

Introduction to optimal transport

Work in progress

Michel Pain

Largely based on lecture notes by François Chapon

KMAS9AG1

Université de Toulouse

M2RI 2026–2027

Contents

1	The Monge–Kantorovich problem	3
1.1	The Monge problem	3
1.2	The Kantorovich relaxation	4
1.3	Discrete example	6
2	Analysis of the Kantorovich’s problem	8
2.1	Existence of minimizers	9
2.2	c -cyclical monotonicity	10
2.2.1	General cost functions	10
2.2.2	The case of quadratic cost	12
2.3	Kantorovich duality	16
2.4	Uniqueness of optimal transport maps for the quadratic cost	18
3	Wasserstein distances	20
A	Lower semicontinuity	22
B	Legendre duality	25

Preamble

This notes correspond to the second part of the course *Convergence of probability measures, limit theorems and optimal transport*. I am deeply grateful to François Chapon for sending me his lecture notes, on which these notes are largely based. The main references for this part of the course are the books [1] by Figalli and Glaudo, and [3] by Villani. We also make use of convex analysis, for which the book [2] is a well-known reference.

References

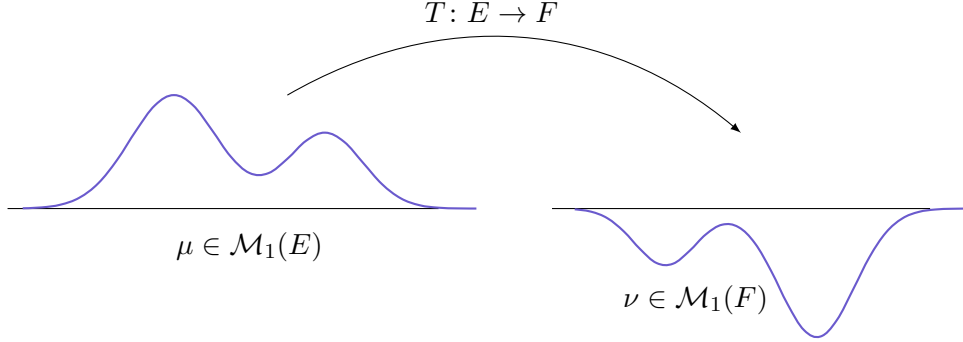
- [1] Alessio Figalli and Federico Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. EMS Textbooks in Mathematics. EMS Press, Berlin, 2023. Second edition.
- [2] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, NJ, 1970.
- [3] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

1 The Monge–Kantorovich problem

1.1 The Monge problem

The theory of optimal transport goes back to the work of Monge (Mémoire sur la théorie des déblais et des remblais, 1781). The basic idea is to determine the most efficient way to move one distribution of mass onto another while minimizing a cost function that measures the “effort” of transportation. Imagine that you have a sandpile and a hole at a distance away. The question is:

How can you move the sand to fill the hole while doing the least amount of work?



Each grain of sand can be moved from where it is to a new position, but moving it farther costs more effort. The problem asks for a map T that tells every grain of sand where to go. Your plan is then to minimise the total transportation cost.

In modern terms, the sandpile is represented by some probability measure $\mu \in \mathcal{M}_1(E)$, and the hole by another probability measure $\nu \in \mathcal{M}_1(F)$. Here and throughout, E and F are assumed to be Polish spaces, equipped with their Borel σ -algebras.

Definition 1.1. Let $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$. A measurable map $T: E \rightarrow F$ is called a **transport map** from μ to ν if

$$\nu = T_{\#}\mu,$$

i.e. if ν is the pushforward of μ by T .

The **cost** of transportation is then given by some measurable function $c: E \times F \rightarrow [0, \infty]$, which is typically assumed to be lower semicontinuous (see Appendix A for definition and basic properties of lower semicontinuity). For instance on $\mathbb{R}^d$, the cost can be given by the Euclidean distance $c(x, y) = |x - y|$ for $x, y \in \mathbb{R}^d$. Another important choice on $\mathbb{R}^d$ is $c(x, y) = |x - y|^2$, which we call the **quadratic cost**. The quantity $c(x, T(x))$ represents the cost of moving one unit of mass x to $T(x)$. The total cost of the transport map T is thus $\int c(x, T(x))\mu(dx)$, and we seek to minimize this quantity among all transport maps.

Definition 1.2. Let $c: E \times F \rightarrow [0, \infty]$ be measurable. **Monge’s problem** for transporting $\mu \in \mathcal{M}_1(E)$ to $\nu \in \mathcal{M}_1(F)$ consists in finding the following infimum:

$$\mathcal{M}_c(\mu, \nu) := \inf_{T_{\#}\mu=\nu} \int_E c(x, T(x))\mu(dx),$$

where the infimum runs over transport maps T from μ to ν .

Monge’s formulation is an optimization problem with nonlinear constraints. As such, it is in general a hard problem.

Remark 1.3. It might exist no transport map between μ and ν and in that case $\mathcal{M}_c(\mu, \nu) = \infty$. For example, if $\mu = \delta_x$ for some $x \in E$, then for any measurable $T: E \rightarrow F$, $T_{\#}\mu = \delta_{T(x)}$. Indeed, for any $A \in \mathcal{B}(F)$, we have

$$\begin{aligned} T_{\#}\mu(A) &= \delta_x(T^{-1}(A)) \\ &= \begin{cases} 1, & \text{if } x \in T^{-1}(A), \\ 0, & \text{if not,} \end{cases} \\ &= \delta_{T(x)}(A). \end{aligned}$$

Thus, if ν is not a single Dirac mass, then there exists no transport map between μ and ν .

Exercise 1 (Monotone rearrangement). For μ a probability measure on $\mathbb{R}$, we denote by F_μ its cumulative distribution function, defined by $F_\mu(x) := \mu((-\infty, x])$ for $x \in \mathbb{R}$. The *quantile function* of μ is the function $q_\mu: (0, 1) \rightarrow \mathbb{R}$ defined by

$$\forall \alpha \in (0, 1), \quad q_\mu(\alpha) = \inf\{x \in \mathbb{R} : \alpha < F_\mu(x)\}.$$

1. Prove that q_μ is well-defined, non-decreasing and right-continuous.
2. For any $\mu \in \mathcal{M}_1(\mathbb{R})$, prove that q_μ is a transport map from the uniform measure on $(0, 1)$ to μ .
3. For any $\mu \in \mathcal{M}_1(\mathbb{R})$ without atoms, prove that F_μ is a transport map from μ to the uniform measure on $(0, 1)$.
4. Let $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$. If μ has no atoms, prove that $q_\nu \circ F_\mu$ is a transport map from μ to ν .

1.2 The Kantorovich relaxation

In [On the Translocation of Masses, 1942], Kantorovich proposed a relaxation of Monge's problem. The key idea is to allow "mass splitting", in contrast to Monge's formulation, where each point x is sent to a single destination $T(x)$.

Definition 1.4. Let $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$. A *coupling or transport plan* between μ and ν is a probability measure on $E \times F$ whose first marginal is equal to μ and second marginal equal to ν . We denote by $\Pi(\mu, \nu)$ the set of transport plans between μ and ν , that is:

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{M}_1(E \times F) \mid (p_E)_{\#}\pi = \mu \text{ and } (p_F)_{\#}\pi = \nu\},$$

where p_E and p_F are the canonical projections onto E and F respectively.

The term coupling is usually used in probability, whereas transport plan is used in the optimal transport literature. In probabilistic terms, a coupling between μ and ν is the joint distribution of a couple of random variables (X, Y) , such that $X \sim \mu$ and $Y \sim \nu$.

Remark 1.5. We did not specify the σ -algebra which equips $E \times F$: it could be the Borel σ -algebra of the product $\mathcal{B}(E \times F)$ or the product of the Borel σ -algebras $\mathcal{B}(E) \otimes \mathcal{B}(F)$. In general, it is more convenient to consider $\mathcal{B}(E) \otimes \mathcal{B}(F)$, on which for example the product measure $\mu \otimes \nu$ is well-defined. However, with our separability assumption on E and F , one has $\mathcal{B}(E \times F) = \mathcal{B}(E) \otimes \mathcal{B}(F)$, so we do not need to specify the choice.

Remark 1.6. We explain here why a coupling $\pi \in \Pi(\mu, \nu)$ provides a way to transport μ to ν . By the disintegration theorem (this uses that E and F are Polish), there exists a family $(\pi_x)_{x \in E}$ of probability measures on F such that, for any $B \in \mathcal{B}(F)$, $x \in E \mapsto \pi_x(B)$ is measurable, and

$$\pi(dx, dy) = \mu(dx)\pi_x(dy).$$

Furthermore, the measures π_x are unique for μ -a.e. $x \in E$. Then, the interpretation of the transport plan π is the following: the mass located at $x \in E$ is transported to F according to the distribution π_x (instead of being transported at a single point $T(x)$ as in Monge's problem). Note that, in probabilistic term, if $(X, Y) \sim \pi$, then $\pi_X(dy)$ is the conditional distribution of Y given X .

Example 1.7. Let $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$.

- The product measure $\mu \otimes \nu$ is a transport plan between μ and ν , which corresponds to the coupling (X, Y) where X and Y are independent.

In particular, the set of transport plans $\Pi(\mu, \nu)$ is never empty.

- Let T be a transport map between μ and ν and define

$$\gamma_T = (\text{Id} \times T)_\# \mu,$$

where $\text{Id} \times T: E \rightarrow E \times F$ is the map $x \mapsto (x, T(x))$. Then, for any $A \in \mathcal{B}(E)$,

$$\gamma_T(A \times F) = \mu((\text{Id} \times T)^{-1}(A \times F)) = \mu(A),$$

so $(p_E)_\# \gamma_T = \mu$. On the other hand, for any $B \in \mathcal{B}(F)$,

$$\gamma_T(E \times B) = \mu((\text{Id} \times T)^{-1}(E \times B)) = \mu(T^{-1}(B)) = \nu(B),$$

because $\nu = T_\# \mu$. So we get $(p_F)_\# \gamma_T = \nu$. Thus, γ_T is a transport plan between μ and ν . This proves that transport maps can be seen as a particular case of transport plans.

Definition 1.8. Let $c: E \times F \rightarrow [0, \infty]$ be measurable. **Kantorovich's problem** for transporting $\mu \in \mathcal{M}_1(E)$ to $\nu \in \mathcal{M}_1(F)$ consists in finding the following infimum:

$$\mathcal{K}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{E \times F} c(x, y) \pi(dx, dy).$$

In terms of random variables, we can rewrite

$$\mathcal{K}_c(\mu, \nu) = \inf_{\substack{X \sim \mu \\ Y \sim \nu}} \mathbb{E}[c(X, Y)].$$

Unlike Monge's problem, Kantorovich's problem is a linear optimization problem over a convex set of measures, which makes it much more tractable.

Definition 1.9. Let S be a vector space. A subset C of S is **convex** if, for all $x, y \in C$ and $t \in [0, 1]$, one has $tx + (1 - t)y \in C$.

Exercise 2. Let $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$. Prove that $\Pi(\mu, \nu)$ is convex.

We now compare both problems.

Proposition 1.10. For any $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$,

$$\mathcal{K}_c(\mu, \nu) \leq \mathcal{M}_c(\mu, \nu).$$

Moreover, if a minimizer for Kantorovich's problem can be written as $(\text{Id} \times T)_\# \mu$ for some transport map T between μ and ν , then T is a minimizer for Monge's problem and

$$\mathcal{K}_c(\mu, \nu) = \mathcal{M}_c(\mu, \nu).$$

Proof. Let T be a transport map between μ and ν . Consider the transport plan $\gamma_T = (\text{Id} \times T)_\# \mu$ (see Example 1.7). Then, by the change of variable formula for pushforward measures,

$$\int_{E \times F} c(x, y) \gamma_T(dx, dy) = \int_E c(x, T(x)) \mu(dx).$$

Therefore,

$$\mathcal{K}_c(\mu, \nu) \leq \int_E c(x, T(x)) \mu(dx),$$

and optimizing over transport maps, one obtains $\mathcal{K}_c(\mu, \nu) \leq \mathcal{M}_c(\mu, \nu)$.

Now assume that a minimizer for Kantorovich's problem can be written as $\gamma_T = (\text{Id} \times T)_\# \mu$ for some transport map T between μ and ν . Then,

$$\mathcal{K}_c(\mu, \nu) = \int_{E \times F} c(x, y) \gamma_T(dx, dy) = \int_E c(x, T(x)) \mu(dx) \geq \mathcal{M}_c(\mu, \nu),$$

so, using $\mathcal{K}_c(\mu, \nu) \leq \mathcal{M}_c(\mu, \nu)$, the inequality in the displayed equation above has to be an equality: this means that $\mathcal{K}_c(\mu, \nu) = \mathcal{M}_c(\mu, \nu)$ and T is a minimizer for Monge's problem. $\square$

1.3 Discrete example

Let $x_1, \dots, x_n \in E$ be pairwise distinct points and $y_1, \dots, y_n \in F$ be pairwise distinct points. Consider

$$\mu := \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \in \mathcal{M}_1(E) \quad \text{and} \quad \nu := \frac{1}{n} \sum_{j=1}^n \delta_{y_j} \in \mathcal{M}_1(F). \quad (1.1)$$

We discuss in this section the question of optimal transport from μ to ν

We first consider *Monge's problem*. For any measurable $T: E \rightarrow F$, one has

$$T_\# \mu = \frac{1}{n} \sum_{i=1}^n \delta_{T(x_i)},$$

therefore T is a transport map from μ to ν if and only if there exists a permutation $\sigma \in S_n$ such that, for all $i \in \{1, \dots, n\}$, $T(x_i) = y_{\sigma(i)}$. Furthermore, in that case, the cost of the transport map T is

$$\int_E c(x, T(x)) \mu(dx) = \frac{1}{n} \sum_{i=1}^n c(x_i, y_{\sigma(i)}).$$

Therefore, Monge's problem can be rewritten as follows:

$$\mathcal{M}_c(\mu, \nu) = \min_{\sigma \in S_n} \frac{1}{n} \sum_{i=1}^n c(x_i, y_{\sigma(i)}),$$

where the infimum becomes a minimum because the set S_n is finite.

We now consider *Kantorovich's problem*. A transport plan $\pi \in \Pi(\mu, \nu)$ between μ and ν takes the form

$$\pi = \frac{1}{n} \sum_{i,j=1}^n \pi_{ij} \delta_{(x_i, y_j)},$$

for some coefficients $\pi_{ij} \geq 0$. Furthermore, the fact that the first marginal of π equals μ implies that, for any $i \in \{1, \dots, n\}$,

$$\frac{1}{n} \sum_{j=1}^n \pi_{ij} = \pi(\{x_i\} \times F) = \mu(\{x_i\}) = \frac{1}{n},$$

so that $\sum_{j=1}^n \pi_{ij} = 1$. Likewise, the fact that the second marginal of π equals ν implies that, for any $j \in \{1, \dots, n\}$, $\sum_{i=1}^n \pi_{ij} = 1$. Therefore $(\pi_{ij})_{1 \leq i, j \leq n}$ is a *doubly stochastic matrix*. Let

$$\mathcal{B}_n := \{n \times n \text{ doubly stochastic matrices}\},$$

which is called the *Birkhoff polytope*. Then, the Kantorovich problem takes the form:

$$\begin{aligned} \mathcal{K}_c(\mu, \nu) &= \inf_{\pi \in \Pi(\mu, \nu)} \int_{E \times F} c(x, y) \pi(dx, dy) \\ &= \inf_{\pi \in \mathcal{B}_n} \frac{1}{n} \sum_{i, j=1}^n c(x_i, y_j) \pi_{ij}. \end{aligned}$$

Note that this is still a linear optimization problem over a convex set, which is now $\mathcal{B}_n$. We now recall the definition of the convex hull of a set.

Definition 1.11. Let S be a vector space and $A \subset S$. The *convex hull* of A , denoted by $\text{conv}(A)$ is the set of all convex combinations of points in A , that is

$$\text{conv}(A) = \left\{ \sum_{i=1}^k t_i a_i \mid n \geq 1, a_1, \dots, a_k \in A, t_1, \dots, t_k \geq 0 \text{ such that } \sum_{i=1}^k t_i = 1 \right\}.$$

In the case of $\mathcal{B}_n$, we accept without proof the following result.

Theorem 1.12 (Birkhoff–von Neumann theorem (1946)). The set $\mathcal{B}_n$ is the convex hull of the set of $n \times n$ permutation matrices.

Exercise 3. Let S be a vector space, $A \subset S$ and $C = \text{conv}(A)$. Let $f: S \rightarrow \mathbb{R}$ be a linear form. Assume that the infimum of f on C is attained (at some point of C). Then, prove that the infimum of f on C is attained at some point of A .

With this fact in hand, we can now prove that Monge's and Kantorovich's problems coincide for this example.

Proposition 1.13. Let μ and ν be defined as in (1.1). Then, $\mathcal{M}_c(\mu, \nu) = \mathcal{K}_c(\mu, \nu)$.

Proof. Recall that $\mathcal{K}_c(\mu, \nu) = \inf_{\pi \in \mathcal{B}_n} f(\pi)$, where f is the linear form

$$f: \pi \in \mathbb{R}^{n \times n} \mapsto \frac{1}{n} \sum_{i, j=1}^n c(x_i, y_j) \pi_{ij}.$$

This function is continuous (as a linear form on a finite-dimensional vector space). Let $\mathcal{P}_n$ be the set of $n \times n$ permutation matrices so that $\mathcal{B}_n = \text{conv}(\mathcal{P}_n)$ by Birkhoff–von Neumann theorem. Furthermore, $\mathcal{B}_n$ is a closed subset of the compact $[0, 1]^{n \times n}$, so $\mathcal{B}_n$ is compact and so the infimum of the continuous function f on $\mathcal{B}_n$ is attained. Therefore, by Exercise 3, the infimum of the continuous function f on $\mathcal{B}_n$ is attained on $\mathcal{P}_n$. This means that

$$\mathcal{K}_c(\mu, \nu) = \min_{\pi \in \mathcal{P}_n} \frac{1}{n} \sum_{i, j=1}^n c(x_i, y_j) \pi_{ij} = \min_{\sigma \in S_n} \frac{1}{n} \sum_{i, j=1}^n c(x_i, y_j) \mathbb{1}_{\sigma(i)=j} = \min_{\sigma \in S_n} \frac{1}{n} \sum_{i=1}^n c(x_i, y_{\sigma(i)}),$$

which is exactly the formula we obtained for $\mathcal{M}_c(\mu, \nu)$. $\square$

Now, we completely solve the problem in the case of the quadratic cost on $\mathbb{R}$, by obtaining an explicit value for the optimal transport cost.

Proposition 1.14. Assume $E = F = \mathbb{R}$, $x_1 < \dots < x_n$ and $y_1 < \dots < y_n$. Consider the quadratic cost $c(x, y) = |x - y|^2$ for $x, y \in \mathbb{R}$. Let μ and ν be defined as in (1.1). Then,

$$\mathcal{M}_c(\mu, \nu) = \mathcal{K}_c(\mu, \nu) = \frac{1}{n} \sum_{i=1}^n c(x_i, y_i).$$

Furthermore, any minimizing transport map T for Monge's problem has to satisfy $T(x_i) = y_i$ for each $i \in \{1, \dots, n\}$.

Proof. By the previous discussions, it is enough to prove that if $\sigma \in S_n$ is a minimizer of

$$\min_{\sigma \in S_n} \sum_{i=1}^n c(x_i, y_{\sigma(i)}),$$

then $\sigma = \text{Id}$. By contradiction, assume $\sigma \neq \text{Id}$. Then, there exists $j \in \{1, \dots, n\}$ such that $\sigma(j+1) < \sigma(j)$ because the identity is the unique increasing permutation. Then

$$\begin{aligned} c(x_j, y_{\sigma(j)}) + c(x_{j+1}, y_{\sigma(j+1)}) &= (x_j - y_{\sigma(j)})^2 + (x_{j+1} - y_{\sigma(j+1)})^2 \\ &= (x_j - y_{\sigma(j+1)})^2 + (x_{j+1} - y_{\sigma(j)})^2 + 2(x_{j+1} - x_j)(y_{\sigma(j)} - y_{\sigma(j+1)}) \\ &> (x_j - y_{\sigma(j+1)})^2 + (x_{j+1} - y_{\sigma(j)})^2, \end{aligned}$$

because $x_{j+1} - x_j > 0$ and $y_{\sigma(j)} - y_{\sigma(j+1)} > 0$. Denote by $\psi \in S_n$ the transposition of j and $j+1$, i.e. $\psi(j) = j+1$, $\psi(j+1) = j$ and $\psi(i) = i$ for $i \notin \{j, j+1\}$. Then, we proved that

$$c(x_j, y_{\sigma(j)}) + c(x_{j+1}, y_{\sigma(j+1)}) > c(x_j, y_{\sigma \circ \psi(j)}) + c(x_{j+1}, y_{\sigma \circ \psi(j+1)}).$$

Since for $i \notin \{j, j+1\}$, $c(x_i, y_{\sigma(i)}) = c(x_i, y_{\sigma \circ \psi(i)})$, this yields

$$\sum_{i=1}^n c(x_i, y_{\sigma(i)}) > \sum_{i=1}^n c(x_i, y_{\sigma \circ \psi(i)}),$$

contradicting the minimality of σ . □

Exercise 4. We consider the framework of Proposition 1.14 but with the more general cost $c(x, y) = |x - y|^q$ for $q > 0$.

1. If $q > 1$, prove that the conclusions of Proposition 1.14 still hold.
2. If $q = 1$, prove that the formula for $\mathcal{M}_c(\mu, \nu)$ still holds, but that there can exist minimizing transport map T for Monge's problem such that $T(x_i) \neq y_i$ for some $i \in \{1, \dots, n\}$.
3. If $q < 1$, prove that the formula for $\mathcal{M}_c(\mu, \nu)$ do not necessarily hold.

2 Analysis of the Kantorovich's problem

In this section, we study further Kantorovich's problem by proving the existence of minimizers and proving geometric properties for the support of those minimizers. We pursue the analysis further in the case of a quadratic cost by proving a dual formulation for Kantorovich's problem and by using it to prove uniqueness of the minimizer under general assumption. Recall E and F are assumed to be Polish spaces.

2.1 Existence of minimizers

For $\pi \in \mathcal{M}_1(E \times F)$, we denote by $I(\pi)$ the cost of the transport plan π , i.e.

$$I(\pi) := \int_{E \times F} c(x, y) \pi(dx, dy),$$

so that $\mathcal{K}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} I(\pi)$ for $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$. In this subsection, we prove that this infimum is attained when E and F are Polish.

Theorem 2.1. *Let $c: E \times F \rightarrow [0, \infty]$ be lower semicontinuous. Let $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$. There exists $\pi^* \in \Pi(\mu, \nu)$, such that*

$$I(\pi^*) = \inf_{\pi \in \Pi(\mu, \nu)} I(\pi).$$

Such a minimizer π^* is called an *optimal* transport plan between μ and ν . Note that in general, it is not unique.

Proof. As usual $\mathcal{M}_1(E \times F)$ is equipped with the weak convergence, which is metrizable because we assumed that E and F are separable so $E \times F$ is separable as well. The strategy consists in showing that the set $\Pi(\mu, \nu) \subset \mathcal{M}_1(E \times F)$ is compact (by proving that it is closed and relatively compact) and that I is lower semicontinuous. Since a lower semicontinuous function on a compact set attains its infimum (see Proposition A.4), this is enough to prove the result.

Step 1: $\Pi(\mu, \nu)$ is closed. Let $(\pi_n)_{n \geq 0} \in \Pi(\mu, \nu)^\mathbb{N}$ converging weakly to some $\pi \in \mathcal{M}_1(E \times F)$. Then, for all $g \in \mathcal{C}_b(E \times F)$, the function $(x, y) \in E \times F \mapsto g(x)$ is in $\mathcal{C}_b(E \times F)$, so we have

$$\int_E g d\mu = \int_{E \times F} g(x) d\pi_n(x, y) \xrightarrow{n \rightarrow \infty} \int_{E \times F} g(x) d\pi(x, y),$$

hence

$$\int_E g d\mu = \int_{E \times F} g(x) d\pi(x, y),$$

which implies that $(p_E)_\# \pi = \mu$. Similarly, we have $(p_F)_\# \pi = \nu$. Hence, $\pi \in \Pi(\mu, \nu)$, so $\Pi(\mu, \nu)$ is closed.

Step 2: $\Pi(\mu, \nu)$ is relatively compact. By Prokhorov's theorem (and the fact that $\mathcal{M}_1(E \times F)$ is metric), it is enough to prove that $\Pi(\mu, \nu)$ is tight. Let $\varepsilon > 0$. Since E and F are Polish spaces, μ and ν are tight, hence there exists a compact set $K_E \subset E$ and a compact set $K_F \subset F$ such that

$$\mu(K_E^c) \leq \varepsilon/2 \quad \text{and} \quad \nu(K_F^c) \leq \varepsilon/2.$$

Set $K := K_E \times K_F$. Then K is compact as a product of compact spaces, and for all $\pi \in \Pi(\mu, \nu)$,

$$\begin{aligned} \pi(K^c) &= \pi((K_E \times K_F)^c) \\ &= \pi((K_E^c \times F) \cup (E \times K_F^c)) \\ &\leq \pi(K_E^c \times F) + \pi(E \times K_F^c) \\ &= \mu(K_E^c) + \nu(K_F^c) \\ &\leq \varepsilon. \end{aligned}$$

Hence, $\Pi(\mu, \nu)$ is relatively compact.

Step 3: I is lower semicontinuous on $\mathcal{M}_1(E \times F)$. Let $(\pi_n)_{n \geq 0} \in \mathcal{M}_1(E \times F)^\mathbb{N}$ converging weakly to some $\pi \in \mathcal{M}_1(E \times F)$. We aim at proving

$$I(\pi) \leq \liminf_{n \rightarrow \infty} I(\pi_n).$$

By assumption, c is a lower semicontinuous function, hence (see Proposition A.5) there exists a nondecreasing sequence of bounded and uniformly continuous functions $(c_k)_{k \geq 0}$ from $E \times F$ to

$[0, \infty)$ such that $c_k \nearrow c$ as $k \rightarrow \infty$. Then, by the weak convergence of $(\pi_n)_{n \geq 0}$ towards π , for any $k \geq 1$, we have

$$\int c_k d\pi = \lim_{n \rightarrow \infty} \int c_k d\pi_n \leq \liminf_{n \rightarrow \infty} \int c d\pi_n,$$

because $c_k \leq c$. But using monotone convergence theorem, we also have

$$\lim_{k \rightarrow \infty} \int c_k d\pi = \int c d\pi,$$

hence we get

$$\int c d\pi \leq \liminf_{n \rightarrow \infty} \int c d\pi_n,$$

which proves $I(\pi) \leq \liminf_{n \rightarrow \infty} I(\pi_n)$ as desired. This concludes the proof. $\square$

Exercise 5. On $\mathbb{R}^2 \times \mathbb{R}^2$, consider the cost function $c(x, y) = |x - y|^2$, where $|\cdot|$ denotes the Euclidean norm. Let

$$x_1 = (0, 0), \quad x_2 = (1, 1), \quad y_1 = (1, 0), \quad y_2 = (0, 1),$$

and consider the probability measures

$$\mu = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2} \quad \text{and} \quad \nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}.$$

Show that

$$\Pi(\mu, \nu) = \left\{ \pi_\alpha = \alpha\delta_{(x_1, y_1)} + \left(\frac{1}{2} - \alpha\right)\alpha\delta_{(x_1, y_2)} + \alpha\delta_{(x_2, y_2)} + \left(\frac{1}{2} - \alpha\right)\delta_{(x_2, y_1)} \mid \alpha \in \left[0, \frac{1}{2}\right] \right\},$$

and show that

$$\int_{\mathbb{R}^2 \times \mathbb{R}^2} c d\pi_\alpha = 1, \quad \text{for all } \alpha \in \left[0, \frac{1}{2}\right].$$

In particular, there is no unicity in general for the minimizer of the functional I .

Having established the existence of optimal transport plans, our next goal is to investigate the geometric conditions characterizing optimality, through the notion of c -cyclical monotonicity.

2.2 c -cyclical monotonicity

2.2.1 General cost functions

Definition 2.2. Let $\mu \in \mathcal{M}_1(E)$. The **support** of μ is the (closed) set

$$\text{supp } \mu = \{x \in E \mid \forall \varepsilon > 0, \mu(B(x, \varepsilon)) > 0\}.$$

Hence, the support of μ is the largest closed subset of E for which every open neighbourhood of every point of the set has positive measure. For instance, the support of Lebesgue measure on $\mathbb{R}$ is $\mathbb{R}$, and the support of δ_x is $\{x\}$.

We will now characterize the support of an optimal transport plan.

Recall that in the problem of transporting the discrete measure $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ to the measure $\nu = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$, we have found that for the quadratic cost, the minimum of

$$\sum_{i=1}^n |x_i - y_{\sigma(i)}|^2$$

over $\sigma \in \mathcal{S}_n$ is attained when the two supports of μ and ν satisfy

$$x_1 < \cdots < x_n \quad \text{and} \quad y_1 < \cdots < y_n.$$

Hence, locally, any "reshuffling" of indices will increase the cost. For instance, for three points, we have

$$\sum_{i=1}^3 |x_i - y_i|^2 \leq \sum_{i=1}^3 |x_{i+1} - y_i|^2,$$

with the convention that $x_4 = x_1$.

This motivates the following definition.

Definition 2.3. A subset $\Gamma \subset E \times F$ is said to be *c-cyclically monotone* if for any finite sequence of points $(x_i, y_i)_{i=1, \dots, n}$ in Γ , we have

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

with the convention that $x_{n+1} = x_1$.

We have then the following.

Theorem 2.4. Consider a continuous cost c and let $\pi \in \Pi(\mu, \nu)$ be an optimal transport plan. Then, $\text{supp } \pi$ is *c-cyclically monotone*.

Proof. By contradiction, suppose that $\text{supp } \pi$ is not *c-cyclically monotone*. Then, there exists $\varepsilon > 0$ and n points $(x_1, y_1), \dots, (x_n, y_n)$ in $\text{supp } \pi$ such that

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

By continuity of c , for all $i \in \{1, \dots, n\}$, there exists an open neighbourhood U_i of x_i and an open neighbourhood V_i of y_i such that

$$|c(x, y) - c(x_i, y_i)| \leq \frac{\varepsilon}{4n}, \quad \forall (x, y) \in U_i \times V_i,$$

and

$$|c(x, y) - c(x_{i+1}, y_i)| \leq \frac{\varepsilon}{4n}, \quad \forall (x, y) \in U_{i+1} \times V_i.$$

Let $\alpha_i = \pi(U_i \times V_i)$. Since (x_i, y_i) belongs to $\text{supp } \pi$, we have $\alpha_i > 0$ for all i . Let $\alpha = \min(\alpha_1, \dots, \alpha_n)$ and define π_i the probability measure on $E \times F$ defined by

$$\pi_i(A) = \frac{1}{\alpha_i} \pi(A \cap (U_i \times V_i)),$$

for all Borel set $A \subset E \times F$, that π_i is the restriction of π to $U_i \times V_i$. Define also $\mu_i = (p_E)_\# \pi_i \in \mathcal{M}_1(E)$ and $\nu_i = (p_F)_\# \pi_i \in \mathcal{M}_1(F)$. Finally, define

$$\pi' = \pi - \frac{\alpha}{n} \sum_{i=1}^n \pi_i + \frac{\alpha}{n} \sum_{i=1}^n \mu_{i+1} \otimes \nu_i.$$

Then π' is a probability measure: it is a measure as a sum of measures, and we have

$$\begin{aligned} \pi' &\geq \pi - \frac{\alpha}{n} \sum_{i=1}^n \pi_i \\ &\geq \pi - \frac{\alpha}{n} \sum_{i=1}^n \frac{1}{\alpha_i} \pi|_{U_i \times V_i} \\ &\geq \pi - \frac{1}{n} \sum_{i=1}^n \pi|_{U_i \times V_i} \\ &\geq \pi - \frac{1}{n} \sum_{i=1}^n \pi = 0, \end{aligned}$$

so π' is a nonnegative measure. The mass of π' is clearly 1. Moreover, it is easy to see that $\pi' \in \Pi(\mu, \nu)$ by linearity.

We have, as $\mu_{i+1} \otimes \nu_i \in \mathcal{M}_1(U_{i+1} \times V_i)$,

$$\begin{aligned} \int_{E \times F} c d(\mu_{i+1} \otimes \nu_i) &= \int_{U_{i+1} \times V_i} c d(\mu_{i+1} \otimes \nu_i) \\ &\leq \int_{U_{i+1} \times V_i} \left(c(x_{i+1}, y_i) + \frac{\varepsilon}{4n} \right) d(\mu_{i+1} \otimes \nu_i) \\ &= c(x_{i+1}, y_i) + \frac{\varepsilon}{4n}. \end{aligned}$$

As $\pi_i \in \mathcal{M}_1(U_i \times V_i)$, we also have

$$\begin{aligned} \int_{E \times F} c d\pi_i &= \int_{U_i \times V_i} c d\pi_i \\ &\geq \int_{U_{i+1} \times V_i} \left(c(x_i, y_i) - \frac{\varepsilon}{4n} \right) d\pi_i \\ &= c(x_i, y_i) - \frac{\varepsilon}{4n}. \end{aligned}$$

Finally, we get

$$\begin{aligned} \int_{E \times F} c d\pi - \int_{E \times F} c d\pi' &= \frac{\alpha}{n} \sum_{i=1}^n \left(\int_{E \times F} c d\pi_i - \int_{E \times F} c d(\mu_{i+1} \otimes \nu_i) \right) \\ &\geq \frac{\alpha}{n} \sum_{i=1}^n \left(c(x_i, y_i) - c(x_{i+1}, y_i) - \frac{\varepsilon}{2n} \right) \\ &\geq \frac{\alpha}{n} \varepsilon - \frac{\alpha \varepsilon}{n 2} \\ &= \frac{\alpha \varepsilon}{n 2} \\ &> 0, \end{aligned}$$

where we have used our assumption that $\text{supp } \pi$ is not c -cyclically monotone. Thus $\int c d\pi' < \int c d\pi$, which contradicts the optimality of π . $\square$

2.2.2 The case of quadratic cost

For the quadratic cost $c(x, y) = |x - y|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$, the c -cyclical monotonicity of a subset $\Gamma \subset \mathbb{R}^d \times \mathbb{R}^d$ is equivalent to (using $|x - y|^2 = |x|^2 + |y|^2 - 2\langle x, y \rangle$)

$$\sum_{i=1}^n \langle y_i, x_{i+1} - x_i \rangle \leq 0,$$

for any finite sequence $(x_i, y_i)_{i=1, \dots, n} \subset \Gamma$. We will say in that case that Γ is *cyclically monotone* (i.e. we drop the c).

Before giving a characterization of cyclically monotone sets in $\mathbb{R}^d \times \mathbb{R}^d$, we collect a few results on convex functions defined on $\mathbb{R}^d$.

Let $\varphi: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper convex function. We recall that φ *proper* means that $\varphi \not\equiv +\infty$, so the domain of φ , i.e. $D_\varphi = \{x \mid \varphi(x) < \infty\}$, is non-empty. Recall that φ is *convex* if for all $x, y \in \mathbb{R}^d$, for all $t \in [0, 1]$,

$$\varphi(tx + (1-t)y) \leq t\varphi(x) + (1-t)\varphi(y),$$

and that a subset $C \subset E$ is called *convex* if for all $x, y \in C$, the line segment $[x, y]$ is included in C , i.e.

$$\forall x, y \in C, \forall t \in [0, 1], \quad tx + (1 - t)y \in C.$$

Note that the domain D_φ of a convex function is then a convex set (it can be open, closed or neither), and thus $\lambda(\partial D_\varphi) = 0$, where λ denotes Lebesgue measure on $\mathbb{R}^d$ by the following lemma:

Lemme 2.5. *Let $C \subset \mathbb{R}^d$ be a convex set. Then, $\lambda(\partial C) = 0$.*

Proof. If $\mathring{C} = \emptyset$, then C lies in an affine subspace of dimension strictly less than d , and thus has zero Lebesgue measure.

Suppose that $\mathring{C} \neq \emptyset$. By invariance by translation of Lebesgue measure, one can suppose that $0 \in \mathring{C}$. Moreover, by intersecting with the ball centered at 0 with radius R and letting $R \uparrow +\infty$, one can suppose that C is bounded. Let $t \in (0, 1)$. We claim that

$$\partial C \subset \frac{1}{t}\mathring{C}.$$

Indeed, since $0 \in \mathring{C}$, there exists $r > 0$ such that $B(0, r) \subset C$. By convexity of C , for all $q \in C$, for all $t \in (0, 1)$, for all $x \in B(0, r)$,

$$tq + (1 - t)x \in C,$$

i.e.

$$B(tq, (1 - t)r) \subset C.$$

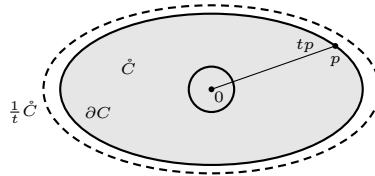
Now let $p \in \partial C$ and let $(p_n)_n \subset C$ such that $p_n \rightarrow p$. Hence, for all n , we have

$$B(tp_n, (1 - t)r) \subset C.$$

As $p_n \rightarrow p$, for n large enough, we have that $tp \in B(tp_n, (1 - t)r/2)$ and thus

$$B(tp, (1 - t)r/2) \subset B(tp_n, (1 - t)r).$$

Hence, $B(tp, (1 - t)r/2) \subset C$, so $tp \in \mathring{C}$, that is $\partial C \subset \frac{1}{t}\mathring{C}$, which proves the claim.



Note that by convexity of $\mathring{C}$ (exercise), we also have that $\mathring{C} \subset \frac{1}{t}\mathring{C}$, since for all $t \in (0, 1)$, and all $q \in \mathring{C}$, $tq + (1 - t)0 = tq \in \mathring{C}$.

Since the interior and the boundary of any set are disjoint, we get that

$$\partial C \subset \frac{1}{t}\mathring{C} \setminus \mathring{C},$$

and thus

$$\begin{aligned} \lambda(\partial C) &\leq \lambda\left(\frac{1}{t}\mathring{C} \setminus \mathring{C}\right) \\ &= \lambda\left(\frac{1}{t}\mathring{C}\right) - \lambda(\mathring{C}) \\ &= \frac{1}{t^d}\lambda(\mathring{C}) - \lambda(\mathring{C}). \end{aligned}$$

Letting $t \uparrow 1$ gives that $\lambda(\partial C) = 0$. □

A proper convex function φ is continuous and locally Lipschitz on $\mathring{D}_\varphi$, hence by Rademacher's theorem, φ is differentiable λ -almost everywhere.

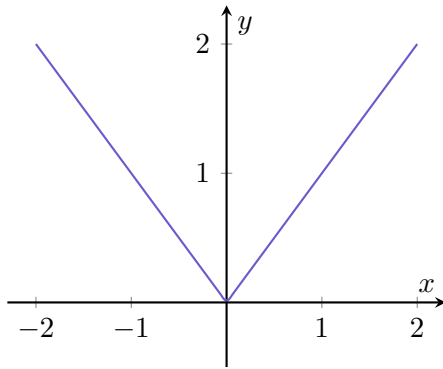
If φ is differentiable at x , then, for all $z \in \mathbb{R}^d$,

$$\varphi(z) \geq \varphi(x) + \langle \nabla \varphi(x), z - x \rangle,$$

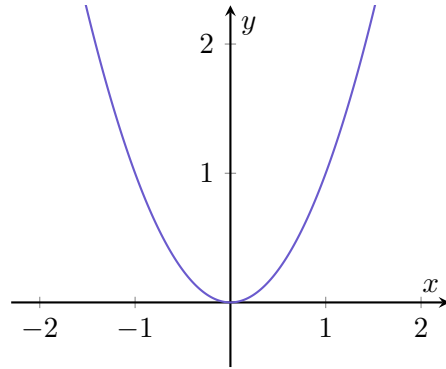
where $\nabla \varphi(x)$ is the gradient of φ at x . It says that the graph of φ lies above its tangent at x . When φ is not differentiable, we can generalize this idea by introducing the *subdifferential* $\partial\varphi(x)$ at x , which is the set:

$$\partial\varphi(x) = \left\{ y \in \mathbb{R}^d \mid \varphi(z) \geq \varphi(x) + \langle y, z - x \rangle, \text{ for all } z \in \mathbb{R}^d \right\}.$$

Thus, $\partial\varphi: x \mapsto \partial\varphi(x)$ is a set-valued function, and when φ is differentiable at x , one has $\partial\varphi(x) = \{\nabla\varphi(x)\}$. One can compare the two convex functions $|x|$ and x^2 to understand the difference (for $\varphi(x) = |x|$, one has $\partial\varphi(0) = [-1, 1]$).



(a) $y = |x|$



(b) $y = x^2$

We also define the *subdifferential* of φ as the following subset of $\mathbb{R}^d \times \mathbb{R}^d$:

$$\partial\varphi = \bigcup_{x \in \mathbb{R}^d} \{x\} \times \partial\varphi(x).$$

The *Legendre transform* (or convex conjugate) of a proper function φ is:

$$\varphi^*(y) = \sup_{x \in \mathbb{R}^n} (\langle x, y \rangle - \varphi(x)), \quad \text{for all } y \in \mathbb{R}^d.$$

From a geometric interpretation, the Legendre transform φ^* describes the family of all affine functions that lie below φ . For a given slope $y \in \mathbb{R}^d$, the best affine function of slope y that lies below φ is $x \mapsto \langle y, x \rangle - \varphi^*(y)$. For example, the Legendre transform of $x \mapsto \frac{1}{2}x^2$ is $y \mapsto \frac{1}{2}y^2$, while the Legendre transform of $x \mapsto |x|$ is the convex indicator function of $[-1, 1]$, i.e. $\infty \cdot \mathbb{1}_{[-1,1]^c}$.

Note that obviously, for all $x, y \in \mathbb{R}^d$, one has

$$\langle x, y \rangle \leq \varphi(x) + \varphi^*(y).$$

Note also that the Legendre transform of a proper convex function is a proper convex lower semicontinuous function (as the supremum of a family of affine functions).

We now give a characterization for a point to be in the subdifferential of a convex function.

Lemme 2.6. *Let φ be a proper convex function on $\mathbb{R}^d$. Then, for all $x, y \in \mathbb{R}^d$,*

$$\langle x, y \rangle = \varphi(x) + \varphi^*(y) \Leftrightarrow y \in \partial\varphi(x).$$

Proof. We already know that for all $x, y \in \mathbb{R}^d$, $\langle x, y \rangle \leq \varphi(x) + \varphi^*(y)$. Hence,

$$\begin{aligned} \langle x, y \rangle = \varphi(x) + \varphi^*(y) &\Leftrightarrow \langle x, y \rangle \geq \varphi(x) + \varphi^*(y) \\ &\Leftrightarrow \forall z \in \mathbb{R}^d, \langle x, y \rangle \geq \varphi(x) + \langle z, y \rangle - \varphi(z) \\ &\Leftrightarrow \forall z \in \mathbb{R}^d, \varphi(z) \geq \varphi(x) + \langle y, z - x \rangle \\ &\Leftrightarrow y \in \partial\varphi(x), \end{aligned}$$

hence proving the claimed equivalence. $\square$

We have then the following characterization of cyclically monotone subsets of $\mathbb{R}^d \times \mathbb{R}^d$.

Theorem 2.7. *A set $\Gamma \subset \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone if and only if there exists a proper convex lower semicontinuous function $\varphi: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ such that $\Gamma \subset \partial\varphi$.*

Proof. ($\Leftarrow$) Suppose that $\Gamma \subset \partial\varphi$ for some convex function φ . Let $(x_i, y_i)_{i=1, \dots, n} \subset \Gamma$. Then, for all $i \in \{1, \dots, n\}$, $y_i \in \partial\varphi(x_i)$, therefore,

$$\forall z \in \mathbb{R}^d, \varphi(z) \geq \varphi(x_i) + \langle y_i, z - x_i \rangle.$$

In particular, for $z = x_{i+1}$,

$$\varphi(x_{i+1}) \geq \varphi(x_i) + \langle y_i, x_{i+1} - x_i \rangle.$$

Summing these inequalities over i gives that

$$\sum_{i=1}^n \varphi(x_{i+1}) \geq \sum_{i=1}^n \varphi(x_i) + \sum_{i=1}^n \langle y_i, x_{i+1} - x_i \rangle.$$

But the first two sums being equal, one obtains that

$$\sum_{i=1}^n \langle y_i, x_{i+1} - x_i \rangle \leq 0,$$

i.e. Γ is cyclically monotone.

($\Rightarrow$) Assume Γ is cyclically monotone. We will construct φ explicitly. Suppose that we have some convex function φ with $\Gamma \subset \partial\varphi$. Let $(x_0, y_0) \in \Gamma$, and suppose that $\varphi(x_0) = 0$. By induction, for all $(x_i, y_i)_{i=1, \dots, n} \subset \Gamma$, we have that for all $x \in \mathbb{R}^d$,

$$\varphi(x) \geq \langle y_n, x - x_n \rangle + \langle y_{n-1}, x_n - x_{n-1} \rangle + \dots + \langle y_0, x_1 - x_0 \rangle.$$

Indeed, $(x_0, y_0) \in \partial\varphi$, hence for all x ,

$$\varphi(x) \geq \varphi(x_0) + \langle y_0, x - x_0 \rangle = \langle y_0, x - x_0 \rangle,$$

since $\varphi(x_0) = 0$, so the base case is proven. Suppose that for all $(x_i, y_i)_{i=1, \dots, n} \subset \Gamma$, we have that for all $x \in \mathbb{R}^d$,

$$\varphi(x) \geq \langle y_n, x - x_n \rangle + \langle y_{n-1}, x_n - x_{n-1} \rangle + \dots + \langle y_0, x_1 - x_0 \rangle.$$

Let $(x_{n+1}, y_{n+1}) \in \Gamma \subset \partial\varphi$. Then, for all $x \in \mathbb{R}^d$,

$$\varphi(x) \geq \varphi(x_{n+1}) + \langle y_{n+1}, x - x_{n+1} \rangle.$$

Applying the induction hypothesis to $x = x_{n+1}$ gives the result.

Hence, we define, for all $x \in \mathbb{R}^d$,

$$\varphi(x) = \sup_{\substack{n \geq 1 \\ (x_i, y_i)_{i=1, \dots, n} \subset \Gamma}} \{ \langle y_n, x - x_n \rangle + \langle y_{n-1}, x_n - x_{n-1} \rangle + \dots + \langle y_0, x_1 - x_0 \rangle \}.$$

Then we have that φ is a convex lower semicontinuous function as a supremum of affine (and thus convex) functions. Choosing $n = 1$ and $(x_1, y_1) = (x_0, y_0)$ yield that

$$\varphi(x) \geq \langle y_0, x - x_0 \rangle,$$

for all x , hence $\varphi(x_0) \geq 0$. Moreover, by cyclic monotonicity, for all $(x_i, y_i)_{i=1, \dots, n} \subset \Gamma$, we have

$$\langle y_n, x_0 - x_n \rangle + \langle y_{n-1}, x_n - x_{n-1} \rangle + \dots + \langle y_0, x_1 - x_0 \rangle \leq 0,$$

hence $\varphi(x_0) \leq 0$. Finally $\varphi(x_0) = 0$, so φ is a proper function. It remains to prove that $\Gamma \subset \partial\varphi$. Let $(x', y') \in \Gamma$ and let $\alpha < \varphi(x')$. Hence, there exists n and $(x_i, y_i)_{i=1, \dots, n} \subset \Gamma$ such that

$$\alpha \leq \langle y_n, x' - x_n \rangle + \dots + \langle y_0, x_1 - x_0 \rangle.$$

Now consider the sequence $(x_i, y_i)_{i=1, \dots, n+1} \subset \Gamma$ with $(x_{n+1}, y_{n+1}) = (x', y')$. We get that, for all $z \in \mathbb{R}^d$,

$$\begin{aligned} \varphi(z) &\geq \langle y_{n+1}, z - x_{n+1} \rangle + \langle y_n, x_{n+1} - x_n \rangle + \dots + \langle y_0, x_1 - x_0 \rangle \\ &\geq \langle y', z - x_{n+1} \rangle + \alpha. \end{aligned}$$

Letting $\alpha \uparrow \varphi(x')$ gives that for all $z \in \mathbb{R}^d$,

$$\varphi(z) \geq \varphi(x') + \langle y', z - x_{n+1} \rangle,$$

i.e. $y' \in \partial\varphi(x')$, hence $(x', y') \in \partial\varphi$. □

2.3 Kantorovich duality

Let E and F be Polish spaces. Let $\mu \in \mathcal{M}_1(E)$ and $\nu \in \mathcal{M}_1(F)$. Let $c: E \times F \rightarrow [0, +\infty]$ be a measurable function. We assume that c is lower semicontinuous (see Appendix A).

Let $\pi \in \Pi(\mu, \nu)$ be a transport plan between μ and ν and define:

$$I(\pi) := \int_{E \times F} c(x, y) \pi(dx, dy).$$

For $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$, define:

$$J(\varphi, \psi) := \int_E \varphi d\mu + \int_F \psi d\nu.$$

Denote

$$\begin{aligned} \Phi_c &= \{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) \mid \varphi(x) + \psi(y) \leq c(x, y) \\ &\quad \text{for } \mu\text{-almost all } x \in E \text{ and } \nu\text{-almost all } y \in F\}. \end{aligned}$$

We denote by $\Phi_c \cap \mathcal{C}_b$ the same set than above with $L^1(\mu) \times L^1(\nu)$ replaced by $\mathcal{C}_b(E) \times \mathcal{C}_b(F)$. Note that $\Phi_c \cap \mathcal{C}_b \subset \Phi_c$.

Theorem 2.8 (Kantorovich duality). *We have*

$$\inf_{\pi \in \Pi(\mu, \nu)} I(\pi) = \sup_{(\varphi, \psi) \in \Phi_c} J(\varphi, \psi) = \sup_{(\varphi, \psi) \in \Phi_c \cap \mathcal{C}_b} J(\varphi, \psi).$$

In this section, we only prove the Kantorovich duality in the case of a quadratic cost $c(x, y) = \frac{1}{2}|x - y|^2$ defined on $\mathbb{R}^d \times \mathbb{R}^d$, where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$.

Let μ and ν be probability measures on $\mathbb{R}^d$ with finite second moments:

$$\int_{\mathbb{R}^d} |x|^2 \mu(dx) < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} |x|^2 \nu(dx) < \infty.$$

Define

$$M := \frac{1}{2} \int_{\mathbb{R}^d} |x|^2 \mu(dx) + \frac{1}{2} \int_{\mathbb{R}^d} |y|^2 \nu(dy) < \infty.$$

Since $|x - y|^2 = |x|^2 + |y|^2 - 2\langle x, y \rangle$, we have

$$I(\pi) = \frac{1}{2} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \pi(dx, dy) = M - \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle \pi(dx, dy).$$

Note that an optimal transport plan π^* is then also optimal for the cost $-\langle x, y \rangle$, i.e. if

$$\min_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2} |x - y|^2 d\pi(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2} |x - y|^2 d\pi^*(x, y),$$

then

$$\max_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi^*(x, y).$$

Now let $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$, such that for μ -almost all $x \in \mathbb{R}^d$ and ν -almost all $y \in \mathbb{R}^d$,

$$\varphi(x) + \psi(y) \leq \frac{1}{2} |x - y|^2.$$

Define, for almost all $x, y \in \mathbb{R}^d$,

$$\begin{aligned} \tilde{\varphi}(x) &= \frac{1}{2} |x|^2 - \varphi(x) \\ \tilde{\psi}(y) &= \frac{1}{2} |y|^2 - \psi(y). \end{aligned}$$

Since μ and ν have finite second moment, we have $(\tilde{\varphi}, \tilde{\psi}) \in L^1(\mu) \times L^1(\nu)$, and for almost all x, y ,

$$\begin{aligned} \tilde{\varphi}(x) + \tilde{\psi}(y) &= \frac{1}{2} |x|^2 + \frac{1}{2} |y|^2 - \varphi(x) - \psi(y) \\ &\geq \frac{1}{2} |x|^2 + \frac{1}{2} |y|^2 - \frac{1}{2} |x - y|^2 = \langle x, y \rangle. \end{aligned}$$

Redefining $\tilde{\varphi}$ and $\tilde{\psi}$ to be $+\infty$ on μ -negligible and ν -negligible sets respectively, we can assume that the above inequality holds true for all $x, y \in \mathbb{R}^d$ without changing the values of the integrals of $\tilde{\varphi}$ and $\tilde{\psi}$. Moreover, we have

$$J(\tilde{\varphi}, \tilde{\psi}) = M - J(\varphi, \psi).$$

Hence, Kantorovich duality takes the following form:

Theorem 2.9. *Let μ and ν be two probability measures on $\mathbb{R}^d$ with finite second moments. We have,*

$$\max_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y) = \inf_{\varphi(x) + \psi(y) \geq \langle x, y \rangle} J(\varphi, \psi).$$

Moreover, the above infimum is attained at a pair (φ, φ^*) , where φ is a proper convex function. Any pair at which the infimum is attained is called a pair of optimal Kantorovich potentials.

Proof. Let $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ such that

$$\varphi(x) + \psi(y) \geq \langle x, y \rangle,$$

for all $x, y \in \mathbb{R}^d$. Integrating over π gives that

$$\begin{aligned} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y) &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(x) d\pi(x, y) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \psi(y) d\pi(x, y) \\ &= \int_{\mathbb{R}^d} \varphi(x) d\mu(x) + \int_{\mathbb{R}^d} \psi(y) d\nu(y). \end{aligned}$$

Taking the maximum over transport plans and the infimum over (φ, ψ) gives the inequality

$$\max_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y) \leq \inf_{\varphi(x) + \psi(y) \geq \langle x, y \rangle} J(\varphi, \psi).$$

Now let π be an optimal transport plan, i.e.

$$\max_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\gamma(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y).$$

The support of π is then cyclically monotone, hence there exists a proper convex function φ such that

$$\text{supp } \pi \subset \partial\varphi.$$

Thus, by the subdifferential characterization lemma, we get that for π -almost all x, y ,

$$\langle x, y \rangle = \varphi(x) + \varphi^*(y).$$

Integrating over π (φ and φ^* being proper convex functions, they are bounded below by some affine functions, hence their integrals are well defined), one gets that,

$$\begin{aligned} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle d\pi(x, y) &= \int_{\mathbb{R}^d} (\varphi(x) + \varphi^*(y)) d\pi(x, y) \\ &= \int_{\mathbb{R}^d} \varphi(x) d\mu(x) + \int_{\mathbb{R}^d} \varphi^*(y) d\nu(y). \end{aligned}$$

Hence, the infimum is attained at the pair (φ, φ^*) , which proves the theorem. $\square$

Remark 2.10. The idea of the proof is based on the so-called "double convexification trick". If (φ, ψ) are such that

$$\varphi(x) + \psi(y) \geq \langle x, y \rangle, \quad \forall x, y \in \mathbb{R}^d$$

then $\forall y, \psi(y) \geq \langle x, y \rangle - \varphi(x)$, for all x , hence $\psi \geq \varphi^*$ and thus

$$\int \varphi d\mu + \int \psi d\nu \geq \int \varphi d\mu + \int \varphi^* d\nu.$$

We can thus reduce the functional J by replacing ψ by φ^* . Moreover, since $\varphi(x) \geq \langle x, y \rangle - \varphi^*(y)$, $\forall x, y$, we have $\varphi \geq \varphi^{**}$, so we can further reduce the functional by replacing φ by φ^{**} :

$$\int \varphi d\mu + \int \psi d\nu \geq \int \varphi d\mu + \int \varphi^* d\nu \geq \int \varphi^{**} d\mu + \int \varphi^* d\nu.$$

There's no point in repeating this process because of Legendre duality (see Appendix B): f is a proper convex and lower semicontinuous function if and only if $f^{**} = f$.

Hence, the pair of optimal Kantorovich potentials (φ, ψ) can be taken to be a pair of proper convex and lower semicontinuous functions which are convex conjugates to each other, i.e. $\varphi = \psi^*$ and $\psi = \varphi^*$.

2.4 Uniqueness of optimal transport maps for the quadratic cost

We now prove Brenier's theorem which is at the core of the theory of optimal transport.

Theorem 2.11 (Brenier theorem). *Let μ and ν be two probability measures on $\mathbb{R}^d$ with finite second moments. We consider the Monge-Kantorovich problem associated with the quadratic cost $c(x, y) = \frac{1}{2}|x - y|^2$, for $x, y \in \mathbb{R}^d$. Assume that μ is absolutely continuous with respect to Lebesgue measure. Then, there exists a unique optimal transport plan $\pi \in \Pi(\mu, \nu)$ which is given by*

$$\pi = (\text{Id} \times \nabla\varphi)_{\#}\mu,$$

where $\nabla\varphi$ is the unique (i.e. uniquely determined μ -almost everywhere) gradient of a convex function φ which pushes μ forward to ν , i.e. $\nabla\varphi_{\#}\mu = \nu$. In particular, the Monge problem admits a unique solution.

Proof. Let π be an optimal transport plan and let (φ, φ^*) be optimal Kantorovich convex potentials, so that

$$\begin{aligned} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle \pi(dx, dy) &= \int_{\mathbb{R}^d} \varphi d\mu + \int_{\mathbb{R}^d} \varphi^* d\nu \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} (\varphi(x) + \varphi^*(y)) \pi(dx, dy). \end{aligned}$$

Hence, we have

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} (\varphi(x) + \varphi^*(y) - \langle x, y \rangle) \pi(dx, dy) = 0.$$

But since $\varphi(x) + \varphi^*(y) \geq \langle x, y \rangle$, for all x, y , the integrand is nonnegative, hence we get that

$$\varphi(x) + \varphi^*(y) = \langle x, y \rangle, \quad \text{for } \pi\text{-almost all } x, y \in \mathbb{R}^d.$$

In particular, φ and φ^* are finite π -almost everywhere, hence $\mu(D_\varphi) = 1$. Moreover, since φ is convex, we have $\mu(\partial D_\varphi) = 0$ since μ is absolutely continuous with respect to Lebesgue measure. Hence, we get

$$\mu(\mathring{D}_\varphi) = 1.$$

In particular, $\mathring{D}_\varphi$ is non-empty. Moreover, on $\mathring{D}_\varphi$, φ is differentiable almost everywhere, hence φ is μ -almost everywhere differentiable.

Now by the characterization of subdifferential lemma, we have that

$$y \in \partial\varphi(x), \quad \text{for } \pi\text{-almost all } x, y \in \mathbb{R}^d.$$

But since φ is μ -almost everywhere differentiable, we have that

$$\partial\varphi(x) = \{\nabla\varphi(x)\}, \quad \mu\text{-almost everywhere.}$$

Since a property that is true μ -almost everywhere is also true π -almost everywhere, we finally obtain that for π -almost all $x, y \in \mathbb{R}^d$, $y \in \partial\varphi(x) = \{\nabla\varphi(x)\}$, i.e.

$$y = \nabla\varphi(x), \quad \text{for } \pi\text{-almost all } x, y \in \mathbb{R}^d.$$

This implies that

$$\pi = (\text{Id} \times \nabla\varphi)_\# \mu,$$

or equivalently that $\pi(dx, dy) = \mu(dx) \delta_{\nabla\varphi(x)}(dy)$. Indeed, for any A, B Borel sets, we have

$$\begin{aligned} \pi(A \times B) &= \pi(\{(x, y) \mid x \in A, y \in B\}) \\ &= \pi(\{(x, y) \mid x \in A, y \in B\} \cap \{y = T(x)\}) \\ &= \pi(\{(x, T(x)) \mid x \in A, T(x) \in B\} \cap \{y = T(x)\}) \\ &= \mu(A \cap T^{-1}(B)) \\ &= (\text{Id} \times \nabla\varphi)_\# \mu(A \times B), \end{aligned}$$

where we have denoted $T = \nabla\varphi$.

It remains to prove unicity. Let π_1 and π_2 be two optimal transport plans for the Kantorovich problem. Define $\pi = \frac{1}{2}\pi_1 + \frac{1}{2}\pi_2$. Then π is also an optimal solution (easy). By the above, there exists $\varphi_1, \varphi_2, \varphi$ three convex functions such that

- (i) $\pi_1 = (\text{Id} \times \nabla\varphi_1)_\# \mu$, i.e. $(x, y) = (x, \nabla\varphi_1(x))$ π_1 -a.e.
- (ii) $\pi_2 = (\text{Id} \times \nabla\varphi_2)_\# \mu$, i.e. $(x, y) = (x, \nabla\varphi_2(x))$ π_2 -a.e.
- (iii) $\pi = (\text{Id} \times \nabla\varphi)_\# \mu$, i.e. $(x, y) = (x, \nabla\varphi(x))$ π -a.e.

But then (iii) holds also π_1 -almost everywhere. Therefore we get that

$$(x, \nabla \varphi_1(x)) = (x, \nabla \varphi(x)) \quad \pi_1\text{-a.e.},$$

hence

$$\nabla \varphi_1 = \nabla \varphi \quad \mu\text{-a.e.}$$

The same holds for $\nabla \varphi_2$. Finally, $\nabla \varphi$ is unique μ -a.e.

Finally, we have seen that if π is an optimal transport plan that can be written $\pi = (\text{Id} \times T)_\# \mu$ for some T , then T is optimal in Monge's problem and both problems coincide, thus we have that

$$\inf_{T \mid T_\# \mu = \nu} \int_{\mathbb{R}^d} \frac{1}{2} |x - T(x)|^2 \mu(dx) = \int_{\mathbb{R}^d} \frac{1}{2} |x - \nabla \varphi(x)|^2 \mu(dx),$$

and $\nabla \varphi$ is the (μ -almost surely) unique solution of Monge's problem. $\square$

3 Wasserstein distances

Let $p \in [1, +\infty)$. Define $\mathcal{P}_p(E)$ the set of probability measures which admit a p th moment:

$$\mathcal{P}_p(E) = \left\{ \mu \in \mathcal{M}_1(E) \mid \int_E d(x, x_0)^p d\mu(x) < \infty, \text{ for some } x_0 \in E \right\}.$$

Note that by the triangle inequality,

$$d(x, y) \leq d(x, x_0) + d(x_0, y),$$

so if $\mu \in \mathcal{P}_p(E)$, then for all $y \in E$,

$$\int_E d(x, y)^p d\mu(x) < \infty.$$

For the particular cost $c(x, y) = d(x, y)^p$ we define:

Definition 3.1. For μ and ν in $\mathcal{P}_p(E)$, the Wasserstein p -distance between μ and ν is defined by,

$$W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{E \times E} d(x, y)^p \pi(dx, dy) \right)^{1/p}.$$

One can prove that W_p is indeed a metric (for $p > 1$, the triangle inequality is not completely trivial to prove).

We are interested in the following dual formulation for the case $p = 1$. Define

$$\|f\|_{\text{Lip}} = \sup_{\substack{x, y \in E \\ x \neq y}} \frac{|f(x) - f(y)|}{d(x, y)},$$

so that f is 1-Lipschitz if and only if $\|f\|_{\text{Lip}} \leq 1$.

Proposition 3.2. For all μ and ν in $\mathcal{P}_1(E)$, one has

$$W_1(\mu, \nu) = \sup \left\{ \left| \int f d\mu - \int f d\nu \right| \mid \|f\|_{\text{Lip}} \leq 1 \right\}.$$

Note that it resembles the Rubinstein distance without the bounded condition on functions.

Proof. Let $\pi \in \Pi(\mu, \nu)$. First remark that

$$\int_{E \times E} d(x, y) \pi(dx, dy) < \infty,$$

by the triangle inequality since $\mu, \nu \in \mathcal{P}_1(E)$. Now let f be a 1-Lipschitz function on E . Then,

$$\begin{aligned} \left| \int f d\mu - \int f d\nu \right| &= \left| \int (f(x) - f(y)) \pi(dx, dy) \right| \\ &\leq \int |f(x) - f(y)| \pi(dx, dy) \\ &\leq \int d(x, y) \pi(dx, dy), \end{aligned}$$

since $\|f\|_{\text{Lip}} \leq 1$. Taking the infimum over transport plans and the supremum over 1-Lipschitz functions gives the inequality

$$\sup_{\|f\|_{\text{Lip}} \leq 1} \left| \int f d\mu - \int f d\nu \right| \leq W_1(\mu, \nu)$$

For the converse inequality, we will use the Kantorovich duality. Let $\varepsilon > 0$. Then, there exists $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ such that $\varphi(x) + \psi(y) \leq d(x, y)$, for μ -almost all $x \in E$ and ν -almost all $y \in E$, and such that

$$\int \varphi d\mu + \int \psi d\nu \geq W_1(\mu, \nu) - \varepsilon.$$

Now define, for all $x \in E$,

$$\psi^d(x) = \sup_{y \in E} (\psi(y) - d(x, y)).$$

Then, for all x, x' in E , using the triangle inequality $d(x', y) \leq d(x', x) + d(x, y)$, we have

$$\begin{aligned} \psi^d(x) &\leq \sup_{y \in E} (\psi(y) - d(x', y) + d(x', x)) \\ &= \sup_{y \in E} (\psi(y) - d(x', y)) + d(x', x) \\ &= \psi^d(x') + d(x', x), \end{aligned}$$

hence ψ^d is 1-Lipschitz. Note also that $\psi^d(x) \geq \psi(x)$ (taking $y = x$ in the definition of ψ^d), and from the inequality $\varphi(x) + \psi(y) \leq d(x, y)$, we also have $\psi^d(x) \leq -\varphi(x)$. Thus, one obtains

$$\begin{aligned} \sup_{\|f\|_{\text{Lip}} \leq 1} \left| \int f d\mu - \int f d\nu \right| &\geq - \int \psi^d d\mu + \int \psi^d d\nu \\ &\geq \int \varphi d\mu + \int \psi d\nu \\ &\geq W_1(\mu, \nu) - \varepsilon. \end{aligned}$$

Letting $\varepsilon \rightarrow 0$ concludes the proof. □

One can prove:

Theorem 3.3. *Let $(\mu_n)_n$ and μ be probability measures in $\mathcal{P}_p(E)$. Then,*

$$W_p(\mu_n, \mu) \xrightarrow{n \rightarrow \infty} 0$$

if and only if $\mu_n \xrightarrow[n \rightarrow \infty]{\text{weakly}} \mu$ and $\int_E d(x, x_0)^p d\mu_n(x) \xrightarrow[n \rightarrow \infty]{} \int d(x, x_0)^p d\mu(x)$, for some $x_0 \in E$.

We refer for instance to [2] for a proof.

A Lower semicontinuity

Let (E, d) be a metric space.

Definition A.1. A function $f: E \rightarrow \mathbb{R} \cup \{+\infty\}$ is said to be **lower semicontinuous** at $x \in E$ if for every sequence $x_n \rightarrow x$ in E , we have:

$$f(x) \leq \liminf_{n \rightarrow \infty} f(x_n).$$

We say that f is **lower semicontinuous on E** if f is lower semicontinuous at every point $x \in E$.

Equivalently, f is lower semicontinuous at $x \in E$ if

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ such that if } d(x, y) < \delta \text{ then } f(x) \leq f(y) + \varepsilon.$$

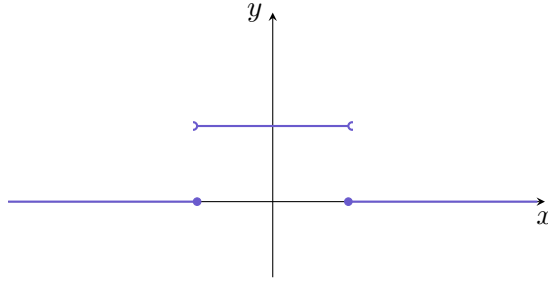


Figure 2: The characteristic function $\mathbb{1}_{(-1,1)}$ of the open set $(-1, 1)$ is a lower semicontinuous function.

Example A.2. Let $A \subset E$. Then $\mathbb{1}_A$ is lower semicontinuous if and only if A is open.

Proposition A.3 (Level set characterization). A function $f: E \rightarrow \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous if and only if for all $\alpha \in \mathbb{R}$, the sublevel set $\{x \in E \mid f(x) \leq \alpha\}$ is closed.

Proof. Let $\alpha \in \mathbb{R}$. Let $(x_n)_n \subset \{f \leq \alpha\}$ such that $x_n \rightarrow x$. Hence, for all n , $f(x_n) \leq \alpha$, so

$$\liminf_{n \rightarrow \infty} f(x_n) \leq \alpha.$$

Since f is lower semicontinuous, we get:

$$f(x) \leq \liminf_{n \rightarrow \infty} f(x_n) \leq \alpha,$$

so $x \in \{f \leq \alpha\}$. Hence, $\{f \leq \alpha\}$ is closed.

Now, suppose that for all α , $\{f \leq \alpha\}$ is closed, or equivalently that $\{f > \alpha\}$ is open. Let $(x_n)_n$ such that $x_n \rightarrow x$.

Suppose that $f(x) < \infty$. Let $\varepsilon > 0$, and consider the open set $\{f > f(x) - \varepsilon\}$. Then obviously, $x \in \{f > f(x) - \varepsilon\}$, hence there exists n_0 such that for all $n \geq n_0$, $x_n \in \{f > f(x) - \varepsilon\}$, that is

$$\exists n_0, \forall n \geq n_0, f(x_n) > f(x) - \varepsilon.$$

Hence,

$$\sup_n \inf_{k \geq n} f(x_k) \geq f(x) - \varepsilon,$$

and letting $\varepsilon \rightarrow 0$, we get, $\liminf_{n \rightarrow \infty} f(x_n) \geq f(x)$.

If now $f(x) = +\infty$, then for all $M > 0$, $\{f > M\}$ is open and contains x , so again we get that

$$\liminf_{n \rightarrow \infty} f(x_n) \geq M, \quad \text{for all } M,$$

hence $\liminf_{n \rightarrow \infty} f(x_n) = +\infty = f(x)$. □

Proposition A.4. Let $f: E \rightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function and suppose that E is compact. Then:

- (i) f is bounded below,
- (ii) f attains its infimum on E , i.e., $\exists x_0 \in E$ such that $f(x_0) = \inf_{x \in E} f(x)$.

Proof. Part (i). By contradiction, suppose that f is not bounded below. Then there exists a sequence $(x_n)_n \in E$ such that $f(x_n) \rightarrow -\infty$. Since E is compact, up to extracting a subsequence, we can suppose that $(x_n)_n$ converges to some $x \in E$. Since f is lower semicontinuous, we have,

$$f(x) \leq \liminf_{n \rightarrow \infty} f(x_n) = -\infty,$$

which is a contradiction since f takes values in $\mathbb{R} \cup \{+\infty\}$.

Part (ii). Let $m = \inf_{x \in E} f(x)$. Let $(x_n)_n \subset E$ be a minimizing sequence such that $f(x_n) \rightarrow m$. Again, since E is compact, up to extracting a subsequence, we can suppose that $(x_n)_n$ converges to some $x \in E$, and by lower semicontinuity:

$$f(x) \leq \liminf_{n \rightarrow \infty} f(x_n) = m.$$

But since $m \leq f(x)$ by definition, we get that $f(x) = m$, and the infimum is attained. $\square$

Proposition A.5. (i) Let $(f_i)_{i \in I}$ be a family of lower semicontinuous functions on E . Then, $f := \sup_{i \in I} f_i$ is lower semicontinuous.

- (ii) If $f: E \rightarrow \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous and bounded below, then there exists a nondecreasing sequence $(f_n)_{n \geq 0}$ of bounded, uniformly continuous functions (even Lipschitz) such that $f = \sup_{n \geq 0} f_n$.

Proof. Part (i). For any $\alpha \in \mathbb{R}$, we have,

$$\{f \leq \alpha\} = \{\sup_{i \in I} f_i \leq \alpha\} = \bigcap_{i \in I} \{f_i \leq \alpha\}.$$

But for all $i \in I$, $\{f_i \leq \alpha\}$ is closed since f_i is lower semicontinuous, hence $\{f \leq \alpha\}$ is closed as an intersection of closed sets.

Part (ii). If $f \equiv +\infty$, then $f_n = n$ works. So suppose that there exists x_0 such that $f(x_0) < \infty$. Define for each $n \in \mathbb{N}$:

$$f_n(x) := \inf_{y \in E} (f(y) + nd(x, y)), \quad \text{for all } x \in E.$$

Then:

- It is clear that for all $x \in E$, $f_n(x) \leq f_{n+1}(x)$, so the sequence is nondecreasing.
- From the previous point, we get that for all n , $f_n \geq f_0 = \inf_{y \in E} f(y)$. Since f is lower bounded, so is f_n .
- $\forall x \in E$, $f_n(x) \leq f(x)$. Indeed, by definition,

$$f_n(x) \leq f(y) + nd(x, y), \quad \text{for all } y \in E,$$

in particular for $y = x$, we get $f_n(x) \leq f(x)$.

- $\forall x \in E$, $f_n(x) < \infty$. Indeed,

$$f_n(x) \leq f(x_0) + nd(x, x_0) < \infty.$$

- Each f_n is n -Lipschitz: let $x, x' \in E$. Then,

$$\begin{aligned}
f_n(x) &= \inf_{y \in E} (f(y) + nd(x, y)) \\
&\leq \inf_{y \in E} (f(y) + nd(x, x') + nd(x', y)) \\
&= \inf_{y \in E} (f(y) + nd(x', y)) + nd(x, x') \\
&= f_n(x') + nd(x, x').
\end{aligned}$$

Exchanging the roles of x and x' , we get that

$$|f_n(x) - f_n(x')| \leq nd(x, x'),$$

that is f_n is n -Lipschitz (and in particular uniformly continuous on E).

- Now we prove that $f_n(x) \rightarrow f(x)$ for all x . By definition, for all $\varepsilon > 0$, there exists $y \in E$, such that

$$f(y) + nd(x, y) \leq f_n(x) + \varepsilon.$$

Hence, there exists a sequence $(y_n)_n$ such that

$$f(y_n) + nd(x, y_n) \leq f_n(x) + \frac{1}{n}.$$

Suppose that $f(x) < \infty$. Let m be a lower bound of f . Hence,

$$m + nd(x, y_n) \leq f_n(x) + \frac{1}{n} \leq f(x) + \frac{1}{n}$$

so $nd(x, y_n)$ is bounded from above. It implies that $d(y_n, x) \rightarrow 0$, so $y_n \rightarrow x$. Using

$$f(y_n) \leq f(y_n) + nd(x, y_n) \leq f_n(x) + \frac{1}{n} \leq f(x) + \frac{1}{n},$$

and taking the $\liminf$, we get that

$$\liminf_{n \rightarrow \infty} f(y_n) \leq \liminf_{n \rightarrow \infty} f_n(x) \leq f(x).$$

But since f is lower semicontinuous, and $y_n \rightarrow x$, we also have $f(x) \leq \liminf f(y_n)$. Hence, $\liminf f_n(x) = f(x)$. But since the sequence $(f_n(x))_n$ is increasing, we have that

$$\lim_{n \rightarrow \infty} f_n(x) = f(x).$$

Now if $f(x) = +\infty$, suppose by contradiction that $(f_n(x))_n$ is bounded from above by some M . We still have that $y_n \rightarrow x$, so

$$M \geq \liminf_{n \rightarrow \infty} f(y_n) = +\infty,$$

which is a contradiction. Hence $(f_n(x))_n$ is not bounded from above, and since it is increasing, one has $f_n(x) \rightarrow +\infty$.

Finally, replacing f_n by $f_n \wedge n$, one obtains an increasing sequence $(f_n)_n$ of bounded, uniformly continuous functions such that $f = \sup_n f_n$. $\square$

B Legendre duality

Theorem B.1 (Legendre duality). *Let $\varphi: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper function. The following assertions are equivalent:*

- (i) φ is convex and lower semicontinuous,
- (ii) $\varphi = \psi^*$ for some proper function ψ ,
- (iii) $\varphi^{**} = \varphi$.

Proof. (iii) $\Rightarrow$ (ii) is obvious.

(ii) $\Rightarrow$ (i) is easy.

It remains to show that (i) $\Rightarrow$ (iii). So let φ be a convex and lower semicontinuous function. By definition of the Legendre transform, one has

$$\varphi^*(y) \geq \langle x, y \rangle - \varphi(x), \quad \forall x, y$$

hence,

$$\varphi(x) \geq \langle x, y \rangle - \varphi^*(y), \quad \forall x, y$$

and thus

$$\begin{aligned} \varphi(x) &\geq \sup_y (\langle x, y \rangle - \varphi^*(y)), \quad \forall x \\ &= \varphi^{**}(x), \quad \forall x. \end{aligned}$$

Hence, it remains to prove that $\varphi^{**} \geq \varphi$. Assume first that $D_\varphi = \mathbb{R}^d$. Then, for all x , $\partial\varphi(x) \neq \emptyset$ by convexity of φ . So let $y_0 \in \partial\varphi(x)$. By the subdifferential characterization lemma,

$$\varphi(x) + \varphi^*(y_0) = \langle x, y_0 \rangle.$$

Hence,

$$\begin{aligned} \varphi(x) &= \langle x, y_0 \rangle - \varphi^*(y_0) \\ &\leq \sup_y (\langle x, y \rangle - \varphi^*(y)) = \varphi^{**}(x). \end{aligned}$$

Hence, $\varphi^{**} \geq \varphi$ if $D_\varphi = \mathbb{R}^d$. Now if $D_\varphi \subsetneq \mathbb{R}^d$, introduce the "infimal convolution"

$$\varphi_\varepsilon(x) = \inf_y (\varphi(x - y) + \frac{1}{\varepsilon}|y|) = \inf_y (\varphi(y) + \frac{1}{\varepsilon}|x - y|),$$

for all x . Since φ is proper, there exists x_0 such that $\varphi(x_0) < \infty$, hence, for all x ,

$$\varphi_\varepsilon(x) \leq \varphi(x_0) + \frac{1}{\varepsilon}|x - x_0| < \infty,$$

i.e. $D_{\varphi_\varepsilon} = \mathbb{R}^d$.

Moreover, φ_ε is easily seen to be convex (exercise). Now we prove that φ_ε is lower semicontinuous. Let us show that for all $\alpha \in \mathbb{R}$, the sublevel set $\{\varphi_\varepsilon \leq \alpha\}$ is closed. So let $x_n \in \{\varphi_\varepsilon \leq \alpha\}$ with $x_n \rightarrow x$. By definition, for all n , there exists y_n such that

$$\varphi_\varepsilon(x_n) \geq \varphi(x_n) + \frac{1}{\varepsilon}|x_n - y_n| - \frac{1}{n}.$$

Now, since,

$$\varphi_\varepsilon(x_n) \leq \varphi(x_0) + \frac{1}{\varepsilon}|x_n - x_0|.$$

and since $x_n \rightarrow x$, then for n large enough, the right-hand side is bounded by some constant $M > 0$. Combining the two above inequalities gives that

$$\varphi(x_n) + \frac{1}{\varepsilon}|x_n - y_n| \leq M + \frac{1}{n}$$

for n large enough. Hence, the sequence $(y_n)_n$ is bounded (if not, this imposes that $\varphi(x_n) \rightarrow -\infty$ which is not possible since φ is lower semicontinuous), so one can extract a convergent subsequence $y_{n_k} \rightarrow y$, for some y . Thus, one obtains, using the fact that φ is lower semicontinuous, that

$$\begin{aligned} \varphi(x) + \frac{1}{\varepsilon}|y - x| &\leq \liminf_k \left(\varphi(x_{n_k}) + \frac{1}{\varepsilon}|y_{n_k} - x_{n_k}| - \frac{1}{n_k} \right) \\ &\leq \liminf_k \varphi_\varepsilon(x_{n_k}) \\ &\leq \alpha. \end{aligned}$$

Hence, one obtains that $\varphi_\varepsilon(x) \leq \alpha$, so $x \in \{\varphi_\varepsilon \leq \alpha\}$, hence the sublevel set $\{\varphi_\varepsilon \leq \alpha\}$ is closed. Finally, φ_ε is lower semicontinuous.

Thus we get that φ_ε is a convex and lower semicontinuous function with domain equal to the whole of $\mathbb{R}^d$, so $\varphi_\varepsilon^{**} = \varphi_\varepsilon$. Moreover (exercise), for all x

$$\liminf_{\varepsilon \rightarrow 0} \varphi_\varepsilon(x) \geq \varphi(x).$$

Finally, using that $\varphi_\varepsilon \leq \varphi$, we get that

$$\begin{aligned} \varphi^{**}(x) &= \sup_y (\langle x, y \rangle - \varphi^*(y)) = \sup_y \inf_z (\langle y, x - z \rangle + \varphi(z)) \\ &\geq \sup_y \inf_z (\langle y, x - z \rangle + \varphi_\varepsilon(z)) \\ &= \varphi_\varepsilon^{**}(x) \\ &= \varphi_\varepsilon(x). \end{aligned}$$

Taking the $\liminf$ as $\varepsilon \rightarrow 0$, gives that $\varphi^{**} \geq \varphi$. □