(x,y) \coloneqq h_k(x-y)$. Note that $\gamma_{mon}(\mu,\nu)$ is independent of the cost c_k. Then, by using (4.20)

$$\begin{aligned}
\int_{\mathbb{R}\times\mathbb{R}} h(y-x)\, d\gamma_{mon}(\mu,\nu) &\leq \int_{\mathbb{R}\times\mathbb{R}} h_k(y-x)\, d\gamma_{mon}(\mu,\nu) \\
&\leq \int_{\mathbb{R}\times\mathbb{R}} h(y-x)\, d\gamma \\
&\leq \left(1+\frac{1}{k}\right) \int_{\mathbb{R}\times\mathbb{R}} h(y-x)\, d\gamma + \frac{1}{k},
\end{aligned}$$

where in the last step we used the fact that $\gamma(\mathbb{R}\times\mathbb{R}) = 1$. Thus, by passing to the limit as $k \to \infty$, we get that

$$\int_{\mathbb{R}\times\mathbb{R}} h(y-x)\, d\gamma_{mon}(\mu,\nu) = \int_{\mathbb{R}\times\mathbb{R}} h(y-x)\, d\gamma.$$

Thus $\gamma_{mon}(\mu,\nu)$ is also a solution to the Kantorovich problem (not necessarily the only one). □

As an easy consequence of the above result we have a way to compute the transport cost from μ to ν.

Corollary 4.41. *Let $\mu,\nu \in \mathcal{P}(\mathbb{R})$ and consider a cost $c : \mathbb{R}\times\mathbb{R} \to [0,\infty)$ of the form*

$$c(x,y) = h(y-x)$$

where $h : \mathbb{R} \to [0,\infty)$ is a convex function. Then

$$\inf\left\{ \int_{\mathbb{R}\times\mathbb{R}} c\, d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\} = \int_0^1 h\left(F_\nu^{[-1]}(t) - F_\mu^{[-1]}(t)\right) d\mathcal{L}^1(t).$$

Remark 4.42. What would be the analogous of Theorem 4.40 in the case where $c(x,y) = l(|x-y|)$ with l is (strictly) concave?

5. Further properties

5.1. inf **Monge** = min **Kantorovich.** We now turn our attention to the following question: when is it true that

$$\min\left\{\int_{X\times Y} c\,d\gamma \,:\, \gamma\in\Pi(\mu,\nu)\right\} = \inf\left\{\int_X c\,(x,T(x))\;d\mu \,:\, T_\#\mu=\nu\right\}? \tag{5.1}$$

Note that the two quantities are always related by $\leq$. We are already aware of the fact that this equality is not true in general. Indeed, in the case μ is a Diract delta, but ν is not, the infimum of the right-hand side is $+\infty$ since there are no transport maps. On the other hand, the left-hand side is always finite under reasonable assumptions on the cost c (lower semi-continuity) and the space X (compactness).

As we will see here, having μ concentrated on points is the only issue that can cause the above equality to fail. To prove (5.1) we need to construct a sequence $\{T_n\}_{n\in\mathbb{N}}$ of transport maps such that

$$\lim_{n\to\infty}\int_X c\,(x,T_n(x))\;d\mu = \min\left\{\int_{X\times Y} c\,d\gamma \,:\, \gamma\in\Pi(\mu,\nu)\right\}.$$

What we will prove is actually stronger than that! We will show that, if μ is atomoless, given *any* transport plan $\gamma\in\Pi(\mu,\nu)$, we can find a sequence of transport maps $\{T_n\}_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty}\int_X c\,(x,T_n(x))\;d\mu = \int_{X\times Y} c\,d\gamma.$$

In the language of Γ-convergence, this is the so called *recovery sequence*, the second ingredient needed to prove that the Kantorovich problem is the relaxation in the space of measures of the Monge problem. In other words, it is the smallest extension of the Monge problem that admits the existence of a generalized solution (a transport plan). See the discussion at the end of Chapter 3 for more on this.

The construction of such a sequence of transport maps is based on a result of independent interest that we state separately with a sketch of the proof. This result goes back to the work of Cantor on the cardinality of real numbers. In particular, he proved that the cardinality of points of the square is the same as the cardinality of points of a segment. Shocked by this fact, he wrote in 1877 to Dedekind (in a letter, you know, pre-Twitter time!): '*Je le vois mais je n'y crois pas!*' (I see it, but I don't believe it!). Well, thanks to Cantor's incredible genius, we now do believe it!

This result allows to reduce the problem to the one dimensional case. It is also useful to extend the result to more general metric space.

Lemma 5.1. *There exists a Borel map $\sigma_N:\mathbb{R}^N\to\mathbb{R}$ such that*

(i) σ_N is injective;
(ii) $\sigma_N(\mathbb{R}^N)$ is a Borel subset of $\mathbb{R}$;
(iii) $(\sigma_N)^{-1}:\sigma_N(\mathbb{R}^N)\to\mathbb{R}^N$ is Borel.

Proof. Step 1. First of all we show that if we know the result for $N=2$, then we can conclude. Indeed, it is possible to construct σ_N by induction as follows:

$$\sigma_N(x_1,\dots,x_N) \coloneqq \sigma_2\left(x_1,\sigma_{N-1}(x_2,\dots,x_N)\right).$$

Properties (i), (ii), and (iii) that are valid for σ_2 and σ_{N-1} are then inherited by σ_N.

Step 2. Thanks to step 1, we can assume $N=2$. We now show that it suffices to construct a map $\tau:(0,1)^2\to\mathbb{R}$ satisfying properties (i), (ii), and (iii). Indeed, once we have this map, we can define σ_2 by

$$\sigma_2(x,y) \coloneqq \tau\left(\frac{1}{2}+\frac{1}{\pi}\arctan x, \frac{1}{2}+\frac{1}{\pi}\arctan y\right).$$

Also in this case properties (i), (ii), and (iii) that are valid for τ are then inherited by σ_2.

Step 3. We now face the problem of constructing a map $\tau : (0,1)^2 \to (0,1)$ satisfying properties (i), (ii), and (iii). The beautiful idea is this one. given a point $(x,y) \in (0,1)$, we *merge* the coordinates x and y as follows: write in decimal notation

$$x = 0.x_1x_2x_3x_4\ldots, \qquad y = 0.y_1y_2y_3y_4\ldots,$$

and send the point (x,y) to the point

$$\tau(x,y) \coloneqq 0.x_1y_1x_2y_2x_3y_3x_4y_4\ldots$$

There is some ambiguity to take care of: the point $0.1999999\ldots$ is the point 0.2. How to uniquely define its image? For points ending with a periodic 9, we decide to consider the notation without the period. Namely $0.499999\ldots$ will be 0.45, etc. . This implies that some points in $(0,1)$ will not be in the image of τ. It is possible to prove that the set of points not in the image of τ is a Borel set.

Injectivity of the map τ is immediate. Moreover, it is also possible to prove that τ and its inverse are Borel maps. □

We need a technical result regarding finite Radon measures. The proof of this result goes along similar lines of step 2 in the proof of Lemma 2.81.

Lemma 5.2. *Let $\lambda \in \mathcal{M}(\mathbb{R}^N)$ be a finite Radon measure. Fix an open cube $Q \subset \mathbb{R}^N$. Then for almost all unit vectors $v \in \mathbb{S}^{N-1}$ and for $\mathcal{L}^1$-almost all $t \in \mathbb{R}$ it holds that*

$$\lambda\left(\partial(Q + tv)\right) = 0.$$

We are now in position to prove the main result of this section.

Theorem 5.3. *Let $X \subset \mathbb{R}^N$ and $Y \subset \mathbb{R}^M$ be two compact sets, and $\mu \in \mathcal{P}(X)$ be an atomless probability measure. Given $\gamma \in \Pi(\mu,\nu)$ there exists a sequence of transport maps $\{T_n\}_{n\in\mathbb{N}}$ with $(T_n)_\#\mu = \nu$ for each $n \in \mathbb{N}$, such that $\gamma_{T_n} \overset{w*}{\rightharpoonup} \gamma$. In particular, we get*

$$\lim_{n\to\infty} \int_X c\left(x, T_n(x)\right)\, d\mu = \int_{X\times Y} c\, d\gamma.$$

Proof. Step 1. Fix $n \in \mathbb{N}$. Let $\{\widetilde{Q}_i^n\}_{i\in\mathbb{N}}$ be a grid of open cubes in $\mathbb{R}^N$ of side length $1/n$ such that

$$\mu\left(\partial Q_i^n\right) = 0 \tag{5.2}$$

for all $i \in \mathbb{N}$. To be precise, we have $\widetilde{Q}_i^n = \frac{1}{n}(0,1)^N + \frac{1}{n}z_i^n + v_n$, where $\{z_i^n\}_{i\in\mathbb{N}} = \mathbb{Z}^N$, and $v_n \in \mathbb{R}^N$ is the translation ensuring the validity of (5.2). This is possible by using Lemma 5.2. For each $i \in \mathbb{N}$, let $Q_i^n \coloneqq \widetilde{Q}_i^n \cap X$. Note that since X is compact, $Q_i^n = \emptyset$ for all but finitely many indexes i's. For each $i \in \mathbb{N}$ define

$$\gamma_i^n \coloneqq \gamma \llcorner (Q_i^n \times Y),$$

the restriction of γ to $Q_i^n \times Y$ and set

$$\mu_i^n \coloneqq (\pi_1)_\#\gamma_i^n, \qquad \nu_i^n \coloneqq (\pi_2)_\#\gamma_i^n.$$

Note that $\mu_i^n = \mu \llcorner Q_i^n$. Since μ is atomless, so is μ_i^n for each $i \in \mathbb{N}$. Let $\sigma_N : \mathbb{R}^N \to \mathbb{R}$ and $\sigma_M : \mathbb{R}^M \to \mathbb{R}$ be the maps given by Lemma 5.1. Consider the probability measures on $\mathbb{R}$ given by

$$\widetilde{\mu}_i^n \coloneqq (\sigma_N)_\#\mu_i^n, \qquad \widetilde{\nu}_i^n \coloneqq (\sigma_M)_\#\nu_i^n.$$

It holds that $\widetilde{\mu}_i^n$ is atomless. Therefore, thanks to results in the proof of Theorem 4.40 we get that the existence of a non-decreasing map S_i^n that pushes $\widetilde{\mu}_i^n$ to $\widetilde{\nu}_i^n$. Define, for each $i \in \mathbb{N}$, the map

$$T_i^n \coloneqq (\sigma_M)^{-1} \circ S_i^n \circ \sigma_N.$$

It is easy to see that $(T_i^n)_{\#}\mu_i^n = \nu_i^n$. Then, define the map $T_n : X \to Y$ by

$$T_n(x) \coloneqq T_i^n(x)$$

if $x \in Q_i^n$.

Step 2. We now take a grid of open cubes in $\mathbb{R}^M$ of side length $1/n$ as we did in the previous step. We call $\{U_j^n\}_{j\in\mathbb{N}}$ its intersection with Y, and we assume that

$$\nu\left(\partial U_j^n\right) = 0 \tag{5.3}$$

for all $j \in \mathbb{N}$. Again, note that $U_j^n = \emptyset$ for all but finitely many indexes j's since Y is compact. We claim that

$$\gamma_{T_n}(Q_i^n \times U_i^n) = \gamma(Q_i^n \times U_i^n), \tag{5.4}$$

for all $i, j \in \mathbb{N}$. This is possible by using Lemma 5.2. Indeed

$$\begin{aligned}
\gamma_{T_n}(Q_i^n \times U_i^n) &= \mu_i^n\left(\{x \in Q_i^n \,:\, T_n(x) \in U_j^n\}\right) \\
&= \mu_i^n\left(\{x \in \mathbb{R}^N \,:\, T_n(x) \in U_j^n\}\right) \\
&= \mu_i^n\left((T_n)^{-1}\left(U_j^n\right)\right\}) \\
&= \nu_i^n(U_j^n) \\
&= (\pi_2)_{\#}\gamma_i^n(U_j^n) \\
&= \gamma_i^n(X \times U_j^n) \\
&= \gamma\left((X \times U_j^n) \cap (Q_i^n \times Y)\right) \\
&= \gamma(Q_i^n \times U_i^n).
\end{aligned}$$

This proves the claim.

Step 3. We claim that $\gamma_{T_n} \overset{w*}{\rightharpoonup} \gamma$. Let $\varphi \in C(X \times Y)$ and let $\omega : (0,\infty) \to [0,\infty)$ its modulus of continuity. For each $i, j, n \in \mathbb{N}$, let $x_{i,j}^n \in Q_i^n \times U_i^n$. We have that

$$|\varphi(z) - \varphi(x_{i,j}^n)| \le \omega(|z - x_{i,j}^n|) \le \omega(\operatorname{diam}(Q_i^n \times U_i^n)) \tag{5.5}$$

for all $z \in Q_i^n \times U_i^n$. Therefore

$$\begin{aligned}
\left|\int_{X\times Y} \varphi\, d\gamma_{T_n} - \int_{X\times Y} \varphi\, d\gamma\right| &= \left|\sum_{i,j}\left[\int_{Q_i^n\times U_i^n}\left(\varphi - \varphi(x_{i,j}^n)\right)\, d\gamma_{T_n} - \int_{Q_i^n\times U_i^n}\left(\varphi - \varphi(x_{i,j}^n)\right)\, d\gamma\right]\right| \\
&\le \sum_{i,j}\left|\int_{Q_i^n\times U_i^n}\left(\varphi - \varphi(x_{i,j}^n)\right)\, d\gamma_{T_n} - \int_{Q_i^n\times U_i^n}\left(\varphi - \varphi(x_{i,j}^n)\right)\, d\gamma\right| \\
&\le \sum_{i,j} 2\omega(\operatorname{diam}(Q_i^n \times U_i^n))\,\gamma(Q_i^n \times U_i^n) \\
&\le 2\omega\left(\frac{\sqrt{NM}}{n}\right)\sum_{i,j}\gamma(Q_i^n \times U_i^n) \\
&= 2\omega\left(\frac{\sqrt{NM}}{n}\right),
\end{aligned}$$

where we used (5.2) and (5.3) in the first step in the second step, and (5.4) to add and subtract the term $\varphi(x_{i,j}^n)$ in the second step. Moreover, in the third step we used (5.5). Note that we did not specify the exact range of the indexes i's and j's that changes with n (the grids get more and more fine), not to introduce a too heavy notation. Since $\omega\left(\frac{\sqrt{NM}}{n}\right) \to 0$ as $n \to \infty$, we conclude.

Step 4. Finally, since the cost c is continuous on $X \times Y$, by using step 3 we get

$$\lim_{n\to\infty} \int_X c\,(x, T_n(x))\; d\mu = \lim_{n\to\infty} \int_{X\times Y} c\, d\gamma_{T_n} = \int_{X\times Y} c\, d\gamma.$$

This concludes the proof. □

As a consequence of the above result (together with the fact that the Kantorovich problem admits a solution) we get the validity of equality (5.1).

Corollary 5.4. *Let $X \subset \mathbb{R}^N$ and $Y \subset \mathbb{R}^M$ be two compact sets. Let $\mu \in \mathcal{P}(X)$ be an atomoless probability measure. Then*

$$\min\left\{ \int_{X\times Y} c\, d\gamma \;:\; \gamma \in \Pi(\mu,\nu) \right\} = \inf\left\{ \int_X c\,(x, T(x))\; d\mu \;:\; T_\#\mu = \nu \right\}$$

for any probability measure $\nu \in \mathcal{P}(Y)$ and every continuous cost $c : X \times Y \to [0,\infty)$.

Remark 5.5. We know (see Theorem **??**) that the Kantorovich problem admits a solution also when the costs c is lower semi-continuous. Does the result of Corollary 5.4 hold also in this case? If not, can you construct a sequence of costs $\{c_k\}_{k\in\mathbb{N}}$ and a sequence of maps $\{T_k\}_{k\in\mathbb{N}}$ with $(T_k)_\#\mu = \nu$ such that

$$\lim_{k\to\infty} \int_X c_k\,(x, T_k(x))\; d\mu = \min\left\{ \int_{X\times Y} c\, d\gamma \;:\; \gamma \in \Pi(\mu,\nu) \right\}?$$

And what about the case where X and Y are not compact?

5.2. **Stability of the Kantorovich problem under approximations.** Suppose you want to compute the optimal cost of pushing a probability measure μ to another probability measure ν. Often the two probability distributions μ and ν are too complicated to allow for explicit computations. Therefore, one way to overcome this difficulty is to *approximate* μ and ν with sequences $\{\mu_k\}_{k\in\mathbb{N}}$ and $\{\nu_k\}_{k\in\mathbb{N}}$. Think of the case where $\mu = f\mathcal{L}^N$, with f supported on a set $X \subset \mathbb{R}^N$, and $\nu = g\mathcal{L}^N$ with g supported on a set $Y \subset \mathbb{R}^N$. A possible way to construct the approximation μ_k (and similarly the approximation ν_k) is to split X in cubes $\{Q_i^k\}_{i\in\mathbb{N}}$ of side length $1/k$ and center x_i^k and consider in each of these small cubes the average

$$M_i^k := \frac{1}{\mathcal{L}^N(Q_i^k)} \int_{Q_i^k} f\, dx.$$

Then the measures μ_k will be the defined by

$$\mu_k := \sum_i M_i^k \delta_{x_i^k}.$$

This allows to consider the problem in the fully discrete setting, for which there are several algorithms to deal with it. The question is then the following: suppose we are able to compute

$$m_k := \min\left\{ \int_{X\times Y} c\, d\gamma \;:\; \gamma \in \Pi(\mu_k,\nu_k) \right\} = \int_{X\times Y} c\, d\gamma_k.$$

Is it true that

$$\lim_{k\to\infty} m_k = \min\left\{ \int_{X\times Y} c\, d\gamma \;:\; \gamma \in \Pi(\mu,\nu) \right\} = \int_{X\times Y} c\, d\overline{\gamma}\,?$$

And that $\gamma_k \overset{w^*}{\rightharpoonup} \overline{\gamma}$?

The purpose of this section is to give an answer to the above questions. We start by proving an optimality criterion that is interesting in its own: c-cyclically monotonicity of the support of a transport plan implies its optimality for its own marginals. In all the section, X and Y will be compact metric spaces.

Theorem 5.6. *Let $\gamma \in \mathcal{P}(X \times Y)$. Assume that* $\mathrm{supp}(\gamma)$ *is c-cyclically monotone for a continuous cost $c : X \times Y \to [0, \infty)$. Then γ is a solution to the Kantorovich problem with cost c and with marginals $(\pi_1)_\#\gamma$ and $(\pi_2)_\#\gamma$ respectively.*

Proof. Let $\mu := (\pi_1)_\#\gamma$ and $\nu := (\pi_2)_\#\gamma$. Thanks to Theorem 4.16 the c-cyclically monotonicity of $\mathrm{supp}(\gamma)$ yields the existence of a c-concave function $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ such that

$$\Gamma \subset \{\, (x, y) \in X \times Y \,:\, \varphi(x) + \varphi^c(y) = c \,\}.$$

Therefore

$$\begin{aligned}
\min\left\{ \int_{X\times Y} c\, d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\} &\le \int_{X\times Y} c\, d\gamma \\
&= \int_{X\times Y} [\,\varphi(x) + \varphi^c(y)\,]\, d\gamma \\
&= \int_X \varphi\, d\mu + \int_Y \varphi^c\, d\nu \\
&\le \max\left\{ \int_X \varphi\, d\mu + \int_Y \psi\, d\nu \,:\, \varphi \oplus \psi \le c \right\} \\
&= \min\left\{ \int_{X\times Y} c\, d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\},
\end{aligned}$$

where in the last step we used the duality formula (see Theorem 4.18). This proves the optimality of γ. □

The idea to prove the stability of optimal transport plans is the following. Let $\{\mu_k\}_{k\in\mathbb{N}}$ and $\{\nu_k\}_{k\in\mathbb{N}}$ be sequences approximating μ and ν respectively, and $\{\gamma_k\}_{k\in\mathbb{N}}$ be a sequence of optimal transport planes for the Kantorovich problem with marginals μ_k and ν_k. We know that the support of each γ_k is c-cyclically monotone. By compactness, up to a subsequence,

$$\gamma_k \overset{w^*}{\rightharpoonup} \overline{\gamma}$$

for some $\overline{\gamma} \in \mathcal{P}(X \times Y)$. We already know that $\overline{\gamma} \in \Pi(\mu, \nu)$. In order to prove that $\overline{\gamma}$ is an optimal transport plan for the marginals μ and ν, we would like to be able to say that $\mathrm{supp}(\overline{\gamma})$ is c-cyclically monotone. Thus, Theorem 5.6 would allow us to conclude. To achieve our goal we first need to understand if there is a topology for which (a subsequence of) $\{\mathrm{supp}(\gamma_k)\}_{k\in\mathbb{N}}$ converges to a limiting object Γ. Granted this, we need to prove that Γ is still c-cyclically monotone, and that it contains $\mathrm{supp}(\overline{\gamma})$.

Let us start by introducing the proper notion of convergence for compact sets (see Figure 22).

Definition 5.7. Let (Z, d) be a compact metric space. Given two compact sets $A, B \subset Z$ we define the *Hausdorff distance* between them by

$$\mathrm{d_H}(A, B) := \max\{\, \max\{\mathrm{d}(x, B) \,:\, x \in A\},\, \max\{\mathrm{d}(A, y) \,:\, y \in B\}\,\}.$$

Remark 5.8. In the above definition, the requirement that A and B are closed can be dropped at the cost of substituting max with sup. The advantage of taking compact sets is to have the implication

$$\mathrm{d}_H(A, B) = 0 \quad \Rightarrow \quad A = B$$

in force. Moreover, note that, given two sets $A, B \subset Z$, if $r > 0$ is such that $B \subset A + B_r$, then it does not necessarily follows that $A \subset B + B_r$.

The following results ensures that the above definition gives rise to a distance for which the space of compact sets is pre-compact.

Theorem 5.9 (Blascke Theorem)**.** *The function* $\mathrm{d_H}$ *defines a distance on the space of compact subsets of a compact metric space* (Z, d).

Moreover, if $\{K_n\}_{n\in\mathbb{N}}$ is a sequence of compact subsets of Z, then there exists a subsequence $\{Z_{n_i}\}_{i\in\mathbb{N}}$ and a compact set $K \subset Z$ such that $\mathrm{d_H}(K_{n_i}, K) \to 0$ *as $i \to \infty$.*

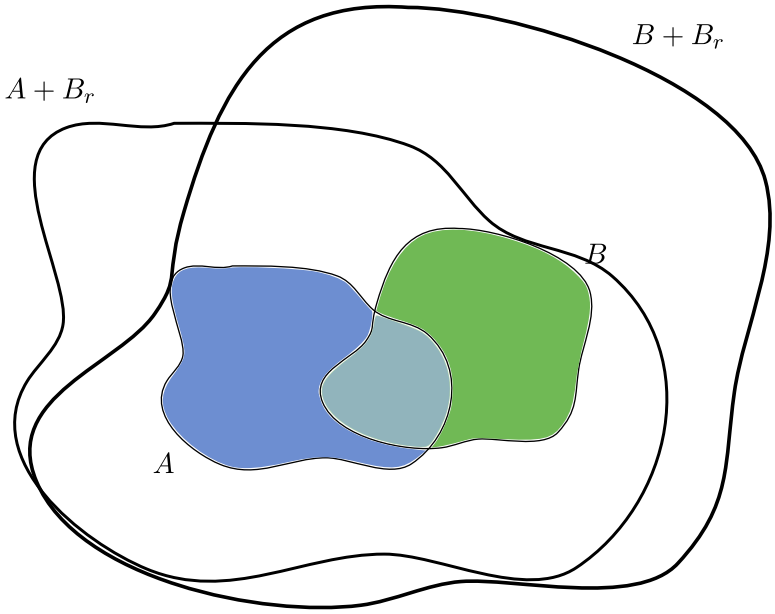

FIGURE 22. The idea of the Hausdorff convergence between the set A and the set B is the smallest number $r > 0$ such that A is contained in $B + B_r$, the r-enlargement of B and B is contained in $A + B_r$ the r-enlargement of A, where B_r is the ball of radius r.

To better understand the behaviour of a sequence of compact sets converging in the Hausdorff metric, we state the following two properties.

Lemma 5.10. *Let $\{K_n\}_{n\in\mathbb{N}}$ be a sequence of compact subsets of Z such that $\mathrm{d_H}(K_n, K) \to 0$ for some compact set $K \subset Z$. Then*

(i) For each $x \in K$ there exists a sequence $\{x_n\}_{n\in\mathbb{N}}$ with $x_n \in K_n$ for each $n \in \mathbb{N}$, such that $x_n \to x$;
(ii) If $A \subset Z$ is a compact set such that $A \cap K = \emptyset$, then there exists $n_0 \in \mathbb{N}$ such that $A \cap K_n = \emptyset$ for all $n \geq n_0$.

Next result shoes that the Hausdorff convergence of compact sets containing the support of a sequence of measures behaves nicely.

Lemma 5.11. *Let $\{\lambda_k\}_{k\in\mathbb{N}}$ be a sequence of measures on a compact metric space (Z, d), such that $\lambda_k \overset{w*}{\rightharpoonup} \lambda$. Assume that $\mathrm{supp}(\lambda_k) \subset A_k$ for some compact set $A_k \subset Z$, such that A_k converges to a compact set $A \subset Z$ in the Hausdorff topology. Then $\mathrm{supp}(\lambda) \subset A$.*

Proof. To prove that $\mathrm{supp}(\lambda) \subset A$, we consider a continuous function $\varphi \in C(Z)$ (recall that since Z is compact, $C_0(Z) = C_b(Z) = C(Z)$) with $\mathrm{supp}(\varphi) \subset Z \setminus A$. We would like to prove that

$$\int_Z \varphi \, d\lambda = 0.$$

Since $\mathrm{supp}(\varphi)$ and A are disjoint compact sets it holds

$$\mathrm{d_H}\left(\mathrm{supp}(\varphi), A\right) > 0.$$

Thus, thanks to Lemma 5.10 (ii), there exists $\bar{k} \in \mathbb{N}$ such that for all $k \geq k_0$ it holds

$$\mathrm{d_H}\left(\mathrm{supp}(\varphi), A_k\right) > 0$$

In particular, this implies that $\operatorname{supp}(\varphi) \cap A_k = \emptyset$. Thus

$$0 = \lim_{k\to\infty} \int_Z \varphi \, d\gamma_k = \int_Z \varphi \, d\gamma,$$

where the convergence of the integrals follows from the weak* convergence of μ_k to μ and the fact that φ is continuous. This concludes the proof. □

We are now in position to prove the stability result.

Theorem 5.12. *Let X and Y be compact metric spaces and let $c : X \times Y \to [0,\infty)$ be a continuous cost. Let $\{\mu_k\}_{k\in\mathbb{N}} \subset \mathcal{P}(X)$ and $\{\nu_k\}_{k\in\mathbb{N}} \subset \mathcal{P}(Y)$ be such that*

$$\mu_k \overset{w^*}{\rightharpoonup} \mu\,, \qquad \nu_k \overset{w^*}{\rightharpoonup} \nu$$

for some $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then

(i) It holds

$$\min\left\{ \int_{X\times Y} c\, d\gamma \,:\, \gamma \in \Pi(\mu_k,\nu_k) \right\} \to \min\left\{ \int_{X\times Y} c\, d\gamma \,:\, \gamma \in \Pi(\mu,\nu) \right\}$$

as $k \to \infty$;

(ii) Let $\{\gamma_k\}_{k\in\mathbb{N}} \subset \Pi(\mu_k,\nu_k)$ be a sequence of solutions to the Kantorovich problem for the cost c with marginals μ_k and ν_k such that $\gamma_k \overset{w^*}{\rightharpoonup} \overline{\gamma}$ for some $\overline{\gamma} \in \mathcal{P}(X\times Y)$. Then $\overline{\gamma}$ is a solution to the Kantorovich problem for the cost c with marginals μ and ν;

(iii) Let $\{(\varphi_k,\psi_k)\}_{k\in\mathbb{N}}$ be a sequence of pairs of c-concave Kantorovich potentials for the cost c with marginals μ_k and ν_k such that $\varphi \to \overline{\varphi}$ and $\psi_k \to \overline{\psi}$ uniformly. Then $(\overline{\varphi},\overline{\psi})$ is a pair of Kantorovich potentials for the cost c with marginals μ and ν.

Finally, if the Kantorovich problem for the cost c with marginals μ and ν admits a unique solution, then it is not needed to extract a subsequence in (ii). Similarly, if the dual problem admits a unique solution, there is no need to extract a subsequence in (iii).

Proof. We will prove (ii), then (i), and finally (iii).

Step 1: Proof of (ii). Let $\Gamma_k := \operatorname{supp}(\gamma_k)$. By definition, Γ_k is a compact subset of the compact space $X \times Y$. Therefore, Theorem 5.9 ensures that, up to a subsequence, Γ_k converges in the Hausdorff metric to a compact set $\Gamma \subset X \times Y$. By Lemma 5.11 we get that

$$\operatorname{supp}(\gamma) \subset \Gamma.$$

We now prove that Γ is c-cyclically monotone, and thus conclude thanks to Theorem 5.6.

Fix $(x_1,y_1),\dots,(x_n,y_n) \in \Gamma$ and a permutation $\sigma : \{1,\dots,n\} \to \{1,\dots,n\}$. Since Γ is the Hausdorff limit of Γ_k, thanks to Lemma 5.10 (i), for each $i = 1,\dots,n$ we get a sequence of points $\left\{(x_i^k,y_i^k)\right\}_{k\in\mathbb{N}}$ with $(x_i^k,y_i^k) \subset \Gamma_k$ such that

$$x_i^k \to x_i \qquad\qquad y_i^k \to y_i$$

as $k \to \infty$. By using the fact that each Γ_k is c-cyclically monotone (see Theorem 4.15), for all $k \in \mathbb{N}$, we get that

$$\sum_{i=1}^n c(x_i^k,y_i^k) \le \sum_{i=1}^n c(x_i^k,y_{\sigma(i)}^k).$$

By passing to the limit as $k \to \infty$ in the above inequality and using the continuity of the cost c, we get

$$\sum_{i=1}^n c(x_i,y_i) \le \sum_{i=1}^n c(x_i,y_{\sigma(i)}).$$

This proves that Γ is c-cyclically monotone and concludes the proof of this first part.

Step 2: Proof of (i). We now prove (i). Let $\{\gamma_k\}_{k\in\mathbb{N}} \subset \Pi(\mu_k, \nu_k)$ be a sequence of solutions to the Kantorovich problem for the cost c with marginals μ_k and ν_k. Up to a subsequence (here not relabelled because it doesn't matter), we have that

$$\gamma_k \overset{w^*}{\rightharpoonup} \overline{\gamma}$$

for some $\overline{\gamma} \in \mathcal{P}(X \times Y)$. Thanks to step 1 we know that γ is a solution to the Kantorovich problem for the cost c with marginals μ and ν. Thus

$$\begin{aligned}\lim_{k\to\infty} \min\left\{ \int_{X\times Y} c\, d\gamma \,:\, \gamma \in \Pi(\mu_k, \nu_k) \right\} &= \lim_{k\to\infty} \int_{X\times Y} c\, d\gamma_k \\ &= \int_{X\times Y} c\, d\overline{\gamma} \\ &= \min\left\{ \int_{X\times Y} c\, d\gamma \,:\, \gamma \in \Pi(\mu, \nu) \right\}\end{aligned}$$

where the convergence of the integrals follows from the fact that $\gamma_k \overset{w^*}{\rightharpoonup} \gamma$ and that the cost c is continuous.

Step 3: Proof of (iii). Since each pair (φ_k, ψ_k) is admissible for the dual problem, we have that

$$\varphi_k(x) + \psi_k(y) \leq c(x, y)$$

for all $(x, y) \in X \times Y$. By uniform convergence of φ_k to $\overline{\varphi}$ and of ψ_k to $\overline{\psi}$, we get

$$\overline{\varphi}(x) + \overline{\psi}(y) \leq c(x, y)$$

for all $(x, y) \in X \times Y$. Thus

$$\begin{aligned}\min\left\{ \int_{X\times Y} c\, d\gamma \,:\, \gamma \in \Pi(\mu_k, \nu_k) \right\} &= \max\left\{ \int_X \varphi\, d\mu_k + \int_Y \psi\, d\nu_k \,:\, \varphi \oplus \psi \leq c \right\} \\ &= \int_X \varphi_k\, d\mu_k + \int_Y \psi_k\, d\nu_k \end{aligned} \tag{5.6}$$

where in the first step we used the duality formula (see Theorem 4.18). Note that

$$\begin{aligned}\left| \int_X \varphi_k\, d\mu_k - \int_X \overline{\varphi}\, d\mu \right| &\leq \left| \int_X \varphi_k\, d\mu_k - \int_X \overline{\varphi}\, d\mu_k \right| + \left| \int_X \overline{\varphi}\, d\mu_k - \int_X \overline{\varphi}\, d\mu \right| \\ &\leq \|\varphi - \varphi_k\|_\infty \, |\mu_k(X)| + \left| \int_X \overline{\varphi}\, d\mu_k - \int_X \overline{\varphi}\, d\mu \right| \\ &\to 0\end{aligned}$$

as $k \to \infty$: the first term because $\mu_k(X) = 1$ for all $k \in \mathbb{N}$ and $\varphi_k \to \varphi$ uniformly, while the second term vanishes in the limit because by assumption $\mu_k \overset{w*}{\rightharpoonup} \mu$ and $\overline{\varphi} \in C(X) = C_0(X)$ (since X is compact). A similar convergence holds for the other integral. Thus

$$\int_X \varphi_k\, d\mu_k + \int_Y \psi_k\, d\nu_k \to \int_X \overline{\varphi}\, d\mu + \int_Y \overline{\psi}\, d\nu. \tag{5.7}$$

Therefore, from (5.6), (5.7), and step 2, we conclude.

Step 4: Uniqueness. Finally, uniqueness of a solution to the Kantorovich problem yields that every subsequence of $\{\gamma_k\}_{k\in\mathbb{N}}$ converge to the unique solution. Similar argument for the dual problem. □

6. Wasserstein spaces

The Kantorovich problem can be used to define a distance between measures.

Definition 6.1. Let $p \geq 1$. For $\mu, \nu \in \mathcal{P}(\Omega)$ define the *p-Wasserstain distance* between μ and ν by

$$W_p(\mu,\nu) \coloneqq \left[\min \left\{ \int_{\Omega\times\Omega} |x-y|^p \, d\gamma \, : \, \gamma \in \Pi(\mu,\nu) \right\} \right]^{\frac{1}{p}}.$$

This is also known as the *p-Monge-Kantorovich distance*, or the *p-Earth-Mover distance*.

Giving for grant for a moment that W_p is actually a distance, and thus that it deserves its name, we would like to compare it with another distance in the subclass of absolutely continuous measures: the L^p distance. Consider a functions $f, g \in L^1(\Omega)$ and the associated measures $\mu \coloneqq f\mathcal{L}^N$ and $\nu \coloneqq g\mathcal{L}^N$. The L^p distance is a *vertical* distance: for each $x \in \Omega$ we consider the quantity

$$|f(x) - g(x)|^p,$$

that corresponds to how much movement in the vertical direction we have to make to transform f into g.

On the other hand, the Monge-Kantorovich distance $W_p(\mu,\nu)$ takes also in consideration the *horizontal* movements to transform μ into ν. This can be better understood if, for $h > 0$, the function g is given by

$$g(x) \coloneqq f(x-h),$$

namely g is the function f horizontally translated by the amount h. We would like to have a distance that is of the same order as h. It is easy to see that the L^p distance does not satisfies such a property. Assume that f is compactly supported. Then for h large enough, the support of f and g are disjoint, and the L^p distance between them is

$$\|f-g\|_{L^p} = 2\|f\|_{L^p},$$

regardless of how large h is. Moreover, for h small, L^p is of the order h only for functions f that are smooth enough (having $f' \in L^p(\Omega)$ for instance). This is not the case for the function $f(t) \coloneqq t^{-\alpha}\mathbb{1}_{(0,1)}(t)$, for $\alpha \in (0,1)$, we have that

$$\|f-g\|_{L^p} \geq Ch^{1-\alpha p},$$

for $h \ll 1$, and $\alpha p < 1$.

On the other hand, by using Theorem 4.40, it is easy to see that

$$W_p(\mu,\nu) = h,$$

regardless of how large or small the parameter h is. This is a property that is crucial in certain applications, and this is why Monge-Kantorovich distances are used.

Remark 6.2. Let $\mu, \nu \in \mathcal{P}(\Omega)$. It is easy to see that, for $p \leq q$ it holds

$$W_p(\mu,\nu) \leq W_q(\mu,\nu). \tag{6.1}$$

Moreover, if Ω is bounded, we get

$$W_p(\mu,\nu) \leq [\operatorname{diam}(\Omega)]^{\frac{p-1}{p}} \, [W_1(\mu,\nu)]^{\frac{1}{p}} \tag{6.2}$$

where $\operatorname{diam}(\Omega)$ denotes the diameter of Ω.

Let us now prove that the function W_p defines a distance on $\mathcal{P}(\Omega)$. There are usually two ways to prove it (in particular to prove the triangle inequality): using mollifiers or the disintegration theorem for measures. We choose the former because it is easier to introduce from the technical point of view, other that giving the possibility to talk about convolution.

Proposition 6.3. *The function W_p is a distance on $\mathcal{P}(\Omega)$.*

Proof. We have to prove that

(i) $W_p(\mu,\nu) = W_p(\nu,\mu)$;
(ii) $W_p(\mu,\nu) \geq 0$, with equality if and only if $\mu = \nu$;
(iii) $W_p(\mu,\nu) \leq W_p(\mu,\lambda) + W_p(\lambda,\nu)$,

for all $\mu,\nu\lambda \in \mathcal{P}(\Omega)$. The first item is clear. For (ii): if $\mu = \nu$, clearly we have that $W_p(\mu,\nu) = 0$. To prove the other way round, assume that $W_p(\mu,\nu)$ and let $\gamma \in \Pi(\mu,\nu)$ be an optimal transport plan achieving $W_p(\mu,\nu)$. Then we get that $|x-y| = 0$ for all $(x,y) \in \operatorname{supp}(\gamma)$, which means that $x = y$. Thus, $\mu = (\pi_1)_\#\gamma = (\pi_2)_\#\gamma = \nu$, as wanted.

Finally, we prove the triangle inequality. Assume initially that μ and λ are non-atomic. Therefore, by Theorem 4.22 there exist $S,T : \Omega \to \Omega$ with $T_\#\lambda = \nu$ and $S_\#\mu = \lambda$ such that

$$W_p^p(\mu,\lambda) = \int_\Omega |x - S(x)|^p \, d\mu, \qquad W_p^p(\lambda,\nu) = \int_\Omega |x - T(x)|^p \, d\lambda. \tag{6.3}$$

Moreover, the map $T\circ S : \Omega \to \Omega$ is such that $(T\circ S)_\#\mu = \nu$, and thus and admissible competitor for the minimization problem defining $W_p(\mu,\nu)$. Therefore

$$\begin{aligned} W_p(\mu,\nu) &\leq \left[\int_\Omega |T\circ S(x) - x|^p \, d\mu\right]^{\frac{1}{p}} \\ &= \|T\circ S - \mathrm{Id}\|_{L^p(\mu)} \\ &\leq \|T\circ S - S\|_{L^p(\mu)} + \|S - \mathrm{Id}\|_{L^p(\mu)} \\ &= \left[\int_\Omega |T\circ S(x) - S(x)|^p \, d\mu\right]^{\frac{1}{p}} + W_p(\mu,\lambda) \\ &= \left[\int_\Omega |T(y) - y|^p \, d\lambda\right]^{\frac{1}{p}} + W_p(\mu,\lambda) \\ &= W_p(\lambda,\nu) + W_p(\mu,\lambda), \end{aligned}$$

where in the third and in the last step we used (6.3), while in the previous to last one we used the change of variable related to $S_\#\mu = \lambda$ (see Lemma 2.73).

In the general case, namely without the assumption that μ and λ are non-atomic, we use an approximation. Namely we take sequences $\{\mu_k\}_{k\in\mathbb{N}}$ and $\{\lambda_k\}_{k\in\mathbb{N}}$ of non-atomic probability measures in Ω such that

$$\lim_{k\to\infty} W_p(\mu_k,\nu_k) = W_p(\mu,\nu), \ \lim_{k\to\infty} W_p(\nu_k,\lambda_k) = W_p(\nu,\lambda), \ \lim_{k\to\infty} W_p(\mu_k,\lambda_k) = W_p(\mu,\lambda). \tag{6.4}$$

The existence of such a sequence is postponed to later (even if usually it is custom to prove the technical lemmata before, here it is more instructive to do the opposite, in order to understand why this approximation is needed). Assuming that such a sequence exists, for each $k \in \mathbb{N}$, we just proved that

$$W_p(\mu_k,\nu_k) \leq W_p(\lambda_k,\nu_k) + W_p(\mu_k,\lambda_k).$$

By passing to the limit in he above inequality as $k \to \infty$ and using (6.4) we get the desired result. □

How to get the approximations needed to conclude the second step of the proof? The technical tool needed is called *convolution*, that are going to quickly introduce.

Definition 6.4. Fix an L^1 function $\xi : \mathbb{R}^N \to [0,\infty)$ with support in $B_1(0)$ such that

$$\int_{B_1(0)} \xi(x)\, dx = 1.$$

For $\varepsilon > 0$ we define $\xi_\varepsilon : \mathbb{R}^N \to [0,\infty)$ by

$$\xi_\varepsilon(x) := \frac{1}{\varepsilon^N}\xi\left(\frac{x}{\varepsilon}\right).$$

The function ξ is called a *convolution kernel*, while the family $\{\xi_\varepsilon\}_{\varepsilon>0}$ a *family of mollifiers*.

Mollifiers are used to *smooth out* rough objects. For instance, they can be used to approximate a finite Radon measure with absolutely continuous measures, as the following results shows. The proof is left as an exercise to the reader.

Lemma 6.5. *Let $\mu \in \mathcal{P}(\mathbb{R}^N)$ and define and let $\{\xi_\varepsilon\}_{\varepsilon>0}$ be a family of mollifiers. Define, for $\varepsilon > 0$, the absolutely continuous measure $\mu * \xi_\varepsilon \in \mathcal{P}(\mathbb{R}^N)$ by*

$$\mu * \xi_\varepsilon := f_\varepsilon \mathcal{L}^N,$$

where

$$f_\varepsilon(x) := \int_{\mathbb{R}^N} \xi_\varepsilon(y-x) d\mu.$$

Then $\mu_\varepsilon \overset{w*}{\rightharpoonup} \mu$.

We are now in position to provide the existence of the families required to complete the proof of Proposition 6.3.

Lemma 6.6. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^N)$. Then there exist two sequences $\{\mu_k\}_{k\in\mathbb{N}}$ and $\{\nu_k\}_{k\in\mathbb{N}}$ of absolutely continuous measures in $\mathcal{P}(\mathbb{R}^N)$ such that*

$$\lim_{k\to\infty} W_p(\mu_k, \nu_k) = W_p(\mu, \nu).$$

Proof. Let $\{\varepsilon_k\}_{k\in\mathbb{N}} \subset (0,1)$ be such that $\varepsilon_k \to 0$ as $k \to \infty$, and let ξ a *symmetric* mollifier, namely such that $\xi(-z) = \xi(z)$ for all $z \in \mathbb{R}^N$. Set $\mu_k := \mu * \xi_{\varepsilon_k}$, and $\nu_k := \nu * \xi_{\varepsilon_k}$.

Step 1. We first prove that

$$\limsup_{k\to\infty} W_p(\mu_k, \nu_k) \le W_p(\mu, \nu).$$

Let $\gamma \in \Pi(\mu, \nu)$ be an optimal transport plan between μ and ν relative to the cost $c(x,y) := |x-y|^p$. For $k \in \mathbb{N}$, define $\gamma_k \in \mathcal{P}(\mathbb{R}^N \times \mathbb{R}^N)$ by duality as follows:

$$\int_{\mathbb{R}^N\times\mathbb{R}^N} \varphi \, d\gamma_k := \int_{\mathbb{R}^N\times\mathbb{R}^N} \left[\int_{\mathbb{R}^N} \varphi(x+z, y+z) \xi_{\varepsilon_k}(z) dz \right] d\gamma,$$

for every $\varphi \in C_0(\mathbb{R}^N \times \mathbb{R}^N)$. Then it is easy to see that

$$\int_{\mathbb{R}^N\times\mathbb{R}^N} |x-y|^p \, d\gamma_k = \int_{\mathbb{R}^N\times\mathbb{R}^N} |x-y|^p \, d\gamma. \tag{6.5}$$

Moreover, for any $\varphi \in C_0(\mathbb{R}^N)$ it holds

$$\begin{aligned}
\int_{\mathbb{R}^N\times\mathbb{R}^N} \varphi(x) \, d\gamma_k(x,y) &= \int_{\mathbb{R}^N\times\mathbb{R}^N} \left[\int_{\mathbb{R}^N} \varphi(x+z) \xi_{\varepsilon_k}(z) dz \right] d\gamma \\
&= \int_{\mathbb{R}^N\times\mathbb{R}^N} \varphi * \xi_{\varepsilon_k}(x) \, d\gamma \\
&= \int_{\mathbb{R}^N} \varphi * \xi_{\varepsilon_k} \, d\mu \\
&= \int_{\mathbb{R}^N} \varphi \, d\mu_k,
\end{aligned}$$

where in the previous to last step we used the fact that $(\pi_1)_\# \gamma = \mu$, while last step follows from the symmetry of the mollifier ξ. This implies that $(\pi_1)_\# \gamma_k = \mu_k$. Similar computations yield that $(\pi_2)_\# \gamma_k = \nu_k$. Therefore $\gamma_k \in \Pi(\mu_k, \nu_k)$ and

$$W_p^p(\mu_k, \nu_k) \le \int_{\mathbb{R}^N\times\mathbb{R}^N} |x-y|^p \, d\gamma_k = \int_{\mathbb{R}^N\times\mathbb{R}^N} |x-y|^p \, d\gamma \,,$$

where in the last step we used (6.5).

Step 2. We now prove that

$$\liminf_{k\to\infty} W_p(\mu_k,\nu_k) \geq W_p(\mu,\nu).$$

By Lemma 6.6 we have that $\mu_k \overset{w^*}{\rightharpoonup} \mu$ and $\nu_k \overset{w^*}{\rightharpoonup} \nu$ as $k\to\infty$. For each $k\in\mathbb{N}$, let $\overline{\gamma}_k \in \Pi(\mu_k,\nu_k)$ be an optimal transport plan between μ_k and ν_k relative to the cost $c(x,y) \coloneqq |x-y|^p$. Up to a subsequence (not relabeled), $\overline{\gamma}_k \overset{w^*}{\rightharpoonup} \gamma$, for some $\gamma\in\Pi(\mu,\nu)$. Thus

$$W_p^p(\mu,\nu) \leq \int_{\mathbb{R}^N\times\mathbb{R}^N} |x-y|^p d\gamma \leq \liminf_{k\to\infty} \int_{\mathbb{R}^N\times\mathbb{R}^N} |x-y|^p d\overline{\gamma}_k = \liminf_{k\to\infty} W_p^p(\mu_k,\nu_k).$$

Note that if all μ_k's and ν_k's were supported on a compact set $K\subset\mathbb{R}^N$, by using the weak* convergence of measures we would have equality in the second step above. Since we do not know that a priori, we can only get a lower semi-continuity of that quantity. In any case, this is enough to conclude the proof. □

6.1. **Induced convergence.** Now that we know that $(\mathcal{P}(\Omega),W_p)$ is a metric space, it is interesting to understand what notion of convergence is induced by the W_p distance. We first need to refine (1.4): an optimality criterion for $p=1$. We recall that $\mathrm{Lip}_1(\Omega)$ denotes the space of Lipschitz maps $u:\Omega\to\mathbb{R}$ with Lipschitz constants 1.

Lemma 6.7. *Let $c:\Omega\times\Omega\to\mathbb{R}$ be a distance. Then a function $\varphi:\Omega\to\mathbb{R}$ is c-concave if and only if $\varphi\in\mathrm{Lip}_1(\Omega)$. In particular, if $\varphi\in\mathrm{Lip}_1(\Omega)$, then $\varphi^c=-\varphi$. Finally, for all $\mu,\nu\in\mathcal{P}(\Omega)$, it holds*

$$\min\left\{\int_{\Omega\times\Omega} c(x,y)\,d\gamma \,:\, \gamma\in\Pi(\mu,\nu)\right\} = \max\left\{\int_\Omega \varphi\,d\mu - \int_\Omega \varphi\,d\nu \,:\, \varphi\in\mathrm{Lip}_1(\Omega)\right\}.$$

Proof. Let $\varphi:\Omega\to\mathbb{R}$ be c-concave. Then there exists a function $\chi:\Omega\to\mathbb{R}$ such that

$$\varphi(x) = \chi^c(x) = \inf_{y\in\Omega}\left[\,c(x,y)-\chi(y)\,\right].$$

Since the function c is a distance, we have that $x\mapsto c(x,y)-\chi(y)$ is 1-Lipschitz for all $y\in\Omega$. Therefore, φ is the infimum of a family of functions in $\mathrm{Lip}_1(\Omega)$. Thus, from Lemma 4.9 we have that also $\varphi\in\mathrm{Lip}_1(\Omega)$.

Let us now take $\varphi\in\mathrm{Lip}_1(\Omega)$. We claim that it is possible to write

$$u\varphi(x) = \inf_{y\in\Omega}\left[\,c(x,y)+\varphi(y)\,\right].$$

Indeed $\varphi(x)-\varphi(y)\leq c(x,y)$ because $\varphi\in\mathrm{Lip}_1(\Omega)$, while the other inequality follows by taking $y=x$. This proves that u is c-concave and that $\varphi^c=-\varphi$.

Finally, by the duality formula Theorem 4.18 we get that

$$\begin{aligned}\min\left\{\int_{\Omega\times\Omega} c\,d\gamma \,:\, \gamma\in\Pi(\mu,\nu)\right\} &= \max\left\{\int_\Omega \varphi\,d\mu + \int_\Omega \psi\,d\nu \,:\, \varphi\oplus\psi\leq c\right\}\\ &= \max\left\{\int_\Omega \varphi\,d\mu + \int_\Omega \varphi^c\,d\nu \,:\, \varphi\in C(\Omega),\,\varphi\ c-\text{concave}\right\}\\ &= \max\left\{\int_\Omega \varphi\,d\mu - \int_\Omega \varphi\,d\nu \,:\, \varphi\in C(\Omega),\,\varphi\ c-\text{concave}\right\}.\end{aligned}$$

This concludes the proof. □

We are now in position to prove the the main result of this section. To avoid focusing on technicalities, the proof will be given only for compact sets.

Theorem 6.8. *Let $\Omega\subset\mathbb{R}^N$, $\{\mu_k\}_{k\in\mathbb{N}}\subset\mathcal{P}(\Omega)$, $\mu\in\mathcal{P}(\Omega)$. Then $\mu_k \overset{w^*}{\rightharpoonup} \mu$ if and only if*

$$W_p(\mu_k,\mu)\to 0, \qquad \int_\Omega |x|^p\,d\mu_k \to \int_\Omega |x|^p\,d\mu$$

as $k\to\infty$.

Proof. As anticipated above, we will give the proof only in the case Ω compact.

Step 1: $p = 1$. Assume that $W_1(\mu_k, \mu) \to 0$. To prove that $\mu \overset{w*}{\rightharpoonup} \mu$ we fix $\varphi \in C_c(\Omega)$ and we prove that

$$\lim_{k\to\infty} \int_\Omega \varphi \, d\mu_k = \int_\Omega \varphi \, d\mu.$$

If $\varphi \in \mathrm{Lip}_1(\Omega) \subset C_c(\Omega)$, the conclusion follows from Lemma 6.7. If $\varphi \in \mathrm{Lip}(\Omega)$, the desired convergence follows from the linearity of the integral together with the fact that it is possible to write a generic Lipschitz function as a finite sum of Lipschitz functions with Lipschitz constant 1. Finally, in the general case, we use the density of Lipschitz functions in $C_c(\Omega)$ to conclude.

Let us now prove the reverse implication. Assume that $\mu_k \overset{w*}{\rightharpoonup} \mu$. Since Ω is compact, we have that

$$\lim_{n\to\infty} \int_\Omega |x|^p \, d\mu_n = \int_\Omega |x|^p \, d\mu.$$

In order to prove that $W_1(\mu_n, \mu) \to 0$, thanks to Lemma 6.7 we will prove that

$$\max\left\{ \int_\Omega \varphi \, d\mu_k - \int_\Omega \varphi \, d\nu_k \, : \, \varphi \in \mathrm{Lip}_1(\Omega) \right\} \to 0.$$

For each $k \in \mathbb{N}$ let $\varphi_k \in \mathrm{Lip}_1(\Omega)$ be a solution to the above minimization problem for μ_k and ν_k. Since

$$\int_\Omega (\varphi_k - t) \, d\mu_k - \int_\Omega (\varphi_k - t) \, d\mu$$

for every $t \in \mathbb{R}$, we can assume without loss of generality that there exists $x_0 \in \Omega$ such that $\varphi_k(x_0) = 0$ for all $k \in \mathbb{N}$. This implies that, for all $y \in \Omega$,

$$|\varphi_k(y)| = |\varphi_k(y) - \varphi_k(x_0)| \leq |y - x_0| \leq \mathrm{diam}(\Omega),$$

where in the previous to last step we used the fact that $\mathrm{Lip}(\varphi_k) = 1$. Thus, $\{\varphi_k\}_{k\in\mathbb{N}}$ is a sequence of uniformly bounded continuous functions with the same modulus of continuity. By the Ascoli-Arzelà Theorem they converge uniformly, up to a not relabeld subsequence, to $\varphi \in \mathrm{Lip}_1(\Omega)$. Therefore

$$\begin{aligned} W_1(\mu_k, \nu_k) &= \int_\Omega \varphi_k \, d\mu_k - \int_\Omega \varphi_k \, d\mu \\ &= \int_\Omega (\varphi_k - \varphi) \, d\mu_k + \left[\int_\Omega \varphi \, d\mu_k - \int_\Omega \varphi \, d\mu \right] - \int_\Omega (\varphi_k - \varphi) \, d\mu. \end{aligned}$$

By using the uniform continuity of φ_k to φ and the assumption that $\mu \overset{w*}{\rightharpoonup} \mu$, we get that the right-hand side vanishes as $k \to \infty$. This concludes the proof in the case $p = 1$.

Step 2: $p > 1$. Since by (6.1) and (6.2) it holds

$$W_1(\mu, \mu_k) \leq W_p(\mu, \mu_k) \leq (\mathrm{diam}(\Omega))^{\frac{p-1}{p}} W_1(\mu, \mu_k)^{\frac{1}{p}}$$

the conclusion follows directly from step 1. □

References

[1] L. Ambrosio, *Lecture notes on optimal transport problems*, Mathematical Aspects of Evolving Interfaces, (2003), pp. 1–52.

[2] L. Ambrosio and N. Gigli, *A user's guide to optimal transport*, Modelling and Optimisation of Flows on Networks, 2062 (2013), pp. 1–155.

[3] L. Ambrosio, N. Gigli, and G. Savarè, *Gradient flows: In metric spaces and in the space of probability measures*, (2008).

[4] D. P. Bourne, *A brief introduction to optimal transport theory*. Lecture notes from the 2018 LMS Undergraduate Summer School http://www.maths.gla.ac.uk/ gbellamy/LMS/BourneLectures.pdf, 2018.

[5] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Commun. Pure Appl. Math., 44 (1991), pp. 375–417.
[6] G. Buttazzo and F. Santambrogio, *A mass transportation model for the optimal planning of an urban region*, SIAM review, 51 (2009), pp. 593–610.
[7] G. Carlier, *Lecture notes optimal transportation and economic applications*. IMA, New Mathematical Models in Economics and Finance, 2010.
[8] L. Evans, *Partial differential equations and monge-kantorovich mass transfer*, (2001).
[9] L. Evans and R. Gariepy, *Measure Theory and Fine properties of functions*, CRC Press, Studies in Advanced mathematics, 1992.
[10] I. Fonseca and G. Leoni, *Modern methods in the calculus of variations: L^p spaces*, Springer Monographs in Mathematics, Springer, New York, 2007.
[11] A. Galichon, *Optimal Transport Methods in Economics*, Princeton University Press, 2016.
[12] M. R. Gangbo, W., *The geometry of optimal transportation*, Acta Math., 177 (1996), pp. 113–161.
[13] R. F. Gariepy and W. P. Ziemer, *Modern Real Analysis*, Brooks/Cole Pub Co, 1994.
[14] L. Kantorovich, *On the transfer of masses*, Dokl. Acad. Nauk. USSR, 7 (1943), pp. 7–8.
[15] R. McCann and N. Guillen, *Five lectures on optimal transportation: geometry, regularity and applications*. http://www.math.toronto.edu/mccann/papers/FiveLectures.pdf.
[16] G. Peyré and M. Cuturi, *Computational optimal transport*. https://arxiv.org/abs/1803.00567, 2018.
[17] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Birkhäuser, 2015.
[18] V. Sudakov, *Geometric properties in the theory of infinite-dimensional probability distributions*, Proc. Steklov Inst. Math., 141 (1979), pp. 1–178.
[19] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.
[20] ———, *Optimal Transport: Old and New*, Springer, 2009.