INTRODUCTION TO OPTIMAL TRANSPORT THEORY

FRANÇOIS CHAPON

Université de Toulouse

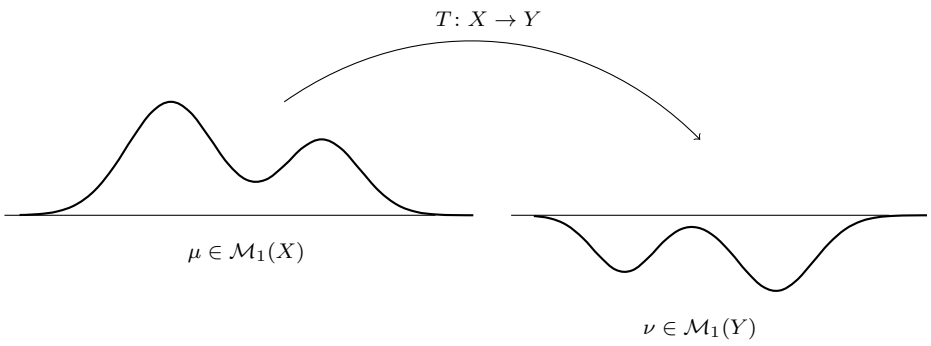
2025–2026

Contents

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 $\mu \in \mathcal{M}_1(X)$, and the hole by another probability measure $\nu \in \mathcal{M}_1(Y)$. Here and throughout, X and Y are assumed to be Polish spaces.

Definition 1.1. Let $\mu \in \mathcal{M}_1(X)$ and $\nu \in \mathcal{M}_1(Y)$. A measurable map $T\colon X \to Y$ 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 function $c\colon X \times Y \to \mathbb{R} \cup \{+\infty\}$, 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 $x, y \in \mathbb{R}^d$. 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.

The Monge's problem formulation is then:

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

Monge's formulation is an optimization problem with nonlinear constraints. As such, it is in general a hard problem. Moreover, it might exist no transport map between μ and ν. For instance, if $\mu = \delta_x$, for some $x \in X$, what can be $\nu = T_{\#}\mu$? For any Borel set $A \in \mathcal{B}(Y)$, we thus have

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

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

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 1.2. Let $\mu \in \mathcal{M}_1(X)$ and $\nu \in \mathcal{M}_1(Y)$. *A transport plan* between μ and ν is a probability measure on $X \times Y$ 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:

$$\Pi(\mu, \nu) = \left\{\pi \in \mathcal{M}_1(X \times Y) \mid (p_X)_{\#}\pi = \mu \text{ and } (p_Y)_{\#}\pi = \nu\right\},$$

where p_X and p_Y are the canonical projections onto X and Y respectively.

In probabilistic terms, the set of transport plans between μ and ν is the set of distributions of random vectors (X, Y), such that $X \sim \mu$ and $Y \sim \nu$. Any such pair (X, Y) is called a coupling of X and Y.
Note that the set of transport plans is never empty, since we always have

$$\mu \otimes \nu \in \Pi(\mu, \nu),$$

which corresponds to the independent coupling (X, Y) where X and Y are independent.

The Kantorovich problem formulation is then

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

In terms of random variables, we have

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

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

Now assume that T is a transport map between μ and ν. Define

$$\gamma_T(dx,dy) = \mu(dx)\delta_{T(x)}(dy),$$

i.e.

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

where $Id\times T\colon X\to X\times Y$ is the map $x\mapsto(x,T(x))$. Then,

$$\begin{aligned}\forall A\in\mathcal{B}(X),\ \gamma_T(A\times Y) &= \int_{A\times Y}\mu(dx)\delta_{T(x)}(dy)\\ &= \int_A \mu(dx) = \mu(A),\end{aligned}$$

and

$$\begin{aligned}\forall B\in\mathcal{B}(Y),\ \gamma_T(X\times B) &= \int_{X\times B}\mu(dx)\delta_{T(x)}(dy)\\ &= \int_X \mu(dx)\Bigg(\underbrace{\int_B \delta_{T(x)}(dy)}_{=\mathbb{1}_{T^{-1}(B)}(x)}\Bigg)\\ &= \int_X \mathbb{1}_{T^{-1}(B)}(x)\mu(dx)\\ &= \mu\left(T^{-1}(B)\right) = T_{\#}\mu(B) = \nu(B).\end{aligned}$$

Thus, γ_T is a transport plan between μ and ν. Moreover,

$$\begin{aligned}\int_{X\times Y} c(x,y)\gamma_T(dx,dy) &= \int_{X\times Y} c(x,y)\mu(dx)\delta_{T(x)}(dy)\\ &= \int_X c(x,T(x))\mu(dx).\end{aligned}$$

Therefore,

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

and optimizing over transport maps, one obtains that we always have:

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

Moreover, assume that π is optimal in the Kantorovich problem (i.e. the infimum is attained at π), and that π can be written $\pi=\gamma_T$, for some T. Then T is optimal for the Monge's problem and both problems coincide:

$$\mathcal{K}_c(\mu,\nu) = \int_{X\times Y} c(x,y)\gamma_T(dx,dy) = \int_X c(x,T(x))\mu(dx)\ge\mathcal{M}_c(\mu,\nu),$$

so in that case,

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

In the next subsection, we illustrate when this equality occurs with a discrete example.

1.3. **Discrete example.** Let $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_n\}$ two finite discrete spaces. Consider

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \in \mathcal{M}_1(X)$$
$$\nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j} \in \mathcal{M}_1(Y).$$

The measures μ and ν can be identified with row vectors in $\mathbb{R}^n$. Now a transport plan $\pi \in \Pi(\mu, \nu)$ between μ and ν takes the form

$$\pi = \left(\frac{1}{n}\pi_{ij}\right)_{1\le i,j\le n}.$$

Then, the first marginal of π being μ implies that

$$\forall i \in \{1, \dots, n\},\ \sum_{j=1}^{n} \frac{1}{n}\pi_{ij} = \mu_i = \frac{1}{n}, \quad \text{hence} \quad \sum_{j=1}^{n} \pi_{ij} = 1.$$

Likewise, the second marginal of π being ν implies that

$$\forall j \in \{1, \dots, n\},\ \sum_{i=1}^{n} \frac{1}{n}\pi_{ij} = \nu_j = \frac{1}{n}, \quad \text{hence} \quad \sum_{i=1}^{n} \pi_{ij} = 1.$$

We thus have that $(\pi_{ij})_{1\le i,j\le n}$ is a doubly stochastic matrix. Let

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

which is called the Birkhoff polytope. It is compact (as a closed subset of $M_n([0,1]) \simeq [0,1]^{n^2}$) and convex. Then, the Kantorovich problem takes the form:

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

It is well known that we have:

Theorem 1.1 (Birkhoff-von Neumann theorem (1946))**.** *The $n \times n$ permutation matrices constitue the extreme points of $\mathcal{B}_n$. Moreover, $\mathcal{B}_n$ is the convex hull of the set of $n \times n$ permutation matrices.*

We recall that if K is a convex set, the extreme points of K are the points $x \in K$ such that for all $a, b \in K$, for all $\lambda \in [0, 1]$,

$$x = \lambda a + (1 - \lambda)b \ \Rightarrow\ a = b \text{ or } \lambda \in \{0, 1\}.$$

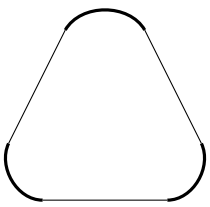


FIGURE 1. A convex set K and its extreme points (thick lines).

Now, a linear form on a compact convex subset K of a Euclidean space attains its minimum on an extreme point of K (this is known as Choquet's theorem).

Hence, we get

$$\begin{aligned}\mathcal{K}_c(\mu,\nu) &= \min_{\sigma\in S_n}\frac{1}{n}\sum_{i,j=1}^{n} c(x_i,y_j)\mathbb{1}_{\sigma(i)=j}\\ &= \min_{\sigma\in S_n}\frac{1}{n}\sum_{i=1}^{n} c(x_i,y_{\sigma(i)}).\end{aligned}$$

In this example, one has that the Monge problem coincides with the Kantorovich problem. Now as an example, take the cost to be the square of the Euclidean distance, $c(x,y)=|x-y|^2$ on $\mathbb{R}\times\mathbb{R}$, so our problem consists of finding

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

One can see that the minimum is attained when the two sequences $(x_i)_{1\le i\le n}$ and $(y_i)_{1\le i\le n}$ are monotone ordering, i.e.

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

To see this, one may assume that the sequence $(x_i)_{1\le i\le n}$ is increasing. Now, if there is some i such that $y_i>y_{i+1}$, then

$$\begin{aligned}(x_i-y_i)^2+(x_{i+1}-y_{i+1})^2 &= (x_i-y_{i+1})^2+(x_{i+1}-y_i)^2+2(x_{i+1}-x_i)(y_i-y_{i+1})\\ &\ge (x_i-y_{i+1})^2+(x_{i+1}-y_i)^2.\end{aligned}$$

Hence, one can reorder the sequence of $(y_i)_{1\le i\le n}$ using successive transpositions while keeping the total cost smaller. Thus, we get that

$$\min_{\sigma\in S_n}\sum_{i=1}^{n}|x_i-y_{\sigma(i)}|^2=\sum_{i=1}^{n}|x_i-y_{\sigma\circ\psi^{-1}(i)}|^2,$$

where σ and ψ are the permutations defined by

$$x_{\psi(1)}<\cdots<x_{\psi(n)} \quad \text{and} \quad y_{\sigma(1)}<\cdots<y_{\sigma(n)}.$$

Here, the solution of Monge problem is unique and the optimal transport map is $T\colon X\to Y$ given by $T\colon x_i\mapsto y_{\sigma\circ\psi^{-1}(i}$, for $i\in\{1,\dots,n\}$.

The above discrete optimization problem is an example of a linear programming, or linear optimization. The standard formulation of a linear programming consists of, given $c\in\mathbb{R}^N$, $b\in\mathbb{R}^M$ and a $M\times N$ matrix A,

$$\begin{cases}\text{minimize } \langle c,x\rangle \text{ subject to the}\\ \text{constraints } Ax=b,\ x\ge 0.\end{cases}$$

This is known as the *primal* problem. The *dual* problem is then to

$$\begin{cases}\text{maximize } \langle b,y\rangle \text{ subject to the}\\ \text{constraints } A^{\intercal}y\le c.\end{cases}$$

Intuitively, the "constraints become the variables" and the "variables become the constraints". Indeed, if $x\in\mathbb{R}^n$ is a solution of the primal problem subject to the constraints

$$\forall i\in\{1,\dots,M\},\quad (Ax)_i=b_i,$$

then multiplying each constraint by a scaling factor $y_i\in\mathbb{R}$ and summing over all the constraints give that

$$\sum_{i=1}^{M}y_i(Ax)_i=\sum_{i=1}^{M}y_ib_i,$$

i.e.

$$\langle A^{\intercal}y,x\rangle=\langle b,y\rangle.$$

Thus, if we can find $y \in \mathbb{R}^M$ such that $A^\intercal y \leq c$, then since $x \geq 0$, we get that

$$\langle b, y\rangle = \langle A^\intercal y, x\rangle \leq \langle c, x\rangle.$$

Thus, each $y \in \mathbb{R}^M$ such that $A^\intercal y \leq c$ gives a possible lower bound for $\langle c, x\rangle$. As we seek for the best lower bound, we have to maximize over such y, giving the dual problem formulation. Moreover, under some mild conditions, we have the duality

$$\min\left\{\langle c, x\rangle \,|\, Ax = b, x \geq 0\right\} = \max\left\{\langle b, y\rangle \,|\, A^\intercal y \leq c\right\}.$$

Our initial problem of transporting the discrete measure $\mu = (\mu_i)_{1\leq i\leq n}$ to the measure $\nu = (\nu_j)_{1\leq j\leq m}$ with cost $c = (c_{ij})_{1\leq i\leq n, 1\leq j\leq m}$, with $c_{ij} = c(x_i, y_j)$, is

$$\text{minimize} \sum_{\substack{i=1,\dots,n\\ j=1,\dots,m}} c_{ij}\pi_{ij}$$

over π subject to the constraints

$$\sum_{j=1}^{m} \pi_{ij} = \mu_i,\ \sum_{i=1}^{n} \pi_{ij} = \nu_j,\ \pi_{ij} \geq 0.$$

This is translated into a primal linear program by putting $N = nm$, $M = n+m$, and "vectorize" π, c, b as

$$\begin{aligned}
\pi &= (\pi_{11}, \dots, \pi_{1m}, \pi_{21}, \dots, \pi_{2m}, \dots)^\intercal \in \mathbb{R}^N\\
c &= (c_{11}, \dots, c_{1m}, c_{21}, \dots, c_{2m}, \dots)^\intercal \in \mathbb{R}^N\\
b &= (\mu_1, \dots, \mu_n, \nu_1, \dots, \nu_m)^\intercal \in \mathbb{R}^M,
\end{aligned}$$

and define the $(n+m)\times nm$ matrix A by

$$A = \begin{pmatrix}
\overbrace{\mathbf{1}}^{m} & \overbrace{0}^{m} & \cdots & \overbrace{0}^{m}\\
0 & \mathbf{1} & \cdots & 0\\
\vdots & & & \\
0 & 0 & \cdots & \mathbf{1}\\
e_1 & e_1 & \cdots & e_1\\
\vdots & & & \\
e_m & e_m & \cdots & e_m
\end{pmatrix}$$

where, in the first n rows, $\mathbf{1}$ is the row vector $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^m$, and the m last rows are given by n times the ith basis row vector $e_i = (0, \dots, 1, \dots, 0)$ of $\mathbb{R}^m$.
Now, letting $y = (\varphi_1, \dots, \varphi_n, \psi_1, \dots, \psi_m) \in \mathbb{R}^M$, and using the explicit form of the matrix A, the dual problem writes:

$$\text{maximize} \sum_{i=1}^{n} \varphi_i\mu_i + \sum_{j=1}^{m} \psi_j\nu_j$$

over (φ, ψ) subject to the constraints

$$\varphi_i + \psi_j \leq c_{ij}, \quad \text{for all } i = 1, \dots, n \text{ and } j = 1, \dots, m.$$

Note that from a computational point of view, exploring the set of $n\times m$ matrices (π_{ij}) is much more demanding than exploring the set of vectors $(\varphi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m$. Hence, an algorithmic resolution should prefer the dual problem!
In fact, there is a general dual formulation for the Kantorovich problem.

1.4. **Kantorovich dual formulation.** Let X and Y be Polish spaces. Let $\mu \in \mathcal{M}_1(X)$ and $\nu \in \mathcal{M}_1(Y)$. Let $c\colon X \times Y \to [0, +\infty]$ 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:

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

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

$$J(\varphi, \psi) := \int_X \varphi d\mu + \int_Y \psi d\nu.$$

Denote

$$\Phi_c = \Big\{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) \,|\, \varphi(x) + \Psi(y) \leq c(x, y) \quad \text{for } \mu\text{-almost all } x \in X \text{ and } \nu\text{-almost all } y \in Y\Big\}.$$

We denote by $\Phi_c \cap C_b$ the same set than above with $L^1(\mu) \times L^1(\nu)$ replaced by $C_b(X) \times C_b(Y)$. Note that $\Phi_c \cap C_b \subset \Phi_c$.

Theorem 1.2 (Kantorovich duality)**.** *We have*

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

We will prove later the Kantorovich duality but only for the quadratic cost $c(x, y) = \frac{1}{2}|x - y|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$. We refer to [2] for a proof in the general case.
We now turn to a closer analysis of the Kantorovich formulation, focusing on the existence and structure of optimal transport plans.

2. Analysis of the Kantorovich formulation

2.1. **Existence of minimizers.** First we prove that the infimum of I over transport plans is attained.

Theorem 2.1. *There exists $\pi^* \in \Pi(\mu, \nu)$, such that*

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

We say that π^* is an *optimal* transport plan. Note that in general, it is not unique.

Proof. We first prove that the set of transport plans $\Pi(\mu, \nu)$ is compact for the weak topology. It is easily seen to be closed: let $(\pi_n)_n \subset \Pi(\mu, \nu)$ such that $\pi_n \overset{\text{weakly}}{\longrightarrow} \pi$. Then for all bounded continuous function f on $X \times Y$,

$$\int_{X\times Y} f d\pi_n \underset{n\to\infty}{\longrightarrow} \int_{X\times Y} f d\pi.$$

Then, for all $g \in C_b(X)$, we have

$$\int_X g d\mu = \int_{X\times Y} g d\pi_n \underset{n\to\infty}{\longrightarrow} \int_{X\times Y} g d\pi,$$

hence

$$\int_X g d\mu = \int_{X\times Y} g d\pi,$$

which implies that $(p_X)_\# \pi = \mu$. Similarly, we have $(p_Y)_\# \pi = \nu$. Hence, $\pi \in \Pi(\mu, \nu)$, so $\Pi(\mu, \nu)$ is closed.
Now we prove that $\Pi(\mu, \nu)$ is relatively compact, which is equivalent to $\Pi(\mu, \nu)$ being tight by Prokhorov's theorem. Since X and Y are Polish spaces, μ and ν are tight, hence, for all $\varepsilon > 0$, there exists a compact set $K_X \subset X$, a compact set $K_Y \subset Y$ such that

$$\mu\left(K_X^c\right) \leq \varepsilon/2 \quad \text{and} \quad \nu\left(K_Y^c\right) \leq \varepsilon/2.$$

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

$$\begin{aligned}
\pi\left(K^c\right) &= \pi\left(\left(K_X \times K_Y\right)^c\right) \\
&= \pi\left(\left(K_X^c \times Y\right) \cup \left(X \times K_Y^c\right)\right) \\
&\leq \pi\left(K_X^c \times Y\right) + \pi\left(X \times K_Y^c\right) \\
&= \mu\left(K_X^c\right) + \nu\left(K_Y^c\right) \\
&\leq \varepsilon/2 + \varepsilon/2 \\
&= \varepsilon.
\end{aligned}$$

Hence, $\Pi(\mu, \nu)$ is relatively compact, and thus compact.
Now we prove that I is lower semicontinuous, that is for all sequence $(\pi_n)_n \subset \mathcal{M}_1(X \times Y)$ such that $\pi_n \xrightarrow[n\to\infty]{\text{weakly}} \pi$,

$$I(\pi) \leq \liminf_n I(\pi_n).$$

By assumption, c is a lower semicontinuous function, hence (see appendix A) there exists a nondecreasing sequence of bounded uniformly continuous functions $(c_k)_k$ such that $c_k \nearrow c$ as $k \to \infty$. Hence, we have,

$$\int c_k d\pi = \lim_{n\to\infty} \int c_k d\pi_n = \liminf_n \int c_k d\pi_n \leq \liminf_n \int c d\pi_n.$$

But using monotone convergence theorem, we also have

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

hence we get

$$\int c d\pi \leq \liminf_n \int c d\pi_n.$$

To conclude, we known (see appendix A) that a lower semicontinuous function on a compact set attains its infimum, hence there exists $\pi^* \in \Pi(\mu, \nu)$, such that $I(\pi^*) = \inf_{\pi \in \Pi(\mu,\nu)} I(\pi)$. □

Exercise 1. 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

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

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. **Cyclical monotonicity.**

Definition 2.1. Let $\mu \in \mathcal{M}_1(X)$. The *support* of μ is the (closed) set

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

Hence, the support of μ is the largest closed subset of X 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 2.2. A subset $\Gamma \subset X \times Y$ 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 2.2. *Consider a continuous cost c and let $\pi \in \Pi(\mu,\nu)$ be an optimal transport plan. Then,* $\operatorname{supp}\pi$ *is c-cyclically monotone.*

Proof. By contradiction, suppose that $\operatorname{supp}\pi$ is not c-cyclically monotone. Then, there exists $\varepsilon > 0$ and n points $(x_1, y_1), \dots, (x_n, y_n)$ in $\operatorname{supp}\pi$ 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 $\operatorname{supp}\pi$, we have $\alpha_i > 0$ for all i. Let $\alpha = \min(\alpha_1, \dots, \alpha_n)$ and define π_i the probability measure on $X \times Y$ defined by

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

for all Borel set $A \subset X \times Y$, that π_i is the restriction of π to $U_i \times V_i$. Define also $\mu_i = (p_X)_\# \pi_i \in \mathcal{M}_1(X)$ and $\nu_i = (p_Y)_\# \pi_i \in \mathcal{M}_1(Y)$. 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 $\mu_{i+1} \otimes \nu_i \in \mathcal{M}_1(U_{i+1} \times V_i)$,

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

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

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

Finally, we get

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

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

For the quadratic cost $c(x, y) = \frac{1}{2}|x - y|^2$ 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 $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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:

Lemma 2.1. Let $C \subset \mathbb{R}^d$ be a convex set. Then, $\lambda(\partial C) = 0$.

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

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

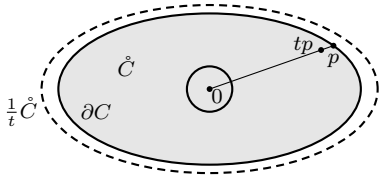
Now let $p \in \partial C$ and let $(p_n)_n \subset C$ such that $p_n \to p$. Hence, for all n, we have

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

As $p_n \to p$, 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

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

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

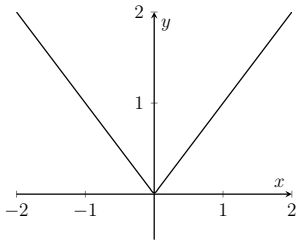
A proper convex function φ is continuous and locally Lipschitz on $\mathring{D}_\varphi$, hence by Rademacher's theorem, φ is differentiable λ-almost everywhere.
If φ is differentiable at x, then, for all $z \in \mathbb{R}^d$,

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

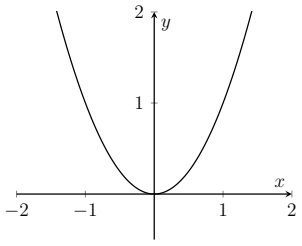
where $\nabla\varphi(x)$ is the gradient of φ at x. It says that the graph of φ lies above its tangent at x. When φ is not differentiable, we can generalize this idea by introducing the *subdifferential* $\partial\varphi(x)$ at x, which is the set:

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

Thus, $\partial\varphi\colon x \mapsto \partial\varphi(x)$ 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:

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

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.

Lemma 2.2. 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,

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

□

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

Theorem 2.3. *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 $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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 $x \in \mathbb{R}^d$,

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

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 $x \in \mathbb{R}^d$,

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

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 $x \in \mathbb{R}^d$,

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

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

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

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

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

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 $z \in \mathbb{R}^d$,

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

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 $(x', y') \in \partial\varphi$. □

3. The case of the quadratic cost

3.1. **Kantorovich duality for the quadratic cost.** In this section, we 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 $x, y \in \mathbb{R}^d$,

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

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

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

Redefining $\widetilde{\varphi}$ and $\widetilde{\psi}$ 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 $\widetilde{\varphi}$ and $\widetilde{\psi}$. Moreover, we have

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

Hence, Kantorovich duality takes the following form:

Theorem 3.1. *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

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

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,

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

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

Remark 3.1. 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 $\psi = \varphi^*$.

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

Theorem 3.2 (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*

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

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

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

Hence, we have

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

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

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

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

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

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

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

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) $\pi_1 = (Id \times \nabla\varphi_1)_{\#}\mu$, i.e. $(x, y) = (x, \nabla\varphi_1(x))$ π_1-a.e.
(ii) $\pi_2 = (Id \times \nabla\varphi_2)_{\#}\mu$, i.e. $(x, y) = (x, \nabla\varphi_2(x))$ π_2-a.e.
(iii) $\pi = (Id \times \nabla\varphi)_{\#}\mu$, 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 $\pi = (Id \times T)_{\#}\mu$ for some T, then T is optimal in Monge's problem and both problems coincide, thus we have that

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

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

4. Wasserstein distances

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

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

Note that by the triangle inequality,

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

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

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

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

Definition 4.1. For μ and ν in $\mathcal{P}_p(X)$, the *Wasserstein p-distance* between μ and ν is defined by,

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

One can proved 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

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

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

Proposition 4.1. *For all μ and ν in $\mathcal{P}_1(X)$, one has*

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

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

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

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

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

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

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

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

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 $x \in X$ and ν-almost all $y \in X$, and such that

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

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

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

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

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

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

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

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

One can prove:

Theorem 4.1. *Let $(\mu_n)_n$ and μ be probability measures in $\mathcal{P}_p(X)$. Then,*

$$W_p(\mu_n,\mu) \underset{n\to\infty}{\longrightarrow} 0$$

if and only if $\mu_n \overset{weakly}{\underset{n\to\infty}{\longrightarrow}} \mu$ and $\int_X d(x,x_0)^p d\mu_n(x) \underset{n\to\infty}{\longrightarrow} \int d(x,x_0)^p d\mu(x)$, for some $x_0 \in X$.

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

References

This lecture notes are based on:

[1] Figalli, A., and Glaudo F., *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*, European Mathematical Society, (2023).
[2] Villani, C., *Topics in Optimal Transportation*, Vol. 58, American Mathematical Soc., (2021).

Appendix A. Lower semicontinuity

Let (X, d) be a metric space.

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

$$f(x) \leq \liminf_n f(x_n).$$

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

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

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

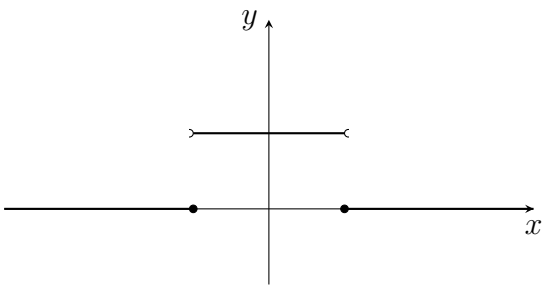


Figure 3. The characteristic function $\mathbb{1}_{(-1,1)}$ of the open set $(-1, 1)$ is a lower semicontinuous function.

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

Proposition A.1 (Level set characterization)**.** *A function $f\colon X \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous if and only if for all $\alpha \in \mathbb{R}$, the sublevel set $\{x \in X \mid f(x) \leq \alpha\}$ is closed.*

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

$$\liminf_n f(x_n) \leq \alpha.$$

Since f is lower semicontinuous, we get:

$$f(x) \leq \liminf f(x_n) \leq \alpha,$$

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

$$\liminf_n f(x_n) \geq M, \quad \text{for all } M,$$

hence $\liminf_n f(x_n) = +\infty = f(x)$. □

Proposition A.2. *Let $f\colon X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function and suppose that X is compact. Then:*

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

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

$$f(x) \leq \liminf_n f(x_n) = -\infty,$$

which is a contradiction since f takes values in $\mathbb{R} \cup \{+\infty\}$.
(2) Let $m = \inf_{x\in X} f(x)$. Let $(x_n)_n \subset X$ be a minimizing sequence such that $f(x_n) \to m$. Again, since X is compact, up to extracting a subsequence, we can suppose that $(x_n)_n$ converges to some $x \in X$, and by lower semicontinuity:

$$f(x) \leq \liminf_n f(x_n) = m.$$

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

Proposition A.3. *(i) Let $(f_i)_{i\in I}$ be a family of lower semicontinuous functions on X. Then, $f := \sup_{i\in I} f_i$ is lower semicontinuous.*
(ii) If $f\colon X \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous and bounded below, then there exists a nondecreasing sequence $(f_n)_n$ of bounded, uniformly continuous functions (even Lipschitz) such that $f = \sup_n f_n$.

Proof. (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.
(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 $n \in \mathbb{N}$:

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

Then:

- It is clear that for all $x \in X$, $f_n(x) \leq f_{n+1}(x)$, so the sequence is nondecreasing.
- From the previous point, we get that for all n, $f_n \geq f_0 = \inf_{y\in X} f(y)$. Since f is lower bounded, so is f_n.
- $\forall x \in X$, $f_n(x) \leq f(x)$. Indeed, by definition,
$$f_n(x) \leq f(y) + nd(x, y), \quad \text{for all } y \in X,$$
in particular for $y = x$, we get $f_n(x) \leq f(x)$.
- $\forall x \in X$, $f_n(x) < \infty$. Indeed,
$$f_n(x) \leq f(x_0) + nd(x, x_0) < \infty.$$
- Each f_n is n-Lipschitz: let $x, x' \in X$. Then,
$$\begin{aligned} f_n(x) &= \inf_{y\in X} \Big(f(y) + nd(x, y) \Big) \\ &\leq \inf_{y\in X} \Big(f(y) + nd(x, x') + nd(x', y) \Big) \\ &= \inf_{y\in X} \Big(f(y) + nd(x', y) \Big) + nd(x, x') \\ &= f_n(x') + nd(x, x'). \end{aligned}$$

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 X).

- Now we prove that $f_n(x) \to f(x)$ for all x. By definition, for all $\varepsilon > 0$, there exists $y \in X$, 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 $d(y_n, x) \to 0$, so $y_n \to x$. 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

$$\liminf_n f(y_n) \leq \liminf_n f_n(x) \leq f(x).$$

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

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

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

$$M \geq \liminf_n f(y_n) = +\infty,$$

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

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

Appendix B. Legendre duality

Theorem B.1 (Legendre duality)**.** *Let $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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)⇒(ii) is obvious.
(ii)⇒(i) is easy.
It remains to show that (i)⇒(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,

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

and thus

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

Hence, it remains to prove that $\varphi^{**} \geq \varphi$. Assume first that $D_\varphi = \mathbb{R}^d$. Then, for all x, $\partial\varphi(x) \neq \varnothing$ 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,

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

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

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

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 $x_n \to x$. 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 $x_n \to x$, 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 $\varphi(x_n) \to -\infty$ which is not possible since φ is lower semicontinuous), so one can extract a convergent subsequence $y_{n_k} \to y$, for some y. Thus, one obtains, using the fact that φ is lower semicontinuous, that

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

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

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

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

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

Taking the lim inf as $\varepsilon \to 0$, gives that $\varphi^{**} \geq \varphi$. $\square$