Gradient flows in measure spaces
(Topics in Analysis 2011)

Onno van Gaans

Version of 11 May 2011

These notes are the lecture notes of the course Topics in Analysis given at Leiden University in Spring 2011.

Contents

1 Topics of the course – an informal introduction 2

2 Probability measures on metric spaces 5

2.1 Borel sets 5

2.2 Borel probability measures 6

2.3 Narrow convergence of measures 10

2.4 The bounded Lipschitz metric 13

2.5 Measures as functionals 16

2.6 Prokhorov's theorem 18

2.7 Marginals and disintegration 22

2.8 Conditional probabilities 24

3 Optimal transportation problems 31

3.1 Introduction 31

3.2 Existence for the Kantorovich problem 32

3.3 Structure of optimal transportation plans 33

3.4 Kantorovich potentials 35

3.5 Existence and uniqueness for the Monge problem 44

4 Gradient flows in Hilbert spaces 48

4.1 Existence and uniqueness 49

4.2 Evolution variational inequality 52

5 Gradient flows in metric spaces 54

5.1 Absolutely continuous curves 54

5.2 Gradient flows 56

6 Wasserstein spaces 58

6.1 The Wasserstein metric 58

6.2 Geodesics in metric spaces 65

1 Topics of the course – an informal introduction

Currently, the study of gradient flows in metric spaces and in particular gradient flows in metric spaces consisting of probability measures is a very active field of research. The research extends three different areas of mathematics, each of them with a long tradition and many interesting results. Fairly recently, connections between these areas have been established and many new questions and results have emerged. This chapter briefly discusses the three areas and how they connect. The discussion is only an informal sketch, no historical or mathematical correctness is claimed!

Gradient flows

If water streams downhill on a mountain, it roughly follows the direction of steepest descent. If $H(x, y)$ is the height of the mountain above the point (x, y), then the direction of steepest descent is $-\nabla H(x, y)$, where $\nabla H = \frac{\partial H}{\partial x} + \frac{\partial H}{\partial y}$ denotes the gradient of H. If the speed of the water flow is proportional to the steepness of the hill, then the position of the water which is at time 0 above the point u_0 is $(u(t), H(u(t))$, where u satisfies

$$\begin{aligned} &u'(t) = -\nabla H(u(t)), \quad t \geq 0, \\ &u(0) = u_0. \end{aligned}$$

More generally, for a function $V\colon \mathbb{R}^d \to \mathbb{R}$ the differential equation

$$u'(t) = \nabla V(u(t))$$

is called a *gradient flow equation*, its solution is called a *gradient flow* and the function V is called a *potential*. The special structure with the gradient in the equation allows for a different type of existence theorems than the usual Lipschitz conditions. It turns out that if V is differentiable and *convex*, then for each initial value u_0 there exists a unique solution $u\colon [0, \infty) \to \mathbb{R}$ of the gradient flow differential equation for V. (What can go wrong with the flow of water if $-H$ is not convex?)

More general settings of gradient flows have been studied. For instance, there is a well developed theory of gradient flows for potential function $V\colon X \to \mathbb{R}$, where X is a Hilbert space. There are also generalisation for Banach spaces X. In a different direction, an extensive theory of gradient flows has been developed for potential functions only defined on manifolds in $\mathbb{R}^d$ or even in Hilbert spaces.

Encouraged by the generality of the theory one could ask for a generalisation to the general setting of a metric space X. However, this leads immediately to a major difficulty: what should $u'(t)$ mean if u takes values in a metric space? In Hilbert spaces and Banach spaces the linear structure allows to define $u'(t)$ as the limit of $(1/h)(u(t + h) - u(t)$ for $h \to 0$ and for curves on manifolds there is a sophisticated theory of differentiation. There is no such theory for general metric spaces.

There is a clever way to avoid differentiation in metric spaces. If V is a differentiable convex function on a Hilbert space X, then its gradient flow equation

$$u'(t) = \nabla V(u(t)), \quad t \geq 0$$

is equivalent to

$$\frac{1}{2}\frac{d}{dt}\|u(t) - z\|^2 + V(u(t)) \leq V(z) \text{ for all } z \in X.$$

In the latter inequalities we only differentiate the real valued function $\|u(t) - z\|^2$. The expression $\|u(t) - z\|$ can be written as $d(u(t), z)$, if we denote by d the metric induced by the norm. The ensuing formulation only uses the metric of the Hilbert space and can thus be formulated in any metric space! It has been very difficult to develop a satisfactory theory on existence and uniqueness of solutions of such equations and other properties like stability and regularity. The main issue was to find a suitable notion of convexity on metric spaces. Much of the work has been done in Italy, initiated by De Giorgi, leading to the first book [1] on this topic by Ambrosio, Gigli, and Savaré, who developed themselves most of the theory. There are still many questions open and research on this topic is ongoing.

Optimal transportation

Suppose a certain amount of goods is located at several distribution centers and has to be shipped to several retail outlets, each needing its own amount of the goods. Suppose moreover that the transportation costs on each route from distribution center to outlet are proportional to the amount of goods transported there. What is the optimal way to transport the goods from centers to outlets? This is typical question about optimal transportation.

In 1781 Gaspard Monge started a long tradition of research on optimal mass transportation by his paper "Mémoire sur la théorie des déblais et des remblais". One of the difficult problems, called *Monge's problem*, is the following. Suppose an initial distribution of mass is given by a mass density function $f\colon \mathbb{R}^d \to [0, \infty)$ and a desired end distribution is given by a function $g\colon \mathbb{R}^d \to [0, \infty)$. Let us scale the total mass to be 1: $\int f(x)dx = \int g(x)dx = 1$. Suppose the 'cost' of shipping one unit of mass from the initial position x to the end position y is $c(x, y)$. The function $c\colon \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is called the *cost function.* For example, c could be a multiple of the distance between x and y, or a function of the distance. The problem is to find the cheapest way of transporting the mass such that initial distribution f turns into the desired end distribution g. A 'way of transporting' here is supposed to be given by a function $r\colon \mathbb{R}^d \to \mathbb{R}^d$, which says that all mass from initial point x is moved to end point $r(x)$. As the amount of mass at x is $f(x)$, the total transportation cost is then

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

This quantity should be minimized over all possible transport functions r. The requirement on r is that it moves the initial mass distribution f to g. That means, if B is a subset of $\mathbb{R}^d$, then the total amount of mass in the end distribution contained in B, which equals $\int_B g(x)\,dx$, should equal the total amount of mass that was present on all points that are mapped to points in B. Hence r should satisfy

$$\int_B g(x)\,dx = \int_{r^{-1}(B)} f(x)\,dx.$$

(Monge considered the case where f and g are indicator functions of open sets U and V and $c(x, y) = \|x - y\|$.)

It turned out to be very difficult to determine existence of an optimal transport map r under satisfactory conditions.

Kantorovich in 1942 formulated a more abstract optimal transportation problem. Now the initial and end distributions are given by probability measures μ and ν, respectively, on $\mathbb{R}^d$. The way of transporting is now not described by a function, but by a measure η on $\mathbb{R}^d \times \mathbb{R}^d$. The interpretation is that $\eta(A \times B)$ is the amount of mass that is moved from the set A into the set B. As all mass is preserved we have $\eta(A \times \mathbb{R}^d) = \mu(A)$ and $\eta(\mathbb{R}^d \times B) = \nu(B)$ for $A, B \subset \mathbb{R}^d$. Such a measure η is called a *transport plan*. The total cost is

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\eta(x,y)$$

and this quantity should be minimized over all transport plans η.

Kantorovich's problem is not as difficult to solve as Monge's problem. Prokhorov proved a characterisation of compactness in sets of probability measures in 1956 with which existence of an optimal transport plan is easily proved in a very general setting.

It took longer to settle Monge's problem. In 1976 Sudakov dealt with the case $d = 2$ and $c(x, y) = \|x - y\|$. Gangbo and MacCann (1995) did general dimension d for $c(x, y) = \|x - y\|^p$ $(p > 1)$ and Evans and Gangbo in 1999 settled $d \geq 2$ for $c(x, y) = \|x - y\|$ under more restrictive conditions on f and g. These results have been gradually improved since then.

Closely related to Kantorovich's problem is the notion of the **Wasserstein metric** (which is claimed to be due to Kantorovich and Rubinstein rather than Wasserstein, which should more correctly be transcribed as Vasherstein). The Wasserstein distance between μ and ν is defined as

$$d_W(\mu,\nu) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\eta(x,y) \right)^{1/2},$$

where η is an optimal transport plan in Kantorovich's problem. The Wasserstein metric turns out to be extremely useful even in areas as applied as data analysis and pattern recognition.

Stochastic differential equations

Many systems in physics, economics, engineering, etc. can be approximately described by a system of linear differential equations

$$u'(t) = -Au(t), \quad t \geq 0,$$

where A is a $d \times d$ matrix. More realistically, such systems are often subject to small unpredictable perturbations from the outside world, called *noise*. An effective mathematical description of noise uses Brownian motion, which is a stochastic process that is a suitable continuous time limit of random walks. *White noise* is the derivative of Brownian motion W, so that we arrive at the stochastic differential equation

$$u'(t) = -Au(t) + W'(t).$$

(A correct mathematical formulation is more involved, as W is not differentiable, at any t, with probability one). A solution $t \mapsto u(t)$ of the stochastic differential equation is a stochastic process. That is, for each t, $u(t)$ is a random variable, which has a distribution $\mu(t)$. Thus, the solution gives a curve $t \mapsto \mu(t)$ in the set of probability measures on $\mathbb{R}^d$.

The same phenomenon can be described by means of partial differential equations. For $t > 0$, the measure $\mu(t)$ has a density $\rho(t)$ with respect to the Lebesgue measure. This density satisfies the so-called Fokker-Planck equation

$$\frac{\partial \rho}{\partial t}(t,x) = \frac{1}{2}\sum_{i=1}^{d} \frac{\partial^2 \rho}{\partial x_i^2}(t,x) + \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\Big(\rho(t,x)(Ax)_i\Big), \quad t > 0,\ x \in \mathbb{R}^d.$$

Instead of solving the stochastic differential equation and consider the distributions $\mu(t)$ of $u(t)$, one could solve the Fokker-Planck equation and consider the measures $\mu(t)$ with density $\rho(t)$. Both ways yield the same curve $t \mapsto \mu(t)$ in the set of probability measures on $\mathbb{R}^d$.

Connection

How are these three areas of mathematics connected? Felix Otto (1998) showed that the curve of measures $\mu(t)$ given by the stochastic differential equation (or Fokker-Planck equation) is a gradient flow in the metric space of probability measures endowed with the Wasserstein metric. Important contributions in this direction have been made by Yann Brenier. The recent progress in gradient flows in metric spaces and optimal transportation problems and the new connections of these fields have sparked a new interest in these topics. In 2010 Cédric Villani won the famous Fields Medal for, among others, his work on geometric aspects of gradient flows in metric spaces.

It is the aim of this course to understand the connection of gradient flows, optimal transportation and stochastic differential equations. This includes understanding the basics of each of the three ingredients. On our way we will encounter tools that are widely used in other fields of mathematics as well.

We will first consider the set of probability measures on $\mathbb{R}^d$ with a suitable topology. Instead of $\mathbb{R}^d$ we extend our scope to general metric spaces that are separable and complete and we enter the theory of probability measures on metric spaces.

2 Probability measures on metric spaces

When we study curves in spaces of probability measures we will be faced with continuity and other regularity properties and therefore with convergence of probability measures. The probability measures will be defined on the Borel σ-algebra of a metric space. Since we want to be able to apply the results to probability measures on a Hilbert space, it is not too restrictive to assume separability and completeness but we should avoid assuming compactness of the metric space.

We will consider Borel probability measures on metric spaces, narrow convergence of such measures, a metric for narrow convergence, and Prokhorov's theorem on compactness relative to the narrow convergence.

2.1 Borel sets

Let (X,d) be a metric space. The *Borel σ-algebra* (σ-field) $\mathcal{B} = \mathcal{B}(X)$ is the smallest σ-algebra in X that contains all open subsets of X. The elements of $\mathcal{B}$ are called the *Borel sets* of X.

The metric space (X,d) is called *separable* if it has a countable dense subset, that is, there are $x_1, x_2, \ldots$ in X such that $\overline{\{x_1, x_2, \ldots\}} = X$. ($\overline{A}$ denotes the closure of $A \subset X$.)

Lemma 2.1. *If X is a separable metric space, then $\mathcal{B}(X)$ equals the σ-algebra generated by the open (or closed) balls of X.*

Proof. Denote

$$\mathcal{A} := \sigma\text{-algebra generated by the open (or closed) balls of } X.$$

Clearly, $\mathcal{A} \subset \mathcal{B}$.

Let D be a countable dense set in X. Let $U \subset X$ be open. For $x \in U$ take $r > 0$, $r \in \mathbb{Q}$ such that $B(x,r) \subset U$ ($B(x,r)$ open or closed ball with center x and radius r) and take $y_x \in D \cap B(x, r/3)$. Then $x \in B(y_x, r/2) \subset B(x,r)$. Set $r_x := r/2$. Then

$$U = \bigcup \{B(y_x, r_x) : x \in U\},$$

which is a countable union. Therefore $U \in \mathcal{A}$. Hence $\mathcal{B} \subset \mathcal{A}$. □

Lemma 2.2. *Let (X,d) be a separable metric space. Let $\mathcal{C} \subset \mathcal{B}$ be countable. If $\mathcal{C}$ separates closed balls from points in the sense that for every closed ball B and every $x \in X \setminus B$ there exists $C \in \mathcal{C}$ such that $B \subset C$ and $x \notin C$, then the σ-algebra generated by $\mathcal{C}$ is the Borel σ-algebra.*

Proof. Clearly $\sigma(\mathcal{C}) \subset \mathcal{B}$, where $\sigma(\mathcal{C})$ denotes the σ-algebra generated by $\mathcal{C}$. Let B be a closed ball in X. Then $B = \bigcap\{C \in \mathcal{C} : B \subset C\}$, which is a countable intersection and hence a member of $\sigma(\mathcal{C})$. By the previous lemma we obtain $\mathcal{B} \subset \sigma(\mathcal{C})$. □

If $f : S \to T$ and $\mathcal{A}_S$ and $\mathcal{A}_T$ are σ-algebras in S and T, respectively, then f is called *measurable* (w.r.t. $\mathcal{A}_S$ and $\mathcal{A}_T$) if

$$f^{-1}(A) = \{x \in S : f(x) \in A\} \in \mathcal{A}_S \text{ for all } A \in \mathcal{A}_T.$$

Proposition 2.3. *Let (X,d) be a metric space. $\mathcal{B}(X)$ is the smallest σ-algebra with respect to which all (real valued) continuous functions on X are measurable (w.r.t. $\mathcal{B}(X)$ and $\mathcal{B}(\mathbb{R})$).* (See [14, Theorem I.1.7, p. 4].)

2.2 Borel probability measures

Let (X,d) be a metric space. A *finite Borel measure* on X is a map $\mu : \mathcal{B}(X) \to [0,\infty)$ such that

$$\begin{aligned}&\mu(\emptyset) = 0, \text{ and}\\ &A_1, A_2, \ldots \in \mathcal{B} \text{ mutually disjoint } \implies \mu(\textstyle\bigcup_{i=1}^\infty B_i) = \sum_{i=1}^\infty \mu(B_i).\end{aligned}$$

μ is called a *Borel probabiliy measure* if in addition $\mu(X) = 1$.

The following well known continuity properties will be used several times.

Lemma 2.4. *Let X be a metric space and μ a finite Borel measure on X. Let $A_1, A_2, \ldots$ be Borel sets.*

(1) If $A_1 \subset A_2 \subset \cdots$ and $A = \bigcup_{i=1}^\infty A_i$, then $\mu(A) = \lim_{n\to\infty} \mu(A_n)$.

(2) If $A_1 \supset A_2 \supset \cdots$ and $A = \bigcap_{i=1}^\infty$, then $\mu(A) = \lim_{n\to\infty} \mu(A_n)$.

The next observation is important in the proof of Theorem 2.13 (the Portmanteau theorem).

Lemma 2.5. *If μ is a finite Borel measure on X and $\mathcal{A}$ is a collection of mutually disjoint Borel sets of X, then at most countably many elements of $\mathcal{A}$ have nonzero μ-measure.*

Proof. For $m \geq 1$, let $\mathcal{A}_m := \{A \in \mathcal{A} : \mu(A) > 1/m\}$. For any distinct $A_1, \dots, A_k \in \mathcal{A}_m$ we have

$$\mu(X) \geq \mu(\bigcup_{i=1}^{k} A_i) = \mu(A_1) + \cdots + \mu(A_k) > k/m,$$

hence $\mathcal{A}_m$ has at most $m\mu(X)$ elements. Thus

$$\{A \in \mathcal{A} : \mu(A) > 0\} = \bigcup_{m=1}^{\infty} \mathcal{A}_m$$

is countable. □

Example. If μ is a finite Borel measure on $\mathbb{R}$, then $\mu(\{t\}) = 0$ for all except at most countably many $t \in \mathbb{R}$.

Proposition 2.6. *Any finite Borel measure on X is* regular*, that is, for every $B \in \mathcal{B}$*

$$\begin{aligned} \mu(B) &= \sup\{\mu(C) : C \subset B,\ C \textit{ closed}\} \quad \textit{(inner regular)} \\ &= \inf\{\mu(U) : U \supset B,\ U \textit{ open}\} \quad \textit{(outer regular).} \end{aligned}$$

Proof. Define the collection $\mathcal{R}$ by

$$A \in \mathcal{R} \iff \begin{array}{l} \mu(A) = \sup\{\mu(C) : C \subset A,\ C\text{closed}\} \text{ and} \\ \mu(A) = \inf\{\mu(U) : U \supset A,\ U \text{ open}\}. \end{array}$$

We have to show that $\mathcal{R}$ contains the Borel sets. *step 1: $\mathcal{R}$ is a σ-algebra:*
$\emptyset \in \mathcal{R}$. Let $A \in \mathcal{R}$, let $\varepsilon > 0$. Take C closed and U open with $C \subset A \subset U$ and $\mu(A) < \mu(C) + \varepsilon$, $\mu(A) > \mu(U) - \varepsilon$. Then $U^c \subset A^c \subset C^c$, U^c is closed, C^c is open, and

$$\begin{aligned} \mu(A^c) = \mu(X) - \mu(A) > \mu(X) - \mu(C) - \varepsilon = \mu(C^c) - \varepsilon, \\ \mu(A^c) = \mu(X) - \mu(A) < \mu(X) - \mu(U) + \varepsilon = \mu(U^c) + \varepsilon. \end{aligned}$$

Hence $A^c \in \mathcal{R}$.

Let $A_1, A_2, \ldots \in \mathcal{R}$ and let $\varepsilon > 0$. Take for each i

U_i open ,C_i closed with
$C_i \subset A_i \subset U_i$,
$\mu(U_i) - \mu(A_i) < 2^{-i}\varepsilon,\ \mu(A_i) - \mu(C_i) < 2^{-i}\varepsilon/2.$

Then $\bigcup_i C_i \subset \bigcup_i A_i \subset \bigcup_i U_i$ and $\bigcup_i U_i$ is open, and

$$\begin{aligned} \mu(\bigcup_i U_i) - \mu(\bigcup_i A_i) &\leq \mu\Big(\bigcup_{i=1}^{\infty} U_i \setminus \bigcup_{i=1}^{\infty} A_i\Big) \\ &\leq \mu\Big(\bigcup_{i=1}^{\infty} (U_i \setminus A_i)\Big) \leq \sum_{i=1}^{\infty} \mu(U_i \setminus A_i) \\ &= \sum_{i=1}^{\infty} (\mu(U_i) - \mu(A_i)) < \sum_{i=1}^{\infty} 2^{-i}\varepsilon = \varepsilon. \end{aligned}$$

Further, $\mu(\bigcup_{i=1}^\infty C_i) = \lim_{k\to\infty}\mu(\bigcup_{i=1}^k C_i)$, hence for some large k, $\mu(\bigcup_{i=1}^\infty C_i) - \mu(\bigcup_{i=1}^k C_i) < \varepsilon/2$. Then $C := \bigcup_{i=1}^k C_i \subset \bigcup_{i=1}^\infty A_i$, C is closed, and

$$\begin{aligned} \mu(\bigcup_{i=1}^\infty A_i) - \mu(C) &< \mu(\bigcup_{i=1}^\infty A_i) - \mu(\bigcup_{i=1}^\infty C_i) + \varepsilon/2 \\ &\le \mu\Big(\bigcup_{i=1}^\infty A_i \setminus \bigcup_{i=1}^\infty C_i\Big) + \varepsilon/2 \\ &\le \mu\Big(\bigcup_{i=1}^\infty (A_i \setminus C_i)\Big) + \varepsilon/2 \\ &\le \sum_{i=1}^\infty \mu(A_i \setminus C_i) + \varepsilon/2 \\ &= \sum_{i=1}^\infty \Big(\mu(A_i) - \mu(C_i)\Big) + \varepsilon/2 \le \varepsilon/2 + \varepsilon/2. \end{aligned}$$

Hence $\bigcup_{i=1}^\infty A_i \in \mathcal{R}$. Thus $\mathcal{R}$ is a σ-algebra.

step2: $\mathcal{R}$ contains all open sets: We prove: $\mathcal{R}$ contains all closed sets. Let $A \subset X$ be closed. Let $U_n := \{x \in X : d(x, A) < 1/n\} = \{x \in X : \exists a \in A \text{ with } d(a, x) < 1/n\}$, $n = 1, 2, \ldots$. Then U_n is open, $U_1 \supset U_2 \supset \cdots$, and $\bigcap_{i=1}^\infty U_i = A$, as A is closed. Hence $\mu(A) = \lim_{n\to\infty}\mu(U_n) = \inf_n \mu(U_n)$. So

$$\mu(A) \le \inf\{\mu(U) : U \supset A,\ U \text{ open}\} \le \inf_n \mu(U_n) = \mu(A).$$

Hence $A \in \mathcal{R}$.

Conclusion: $\mathcal{R}$ is a σ-algebra that contains all open sets, so $\mathcal{R} \supset \mathcal{B}$. □

Corollary 2.7. *If μ and ν are finite Borel measures on the metric space X and $\mu(A) = \nu(A)$ for all closed A (or all open A), then $\mu = \nu$.*

A finite Borel measure μ on X is called *tight* if for every $\varepsilon > 0$ there exists a compact set $K \subset X$ such that $\mu(X \setminus K) < \varepsilon$, or, equivalently, $\mu(K) \ge \mu(X) - \varepsilon$. A tight finite Borel measure is also called a *Radon measure.*

Corollary 2.8. *If μ is a tight finite Borel measure on the metric space X, then*

$$\mu(A) = \sup\{\mu(K) : K \subset A,\ K \textit{ compact}\}$$

for every Borel set A in X.

Proof. Take for every $\varepsilon > 0$ a compact set K_ε such that $\mu(X \setminus K_\varepsilon) < \varepsilon$. Then

$$\mu(A \cap K_\varepsilon) = \mu(A \setminus K_\varepsilon^c) \ge \mu(A) - \mu(K_\varepsilon^c) > \mu(A) - \varepsilon$$

and

$$\begin{aligned} \mu(A \cap K_\varepsilon) &= \sup\{\mu(C) : C \subset K_\varepsilon \cap A,\ C \text{ closed}\} \\ &\le \sup\{\mu(K) : K \subset A,\ K \text{ compact}\}, \end{aligned}$$

because each closed subset contained in a compact set is compact. Combination completes the proof. □

Of course, if (X, d) is a compact metric space, then every finite Borel measure on X is tight. There is another interesting case. A complete separable metric space is sometimes called a *Polish space.*

Theorem 2.9. *If (X, d) is a complete separable metric space, then every finite Borel measure on X is tight.*

We need a lemma from topology.

Lemma 2.10. *If (X, d) is a complete metric space, then a closed set K in X is compact if and only if it is totally bounded, that is, for every $\varepsilon > 0$ the set K is covered by finitely many balls (open or closed) of radius less than or equal to ε.*

Proof. $\Rightarrow$) Clear: the covering with all ε-balls with centers in K has a finite subcovering.

$\Leftarrow$) Let $(x_n)_n$ be a sequence in K. For each $m \geq 1$ there are finitely many $1/m$-balls that cover K, at least one of which contains x_n for infinitely many n. For $m = 1$ take a ball B_1 with radius ≤ 1 such that $N_1 := \{n : x_n \in B_1\}$ is infinite, and take $n_1 \in N_1$. Take a ball B_2 with radius $\leq 1/2$ such that $N_2 := \{n > n_1 : x_n \in B_2 \cap B_1\}$ is infinite, and take $n_2 \in N_2$. Take B_3, radius $\leq 1/3$, $N_3 := \{n > n_2 : x_n \in B_3 \cap B_2 \cap B_1\}$ infinite, $n_3 \in N_3$. And so on.

Thus $(x_{n_k})_k$ is a subsequence of $(x_n)_n$ and since $x_{n_\ell} \in B_k$ for all $\ell \geq k$, $(x_{n_k})_k$ is a Cauchy sequence. As X is complete, $(x_n)_n$ converges in X and as K is closed, the limit is in K. So $(x_n)_n$ has a convergent subsequence and K is compact. □

Proof of Theorem 2.9. We have to prove that for every $\varepsilon > 0$ there exists a compact set K such that $\mu(X \setminus K) < \varepsilon$. Let $D = \{a_1, a_2, \ldots\}$ be a countable dense subset of X. Then for each $\delta > 0$, $\bigcup_{k=1}^{\infty} B(a_k, \delta) = X$. Hence $\mu(X) = \lim_{n\to\infty} \mu(\bigcup_{k=1}^{n} B(a_k, \delta))$ for all $\delta > 0$. Let $\varepsilon > 0$. Then there is for each $m \geq 1$ an n_m such that

$$\mu\Big(\bigcup_{k=1}^{n_m} B(a_k, 1/m)\Big) > \mu(X) - 2^{-m}\varepsilon.$$

Let

$$K := \bigcap_{m=1}^{\infty} \bigcup_{k=1}^{n_m} \overline{B}(a_k, 1/m).$$

Then K is closed and for each $\delta > 0$,

$$K \subset \bigcup_{k=1}^{n_m} \overline{B}(a_k, 1/m) \subset \bigcup_{k=1}^{n_m} B(a_k, \delta)$$

if we choose $m > 1/\delta$. So K is compact, by the previous lemma. Further,

$$\begin{aligned} \mu(X \setminus K) &= \mu\Big(\bigcup_{m=1}^{\infty} (X \setminus \bigcup_{k=1}^{n_m} \overline{B}(a_k, 1/m))\Big) \leq \sum_{m=1}^{\infty} \mu\Big(X \setminus \bigcup_{k=1}^{n_m} \overline{B}(a_k, 1/m)\Big) \\ &= \sum_{m=1}^{\infty} \Big(\mu(X) - \mu(\bigcup_{k=1}^{n_m} \overline{B}(a_k, 1/m))\Big) < \sum_{m=1}^{\infty} 2^{-m}\varepsilon = \varepsilon. \end{aligned}$$

□

2.3 Narrow convergence of measures

Let (X, d) be a metric space and denote

$$C_b(X) := \{f : X \to \mathbb{R} : f \text{ is continuous and bounded}\}.$$

Each $f \in C_b(X)$ is integrable with respect to any finite Borel measure on X.

Definition 2.11. Let $\mu, \mu_1, \mu_2, \dots$ be finite Borel measures on X. We say that the sequence $(\mu_i)_i$ *converges narrowly* to μ if

$$\int f \,\mathrm{d}\mu_i \to \int f \,\mathrm{d}\mu \text{ as } i \to \infty \text{ for all } f \in C_b(X).$$

We will simply use the notation $\mu_i \to \mu$. (There is at most one such a limit μ, as follows from the metrization by the bounded Lipschitz metric, which is discussed in the next section.)

Narrow convergence can be described by means of other classes of functions than the bounded continuous ones. Recall that a function f from a metric space (X, d) into $\mathbb{R}$ is called *lower semicontinuous* (l.s.c.) if for every $x, x_1, x_2, \dots$ with $x_i \to x$ one has

$$f(x) \leq \liminf_{i\to\infty} f(x_i)$$

and *upper semicontinuous* (u.s.c.) if

$$f(x) \geq \limsup_{i\to\infty} f(x_i).$$

The limits here may be ∞ or $-\infty$ and then the usual order on $[-\infty, \infty]$ is considered. The indicator function of an open set is l.s.c. and the indicator function of a closed set is u.s.c.

Proposition 2.12. *Let (X, d) be a metric space and let $\mu, \mu_1, \mu_2, \dots$ be Borel probability measures on X. The following four statements are equivalent:*

- *(a) $\mu_i \to \mu$, that is, $\int f \,\mathrm{d}\mu_i \to \int f \,\mathrm{d}\mu$ for every $f \in C_b(X)$*
- *(b) $\int f \,\mathrm{d}\mu_i \to \int f \,\mathrm{d}\mu$ for every bounded Lipschitz function $f : X \to \mathbb{R}$*
- *(c) $\liminf_{i\to\infty} \int f \,\mathrm{d}\mu_i \geq \int f \,\mathrm{d}\mu$ for every l.s.c. function $f : X \to \mathbb{R}$ that is bounded from below*
- *(c') $\limsup_{i\to\infty} \int f \,\mathrm{d}\mu_i \leq \int f \,\mathrm{d}\mu$ for every u.s.c. function $f : X \to \mathbb{R}$ that is bounded from above.*

Proof. (a)⇒(b) is clear.

(b)⇒(c): First assume that f is bounded. Define for $n \in \mathbb{N}$ the *Moreau-Yosida* approximation

$$f_n(x) := \inf_{y\in X} \Big(f(y) + nd(x, y)\Big), \quad x \in X.$$

Then clearly $\inf f \leq f_0 \leq f_1 \leq f_2 \leq \cdots \leq f$, so that, in particular, f_n is bounded for each n. Further, f_n is Lipschitz. Indeed, let $u, v \in X$ and observe that for $y \in X$ we have

$$\begin{aligned} f_n(u) - \Big(f(y) + nd(v, y)\Big) &\leq \Big(f(y) + nd(u, y)\Big) - \Big(f(y) + nd(v, y)\Big) \\ &\leq nd(u, v). \end{aligned}$$

If we take supremum over y we obtain $f_n(u) - f_n(v) \leq nd(u,v)$. By changing the role of u and v we infer

$$|f_n(u) - f_n(v)| \leq nd(u,v),$$

so f_n is Lipschitz.

Next we show that $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in X$. For $x \in X$ and $n \geq 1$ there is a $y_n \in X$ such that

$$f_n(x) \geq f(y_n) + nd(x,y_n) - 1/n \geq \inf f + nd(x,y_n) - 1, \tag{1}$$

so

$$nd(x,y_n) \leq f_n(x) - \inf f + 1 \leq f(x) - \inf f + 1 \quad \text{for all } n,$$

hence $y_n \to x$ as $n \to \infty$. Then (1) yields

$$\liminf_{n\to\infty} f_n(x) \geq \liminf_{n\to\infty} f(y_n) \geq f(x),$$

as f is l.s.c. Since $f_n(x) \leq f(x)$ for all n, we obtain that $f_n(x)$ converges to $f(x)$.

Due to the monotone convergence, $\int f_n \,\mathrm{d}\mu \uparrow \int f \,\mathrm{d}\mu$. As $f \geq f_n$,

$$\liminf_{i\to\infty} \int f \,\mathrm{d}\mu_i \geq \liminf_{i\to\infty} \int f_n \,\mathrm{d}\mu_i = \int f_n \,\mathrm{d}\mu$$

for all n, by (b). Hence $\liminf_{i\to\infty} \int f \,\mathrm{d}\mu_i \geq \int f \,\mathrm{d}\mu$.

If f is not bounded from above, let $m \in \mathbb{N}$ and truncate f at m: $f \wedge m = x \mapsto \min\{f(x), m\}$. The above conclusion applied to $f \wedge m$ yields,

$$\int f \wedge m \,\mathrm{d}\mu \leq \liminf_{i\to\infty} \int f \wedge m \,\mathrm{d}\mu_i \leq \liminf_{i\to\infty} \int f \,\mathrm{d}\mu_i$$

and $\int f \,\mathrm{d}\mu = \lim_{m\to\infty} \int f \wedge m \,\mathrm{d}\mu \leq \liminf_{i\to\infty} \int f \,\mathrm{d}\mu_i$.

(c)⇔(c'): multiply by -1.

(c)⇒(a): if f is continuous and bounded, we have (c) both for f and $-f$. □

Narrow convergence can also be described as convergence on sets.

Theorem 2.13 (Portmanteau theorem)**.** *Let (X,d) be a metric space and let $\mu, \mu_1, \mu_2, \ldots$ be Borel probability measures on X. The following four statements are equivalent:*

(a) $\mu_i \to \mu$ (narrow convergence)

(b) $\liminf_{i\to\infty} \mu_i(U) \geq \mu(U)$ for all open $U \subset X$

(b') $\limsup_{i\to\infty} \mu_i(C) \leq \mu(C)$ for all closed $C \subset X$

(c) $\mu_i(A) \to \mu(A)$ for every Borel set A in X with $\mu(\partial A) = 0$. (Here$\partial A = \overline{A} \setminus A^\circ$.)

Proof. (a)⇒(b): If U is open, then the indicator function 1_U of U is l.s.c. So by the previous proposition,

$$\liminf_{i\to\infty} \mu_i(U) = \liminf_{i\to\infty} \int 1_U \,\mathrm{d}\mu_i \geq \int 1_U \,\mathrm{d}\mu = \mu(U).$$

(b)⇒(b'): By complements,

$$\begin{aligned}\limsup_{i\to\infty}\mu_i(C) &= \limsup_{i\to\infty}\Big(\mu_i(X)-\mu_i(C^c)\Big) = 1-\liminf_{i\to\infty}\mu_i(C^c)\\ &\geq 1-\mu(C^c) = \mu(X)-\mu(C^c)=\mu(C).\end{aligned}$$

(b')⇒(b): Similarly.

(b)+(b')⇒(c): $A^\circ \subset A \subset \overline{A}$, A° is open and $\overline{A}$ is closed, so by (b) and (b'),

$$\begin{aligned}\limsup\mu_i(A) &\leq \limsup\mu_i(\overline{A}) \leq \mu(\overline{A}) = \mu(A\cup\partial A)\\ &\leq \mu(A)+\mu(\partial A) = \mu(A),\\ \liminf\mu_i(A) &\geq \liminf\mu_i(A^\circ) \geq \mu(A^\circ) = \mu(A\setminus\partial A)\\ &\geq \mu(A)-\mu(\partial A)=\mu(A),\end{aligned}$$

hence $\mu_i(A)\to\mu(A)$.

(c)⇒(a): Let $g\in C_b(X)$. Idea: we have $\int f\,\mathrm{d}\mu_i \to \int f\,\mathrm{d}\mu$ for suitable simple functions; we want to approximate g to get $\int g\,\mathrm{d}\mu_i\to\int g\,\mathrm{d}\mu$.

Define

$$\nu(E) := \mu(\{x : g(x)\in E\}) = \mu(g^{-1}(E)), \quad E \text{ Borel set in } \mathbb{R}.$$

Then ν is a finite Borel measure (probability measure) on $\mathbb{R}$ and if we take $a < -\|g\|_\infty$, $b>\|g\|_\infty$, then $\nu(\mathbb{R}\setminus(a,b))=0$. As ν is finite, there are at most countably many α with $\nu(\{\alpha\})>0$ (see Lemma 2.5). Hence for $\varepsilon>0$ there are $t_0,\dots,t_m\in\mathbb{R}$ such that

(i) $a=t_0<t_1<\cdots<t_m=b$,
(ii) $t_j-t_{j-1}<\varepsilon,\ j=1,\dots,m$,
(iii) $\nu(\{t_j\})=0, \text{i.e., } \mu(\{x : g(x)=t_j\})=0,\ j=0,\dots,m$.

Take

$$A_j := \{x\in X : t_{j-1}\leq g(x)<t_j\} = g^{-1}([t_{j-1},t_j)),\ j=1,\dots,m.$$

Then $A_j\in\mathcal{B}(X)$ for all j and $X=\bigcup_{j=1}^m A_j$. Further,

$$\begin{aligned}\overline{A}_j &\subset \{x : t_{j-1}\leq g(x)\leq t_j\} \text{ (since this set is closed and } \supset A_j),\\ A_j^\circ &\supset \{x : t_{j-1}< g(x)< t_j\} \text{ (since this set is open and } \subset A_j),\end{aligned}$$

so

$$\begin{aligned}\mu(\partial A_j) &= \mu(\overline{A}_j\setminus A_j^\circ) \leq \mu(\{x : g(x)=t_{j-1} \text{ or } g(x)=t_j\})\\ &= \mu(\{x : g(x)=t_{j-1}\})+\mu(\{x : g(x)=t_j\}) = 0+0.\end{aligned}$$

Hence by (e), $\mu_i(A_j)\to\mu(A_j)$ as $i\to\infty$ for $j=1,\dots,m$. Put

$$h := \sum_{j=1}^m t_{j-1}1_{A_j},$$

then $h(x)\leq g(x)\leq h(x)+\varepsilon$ for all $x\in X$. Hence

$$\begin{aligned}\Big|\int g\,\mathrm{d}\mu_i - \int g\,\mathrm{d}\mu\Big| &= \Big|\int(g-h)\,\mathrm{d}\mu_i + \int h\,\mathrm{d}\mu_i - \int (g-h)\,\mathrm{d}\mu - \int h\,\mathrm{d}\mu\Big|\\ &\leq \int|g-h|\,\mathrm{d}\mu_i + \Big|\int h\,\mathrm{d}\mu_i - \int h\,\mathrm{d}\mu\Big| + \int|g-h|\,\mathrm{d}\mu\\ &\leq \varepsilon\mu_i(X) + \Big|\sum_{j=1}^m t_{j-1}\Big(\mu_i(A_j)-\mu(A_j)\Big)\Big| + \varepsilon\mu(X).\end{aligned}$$

It follows that $\limsup_{i\to\infty} |\int g \,\mathrm{d}\mu_i - \int g \,\mathrm{d}\mu| \le 2\varepsilon$. Thus $\int g \,\mathrm{d}\mu_i \to \int g \,\mathrm{d}\mu$ as $i \to \infty$. □

2.4 The bounded Lipschitz metric

Let (X, d) be a metric space. Denote

$$\mathcal{P} = \mathcal{P}(X) := \text{all Borel probability measures on } X.$$

We have defined the notion of narrow convergence in $\mathcal{P}$. We will show next that narrow convergence is induced by a metric, provided that X is separable. This results goes back to Prokhorov [15]. Instead of Prokhorov's metric, we will consider the "bounded Lipschitz metric" due to Dudley [6], as it is easier to work with. (See also [7, 18].) Denote

$$\mathrm{BL}(X, d) := \{f : X \to \mathbb{R} \colon\ f \text{ is bounded and Lipschitz}\}.$$

Define for $f \in \mathrm{BL}(X, d)$

$$\|f\|_{\mathrm{BL}} = \|f\|_\infty + \mathrm{Lip}(f),$$

where

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

and

$$\mathrm{Lip}(f) := \sup_{x,y\in X,\ x\neq y} \frac{|f(x) - f(y)|}{d(x, y)} = \inf\{L : |f(x) - f(y)| \le Ld(x, y)\ \forall x, y \in X\}.$$

Then $\|\cdot\|_{\mathrm{BL}}$ is a norm on $\mathrm{BL}(X, d)$. Define for $\mu, \nu \in \mathcal{P}(X)$

$$d_{\mathrm{BL}}(\mu, \nu) := \sup\{|\int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu| \colon\ f \in \mathrm{BL}(X, d),\ \|f\|_{\mathrm{BL}} \le 1\}.$$

The function d_{BL} is called the *bounded Lipschitz metric on* $\mathcal{P}$ (induced by d), which makes sense because of the next theorem.

Theorem 2.14 (Dudley, 1966)**.** *Let (X, d) be a metric space.*

(1) d_{BL} is a metric on $\mathcal{P} = \mathcal{P}(X)$.

(2) If X is separable and $\mu, \mu_1, \mu_2, \ldots \in \mathcal{P}$, then

$$\mu_i \to \mu \text{ (narrowly)} \quad \Longleftrightarrow \quad \mu_{\mathrm{BL}}(\mu_i, \mu) \to 0.$$

Proof. (See [7, Theorem 11.3.3, p. 395].)

(1): To show the triangle inequality, let $\mu, \nu, \eta \in \mathcal{P}(X)$ and observe that

$$|\int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\eta| \le |\int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu| + |\int f \,\mathrm{d}\nu - \int f \,\mathrm{d}\eta| \quad \forall f \in \mathrm{BL}(X, d),$$

so $d_{\mathrm{BL}}(\mu, \eta) \le d_{\mathrm{BL}}(\mu, \nu) + d_{\mathrm{BL}}(\nu, \eta)$. Clearly, $d_{\mathrm{BL}}(\mu, \nu) = d_{\mathrm{BL}}(\nu, \mu)$ and $d_{\mathrm{BL}}(\mu, \mu) = 0$. If $d_{\mathrm{BL}}(\mu, \nu) = 0$, then $\int f \,\mathrm{d}\mu = \int f \,\mathrm{d}\nu$ for all $f \in \mathrm{BL}(X, d)$. Therefore the constant sequence $\mu, \mu, \ldots$ converges narrowly to ν and $\nu, \nu, \ldots$ converges to μ. The Portmanteau theorem then yields $\nu(U) \leq \mu(U)$ and $\mu(U) \leq \nu(U)$ hence $\mu(U) = \nu(U)$ for any open $U \subseteq X$. By outer regularity of both μ and ν it follows that $\mu = \nu$. Thus d_{BL} is a metric on $\mathcal{P}$.

(2): If $d_{\mathrm{BL}}(\mu_i, \mu) \to 0$, then $\int f \, \mathrm{d}\mu_i \to \int f \, \mathrm{d}\mu$ for all $f \in \mathrm{BL}(X, d)$ with $\|f\|_{\mathrm{BL}} \leq 1$ and hence for all $f \in \mathrm{BL}(X, d)$. With the aid of Proposition 2.12 we infer that μ_i converges narrowly to μ.

Conversely, assume that μ_i converges narrowly to μ, that is, $\int f \, \mathrm{d}\mu_i \to \int f \, \mathrm{d}\mu$ for all $f \in C_b(X)$. Denote

$$B := \{f \in \mathrm{BL}(X, d) : \|f\|_{\mathrm{BL}} \leq 1\}.$$

In order to show that $d_{\mathrm{BL}}(\mu_i, \mu) \to 0$ we have to show that $\int f \, \mathrm{d}\mu_i$ converges uniformly in $f \in B$. If X were compact, we could use the Arzela-Ascoli theorem and reduce to a finite set of functions f. As X may not be compact, we will first call upon Theorem 2.9.

Let $\hat{X}$ be the completion of the metric space (X, d). Every $f \in B$ extends uniquely to an $\hat{f} : \hat{X} \to \mathbb{R}$ with $\|\hat{f}\|_{\mathrm{BL}} = \|f\|_{\mathrm{BL}}$. Also μ extends to $\hat{X}$:

$$\hat{\mu}(A) := \mu(A \cap X), \quad A \subseteq \hat{X} \text{ Borel.}$$

Let $\varepsilon > 0$. By the lemma, there exists a compact set $K \subseteq \hat{X}$ such that $\hat{\mu}(K) \geq 1 - \varepsilon$. The set $G := \{\hat{f}|_K : f \in B\}$ is equicontinuous and uniformly bounded, so by the Arzela-Ascoli theorem (see [7, Theorem 2.4.7, p. 52]) it is relatively compact in $(C(K), \|\cdot\|_\infty)$. Hence there are $f_1, \ldots, f_m \in B$ such that

$$\forall f \in B \ \exists \ell \text{ such that } \|\hat{f}|_K - \hat{f}_\ell|_K\|_\infty < \varepsilon \tag{2}$$

(the ε-balls around the f_i cover B). Take N such that

$$\left|\int_X f_\ell \, \mathrm{d}\mu_i - \int_X f_\ell \, \mathrm{d}\mu\right| < \varepsilon$$

for $k = 1, \ldots, N$ and $i \geq N$. Let $f \in B$ and choose a corresponding ℓ as in (2). Denote

$$K_\varepsilon = \{x \in X : \mathrm{dist}(x, K) < \varepsilon\},$$

which is an open set in X. (Here $\mathrm{dist}(x, K) := \inf\{d(x, y) \colon y \in K\}$.) For $x \in K_\varepsilon$, take $y \in K$ with $d(x, y) < \varepsilon$, then

$$\begin{aligned} |f(x) - f_\ell(x)| &\leq |f(x) - \hat{f}(y)| + |\hat{f}(y) - \hat{f}_\ell(y)| + |\hat{f}_\ell(y) - f_\ell(x)| \\ &< \mathrm{Lip}(\hat{f}) d(x, y) + \varepsilon + \mathrm{Lip}(\hat{f}_\ell) d(y, x) \\ &< 3\varepsilon. \end{aligned}$$

Further, $X \setminus K_\varepsilon$ is closed, so

$$\limsup_{i \to \infty} \mu_i(X \setminus K_\varepsilon) \leq \mu(X \setminus K_\varepsilon) \leq \mu(X \setminus K) = \hat{\mu}(\hat{X} \setminus K) \leq \varepsilon,$$

so there is an M with $\mu_i(X \setminus K_\varepsilon) \leq \varepsilon$ for all $i \geq M$. Hence for $i \geq N \vee M$,

$$\begin{aligned} \left|\int_X f \, \mathrm{d}\mu_i - \int_X f \, \mathrm{d}\mu\right| &\leq \left|\int_X f_\ell \, \mathrm{d}\mu_i - \int_X f_\ell \, \mathrm{d}\mu\right| + \int_{K_\varepsilon} |f_\ell - f| \, \mathrm{d}(\mu_i + \mu) \\ &\quad + \int_{X \setminus K_\varepsilon} |f_\ell - f| \, \mathrm{d}(\mu_i + \mu) \\ &< \varepsilon + 6\varepsilon + \int_{X \setminus K_\varepsilon} 2 \, \mathrm{d}\mu_i + \int_{X \setminus K_\varepsilon} 2 \, \mathrm{d}\mu \\ &\leq 11\varepsilon, \end{aligned}$$

hence $d_{\mathrm{BL}}(\mu_i, \mu) \leq 11\varepsilon$ for $i \geq N \vee M$. Thus, $d_{\mathrm{BL}}(\mu_i, \mu) \to 0$ as $i \to \infty$. □

Proposition 2.15. *Let (X, d) be a separable metric space. Then $\mathcal{P} = \mathcal{P}(X)$ with the bounded Lipschitz metric d_{BL} is separable.*

Proof. Let $D := \{a_1, a_2, \ldots\}$ be a countable set in X. Let

$$\mathcal{M} := \{\alpha_1\delta_{a_1} + \cdots + \alpha_k\delta_{a_k} : \alpha_1, \ldots, \alpha_k \in \mathbb{Q} \cap [0,1],\ \sum_{j=1}^{k} \alpha_j = 1,\ k = 1, 2, \ldots\}.$$

(Here δ_a denotes the Dirac measure at $a \in X$: $\delta_a(A) = 1$ if $a \in A$, 0 otherwise.) Clearly, $\mathcal{M} \subset \mathcal{P}$ and $\mathcal{M}$ is countable.

Claim: $\mathcal{M}$ is dense in $\mathcal{P}$. Indeed, let $\mu \in \mathcal{P}$. For each $m \geq 1$, $\bigcup_{j=1}^{\infty} B(a_j, 1/m) = X$. Take k_m such that

$$\mu(\bigcup_{j=1}^{k_m} B(a_j, 1/m)) \geq 1 - 1/m.$$

Modify the balls $B(a_j, 1/m)$ into disjoint sets by taking $A_1^m := B(a_1, 1/m)$, $A_j^m := B(a_j, 1/m) \setminus \left[\bigcup_{i=1}^{j-1} B(a_i, 1/m)\right]$, $j = 2, \ldots, k_m$. Then $A_1^m, \ldots, A_{k_m}^m$ are disjoint and $\bigcup_{i=1}^{j} A_i^m = \bigcup_{i=1}^{j} B(a_i, 1/m)$ for all j. In particular, $\mu(\bigcup_{j=1}^{k_m} A_j^m) \geq 1 - 1/m$, so

$$\sum_{j=1}^{k_m} \mu(A_j^m) \in [1 - 1/m, 1].$$

We approximate

$$\mu(A_1^m)\delta_{a_1} + \cdots + \mu(A_{k_m}^m)\delta_{a_{k_m}}$$

by

$$\mu_m := \alpha_1^m\delta_{a_1} + \cdots + \alpha_{k_m}^m\delta_{a_{k_m}},$$

where we choose $\alpha_j^m \in [0,1] \cap \mathbb{Q}$ such that $\sum_{j=1}^{k_m} \alpha_j^m = 1$ and

$$\sum_{j=1}^{k_m} |\mu(A_j^m) - \alpha_j^m| < 2/m.$$

(First take $\beta_j \in [0,1] \cap \mathbb{Q}$ with $\sum_{j=1}^{k_m} |\beta_j - \mu(A_j^m)| < 1/2m$, then $\sum_j \beta_j \in [1 - 3/2m, 1 + 1/2m]$. Take $\alpha_j := \beta_j / \sum_i \beta_i \in [0,1] \cap \mathbb{Q}$, then $\sum_j \alpha_j = 1$ and $\sum_{j=1}^{k_m} |\beta_j - \alpha_j| = |1 - 1/\sum_i \beta_i| \sum_{j=1}^{k_m} \beta_j = |\sum_i \beta_j - 1| \leq 3/2m$, so $\sum_{j=1}^{k_m} |\alpha_j - \mu(A_j^m)| < 1/2m + 3/2m = 2/m$.)

Then for each m, $\mu_m \in \mathcal{M}$. To show: $\mu_m \to \mu$ in $\mathcal{P}$, that is, $\mu_n \to \mu$ narrowly. Let

$g \in \mathrm{BL}(X,d)$. Then

$$\begin{aligned}
\Big|\int g\,\mathrm{d}\mu_m - \int g\,\mathrm{d}\mu\Big| &= \Big|\sum_{j=1}^{k_m} \alpha_j^m g(a_j) - \int g\,\mathrm{d}\mu\Big| \\
&\le \Big|\sum_{j=1}^{k_m} \mu(A_j^m) g(a_j) - \int g\,\mathrm{d}\mu\Big| + (2/m)\sup_j |g(a_j)| \\
&\le \Big|\int \sum_{j=1}^{k_m} g(a_j) 1_{A_j^m}\,\mathrm{d}\mu - \int g\,\mathrm{d}\mu\Big| + (2/m)\|g\|_\infty \\
&\le \Big|\sum_{j=1}^{k_m} \int \Big(g(a_j)1_{A_j^m} - g1_{A_j^m}\Big)\,\mathrm{d}\mu - \int g1_{(\bigcup_{j=1}^{k_m})^c}\,\mathrm{d}\mu\Big| + (2/m)\|g\|_\infty \\
&\le \sum_{j=1}^{k_m} \sup_{x\in A_j^m} |g(a_j) - g(x)|\mu(A_j^m) + \|g\|_\infty \mu\Big(\big(\bigcup_{j=1}^{k_m} A_j^m\big)^c\Big) + (2/m)\|g\|_\infty \\
&\le \sum_{j=1}^{k_m} \mathrm{Lip}(g)(1/m)\mu(A_j^m) + (3/m)\|g\|_\infty \\
&\le (3/m)\|g\|_{\mathrm{BL}}.
\end{aligned}$$

Hence $\int g\,\mathrm{d}\mu_m \to \int g\,\mathrm{d}\mu$ as $m \to \infty$. Thus, $\mu_m \to \mu$. □

Conclusion. If (X,d) is a separable metric space, then so is $\mathcal{P}(X)$ with the induced bounded Lipschitz metric. Moreover, a sequence in $\mathcal{P}(X)$ converges in metric if and only if it converges narrowly and then in both senses to the same limit.

2.5 Measures as functionals

Let (X,d) be a metric space. The space of real valued bounded continuous functions $C_b(X)$ endowed with the supremum norm $\|\cdot\|$ is a Banach space. It is sometimes convenient to apply functional analytic results about the Banach space $(C_b(X), \|\cdot\|_\infty)$ to the set of Borel probability measures on X. We will for instance need the Riesz representation theorem in the proof of Prokhorov's theorem. Let us consider the relation between measures and functionals.

Recall that a linear map $\varphi : C_b(X) \to \mathbb{R}$ is called a *bounded functional* if

$$|\varphi(f)| \le M\|f\|_\infty \quad \text{for all } f \in C_b(X)$$

for some constant M. The space of all bounded linear functionals on $C_b(X)$ is denoted by

$$C_b(X)' := \{\varphi : C_b(X) \to \mathbb{R} \colon \varphi \text{ is linear and bounded}\}$$

and called the (Banach) *dual space* of $C_b(X)$. A norm on $C_b(X)'$ is defined by

$$\|\varphi\| = \sup\{|\varphi(f)| : f \in C_b(X),\ \|f\|_\infty \le 1\}, \quad \varphi \in C_b(X)'.$$

A functional $\varphi \in C_b(X)'$ is called *positive* if $\varphi(f) \ge 0$ for all $f \in C_b(X)$ with $f \ge 0$.

For each finite Borel measure μ on a metric space (X, d), the map φ_μ defined by

$$\varphi_\mu(f) := \int f \,\mathrm{d}\mu, \quad f \in C_b(X),$$

is linear from $C_b(X)$ to $\mathbb{R}$ and

$$|\varphi_\mu(f)| \leq \int |f| \,\mathrm{d}\mu \leq \|f\|_\infty \mu(X).$$

Hence $\varphi_\mu \in C_b(X)'$. Further, $\|\varphi_\mu\| \leq \mu(X)$ and since $\varphi_\mu(1) = \mu(X) = \|1\|_\infty \mu(X)$ we have

$$\|\varphi_\mu\| = \mu(X).$$

Moreover, φ_μ is positive.

Conversely, if X is compact, then $C_b(X) = C(X) = \{f : X \to \mathbb{R} : f \text{ is continuous}\}$ and every positive bounded linear functional on $C(X)$ is represented by a finite Borel measure on X. The truth of this statement does not depend on X being a metric space. Therefore we state it in its usual general form, although we have not formally defined Borel sets, Borel measures, $C_b(X)$, etc. for topological spaces that are not metrizable. We denote by 1 the function on X that is identically 1.

Theorem 2.16 (Riesz representation theorem)**.** *If (X, d) is a compact Hausdorff space and $\varphi \in C(X)'$ is positive (that is, $\varphi(f) \geq 0$ for every $f \in C(X)$ with $f \geq 0$) and $\varphi(1) = 1$, then there exists a unique Borel probability measure μ on X such that*

$$\varphi(f) = \int f \,\mathrm{d}\mu \quad \textit{for all } f \in C(X).$$

(See [16, Theorem 2.14, p. 40].)

Let us next observe that narrow convergence in $\mathcal{P}(X)$ corresponds to weak* convergence in $C_b(X)'$. The weak* topology on $C_b(X)'$ is the coarsest topology such that the function $\varphi \to \varphi(f)$ on $C_b(X)'$ is continuous for every $f \in C_b(X)'$. A sequence $\varphi_1, \varphi_2, \ldots$ in $C_b(X)'$ *converges weak** to φ in $C_b(X)'$ if and only if

$$\varphi_i(f) \to \varphi(f) \quad \text{as } i \to \infty \text{ for all } f \in C_b(X).$$

If $\mu, \mu_1, \mu_2, \ldots$ are Borel probability measures on X, it is immediately clear that

$$\mu_i \to \mu \text{ narrowly in } \mathcal{P}(X) \quad \Longleftrightarrow \quad \varphi_{\mu_i} \to \varphi_\mu \text{ weak* in } C_b(X)',$$

where, as before, $\varphi_{\mu_i}(f) = \int f \,\mathrm{d}\mu_i$ and $\varphi_\mu(f) = \int f \,\mathrm{d}\mu$, $f \in C_b(X)$, $i \geq 1$.

For the next two theorems see [10, Exercise V.7.17, p. 437] and [17, Theorem 8.13].

Theorem 2.17. *If (X, d) is a metric space, then*

$$C_b(X) \textit{ is separable} \quad \Longleftrightarrow \quad X \textit{ is compact.}$$

Theorem 2.18. *If E is a separable Banach space, then $\{\varphi \in E' \colon \|\varphi\| \leq 1\}$ is weak* sequentially compact.*

Consequently, if (X,d) is a compact metric space, then the closed unit ball of $C_b(X)'$ is weak* sequentially compact. In combination with the Riesz representation theorem we obtain the following statements for sets of Borel probability measures.

Proposition 2.19. *Let (X,d) be a metric space. If (X,d) is compact, then $(\mathcal{P}(X), d_{\mathrm{BL}})$ is compact, where d_{BL} is the bounded Lipschitz metric induced by d. (Note that any compact metric space is separable.)*

Proof. Assume that (X,d) is compact. Then $C_b(X) = C(X) := \{f : X \to \mathbb{R}\colon f \text{ is continuous}\}$. The unit ball $B' := \{\varphi \in C_b(X)'\colon \|\varphi\| \leq 1\}$ of $C_b(X)'$ is weak* sequentially compact. As $(\mathcal{P}(X), d_{\mathrm{BL}})$ is a metric space, sequentially compactness is equivalent to compactness. Let $(\mu_n)_n$ be a sequence in $\mathcal{P}(X)$ and let

$$\varphi_n(f) := \int f \, \mathrm{d}\mu_n, \quad n \in \mathbb{N}.$$

Then $\varphi_n \in B'$ for all n. As B' is weak* sequentially compact, hence there exists a $\varphi \in B'$ and a subsequence $(\varphi_{n_k})_k$ such that $\varphi_{n_k} \to \varphi$ in the weak* topology. Then for each $f \in C_b(X)$ with $f \geq 0$,

$$\varphi(f) = \lim_{k\to\infty} \varphi_{n_k}(f) \geq 0,$$

so φ is positive. Further, $\varphi(1) = \lim_{k\to\infty} \varphi_{n_k}(1) = 1$. Due to the Riesz representation theorem there exists a $\mu \in \mathcal{P}(X)$ such that $\varphi(f) = \int f \, \mathrm{d}\mu$ for all $f \in C(X) = C_b(X)$. Since $\varphi_{n_k} \to \varphi$ weak*, it follows that $\mu_{n_k} \to \mu$ narrowly. Thus $\mathcal{P}(X)$ is sequentially compact. □

2.6 Prokhorov's theorem

Let (X,d) be a metric space and let $\mathcal{P}(X)$ be the set of Borel probability measures on X. Endow $\mathcal{P}(X)$ with the bounded Lipschitz metric induced by d.

In the study of measure valued functions or the limit behavior of stochastic processes one often needs to know when a sequence of random variables is convergent in distribution or, at least, has a subsequence that converges in distribution. This comes down to finding a good description of the sequences in $\mathcal{P}(X)$ that have a convergent subsequence or rather of the relatively compact sets of $\mathcal{P}(X)$. Recall that a subset S of a metric space is called *relatively compact* if its closure $\overline{S}$ is compact. The following theorem by Yu.V. Prokhorov [15] gives a useful description of the relatively compact sets of $\mathcal{P}(X)$ in case X is separable and complete. Let us first attach a name to the equivalent condition.

Definition 2.20. A set Γ of Borel probability measures on X is called *tight* if for every $\varepsilon > 0$ there exists a compact subset K of X such that

$$\mu(K) \geq 1 - \varepsilon \quad \text{for all } \mu \in \Gamma.$$

(Also other names and phrases are in use instead of 'Γ is tight': 'Γ is uniformly tight', 'Γ satisfies Prokhorov's condition', 'Γ is uniformly Radon', and maybe more).

Remark. We have shown already: if (X,d) is a complete separable metric space, then $\{\mu\}$ is tight for each $\mu \in \mathcal{P}(X)$ (see Theorem 2.9).

Theorem 2.21 (Prokhorov, 1956)**.** *Let (X,d) be a complete separable metric space and let Γ be a subset of $\mathcal{P}(X)$. Then the following two statements are equivalent:*

(a) $\overline{\Gamma}$ *is compact in* $\mathcal{P}(X)$.

(b) Γ *is tight.*

Let us first remark here that completeness of X is not needed for the implication (b)⇒(a). The proof of the theorem is quite involved. We start with the more straightforward implication (a)⇒(b).

Proof of (a)⇒(b). Claim: If $U_1, U_2, \ldots$ are open sets in X that cover X and if $\varepsilon > 0$, then there exists a $k \geq 1$ such that

$$\mu(\bigcup_{i=1}^{k} U_i) > 1 - \varepsilon \quad \text{for all } \mu \in \Gamma.$$

To prove the claim by contradiction, suppose that for every $k \geq 1$ there is a $\mu_k \in \Gamma$ with $\mu_k(\bigcup_{i=1}^{k} U_i) \leq 1 - \varepsilon$. As $\overline{\Gamma}$ is compact, there is a $\mu \in \overline{\Gamma}$ and a subsequence with $\mu_{k_j} \to \mu$. For any $n \geq 1$, $\bigcup_{i=1}^{n} U_i$ is open, so

$$\begin{aligned} \mu(\bigcup_{i=1}^{n} U_i) &\leq \liminf_{j\to\infty} \mu_{k_j}(\bigcup_{i=1}^{n} U_i) \\ &\leq \liminf_{j\to\infty} \mu_{k_j}(\bigcup_{i=1}^{k_j} U_i) \leq 1 - \varepsilon. \end{aligned}$$

But $\bigcup_{i=1}^{\infty} U_i = X$, so $\mu(\bigcup_{i=1}^{n} U_i) \to \mu(X) = 1$ as $n \to \infty$, which is a contradiction. Thus the claim is proved.

Now let $\varepsilon > 0$ be given. Take $D = \{a_1, a_2, \ldots\}$ dense in X. For every $m \geq 1$ the open balls $B(a_i, 1/m)$, $i = 1, 2, \ldots$, cover X, so by the claim there is a k_m such that

$$\mu\Big(\bigcup_{i=1}^{k_m} B(a_i, 1/m)\Big) > 1 - \varepsilon 2^{-m} \quad \text{for all } \mu \in \Gamma.$$

Take

$$K := \bigcap_{m=1}^{\infty} \bigcup_{i=1}^{k_m} \overline{B}(a_i, 1/m).$$

Then K is closed and for each $\delta > 0$ we can take $m > 1/\delta$ and obtain $K \subset \bigcup_{i=1}^{k_m} B(a_i, \delta)$, so that K is totally bounded. Hence K is compact, since X is complete. Moreover, for each $\mu \in \Gamma$

$$\begin{aligned} \mu(X \setminus K) &= \mu\Big(\bigcup_{m=1}^{\infty} \Big[\bigcup_{i=1}^{k_m} \overline{B}(a_i, 1/m)\Big]^c\Big) \\ &\leq \sum_{m=1}^{\infty} \mu\Big(\Big[\bigcup_{i=1}^{k_m} \overline{B}(a_i, 1/m)\Big]^c\Big) \\ &= \sum_{m=1}^{\infty} \Big(1 - \mu\Big(\bigcup_{i=1}^{k_m} \overline{B}(a_i, 1/m)\Big)\Big) \\ &< \sum_{m=1}^{\infty} \varepsilon 2^{-m} = \varepsilon. \end{aligned}$$

Hence Γ is tight. □

The proof that condition (b) implies (a) is more difficult. We will follow the proof from [14], which is based on compactifications. We have shown already that if X is compact, then $\mathcal{P}(X)$ is compact (see Proposition 2.19). In that case (a) trivially holds. In the cases that we want to consider, X will not always be compact. We can reduce to the compact case by considering a compactification of X.

Lemma 2.22. *If (X, d) is a separable metric space, then there exist a compact metric space (Y, δ) and a map $T : X \to Y$ such that T is a homeomorphism from X onto $T(X)$.*

(T is in general not an isometry. If it were, then X complete $\Rightarrow T(X)$ complete $\Rightarrow T(X) \subset Y$ closed $\Rightarrow T(X)$ compact, which is not true for, e.g., $X = \mathbb{R}$.)

Proof. Let $Y := [0,1]^{\mathbb{N}} = \{(\xi_i)_{i=1}^{\infty} : \xi_i \in [0,1]\ \forall i\}$ and

$$\delta(\xi, \eta) := \sum_{i=1}^{\infty} 2^{-i}|\xi_i - \eta_i|, \qquad \xi, \eta \in Y.$$

Then δ is a metric on Y, its topology is the topology of coordinatewise convergence, and (Y, δ) is compact.

Let $D = \{a_1, a_2, \ldots\}$ be dense in X and define

$$\alpha_i(x) := \min\{d(x, a_i), 1\}, \qquad x \in X,\ i = 1, 2, \ldots.$$

Then for each k, $\alpha_k : X \to [0,1]$ is continuous. For $x \in X$ define

$$T(x) := (\alpha_i(x))_{i=1}^{\infty} \in Y.$$

Claim: for any $C \subset X$ closed and $x \notin C$ there exist $\varepsilon > 0$ and i such that

$$\alpha_i(x) \le \varepsilon/3, \quad \alpha_i(y) \ge 2\varepsilon/3 \quad \text{for all } y \in C.$$

To prove the claim, take $\varepsilon := \min\{d(x, C), 1\} \in (0, 1]$. Take i such that $d(a_i, x) < \varepsilon/3$. Then $\alpha_i(x) \le \varepsilon/3$ and for $y \in C$ we have

$$\begin{aligned} \alpha_i(y) &= \min\{d(y, a_i), 1\} \ge \min\{(d(y, x) - d(x, a_i)), 1\} \\ &\ge \min\{(d(x, C) - \varepsilon/3), 1\} \\ &\ge \min\{2\varepsilon/3, 1\} = 2\varepsilon/3. \end{aligned}$$

In particular, if $x \ne y$ then there exists an i such that $\alpha_i(x) \ne \alpha_i(y)$, so T is injective. Hence $T : X \to T(X)$ is a bijection. It remains to show that for $(x_n)_n$ and x in X:

$$x_n \to x \quad \Longleftrightarrow \quad T(x_n) \to T(x).$$

If $x_n \to x$, then $\alpha_i(x_n) \to \alpha_i(x)$ for all i, so $\delta(T(x_n), T(x)) \to 0$ as $n \to \infty$.

Conversely, suppose that $x_n \not\to x$. Then there is a subsequence such that $x \notin \overline{\{x_{n_1}, x_{n_2}, \ldots\}}$. Then by the claim there is an i such that $\alpha_i(x) \le \varepsilon/3$ and $\alpha_i(x_{n_k}) \ge 2\varepsilon/3$ for all k, so that $\alpha_i(x_{n_k}) \not\to \alpha_i(x)$ as $k \to \infty$ and hence $T(x_{n_k}) \not\to T(x)$. □

We can now complete the proof of Prokhorov's theorem.

Proof of (b)⇒(a). We will show more: *If (X,d) is a separable metric space and $\Gamma \subset \mathcal{P}(X)$ is tight, then $\overline{\Gamma}$ is compact.* Let $\Gamma \subset \mathcal{P}(X)$ be tight. First observe that $\overline{\Gamma}$ is tight as well. Indeed, let $\varepsilon > 0$ and let K be a compact subset of X such that $\mu(K) \geq 1-\varepsilon$ for all $\mu \in \Gamma$. Then for every $\mu \in \overline{\Gamma}$ there is a sequence $(\mu_n)_n$ in Γ that converges to μ and then we have $\mu(K) \geq \limsup_{n\to\infty} \mu_n(K) \geq 1-\varepsilon$.

Let $(\mu_n)_n$ be a sequence in $\overline{\Gamma}$. We have to show that it has a convergent subsequence. Let (Y,δ) be a compact metric space and $T : X \to Y$ be such that T is a homeomorphism from X onto $T(X)$. For $B \in \mathcal{B}(Y)$, $T^{-1}(B)$ is Borel in X. Define

$$\nu_n(B) := \mu_n(T^{-1}(B)), \quad B \in \mathcal{B}(Y),\ n = 1,2,\ldots.$$

Then $\nu \in \mathcal{P}(Y)$ for all n. As Y is a compact metric space, $\mathcal{P}(X)$ is a compact metric space, hence there is a $\nu \in \mathcal{P}(Y)$ and a subsequence such that $\nu_{n_k} \to \nu$ in $\mathcal{P}(Y)$. We want to translate ν back to a measure on X. Set $Y_0 := T(X)$.

Claim: ν is concentrated on Y_0 in the sense that there exists a set $E \in \mathcal{B}(Y)$ with $E \subset Y_0$ and $\nu(E) = 1$.

If we assume the claim, define

$$\nu_0(A) := \nu(A \cap E), \quad A \in \mathcal{B}(Y_0).$$

(Note: $A \in \mathcal{B}(Y_0) \Rightarrow A \cap E$ Borel in $E \Rightarrow A \cap E$ Borel in Y, since E is a Borel subset of Y.) The measure ν_0 is a finite Borel measure on Y_0 and $\nu_0(E) = \nu(E) = 1$. Now we can translate ν_0 back to

$$\mu(A) := \nu_0(T(A)) = \nu_0((T^{-1})^{-1}(A)), \quad A \in \mathcal{B}(X).$$

Then $\mu \in \mathcal{P}(X)$. We want to show that $\mu_{n_k} \to \mu$ in $\mathcal{P}(X)$. Let C be closed in X. Then $T(C)$ is closed in $T(X) = Y_0$. ($T(C)$ need not be closed in Y.) Therefore there exists $Z \subset Y$ closed with $Z \cap Y_0 = T(C)$. Then $C = \{x \in X : T(x) \in T(C)\} = \{x \in X : T(x) \in Z\} = T^{-1}(Z)$, because there are no points in $T(C)$ outside Y_0, and $Z \cap E = T(C) \cap E$. Hence

$$\begin{aligned}
\limsup_{k\to\infty} \mu_{n_k}(C) &= \limsup_{k\to\infty} \nu_{n_k}(Z) \\
&\leq \nu(Z) \\
&= \nu(Z \cap E) + \nu(Z \cap E^c) = \nu(T(C) \cap E) + 0 \\
&= \nu_0(T(C)) = \mu(C).
\end{aligned}$$

So $\mu_{n_k} \to \mu$.

Finally, to prove the claim we use tightness of $\overline{\Gamma}$. For each $m \geq 1$ take K_m compact in X such that $\mu(K_m) \geq 1 - 1/m$ for all $\mu \in \Gamma$. Then $T(K_m)$ is a compact subset of Y hence closed in Y, so

$$\begin{aligned}
\nu(T(K_m)) &\geq \limsup_{k\to\infty} \nu_{n_k}(T(K_m)) \\
&\geq \limsup_{k\to\infty} \mu_{n_k}(K_m) \geq 1 - 1/m.
\end{aligned}$$

Take $E := \bigcup_{m=1}^{\infty} K_m$. Then $E \in \mathcal{B}(Y)$ and $\nu(E) \geq \nu(K_m)$ for all m, so $\nu(E) = 1$. □

Example. Let $X = \mathbb{R}$, $\mu_n(A) := n^{-1}\lambda(A \cap [0,n])$, $A \in \mathcal{B}(\mathbb{R})$. Here λ denotes Lebesgue measure on $\mathbb{R}$. Then $\mu_n \in \mathcal{P}(\mathbb{R})$ for all n. The sequence $(\mu_n)_n$ has no convergent subsequence. Indeed, suppose $\mu_{n_k} \to \mu$, then

$$\begin{aligned} \mu((-N,N)) &\leq \liminf_{n\to\infty} \mu_n((-N,N)) \\ &= \liminf_{n\to\infty} n^{-1}\lambda([0,N]) = \liminf_{n\to\infty} N/n = 0, \end{aligned}$$

so $\mu(\mathbb{R}) = \sup_{N\geq 1} \mu((-N,N)) = 0$. There is leaking mass to infinity; the set $\{\mu_n : n = 1,2,\ldots\}$ is not tight.

2.7 Marginals and disintegration

This section contains some basic and some advanced concepts from measure theory. We begin by recalling image measures and product measures.

Let $(X, \mathcal{A}, \mu)$ be a measue space and let $(T, \mathcal{B})$ be a measurable space. Recall that a map $r\colon X \to T$ is called measurable if $r^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$. The measure μ is mapped under r to a measure $r_\#\mu$ on T given by

$$r_\#\mu(B) := \mu(r^{-1}(B)), \quad B \in \mathcal{B}.$$

The measure $r_\#\mu$ is called the *push forward* or *image measure* of μ under r. Then for $B \in \mathcal{B}$,

$$\int_T 1_B(t)\, dr_\#\mu(t) = r_\#\mu(B) = \mu(r^{-1}(B)) = \int 1_{r^{-1}(B)}(x)\, d\mu(x) = \int_X 1_B(r(x))\, d\mu(x),$$

and by linear combinations and monotone convergence theorem this generalizes to

$$\int_T f(t)\, dr_\#\mu(t) = \int_X f(r(x))\, d\mu(x)$$

for each Borel measurable $f\colon T \to [0,\infty]$ and then also for each Borel function f which is integrable with respect to $r_\#\mu$.

If $\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, then a (real valued) *random variable* is a Borel measurable function $f\colon \Omega \to \mathbb{R}$. The image measure $f_\#\mathbb{P}$ is called the *(probability) law or distribution* of f.

Lemma 2.23. *If $(X_i, \mathcal{A}_i)$, $i = 1,2,3$, are measurable spaces, $f^i\colon X_i \to X_{i+1}$, $i = 1,2$, are measurable maps, and μ is a measure on $\mathcal{A}_1$, then*

$$f^2_\#(f^1_\#\mu) = (f^2 \circ f^1)_\#\mu.$$

Given two measure spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$, one can construct the *product space* of these two. A subset of $X \times Y$ is called a *rectangle* if it is of the form $A \times B$ for some $A \in \mathcal{A}$ and $B \in \mathcal{B}$. The σ-algebra in $X \times Y$ generated by the rectangles is called the *product σ-algebra*, denoted $\mathcal{A} \otimes \mathcal{B}$. On the rectangles we can define

$$\gamma(A \times B) = \mu(A)\nu(B)$$

and extend γ by Carathéodory's extension theorem to a measure on $\mathcal{A} \otimes \mathcal{B}$, called the *product measure* of μ and ν and denoted $\mu \otimes \nu$.

If γ is a measure on $X \times Y$ (more correctly, on $\mathcal{A} \otimes \mathcal{B}$), then its *first marginal* or *X-marginal* is the measure μ defined by

$$\mu(A) = \gamma(A \times Y), \quad A \in \mathcal{A},$$

and the *second marginal* or *Y-marginal* of γ is given by

$$\nu(B) = \gamma(X \times B), \quad B \in \mathcal{B}.$$

Clearly the product measure $\mu \otimes \nu$ has marginals μ and ν, but there may be many more measures with these marginals.

The marginals of a measure γ on $X \times Y$ are image measures under the *coordinate projections*. Define $\pi^1(x, y) = x$ and $\pi^2(x, y) = y$, $(x, y) \in X \times Y$, then $\pi^1_\# \gamma$ is the first marginal of γ and $\pi^2_\# \gamma$ the second marginal.

Obviously, the concept of marginals applies to larger products. If $(X_i, \mathcal{A}_i)$, $i = 1, \ldots, n$ are measurable spaces and if γ is a measure on $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$, then $\pi^j_\# \gamma$ is the jth marginal of γ, where $\pi^j(x_1, \ldots, x_n) = x_j$. From Lemma 2.23 it is clear that

$$\pi^1_\# \pi^{1,3}_\# \gamma = \pi^1_\# \gamma,$$

where $\pi^{1,3}(x_1, \ldots, x_n) = (x_1, x_3)$, and that the obvious similar formulas hold as well.

Next we will consider disintegration. Consider two probability spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$. If for each $x \in X$ a probability measure ν_x on $\mathcal{B}$ is given such that $x \mapsto \nu_x(B)$ is measurable for all $B \in \mathcal{B}$, then

$$\gamma(C) := \int_X 1_C(x, y)\, d\nu_x(y)\, d\mu(x), \quad C \in \mathcal{A} \otimes \mathcal{B},$$

defines a probability measure on the product $X \times Y$. How many of the probability measures on $X \times Y$ can we construct this way? The answer is: all of them, provided the spaces X and Y are sufficiently nice.

Theorem 2.24 (Disintegration, product form)**.** *Let (X, d_X) and (Y, d_Y) be two separable complete metric spaces. Let $\gamma \in \mathcal{P}(X \times Y)$ and $\mu(A) = \gamma(A \times Y)$ for $A \subseteq X$ Borel. Then for every $x \in X$ there exists a $\nu_x \in \mathcal{P}(Y)$ such that*

(i) $x \to \nu_x(B) : X \to \mathbb{R}$ is $\mathcal{B}_X$-measurable for every $B \in \mathcal{B}_Y$, and

(ii) $\displaystyle\int_{X \times Y} f(x, y)\, \mathrm{d}\gamma(x, y) = \int_X \left(\int_Y f(x, y)\, \mathrm{d}\nu_x(y) \right) \mathrm{d}\mu(x)$ for every Borel measurable $f : X \times Y \to [0, \infty]$.

If we make a picture of the situation of the theorem with X a horizontal line segment, Y a vertical line segment and $X \times Y$ a rectangle, then γ is a measure on the rectangle. For each subset A of X we can consider the vertical 'strip' $A \times Y$. Its γ-measure will be the marginal measure μ of the set A. The measure ν_x can be viewed as a measure on the vertical line above x. The theorem says that if we integrate for each x over the vertical line above x with respect to ν_x and then integrate these values over X with respect to μ, we retrieve the integral over the rectangle with respect to γ. It is probably superfluous to mention that an entirely similar theorem holds true where the inner integral runs over X and the outer integral over Y.

The above disintegration theorem for product spaces is a special case of the next theorem. In a picture we can imagine that instead of considering a rectangle divided into vertical lines we can take a different shape composed of disjoint curved lines. The next theorem gives a rigorous formulation of such a situation. The set Z replaces the product $X \times Y$ and the map π replaces the coordinate projection on X. As we do not have the second coordinate space Y anymore, the measures ν_x will be measures on the whole space Z, but concentrated on $\pi^{-1}(\{x\})$. This more general form also includes the case similar to the above theorem with the order of integration interchanged.

Theorem 2.25 (Disintegration, general form)**.** *Let (Z, d_Z) and (X, d_X) be separable complete metric spaces, let $\pi : Z \to X$ be a Borel map, let $\gamma \in \mathcal{P}(Z)$, and let $\mu(A) := \gamma(\pi^{-1}(A))$, $A \subseteq X$ Borel. Then for every $x \in X$ there exists a $\nu_x \in \mathcal{P}(Z)$ such that*

(i) ν_x is concentrated on $\pi^{-1}(\{x\})$, that is, $\nu_x(Z \setminus \pi^{-1}(\{x\})) = 0$ for μ-almost every $x \in X$,

(ii) $z \mapsto \nu_{\pi(x)}(C) : Z \to \mathbb{R}$ is Borel measurable for every Borel $C \subseteq Z$, and

(iii) $\displaystyle\int_Z f(z)\,\mathrm{d}\eta(z) = \int_X \left(\int_{\pi^{-1}(\{x\})} f(y)\,\mathrm{d}\nu_x(y) \right) \mathrm{d}\mu(x).$

The double integral at the right hand side can be rewritten. Firstly, the measure ν_x is concentrated on $\pi^{-1}(\{x\})$, so that the set of integration in the inner integral may be replaced by Z. Secondly, μ is the push forward of γ under π and the outer integral can therefore be transformed to an integral over Z, Thus the equality of (iii) becomes

$$\int_Z f(z)\,\mathrm{d}\eta(z) = \int_Z \left(\int_Z f(y)\,\mathrm{d}\nu_{\pi(z)}(y) \right) \mathrm{d}\gamma(z).$$

In this form it is clear that the inner integral at the right hand side is an integrable function of the variable z.

A dsicussion with proofs can be found in [7, Section 10.2, p. 341–351] or [5, III.70–74]. We will prove the general form at the end of the next section.

2.8 Conditional probabilities

Disintegration is closely related to the concept of *regular conditional probabilities* in probability theory. In fact, the disintegration theorems of the previous section follow from a theorem on existence of regular conditional probabilities. Since it is a fundamental concept in advanced probability theory and clarifies the importance of the underlying spaces being complete separable metric spaces, we include a discussion of regular conditional probabilities and an existence proof here. We begin with a discussion on conditional expectation.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider a (real valued) random variable f on Ω, that is, f is an $\mathcal{F}$-Borel measurable function from Ω to $\mathbb{R}$. For an interpretation, the probability space may be seen as a lottery. It draws outcomes ω from Ω inexhaustibly, some of them being more likely than others. The distribution of the likelyness is given by the measure $\mathbb{P}$. The probability at each time that the outcome ω will lie in a set A from $\mathcal{F}$ is $\mathbb{P}(A)$. The value of the random variable f will be $f(\omega)$, if the outcome of the lottery is ω. The *expected* (average) *value* of f (after very many drawings of the lottery) is

$$\mathbb{E}f := \int f(\omega)\,d\mathbb{P}(\omega) = \int_{\mathbb{R}} x\,df_{\#}\mathbb{P}(x).$$

The probability that the value of f will be in a Borel set B of $\mathbb{R}$ is $\mathbb{P}(\{\omega \in \Omega\colon\ f(\omega) \in B\})$, which equals $f_{\#}\mathbb{P}(B)$. Often one abbreviates

$$\{f \in B\} := \{\omega \in \Omega\colon\ f(\omega) \in B\}.$$

The probability that f is in B can be recovered from the expected value of the random variable 1_B, since

$$\mathbb{P}(f \in B) = \int_{\Omega} 1_{\{f\in B\}}\,d\mathbb{P} = \mathbb{E}1_{\{f\in B\}}.$$

Consider a set $A \in \mathcal{F}$ and its complement $A^c = \Omega \setminus A$. Suppose that for each drawing of the lottery someone tells us whether ω is in A or not. Then we still don't know what the value of f will be, but we can give a more precise expected average than $\mathbb{E}f$. Indeed, if ω is in A, then the expected value of f will be the average value of f on A, which is $\int_A f(\xi)\,d\mathbb{P}(\xi)/\mathbb{P}(A)$. If ω is said to be in A^c the expected value is $\int_{A^c} f(\xi)\,d\mathbb{P}(\xi)$. The combination of these two values is called the conditional expectation of f conditional on the information $\omega \in A$ or $\omega \in A^c$. Notice that the conditional expectation is not just a number, but depends on ω. It is itself a random variable.

Of course one can easily extend to the case that the given information is whether ω is in A_1 or A_2 or ... in A_n, where $A_1 \dots, A_n$ are disjoints sets with $A_1 \cup \cdots \cup A_n = \Omega$. The conditional expectation g of f is now equal to $\int_{A_i} f(\omega)\,d\mathbb{P}(\omega)/\mathbb{P}(A_i)$ if $\omega \in A_i$. This can be written in one formula as

$$g(\omega) = \sum_{i=1}^{n} \int_{A_i} f(\xi)\,d\mathbb{P}(\xi)/\mathbb{P}(A_i)1_{A_i}(\omega).$$

It is easily observed that the conditional expectation g is the unique function that g such that g is constant on each A_i and on that set the average of g equals that of f, which comes down to $\int_{A_i} g\,d\mathbb{P} = \int_{A_i} f\,d\mathbb{P}$ for $i = 1, \dots, n$.

The situation is more difficult if there are infinitely many sets A involved. For instance, if $\Omega = [0, 1]$ it could be that someone tells us if ω is between 0 and 1/3, between 1/3 and 2/3 or between 2/3 and 1. In the first case we are also told whether ω is in $[0, 1/9)$, $[1/9, 2/9)$ or $[2/9, 1/3)$ and in the third case whether ω is in $[2/3, 2/3 + 1/9)$, $[2/3 + 1/9, 2/3 + 2/9)$ or $[2/3 + 2/9, 1]$. In each of the cases $[0, 1/9)$, $[2/9, 1/3)$, $[2/3, 2/3 + 1/9)$, and $[2/3 + 2/9, 1]$ we get the information in which of the three equal parts ω is and then for each first and third part of the subinterval again information in which of the three equal parts, and so on. An explicit formula for the conditional expectation is now more difficult. In the general setting, the information given is described by a sub-σ-algebra $\mathcal{C}$ of $\mathcal{F}$. We suppose that for each $A \in \mathcal{C}$ we are told whether ω is in A or not. In the case of the finite partition $A_1, \dots, A_n$, $\mathcal{C}$ will simply be the (finite) σ-algebra generated by these sets. Knowing whether $\omega \in A_i$ or nor for each i is equivalent to knowing it for each set of the σ-algebra. It may not be possible to partition Ω into smallest disjoint pieces of $\mathcal{C}$. Instead of saying that the conditional expectation should be constant on each of the pieces of a disjoint partition we require instead that it should be $\mathcal{C}$ measurable. In the above cases of finitely many sets this condition is equivalent. Thus we arrive at the following definition.

Definition 2.26. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $f\colon \Omega \to \mathbb{R}$ be a random variable, and let $\mathcal{C}$ be a sub-σ-algebra of $\mathcal{F}$. The *conditional expectation* of f given $\mathcal{C}$, denoted $\mathbb{E}[f|\mathcal{C}]$ is a random variable on Ω such that

(a) $\mathbb{E}[f|\mathcal{C}]$ is $\mathcal{C}$-measurable

(b) $\displaystyle\int_A \mathbb{E}[f|\mathcal{C}]\, d\mathbb{P} = \int_A f\, d\mathbb{P}$ for all $A \in \mathcal{C}$.

It is a consequence of the Radon-Nikodym theorem that $\mathbb{E}[f|\mathcal{C}]$ exists for each $\mathcal{F}$-measurable f with $\int |f|\, d\mathbb{P} < \infty$ and that it is unique up to $\mathbb{P}$-almost everywhere equality. Furthermore, $f \mapsto \mathbb{E}[f|\mathcal{C}]$ in linear from the vector space of $\mathcal{F}$-measurable $\mathbb{P}$-integrable functions to the vector space of $\mathcal{C}$-measurable $\mathbb{P}$-integrable functions, $\mathbb{E}[1|\mathcal{C}] = 1$ a.e. on Ω, and if $f \le g$ a.e. on Ω, then $\mathbb{E}[f|\mathcal{C}] \le \mathbb{E}[g|\mathcal{C}]$ a.e. on Ω. The following lemma is sometimes useful.

Lemma 2.27 (Monotone convergence for conditional expectations)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $f_1, f_2, \dots$ be positive $\mathcal{F}$-measurable $\mathbb{P}$-integrable functions on Ω and let $\mathcal{C}$ be a sub-σ-algebra of $\mathcal{F}$. If $f_k \le f_{k+1}$ a.e. for each k and $f = \lim_{k\to\infty} f_k$ a.e., then*

$$\mathbb{E}[f|\mathcal{C}] = \lim_{k\to\infty} \mathbb{E}[f_k|\mathcal{C}] \ \textit{a.e. on}\ \Omega.$$

Proof. Denote $g_k = \mathbb{E}[f_k|\mathcal{C}]$, $k \in \mathbb{N}$. We have $0 \le g_k \le g_{k+1}$ a.e. for all k. Let $g := \sup_{k\in\mathbb{N}} g_k$. Then g is $\mathcal{C}$-measurable and positive a.e. By means of the monotone convergence theorem applied to f and to g we obtain that

$$\int_C g\, d\mathbb{P} = \lim_k \int_C g_k\, d\mathbb{P} = \int_C f_k\, d\mathbb{P} = \int_C f\, d\mathbb{P},$$

for every $C \in \mathcal{C}$, which means that g is (a.e. equal to) the conditional expectation of f given $\mathcal{C}$. □

Is there a similar concept of conditional probability? If we throw a dice and someone tells us that the outcome is odd, we know what the probability is that the outcome is 1, 2, 3, 4, 5 or 6. In the general setting of a random variable f on a probabiity space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-algebra $\mathcal{C}$ of $\mathcal{F}$ a reasonable definition of the conditional probability that $f \in B$ given $\mathcal{C}$ would be

$$P_{f|\mathcal{C}}(B) = \mathbb{E}[1_{\{f\in B\}}|\mathcal{C}].$$

This leads to serious technical difficulties. To make sense, the conditional probabilities that f is in B or not in B should add up to 1. More generally, given that we know for each ω whether it is in C or not for every $C \in \mathcal{C}$, the probability that $f \in B$ should give a probability measure on the collection of Borel sets B of $\mathbb{R}$. The random variable $\mathbb{E}[1_{\{f\in B\}}|\mathcal{C}]$, however, is only determined up to $\mathbb{P}$-almost everywhere equality. This means that for a Borel set $B \subseteq \mathbb{R}$ the conditional probabilities that $f \in B$ or $f \notin B$ need not add up to 1 on a subset of Ω of measure 0. Of course we can ignore this set. However, the collection of all Borel sets B of $\mathbb{R}$ is uncountable, so we can not simply take union of all exception sets of measure 0 to make sure that the conditional probability is a probablity measure for almost every ω in Ω. We will list the desired properties of conditional probabilities in the next definition and then study its existence. Instead of only real-valued random variables we may as well consider random variables with values in $\mathbb{R}^n$ or even more general metric spaces.

Definition 2.28. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(T, \mathcal{B})$ a measure space, $F\colon \Omega \to T$ a measurable map, and $\mathcal{C} \subseteq \mathcal{F}$ a sub-σ-algebra. A *(regular) conditional probability distribution* for F given $\mathcal{C}$ is a function $P_{f|\mathcal{C}}\colon \mathcal{B} \times \Omega \to [0,1]$ such that

(a) $B \mapsto P_{f|\mathcal{C}}(B, \omega)$ is a probability measure on $\mathcal{B}$ for $\mathbb{P}$-almost every $\omega \in \Omega$

(b) For every $B \in \mathcal{B}$ the map $\omega \mapsto P_{f|\mathcal{C}}(B, \omega)$ is $\mathcal{C}$-measurable and

$$P_{f|\mathcal{C}}(B, \omega) = \mathbb{E}[1_{\{F \in B\}} | \mathcal{C}](\omega) \text{ for } \mathbb{P}\text{-almost every } \omega \in \Omega.$$

A regular conditional probability distribution does not always exist. As pointed out above, the difficulty in the definition is that there may be uncountably many sets B in $\mathcal{B}$. It turns out that a regular conditional probability can be constructed if the sets of $\mathcal{B}$ can suitably be approximated by sets of a countable subcollection. It is here where topology starts playing a role. Such an approximation by a countable collection can be carried out if T is a separable complete metric space.

Theorem 2.29 (Existence of regular conditional probabilities)**.** *If T is a separable complete metric space, $\mathcal{B}$ its Borel σ-algebra, $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $F\colon \Omega \to T$ a measurable map, and $\mathcal{C}$ a sub-σ-algebra of $\mathcal{F}$, then there exists a regular conditional probability distribution $P_{f|\mathcal{C}}$ for F given $\mathcal{C}$ on $\mathcal{B} \times \Omega$. It is unique in the following sense: if P' is another regular conditional probability distribution for F given $\mathcal{C}$, then for $\mathbb{P}$-a.e. $\omega \in \Omega$ we have*

$$P'(B, \omega) = P_{f|\mathcal{C}}(B, \omega) \textit{ for all } B \in \mathcal{B}.$$

In the proof of the theorem we will need the following lemma. Recall that a collection $\mathcal{V}$ of subsets of a set T is called an *algebra* if it contains the empty set, $T \setminus A \in \mathcal{V}$ for all $A \in \mathcal{V}$ and $A \cup B \in \mathcal{V}$ for all $A, B \in \mathcal{V}$.

Lemma 2.30. *If $\mathcal{V}$ and $\mathcal{D}$ are two algebras of subsets of a separable complete metric space T, $\mathcal{V} \subseteq \mathcal{D}$ and $\mu\colon \mathcal{D} \to [0, \infty)$ is (finitely) additive and for every $B \in \mathcal{V}$ we have*

$$\mu(B) = \sup\{\mu(B)\colon\ K \subseteq B,\ K \in \mathcal{D},\ K \textit{ compact}\},$$

then μ is σ-additive on $\mathcal{V}$.

Proof. Suppose not: there are $B_1, B_2, \dots$ in $\mathcal{V}$ disjoint and $B \in \mathcal{V}$ such that $B = \bigcup_{i=1}^\infty B_i$ and $\delta := \mu(B) - \sum_{i=1}^\infty \mu(B_i) > 0$. Let $C_k := B \setminus \bigcup_{j=1}^k B_j$. Then $C_k \in \mathcal{V}$, $C_k \supseteq C_{k+1}$, $\bigcap_{k=1}^\infty C_k = \emptyset$, and $\mu(C_k) \geq \delta$ for all k. Take for each k a compact set $K_k \in \mathcal{D}$ with $K_k \subseteq C_k$ such that $\mu(C_k \setminus K_k) < 2^{-k}\delta/2$. Then

$$\begin{aligned} \mu(K_1 \cap \cdots \cap K_n) &= \mu((C_1 \cap \cdots \cap C_n) \setminus \bigcup_{k=1}^n (C_k \setminus K_k)) \\ &\geq \mu(C_n) - \sum_{k=1}^n \mu(C_k \setminus K_k) \\ &\geq \mu(C_n) - \delta/2 \geq \delta/2, \end{aligned}$$

so $K_1 \cap \cdots \cap K_n \neq \emptyset$ for all n. Since the sets K_k are compact it follows that $\bigcap_{k=1}^\infty K_k \neq \emptyset$. This contradicts $C_k \supseteq K_k$ and $\bigcap_{k=1}^\infty C_k = \emptyset$. □

Proof of Theorem 2.29. Choose a countable dense subset $\{t_1, t_2, \ldots\}$ of T. Let

$$\mathcal{U} := \{B_r(t_k)\colon\ k \in \mathbb{N},\ r \in \mathbb{Q}, r \geq 0\},$$

where $B_r(t)$ denotes the open ball in T with center t and radius r. The collection $\mathcal{U}$ is countable and generates the σ-algebra $\mathcal{B}$. Let $\mathcal{V}$ be the algebra generated by $\mathcal{U}$. Then also $\mathcal{V}$ is countable. (Indeed, there are finite $\mathcal{U}_1 \subseteq \mathcal{U}_2 \subseteq \cdots$ such that $\mathcal{U} = \bigcup_{k=1}^\infty \mathcal{U}_k$. The algebra $\mathcal{V}_k$ genereted by $\mathcal{U}_k$ is also finite and $\bigcup_{k=1}^\infty \mathcal{V}_k$ is an algebra and equals $\mathcal{V}$. Hence $\mathcal{V}$ is countable.) The image measure $\mu_F = F_\# \mathbb{P}$ of $\mathbb{P}$ under F is a Borel probability measure on T and since T is separable and complete μ_F is tight. Hence for every $B \in \mathcal{V}$ we can choose a sequence $B_1 \subseteq B_2 \subseteq B_3 \subseteq \cdots$ of compact sets in T with $B_k \subseteq B$ for every k such that $\mu_F(B) = \lim_{k\to\infty} \mu_F(B_k)$. Then the functions $1_{\{F \in B_k\}}$ increase in k and converge $\mathbb{P}$-a.e. to $1_{\{F \in B\}}$. Due to mononote convergence for conditional expectations,

$$\mathbb{E}[1_{\{F \in B_k\}} | \mathcal{C}] \to \mathbb{E}[1_{\{F \in B\}} | \mathcal{C}].$$

Let $\mathcal{D}$ be the algebra of subsets of T generated by $\mathcal{V}$ and by the compact sets of each of the sequences $B_1 \subseteq B_2 \subseteq$ that have chosen above for each $B \in \mathcal{V}$. Then $\mathcal{D}$ is countable.

For each set $D \in \mathcal{D}$ the conditional expectation $\mathbb{E}[1_{\{F \in D\}} | \mathcal{C}]$ is determined up to $\mathbb{P}$-a.e. equality. Let us fix for each $D \in \mathcal{D}$ a particular choice of $\mathbb{E}[1_{\{F \in D\}} | \mathcal{C}]$ on Ω and define

$$P_{f|\mathcal{C}}(D, \omega) := \mathbb{E}[1_{\{F \in D\}} | \mathcal{C}],\ \omega \in \Omega.$$

We claim that there exists a subset $W \in \mathcal{F}$ with $\mathbb{P}(W) = 0$ such that

(1) For every $D \in \mathcal{D}$, $\omega \mapsto P_{F|\mathcal{C}}(D, \omega)$ is $\mathcal{C}$-measurable;

(2) For every $D \in \mathcal{D}$, $P_{F|\mathcal{C}}(D, \omega) \geq 0$ for all $\omega \in \Omega \setminus W$;

(3) $P_{F|\mathcal{C}}(T, \omega) = 1$ and $P_{F|\mathcal{C}}(\emptyset, \omega) = 0$ for all $\omega \in \Omega \setminus W$;

(4) For every $D_1, \ldots, D_n$ in $\mathcal{D}$ disjoint,

$$P_{F|\mathcal{C}}(D_1 \cup \cdots \cup D_n, \omega) = \sum_{j=1}^n P_{F|\mathcal{C}}(D_j, \omega) \text{ for all } \omega \in \Omega \setminus W;$$

(5) For each $B \in \mathcal{V}$ with the sequence(B_j) corresponding to B as chosen above,

$$P_{F|\mathcal{C}}(B, \omega) = \lim_{j\to\infty} \mathbb{E}[1_{\{F \in B_j\}} | \mathcal{C}](\omega) \text{ for all } \omega \in \Omega \setminus W.$$

Indeed, the functions $\omega \mapsto P_{F|\mathcal{C}}(D, \omega)$, $D \in \mathcal{D}$, defined above satisfy all these properties if 'for all $\omega \in \Omega \setminus W$' is replaced by 'for $\mathbb{P}$-almost every ω in Ω'. Since there are only countably many sets D in $\mathcal{D}$, we can take W to be the union of all the exception sets in the '$\mathbb{P}$-almost everywhere' relations. Then $\mathbb{P}(W) = 0$ and (1)–(5) hold.

Next we extend the definition of $P_{F|\mathcal{C}}(B, \omega)$ to all $B \in \mathcal{B}$ and show it to be a probability measure. Because of (2), (3), and (4), the map $D \mapsto P_{F|\mathcal{C}}(D, \omega)$ is additive and positive. Because of the lemma and property (5), for every $\omega \in \Omega \setminus W$ the map $D \mapsto P_{F|\mathcal{C}}(D, \omega)$

is σ-additive on the algebra $\mathcal{V}$. By the Carathéodory extension theorem, it extends to a σ-additive measure μ_ω on the σ-algebra generated by $\mathcal{V}$, which is $\mathcal{B}$ as $\mathcal{B} \supseteq \mathcal{V} \supseteq \mathcal{U}$ and $\mathcal{U}$ generates $\mathcal{B}$. Clearly $\mu_\omega(T) = 1$, by (3). We show that $(B, \omega) \mapsto \mu_\omega(B)$ has the desired properties of the regular conditional expectation.

Let

$$\begin{aligned} \mathcal{E} := \{B \in \mathcal{B} \colon\ & \omega \mapsto \mu_\omega(B) \text{ is } \mathcal{C}\text{-measurable and} \\ & \mu_\omega(B) = \mathbb{E}[1_{\{F \in B\}} | \mathcal{C}](\omega) \text{ for } \mathbb{P}\text{-a.e. } \omega \in \Omega\}. \end{aligned}$$

Then $\mathcal{E} \supseteq \mathcal{V}$ since $\mu_\omega(B) = P_{F|\mathcal{C}}(B, \omega) = \mathbb{E}[1_{\{F \in B\}} | \mathcal{C}]$ for $B \in \mathcal{V}$ and we have (1). Also, $\mathcal{E}$ is a σ-algebra. For a proof, notice that $\emptyset \in \mathcal{V} \subseteq \mathcal{E}$ and that for $B \in \mathcal{E}$ we have that $\omega \mapsto \mu_\omega(T \setminus B) = 1 - \mu_\omega(B)$ is $\mathcal{C}$-measurable and

$$\begin{aligned} \mu_\omega(T \setminus B) &= 1 - \mu_\omega(B) = 1 - \mathbb{E}[1_{\{F \in B\}} | \mathcal{C}](\omega) \\ &= \mathbb{E}[1_\Omega - 1_{\{F \in B\}} | \mathcal{C}](\omega) = \mathbb{E}[1_{\Omega \setminus \{F \in B\}} | \mathcal{C}](\omega) \\ &= \mathbb{E}[1_{\{F \in T \setminus B\}} | \mathcal{C}](\omega) \text{ a.e. } \omega \in \Omega, \end{aligned}$$

so that $T \setminus B \in \mathcal{E}$. Further, if $B_1, B_2, \ldots \in \mathcal{E}$ are disjoint, $B = \bigcup_{k=1}^\infty B_k$, then $\mu_\omega(B) = \sum_{k=1}^\infty \mu_\omega(B_k)$ for almost every $\omega \in \Omega$, so $\omega \to \mu_\omega(B)$ is $\mathcal{C}$-measurable, and by the monotone convergence theorem for conditional expectations,

$$\begin{aligned} \mu_\omega(B) &= \sum_{k=1}^\infty \mathbb{E}[1_{\{F \in B_k\}} | \mathcal{C}](\omega) = \mathbb{E}[\sum_{k=1}^\infty 1_{\{F \in B_k\}} | \mathcal{C}](\omega) \\ &= \mathbb{E}[1_{\bigcup_{k=1}^\infty \{F \in B_k\}} | \mathcal{C}](\omega) = \mathbb{E}[1_{\{F \in B\}} | \mathcal{C}](\omega) \text{ a.e. } \omega \in \Omega, \end{aligned}$$

so $B \in \mathcal{E}$. Hence $\mathcal{E}$ is a σ-algebra. Since $\mathcal{B}$ is the smallest σ-algebra containing $\mathcal{U}$ and $\mathcal{E}$ is a σ-algebra containing $\mathcal{V} \supseteq \mathcal{U}$, we conclude that $\mathcal{E} \supseteq \mathcal{B}$. Hence every $B \in \mathcal{B}$ satisfies the two properties in the definition of $\mathcal{E}$, which means that

$$(B, \omega) \mapsto P_{F|\mathcal{C}}(B, \omega) := \mu_\omega(B)$$

is a regular conditional probability distribution for F given $\mathcal{C}$.

To see the uniqueness, we use that $P'(B, \omega) = \mathbb{E}[1_{\{F \in B\}} | \mathcal{C}](\omega)$ for almost every $\omega \in \Omega$ and every $B \in \mathcal{B}$. Since $\mathcal{V}$ is countable we can combine the exception sets for $B \in \mathcal{V}$ and obtain a $W' \in \mathcal{F}$ with $\mathbb{P}(W') = 0$ such that $P'(B, \omega) = P_{F|\mathcal{C}}(B, \omega)$ for every $\omega \in \Omega \setminus (W \cup W')$ for every $B \in \mathcal{V}$. Now fix $\omega \in \Omega \setminus (W \cup W')$. Since $P'(\cdot, \omega)$ and $P_{F|\mathcal{C}}(\cdot, \omega)$ are both probability measures on $\mathcal{B}$ and $\mathcal{V}$ is an algebra generating $\mathcal{B}$ on which they coincide, they must be equal on $\mathcal{B}$ (by a uniqueness theorem related to Carathéodory's extension). Hence for $\mathbb{P}$-almost every $\omega \in \Omega$ we have

$$P'(B, \omega) = P_{F|\mathcal{C}}(B, \omega) \text{ for all } B \in \mathcal{B}.$$

□

Let us next show how the disintegration theorem follows from the existence of regular conditional probabilities.

Proof of disintegration theorem, general form. Take $(\Omega, \mathcal{F}, \mathbb{P}) = (Z, \mathcal{B}_Z, \gamma)$, $\mathcal{C} = \{\pi^{-1}(A) \colon A \in \mathcal{B}_X\}$, $(T, \mathcal{B}) = (Z, \mathcal{B}_Z)$, and $F(z) = z$ for all $z \in Z$. Then the regular conditional probability $P_{F|\mathcal{C}} \colon \mathcal{B}_Z \times Z \to [0, 1]$ exists.

For each $B \in \mathcal{B}_Z$, $z \mapsto P_{F|\mathcal{C}}(B, z)$ and $z \mapsto \mathbb{E}[1_{\{F \in B\}}|\mathcal{C}](z)$ are equal γ-almost everywhere on Z. Both are $\mathcal{C}$-measurable and therefore constant on each set of the form $\pi^{-1}(\{x\})$, where $x \in X$. (Otherwise the subset of $\pi^{-1}(\{x\})$ where it equals one of the different values would be a non-empty strict subset, but there is no Borel set A in $\mathbb{R}$ for which $\pi^{-1}(A)$ is a non-empty strict subset of $\pi^{-1}(\{x\})$, which contradicts the $\mathcal{C}$-measurability.) From constant and almost everywhere equal, it also follows that $P_{F|\mathcal{C}}(B, z) = \mathbb{E}[1_{\{F \in B\}}|\mathcal{C}](z)$ for *all* $z \in \pi^{-1}(\{x\})$ (and not just for almost all z).

Let Z_1 be the set of all $z \in Z$ for which $B \mapsto P_{F|\mathcal{C}}(B, z)$ is a probability measure. Then $\gamma(Z_1) = 1$. Let X_1 be set of all $x \in X$ for which there exists a $z \in \pi^{-1}(\{x\})$ with $z \in Z_1$. Then $\mu(X_1) = 1$. Indeed, for every $z \in \pi^{-1}(X \setminus X_1)$ the set $\pi^{-1}\{\pi(z)\}$ doesnot contain an element of Z_1, so $z \in Z \setminus Z_1$, hence $\pi^{-1}(X \setminus X_1) \subseteq Z \setminus Z_1$ and therefore $\mu(X \setminus X_1) = \gamma(\pi^{-1}(X \setminus X_1)) \leq \gamma(Z \setminus Z_1) = 0$.

For $x \in X_1$ we can now unambiguously define

$$\nu_x(B) := P_{F|\mathcal{C}}(B, z), \quad B \in \mathcal{B}_Z,$$

for some $z \in \pi^{-1}(\{x\}) \cap Z_1$. Fix any probability measure ν' on $\mathcal{B}_Z$. For $x \in X \setminus X_1$ we define $\nu_x := \nu'$.

Then clearly for $x \in X$, ν_x is a Borel probability measure on Z. For $x \in X_1$, we have for some $z \in \pi^{-1}(\{x\})$ that

$$\begin{aligned}\nu_x(Z \setminus \pi^{-1}(\{x\}) &= P_{F|\mathcal{C}}(Z \setminus \pi^{-1}(\{x\}), z) \\ &= \mathbb{E}[1_{\{F \in Z \setminus \pi^{-1}(\{x\})\}}|\mathcal{C}](z) \\ &= \mathbb{E}[1_{Z \setminus \pi^{-1}(\{x\})}|\mathcal{C}](z) \\ &= 1_{Z \setminus \pi^{-1}(\{x\})}(z) = 0,\end{aligned}$$

where we used that $Z \setminus \pi^{-1}(\{x\}) = \pi^{-1}(X \setminus \{x\}) \in \mathcal{C}$. Hence ν_x is concentrated on $Z \setminus \pi^{-1}(\{x\})$ for all $x \in X_1$.

For any $B \in \mathcal{B}_Z$ we have that $z \mapsto \nu_{\pi(z)}(B) = P_{F|\mathcal{C}}(B, z)1_{Z_1}(z) + \nu' 1_{Z \setminus Z_1}(z)$, which is $\mathcal{B}_Z$ measurable.

Finally, for $B \in \mathcal{B}_Z$,

$$\begin{aligned}\int_X \int_{\pi^{-1}(\{x\})} 1_B(y)\, d\nu_x(y)\, d\mu(x) &= \int_Z \int_Z 1_B(y)\, d\nu_{\pi(z)}(y)\, d\gamma(z) \\ &= \int_Z \nu_{\pi(z)}(B)\, d\gamma(z) = \int_Z P_{F|\mathcal{C}}(B, z)\, d\gamma(z) \\ &= \int_Z \mathbb{E}[1_{\{F \in B\}}|\mathcal{C}](z)\, d\mathbb{P}(z) \\ &= \int_Z 1_{\{F \in B\}}\, d\mathbb{P} \\ &= \int_Z 1_B(z)\, d\gamma(z).\end{aligned}$$

Hence

$$\int_X \int_{\pi^{-1}(\{x\})} f(y)\, d\nu_x(y)\, d\mu(x) = \int_Z f(z)\, d\gamma(z)$$

holds for $f = 1_B$ for any $B \in \mathcal{B}_Z$. Then by linear combination also for functions of the from $f = \sum_{k=1}^n \alpha_k 1_{B_k}$. Then by the monotone convergence theorem the formula also holds for any positive Borel measurable function $f\colon Z \to [0,\infty]$. By taking difference of two positive Borel functions we also find the formula for any bounded Borel function f. $\square$

3 Optimal transportation problems

Optimal transportation problems aim to minimize costs or energy needed to transport mass from a given initial state to a given final state. We will consider the Monge and Kantorovich optimal transportation problems in metric spaces and discuss existence and uniqueness of optimal transportation plans.

3.1 Introduction

Monge studied a question concerning transportation of a volume of mass from a given initial position to a given end position in such a way that the total cost of the transportation computed as mass times distance is minimal. In modern formulas such a problem reads as follows. Let V_0, V_1 be open subsets of $\mathbb{R}^d$ with equal volumes (i.e., equal Lebesgue measures). Find a bijective map $r\colon V_0 \to V_1$ such that

$$\int_V V_0 \|x - r(x)\| \, dx$$

is minimal. Without loss of generality the total volume may be assumed equal to 1. Instead of a volume of mass, one could consider a distribution of mass over $\mathbb{R}^d$ given by a probability measure μ (or, more specifically, a density function f). The desired end distribution is geven by a probability measure ν. A transportation map r should be such that all mass that should be in a set B in the end situation should equal all mass transported to B from somewhere in the initial state. Thus we should require that $\nu(B) = \mu(r^{-1}(B))$ for each Borel set B. That is, $\nu = r_{\#}\mu$. Instead of $\mathbb{R}^d$ one could consider arbitrary separable complete metric spaces and instead of the distance one could consider an arbotrary cost function c. This yields the problem that is nowadays referred to as *Monge's problem*:

> Given two separable complete metric spaces X and Y and two measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and given a Borel measurable function $c\colon X \times Y \to [0,\infty)$, find a Borel measurable map $r\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $\nu = r_{\#}\mu$ and
>
> $$\int_X c(x, r(x)) \, d\mu(x)$$
>
> is minimal.

Kantorovich stated a more general version of this problem in 1942. His idea is to describe a transportation by a measure η on the product space $X \times Y$. The amount of mass transportated from A to B is then given by $\eta(A \times B)$. All mass that should be present in the set B in the end situation should come from somewhere, so $\nu(B) = \eta(X \times B)$. Similarly all mass present in A in the begin should go somewehere, so $\mu(A) = \eta(A \times Y)$. Hence μ and ν are the first and second marginals of η Denote the set of all *transport plans* by

$$\Gamma(\mu,\nu) = \{\eta \in \mathcal{P}(X \times Y)\colon\ \eta \text{ has first marginal } \mu \text{ en second mnarginal } \nu\}.$$

Kantorovich's problem is:

Given two separable complete metric spaces X and Y and two measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and given a Borel measurable function $c\colon X \times Y \to [0, \infty)$, find $\gamma \in \Gamma(\mu, \nu)$ such that

$$\int_{X \times Y} c(x, y)\, d\eta(x, y)$$

is minimal.

Definition 3.1. The measure $\eta \in \Gamma(\mu, \nu)$ is called *optimal for c* if

$$\int c\, \mathrm{d}\eta = \min\{\int c(x, y)\, \mathrm{d}\gamma(x, y) \colon\ \gamma \in \Gamma(\mu, \nu)\}$$

(possibly $\infty = \infty$).

3.2 Existence for the Kantorovich problem

The existence of an optimal measure for the Kantorovich problem is a consequence of Prokhorov's theorem.

Lemma 3.2. *Let X and Y be separable complete metric spaces and let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then $\Gamma(\mu, \nu)$ is a compact subset of $\mathcal{P}(X \times Y)$.*

Proof. Notice that $X \times Y$ is also separable and complete. We first show that $\Gamma(\mu, \nu)$ is tight. As X and Y are separable and complete, the measures μ and ν are tight. Let $\varepsilon > 0$. Choose compact sets $K \subseteq X$ and $L \subseteq Y$ such that $\mu(K) \geq 1 - \varepsilon/2$ and $\nu(L) \geq 1 - \varepsilon/2$. Then $K \times L$ is compact in $X \times Y$ and for $\gamma \in \Gamma(\mu, \nu)$,

$$\begin{aligned} \gamma(X \times Y \setminus K \times L) &\leq \gamma((X \setminus K) \times Y) + \gamma(X \times (Y \setminus L)) \\ &= \nu(X \setminus K) + \mu(Y \setminus L) \leq \varepsilon/2 + \varepsilon/2. \end{aligned}$$

Hence $\Gamma(\mu, \nu)$ is tight. By Prokhorov's theorem, $\Gamma(\mu, \nu)$ is relatively compact in $\mathcal{P}(X \times Y)$.

It remains to show that $\Gamma(\mu, \nu)$ is closed in $\mathcal{P}(X \times Y)$. Let $(\gamma_n)_n$ be a sequence in $\Gamma(\mu, \nu)$ and $\eta \in \mathcal{P}(X \times Y)$ be such that $\gamma_n \to \eta$ narrowly. Due to the Portmanteau theorem we have for any $C \subseteq X$ closed,

$$\begin{aligned} \eta(C \times Y) &\geq \limsup_{n \to \infty} \eta_n(C \times Y) \\ &= \limsup_{n \to \infty} \mu(C) = \mu(C) \end{aligned}$$

and for $U \subseteq X$ open,

$$\begin{aligned} \eta(U \times Y) &\leq \liminf_{n \to \infty} \eta_n(U \times Y) \\ &= \liminf_{n \to \infty} \mu(U) = \mu(U). \end{aligned}$$

Let $C \subseteq X$ be closed and let

$$U_m := \{x \in X \colon\ \mathrm{dist}(x, C) < 1/m\}, \quad m \geq 1.$$

Then each U_m is open and $\bigcap_{m \geq 1} U_m = C$ and $\bigcap_{m \geq 1} (U_m \times Y) = C \times Y$. Hence

$$\begin{aligned} \eta(C \times Y) &= \lim_{m \to \infty} \eta(U_m \times Y) \\ &\leq \lim_{m \to \infty} \mu(U_m) = \mu(C). \end{aligned}$$

Thus, $\eta(C \times Y) = \mu(C)$. Hence μ is the marginal on X of η. In a similar way we can show that the marginal of η on Y is ν and therefore $\eta \in \Gamma(\mu, \nu)$. □

Theorem 3.3. *Let X and Y be separable complete metric spaces, let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Let $c\colon X \times Y \to [0, \infty)$ be continuous. Then there exists $\eta \in \Gamma(\mu, \nu)$ such that*

$$\int c \, \mathrm{d}\eta = \min\{\int c \, \mathrm{d}\gamma\colon \ \gamma \in \Gamma(\mu, \nu)\}.$$

Proof. Write $\alpha := \min\{\int c \, \mathrm{d}\gamma\colon \ \gamma \in \Gamma(\mu, \nu)\}$. If $\alpha = \infty$, then $\eta = \mu \otimes \nu \in \Gamma(\mu, \nu)$ satisfies $\int c \, \mathrm{d}\eta = \infty = \alpha$. Otherwise, for $n \geq 1$, take a $\gamma_n \in \Gamma(\mu, \nu)$ with

$$\int c \, \mathrm{d}\eta_n \leq \alpha + 1/n.$$

By the previous lemma, $\Gamma(\mu, \nu)$ is compact. Hence there exists $\eta \in \Gamma(\mu, \nu)$ and a subsequence $eta_{n_k} \to \eta$ as $k \to \infty$. For $k \geq 1$ and $m \geq 1$ we have $\int c \wedge m \, \mathrm{d}\eta_{n_k} \leq \int c \, \mathrm{d}\eta_{n_k} \leq \alpha + 1/n_k$. Hence

$$\int c \, \mathrm{d}\eta = \lim_{m \to \infty} \int c \wedge m \, \mathrm{d}\eta = \lim_{m \to \infty} \lim_{k \to \infty} \int c \wedge m \, \mathrm{d}\eta_{n_k} \leq \alpha.$$

□

Remark. The previous theorem is a special instance of a general principle: a l.s.c. function on a compact metric space has a minimum. The above lemma says that $\Gamma(\mu, \nu)$ is a compact metric space. The map $\eta \mapsto \int c \, \mathrm{d}\eta$ is l.s.c. due to Proposition 2.12(c).

3.3 Structure of optimal transportation plans

The support of an optimal transportation plan is a special type of subset of $X \times Y$. The *support* of a measure η on a metric space Z is defined by

$$\operatorname{supp} \eta = \{z \in Z\colon \ \eta(U) > 0 \text{ for every open neighborhood } U \text{ of } z\}.$$

Suppose that a certain amount of mass is transported from $x1$ to y_1 and from x_2 to y_2. Then transporting that mass from x_2 to y_1 and from x_1 to y_2 (keeping the rest the same) would also yield a transportation plan. If the plan is optimal, such permutations could not decrease the cost. This observation leads to the following definition.

Definition 3.4. A set $S \subseteq X \times Y$ is called *c-monotone* if

$$\sum_{i=1}^{n} c(x_{\sigma(i)}, y_i) \geq \sum_{i=1}^{n} c(x_i, y_i)$$

for every $(x_i, y_i) \in S$, $i = 1, \ldots, n$, and every permutation σ of $\{1, \ldots, n\}$.

The next theorem states that optimal transportation plans have c-monotone supports. We begin with a lemma.

Lemma 3.5. *Let X and Y be separable complete metric metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $\eta \in \Gamma(\mu, \nu)$. Then there exists a μ-full Borel set $A \subseteq X$ such that*

$$\forall x \in A\ \exists y \in Y\colon \quad (x, y) \in \operatorname{supp} \eta.$$

Further, $\mu(\pi^X(\operatorname{supp}\eta)) = 1$ and $\nu(\pi^Y(\operatorname{supp}\eta)) = 1$.

Proof. The set $S := \operatorname{supp}\eta$ is closed hence Borel. As $X \times Y$ is separable and complete, η is tight, so

$$1 = \eta(S) = \sup\{\eta(K)\colon\ K \subseteq S,\ K \text{ compact}\}$$

Choose $K_n \subseteq S$ compact such that $\eta(K_n) \geq 1 - 1/n$, for $n \geq 1$. Then $\pi^X(K_n)$ is compact in X and $\mu(\pi^X(K_n)) = \eta(K_n) \geq 1 - 1/n$. Hence $A := \bigcup_n K_n$ is a μ-full Borel set in X. If $x \in A$ then $x \in \pi^X(K_n)$ for some n, so $(x, y) \in K_n \subseteq S$ for some $y \in Y$. □

Theorem 3.6. *Let X and Y be separable complete metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \to [0, \infty)$ continuous. If $\eta \in \Gamma(\mu, \nu)$ is optimal for c and $\int c \,\mathrm{d}\eta < \infty$, then*

$$\operatorname{supp}\eta := \{z \in X \times Y\colon\ \eta(U) > 0 \textit{ for every neighborhood } U \textit{ of } z\}$$

is a c-monotone set.

Proof. (See [11, Theorem 2.3].) Suppose that $\operatorname{supp}\eta$ is not c-monotone. Then there are $n \in \mathbb{N}$ and a permutation σ of $\{1, \ldots, n\}$ such that the function

$$f(u_1, \ldots, u_n, v_1, \ldots, v_n) := \sum_{i=1}^{n} \Big(c(u_{\sigma(i)}, v_i) - c(u_i, v_i) \Big)$$

is strictly negative at some $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ with $(x_i, y_i) \in \operatorname{supp}\eta$. We will construct a more cost efficient measure than η and thus show that η is not optimal for c.

As f is continuous, we can choose Borel neighborhoods U_i of x_i and V_i of y_i such that $f(u_1, \ldots, u_n, v_1, \ldots, v_n) < 0$ for $u_i \in U_i$ and $v_i \in V_i$, $i = 1, \ldots, n$. As $(x_i, y_i) \in \operatorname{supp}\eta$,

$$\lambda := \min_i \eta(U_i \times V_i) > 0.$$

Define $\eta_i \in \mathcal{P}(X \times Y)$ by

$$\eta_i(W) := \frac{1}{\eta(U_i \times V_i)} \eta((U_i \times V_i) \cap W), \quad W \subseteq X \times Y \text{ Borel.}$$

Consider

$$Z = (X \times Y)^n$$

and $\rho \in \mathcal{P}(Z)$ given by

$$\rho = \eta_1 \otimes \cdots \otimes \eta_n.$$

Let $\pi_i^X\colon Z \to X$ be defined by $\pi_i^X(u_1, v_1, \ldots, u_n, v_n) := u_i$ and $\pi_i^Y\colon Z \to Y$ by $\pi_i^Y(u_1, v_1, \ldots, u_n, v_n) := v_i$. Recall that $\pi_i^X \otimes \pi_j^Y$ denotes the map $(u_1, v_1, \ldots, u_n, v_n) \mapsto (u_i, v_j)$. Define

$$\begin{aligned} \gamma &:= \eta - \frac{\lambda}{n} \sum_{i=1}^{n} (\pi_i^X \otimes \pi_i^Y)_{\#}\rho + \frac{\lambda}{n} \sum_{i=1}^{n} (\pi_{\sigma(i)}^X \otimes \pi_i^Y)_{\#}\rho \\ &= \eta - \frac{\lambda}{n} \sum_{i=1}^{n} \eta_i + \frac{\lambda}{n} \sum_{i=1}^{n} (\pi_{\sigma(i)}^X \otimes \pi_i^Y)_{\#}\rho. \end{aligned}$$

Then

$$\begin{aligned}\gamma(W) &\geq \eta(W) - \frac{\lambda}{n}\sum_{i=1}^{n}\eta_i(W)\\ &\geq \eta(W) - \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda}{\eta(U_i \times V_i)}\eta((U_i \times V_i)\cap W)\\ &\geq \eta(W) - \frac{1}{n}\sum_{i=1}^{n}\eta(W) = 0,\end{aligned}$$

for every Borel set $W \subseteq X \times Y$. So γ is a positive Borel measure. It is easy to check that $\gamma \in \mathcal{P}(X \times Y)$. Further, for $A \subseteq X$ Borel,

$$\begin{aligned}(\pi^X_{\sigma(i)} \otimes \pi^Y_i)_{\#}\rho(A \times Y) &= \rho(\{(u_1, v_1, \ldots, u_n, v_n) \in Z\colon\ (u_{\sigma(i)}, v_i) \in A \times Y\})\\ &= \rho(\{(u_1, v_1, \ldots, u_n, v_n) \in Z\colon\ u_{\sigma(i)} \in A\})\\ &= \eta_{\sigma(i)}(A \times Y),\end{aligned}$$

so

$$\begin{aligned}\gamma(A \times Y) &= \eta(A \times Y) - \frac{\lambda}{n}\sum_{i=1}^{n}\eta_i(A \times Y) + \frac{\lambda}{n}\sum_{i=1}^{n}(\pi^X_{\sigma(i)} \otimes \pi^Y_i)_{\#}\rho(A \times Y)\\ &= \mu(A) - \frac{\lambda}{n}\sum_{i=1}^{n}\eta_i(A \times Y) + \frac{\lambda}{n}\sum_{i=1}^{n}\eta_{\sigma(i)}(A \times Y) = \mu(A)\end{aligned}$$

and similarly $\gamma(X \times B) = \nu(B)$ for $B \subseteq Y$ Borel. Hence $\gamma \in \Gamma(\mu, \nu)$.

Finally,

$$\int_{X\times Y} c\,\mathrm{d}(\pi^X_i \otimes \pi^Y_j)_{\#}\rho = \int_Z c(\pi^X_i(z), \pi^Y_j(z))\,\mathrm{d}\rho,$$

so

$$\begin{aligned}\int c\,\mathrm{d}\gamma &= \int c\,\mathrm{d}\eta + \frac{\lambda}{n}\sum_{i=1}^{n}\int_Z \Big(c(\pi^X_{\sigma(i)}(z), \pi^Y_i(z)) - c(\pi^X_i(z), \pi^Y_i(z)\Big)\,\mathrm{d}\rho(z)\\ &= \int c\,\mathrm{d}\eta + \frac{\lambda}{n}\int_{U_1\times V_1\times\cdots\times U_n\times V_n} f(\pi^X_1(z), \ldots, \pi^X_n(z), \pi^Y_1(z), \ldots, \pi^Y_n(z))\,\mathrm{d}\rho(z)\\ &< \int c\,\mathrm{d}\eta,\end{aligned}$$

since ρ is concentrated on $U_1 \times V_1 \times \cdots \times U_n \times V_n$ and $f < 0$ on this set. Thus we have that γ is more cost efficient than η, so that η is not optimal. □

A converse to this theorem will be proved in the next section.

3.4 Kantorovich potentials

Consider again the Kantorovich problem. Throughout this section let (X, d_X) and (Y, d_Y) be separable complete metric spaces. Let $c\colon X \times Y \to [0, \infty]$ be a cost function and let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Imagine that a supplier has goods distributed at different positions according to the measure μ and that a demander wants the goods to be delivered at different positions given by the distribution ν. A transportation plan $\gamma \in \Gamma(\mu, \nu)$ says that an amount $\gamma(A \times B)$ of the goods located in the set A will be transported to positions in the set B. The total costs of this transportation plan will be $\int_{X \times Y} c(x,y)\, d\gamma(x,y)$. A transportation company wants the costs to be paid by the supplier and/or the demander. Suppose that the company charges $\varphi(x)$ per unit to the supplier to take goods from position x and $\psi(y)$ per unit to the demander to deliver the goods at position y. Of course the supplier and demander are not willing to pay more than the total costs needed for the shipping, so that

$$\varphi(x) + \psi(y) \leq c(x,y) \text{ for all } x \in X,\ y \in Y.$$

The optimazation problem for the transportation company is

$$\max\{\int_X \varphi(x)\, d\mu(x) + \int_Y \psi(y)\, d\nu(y) \colon\ \varphi(x) + \psi(y) \leq c(x,y) \text{ for all } x \in X,\ y \in Y\}.$$

This problem is sometimes called the *dual problem* to the Kantorovich problem. The optimal price functions φ and ψ turn out to play an important role in the analysis of optimal transportation plans, which will yield an answer to Monge's problem. We will first associate a certain function φ to a c-monotone set and then show that that the functions φ and φ^c associated to the support of an optimal transportation plan are optimal prices for the dual problem.

Definition 3.7. Let $c\colon X \times Y \to [0,\infty)$ be Borel measurable. Let $S \subseteq X \times Y$ be a non-empty c-monotone set. Fix $(x_0, y_0) \in S$. The function $\varphi\colon X \to [-\infty, \infty]$ defined by

$$\begin{aligned}\varphi(x) = \inf\{ \sum_{i=1}^{p} \Big(c(x_{i+1}, y) - c(x_i, y_i) \Big) \colon\ p \in \mathbb{N},\\ x_{p+1} = x,\ (x_i, y_i) \in S \text{ for } i = 1, \ldots, p\}\end{aligned}$$

is called the *Kantorovich potential* of S for c fixed at (x_0, y_0).

The relation between the cost functionals φ and ψ will be given by the following generalization of the Legendre-Fenchel transform.

Definition 3.8. For a function $\varphi\colon\ X \to [-\infty, \infty]$, the *c-transform* of φ is defined by

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

The c-transform of a function $\psi\colon\ Y \to [-\infty, \infty]$ is

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

Here we use the convention that $\inf \emptyset = \infty$, $\inf \infty = \infty$, $\inf(-\infty) = -\infty$, and the infimum of a set that is not bounded below is $-\infty$.

In the case $X = Y = \mathbb{R}^d$ and $c(x,y) = \langle x, y \rangle$ the function φ^c is Legendre-Fenchel transform of f or, if $d = 1$, the Legendre transform.

Proposition 3.9. *Let $c\colon X \times Y \to [0,\infty)$ be Borel measurable. Let $S \subseteq X \times Y$ be a non-empty c-monotone set. Fix $(x_0, y_0) \in S$. Let φ be the Kantorovich potential of S for c fixed at (x_0, y_0). Then*

(1) φ is Borel measurable

(2) $\varphi(x_0) = 0$

(3) $\varphi(x) > -\infty$ for all $x \in A := \{u \in X\colon \exists y \in Y \text{ with } (u, y) \in S\}$

(4) $\varphi(x) + \varphi^c(x) = c(x, y)$ for all $(x, y) \in S$

(5) $\varphi^{cc}(x) = \varphi(x)$ for all $x \in A$.

Proof. Define

$$\begin{aligned}\varphi_q(x) \quad := \quad & \inf\Big\{\sum_{i=0}^{p}\Big(c(x_{i+1}, y_i) - c(x_i, y_i)\Big) : \\ & \qquad x_{p+1} = x,\ (x_i, y_i) \in S,\ i = 1, \ldots, p,\ 1 \le p \le q\Big\}.\end{aligned}$$

Clearly $\varphi_q(x) \downarrow \varphi(x)$ for all $x \in X$.

We first show that φ_q is upper semicontinuous for each q. Suppose $u_k \to u$ in X. Let $\varepsilon > 0$. Then

$$\varphi_q(u) \ge c(x, y_p) - c(x_p, y_p) + \sum_{i=0}^{p-1}\Big(c(x_{i+1}, y_i) - c(x_i, y_i)\Big) - \varepsilon$$

for some $p \le q$ and $(x_i, y_i) \in S$. Then

$$\varphi_q(u_k) \le c(u_k, y_p) - c(x_p, y_p) + \sum_{i=0}^{p-1}\Big(c(x_{i+1}, y_i) - c(x_i, y_i)\Big),$$

so

$$\begin{aligned}\limsup_{k\to\infty} \varphi_q(u_k) \quad \le \quad & c(u, y_p) - c(x_p, y_p) + \sum_{i=0}^{p-1}\Big(c(x_{i+1}, y_i) - c(x_i, y_i)\Big) \\ \le \quad & \varphi_q(u) + \varepsilon.\end{aligned}$$

Hence φ_q is u.s.c.

(1): We know that φ_q is u.s.c. hence Borel and $\varphi_q \to \varphi$ pointwise, so φ is Borel measurable.

(2): On one hand, choose $(x_1, y_1) = (x_0, y_0) \in S$. Then $\varphi(x_0) \le c(x_0, y_1) - c(x_1, y_1) + c(x_1, y_0) - c(x_0, y_0) = 0$. On the other hand, as S is c-monotone, for $(x_i, y_i) \in S$, $i = 1, \ldots, p$,

$$\sum_{i=0}^{p} c(x_{\sigma(i)}, y_i) \ge \sum_{i=0}^{p} c(x_i, y_i),$$

in particular with the permutation $\sigma(i) = i + 1$ for $0 \le i \le p - 1$ and $\sigma(p) = 0$. So, with the notation $x_{p+1} = x_0$,

$$\sum_{i=0}^{p}\Big(c(x_{i+1}, y_i) - c(x_i, y_i)\Big) \ge 0,$$

so $\varphi(x_0) \geq 0$. Hence $\varphi(x_0) = 0$.

$(2\frac{1}{2})$: $\varphi(u) \leq \varphi(x) + c(u,y) - c(x,y)$ for all $u \in X$ and $(x,y) \in S$. Indeed, for any $p \in \mathbb{N}$ and $(x_i, y_i) \in S$, $i = 1, \ldots, p$, we have

$$\begin{aligned} \varphi(u) &\leq \varphi_{p+1}(u) \\ &\leq c(u,y) - c(x,y) + \sum_{i=0}^{p} \Big(c(x_{i+1}, y_i) - c(x_i, y_i) \Big), \end{aligned}$$

where $x_{p+1} = x$. So, by taking infimum over $\{(x_i, y_i) \colon 0 \leq i \leq p\}$,

$$\varphi(u) \leq c(u,y) - c(x,y) + \varphi(x).$$

(3): If $(x,y) \in S$, then by $(2\frac{1}{2})$,

$$\begin{aligned} \varphi(x) &\geq \varphi(x_0) - c(x_0, y) + c(x,y) \\ &= c(x,y) - c(x_0,y) \in \mathbb{R}. \end{aligned}$$

(4): Let $\psi := \varphi^c$. Then $\varphi(x) + \psi(y) = c(x,y)$ for all $(x,y) \in S$. Indeed, by definition,

$$\psi(y) = \inf_{u \in X} \Big(c(u,y) - \varphi(u) \Big).$$

We have by $(2\frac{1}{2})$ that $c(u,y) - \varphi(u) \geq c(x,y) - \varphi(x)$, so

$$\psi(y) \geq c(x,y) - \varphi(x).$$

From the definition of ψ we find with $u = x$ also $\psi(y) \leq c(x,y) - \varphi(x)$.

(5): Let $x \in A$. We have

$$\varphi^{cc}(x) = \inf_{y \in Y} \Big(c(x,y) - \varphi^c(y) \Big),$$
$$\varphi^c(y) = \inf_{u \in X} \Big(c(u,y) - \varphi(u) \Big).$$

Let $y \in Y$. Then $c(x,y) - \varphi^c(y) \geq c(x,y) - \Big(c(u,y) - \varphi(u) \Big)$ for all $u \in X$, so (with $u = x$) $c(x,y) - \varphi^c(y) \geq \varphi(x)$. Hence $\varphi^{cc}(x) \geq \varphi(x)$. Conversely, since $x \in A$, there exists a $y \in Y$ such that $\varphi^c(y) = c(x,y) - \varphi(x)$. Then $\varphi^{cc}(x) \leq c(x,y) - \varphi^c(y) = c(x,y) - \Big(c(x,y) - \varphi(x) \Big) = \varphi(x)$. □

We are now ready to prove a characterization of optimal transportation plans in terms of c-monotonicity of their supports.

Theorem 3.10. *Let X and Y be separable complete metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c \colon X \times Y \to [0, \infty)$ continuous.*

(1) *If $\eta \in \Gamma(\mu, \nu)$ is optimal for c and $\int c \, \mathrm{d}\eta < \infty$, then*

$$\operatorname{supp} \eta := \{ z \in X \times Y \colon \eta(U) > 0 \textit{ for every neighborhood } U \textit{ of } z \}$$

is a c-monotone set.

(2) *If $\eta \in \Gamma(\mu,\nu)$ is such that*

- $\operatorname{supp}\eta$ *is c-monotone, and*
- $\mu\Big(\{x \in X\colon\ \int_Y c(x,y)\,\mathrm{d}\nu(y) < \infty\}\Big) > 0$, *and*
- $\nu\Big(\{y \in Y\colon\ \int_X c(x,y)\,\mathrm{d}\mu(x) < \infty\}\Big) > 0$,

then η is optimal for c.

(3) *In the situation of (2), one also has*

$$\begin{aligned}&\min\Big\{\int c\,\mathrm{d}\gamma\colon\ \gamma \in \Gamma(\mu,\nu)\Big\}\\ &\quad= \max\Big\{\int \varphi\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu\colon\ \varphi \in L^1(\mu),\ \ \psi \in L^1(\nu),\\ &\qquad\qquad \varphi(x)+\psi(y) \le c(x,y)\ \forall (x,y) \in X \times Y\Big\}\end{aligned}$$

and the maximum at the right hand side is attained at the Kantorovich potential φ of $\operatorname{supp}\eta$ *for c (fixed at some point (x_0,y_0)) and $\psi = \varphi^c$.*

Proof. (1): has been proved in the previous section.

(See [11, Theorem 2.3].) Suppose that $\operatorname{supp}\eta$ is not c-monotone. Then there are $n \in \mathbb{N}$ and a permutation σ of $\{1,\dots,n\}$ such that the function

$$f(u_1,\dots,u_n,v_1,\dots,v_n) := \sum_{i=1}^n \Big(c(u_{\sigma(i)},v_i) - c(u_i,v_i)\Big)$$

is strictly negative at some $(x_1,\dots,x_n,y_1,\dots,y_n)$ with $(x_i,y_i) \in \operatorname{supp}\eta$. We will construct a more cost efficient measure than η and thus show that η is not optimal for c.

As f is continuous, we can choose Borel neighborhoods U_i of x_i and V_i of y_i such that $f(u_1,\dots,u_n,v_1,\dots,v_n) < 0$ for $u_i \in U_i$ and $v_i \in V_i$, $i = 1,\dots,n$. As $(x_i,y_i) \in \operatorname{supp}\eta$,

$$\lambda := \min_i \eta(U_i \times V_i) > 0.$$

Define $\eta_i \in \mathcal{P}(X \times Y)$ by

$$\eta_i(W) := \frac{1}{\eta(U_i \times V_i)}\eta((U_i \times V_i) \cap W), \quad W \subseteq X \times Y \text{ Borel.}$$

Consider

$$Z = (X \times Y)^n$$

and $\rho \in \mathcal{P}(Z)$ given by

$$\rho = \eta_1 \otimes \cdots \otimes \eta_n.$$

Let $\pi_i^X\colon\ Z \to X$ be defined by $\pi_i^X(u_1,v_1,\dots,u_n,v_n) := u_i$ and $\pi_i^Y\colon\ Z \to Y$ by $\pi_i^Y(u_1,v_1,\dots,u_n,v_n) := v_i$. Recall that $\pi_i^X \otimes \pi_j^Y$ denotes the map $(u_1,v_1,\dots,u_n,v_n) \mapsto (u_i,v_j)$. Define

$$\begin{aligned}\gamma &:= \eta - \frac{\lambda}{n}\sum_{i=1}^n (\pi_i^X \otimes \pi_i^Y)_{\#}\rho + \frac{\lambda}{n}\sum_{i=1}^n (\pi_{\sigma(i)}^X \otimes \pi_i^Y)_{\#}\rho\\ &= \eta - \frac{\lambda}{n}\sum_{i=1}^n \eta_i + \frac{\lambda}{n}\sum_{i=1}^n (\pi_{\sigma(i)}^X \otimes \pi_i^Y)_{\#}\rho.\end{aligned}$$

Then

$$\begin{aligned}\gamma(W) &\geq \eta(W) - \frac{\lambda}{n}\sum_{i=1}^{n}\eta_i(W)\\ &\geq \eta(W) - \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda}{\eta(U_i\times V_i)}\eta((U_i\times V_i)\cap W)\\ &\geq \eta(W) - \frac{1}{n}\sum_{i=1}^{n}\eta(W) = 0,\end{aligned}$$

for every Borel set $W \subseteq X \times Y$. So γ is a positive Borel measure. It is easy to check that $\gamma \in \mathcal{P}(X \times Y)$. Further, for $A \subseteq X$ Borel,

$$\begin{aligned}(\pi^X_{\sigma(i)} \otimes \pi^Y_i)_{\#}\rho(A\times Y) &= \rho(\{(u_1, v_1, \ldots, u_n, v_n) \in Z\colon\ (u_{\sigma(i)}, v_i) \in A \times Y\})\\ &= \rho(\{(u_1, v_1, \ldots, u_n, v_n) \in Z\colon\ u_{\sigma(i)} \in A\})\\ &= \eta_{\sigma(i)}(A\times Y),\end{aligned}$$

so

$$\begin{aligned}\gamma(A\times Y) &= \eta(A\times Y) - \frac{\lambda}{n}\sum_{i=1}^{n}\eta_i(A\times Y) + \frac{\lambda}{n}\sum_{i=1}^{n}(\pi^X_{\sigma(i)}\otimes\pi^Y_i)_{\#}\rho(A\times Y)\\ &= \mu(A) - \frac{\lambda}{n}\sum_{i=1}^{n}\eta_i(A\times Y) + \frac{\lambda}{n}\sum_{i=1}^{n}\eta_{\sigma(i)}(A\times Y) = \mu(A)\end{aligned}$$

and similarly $\gamma(X \times B) = \nu(B)$ for $B \subseteq Y$ Borel. Hence $\gamma \in \Gamma(\mu, \nu)$.

Finally,

$$\int_{X\times Y} c\,\mathrm{d}(\pi^X_i \otimes \pi^Y_j)_{\#}\rho = \int_Z c(\pi^X_i(z), \pi^Y_j(z))\,\mathrm{d}\rho,$$

so

$$\begin{aligned}\int c\,\mathrm{d}\gamma &= \int c\,\mathrm{d}\eta + \frac{\lambda}{n}\sum_{i=1}^{n}\int_Z\Big(c(\pi^X_{\sigma(i)}(z), \pi^Y_i(z)) - c(\pi^X_i(z), \pi^Y_i(z)\Big)\,\mathrm{d}\rho(z)\\ &= \int c\,\mathrm{d}\eta + \frac{\lambda}{n}\int_{U_1\times V_1\times\cdots\times U_n\times V_n} f(\pi^X_1(z), \ldots, \pi^X_n(z), \pi^Y_1(z), \ldots, \pi^Y_n(z))\,\mathrm{d}\rho(z)\\ &< \int c\,\mathrm{d}\eta,\end{aligned}$$

since ρ is concentrated on $U_1 \times V_1 \times \cdots \times U_n \times V_n$ and $f < 0$ on this set. Thus we have that γ is more cost efficient than η, so that η is not optimal.

(2) and (3): Let $S := \operatorname{supp}\eta$, which is a c-monotone subset of $X \times Y$. Fix $(x_0, y_0) \in S$ ($\eta(S) = 1$ so S is non-empty) and let φ be the Kantorovich potential of S for c fixed at (x_0, y_0). The proof is divided into several claims, clustered by topic. We first establish some properties of φ, then of $\psi = \varphi^c$, and then we show that φ and ψ are L^1 functions. Then we derive some more connections between φ, ψ and η, and finally we conclude the proof.

Due to Proposition 3.9 we know that (A1) φ is Borel measurable, (A2) $\varphi(x_0) = 0$, (A3) $\varphi(x) > -\infty$ for all x in

$$A := \{u \in X\colon\ \exists y \in Y \text{ with } (u, y) \in S\},$$

(B1) $\varphi(x)+\varphi^c(x)=c(x,y)$ for all $(x,y)\in S$, and $\varphi^{cc}(x)=\varphi(x)$ for all $x\in A$.

Since $\mu(A)=\eta(S)=1$, it follows that $\varphi(x)>-\infty$ and therefore $\phi(x)\in\mathbb{R}$ for μ-a.e. $x\in X$. Then also $\psi(y)\in\mathbb{R}$ for ν-a.e. $y\in Y$, by (B1).

Claim B2: ψ is ν-measurable. Due to Claim B1,

$$\psi(y)1_S(x,y)=\Big(c(x,y)-\varphi(x)\Big)1_S(x,y)\quad\text{for all }(x,y)\in X\times Y$$

and $(x,y)\mapsto c(x,y)-\varphi(x)$ is a Borel map by A1. Hence $(x,y)\mapsto\psi(y)1_S(x,y)$ is η-measurable. By disintegration, there exist $\eta_y\in\mathcal{P}(X)$, $y\in Y$, such that $y\mapsto\int_X f(x,y)\,\mathrm{d}\eta_y(x)$ is ν-measurable and

$$\int_{X\times Y}f(x,y)\,\mathrm{d}\eta(x,y)=\int_Y\left(\int_X f(x,y)\,\mathrm{d}\eta_y(x)\right)\,\mathrm{d}\nu(y)$$

for every Borel function $f\colon X\times Y\to[0,\infty]$. The set $S=\operatorname{supp}\eta$ is closed and therefore a Borel set. From the disintegration formula with $f=1_S$ we obtain that

$$\int_X 1_S(x,y)\,\mathrm{d}\eta_y(x)=1\quad\text{for }\nu\text{-almost every }y.$$

If we apply now the disintegration to $f(x,y)=\Big(c(x,y)-\varphi(x)\Big)^+1_B(y)$ for some Borel set $B\subseteq Y$, then

$$\begin{aligned}\int_{X\times B}\psi^+(y)1_S(x,y)\,\mathrm{d}\eta(x,y) &= \int_{X\times B}\Big(c(x,y)-\varphi(x)\Big)^+\,\mathrm{d}\eta(x,y)\\ &= \int_B\left(\int_X\Big(c(x,y)-\varphi(x)\Big)^+\,\mathrm{d}\eta_y(x)\right)\,\mathrm{d}\nu(y)\end{aligned}$$

and

$$\begin{aligned}\int_{X\times B}\psi^+(y)1_S(x,y)\,\mathrm{d}\eta(x,y) &= \int_B\left(\int_X\psi^+(y)1_S(x,y)\,\mathrm{d}\eta_y(x)\right)\,\mathrm{d}\nu(y)\\ &= \int_B\psi^+(y)\left(\int_X 1_S(x,y)\,\mathrm{d}\eta_y(x)\right)\,\mathrm{d}\nu(y).\end{aligned}$$

It follows that

$$\psi^+(y)\left(\int_X 1_S(x,y)\,\mathrm{d}\eta_y(x)\right)=\int_X\Big(c(x,y)-\varphi(x)\Big)^+\,\mathrm{d}\eta_y(x)\quad\text{for }\nu\text{-a.e. }y\in Y,$$

so

$$\psi^+(y)=\int_X\Big(c(x,y)-\varphi(x)\Big)^+\,\mathrm{d}\eta_y(x)\quad\text{for }\nu\text{-a.e. }y\in Y.$$

Hence ψ^+ is ν-measurable. Similarly, ψ^- is ν-measurable and thus ψ is ν-measurable.

Claim C1: $\psi^+(y)\le c(x,y)+\varphi^-(x)$ for all $(x,y)\in S$. As $c\ge0$, we have $c(x,y)+\varphi^-(x)\ge0$. Also $c(x,y)+\varphi^-(x)\ge c(x,y)-\varphi(x)=\psi(y)$. Hence $c(x,y)+\varphi^-(x)\ge\psi^+(y)$.

Claim C2: $\varphi^+\in L^1(\mu)$ and $\psi^+\in L^1(\nu)$. By assumption, $\mu(X_1)>0$, where

$$X_1:=\Big\{x\in X\colon\ \int_Y c(x,y)\,\mathrm{d}\nu(y)<\infty\Big\}.$$

Choose $x \in X_1$ such that $\nu(\{y\colon (x,y) \in S\}) = 1$. Then $\psi^+ \le c(x,\cdot) + \varphi^-(x)$ ν-a.e. on Y (by Claim C1), so

$$\int_Y \psi^+ \,\mathrm{d}\nu \le \int_Y \Big(c(x,y) + \varphi^-(x)\Big) \,\mathrm{d}\nu(y) < \infty,$$

since $x \in X_1$ and $\varphi^-(x) \in \mathbb{R}$ (by Claim A4). Similarly, $\varphi^+(x) \le c(x,y) + \psi^-(y)$ for $(x,y) \in S$ and there exists a y s.t. $\mu(\{x\colon (x,y) \in S\}) = 1$, $\psi^-(y) \in \mathbb{R}$ (Claim B1bis), and $\int_X c(x,y)\,\mathrm{d}\mu(x) < \infty$, so

$$\int_X \varphi^+ \,\mathrm{d}\nu \le \int_X c(x,y)\,\mathrm{d}\mu(x) + \int_X \psi^-(y)\,\mathrm{d}\mu(x) < \infty.$$

Claim C3: $\int_{X\times Y} c(x,y)\,\mathrm{d}\eta < \infty$. We have

$$\begin{aligned} \int_{X\times Y} c\,\mathrm{d}\eta &= \int \Big(\varphi(x) + \psi(y)\Big)\,\mathrm{d}\eta(x,y) \\ &= \int \varphi\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu \\ &\le \int \varphi^+\,\mathrm{d}\mu + \int \psi^+\,\mathrm{d}\nu < \infty. \end{aligned}$$

Claim C4: $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$. For $(x,y) \in S$ we have

$$\begin{aligned} \varphi(x) &= c(x,y) - \psi(y) \ge c(x,y) - \psi^+(y) \\ &\ge -c(x,y) - \psi^+(y), \end{aligned}$$

so $\varphi^-(x) \le c(x,y) + \psi^+(y)$. Hence

$$\begin{aligned} \int \varphi^-\,\mathrm{d}\mu &= \int \varphi^-(x)\,\mathrm{d}\eta(x,y) \le \int \Big(c(x,y) + \psi^+(y)\Big)\,\mathrm{d}\eta(x,y) \\ &= \int c\,\mathrm{d}\eta + \int \psi^+\,\mathrm{d}\nu < \infty. \end{aligned}$$

So $\int |\varphi|\,\mathrm{d}\mu \le \int \varphi^+\,\mathrm{d}\mu + \int \varphi^-\,\mathrm{d}\mu < \infty$. Similarly, $\int |\psi|\,\mathrm{d}\nu < \infty$.

Claim D1: $\varphi(x) + \psi(y) \le c(x,y)$ for all $(x,y) \in X \times Y$. We have

$$\psi(y) = \inf_{u\in X} \Big(c(u,y) - \varphi(u)\Big) \le c(x,y) - \varphi(x).$$

Claim D2: For $\gamma \in \Gamma(\mu,\nu)$,

$$\begin{aligned} \int_{X\times Y} c\,\mathrm{d}\gamma &\ge \int_{X\times Y} \Big(\varphi(x) + \psi(y)\Big)\,\mathrm{d}\gamma(x,y) \\ &= \int \varphi(x)\,\mathrm{d}\mu(x) + \int \psi(y)\,\mathrm{d}\nu(y) \\ &= \int_{X\times Y} \Big(\varphi(x) + \psi(y)\Big)\,\mathrm{d}\eta(x,y) \\ &= \int_S \Big(\varphi(x) + \psi(y)\Big)\,\mathrm{d}\eta(x,y) \\ &= \int_S c\,\mathrm{d}\eta = \int_{X\times Y} c\,\mathrm{d}\eta. \end{aligned}$$

Conclusion: From D2 we see that η is optimal for c, that is,

$$\min\Big\{\int_{X\times Y} c\,\mathrm{d}\gamma\colon\ \gamma\in\Gamma(\mu,\nu)\Big\}=\int_{X\times Y} c\,\mathrm{d}\eta.$$

Further,

$$\begin{aligned}&\max\Big\{\int f\,\mathrm{d}\mu+\int g\,\mathrm{d}\nu\colon\ f\in L^1(\mu),\ g\in L^1(\nu),\ f(x)+g(y)\le c(x,y)\ \forall(x,y)\in X\times Y\Big\}\\&\quad=\int\varphi(x)\,\mathrm{d}\mu(x)+\int\psi(y)\,\mathrm{d}\nu(y)=\int c\,\mathrm{d}\eta.\end{aligned}$$

Finally, we have $\psi=\varphi^c$ by definition of ψ. □

Remark. The conditions in the previous theorem that

- $\int c\,\mathrm{d}\eta<\infty$,
- $\mu\Big(\{x\in X\colon\ \int_Y c(x,y)\,\mathrm{d}\nu(y)<\infty\}\Big)>0$, and
- $\nu\Big(\{y\in Y\colon\ \int_X c(x,y)\,\mathrm{d}\mu(x)<\infty\}\Big)>0$,

are implied by the stronger condition that $\int c\,\mathrm{d}\mu\otimes\nu<\infty$, as is easily seen with the aid of Fubini.

Let us next consider a special case and show that there the Kantorovich potential inherits some regularity of the cost function. Recall that a map h from a metric space A into another metric space B is called *locally Lipschitz* if for every $x_0\in X$ and every $r>0$ there exists an $L\in\mathbb{R}$ such that

$$d(h(x),h(x_0))\le Ld(x,x_0)$$

for all x in the open ball $B(x_0,r)$ around x_0 with radius r. This means that the restriction of h to each ball is Lipschitz. For example, the function $t\mapsto t^2$ from $\mathbb{R}$ to $\mathbb{R}$ is locally Lipschitz but not Lipschitz.

The special situation that we are mostly interested in is the case $X=Y=\mathbb{R}^d$ and $c(x,y)=\|x-y\|^2$.

Proposition 3.11. *Consider $X=Y=\mathbb{R}^d$. Let $c(x,y)=h(x-y)$, $x,y\in\mathbb{R}^d$, where $h\colon\mathbb{R}^d\to[0,\infty)$ is differentiable, locally Lipschitz, and such that ∇h from $\mathbb{R}^d$ to its range is bijective with a Borel measurable inverse. Let $S\subseteq X\times Y$ be a non-empty c-monotone set and assume that*

$$A=\{x\in X\colon\ \exists y\in Y\ \text{with}\ (x,y)\in S\}$$

is bounded. Fix $(x_0,y_0)\in S$. The Kantorovich potential φ of S for c fixed at (x_0,y_0) is locally Lipschitz on A and φ^{cc} is locally Lipschitz on X.

Proof. Take $R>0$ such that $A\subset B(0,R)$. Let $r>0$. We show that φ^{cc} is Lipschitz on $B(0,r)$. Since h is locally Lipschitz, h is L-Lipschitz on $B(0,R+r)$ for some L. Let

$x_1, x_2 \in B(0,r)$. Let $\varepsilon > 0$ and choose $y \in B(0,R)$ such that $\varphi^c(y) > -\infty$ (use (3) of Proposition 3.9). Then

$$\begin{aligned}
\varphi^{cc}(x_1) - \varphi^{cc}(x_2) &\leq c(x_1,y) - \varphi^c(y) - \Big(c(x_2,y) - \varphi^c(y) - \varepsilon\Big) \\
&= h(x_1 - y) - h(x_2 - y) + \varepsilon \\
&\leq L\|x_1 - x_2\| + \varepsilon,
\end{aligned}$$

as $x_i + y \in B(0, r+R)$, and hence

$$\varphi^{cc}(x_1) - \varphi^{cc}(x_2) \leq L\|x_1 - x_2\|.$$

Thus, by interchanging the role of x_1 and x_2, $|\varphi^{cc}(x_1) - \varphi^{cc}(x_2)| \leq L\|x_1 - x_2\|$. Hence φ^{cc} is locally Lipschitz.

By (5) of Proposition 3.9 it follows that φ is locally Lipschitz on A. □

3.5 Existence and uniqueness for the Monge problem

We will now address the questions

- When is the optimal transportation plan η of the Kantorovich problem unique?
- When does the optimal η solve the Monge problem, that is, $\eta = (i \otimes r)_{\#}\mu$ for some Borel map $r : X \to Y$? (Recall, $i(x) = x$ for all $x \in X$.)

We begin with a sloppy sketch of the argument and then prove a version of a theorem on uniqueness and the Monge problem and mention a more general theorem.

We will consider the case that $X = Y = \mathbb{R}^d$ and $c(x,y) = h(x-y)$ for some strictly convex function h. In particular, we have $h(x) =$
$x\|^2$ in mind, where $\|\cdot\|$ is the usual Euclidean norm of $\mathbb{R}^d$.

Suppose $\eta \in \Gamma(\mu,\nu)$ is optimal for c and that $c(x,y) = h(x-y)$ with h differentiable. We want to find a Borel map $r\colon X \to Y$ such that $\eta = (i \otimes r)_{\#}\mu$, that is, $\eta(W) = \mu(\{x\colon (x, r(x)) \in W\})$ for Borel sets $W \subseteq X \times Y$. In other words, we want to show that η is concentrated on the graph of a Borel map. We will try to find for each $x \in X$ a unique point y with (x,y) in the support of η.

Let φ be a Kantorovich potential associated to $\operatorname{supp}\eta$. For $(x,y) \in \operatorname{supp}\eta$ we have $\varphi(x) + \varphi^c(y) = h(x-y)$. Since $\varphi^c(y) = \inf_{u \in X}\Big(h(u-y) - \varphi(u)\Big) = h(x-y) - \varphi(x)$, the function $u \mapsto h(u-y) - \varphi(u)$ attains its minimum at $u = x$. Hence, *if φ is differentiable at* x,

$$\nabla h(x-y) = \nabla\varphi(x).$$

If $u \mapsto \nabla h(u)$ is invertible, we obtain $x - y = (\nabla h)^{-1}(\nabla\varphi(x))$, so

$$y = x - (\nabla h)^{-1}(\nabla\varphi(x)).$$

Hence for x such that φ is differentiable at x there is exactly one y with $(x,y) \in \operatorname{supp}\eta$. Thus we can take

$$r(x) := x - (\nabla h)^{-1}(\nabla\varphi(x)).$$

The main mathematical problems to make the argument work are the differentiability of φ and the Borel measurability of r. We will not be able to obtain everywhere differentiability of φ. Instead we will impose conditions that yield that φ is locally Lipschitz and then use Rademacher's theorem to conclude its Lebesgue almost everywhere differentiability. We need the map r at least μ-a.e. defined and therefore require that μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.

Recall that a map $f : \mathbb{R}^d \to \mathbb{R}^m$ is called *differentiable* at $x \in \mathbb{R}^d$ if there exists a linear operator $L_x \colon \mathbb{R}^d \to \mathbb{R}^m$ such that for every $\varepsilon > 0$ there is a $\delta > 0$ with

$$\left| \frac{f(x+u) - f(u) - L_x u}{\|u\|} \right| < \varepsilon \quad \text{for all } u \in \mathbb{R}^d \text{ with } 0 < \|u\| < \delta.$$

If $m = 1$, then L_x is represented by a vector, which is denoted by $\nabla f(x)$, that is, $L_x u = \langle \nabla f(x), u \rangle$.

Denote the Lebesgue measure on $\mathbb{R}^d$ by $\mathcal{L}^d$.

Theorem 3.12 (Rademacher)**.** *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz. Then f is differentiable $\mathcal{L}^d$-almost everywhere. Moreover, $D = \{x \in \mathbb{R}^d \colon f \text{ differentiable at } x\}$ is a Borel set and*

$$x \mapsto \begin{cases} \nabla f(x) & \text{if } x \in D \\ 0 & \text{otherwise} \end{cases}$$

is a Borel map from $\mathbb{R}^d$ to $\mathbb{R}^d$.

Now we are in a position to prove a theorem on uniqueness for the Kantorovich problem and existence for the Monge problem. More sophisticated statements are given in Theorem 3.14.

Theorem 3.13. *Consider $X = Y = \mathbb{R}^d$. Let $c(x, y) = h(x - y)$, $x, y \in \mathbb{R}^d$, where $h \colon \mathbb{R} \to [0, \infty)$ is differentiable, locally Lipschitz, and such that ∇h from $\mathbb{R}^d$ to its range is bijective with a Borel measurable inverse. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be such that*

- $\int_{X \times Y} h(x - y) \,\mathrm{d}\gamma(x, y) < \infty$ *for some* $\gamma \in \Gamma(\mu, \nu)$,
- $\mu\Big(\Big\{x \in X \colon \int_Y h(x - y) \,\mathrm{d}\nu(y) < \infty\Big\}\Big) > 0$,
- $\nu\Big(\Big\{y \in Y \colon \int_X h(x - y) \,\mathrm{d}\mu(x) < \infty\Big\}\Big) > 0$,

and such that

- *μ is absolutely continuous with respect to $\mathcal{L}^d$,*
- $\operatorname{supp} \nu$ *is bounded.*

Then:

(1) *there is a* unique *$\eta \in \Gamma(\mu, \nu)$ that is optimal for c;*

(2) *η is induced by an optimal transport map, that is, there exists a Borel map $r \colon \mathbb{R}^d \to \mathbb{R}^d$ such that $\eta = (i \otimes r)_{\#}\mu$;*

(3) *the map r of (2) satisfies*

$$r(x) = x - (\nabla h)^{-1}(\nabla \varphi^{cc}(x)) \quad \text{for } \mu\text{-a.e. } x \in \mathbb{R}^d,$$

where φ is a Kantorovich potential associated to $\operatorname{supp} \eta$.

Proof. Let $A_1 \subseteq X$ be a μ-full Borel set such that for all $x \in A_1$ there is a $y \in Y$ such that $(x, y) \in \operatorname{supp} \eta$, $\varphi^{cc}(x) = \varphi(x)$ (use Proposition 3.9), and $\varphi(x) \in \mathbb{R}$.

Take $R > 0$ such that $\operatorname{supp} \nu \subset B(0, R)$. Then for $x \in A_1$,

$$\varphi(x) = \inf_{y \in B(0,R)} \Big(c(x, y) - \varphi^c(x) \Big),$$

since

$$\begin{aligned} \varphi(x) &= \varphi^{cc}(x) = \inf_{y \in \mathbb{R}^d} \Big(c(x, y) - \varphi^c(y) \Big) \\ &\leq \inf_{y \in B(0,R)} \Big(c(x, y) - \varphi^c(y) \Big) = \varphi(x), \end{aligned}$$

where the latter equality follows from Proposition 3.9(4).

Due to Proposition 3.11, φ^{cc} is locally Lipschitz.

Let A_2 be an $\mathcal{L}^d$-full Borel set such that φ^{cc} is differentiable at every $x \in A_2$ (by Rademacher's theorem). Then A_2 is also μ-full, as μ is absolutely continuous with rspect to $\mathcal{L}^d$. Let $A := A_1 \cap A_2$. Then A is a μ-full Borel set and for every $x \in A$ we have

- there is a $y \in \mathbb{R}^d$ with $(x, y) \in \operatorname{supp} \eta$ and therefore $y \in B(0, R)$ and $\varphi(x) + \varphi^c(y) = h(x - y)$,
- $\varphi(x) = \varphi^{cc}(x)$,
- $\varphi(x) \in \mathbb{R}$ and φ^{cc} is differentiable at x.

Let $x \in A$. There exists y such that $(x, y) \in \operatorname{supp} \eta$. Consider such a y. The function $u \mapsto h(u - y) - \varphi^{cc}(u)$ then attains its minimum $\varphi^c(y)$ at $u = x$ and is differentiable at x. So

$$\nabla h(x - y) - \nabla \varphi^{cc}(x) = 0.$$

Hence $\nabla \varphi(x)$ is in the range of ∇h and $x - y = (\nabla h)^{-1}(\nabla \varphi^{cc}(x))$, so

$$y = x - (\nabla h)^{-1}(\nabla \varphi^{cc}(x)). \tag{3}$$

Define

$$r(x) := \begin{cases} x - (\nabla h)^{-1}(\nabla \varphi^{cc}(x)) & x \in A \\ 0 & x \notin A. \end{cases}$$

Due to Rademacher's theorem and the assumptions on h, we infer that $r\colon \mathbb{R}^d \to \mathbb{R}^d$ is a Borel map. Moreover, we have $(x, r(x)) \in \operatorname{supp} \eta$ for all $x \in A$, as follows from (3). Further, in the arguments preceding (3) y is an arbitrary element of $\mathbb{R}^d$ with $(x, y) \in \operatorname{supp} \eta$ and thus we obtain that for $x \in A$,

$$(x, y) \in \operatorname{supp} \eta \quad \Longleftrightarrow y = r(x).$$

Consequently,

$$\eta(\{x \in A\colon\ (x, r(x))\}) = \eta(\operatorname{supp} \eta \setminus (A \times Y)) = 1.$$

Next we claim that

$$\eta = (i \otimes r)_{\#} \mu.$$

For a proof, let $U \times V \subseteq \mathbb{R}^d \times \mathbb{R}^d$ with $U \subseteq \mathbb{R}^d$ and $V \subseteq \mathbb{R}^d$ Borel. Then

$$\begin{aligned}
\eta(U \times V) &= \eta\Big(U \times V \cap \{(x, r(x))\colon\ x \in X\}\Big) \\
&= \eta\Big((U \cap \{x\colon\ r(x) \in V\}) \times Y\Big) \\
&= \mu(U \cap \{x\colon\ (x, r(x)) \in U \times V\}\Big) \\
&= (i \otimes r)_{\#}\mu(U \times V).
\end{aligned}$$

.

Finally, we address uniqueness of η. Suppose $\eta_1, \eta_2 \in \Gamma(\mu,\nu)$ are both optimal for c. Then also $\eta := \frac{1}{2}\eta_1 + \frac{1}{2}\eta_2 \in \Gamma(\mu,\nu)$ is optimal for c. By the first part of the proof given above, we obtain Borel maps $r_1, r\colon\ \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\eta_1 = (i \otimes r_1)_{\#}\mu \quad \text{and} \quad \eta = (i \otimes r)_{\#}\mu.$$

As η_1 is absolutely continuous with respect to η, we have

$$\eta_1\Big(\Big\{(x, r(x))\colon\ x \in \mathbb{R}^d\Big\}\Big) = 1.$$

Then

$$\eta_1\Big(\Big\{(x, r(x))\colon\ x \in \mathbb{R}^d\Big\} \cap \Big\{(x, r_1(x))\colon\ x \in \mathbb{R}^d\Big\}\Big) = 1,$$

so

$$\eta_1\Big(\Big\{(x, r(x))\colon\ x \in \mathbb{R}^d, r(x) = r_1(x)\Big\}\Big) = 1.$$

Hence $r = r_1$ μ-a.e. and, consequently, $\eta_1 = (i \otimes r_1)_{\#}\mu = (i \otimes r)_{\#}\mu = \eta$. Therefore $\eta_1 = \eta_2$. □

The conditions in the previous theorem can be relaxed. In particular the condition that the support of ν be bounded and the differentiability of h. An interesting setting is where h is strictly convex. The extension of the result requires more sophisticated knowledge on almost everywhere differentiability, which we will not discuss here. We only mention the result.

Theorem 3.14. *Consider $X = Y = \mathbb{R}^d$. Let $c(x,y) = h(x-y)$, $x, y \in \mathbb{R}^d$, where $h\colon\ \mathbb{R}^d \to [0,\infty)$ is strictly convex. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be such that*

- $\int_{X\times Y} c(x,y)\,\mathrm{d}\gamma(x,y) < \infty$ *for some* $\gamma \in \Gamma(\mu,\nu)$,
- $\mu\Big(\Big\{x \in X\colon\ \int_Y c(x,y)\,\mathrm{d}\nu(y) < \infty\Big\}\Big) > 0$,
- $\nu\Big(\Big\{y \in Y\colon\ \int_X c(x,y)\,\mathrm{d}\mu(x) < \infty\Big\}\Big) > 0$,

and such that

- *μ is absolutely continuous with respect to $\mathcal{L}^d$.*

Then:

(1) *there is a* unique *$\eta \in \Gamma(\mu,\nu)$ that is optimal for c;*

(2) *η is induced by an optimal transport map, that is, there exists a Borel map $r\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $\eta = (i \otimes r)_{\#}\mu$;*

(3) *the map r of (2) satisfies*

$$r(x) = x - (\partial h)^{-1}(\tilde{\nabla}\varphi(x)) \quad \textit{for } \mu\textit{-a.e. } x \in \mathbb{R}^d,$$

where φ is a Kantorovich potential associated to $\operatorname{supp}\eta$.

4 Gradient flows in Hilbert spaces

Let $(H, \langle , \rangle)$ be a Hilbert space over $\mathbb{R}$ and consider a function $\varphi\colon H \to \mathbb{R}$. We view φ as a potential and we are interested in flows in H that stream in the direction of steepest descent of φ. More specifically we will study the differential equation

$$y'(t) = -\nabla\varphi(y(t)), \quad t \geq 0.$$

If there exists for each $x \in H$ a differentiable map $y\colon [0,\infty) \to H$ such that $y_x'(t) = -\nabla\varphi(y_x(t))$, then the map $(t,x) \mapsto y_x(t)$ is called a *gradient flow* in H for the potential φ. Let us first define the notion of differentiability that we will consider.

Definition 4.1. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces, let U be an open subset of X, let $f\colon U \to Y$, and let $x \in X$. Then f is called *(Fréchet) differentiable at x* if there exists a bounded linear map $L_x\colon X \to Y$ such that

$$\lim_{h\to 0} \frac{\|f(x+h) - f(x) - L_x h\|_Y}{\|h\|_X} = 0,$$

that is, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $h \in X$ with $9 < \|h\|_X < \delta$ we have

$$\frac{\|f(x+h) - f(x) - L_x h\|_Y}{\|h\|_X} < \varepsilon.$$

The map L_x is called the *(Fréchet) derivative of f at x* and denoted by $f'(x)$.

If $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, then one often considers $f'(x)$ to be the matrix representation of L_x with respect to the standard basis rather than the linear map itself. If X is a Hilbert space and $Y = \mathbb{R}$, then L_x is a bounded linear functional on X and therefore by the Riesz-Fréchet representation theorem there exists a $y \in X$ such that $L_x h = \langle h, y\rangle$ for all $x \in X$. This representation element y is then called the *gradient* of f at x and denoted by $\nabla f(x)$. Thus, a function $\varphi\colon H \to \mathbb{R}$ (where H is a Hilbert space) is differentiable at x if and only if there exists an element $\nabla\varphi(x) \in H$ such that

$$\lim_{\|h\|\to 0} \frac{|\varphi(x+h) - \varphi(x) - \langle h, \nabla\varphi(x)\rangle|}{\|h\|} = 0.$$

A map $y\colon [0,\infty) \to H$ is differentiable at $t \in (0,\infty)$ if and only if

$$y'(t) := \lim_{s\to t} \frac{1}{s-t}\Big(y(s) - y(t)\Big) \text{ exists in } H$$

and it is called (right) differentiable at 0 if

$$y'(0) := \lim_{s\downarrow 0} \frac{1}{s}\Big(y(s) - y(0)\Big) \text{ exists in } H.$$

4.1 Existence and uniqueness

Definition 4.2. Let H be a Hilbert space and let $\varphi\colon H \to \mathbb{R}$ be differentiable. A *gradient flow curve* for φ is a map $y\colon [0,\infty) \to H$ such that

$$\begin{aligned} &y \text{ is differentiable at each } t \in (0,\infty),\\ &y \text{ is right differentiable at } 0,\\ &y'(t) = -\nabla\varphi(y(t)) \text{ for all } t \geq 0. \end{aligned}$$

If φ is such that $\nabla\varphi$ is a Lipschitz map, then existence and uniqueness of gradient flow curves follow from the "standard" theory of differential equations with Lipschitz coefficients. Let us formulate the result for the Lipschitz case with a sketch of its proof and then discuss what kind of conditions we want to impose on φ. Recall that a map $F\colon X \to X$ (where X is a normed space) is called Lipschitz if there exists a constant L such that

$$\|F(x) - F(y)\| \leq L\|x - y\| \text{ for all } x, y \in H.$$

Theorem 4.3. *Let X be a Banach space and let $F\colon H \to H$ be a Lipschitz map. For every $u \in X$ there exists a unique differentiable $y\colon [0,\infty) \to X$ such that $y'(t) = F(y(t))$ for all $t \geq 0$ and $y(0) = u$.*

Proof. Let us sketch a proof. Consider the integral equation

$$y(t) = u + \int_0^t F(y(s))\,ds, \quad t \geq 0.$$

The integral here is a Banach space valued integral, which can be defined as a limit of Riemann sums. It suffices to show that the integral equation has a unique continuous solution, since it will then be differentiable due to the continuity of F and the fundamental theorem of calculus (in Banach spaces). Fix $T > 0$ and denote $E : C([0,T];X)$, the space of continuous functions on $[0,T]$ with values in E. Endow E with the norm

$$\|f\|_E := \sup_{t\in[0,T]} e^{-\alpha t}\|f(t)\|_X, \quad f \in E,$$

for some sufficiently large α to be determined later on. Then $(E, \|\cdot\|_E)$ is a Banach space. Define

$$(R(f)(t) := u + \int_0^t F(f(s))\,ds, \quad t \in [0,T],\ f \in E.$$

Then R maps E into E and

$$\begin{aligned} \|(R(f))(t) - (R(g))(t)\|_X &= \|\int_0^t F(f(s)) - F(g(s))\,dx\|_X\\ &\leq \int_0^t \|F(f(s)) - F(g(s))\|_X\,ds\\ &\leq L\int_0^t \|f(s) - g(s)\|_X\,ds, \text{ for all } t \in [0,T], \end{aligned}$$

where L is the Lipschitz constant of F. Hence

$$\begin{aligned}
\|R(f)-R(g)\|_E &= \sup_t e^{-\alpha t}\|(R(f))(t)-(R(g))(t)\|_X \leq \sup_t e^{-\alpha t} L\int_0^t \|f(s)-g(s)\|_X\,ds \\
&\leq \sup_t L\int_0^t e^{-alphat}e^{\alpha s}e^{-\alpha s}\|f(s)-g(s)\|_X\,dx \leq L\int_0^T e^{-\alpha(t-s)}\|f-g\|_E\,ds \\
&\leq (L/2\alpha)(1-e^{\alpha T})\|f-g\|_E \leq (L/2\alpha)\|f-g\|_E.
\end{aligned}$$

Choose $\alpha > L$. Then R is a strict contraction E. The Banach fixed point theorem yields existence of a unique fixed point of R, which is easily seen to be a solution of the integral equation for $t \in [0,T]$. If we do the same for T' instead of T we find a unique solution for $t \in [0,T']$. Due to uniqueness the two solutions coincide on $[0,T]\cap[0,T']$. Therefore we can obtain a unique solution defined at each point of $[0,\infty)$. □

Without a Lipschitz condition existence (and/or uniqueness) could fail. For instance, the one dimensional differential equation

$$y'(t) = -y(t)^2, \quad t \geq 0, \tag{4}$$

is satisfied by $y=0$, or can be solved by rewriting

$$-\frac{y'}{y^2} = 1,$$

so $\frac{1}{y} = t+c$, so $y(t) = \frac{1}{t+c}$. With $Y(0)=u$ we obtain $c = -\frac{1}{u}$, so

$$y(t) = \frac{1}{t+\frac{1}{u}}.$$

If $u>0$, then the solution exists for all $t \geq 0$. If $u<0$, then $y(t) = \frac{1}{t+\frac{1}{u}}$ exists for all $t \in [0, -\frac{1}{u})$. If t approaches $-1/u$, then $y(t)\to\infty$, which means that the solution ‘blows up’.

The differential equation

$$y'(t) = F(y(t)), \quad t \geq 0, \tag{5}$$

with

$$F(x) = \begin{cases} -x^2 & \text{if } x \geq 0, \\ x^2 & \text{if } x < 0, \end{cases}$$

has very different behavior. Now

$$y(t) = \begin{cases} \frac{1}{t+\frac{1}{u}} & \text{if } u > 0, \\ 0 & \text{if } u = 0, \\ \frac{1}{\frac{1}{u}-t} & \text{if } u < 0. \end{cases}$$

The solution exists for all $t \geq 0$, whatever the sign of u. Both function $x \mapsto -x^2$ and F are equally non-Lipschitz. The different behavior is due to the difference in sign at the right hand side. In equation (4), if $y(t)$ is positive for some t, then the derivative is negative, so y decreases, so the derivative becomes less negative, so y decreases slower, etc. The solution ‘stabilizes’. If $y(t)$ is negative for some t, however, $y'(t)$ will be negative, so y decreases, so y' becomes even more negative, so y decreases even faster, etc. Hence $y(t)$ goes quicker and quicker to $-\infty$. If the right hand side would be Lipschitz, the ‘growth’ to $-\infty$ would be at most exponential. In the case of $-y(t)^2$ as right hand side, the amplifying effect is so strong that it results in blow up. In (5) the right hand side is such that y would start decreasing as soon as it is positive and increasing as soon as it is negative. So in (5) the right hand side always ‘stabilizes’.

It is therefore no surprise that (5) has a solution $y\colon [0,\infty) \to \mathbb{R}$ for any initial condition. A similar effect holds true for any decreasing continuous right hand side. Is there an analogous effect in higher dimensions?

Phrases as ‘decreasing’ or ‘opposite sign’ do not make sense for functions on $\mathbb{R}^n$. The idea is to involve the notion of convexity. Decreasing functions on $\mathbb{R}$ correspond to derivatives of concave functions, or rather -1 times derivative of convex functions. Convexity is also defined for functions on $\mathbb{R}^n$ or even on arbitrary vector spaces. The right hand sides that we will consider are minus the gradient of a convex function.

Definition 4.4. Let X be a vector space over $\mathbb{R}$. A function $\varphi\colon X \to \mathbb{R}$ is called *convex* if for every $x, y \in X$ we have

$$\varphi((1-t)x + ty) \leq (1-t)\varphi(x) + t\varphi(y) \text{ for all } t \in [0,1].$$

The following theorem is the main theorem on existence and uniqueness of gradient flows for convex functionals on Hilbert spaces. Its proof is long and complicated. We will only consider a very brief sketch. For full details one can consult [4, 8, 9] (Crandall-Liggett, Brezis, Clement).

Theorem 4.5. *Let H be a Hilbert space over $\mathbb{R}$. If $\varphi\colon H \to \mathbb{R}$ is differentiable and convex, then for every $u \in H$ there exists a unique $y\colon [0,\infty) \to H$ such that*

$$\begin{aligned} &y'(t) = -\nabla\varphi(y(t)), \quad t \geq 0,\\ &y(0) = u. \end{aligned}$$

Idea of proof. Fix $h > 0$. Discretize the differential equation with Euler’s implicit scheme,

$$\frac{y((n+1)h) - y(nh)}{h} = -\nabla\varphi(y(nh).$$

Then, with $y_n = y(nh)$,

$$y_{n+1} + h\nabla\varphi(y_{n+1}) = y_n,$$

so that

$$y_{n+1} = (I + h\nabla\varphi)^{-1}(y_n).$$

One of the difficulties is to show that the function $x \mapsto x + h\nabla\varphi(x)$ is invertible. Then

$$J_h(x) := (I + h\nabla\varphi)^{-1}(x), \quad x \in H,$$

is called the *resolvent* associated to $\nabla\varphi$. We obtain

$$y_n = J_h^n(u),$$

and this is hoped to be a good approximation for $y(nh)$. The next steps are to show that

$$y(t) := \lim_{k\to\infty} J_{t/k}^n(u)$$

exists and that the function y thus defined is the unique solution. □

In many practical situations the functional φ will not be entirely convex, for instance the function $x \mapsto (x-1)^2(x+1)^2$. The existence theorem above can be extended to such functions as well. A function $\varphi H \to \mathbb{R}$ is called *α-convex* (for some $\alpha \in \mathbb{R}$) if

$$x \mapsto \varphi + \frac{\alpha}{2}\|x\|^2$$

is convex on H. Instead of φ being convex in the previous theorem it is sufficient that φ is α-convex for some $\alpha \in \mathbb{R}$.

4.2 Evolution variational inequality

The gradient flow differential equation

$$y'(t) = -\nabla\varphi(y(t))$$

for a convex function φ on a Hilbert space H can be rewritten in a form which does not involve the derivative of φ nor any other linear structure of the Hilbert space. It only uses the metric of H. This form then makes sense in any metric space and we will use it to generalize the concept of gradient flow.

Theorem 4.6. *Let H be a Hilbert space over $\mathbb{R}$ and let $\varphi\colon H \to \mathbb{R}$ be differentiable and convex. For a differentiable $y\colon [0,\infty) \to H$ we have*

$$y'(t) = -\nabla\varphi(y(t)) \textit{ for all } t \geq 0$$

if and only if

$$\frac{1}{2}\frac{d}{dt}\|y(t)-z\|^2 + \varphi(y(t)) \leq \varphi(z) \textit{ for all } t \geq 0,\ \ z \in H.$$

For the proof we need the following lemma.

Lemma 4.7. *$\nabla\varphi(x)$ is the unique $a \in H$ such that*

$$\langle a, z-x\rangle \leq \varphi(z) - \varphi(x) \textit{ for all } z \in H.$$

Proof. Let $z \in H$. Define

$$z_t := (1-t)z + tx, \quad 0 \leq t \leq 1.$$

Then convexity of φ yields

$$\varphi(z_t) \leq (1-t)\varphi(z) + t\varphi(x).$$

Since $\|z_t - x\| \to 0$ as $t \to 0$, we have

$$\frac{|\varphi(z_t) - \varphi(x) - \langle\nabla\varphi(x), z_t - x\rangle|}{\|z_t - x\|} \to 0 \text{ as } t \to 0.$$

Observe that $\|z_t - x\| = (1-t)(z-x)$. We obtain

$$\begin{aligned}
\langle\nabla\varphi(x), z-x\rangle &= \lim_{t\downarrow 0}\frac{(1-t)\langle\nabla\varphi(x), z-x\rangle}{(1-t)\|z-x\|}\|z-x\| \\
&= \lim_{t\downarrow 0}\left(\frac{\langle\nabla\varphi(x), z-x\rangle}{(1-t)\|z-x\|} - \frac{\varphi(z_t)-\varphi(x)}{(1-t)\|z-x\|}\right)\|z-x\| + \lim_{t\downarrow 0}\frac{\varphi(z_t)-\varphi(x)}{1-t} \\
&\leq 0 + \lim_{t\downarrow 0}\frac{(1-t)\varphi(z) + t\varphi(x) - \varphi(x)}{1-t} \\
&= \varphi(z) - \varphi(x).
\end{aligned}$$

For the uniqueness, suppose $a \in H$ is such that $\langle a, z - x\rangle \leq \varphi(z) - \varphi(x)$ for all $z \in H$. Fix $y \in H$ with $\|y\| = 1$. Then

$$\begin{aligned} \langle a - \nabla\varphi(x), y\rangle &= \frac{\langle a, ty\rangle}{\|ty\|} \\ &\leq \frac{\varphi(x+ty) - \varphi(x) - \langle \nabla\varphi(x), ty\rangle}{\|ty\|} \to 0,\ t \to 0, \end{aligned}$$

so $\langle a - \nabla\varphi(x).y\rangle \leq 0$. This holds for all $y \in H$, so (consider y and $-y$) $\langle a - \nabla\varphi(x), y\rangle = 0$ for all $y \in H$, so $a - \nabla\varphi(x) = 0$. □

Now we can prove the theorem.

Proof of theorem. Denote $F(x) = \|x\|^2$ and $G(t) = y(t) - z$. Then

$$\frac{d}{dt}F(G(t)) = F'(G(t))G'(t) = \langle 2x, G'(t)\rangle,$$

hence

$$\frac{d}{dt}\|y(t) - z\|^2 = \langle 2(y(t) - z), y'(t)\rangle.$$

If $y'(t) = -\nabla\varphi(y(t))$, then

$$\begin{aligned} \frac{d}{dt}\|y(t) - z\|^2 &= 2\langle y(t) - z, y'(t)\rangle = 2\langle \nabla\varphi(x), z - y(t)\rangle \\ &\leq 2\Big(\varphi(z) - \varphi(y(t))\Big), \end{aligned}$$

due to the lemma.

Conversely, if $\frac{1}{2}\frac{d}{dt}\|y(t) - z\|^2 + \varphi(y(t)) \leq \varphi(z)$ for all $z \in H$, then

$$\langle -y'(t), z - y(t)\rangle \leq \varphi(z)\varphi(y(t))$$

for all $z \in H$, so, by the lemma,

$$y'(t) = -\nabla\varphi(y(t)).$$

□

There is a similar theorem for functions φ that are only α-convex for some $\alpha \in \mathbb{R}$. We state it without proof.

Theorem 4.8. *Let $\varphi\colon H \to \mathbb{R}$ be α-convex for some $\alpha \in \mathbb{R}$ and differentiable. For a differentiable $y\colon [0, \infty) \to H$ we have*

$$y'(t) = -\nabla\varphi(y(t)) \textit{ for all } t \geq 0$$

if and only if

$$\frac{1}{2}\frac{d}{dt}\|y(t) - z\|^2 + \frac{\alpha}{2}\|u(t) - z\|^2 + \varphi(y(t)) \leq \varphi(z)$$

for all $z \in H$, $t \geq 0$.

5 Gradient flows in metric spaces

Let (X, d) be a complete metric space. In view of the equivalent formulation of a gradient flow in a metric space we will call a curve $y\colon [0, \infty) \to X$ a gradient flow for a potential function $\varphi\colon X \to \mathbb{R}$ if

$$\frac{1}{2}\frac{d}{dt}d(y(t), z)^2 + \varphi(y(t)) \leq \varphi(z) \text{ for all } z \in X,\ t \geq 0.$$

There are three issues to take care of. First, some regularity of y is needed to guarantee that $\frac{d}{dt}d(y(t), z)^2$ exists. Second, A suitable condition on φ replacing the convexity of the Hilbert space case will be needed. Third, existence and uniqueness needs to be established.

5.1 Absolutely continuous curves

For a function $y\colon [0, \infty) \to X$ we cannot talk about differentiability because X need not have any linear structure. Instead we can define a suitable notion of absolute continuity. We start by recalling some facts from real analysis.

If $y\colon [a, b] \to \mathbb{R}$ is differentiable almost everywhere, then its derivative y' is measurable, where we set $y'(t) = 0$ at every t where y is not differentiable. If y' is integrable, that is $y' \in L^1[a, b]$, can we recover y from y' in the sense that

$$y(x) = \int_a^x y'(t)\,dt\ ?$$

In general not.

Example. The *Cantor function* $g\colon [0, 1] \to [0, 1]$ is defined as follows. On $[1/3, 2/3]$ g equals $1/2$, on $[1/9, 2/9]$ g equals $1/4$ and on $[7/9, 8/9]$ g equals $3/4$. Then on $[1/27, 2/27]$ g equals $1/8$, on $[7/27, 8/27]$ g equals $3/8$, etc. It can be show that g is continuous, increasing, differentiable almost everywhere, and that $g' = 0$ almost everywhere. Clearly, $\int_0^1 g'(t)\,dt \neq g(1) - g(0)$.

Functions which are anti-derivatives of their almost everywhere derivative are the so-called absolutely continuous functions. There is a more explicit way to describe thid property, which we take as definition.

Definition 5.1. Let (X, d) be a metric space. A function $[a, b] \to X$ is called *absolutely continuous* on $[a, b]$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for any disjoint subintervals $(a_1, b_1), \dots, (a_n, b_n)$ of $[a, b]$ we have

$$\sum_{k=1}^{n} d(y(a_k), y(b_k)) < \varepsilon.$$

An absolutely continuous function is continuous. The Cantor function is continuous but not absolutely continuous. The following theorem is well known in real analysis and we include it without a proof.

Theorem 5.2. *For $y\colon [0, \infty) \to \mathbb{R}$ the following statements are equivalent:*

(a) y is absolutely continuous

(b) there exists a $v \in L^1[a,b]$ such that $y(t) = y(a) + \int_a^t v(s)\,ds$ for all $t \in [a,b]$.

If (a)-(b) hold true, then y is differentiable almost everywhere and $y' = v$ almost everywhere.

The following lemma is very useful.

Lemma 5.3. *Let $y\colon [0,\infty) \to (X,d)$. The following statements are equivalent:*

(a) y is absolutely continuous on $[a,b]$

(b) there exists a $u \in L^1[a,b]$, with $u \geq 0$ a.e. such that $d(y(s),y(t)) \leq \int_s^t u(r)\,dr$ for all $a \leq s \leq t \leq b$.

Proof. (b)⇒(a): Let $U(t) := \int_a^t u(s)\,ds$, $t \in [a,b]$. Then U is absolutely continuous. Let $\varepsilon > 0$. Take $\delta > 0$ such that

$$\left.\begin{array}{r}(a_1,b_1),\ldots,(a_n,b_n) \text{ disjoint}\\ \sum_{k=1}^n (b_k - a_k) < \delta\end{array}\right\} \Longrightarrow \sum_{k=1}^n |U(b_k) - U(a_k)| < \varepsilon.$$

Then

$$\sum_{k=1}^n d(y(a_k),y(b_k)) \leq \sum_{k=1}^n |U(b_k) - U(a_k)| < \varepsilon.$$

(a)⇒(b): Define

$$U(t) := \sup\{\sum_{k=1}^n d(y(t_{k-1}),y(t_k))\colon\ a = t_0 < \cdots < t_n = b\} = \mathrm{Var}\,(y;[a,b]),$$

which is called the total variation of y over $[a,t]$. We first show that $U(t)$ is finite for all t. For $\varepsilon > 0$ take $\delta > 0$ corresponding to the absolute continuity of y. Fix a partition $a = s_0 < \ldots < s_m = b$ such that $|s_k - s_{k-1}| < \delta$ for all k. Then

$$\sum_{j=1}^m d(y(s_{j-1}),y(s_j)) \leq m\varepsilon.$$

Let $a = t_0 < \ldots < t_n = b$. Make a joint refinement $a = r_0 < \ldots < r_l = b$; $\{t_0,\ldots,t_n\} \subseteq \{r_0,\ldots,r_l\}$, $\{s_0,\ldots,s_m\} \subseteq \{r_0,\ldots,r_l\}$. By the triangle inequality,

$$\sum_{k=1}^n d(y(s_{k-1}),y(s_k)) \leq \sum_{i=1}^l d(y(r_{i-1},y(r_i)) = \sum_{k=1}^m \sum_i \in I_k d(y(r_{i-1},y(r_i)),$$

where $I_k = \{i\colon\ r_i \in (s_{k-1},s_k]\}$. Then $\sum_{i \in I_k} d(y(r_{i-1}),y(r_i)) < \varepsilon$ for all k by the choice of δ and $s_0,\ldots s_m$. Hence

$$\sum_{k=1}^n d(y(t_{k-1}),y(t_k)) \leq m\varepsilon.$$

Clearly, $U(t) \geq U(s) + d(y(s,y(t))$, so

$$d(y(s),y(t)) \leq U(t) - U(s).$$

Next we show that U is absolutely continuous. Let $\varepsilon > 0$. Take $\delta > 0$ corresponding to the absolute continuity of y. Let $(a_1, b_1), \dots, (a_n, b_n)$ be disjoint subintervals of $[a, b]$ such that $\sum_{k=1}^{n}(b_k - a_k) < \delta$. Observe that

$$U(b_k) - U(a_k) = \sup\{\sum_{l=1}^{m} d(y(t_{l-1}), y(t_l))\colon\ a_k = t_0 < \cdots < t_m = b\}.$$

Choose $t_0^k, \dots, t_{m_k}^k$ such that

$$\sum_{l=1}^{m_k} d(y(t_{l-1}^k), y(t_l^k)) > U(t_k) - U(t_{k-1}) - \varepsilon/n$$

for each $k = 1, \dots, n$. Then $\sum_{l=1}^{m_k}(t_{l-1}^k - t_l) \leq b_k - a_k$, so

$$\sum_{k=1}^{n} |U(t_k) - U(t_{k-1})| \leq \varepsilon + n\varepsilon/n = 2\varepsilon.$$

Hence there exists a $u \in L^1[a, b]$ such that $U(t) = U(a) + \int_a^t u(s)\, ds$. U is increasing, so $u \geq 0$ a.e. and

$$U(t) - U(s) = \int_s^t u(s)\, ds.$$

□

Definition 5.4. Let (X, d) be a metric space. A function $y\colon (0, \infty) \to X$ is called *absolutely continuous* on $(0, \infty)$, notation $y \in AC((0, \infty); X)$, if y is absolutely continuous on $[a, b]$ for every $0 < a < b$.

Notice that $y(t) = \ln t$ is such that $y'(t) = 1/t \notin L^1[a, b]$, so y is absolutely continuous on $[a, 1]$ for all $a > 0$ but not on $[0, 1]$.

5.2 Gradient flows

Let (X, d) be a metric space.

Definition 5.5. Let $\varphi\colon X \to \mathbb{R}$. The *Evolution Variational Inequality (EVI)* is the inequality

$$\frac{1}{2}\frac{d}{dt} d(y(t), z)^2 + \varphi(y(t)) \leq \varphi(z) \text{ for a.e. } t > 0, \text{ for all } z \in X.$$

We say that a map $y\colon [0, \infty) \to X$ is a solution of (EVI) if

$$\begin{aligned}&y\colon [0, \infty) \to X \text{ is continuous,}\\ &y \in AC((0, \infty); X), \text{ and}\\ &y \text{ satisfies (EVI).}\end{aligned}$$

A solution of (EVI) is called a *gradient flow* for the potential φ.

If X is a Hilbert space, there exists a solution of (EVI) if φ is convex. In an arbitrary metric space there are no lign segment and convexity is not defined. It turns out that there is a similar existence theorem for gradient flows in arbitrary complete metric spaces if φ satisfies a suitable condition replacing the convexity. It actually is a joint condition on φ and the metric of X. As a motivation for the condition we consider an identity in Hilbert spaces first.

Lemma 5.6. *Let $(H, \langle , \rangle)$ be a Hilbert space. For every $x, u, v \in H$ and every $t \in [0, 1]$ we have*

$$\|(1-t)u + tv - x\|^2 = (1-t)\|u-x\|^2 + t\|v-x\|^2 - t(1-t)\|u-v\|^2.$$

Proof.

$$\begin{aligned}\|(1-t)(u-x) + t(v-x)\|^2 &= (1-t)^2\|u-x\|^2 + t^2\|v-x\|^2 + 2t(1-t)\langle u-x, v-x\rangle \\ &= (1-2t+t^2)\|u-x\|^2 + t^2\|v-x\|^2 + 2t(1-t)\langle u-x, v-x\rangle \\ &= (1-t)\|u-x\|^2 + t\|v-x\|^2 - t(1-t)\Big(\|u-x\|^2 - 2\langle u-x, v-x\rangle + \|v-x\|^2\Big) \\ &= (1-t)\|u-x\|^2 + t\|v-x\|^2 - t(1-t)\|u-v\|^2.\end{aligned}$$

□

It follows from the lemma that $\varphi\colon H \to \mathbb{R}$ is convex if and only if for every $\theta > 0$

$$\begin{aligned}\theta\|(1-t)u + tv - x\|^2 + \varphi((1-t)u + tv) \le &(1-t)\Big(\theta\|u-x\|^2 + \varphi(u)\Big) + t\Big(\theta\|u-x\|^2 + \varphi(v)\Big) \\ &- \theta t(1-t)\|u-v\|^2\end{aligned}$$

for all $t \in [0, 1]$.

Let (X, d) be a metric space and let $\varphi\colon X \to \mathbb{R}$. Consider the following to conditions on φ:

(H1) For every $x, u, v \in X$ there exists $\gamma\colon [0, 1] \to X$ continuous such that $\gamma(0) = u$, $\gamma(1) = v$ and such that

$$\begin{aligned}\theta d(\gamma(t), x)^2 + \varphi(\gamma(t)) \le &(1-t)\Big(\theta d(u, x)^2 + \varphi(u)\Big) + t\Big(\theta d(v, x)^2 + \varphi(v)\Big) \\ &- \theta t(1-t)d(u, v)^2\end{aligned}$$

for all $t \in [0, 1]$ and all $\theta > 0$.

(H2) φ is bounded below on some closed ball, that is, there exist $m \in X$, $r > 0$, and $\alpha \in \mathbb{R}$ such that $\varphi(x) \ge \alpha$ for every $x \in X$ with $d(x, m) \le r$.

The following theorem on existence and uniqueness of gradient flows for φ holds true.

Theorem 5.7. *Let (X, d) be a complete metric space, let $\varphi\colon X \to \mathbb{R}$ be lower semicontinuous and such that* (H1) *and* (H2) *hold. Then for every $u \in X$ there exists a unique solution y of (EVI) with $y(0) = u$. That is,*

$y\colon [0, \infty) \to X$ is continuous,
u is absolutely continuous on $(0, \infty)$,

and

$$\frac{1}{2}\frac{d}{dt}d(y(t),z)^2 + \varphi(y(t)) \leq \varphi(z) \text{ a.e. on } (0,\infty) \text{ for all } z \in X.$$

Moreover,

$$t \mapsto \varphi(y(t) \text{ is decreasing}$$

and if $u_1, u_2 \in X$ *with corresponding solutions* y_1, y_2*, then*

$$d(y_1(t), y_2(t)) \leq d(u_1, u_2) \text{ for all } t \geq .$$

Condition (H1) is a mixed condition on the function φ as well as on the metric d. For instance, if $\varphi = 0$, then the remaining property of the metric d needed to satisfy (H1) is quite similar to the structure of a Hilbert space metric. Apart from Hilbert spaces, there is another interesting class of metric spaces that satisfy (H1) for $\varphi = 0$ (and many other functions φ): the Wasserstein spaces. They are spaces of probability measures endowed with a metric defined as an optimal cost in a Kantorovich problem.

6 Wasserstein spaces

6.1 The Wasserstein metric

The set of Borel probability measures on a complete separable metric space can be endowed with the topolgy of narrow convergence. This topology is metrizable, for instance by the bounded Lipschitz metric. With the aid of Kantorovich's optimal transport problem we can define another metric. It turns out that this metric suits well to the conditions concerning existence of gradient flows.

Recall the following existence result.

Theorem 6.1. *Let* (X, d) *be a separable complete metric space and let* $c \colon X \to [0, \infty)$ *be lower semicontinuous. For every* $\mu, \nu \in \mathcal{P}(X)$ *there exists an optimal transport plan* γ *for* c*, that is,*

$$\gamma \in \Gamma(\mu,\nu) := \{\eta \in \mathcal{P}(X) \colon \ eta(A \times X) = \mu(A), \ \eta(X \times A) = \nu(A) \text{ for all } A \subseteq X \text{ Borel}\}$$

and

$$\int_{X \times X} c(x,y)\, d\gamma(x,y) = \inf\{\int_{X\times X} c(x,y)\, d\eta(x,y) \colon \ \eta \in \Gamma(\mu,\nu)\}.$$

The set of all optimal transport maps for c with marginals μ and ν is denoted by Γ_o.

For $p \geq 1$ the map $(x,y) \mapsto d(x,y)^p$ is continuous. Define for $\mu, \nu \in \mathcal{P}(X)$,

$$\begin{aligned} W_p(\mu,\nu) :=& \inf\{\int_{X\times X} d(x,y)^p \, d\eta(x,y) \colon \ \eta \in \Gamma(\mu,\nu)\}^{1/p} \\ & \left(\int_{X\times X} d(x,y)^p \, d\gamma(x,y)\right)^{1/p}, \quad \text{with } \gamma \in \Gamma_o(\mu,\nu). \end{aligned}$$

Note that $W_p(\mu,\nu) = \infty$ may occur. We will restrict W_p to a subset of $\mathcal{P}(X)$ on wich it is finite.

Definition 6.2. Let (X,d) be a separable complete metric space and let $p \geq 1$. A $\mu \in \mathcal{P}(X)$ is said to have *finite pth moment* if there exists an $x_0 \in X$ such that $\int_{X\times X} d(x,y)^p \, d\mu(x) < \infty$. The subset of measures with finite pth moment is denoted by

$$\mathcal{P}_p(X) := \{\mu \in \mathcal{P}(X)\colon \text{ there exists } x_0 \in X \text{ such that } \int_X d(x,y)^p \, d\mu(x) < \infty\}.$$

Observe that $\mathcal{P}_p(X) = \mathcal{P}(X)$ if the metric space has finite diameter in the sense that there exists an $R > 0$ such that $d(x,y) \leq R$ for all $x, y \in X$.

The finiteness of the integral in the definition of $\mathcal{P}_p(X)$ does not depend on the choice of the point x_0.

Lemma 6.3. *Let (X,d) be a separable complete metric space. Let $\mu \in \mathcal{P}_p(X)$. Then*

$$\int_X d(x,z)^p \, d\mu(x) < \infty \textit{ for every } z \in X.$$

Proof. By Hölder's inequality,

$$a + b = \begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \leq (|a|^p + |b|^p)^{1/p}(1^q + 1^q)^{1/q},$$

where $\frac{1}{p} + \frac{1}{q} = 1$. So for $a, b \geq 0$,

$$(a+b)^p \leq 2^{p-1}(a^p + b^p).$$

Hence

$$d(x,y)^p \leq \Big(d(x,x_0) + d(x_0,z)\Big)^p \leq 2^{p-1}\Big(d(x,x_0)^p + d(x_0,z)^p\Big),$$

so that

$$\int_X d(x,z)^p \, d\mu(x) \leq 2^{p-1} \int_X d(x,x_0)^p \, d\mu(x) + 2^{p-1} \int_X d(x_0,z)^p \, d\mu(x),$$

which is finite. □

Theorem 6.4. *Let (X,d) be a separable complete metric space and let $p \geq 1$. Then W_p is a metrix on $\mathcal{P}_p(X)$.*

Proof. Let $\mu, \nu \in \mathcal{P}_p(X)$.

$W_p(\mu,\nu) < \infty$: Since $\mu \otimes \nu \in \Gamma(\mu,\nu)$, Fubini yields

$$\begin{aligned} W_p(\mu,\nu)^p &\leq \int_{X\times X} d(x,y)^p \, d(\mu\otimes\nu)(x,y) = \int_X \left(\int_X d(x,y)^p \, d\mu(x)\right) d\nu(y) \\ &\leq \int_X \left(\int_X 2^{p-1}\Big(d(x,x_0)^p + d(x_0,y)^p\Big) \, d\mu(x)\right) d\nu(y) \\ &= 2^{p-1}\left(\int_X d(x,x_0)^p \, d\mu(x) + \int_X d(x_0,y)^p \, d\nu(y)\right) < \infty. \end{aligned}$$

$W_p(\mu,\mu) = 0$: Define $f\colon X \to X \times X$ by $f(x) := (x,x)$, $x \in X$, and define $\eta := f_{\#}\mu$. Then $\eta(S) = \mu(\{x\colon f(x) \in S\}) = \mu(\{x\colon (x,x) \in S\})$ for any Borel set $S \subseteq X \times X$, so $\eta(A \times X) = \mu(A)$ and $\eta(X \times A) = \nu(A)$ for any Borel set $A \subseteq X$. Hence $\eta \in \Gamma(\mu, nu)$ and

$$0 \leq W_p(\mu,\mu)^p = \int_{X\times X} d(x,y)^p \, d\eta(x,y) = \int_X d(f(x))^p \, d\mu(x) = 0.$$

Thus $W_p(\mu,\mu)=0$.

$W_p(\mu,\nu)=W_p(\nu,\mu)$: Let $\gamma\in\Gamma_o(\mu,\nu)$ (for the cost function $(x,y)\mapsto d(x,y)^p$). Then

$$W_p(\mu,\nu)^p=\int_{X\times X} d(x,y)^p\,d\gamma(x,y)=\int_{X\times X} d(f(x,y))^p\,d\eta(x,y),$$

with $f(x,y):=(y,x)$ and $\eta:=f_{\#}\gamma$. Then $\eta\in\Gamma(\nu,\mu)$, so that

$$W_p(\mu,\nu)^p=\int_{X\times X} d(x,y)^p\,d\eta(x,y)=\int_{X\times X} d(s,y)^p\,d\eta(x,y)\geq W_p(\nu,\mu)^p.$$

The desired equality follows if we do the same computation with the roles of μ and ν reversed.

If $W_p(\mu,\nu)=0$ then $\mu=\nu$: Let $\gamma\in\Gamma(\mu,\nu)$ be optimal, that is,

$$\int_{X\times X} d(x,y)^p\,d\gamma(x,y)=W_p(\mu,\nu)^p=0.$$

Denote the 'diagonal' of $X\times X$ by

$$\Delta:=\{(x,x)\colon\ x\in X\}.$$

Suppose that $S:=\operatorname{supp}\gamma\not\subseteq\Delta$. Then there exists a $z\in S\setminus\Delta$. Take $U\subseteq X\times X$ open such that $x\in U$, $U\cap\Delta=\emptyset$. Then $\gamma(U)>0$ (since $z\in S$) and $d(x,y)^p>0$ for all $(x,y)\in U$. Hence

$$\int_{X\times X} d(x,y)^p\,d\gamma(x,y)\geq\int_U d(x,y)^p\,d\gamma(x,y)>0,$$

which is a contradiction. Therefore $S\subseteq\Delta$. Next, let $X\subseteq X$ be a Borel set and define $D:=\{(x,x)\colon\ x\in A\}$. Then $(A\times X)\cap S=D$, so

$$\mu(A)=\gamma(A\times X)=\gamma((A\times X)\cap S)=\gamma(D)=\gamma((X\times A)\cap S)=\gamma(X\times A)=\nu(A).$$

Finally, we show the triangle inequality for W_p. Let $\mu_1,\mu_2,\mu_3\in\mathcal{P}_p(X)$. Take $\gamma_{12}\in\Gamma_o(\mu_1,\mu_2)$ and $\gamma_{23}\in\Gamma_o(\mu_2,\mu_3)$. Due to the disintegration theorem there exists a family $(\nu_{1,x})_{x\in X}$ in $\mathcal{P}(X)$ such that

$$x\mapsto\nu_{1,x}(A)\text{ is measurable for every }A\subseteq X\text{ Borel}$$

and

$$\int_{X\times X} h(x,y)\,d\gamma_{12}(x,y)=\int_X\left(\int_X h(x,y)\,d\nu_{1,x}(y)\right)d\mu_1(x)$$

for every Borel measurable $h\colon X\times X\to[0,\infty)$ and, similarly, there exists a family $(\nu_{3,y})_{y\in X}$ in $\mathcal{P}(X)$ with $y\mapsto\nu_{3,y}(A)$ measurable for each Borel set $A\subseteq X$ and

$$\int_{X\times X} h(x,y)\,d\gamma_{23}(y,z)=\int_X\left(\int_X h(y,z)\,d\nu_{3,y}(y)\right)d\mu_2(y)$$

for every Borel measurable $h\colon X\times X\to[0,\infty)$. Define for $V\subseteq X\times X\times X$ Borel

$$\gamma(V):=\int_X\int_X\int_X 1_V(x,y,z)\,d\nu_{3,y}(z)\,d\nu_{1,x}(y)\,d\mu_1(x).$$

Then $\gamma \in \mathcal{P}(X \times X \times X)$,

$$\gamma(A \times X \times X) = \int_A \int_X \int_X 1 \, d\nu_{3,y}(z) \, d\nu_{1,x}(y) \, d\mu_1(x) = \mu_1(A),$$

$$\gamma(X \times A \times X) = \int_X \int_A 1 \, d\nu_{1,x}(y) \, d\mu_1(x) = \int_X \int_X 1_{X\times X}(x,y) \, d\nu_{1,x}(y) \, d\mu_1(x)$$
$$\int_{X\times X} 1_{X\times A} \, d\gamma_{12}(x,y) = \gamma_{12}(X \times A) = \mu_2(A),$$

$$\begin{aligned}\gamma(X \times X \times A) &= \int_X \int_X \int_X 1_A(z) \, d\nu_{3,y}(z) \, d\nu_{1,x}(y) \, d\mu_1(x) \\ &= \int_{X\times X} \int_X 1_A(z) \, d\nu_{3,y}(z) \, d\gamma_{12}(x,y) \\ &= \int_X \int_X 1_A(z) \, d\nu_{3,y}(z) \, d\mu_2(y) = \int_{X\times X} 1_A(z) \, d\gamma_{23}(z) \\ &= \gamma_{23}(X \times A) = \mu_3(A).\end{aligned}$$

Denote $\pi_{13}(x,y,z) := (x,z)$ and define $\gamma_{13} := (\pi_1 3)_\# \gamma$, that is, $\gamma_{13}(A \times B) = \gamma(A \times X \times B)$. Then $\gamma_{13} \in \Gamma(\mu_1, \mu_3)$, so

$$\begin{aligned}W_p(\mu_1,\mu_2) &\le \left(\int_{X\times X} d(x,z)^p \, d\gamma_{13}(x,z)\right)^{1/p} = \left(\int_{X\times X\times X} d(\pi_{13}(x,y,z))^p \, d\gamma(x,y,z)\right)^{1/p} \\ &\le \left(\int_{X\times X\times X} \Big(d(x,y) + d(y,z)\Big)^p \, d\gamma(x,y,z)\right)^{1/p} \\ &= \|d(x,y,z) \mapsto d(x,y) + d(y,z)\|_{L^p(\gamma)} \le \|d(x,y)\|_{L^p(\gamma)} + \|d(y,z)\|_{L^p(\gamma)} \\ &= \int_{X\times X\times X} d(x,y)^p \, d\gamma(x,y) + \int_{X\times X} d(y,z)^p \, d\gamma(y,z) \\ &= \left(\int_X \int_X \int_X d(x,y)^p \, d\nu_{3,y}(z) \, d\nu_{1,x}(y) \, d\mu_1(x)\right)^{1/p} \\ &\qquad + \left(\int_X \int_X \int_X d(y,z)^p \, d\nu_{3,y}(z) \, d\nu_{1,x}(y) \, d\mu_1(x)\right)^{1/p} \\ &= \left(\int_{X\times X} d(x,y)^p \, d\gamma(x,y)\right)^{1/p} + \left(\int_x \int_X d(x,z)^p \, d\nu_{3,y}(z) \, d\gamma_{12}(x,y)\right)^{1/p} \\ &= W_p(\mu_1,\mu_2) + \left(\int_X d(y,z)^p \, d\nu_{3,y}(z) \, d\mu_2(y)\right)^{1/p} \\ &= W_p(\mu_1,\mu_2) + \int_X d(y,z)^p \, d\gamma_{23}(y,z) \\ &= W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3).\end{aligned}$$

Thus W_p satisfies the triangle inequality. □

The metric W_p on $\mathcal{P}_p(X)$ is called the *Wasserstein-p metric* on $\mathcal{P}(X)$.

If the metric space (X,d) has finite diameter, then the topology induced by W_p is the topology of narrow convergence. The proof needs a lemma on infinite products of measures, which we cite first.

Lemma 6.5. *Let $(X_0, d_0), (X_1, d_1), \dots$ be separable complete metric spaces, let $X_\infty := \Pi_{i=0}^\infty X_i$ endowed with the metric*

$$d((x_i)_i, (y_i)_i) = \sum_{i=0}^{\infty} 2^{-i} \min\{d(x_i, y_i), 1\},$$

and let $\mu_i \in \mathcal{P}(X_i)$ for all i.

(i) For every $\gamma_i \in \Gamma(\mu_i, \mu_{i+1})$, $i = 0, 1, \dots$, there exists a $\nu \in \mathcal{P}(X_\infty)$ such that

$$(\pi_{i,i+1})_\# \nu = \gamma_i \text{ for all } i = 1, 2, \dots,$$

where $\pi_{i,i+1}((x_j)_{j=0}^\infty) = (x_i, x_{i+1})$, $(x_j)_{j=0}^\infty \in X_\infty$.

(ii) For every $\gamma_i \in \Gamma(\mu_0, \mu_i)$, $i = 1, 2, \dots$, there exists a $\nu \in \mathcal{P}(X_\infty)$ such that

$$(\pi_{0,i})_\# \nu = \gamma_i \text{ for all } i = 1, 2, \dots.$$

Also the next lemma on supports of measures is needed.

Lemma 6.6. *Let (X, d) be a separable complete metric space and let $\mu \in \mathcal{P}(X)$. If U is an open subset of X with $\mu(U) > 0$, then $U \cap \operatorname{supp} \mu \neq \emptyset$.*

Proof. μ is tight, so there is a compact set $K \subseteq X$ such that $\mu(K) > 0$. We argue by contradiction. Suppose that for all $x \in K$ and for all U_x open with $x \in U_x$ we have $\mu(U_x) = 0$. Then $\bigcup_x U_x$ is an open cover of K, so $K \subseteq \bigcup_{i=1}^n U_{x_i}$ for some $x_1, \dots, x_n \in K$. Then $\mu(K) \leq \sum_{i=1}^n \mu(U_{x_i}) = 0$, which is a contradiction. □

Proposition 6.7. *Let (X, d) be a separable complete metric space such that* $\sup\{d(x, y) \colon x, y \in X\} < \infty$. *Let $(\mu_n)_n$ be a sequence in $\mathcal{P}_p(X)$ and let $\mu \in \mathcal{P}_p(X)$. Then*

$$W_p(\mu_n, \mu) \to 0 \quad \iff \quad \begin{array}{l} \mu_n \to \mu \textit{ narrowly, i.e.} \\ \int_X f \, d\mu_n \to \int_X f \, d\mu \textit{ for all } f \in C_b(X). \end{array}$$

Proof. $\Longrightarrow$) Take for each $i = 1, 2, \dots$ a $\gamma_i \in \Gamma_o(\mu, \mu_i)$ and let ν be as in (ii) of Lemma 6.5. Then

$$\begin{aligned} \int_{X_\infty} d(\pi_n(x), \pi_0(x))^p \, d\nu(x) &= \int_{X_\infty} d(\pi_{0,n}(x))^p \, d\nu(x) \\ &= \int_{X \times X} d(x, y)^p \, d(\pi_{0,n})_\# \nu(x, y) = \int_{X \times X} d(x, y)^p \, d\gamma_n(x, y) \\ &= W_p(\mu, \mu_n)^p \to 0, \ n \to \infty, \end{aligned}$$

so for $f \colon X \to \mathbb{R}$ bounded Lipschitz

$$\begin{aligned} \left| \int_X f(x) \, d\mu_n(x) - \int_X f(x) \, d\mu(x) \right| &= \left| \int_{X_\infty} f(\pi_n(x)) \, d\nu(x) - \int_{X_\infty} f(\pi_0(x)) \, d\nu(x) \right| \\ &\leq \int_{X_\infty} |f(\pi_n(x)) - f(\pi_0(x))| \, d\nu(x) \leq \left(\int_{X_\infty} |f(\pi_n(x)) - f(\pi_0(x))|^p \, d\nu(x) \right)^{1/p} \\ &\leq \operatorname{Lip}(f) \left(\int_{X_\infty} d(\pi_n(x), \pi_0(x))^p \, d\nu(x) \right)^{1/p} \\ &\leq \operatorname{Lip}(f) W_p(\mu, \mu_n) \to, \ n \to \infty. \end{aligned}$$

Hence $\mu_n \to \mu$ narrowly (as convergence with respect to bounded Lipschitz functions is sufficient for narrow convergence).

$\Longleftarrow$) Choose $\gamma_i \in \Gamma_o(\mu, \mu_i)$, for each $i = 1, 2, \ldots$. Since $\{\mu_i \colon , i = 1, 2, \ldots\} \cup \{\mu\}$ is compact in $\mathcal{P}(X)$, Prohorov's theorem yields that this set is tight. Let $\varepsilon > 0$. Then there exists a compact $K \subseteq X$ such that

$$\mu_i(K) \geq 1 - \varepsilon/2 \text{ for all } i \text{ and } \mu(K) \geq 1 - \varepsilon/2.$$

The set $L := K \times K$ is compact in $X \times X$ and

$$\begin{aligned}\gamma_i(X \times X \setminus L) &\leq \gamma_i((X \setminus K) \times X) + \gamma_i(X \times (X \setminus K)) \\ &= \mu(X \setminus K) + \mu_i(X \setminus K) < \varepsilon,\end{aligned}$$

so $\{\gamma_i \colon i = 1, 2, \ldots\}$ is tight. By Prohorov's theorem this set is then compact in the metric space $\mathcal{P}(X \times X)$ and therefore there is a subsequence and a $\gamma \in \mathcal{P}(X \times X)$ such that $\gamma_{i_k} \to \gamma$ in $\mathcal{P}(X \times X)$. Since $(x, y) \mapsto d(x, y)^p$ is a bounded continuous function on $X \times X$ (by the assumption that the diameter is finite) it follows that

$$\int_{X \times X} d(x, y)^p \, d\gamma(x, y) = \lim_{k \to \infty} \int_{X \times X} d(x, y)^p \, d\gamma_{i_k}(x, y) = \lim_{k \to \infty} W_p(\mu, \mu_{i_k}).$$

Also, the coordinate projections π_1 and π_2 from $X \times X$ to X are continuous, so that $\mu = (\pi_1)_\# \gamma_{i_k} \to (\pi_1)_\# \gamma$ and $\mu_{i_k} = (\pi_2)_\# \gamma_{i_k} \to (\pi_2)_\# \gamma$. Hence $\gamma \in \Gamma(\mu, \mu)$.

We show next that γ is an optimal transport plan from μ to μ for the cost function $c \colon (x, y) \mapsto d(x, y)^p$. We first show that $S := \operatorname{supp} \gamma$ is c-monotone. Let $(x_i, y_i) \in S$ for $i = 1, \ldots, n$ and let σ be a permutation of $\{1, \ldots, n\}$. For each $r > 0$ we have $\gamma(B_r(x_i, y_i)) > 0$, where $B_r(x_i, y_i)$ denotes the open ball in $X \times X$ with radius r centered around (x_i, y_i). The Portmanteau theorem yields that

$$\gamma(B_r(x_i, y_i)) \leq \liminf_{k \to \infty} \gamma_{i_k}(B_r(x_i, y_i)),$$

So for k large we have $\gamma_{i_k}(B_r(x_i, y_i)) > 0$ for $i = 1, \ldots, n$. Hence for every $r > 0$ there exists a $K \in \mathbb{N}$ such that $\gamma_{i_k}(B_r(x_i, y_i)) > 0$ for all $k \geq K$. With the aid of the previous lemma it follows that for each $r > 0$ and each $i \in \{1, \ldots, n\}$ there is a point $(x_i^r, y_i^r) \in \operatorname{supp} \gamma_{i_k}$ with $d(x_i^r, x_i) \leq r$ and $d(y_i^r, y_i) \leq r$.

Since γ_{i_k} is optimal, its support $\operatorname{supp} \gamma_{i_k}$ is c-monotone, so

$$\sum_{i=1}^n c(x_{\sigma(i)}^r, y_i^r) \geq \sum_{i=1}^n c(x_i^r, y_i^r).$$

Let $r \downarrow 0$. Since c is continuous we get

$$\sum_{i=1}^n c(x_{\sigma(i)}, y_i) \geq \sum_{i=1}^n c(x_i, y_i).$$

Hence $\operatorname{supp} \gamma$ is c-monotone. By a theorem from optimal transportation theory is then follows that γ is an optimal transport plan for c.

Next we infer that $\int_{X\times X} d(x,y)^p \, d\gamma(x,y) = 0$. We define $f\colon X \to X \times X$ by $f(x) = (x,x)$. Then $f_{\#}\mu \in \Gamma(\mu,\mu)$, so

$$\begin{aligned}\int_{X\times X} d(x,y)^p \, d\gamma(x,y) &\leq \int_{X\times X} d(x,y)^p \, df_{\#}\mu(x,y) \\ &= \int_X d(x,x)^p \, d\mu(x) = 0.\end{aligned}$$

Thus,

$$\begin{aligned}\lim_{k\to\infty} W_p(\mu,\mu_{i_k}) &= \lim_{k\to\infty} \int_{X\times X} d(x,y)^p \, d\gamma_{i_k} \\ &= \int_{X\times X} d(x,y)^p \, d\gamma(x,y) = 0.\end{aligned}$$

By the same arguments we can show that for every subsequence of $(W_p(\mu,\mu_i))_i$ there is a subsequence which converges to 0. That is sufficient to conclude that $W_p(\mu,\mu_i) \to 0$. □

If the metrix space (X,d) does not have finite diameter then the topology on $\mathcal{P}_p(X)$ induced by W_p is a little stronger than the topology of narrow convergence. For sake of completeness we cite the equivalence. A sequence $(\mu_n)_n$ in $\mathcal{P}_p(X)$ is said to have *uniformly integrable p-moments* if there exists an $x_0 \in X$ such that

$$\lim_{r\to\infty} \sup_{n\in\mathbb{N}} \int_{X\setminus B_r(x_0)} d(x,x_0)^p \, d\mu_n(x) = 0,$$

where $B_r(x_0)$ denotes the open ball in X with radius r and center x_0.

Proposition 6.8. *Let (X,d) be a separable complete metric space and let $(\mu_n)_n$ be a sequence in $\mathcal{P}_p(X)$ and $\mu \in \mathcal{P}_p(X)$. Then*

$$W_p(\mu_n,\mu) \to 0 \quad\iff\quad \begin{array}{l}\mu_n \to \mu \textit{ narrowly and} \\ (\mu_n)_n \textit{ has uniformly integrable } p\textit{-moments.}\end{array}$$

Theorem 6.9. *Let (X,d) be a separable complete metric space. Then $(\mathcal{P}_p(X), W_p)$ is a separable complete metric space.*

Proof. To show completeness, let $(\mu_n)_n$ be a Cauchy sequence in $(\mathcal{P}_p(X), W_p)$. Choose a subsequence $(\mu_{n_k})_k$ such that $W_p(\mu_{n_k}, \mu_{n_{k+1}}) < 2^{-k}$. We can relabel the subsequence and thus assume that $W_p(\mu_n,\mu_{n+1}) < 2^{-n}$. Choose $\gamma_i \in \Gamma_o(\mu_i,\mu_{i+1})$ and make $\nu \in \mathcal{P}(X_\infty)$ with $(\pi_{i,i+1})_{\#}\nu = \gamma_i$ for all i, as in Lemma 6.5. Then

$$\sum_{n=1}^{\infty} \int_{X_\infty} d(\pi_n(x), \pi_{n+1}(x))^p \, d\nu(x) = \sum_{k=1}^{\infty} \int_{X\times X} d(x,y)^p \, d\gamma_n(x,y) < \infty.$$

As in the proof of completeness of L^p spaces (with functions taking values in $\mathbb{R}$), one can deduce that there exists a $\pi_\infty\colon X_\infty \to X$ such that

$$\int_{X_\infty} d(\pi_n(x), \pi_\infty(x))^p \, d\nu(x) \to 0.$$

Let $\mu_\infty := (\pi_\infty)_\# \nu$. Then $\mu_\infty \in \mathcal{P}_p(X)$ and

$$\begin{aligned} W_p(\mu_n, \mu_\infty) &\le \int_{X\times X} d(x,y)^p \, d(\pi_n, \pi_\infty)_\# \nu \\ &= \int_{X_\infty} d(\pi_n(x), \pi_\infty(x))^p \, d\nu(x) \to 0, \ n \to \infty. \end{aligned}$$

The separability can be shown in the same way as the separability of $(\mathcal{P}(X), d_{BL})$: the set of convex combinations with rational coefficients of point measures at point of a countable dense subset is dense in $\mathcal{P}_p(X)$. □

6.2 Geodesics in metric spaces

In the definition of convexity of real valued functions on vector spaces (in particular Hilbert spaces) the notion of line segment plays a role. In an arbitrary metric space, there is no linear structure and therefore there are no line segments. There is an alternative notion. Line segments are the shortest paths between two points and shortest pats can be defines in arbitrary metric spaces.

Let (X, d) be a metric space.

Definition 6.10. A *geodesic* in X is a curve $u\colon [a,b] \to X$ such that

$$d(u(r), u(t)) = d(u(r), u(s)) + d(u(s), u(t)) \text{ for all } a \le r \le s \le t \le b.$$

A *constant speed-one geodesic* is a curve $u\colon [a,b] \to X$ such that

$$d(u(s), u(t)) = t - s \text{ for all } a \le s \le t \le b.$$

Usually we will write 'geodesic' if we mean 'constant speed-one geodesic'.

Example. Consider $(\mathbb{R}^2, \|\cdot\|_1)$, that is, $d((u,v),(x,y)) = |u-x| + |v-y|$. Let $A = (0,0)$ and $B = (1,1)$. Then

$$u(t) = (1-t)(0,0) + t(1,1) = (t,t), \ t \in [0,1]$$

is a geodesic from A to B, since

$$d(u(s), u(t)) = \|(t,t) - (s,s)\|_1 = 2|t-s|.$$

In this example this is not the only geodesic connecting A and B. Indeed, also

$$v(t) = \begin{cases} t(1,0), \ 0 \le t \le 1, \\ (t-1)(0,1), \ 1 \le t \le 2. \end{cases}$$

is a geodesic connecting A and B, since

$$d(u(s), u(t)) = \begin{cases} |t-s|, & 0 \le s \le t \le 1, \\ |t-s|, & 1 \le s \le t \le 2, \\ \|(1, t-1) - (s, 0)\|_1 = 1 - s + t - 1 = |t-s|, & 0 \le s \le 1 \le t \le 2. \end{cases}$$

Let us consider geodesics in the Wasserstein spaces over $X = \mathbb{R}^d$. If $\mu_1, \mu_2 \in \mathcal{P}_p(X)$, $(1-t)\mu_1 + t\mu_2$ is an element of $\mathcal{P}_p(X)$ for every $t \in [0,1]$. This segment is however not a geodesic. By means of an optimal transportation plan geodesics can be constructed.

Let $X = \mathbb{R}^d$ with the Euclidean metric $d(x,y) = \|x-y\|_2$. Define the projections $\pi_1, \pi_2 \colon X \times X \to X$ by

$$\pi_1(x,y) = x \text{ and } \pi_2(x,y) = y \text{ for all } x, y \in X.$$

Theorem 6.11. *Let $\mu, \nu in \mathcal{P}_p(\mathbb{R}^d)$. If $\eta \in \Gamma_o(\mu_1, \mu_2)$ and*

$$\alpha_t := (1-t)\pi_1 + t\pi_2.$$

Then the curve u defined by

$$u(t) := (\alpha_t)_{\#}\eta, t \in [0,1]$$

is a constant speed-one geodesic with $\mu(0) = \mu$ and $\mu(1) = \nu$ and

$$W_p(\mu(s), \mu(t)) = |s-t| W_p(\mu, \nu).$$

Proof. Observe that

$$\begin{aligned}\int_X d(x,0)^p \, d\mu(t)(x) &= \int_X \|x\|^p \, d(\alpha_t)_{\#}\eta(x) = \int_{X\times X} \|\alpha_t(x,y)\|^p \, d\eta(x,y) \\ &= \int_{X\times X} \|(1-t)x + ty\|^p \, d\eta(x,y) \\ &\le \int_{X\times X} 2^{p-1}\Big(\|(1-t)x\|^p + \|ty\|^p\Big) \, d\eta(x,y) \\ &= 2^{p-1}\Big(\int_X (1-t)^p \|x\|^p \, d\mu(x) + \int_X t^p \|y\|^p \, d\nu(x)\Big) < \infty,\end{aligned}$$

so $\mu(t) \in \mathcal{P}_p(\mathbb{R}^d)$.

Let $0 \le s \le t \le 1$ and define $\gamma := (\alpha_s, \alpha_t)_{\#}\eta$. Then $\gamma \in \mathcal{P}(X \times X)$,

$$\begin{aligned}\gamma(A \times X) &= \int_{X\times X} 1_{A\times X}(z) \, d\gamma(z) = \int_{X\times X} 1_{A\times X}(\alpha_s(z), \alpha_t(z)) \, d\eta(z) \\ &= \int_{X\times X} 1_A(\alpha_s(z)) \, d\eta(z) = \int_X 1_A(x) d(\ alpha_s)_{\#}\eta(x) \\ &= (\alpha_s)_{\#}\eta(A) = \mu(s)(A)\end{aligned}$$

and similarly

$$\gamma(X \times A) = \mu(t)(A)$$

for $A \subseteq X$ Borel. Hence $\gamma \in \Gamma(\mu(s), \mu(t))$. Therefore,

$$\begin{aligned}W_p(\mu(s), \mu(t))^p &\le \int_{X\times Y} d(x,y)^p \, d\gamma(x,y) = \int_{X\times X} d(x,y)^p \, d(\alpha_s, \alpha_t)_{\#}]eta(x,y) \\ &= \int_{X\times X} d(\alpha_s(x,y), \alpha_t(x,y))^p \, d\eta(x,y) \\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \|(1-s)x + sy - (1-t)x - ty\|^p \, d\eta(x,y) \\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} |t-s|^p \|x-y\|^p \, d\eta(x,y) = |t-s|^p W_p(\mu,\nu)^p,\end{aligned}$$

since $\eta \in \Gamma_o(\mu, \nu)$. □

It follows from the theorem that $t \mapsto \mu(t)$ is absolutely continuous, because

$$W_p(\mu(s), \mu(t)) \leq \int_s^t W_p(\mu, \nu)\, dr.$$

Recall the convexity condition (H1) on a function $\varphi\colon X \to \mathbb{R}$ that was imposed on φ in order to have a gradient flow for φ in the metric space X: For every $x, y_0, y_1 \in X$ there exists a $\gamma\colon [0,1] \to X$ with $\gamma(0) = y_0$, $\gamma(1) = y_1$ and

$$\begin{aligned} d(x,\gamma(t))^2 + \varphi(\gamma(t)) \leq &(1-t)\Big(d(x,y_0)^2 + \varphi(y_0)\Big) \\ +t\Big(d(x,y_1)^2 + \varphi(y_1)\Big) - t(1-t)d(y_0,y_1)^2 \end{aligned}$$

for all $t \in [0,1]$. If $\varphi = 0$ then (H1) is only satisfied if there exists a curve γ such that

$$d(x,\gamma(t))^2 \leq (1-t)d(x,y_0)^2 + td(x,y_1)^2 - t(1-t)d(y_0,y_1)^2.$$

Since in linear spaces the usual convexity property makes use of line segments, it seems reasonably to consider geodesics for γ. It turns out that this may be too restrictive in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$.

Example. Let $X = \mathbb{R}^2$ and $\mu = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(2,1)}$ and $\nu = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(-2,1)}$. Then any Borel map r which maps $(0,0)$ to $(-2,1)$ and $(2,1)$ to $(0,0)$ is optimal for $(x,y) \mapsto \|x-y\|^2$. That means that

$$\eta((0,0),(-2,1)) = 1/2 \text{ and } \eta((2,1),(0,0)) = 1/2$$

defines an optimal transport plan η. So

$$W_2(\mu,\nu)^2 = \int_{\mathbb{R}^2\times\mathbb{R}^2} \|x-y\|^2\, d\eta(x,y) = 5/2 + 5/2 = 5.$$

Let $\mu(t) = (\alpha_t)_{\#}\eta$, where $\alpha_t = (1-t)\pi_1 + t\pi$, $t \in [0,1]$. Then

$$\mu(t)(A) = \eta(\{z\colon\ \alpha_t(z) \in A\}) = \eta(\{(x,y)\colon\ (1-t)x + ty \in A\}).$$

If $(x,y) = ((0,0),(-2,1))$, then $(1-t)x + ty = (-2t,t)$ and if $(x,y) = ((2,1),(0,0))$ then $(1-t)x + ty = (2-2t, 1-t)$. Hence

$$\mu(t) = \tfrac{1}{2}\delta_{(-2t,t)} + \tfrac{1}{2}\delta_{(2-2t,1-t)}.$$

Take $\chi = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(0,-2)}$, then a transport plan with marginals χ and $\mu(t)$ charges $((0,0),(-2t,1))$ and $((0,-2),(2-2t,1-t))$ both with mass $\frac{1}{2}$ or both $((0,0),(2-2t,1-t))$ and $((0,2),(-2t,1))$ with mass $\frac{1}{2}$. Hence

$$W_2(\chi,\mu(t))^2 = \min\{5t^2 - 7t + \tfrac{13}{2}, 5t^2 - 3t + \tfrac{9}{2}\}.$$

This yields that

$$W_2(\chi,\mu(0))^2 = 9/2, \quad W_2(\chi,\mu(1))^2 = 9/2, \quad W_2(\chi,\mu(1/2))^2 = 17/4$$

and it follows that

$$W_2(\chi,\mu(t))^2 \leq (1-t)W_2(\chi,\mu(0))^2 W_2(\chi,\mu(1))^2 - t(t-1)W_2(\mu(0),\mu(1))^2$$

is not satisfied at $t = 1/2$.

The example suggest that in the Wasserstein spaces geodesics are not suitable curves γ to consider in order to verify condition (H1), for instance not for the function $\varphi = 0$. It turns out that a generalization of the geodesics in Wasserstein spaces yields a convenient class of curves.

Definition 6.12. Let (X, d) be a separable complete metric space and let $\mu, \nu, \beta \in \mathcal{P}(X)$. A *generalized geodesic joining μ and ν with base point β* is a curve

$$\mu(t) = (\pi_t^{2\to 3})_{\#}\eta, \ t \in [0, 1]$$

where $\eta \in \Gamma(\mu, \nu\beta)$ (which means that $\eta \in cP(X \times X \times X)$ has marginals μ, ν and β) is such that

$$\pi_{\#}^{1,2}\eta \in \Gamma_o(\beta, \mu), \quad \pi_{\#}^{1,3} \in \Gamma_o(\beta, \nu),$$

and

$$\begin{aligned}&\pi_t^{2\to 3} = (1-t)\pi^2 + t\pi^3,\\ &\pi^{1,2}(x, y, z) = (x, y), \ \pi^{1,3}(x, y, z) = (x, z), \ (x, y, z) \in X \times X \times X.\end{aligned}$$

References

[1] Ambrosio, L., Gigli, N., and Savaré, G., *Gradient flows in metric spaces and in the space of probability measures*, Lectures in Mathematics, ETH Zürich, Birkhäuser Verlag, Basel - Boston - Berlin, 2005.

[2] Billingsley, Patrick, *Convergence of Probability Measures*, Wiley & Sons, New York - London, 1968.

[3] Bourbaki, N., *Éléments de mathématique. Livre 6. Intégration. Chapitre 9*, Hermann, Paris, 1969.

[4] Brézis, H., *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland Mathematics Studies, No. 5, North-Holland Publishing Co. (Elsevier), 1973.

[5] Dellacherie, C.; Meyer, P.-A., *Probabilité et potentiel*, Chapitre I à IV, Édition entièrement refondue, Hermann, Paris, 1973.

[6] Dudley, R.M., Convergence of Baire measures, *Studia Math.* **27** (1966), 251–268.

[7] Dudley, R.M., *Real analysis and probability*, Cambridge Studies in Advanced Mathematics 74, Cambridge University Press, Cambridge, 2002.

[8] Clément, P., An introduction to gradient flows in metric spaces, Report MI-2009-09, Mathematical Institute , Leiden University, www.math.leidenuniv.nl/en/reports/1171, 2009.

[9] Crandall, M.G. and Liggett, T.M., Generation of semi-groups of nonlinear transformations on general Banach spaces, *Amer. J. Math.* **93** (1971), 265–298.

[10] Dunford, Nelson, and Schwartz, Jacob T., *Linear Operators. Part I: General Theory*, Wiley-Interscience, New York, 1957.

[11] Gangbo, W., and R.J. McCann, The geometry of optimal transportation, *Acta Math.* **177**(1996), 113–161.

[12] Halmos, Paul R., *Measure Theory*, Graduate Texts in Mathematics 18, Springer, New York - Berlin - Heidelberg, 1974.

[13] Kelley, John L., *General Topology*, Graduate Texts in Mathematics 27, Springer, New York - Heidelberg - Berlin, 1975.

[14] Parthasarathy, K.R., *Probability Measures on Metric Spaces,* Academic Press, New York - London, 1967.

[15] Prokhorov, Yu.V., Convergence of random processes and limit theorems in probability theory, *Theor. Prob. Appl.* **1** (1956), 157–214.

[16] Rudin, Walter, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1987.

[17] Schechter, Martin, *Principles of Functional Analysis*, 2nd ed., Graduate Studies in Mathematics 36, American Mathematical Society, Providence, Rhode Island, 2002.

[18] van Gaans, Onno, Probability measures on metric spaces, Notes of the seminar 'Stochastic Evolution Equations', Delft University of Technology, 2003.