# Optimal Transport

## Lecture Notes

Lecturer: Prof. Dr. Alessio Figalli

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Version: January 6, 2020

Optimal Transport taught by Prof. Alessio Figalli
during the fall semester 2019 at the ETH Zürich

# Contents

**Abstract**

The aim of the course is to provide a self-contained introduction to optimal transport. The students are expected to know the basic concepts of measure theory. Although not strictly required, some basic knowledge of Riemannian geometry may be useful.
We will first introduce the optimal transport problem and explain how to solve it in some important cases of interest. Then we will show a series of applications to geometry and to gradient flows.

# 1 Introduction

## 1.1 Historical Overview

**1781 - Monge** Concept: Transport map. How to build fortifications from soil in the cheapest possible way. Monge's cost $c(x,y) \coloneqq |x-y|$. Question: Who goes where?

**1940's - Kantorovich** Concept: Transport class. Fix bakeries at $\{x_i\}$ and coffee shops at $\{y_j\}$. Bakeries produce $\alpha_i > 0$ breads and coffee shops require $\beta_j > 0$ breads, where we assume that $\sum_{i\in I}\alpha_i = \sum_{j\in J}\beta_j = 1$ (demand=request, and we normalize them to be equal to 1). For Monge, transport is deterministic: mass located at $x$ is sent to some point $T(x)$. Monge's problem is not compatible with Kantorovich's. He looked for matrices $(\gamma_{ij})_{\substack{i=1,\dots,N\\ j=1,\dots,M}}$, where $N \coloneqq$ #bakeries and $M \coloneqq$ #coffee shops, such that

- $\gamma_{ij} \geq 0$ (convexity constraint);
- $\forall i\colon \alpha_i = \sum_{j=1}^{M}\gamma_{ij}$ (linear constraint);
- $\forall j\colon \beta_j = \sum_{i=1}^{N}\gamma_{ij}$ (linear constraint);
- $\gamma_{ij}$ minimise the cost $\sum_{ij}\gamma_{ij}c(x_i,y_j)$ (linear function).

**Applications** The minimisation problems depend a lot on the cost function:

- $c(x,y) = |x-y|^2$ – Connected to:
  Euler equations; Isoperimetric, Sobolev inequalities; PDE's $\partial_t u = \Delta u$, $\partial_t u = \Delta(u^m)$ and $\partial_t u = \operatorname{div}(\nabla W * u\, u)$.
- $c(x,y) = |x-y|^p$; For $p=1$ it's the most difficult problem (Monge). – Appears in probability and kinetic theory.
- $c(x,y) = d(x,y)^2$ on $(M,g)$ a Riemannian manifold – Applications to curvature.

## 1.2 Basics

*Remark* 1.2.1. All spaces we consider are metric spaces that are separable and complete, for most of the course $X = Y = \mathbb{R}^n$. Also all measures are Borel measures and all maps are Borel $S\colon X \to Y$ ($S^{-1}(A)$ Borel for all $A \subset Y$ Borel).

*Remark* 1.2.2. The set of probability measures over $X$ will be denoted by $\mathcal{P}(X)$, and the class of Borel-measurable sets by $\mathcal{B}(X)$.

**Definition 1.2.3.** Given $X, Y$ spaces, take $T\colon X \to Y$ and $\mu \in \mathcal{P}(X)$. We define the **image measure** $T_{\#}\mu \in \mathcal{P}(Y)$ as

$$\forall A \in \mathcal{B}(Y)\colon \qquad (T_{\#}\mu)(A) \coloneqq \mu(T^{-1}(A)).$$

**Lemma 1.2.4.** *$T_{\#}\mu$ is a probability measure on $Y$.*

*Proof.*
- $(T_{\#}\mu)(\emptyset) = \mu(T^{-1}(\emptyset)) = \mu(\emptyset) = 0$;
- $(T_{\#}\mu)(Y) = \mu(T^{-1}(Y)) = \mu(X) = 1$;

- Let $\{A_i\}_{i\in I} \subset Y$ be a countable family of disjoint sets. We claim first that $\{T^{-1}(A_i)\}_{i\in I}$ are disjoint. Indeed, if that would not be the case and $x \in T^{-1}(A_i) \cap T^{-1}(A_j)$ then $T(x) \in A_i \cap A_j$ which is a contradiction. Thanks to this, using that $\mu$ is a measure (and thus $\sigma$-additive on disjoint sets) we get

$$\begin{aligned} T_{\#}\mu(\cup_{i\in I} A_i) &= \mu(T^{-1}(\cup_{i\in I} A_i)) = \mu(\cup_{i\in I} T^{-1}(A_i)) \\ &= \sum_{i\in I} \mu(T^{-1}(A_i)) = \sum_{i\in I} T_{\#}\mu(A_i). \end{aligned}$$

□

*Remark* 1.2.5. One might be tempted to define the "measure" $S^{\#}\nu(E) \coloneqq \nu(S(E))$ for $S\colon X \to Y$ and $\nu \in \mathcal{P}(Y)$, but this construction does not work in general. Indeed, since the image of two disjoint sets can coincide (consider for instance the case when $S$ is a constant map), $S^{\#}\nu$ may not be additive on disjoint sets.

**Lemma 1.2.6.** *Let $T\colon X \to Y$, $\mu \in \mathcal{P}(X)$, and $\nu \in \mathcal{P}(Y)$. Then*

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

*if and only if, for any $\varphi\colon Y \to \mathbb{R}$ Borel and bounded, we have*

$$\int_Y \varphi(y) d\nu(y) = \int_X \varphi(T(x)) d\mu(x). \tag{1.1}$$

*Proof.* The implication (1.1) $\implies$ $\nu = T_{\#}\mu$ follows choosing $\varphi = \mathbb{1}_A$ (with $A$ Borel). We will focus on the other implication.

For any $A \subset Y$ and $\lambda \in \mathbb{R}$, it holds

$$\begin{aligned} \nu(A) = \mu(T^{-1}(A)) &\Leftrightarrow \int_A d\nu(y) = \int_{T^{-1}(A)} d\mu(x) \\ &\Leftrightarrow \int_Y \mathbb{1}_A(y) d\nu(y) = \int_X \mathbb{1}_{T^{-1}(A)}(x) d\mu(x) \\ &\Leftrightarrow \int_Y \mathbb{1}_A(y) d\nu(y) = \int_X \mathbb{1}_A(T(x)) d\mu(x) \\ &\Leftrightarrow \int_Y \lambda\mathbb{1}_A(y) d\nu(y) = \int_X \lambda\mathbb{1}_A(T(x)) d\mu(x). \end{aligned}$$

Also, by linearity,

$$\int_Y \lambda\mathbb{1}_A(y) d\nu(y) = \int_X \lambda\mathbb{1}_A(T(x)) d\mu(x) \Leftrightarrow \int_Y \sum_{i\in I} \lambda_i \mathbb{1}_{A_i}(y) d\nu(y) = \int_X \sum_{i\in I} \lambda_i \mathbb{1}_{A_i}(T(x)) d\mu(x),$$

where $I$ is a finite set of indices, $\{\lambda_i\}_{i\in I} \subseteq \mathbb{R}$ and $\{A_i\}_{i\in I}$ are Borel sets. To deduce (1.1), we need to show that any bounded Borel function can be approximated by functions of the form $\sum_{i\in I} \lambda_i \mathbb{1}_{A_i}$ where $I$ is a finite set. To prove this, take $M \gg 1$ and for any $i \in \mathbb{Z}$ define the set $A_i \coloneqq \{\frac{i}{M} \le \varphi < \frac{i+1}{M}\}$, and let $\varphi_M \coloneqq \sum_{i\in\mathbb{Z}} \frac{i}{M} \mathbb{1}_{A_i}$ (note that, since $\varphi$ is bounded, we have $A_i = \emptyset$ for $|i| \gg 1$). Then

$$\|\varphi - \varphi_M\|_{L^\infty} \le \max_{i\in\mathbb{Z}} \|\varphi - \varphi_M\|_{L^\infty(A_i)} \le \frac{1}{M}$$

and therefore, for any $\sigma \in \mathcal{P}(Y)$, we have

$$\left| \int_Y (\varphi - \varphi_M)\, d\sigma \right| \le \|\varphi - \varphi_M\|_{L^\infty} \int_Y d\sigma \le \frac{1}{M}.$$

Hence,

$$\int_Y \sum_{i\in I} \lambda_i \mathbb{1}_{A_i}(y) d\nu(y) = \int_X \sum_{i\in I} \lambda_i \mathbb{1}_{A_i}(T(x)) d\mu(x) \qquad \forall\, \lambda_i, A_i,$$
$$\Rightarrow \int_Y \varphi_M(y) d\nu(y) = \int_X \varphi_M(T(x)) d\mu(x) \qquad \forall\, M$$
$$\Rightarrow \int_Y \varphi(y) d\nu(y) = \int_X \varphi(T(x)) d\mu(x).$$

□

**Corollary 1.2.7.** *$\nu = T_\#\mu \Rightarrow \int_Y \varphi\, d(T_\#\mu) = \int_X \varphi \circ T\, d\mu$ for any $\varphi : X \to \mathbb{R}$ Borel and bounded.*

**Lemma 1.2.8.** *Let $T\colon X \to Y$ and $S\colon Y \to Z$ be measurable, then*

$$(S \circ T)_\#\mu = S_\#(T_\#\mu).$$

*Proof.* For any $\varphi : X \to \mathbb{R}$ Borel and bounded we have

$$\int \varphi\, d(S \circ T)_\#\mu = \int \varphi \circ (S \circ T)\, d\mu = \int (\varphi \circ S) \circ T\, d\mu$$
$$= \int \varphi \circ S\, dT_\#\mu = \int \varphi\, dS_\#(T_\#\mu).$$

Then the result follows from Lemma 1.2.6. □

## 1.3 Transport maps

**Definition 1.3.1.** Given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, a map $T\colon X \to Y$ is called a **transport map** from $\mu$ to $\nu$ if $T_\#\mu = \nu$.

*Remark* 1.3.2. Given $\mu$ and $\nu$, the set $\{T \mid T_\#\mu = \nu\}$ may be empty.

**Example 1.3.3.** For $\mu = \delta_{x_0}$ with $x_0 \in X$, given some $T\colon X \to Y$, we have

$$\int_Y \varphi(y) d(T_\#\mu)(y) = \int_Y \varphi \circ T(x) d\mu(x) = \varphi(T(x_0)).$$

Therefore, $T_\#\mu = \delta_{T(x_0)}$. If $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$, then for any $T$ we have $T_\#\mu \neq \nu$. Thus the set $\{T \mid T_\#\mu = \nu\}$ is empty.

**Definition 1.3.4.** We call $\gamma \in \mathcal{P}(X \times Y)$ a **coupling** of $\mu$ and $\nu$ if

$$(\pi_X)_\#\gamma = \mu \text{ and } (\pi_Y)_\#\gamma = \nu,$$

where

$$\pi_X(x, y) = x, \qquad \pi_Y(x, y) = y \qquad \forall\, (x, y) \in X \times Y.$$

This is equivalent to requiring that

$$\forall \varphi\colon X \to \mathbb{R} \quad \int_{X\times Y} \varphi(x) d\gamma(x, y) = \int_{X\times Y} \varphi \circ \pi_X(x, y) d\gamma(x, y) = \int_X \varphi(x) d\mu(x),$$

and

$$\forall \psi\colon Y \to \mathbb{R} \quad \int_{X\times Y} \psi(y) d\gamma(x, y) = \int_{X\times Y} \psi \circ \pi_Y(x, y) d\gamma(x, y) = \int_Y \psi(y) d\nu(y).$$

We denote by $\Gamma(\mu, \nu)$ the set of coupling of $\mu$ and $\nu$.

*Remark* 1.3.5. Given $\mu$ and $\nu$, the set $\Gamma(\mu,\nu)$ is always non-empty. Indeed $\gamma = \mu \otimes \nu$ (defined by $\int \varphi(x,y)d\gamma(x,y) = \iint \varphi(x,y)d\mu(x)d\nu(y)$) is a coupling:

$$\int_{X\times Y} \varphi(x)d\mu(x)d\nu(y) = \int_Y d\nu(y)\int_X \varphi(x)d\mu(x) = 1\cdot\int_X \varphi(x)d\mu(x) = \int_X \varphi(x)d\mu(x),$$
$$\int_{X\times Y} \psi(y)d\mu(x)d\nu(y) = \int_X d\mu(x)\int_Y \psi(y)d\nu(y) = 1\cdot\int_Y \psi(y)d\nu(y) = \int_Y \psi(y)d\nu(y).$$

*Remark* 1.3.6. *(Transport map vs. Coupling)* Let $T\colon X \to Y$ such that $T_\#\mu = \nu$. Consider the map $\mathrm{id}\times T\colon X \to X\times Y$, i.e. $x \mapsto (x, T(x))$, and define

$$\gamma_T \coloneqq (\mathrm{id}\times T)_\#\mu \in \mathcal{P}(X\times Y).$$

Claim: $\gamma_T \in \Gamma(\mu,\nu)$.
Indeed:

- $(\pi_X)_\#\gamma_T = (\pi_X)_\#(\mathrm{id}\times T)_\#\mu = (\pi_X\circ(\mathrm{id}\times T))_\#\mu = \mathrm{id}_\#\mu = \mu$.
- $(\pi_Y)_\#\gamma_T = (\pi_Y)_\#(\mathrm{id}\times T)_\#\mu = (\pi_Y\circ(\mathrm{id}\times T))_\#\mu = T_\#\mu = \nu$.

This proves that any transport map $T$ induces a coupling $\gamma_T$.

**Example 1.3.7.** *(Examples of transport)*

1. Measurable transport ([]):

   **Theorem 1.3.8.** *Let $\mu \in \mathcal{P}(X)$ such that $\mu$ has no atoms (i.e. $\mu(\{x\}) = 0$ for any $x \in X$). Then there exists $T\colon X \to \mathbb{R}$ such that $T$ is injective $\mu$-a.e.,*

   $$T_\#\mu = dx|_{[0,1]}.$$

   *Moreover $T^{-1}\colon [0,1] \to X$ exists Lebesgue-a.e., and $T^{-1}_\# dx = \mu$.*

2. Monotone rearrangement: Given $\mu,\nu \in \mathcal{P}(\mathbb{R})$, let $F(x) \coloneqq \int_{-\infty}^x d\mu(t)$ and $G(y) \coloneqq \int_{-\infty}^y d\nu(t)$ (convention $G(y) \coloneqq \lim_{\varepsilon\to 0} G(y-\varepsilon)$). Set $G^{-1}(y) \coloneqq \inf\{y \mid G(y) > t\}$ and define $T \coloneqq G^{-1}\circ F\colon \mathbb{R}\to\mathbb{R}$.

   **Theorem 1.3.9.** *If $\mu$ has no atoms, then $T_\#\mu = \nu$.*

   To prove this theorem, we need some preliminary results.

   **Lemma 1.3.10.** *If $\mu$ has no atoms, then for all $t \in [0,1]$ we have*

   $$\mu(F^{-1}([0,t])) = t.$$

   *Proof.* The result is easily seen to be true for $t = 0$ and $t = 1$.

   Also, since $\mu$ has no atoms,

   $$|F(t_k) - F(t)| \le \int_t^{t_k} d\mu \xrightarrow[t_k\to t]{} 0,$$

   thus $F \in C^0(\mathbb{R},\mathbb{R})$. In particular $F$ is surjective on $(0,1)$. Thus, given $t \in (0,1)$, let $x \in \mathbb{R}$ be the largest point such that $F(x) = t$ (this point exists by the continuity of $F$). Hence

   $$\mu(F^{-1}([0,t])) = \int_{F^{-1}([0,t])} d\mu = \int_{-\infty}^x d\mu = t.$$

   □

**Corollary 1.3.11.** *If $\mu$ has no atoms, then for all $t \in [0,1]$ we have*

$$\mu(F^{-1}([0,t))) = t.$$

*Proof.* We apply Lemma 1.3.10 to the interval $[0, t-\varepsilon]$:

$$t = \mu(F^{-1}([0,t])) \geq \mu(F^{-1}([0,t))) \geq \mu(F^{-1}([0,t-\varepsilon])) = t - \varepsilon \xrightarrow[\varepsilon\to 0]{} t.$$

□

*Proof of Theorem 1.3.9.* We split the proof in steps.

Step 1. Let $A = (-\infty, a)$ with $a \in \mathbb{R}$. Then, applying Corollary 1.3.11,

$$\begin{aligned} T_{\#}\mu(A) &= \mu(T^{-1}(A)) \\ &= \mu(F^{-1}(G((-\infty,a)))) \\ &= \mu(F^{-1}([0, \lim_{\delta\to 0} G(a-\delta)])) \\ &= G(a^-) = \int_{-\infty}^{a^-} d\nu = \nu((-\infty,a)) = \nu(A). \end{aligned}$$

Step 2. Let $A = [a,b) = (-\infty,b) \setminus (-\infty,a)$. Then, applying Step 1,

$$\begin{aligned} T_{\#}\mu(A) &= T_{\#}\mu((-\infty,b)) - T_{\#}\mu((-\infty,a)) \\ &= \nu((-\infty,b)) - \nu((-\infty,a)) = \nu(A). \end{aligned}$$

Step 3. Let $A = (a,b) = \lim_{\varepsilon\to 0} A_\varepsilon$, with $A_\varepsilon := [a+\varepsilon, b)$. Then, thanks to Step 2,

$$\nu(A) \nwarrow \nu(A_\varepsilon) = T_{\#}\mu(A_\varepsilon) \nearrow T_{\#}\mu(A).$$

Step 4. Let $A$ be open, then $A = \bigcap_{i\in I}(a_i, b_i)$ with $\{(a_i,b_i)\}_{i\in I}$ disjoint and countable.

$$\nu(A) = \sum_{i\in I} \nu((a_i,b_i)) = \sum_{i\in I} T_{\#}\mu((a_i,b_i)) = T_{\#}\mu(A).$$

Since the open sets are generators of the $\sigma$-algebra, Step 4 proves that

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

□

3. The Knothe map (50's):

**Theorem 1.3.12.** (Disintegration Theorem) *Let $\mu \in \mathcal{P}(\mathbb{R}^2)$ and set $\mu_1 := (\pi_1)_{\#}\mu$, where $\pi_1 : \mathbb{R}^2 \to \mathbb{R}$ is defined as $\pi_1(x_1,x_2) = x_1$. Then there exists a family of probability measures $\{\mu_{x_1}\}_{x_1\in\mathbb{R}}$ such that*

$$\mu(dx_1, dx_2) = \mu_{x_1}(dx_2) \otimes \mu_1(dx_1),$$

*that is for any $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ continuous and bounded, we have:*

$$\int_{\mathbb{R}^2} \varphi(x_1,x_2) d\mu(x_1,x_2) = \int_{\mathbb{R}} \left[ \int_{\mathbb{R}} \varphi(x_1,x_2) d\mu_{x_1}(x_2) \right] d\mu(x_1).$$

*Also, the measures $\mu_{x_1}$ are unique $\mu_1$-a.e.*

For a proof of this result, see also the exercises at the end of these lectures.

**Example 1.3.13.** Let $\mu = f(x_1, x_2)dx_1dx_2$ with $\int f dx_1 dx_2 = 1$. Let $\mu_1 := (\pi_1)_{\#}\mu$. Then

$$\begin{aligned}\int_{\mathbb{R}} \varphi(x_1)d\mu_1(x_1) &= \int_{\mathbb{R}^2} \varphi(x_1)d\mu(x_1, x_2) \\ &= \int_{\mathbb{R}^2} \varphi(x_1)f(x_1, x_2)dx_1dx_2 \underbrace{=}_{Fubini} \int_{\mathbb{R}} \varphi(x_1)\big[\int f(x_1, x_2)dx_2\big]dx_1 \\ &= \int_{\mathbb{R}} \varphi(x_1)F_1(x_1)dx_1.\end{aligned}$$

Hence $\mu_1 = F_1 dx_1$.

Also, let $\mu_{x_1}(dx_2)$ be the disintegration provided by the previous theorem. Then

$$\begin{aligned}\iint \varphi(x_1, x_2)d\mu_{x_1}(x_2)d\mu_1(x_1) &= \int_{\mathbb{R}^2} \varphi(x_1, x_2)d\mu(x_1, x_2) \\ &= \int_{\mathbb{R}^2} \varphi(x_1, x_2)f(x_1, x_2)dx_1dx_2 \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \varphi(x_1, x_2)\frac{f(x_1, x_2)}{F_1(x_1)}dx_2\right] F_1(x_1)dx_1.\end{aligned}$$

Hence, by uniqueness of the disintegration,

$$\mu_{x_1}(dx_2) = \frac{1}{F_1(x_1)}f(x_1, x_2)dx_2 \qquad \mu_1 - a.e.$$

Note that $\mu_{x_1}$ are indeed probability measures:

$$\int_{\mathbb{R}} d\mu_{x_1}(x_2) = \frac{1}{F_1(x_1)}\int_{\mathbb{R}} f(x_1, x_2)dx_1 = \frac{1}{F_1(x_1)}F_1(x_1) = 1.$$

*Remark* 1.3.14. Note that $F_1 > 0$ $\mu_1$-a.e. ($\int_{\{F_1=0\}} d\mu_1 = \int_{\{F_1=0\}} F_1 dx_1 = \int_{\{F_1=0\}} 0dx_1 = 0$). "A measure lives where its density is positive".

**Knothe map:** Take two *absolutely continuous* measures on $\mathbb{R}^2$, namely $\mu(x_1, x_2) = f(x_1, x_2)dx_1dx_2 = \frac{f(x_1,x_2)}{F_1(x_2)}dx_2 \otimes F_1(x_1)dx_1$ and $\nu(y_1, y_2) = g(y_1, y_2)dy_1dy_2 = \frac{g(y_1,y_2)}{G_1(y_2)}dy_2 \otimes G_1(y_1)dy_1$, where

$$F_1(x_1) = \int_{\mathbb{R}} f(x_1, x_2)dx_2 \quad \text{and} \quad G_1(y_1) = \int_{\mathbb{R}} g(y_1, y_2)dy_2.$$

Using Theorem 1.3.9, we obtain a map $T_1 \colon \mathbb{R} \to \mathbb{R}$ such that $T_{1\#}(F_1dx_1) = G_1dy_1$. For $F_1dx_1$-a.e. $x_1 \in \mathbb{R}$, consider the monotone rearrangement $T_2(x_1, \cdot) \colon \mathbb{R} \to \mathbb{R}$ such that

$$T_2(x_1, \cdot)_{\#}\left(\frac{f(x_1, \cdot)dx_2}{F_1(x_1)}\right) = \frac{g(T_1(x_1), \cdot)}{G_1(T_1(x_1))}dy_2. \tag{1.2}$$

Claim: $T(x_1, x_2) = (T_1(x_1), T_2(x_1, x_2))$ transports $\mu$ to $\nu$.

*Proof.* For $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ Borel and bounded, we have

$$\begin{aligned}\int_{\mathbb{R}^2} \varphi(y_1,y_2)g(y_1,y_2)dy_1dy_2 &= \int_{\mathbb{R}} \underbrace{\left(\int_{\mathbb{R}} \varphi(y_1,y_2)\frac{g(y_1,y_2)}{G_1(y_1)}dy_2\right)}_{\Psi(y_1)} G(y_1)dy_1 \\ &\overset{(\dagger_1)}{=} \int_{\mathbb{R}} \Psi(T_1(x_1))F_1(x_1)dx_1 \\ &= \int_{\mathbb{R}}\left(\int_{\mathbb{R}} \frac{\varphi(T_1(x_1),y_2)g(T_1(x_1),y_2)}{G_1(T_1(x_1))}dy_2\right)F_1(x_1)dx_1 \\ &\overset{(\dagger_2)}{=} \int_{\mathbb{R}}\int_{\mathbb{R}} \varphi(T_1(x_1),T_2(x_1,x_2))f(x_1,x_2)dx_2dx_1 \\ &= \int_{\mathbb{R}^2} (\varphi\circ T)(x_1,x_2)d\mu(x_1,x_2),\end{aligned}$$

where we used $(T_1)_\#(F_1dx_1) = G_1dy_1$ at $(\dagger_1)$ and (1.2) at $(\dagger_2)$. □

*Remark* 1.3.15. Since monotone rearrangement is an increasing function, we have (under the assumption that the map $T(x_1,x_2) = (T_1(x_1), T_1(x_1,x_2))$ is smooth):

$$\nabla T = \begin{pmatrix} \partial_1 T_1 \geq 0 & * \\ 0 & \partial_2 T_2 \geq 0 \end{pmatrix}$$

One can use the previous construction of the Knothe map in $\mathbb{R}^2$ and iterate it to obtain a Knothe map on $\mathbb{R}^n$: let $\mu(x_1,\dots,x_n) = f(x_1,\dots,x_n)dx_1\dots dx_n$ and $\nu(y_1,\dots,y_n) = g(y_1,\dots,y_n)dy_1\dots dy_n$ be absolutely continuous measures. Using monotone rearrangement, we get a map $T_1\colon \mathbb{R}\to\mathbb{R}$ such that $T_{1\#}(F_1dx_1) = G_1dy_1$, where $F_1(x_1) = \int f dx_2\dots dx_n$ and $G_1(y_1) = \int g dy_2\dots dy_n$. Similarly both measures also yield new measures on $\mathbb{R}^{n-1}$ given by

$$\mu_{x_1}(x_2,\dots,x_n) = \frac{f(x_1,x_2,\dots,x_n)}{F_1(x_1)}dx_2\dots dx_n$$

and

$$\nu_{y_1}(y_2,\dots,y_n) = \frac{g(y_1,\dots,y_n)}{G_1(y_1)}dy_2\dots dy_n.$$

We can then reproduce the same construction on $\mathbb{R}^{n-1}$ and so on to obtain a Knothe map

$$T(x_1,\dots,x_n) = (T_1(x_1), T_2(x_1,x_2),\dots,T_n(x_1,x_2,\dots,x_n)).$$

*Remark* 1.3.16. Suppose again that the map $T$ is "nice", e.g. smooth, then

$$\nabla T = \begin{pmatrix} \partial_1 T_1 & * & * & * & * \\ 0 & \partial_2 T_2 & * & * & * \\ 0 & 0 & \ddots & * & * \\ 0 & 0 & 0 & \ddots & * \\ 0 & 0 & 0 & 0 & \partial_n T_n \end{pmatrix}.$$

Note that this is an *upper triangular matrix* and that all the values on the diagonal are *non-negative*. This is important for the next subsection.

*Remark* 1.3.17. Although we call it "the" Knothe map, note that the map itself is by no means unique. Indeed, by fixing a basis in $\mathbb{R}^n$ but changing the order of integration, one obtains a different Knothe map. Even more, changing the basis of $\mathbb{R}^n$, yields a different map.

## 1.4 Application: Isoperimetric inequalities

**Theorem 1.4.1.** *Let $E \subset \mathbb{R}^n$ be a bounded set with smooth boundary. Then*

$$\text{Area}(\partial E) \geq n|B_1|^{\frac{1}{n}}|E|^{\frac{n-1}{n}},$$

*where $|B_1|$ is the volume of the unit ball.*

To prove this result, consider the probability measures $\mu = \frac{\mathbb{1}_E}{|E|}dx$ and $\nu = \frac{\mathbb{1}_{B_1}}{|B_1|}dy$.

**Proposition 1.4.2.** *Let $T$ be a Knothe map from $\mu$ to $\nu$, and assume it to be smooth. Then:*

1. $\forall x \in E \Rightarrow |T(x)| \leq 1$.
2. $\det \nabla T = \frac{|B_1|}{|E|}$ *in* $E$.
3. $\text{div } T \geq n(\det \nabla T)^{\frac{1}{n}}$.

We will not discuss the regularity of the map $T$. In this example, assume that $T$ is smooth.

*Proof.* 1. If $x \in E$, then $T(x) \in B_1$.

2. Let $A \subset B_1$, then $T^{-1}(A) \subset E$. Since $T_{\#}\mu = \nu$, we have

$$\nu(A) = \mu(T^{-1}(A)) = \int_{T^{-1}(A)} \frac{dx}{|E|},$$

but we also have

$$\nu(A) = \int_A \frac{dy}{|B_1|} \overset{(\dagger)}{=} \int_{T^{-1}(A)} \frac{1}{|B_1|} \det \nabla T(x) dx,$$

where for $(\dagger)$ we used that $T$ is smooth, $\nabla T$ is upper triangular, its diagonal elements are non-negative (and thus $\det \nabla T \geq 0$), and the transformation formula $dy = \det \nabla T dx$. Since $A$ was arbitrary, we obtain

$$\frac{\det \nabla T}{|B_1|} = \frac{1}{|E|}.$$

3.

$$\begin{aligned}\text{div } T(x) &= \sum_{i=1}^n \underbrace{\partial_i T_i(x)}_{:=a_i \geq 0} = n\left(\frac{1}{n}\sum_{i=1}^n a_i\right) \\ &\overset{\diamondsuit}{\geq} n(\Pi_{i=1}^n a_i)^{\frac{1}{n}} = n(\Pi_{i=1}^n \partial_i T_i(x))^{\frac{1}{n}} \overset{\heartsuit}{=} n(\det \nabla T(x))^{\frac{1}{n}},\end{aligned}$$

where in $\diamondsuit$ we used that the arithmetic mean is greater than the geometric one, and in $\heartsuit$ we used that the matrix $\nabla T$ is upper-triangular to deduce that its determinant is given by the product of its diagonal elements.

□

*Proof of Theorem 1.4.1.*

$$\begin{aligned}\text{Area}(\partial E) &= \int_{\partial E} 1 d\sigma(x) \overset{1.}{\geq} \int_{\partial E} |T(x)| d\sigma(x) \geq \int_{\partial E} T(x) \cdot \nu_E(x) d\sigma(x) \\ &\overset{\dagger}{=} \int_E \text{div } T dx \overset{3.}{\geq} n \int_E (\det \nabla T(x))^{\frac{1}{n}} dx \overset{2.}{=} n \int_E \left(\frac{|B_1|}{|E|}\right)^{\frac{1}{n}} dx = n|B_1|^{\frac{1}{n}}|E|^{\frac{n-1}{n}},\end{aligned}$$

where we used *Stokes' Theorem* at †, and the numbering stands for the properties of Proposition 1.4.2. □

### 1.5 A Jacobian equation for transport maps

Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a smooth diffeomorphism with $\det \nabla T > 0$, and assume that $T_\#(f\,dx) = g\,dy$, where $f$ and $g$ are probability densities.

First of all, by the definition of push-forward, for any bounded Borel function $\zeta : \mathbb{R}^n \to \mathbb{R}$

$$\int \zeta(y) g(y) dy = \int \zeta(T(x)) f(x) dx.$$

On the other hand, using the substitution $y = T(x)$ we have $dy = \det \nabla T(x)\, dx$, hence

$$\int \zeta(y) g(y) dy = \int \zeta(T(x)) g(T(x)) \det \nabla T(x) dx.$$

Thus, comparing the two equations above, since $\zeta$ is arbitrary we deduce

$$f(x) = g(T(x)) \det \nabla T(x).$$

# 2 Optimal Transport

## 2.1 Preliminaries in measure theory

In this section $X$ will be a locally compact complete metric space. Again, the model case is $X = \mathbb{R}^n$. Every measure here will be in $\mathcal{P}(X)$.

Note: the assumptions in these notes are far from being sharp, the goal is to emphasize the main ideas in this theory. In particular, the existence of optimal transport plans (Theorem 2.3.2) and the duality theorem (Theorem 2.5.5) hold in arbitrary separable metric spaces. The interested reader may look at [AGS08, Chapters 5.1-5.4 and 6.1].

*Remark* 2.1.1. By the *Riesz representation Theorem* we have the following equalities:

$$\begin{aligned}\mathcal{M}(X) &\coloneqq \{\text{Finite signed measures on X}\}\\ &\cong C_c(X)^* \coloneqq \{\text{continuous compactly supported functions}\}^*\\ &\cong C_0(X)^* \coloneqq \{\text{continuous functions vanishing at } \infty\}^*.\end{aligned}$$

*Remark* 2.1.2. Note that $C_c(X)$ is not closed if $X$ is not compact.
E.g., for $X = \mathbb{R}$, the sequence

$$f_n(x) \coloneqq \frac{1}{1+x^2}\psi_n(x),$$

with $0 \le \psi_n \le 1$ continuous functions satisfying

$$\psi_n(x) = \begin{cases} 1 & \forall x \in [-n, n] \\ 0 & \forall x \in [-n-1, n+1] \end{cases},$$

converges towards $f(x) = \frac{1}{1+x^2} \notin C_c(\mathbb{R})$.

Let $\{\mu_k\}_{k\in\mathbb{N}}$ be a family of probability measures. Then $\mu_k(X) = 1$ and therefore the measures $\{\mu_k\}$ are bounded inside $\mathcal{M}(X)$. Thus it follows by Banach-Alaoglu that there exists a family $\{\mu_{k_j}\}_{j\in\mathbb{N}}$ such that

$$\mu_{k_j} \overset{*}{\rightharpoonup} \mu \in \mathcal{M}(X)$$

($\overset{*}{\rightharpoonup}$ means with respect to the weak-$*$ topology), i.e.,

$$\forall \varphi \in C_c(X)\colon \int \varphi d\mu_{k_j} \to \int \varphi d\mu.$$

Note that, since $\mu_k \ge 0$ (by assumption), we have that $\mu \ge 0$.
Issue: $\mu$ may not be a probability measure.

**Example 2.1.3.** Let $X = \mathbb{R}$ and $\mu_k = \delta_k$ for $k \in \mathbb{Z}$. Then, for any $\varphi \in C_c(\mathbb{R})$,

$$\int \varphi d\mu_k = \varphi(k) \xrightarrow{k\to\infty} 0,$$

since the point $k \in \mathbb{R}$ leaves the compact support of $\varphi$ for $k$ large enough. Hence $\mu_k \overset{*}{\rightharpoonup} 0$. This shows that, in general, the weak$^*$ limit of probability measures may not be a probability measure.

To resolve that issue, we need to introduce a stronger notion of convergence.

**Definition 2.1.4.** We say that $\mu_k$ **converges to** $\mu$ **narrowly** if, for all $\varphi \in C_b(X) \coloneqq \{\text{continuous bounded functions}\}$, we have

$$\int \varphi d\mu_k \to \int \varphi d\mu.$$

We denote this by $\mu_k \rightharpoonup \mu$.

*Remark* 2.1.5. Assume $\mu_k \in \mathcal{P}(X)$ and $\mu_k \rightharpoonup \mu$. Then, taking $\varphi \equiv 1$ yields

$$\mu_k(X) = \int_X 1 d\mu_k \to \int_X 1 d\mu = \mu(X).$$

Hence $\mu \in \mathcal{P}(X)$.

**Example 2.1.6.** Take $X = \mathbb{R}^n$ and $\mu_k = (1 - \frac{1}{k})\delta_0 + \frac{1}{k}\delta_{x_k}$ for some $x_k \in \mathbb{R}^n$. Then, if $\varphi \in C_b(\mathbb{R}^n)$, we have

$$\int \varphi d\mu_k = \left(1 - \frac{1}{k}\right)\varphi(0) + \frac{1}{k}\varphi(x_k) \xrightarrow{k\to\infty} \varphi(0),$$

so $\mu_k \rightharpoonup \mu = \delta_0$.

*Remark* 2.1.7. The issue when $\mu_k \overset{*}{\rightharpoonup} \mu$ is that some mass of $\mu_k$ may escape to $\infty$.

**Definition 2.1.8.** Let $\mathcal{A} \subset \mathcal{P}(X)$ be a family of probability measures. We say that $\mathcal{A}$ is **tight** if for any $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset X$ such that $\mu(X \setminus K_\varepsilon) \leq \varepsilon$ for any $\mu \in \mathcal{A}$.

**Theorem 2.1.9.** (Prokhorov) *$\mathcal{A} \subset \mathcal{P}(X)$ is tight if and only if $\mathcal{A}$ is relatively compact for the narrow convergence, i.e., for any sequence $\{\mu_k\} \subset \mathcal{A}$ there exists a subsequence $\{\mu_{k_j}\}$ and a probability measure $\mu \in \mathcal{P}(X)$ such that*

$$\mu_{k_j} \rightharpoonup \mu.$$

This theorem will not be proven in the lecture. See the exercises at the end of these lecture notes for a sketch of the proof.

*Remark* 2.1.10. The $\Rightarrow$ implication is the one that will be majoritarily used in the lecture.

**Lemma 2.1.11.** (Criterion for tightness) *If there is $\Phi\colon X \to \mathbb{R}_+$ such that $\{\Phi \leq \lambda\}$ is compact for any $\lambda \in \mathbb{R}$ and $\int_X \Phi d\mu \leq C_0$ for all $\mu \in \mathcal{A}$, then $\mathcal{A}$ is tight.*

*Proof.* Fix $\varepsilon > 0$. Then

$$\mu(\{\Phi > \lambda\}) \leq \int_{\{\Phi \geq \lambda\}} \frac{\Phi}{\lambda} d\mu = \frac{1}{\lambda}\int_{\{\Phi \geq \lambda\}} \Phi d\mu \leq \frac{1}{\lambda}\int_X \Phi d\mu \leq \frac{C_0}{\lambda} = \varepsilon,$$

for $\lambda = \frac{C_0}{\varepsilon}$. Since $K_\varepsilon \coloneqq \{\Phi \leq \frac{C_0}{\varepsilon}\}$ is compact, we have proven the tightness of $\mathcal{A}$. □

**Lemma 2.1.12.** (weak-$*$ convergence + mass conservation = narrow convergence) *Let $\{\mu_k\} \subset \mathcal{P}(X)$ be a sequence such that $\mu_k \overset{*}{\rightharpoonup} \mu$ for some $\mu \in \mathcal{P}(X)$. Then the set $\{\mu_k\}$ is tight and $\mu_k \rightharpoonup \mu$.*

*Proof.* Choose $\varepsilon > 0$. Since the singleton $\{\mu\}$ is clearly compact, thanks to Theorem 2.1.9, there is a compact set $K_\varepsilon$ such that $\mu(K_\varepsilon) \geq 1 - \varepsilon$. Using the fact that $X$ is a locally compact metric space[1], we can find a compactly supported function $\eta_\varepsilon \in C_c(X)$ such that $0 \leq \eta_\varepsilon \leq 1$ everywhere and $\eta_\varepsilon \equiv 1$ in $K_\varepsilon$.

[1]To build the function $\eta_\varepsilon$, first we should construct a larger compact set $H_\varepsilon$ such that $K_\varepsilon \subseteq \mathring{H}_\varepsilon$ and then apply Tietze extension theorem. We leave the details to the interested reader.

Hence it holds

$$\int_X \eta_\varepsilon d\mu \geq 1 - \varepsilon.$$

Since $\mu_k \overset{*}{\rightharpoonup} \mu$ and $\eta_\varepsilon \in C_c(X)$, we have

$$\int_X \eta_\varepsilon d\mu_k \to \int_X \eta_\varepsilon d\mu \geq 1 - \varepsilon \qquad \text{as } k \to \infty.$$

Hence there exists $k_\varepsilon$ such that, for any $k \geq k_\varepsilon$,

$$\mu_k(\mathrm{supp}(\eta_\varepsilon)) \geq \int_X \eta_\varepsilon d\mu_k \geq 1 - 2\varepsilon.$$

Also, for each $k < k_\varepsilon$, with the same approach we adopted for $\mu$, we can find a compact set $K_{\varepsilon,k}$ such that

$$\mu_k(K_{\varepsilon,k}) \geq 1 - 2\varepsilon.$$

Set $\hat{K}_\varepsilon \coloneqq \mathrm{supp}(\eta_\varepsilon) \cup \bigcup_{k=1}^{k_\varepsilon} K_{\varepsilon,k}$. Since it is a finite union of compact sets, $\hat{K}_\varepsilon$ is compact and it holds $\mu_k(\hat{K}_\varepsilon) \geq 1 - 2\varepsilon$ for any $k$, thus $\{\mu_k\}$ is tight.

Hence, given any subsequence $\mu_{k_j}$, thanks to Theorem 2.1.9 there exists a subsequence $\mu_{k_{j_l}}$ such that $\mu_{k_{j_l}} \rightharpoonup \nu \in \mathcal{P}(X)$. On the other hand, since $\mu_k \overset{*}{\rightharpoonup} \mu$, hence $\mu_{k_{j_l}} \overset{*}{\rightharpoonup} \mu$. Therefore, for any $\varphi \in C_c(X)$ we have

$$\int \varphi d\nu \leftarrow \int \varphi d\mu_{k_{j_l}} \to \int \varphi d\mu.$$

This implies that $\mu = \nu$. In other words, for any narrowly converging subsequence of $\mu_k$, the limit is independent of the choice of the subsequence. This implies that the whole sequence $\mu_k$ narrowly converges to $\mu$, as desired. □

**Lemma 2.1.13.** (lower semi-continuity of integrals) *Let $\mu_k \rightharpoonup \mu$ and let $\varphi\colon X \to [-C, +\infty]$ for some $C \in \mathbb{R}$ be a lower semi-continuous (l.s.c) function (i.e., if $x_k \to x$ then $\liminf_{k\to\infty} \varphi(x_k) \geq \varphi(x)$). Then*

$$\liminf_{k\to\infty} \int_X \varphi d\mu_k \geq \int_X \varphi d\mu.$$

*Proof.* Up to replacing $\varphi$ by $\varphi + C$, without loss of generality $\varphi \geq 0$. If $\varphi \equiv +\infty$ then the statement is trivial, hence we assume that this is not the case. Given $\lambda \geq 0$, define

$$\varphi_\lambda(x) \coloneqq \inf_{y\in X} \{\varphi(y) + \lambda d(x, y)\}.$$

Facts:

(a) $\forall \lambda < \lambda'\colon \varphi_\lambda \leq \varphi_{\lambda'} \leq \varphi$;

(b) $\varphi_\lambda$ is $\lambda$-Lipschitz;

(c) For each $x \in X$ it holds $\varphi_\lambda(x) \nearrow \varphi(x)$ as $\lambda \to \infty$.

Proof of the Facts:

(a) For any $y \in X$ we have $\varphi_\lambda(x) \leq \varphi(y) + \lambda d(x, y) \leq \varphi(y) + \lambda' d(x, y)$. Take the infimum over $y \in X$.

(b) Let $x, x' \in X$. Then $\varphi(x') \leq \varphi(y) + \lambda d(x,y) \leq \varphi(y) + \lambda d(x,y) + \lambda d(x',x)$. By taking the infimum over $y$, this yields $\varphi_\lambda(x') \leq \varphi_\lambda(x) + \lambda d(x,x')$. Since the argument is symmetric in $x$ and $x'$, we obtain

$$|\varphi_\lambda(x) - \varphi_\lambda(x')| \leq \lambda d(x,x').$$

(c) Fix $x \in X$. Since $\varphi$ is l.s.c., for all $\varepsilon > 0$ there exists a $\delta > 0$ such that $\varphi(y) \geq \varphi(x) - \varepsilon$ for $y \in X$ with $d(x,y) \leq \delta$.[2] Thus, recalling that $\varphi \geq 0$, we have

- $\varphi(y) + \lambda d(x,y) \geq \varphi(x) - \varepsilon$ if $d(x,y) \leq \delta$;
- $\varphi(y) + \lambda d(x,y) \geq \lambda \cdot \delta$ if $d(x,y) > \delta$.

Thus $\varphi_\lambda(x) \geq \min\{\varphi(x) - \varepsilon, \lambda \cdot \delta\}$. Letting $\lambda \to \infty$, this implies that $\liminf_{\lambda\to\infty} \varphi_\lambda(x) \geq \varphi(x) - \epsilon$. Since $\varepsilon$ is arbitrary, we obtain $\liminf_{\lambda\to\infty} \varphi_\lambda(x) \geq \varphi(x)$. Part (a) yields the other inequality, thus

$$\lim_{\lambda\to\infty} \varphi_\lambda(x) = \varphi(x).$$

Note that since $\varphi \geq 0$, we have $\varphi_\lambda \geq 0$. So to prove the Lemma, set

$$\varphi_{\lambda,M}(x) \coloneqq \min\{\varphi_\lambda(x), M\},$$

for $M > 0$. Then $\varphi_{\lambda,M}$ is continuous and bounded, hence an element of $C_b(X)$. By narrow convergence, and since $\varphi_{\lambda,M} \leq \varphi_\lambda \leq \varphi$:

$$\int \varphi_{\lambda,M} d\mu = \lim_{k\to\infty} \int \varphi_{\lambda,M} d\mu_k \leq \liminf_{k\to\infty} \int \varphi d\mu_k.$$

By monotone convergence and since $\varphi_{\lambda,M} \nearrow \varphi$ as $\lambda, M \to \infty$, we obtain

$$\int \varphi d\mu = \lim_{\lambda,M\to\infty} \int \varphi_{\lambda,M} d\mu \leq \liminf_{k\to\infty} \int \varphi d\mu_k.$$

□

## 2.2 Monge vs. Kantorovich

Fix $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c\colon X \times Y \to [0, +\infty]$ l.s.c.
We have the following situation (recall Definition 1.3.4):

(M) $C_M(\mu,\nu) := \inf\{\int_X c(x, T(x)) \mid T_\#\mu = \nu\}$.

(K) $C_K(\mu,\nu) := \inf\{\int_{X\times Y} c(x,y) d\gamma(x,y) \mid \gamma \in \Gamma(\mu,\nu)\}$.

*Remark* 2.2.1. Recall that if $T_\#\mu = \nu$, then $\gamma_T \coloneqq (\mathrm{id} \times T)_\#\mu \in \Gamma(\mu,\nu)$. Also

$$\int_X c(x,T(x)) d\mu(x) = \int_X c \circ (\mathrm{id} \times T)(x) d\mu(x) = \int_{X\times Y} c(x,y) d\gamma_T(x,y),$$

therefore

$$C_M(\mu,\nu) \geq C_K(\mu,\nu).$$

*Remark* 2.2.2. Let $\gamma \in \Gamma(\mu,\nu)$ and assume that $\gamma = (\mathrm{id} \times S)_\#\mu$ for some map $S : X \to Y$. Then

$$\nu = (\pi_Y)_\#\gamma = (\pi_Y)_\#(\mathrm{id} \times S)_\#\mu = (\pi_Y \circ (\mathrm{id} \times S))_\#\mu = S_\#\mu.$$

Thus $S$ is a transport map. In other words, "if you have a transport plan with the structure of a graph, this yields a transport map".

[2] To be precise, we know that $\liminf_{k\to\infty} \varphi(x_k) \geq \varphi(x)$ if $x_k \to x$. Hence, if $\varphi(x) \in \mathbb{R}$, then for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $\varphi(y) \geq \varphi(x) - \varepsilon$ for $d(x,y) \leq \delta$. On the other hand, if $\varphi(x) = +\infty$, then for all $\varepsilon$ there exists $\delta > 0$ such that $\varphi(y) \geq \frac{1}{\varepsilon}$ for $d(x,y) \leq \delta$. So, if $\varphi(x) = +\infty$, then in the proof of (c) one should replace $\varphi(x) - \varepsilon$ by $\frac{1}{\varepsilon}$.

## 2.3 Existence of an optimal coupling

**Lemma 2.3.1.** *The set $\Gamma(\mu,\nu) \subset \mathcal{P}(X \times Y)$ is tight and closed under narrow convergence.*

*Proof.* 1. $\Gamma(\mu,\nu)$ is tight: the singleton $\{\mu\} \subset \mathcal{P}(X)$ is a compact family, thus $\{\mu\}$ is tight. Therefore, for all $\varepsilon > 0$, there exists a set $K_\varepsilon \subset X$ such that $\mu(X \setminus K_\varepsilon) \leq \frac{\varepsilon}{2}$. Analogously for $\nu$, there exists a set $\tilde{K}_\varepsilon$ such that $\nu(Y \setminus \tilde{K}_\varepsilon) \leq \frac{\varepsilon}{2}$. Define $\bar{K}_\varepsilon \coloneqq K_\varepsilon \times \tilde{K}_\varepsilon \subset X \times Y$. Then, for any $\gamma \in \Gamma(\mu,\nu)$, we have that

$$\begin{aligned}
\gamma((X \times Y) \setminus \bar{K}_\varepsilon) &= \gamma((X \setminus K_\varepsilon) \times Y \cup X \times (Y \setminus \tilde{K}_\varepsilon)) \\
&\leq \gamma((X \setminus K_\varepsilon) \times Y) + \gamma(X \times (Y \setminus \tilde{K}_\varepsilon)) \\
&= \int_{X\times Y} \mathbb{1}_{X\setminus K_\varepsilon}(x) d\gamma(x,y) + \int_{X\times Y} \mathbb{1}_{Y\setminus \tilde{K}_\varepsilon}(y) d\gamma(x,y) \\
&= \int_X \mathbb{1}_{X\setminus K_\varepsilon}(x) d\mu(x) + \int_Y \mathbb{1}_{Y\setminus \tilde{K}_\varepsilon}(y) d\nu(y) \\
&= \mu(X \setminus K_\varepsilon) + \nu(Y \setminus \tilde{K}_\varepsilon) \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

Thus $\Gamma(\mu,\nu)$ is tight.

2. $\Gamma(\mu,\nu)$ is closed under narrow convergence: let $\gamma_k \rightharpoonup \gamma$ be a sequence with $\{\gamma_k\} \subset \Gamma(\mu,\nu)$. For any $\varphi \in C_b(X)$, we have

$$\int_X \varphi(x) d\mu(x) = \int_{X\times Y} \varphi(x) d\gamma_k(x,y) \to \int_{X\times Y} \varphi(x) d\gamma(x,y),$$

thus $(\pi_X)_\# \gamma = \mu$. Analogously $(\pi_Y)_\# \gamma = \nu$.

□

**Theorem 2.3.2.** *Let $c\colon X \times Y \to [0,+\infty]$ be l.s.c., $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then there exists a coupling $\bar{\gamma} \in \Gamma(\mu,\nu)$ which is a minimizer for (K).*

*Remark* 2.3.3. If $\inf_{\gamma\in\Gamma(\mu,\nu)} \int c d\gamma = +\infty$, then there is nothing to do.

*Proof.* Without loss of generality $\alpha \coloneqq \inf_{\gamma\in\Gamma(\mu,\nu)} \int c d\gamma < +\infty$. Let $\{\gamma_k\} \subset \Gamma(\mu,\nu)$ be a sequence such that $\int_{X\times Y} c d\gamma_k \to \alpha$ as $k \to +\infty$. Since $\{\gamma_k\} \subset \Gamma(\mu,\nu)$ is tight, there exists a subsequence $\{\gamma_{k_j}\}$ such that $\gamma_{k_j} \rightharpoonup \bar{\gamma}$. Since $c$ is non-negative and l.s.c., it follows from Lemma 2.1.13 that

$$\int_{X\times Y} c d\bar{\gamma} \leq \liminf_{j\to\infty} \int c d\gamma_{k_j} = \alpha = \inf_{\gamma\in\Gamma(\mu,\nu)} \int c d\gamma.$$

Therefore $\int c d\bar{\gamma} = \alpha$ and since $\bar{\gamma} \in \Gamma(\mu,\nu)$ (thanks to Lemma 2.3.1), $\bar{\gamma}$ is a minimizer. □

The natural questions which arise are the following:

1. Is the minimizer $\gamma$ unique?
2. Is it given by a transport map?

Consider the next two examples:

**Example 2.3.4.** Let $\mu = \delta_{x_0}$ and $\nu = \frac{1}{2}\delta_{y_0} + \frac{1}{2}\delta_{y_1}$, then the unique element in $\Gamma(\mu,\nu)$ is $\gamma \equiv \frac{1}{2}\delta_{(x_0,y_0)} + \frac{1}{2}\delta_{(x_0,y_1)}$. It is clearly not induced by a transport map.

As for the first question:

**Example 2.3.5.** Let $X = Y = \mathbb{R}^2$, let $c(x,y) = |x-y|^2$, let $\nu = \frac{1}{2}\delta_{(1,0)} + \frac{1}{2}\delta_{(0,1)} = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$ and $\mu = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1)} = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}$. Then given any $\alpha \in [0, \frac{1}{2}]$,

$$\gamma_\alpha = \alpha\delta_{(x_1,y_1)} + \left(\frac{1}{2} - \alpha\right)\delta_{(x_1,y_2)} + \left(\frac{1}{2} - \alpha\right)\delta_{(x_2,y_1)} + \alpha\delta_{(x_2,y_2)}$$

is a probability measure. In addition

$$\int_{X\times Y} c d\gamma_\alpha = \alpha|x_1 - y_1|^2 + \left(\frac{1}{2} - \alpha\right)|x_1 - y_2|^2 + \left(\frac{1}{2} - \alpha\right)|x_2 - y_1|^2 + \alpha|x_2 - y_2|^2 = 1.$$

Hence all couplings $\gamma_\alpha$ are optimal, and thus the optimum is definitely not unique (in this example at least).

**Definition 2.3.6.** Let $\bar{\gamma}$ be a "*real*" minimizer (i.e., $\int c d\bar{\gamma} = \inf_{\gamma\in\Gamma(\mu,\nu)} \int c d\gamma < +\infty$). We recall that the **support** of $\bar{\gamma}$ is defined as

$$\operatorname{supp}(\bar{\gamma}) \coloneqq \{(x,y) \in X \times Y \mid \forall \varepsilon > 0 \colon \bar{\gamma}(B_\varepsilon(x) \times B_\varepsilon(y)) > 0\}$$

Morally speaking: "If $(x,y) \in \operatorname{supp}(\bar{\gamma})$, then some mass goes from $x$ to $y$".

Suppose, as an example, that $(x_i, y_i)_{i=1,2,3} \in \operatorname{supp}(\bar{\gamma})$. Now imagine to construct another "transport" where one takes the mass from $x_2$ to $y_1$, from $x_3$ to $y_2$, and from $x_1$ to $y_3$. Since $\bar{\gamma}$ is optimal, re-"shuffling" must increase the cost, i.e.

$$\sum_{i=1}^{3} c(x_{i+1}, y_i) \geq \sum_{i=1}^{3} c(x_i, y_i),$$

where $x_4 \equiv x_1$ (we shall prove this fact later, in full rigor). Since this property needs to hold for any collection of points in the support of $\bar{\gamma}$, this idea yields the following definition:

**Definition 2.3.7.** A set $\Lambda \subset X \times Y$ is said to be **$c$-cyclically monotone** if for any finite subset $\{(x_i, y_i)\}_{i=1,\dots,N} \subset \Lambda$, the following holds:

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

where $x_{N+1} \equiv x_1$.

**Theorem 2.3.8.** *Let $\bar{\gamma}$ be optimal and $c\colon X \times Y \to \mathbb{R}$ continuous, then $\operatorname{supp}(\bar{\gamma})$ is $c$-cyclically monotone.*

*Remark* 2.3.9. We will see later that it is actually an *if and only if*, see Theorem 2.5.2.

*Proof.* By contradiction, suppose $\operatorname{supp}(\bar{\gamma})$ is not $c$-cyclically monotone. Then there exist $N$ pairs $(x_1, y_1), \dots, (x_N, y_N)$ such that

$$\sum_{i=1}^{N} c(x_i, y_i) \geq \sum_{i=1}^{N} c(x_{i+1}, y_i) + \eta, \tag{2.1}$$

for some $\eta > 0$. Since $c$ is continuous, there exist open neighbourhoods $x_i \in U_i \subset X$ and $y_i \in V_i \subset Y$ such that, for any $(x,y) \in U_i \times V_i$, we have

$$|c(x,y) - c(x_i, y_i)| \leq \frac{\eta}{4N} \tag{2.2}$$

and for any $(x, y) \in U_{i+1} \times V_i$

$$|c(x, y) - c(x_{i+1}, y_i)| \leq \frac{\eta}{4N}. \tag{2.3}$$

Set $\varepsilon_i := \bar{\gamma}(U_i \times V_i)$. Note that all $\varepsilon_i$ are positive, since $(x_i, y_i)$ belong to the support of $\bar{\gamma}$. Now set $\varepsilon := \min_{i=1,\dots,N} \varepsilon_i$ and $\gamma_i := \frac{\bar{\gamma}|_{U_i \times V_i}}{\varepsilon_i} \in \mathcal{P}(X \times Y)$.[3] Then we define the measures $\mu_i := (\pi_X)_\# \gamma_i \in \mathcal{P}(X)$ and $\nu_i := (\pi_Y)_\# \gamma_i \in \mathcal{P}(Y)$, and we set

$$\gamma' := \bar{\gamma} - \frac{\varepsilon}{N} \sum_{i=1}^N \gamma_i + \frac{\varepsilon}{N} \sum_{i=1}^N \mu_{i+1} \otimes \nu_i.$$

Claim:

1. $\gamma' \geq 0$: Indeed, since $\varepsilon \leq \varepsilon_i$, we have

$$\begin{aligned}\gamma' &\geq \bar{\gamma} - \frac{\varepsilon}{N} \sum_{i=1}^N \gamma_i = \bar{\gamma} - \frac{1}{N} \sum_{i=1}^N \frac{\varepsilon}{\varepsilon_i} \bar{\gamma}|_{U_i \times V_i} \\ &\geq \bar{\gamma} - \frac{1}{N} \sum_{i=1}^N \bar{\gamma}|_{U_i \times V_i} \geq \bar{\gamma} - \frac{1}{N} \sum_{i=1}^N \bar{\gamma} = 0.\end{aligned}$$

2. $\gamma' \in \Gamma(\mu, \nu)$: Since $(\pi_X)_\# \bar{\gamma} = \mu$, $(\pi_X)_\# \gamma_i = \mu_i$, and $(\pi_X)_\#(\mu_{i+1} \otimes \nu_i) = \mu_{i+1}$, we have that

$$(\pi_X)_\# \gamma' = \mu - \frac{\varepsilon}{N} \sum_{i=1}^N \mu_i + \frac{\varepsilon}{N} \sum_{i=1}^N \mu_{i+1} = \mu.$$

   Analogously $(\pi_Y)_\# \gamma' = \nu$.

We now show that $\int c d\gamma' < \int c d\bar{\gamma}$ and this yields our contradiction, since $\bar{\gamma}$ was assumed to be optimal. Note that, since $\mu_i$ is supported inside $U_i$ and $\nu_i$ is supported inside $V_i$, it follows from (2.3) that

$$\begin{aligned}\int c d(\mu_{i+1} \otimes \nu_i) &= \int_{U_{i+1} \times V_i} c(x, y) d(\mu_{i+1} \otimes \nu_i) \\ &\leq \int_{U_{i+1} \times V_i} \left[ c(x_{i+1}, y_i) + \frac{\eta}{4N} \right] d(\mu_{i+1} \otimes \nu_i) \\ &= c(x_{i+1}, y_i) + \frac{\eta}{4N}.\end{aligned}$$

Analogously, since $\gamma_i$ is supported inside $U_i \times V_i$,

$$\int c d\gamma_i = \int_{U_i \times V_i} c d\gamma_i \geq \int_{U_i \times V_i} \left[ c(x_i, y_i) - \frac{\eta}{4N} \right] d\gamma_i = c(x_i, y_i) - \frac{\eta}{4N}.$$

Then, recalling (2.1),

$$\begin{aligned}\int c d\bar{\gamma} - \int c d\gamma' &= \frac{\varepsilon}{N} \sum_{i=1}^N \left[ \int c d\gamma_i - \int c d(\mu_{i+1} \otimes \nu_i) \right] \\ &\geq \frac{\varepsilon}{N} \sum_{i=1}^N \left[ c(x_i, y_i) - \frac{\eta}{4N} - \left( c(x_{i+1}, y_i) + \frac{\eta}{4N} \right) \right] \\ &\geq \frac{\varepsilon}{N} \sum_{i=1}^N [c(x_i, y_i) - c(x_{i+1}, y_i)] - \frac{\varepsilon}{N} \frac{\eta}{2} \\ &\geq \frac{\varepsilon}{N} \eta - \frac{\varepsilon}{N} \frac{\eta}{2} = \frac{\varepsilon}{N} \frac{\eta}{2} > 0,\end{aligned}$$

[3] Here we are using the notation $\bar{\gamma}|_A$ to denote the restriction of the measure $\bar{\gamma}$ to the set $A$: namely, for any Borel set $E \subset X$, $\bar{\gamma}|_A(E) = \bar{\gamma}(A \cap E)$).

a contradiction that concludes the proof. □

## 2.4 The case $c(x,y) = \frac{|x-y|^2}{2}$ on $X = Y = \mathbb{R}^n$

Let $X = Y = \mathbb{R}^n$ and $c(x,y) = \frac{|x-y|^2}{2}$. Let $\gamma \in \Gamma(\mu,\nu)$, then

$$\begin{aligned}\int \frac{|x-y|^2}{2} d\gamma(x,y) &= \int \left( \frac{|x|^2}{2} + \frac{|y|^2}{2} - x \cdot y \right) d\gamma \\ &= \int \frac{|x|^2}{2} d\mu + \int \frac{|y|^2}{2} d\nu + \int -x \cdot y d\gamma.\end{aligned} \tag{2.4}$$

Thus, provided $\int \frac{|x|^2}{2} d\mu + \int \frac{|y|^2}{2} d\nu < \infty$, $\gamma$ is optimal for the cost $c(x,y) = \frac{|x-y|^2}{2}$ if and only if it is optimal for the cost $c(x,y) = -x \cdot y$. In this case, the equation

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

is equivalent to

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

where $\langle \cdot, \cdot \rangle = \cdot$ is the canonical scalar product on $\mathbb{R}^n$. Any set satisfying this last property is simply called *cyclically monotone set*.

**Definition 2.4.1.** Given $\varphi \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ convex, we define the **subdifferential** of $\varphi$ as

$$\partial\varphi(x) \coloneqq \{y \in \mathbb{R}^n \mid \forall z \in \mathbb{R}^n \colon \varphi(z) \geq \varphi(x) + \langle y, z - x \rangle\}.$$

We set $\partial\varphi \coloneqq \bigcup_{x \in \mathbb{R}^n} \{x\} \times \partial\varphi(x) \subset \mathbb{R}^n \times \mathbb{R}^n$.

**Theorem 2.4.2.** (Rockafellar) *$S$ is cyclically monotone if and only if there exists a convex $\varphi$, such that $S \subset \partial\varphi$.*

*Proof.* $\Leftarrow$ Take $(x_i, y_i)_{i=1,\dots,N} \subset S \subset \partial\varphi$. Then for each $i$, since $y_i \in \partial\varphi(x_i)$, for any $z \in \mathbb{R}^n$ the following inequality holds:

$$\varphi(z) \geq \varphi(x_i) + \langle y_i, z - x_i \rangle.$$

Thus for $z = x_{i+1}$, we obtain

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

and summing over $i$ (where $N + 1 \equiv 1$) yields

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

This implies that

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

since the two summands containing $\varphi$ are equal.

$\Rightarrow$ Fix $(x_0, y_0) \in S$ and define

$$\varphi(x) \coloneqq \sup\{\langle y_N, x - x_N\rangle + \langle y_{N-1}, x_N - x_{N-1}\rangle + \dots \langle y_0, x_1 - x_0\rangle \mid (x_i, y_i)_{i=1,\dots,N} \subset S\}.$$

Note that

- $\varphi$ is a supremum of affine functions, thus it is convex;
- The case $N = 1$, i.e. $(x_1, y_1) = (x_0, x_0)$, yields

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

  and in particular $\varphi(x_0) \geq 0$.
- $\varphi \not\equiv +\infty$: for any $(x_i, y_i)_{i=1,\dots,N} \subset S$, because of cyclic monotonicity, we have $\langle y_N, x_0 - x_N\rangle + \dots + \langle y_0, x_1 - x_0\rangle \leq 0$. Hence $\varphi(x_0) \leq 0$, and therefore $\varphi(x_0) = 0$ (by the previous bound $\varphi(x_0) \geq 0$).
- $S \subset \partial\varphi$: Take $(\bar{x}, \bar{y}) \in S$ and let $\alpha < \varphi(\bar{x})$. Then, by definition of $\varphi$, there exists a sequence $(x_i, y_i)_{i=1,\dots,N}$ such that

  $$\langle y_N, \bar{x} - x_N\rangle + \dots + \langle y_0, x_1 - x_0\rangle \geq \alpha.$$

  Consider the sequence $(x_i, y_i)_{i=1,\dots,N+1}$ obtained by taking $(x_{N+1}, y_{N+1}) = (\bar{x}, \bar{y})$. Then for any $z \in \mathbb{R}^n$, we have

  $$\varphi(z) \geq \langle \underbrace{y_{N+1}}_{=\bar{y}}, z - \overbrace{x_{N+1}}^{=\bar{x}}\rangle + \langle y_N, \overbrace{x_{N+1}}^{=\bar{x}} - x_N\rangle + \dots + \langle y_0, x_1 - x_0\rangle \geq \langle \bar{y}, z - \bar{x}\rangle + \alpha,$$

  where the first inequality follows again by the definition of $\varphi$. Thus for all $z \in \mathbb{R}^n$ and any $\alpha < \varphi(\bar{x})$ we have

  $$\varphi(z) \geq \langle \bar{y}, z - \bar{x}\rangle + \alpha.$$

  By letting $\alpha \to \varphi(\bar{x})$, we obtain that for any $z \in \mathbb{R}^n$

  $$\varphi(z) \geq \langle \bar{y}, z - \bar{x}\rangle + \varphi(\bar{x}),$$

  This proves that $\bar{y} \in \partial\varphi(\bar{x})$, or equivalently $(\bar{x}, \bar{y}) \in \partial\varphi$, as desired.

□

### 2.4.1 Duality

**Definition 2.4.3.** Given $\varphi\colon \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ convex (with $\varphi \not\equiv +\infty$), one defines the **Legendre transform**

$$\varphi^*\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$$

of $\varphi$ as

$$\varphi^*(y) \coloneqq \sup_{x\in\mathbb{R}^n} \{x \cdot y - \varphi(x)\}.$$

**Proposition 2.4.4.** (a) $\forall x, y \in \mathbb{R}^n\colon \varphi(x) + \varphi^*(y) \geq x \cdot y$;

(b) $\varphi(x) + \varphi^*(y) = x \cdot y$ *if and only if* $y \in \partial\varphi(x)$.

*Proof.* (a) For any $x \in \mathbb{R}^n$, it follows by the definition of $\varphi^*$ that $\varphi^*(y) \geq x \cdot y - \varphi(x)$, thus $\varphi^*(y) + \varphi(x) \geq x \cdot y$.

(b) $\Rightarrow$ Assume that $\varphi(x) + \varphi^*(y) = x \cdot y$. By (a), we know that for any $z \in \mathbb{R}^n$, we have

$$\varphi^*(y) \geq z \cdot y - \varphi(z).$$

Then $x \cdot y - \varphi(x) = \varphi^*(y) \geq z \cdot y - \varphi(z)$, which is equivalent to

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

Since $z$ is arbitrary, this proves the result.

$\Leftarrow$ If $y \in \partial\varphi(x)$ then for any $z \in \mathbb{R}^n$ we have $\varphi(z) \geq \varphi(x) + \langle y, z - x \rangle$. This is equivalent to

$$x \cdot y - \varphi(x) \geq z \cdot y - \varphi(z).$$

Then by taking the supremum over $z \in \mathbb{R}^n$ we get

$$x \cdot y - \varphi(x) \geq \varphi^*(y),$$

and by (a) we obtain equality.

□

In the next theorem, we prove the so-called Kantorovich duality. Note that the existence of an optimal coupling for the cost function $c(x, y) = -x \cdot y$ is not immediate, since we only proved existence of an optimal coupling for positive cost function. However, we can use that the cost $c(x, y) = -x \cdot y$ is equivalent to the cost $c'(x, y) = \frac{|x-y|^2}{2}$ provided that $\int \frac{|x|^2}{2} d\mu + \int \frac{|y|^2}{2} d\nu < \infty$ . Hence, using this fact, we can apply Theorem 2.3.2 to obtain the existence of an optimal coupling for the cost $c'$, and then use that this coupling is also optimal for our cost $c$. Recall that, in order to avoid "trivial" situations where all couplings are optimal, we assume that $\inf_{\gamma \in \Gamma(\mu,\nu)} \int \frac{|x-y|^2}{2} d\gamma < +\infty$.

**Theorem 2.4.5.** (Kantorovich duality) *Let $c(x, y) = -x \cdot y$, and assume that*

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int \frac{|x - y|^2}{2} d\gamma + \int \frac{|x|^2}{2} d\mu + \int \frac{|y|^2}{2} d\nu < +\infty.$$

*Then for any $\gamma \in \Gamma(\mu, \nu)$ and $\varphi, \psi \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ measurable, the following holds:*

$$\min_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} -x \cdot y d\gamma(x, y) = \max_{\varphi(x)+\psi(y) \geq x \cdot y} \int_{\mathbb{R}^n} -\varphi(x) d\mu(x) + \int_{\mathbb{R}^n} -\psi(y) d\nu(y).$$

*Proof.* Consider $\varphi, \psi \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that for all $x, y \in \mathbb{R}^n$, we have

$$\varphi(x) + \psi(y) \geq x \cdot y.$$

Then this implies that

$$\begin{aligned} \int_{\mathbb{R}^n \times \mathbb{R}^n} -x \cdot y d\gamma(x, y) &\geq \int_{\mathbb{R}^n} -\varphi(x) d\gamma(x, y) + \int_{\mathbb{R}^n} -\psi(y) d\gamma(x, y) \qquad (\star) \\ &= \int_{\mathbb{R}^n} -\varphi(x) d\mu(x) + \int_{\mathbb{R}^n} -\psi(y) d\nu(y). \end{aligned}$$

Note that the left hand side does not depend on $\varphi$ and $\psi$, and that the right hand side does not depend on $\gamma$. Thus

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} -x \cdot y d\gamma(x, y) \geq \sup_{\varphi(x)+\psi(y) \geq x \cdot y} \int_{\mathbb{R}^n} -\varphi(x) d\mu(x) + \int_{\mathbb{R}^n} -\psi(y) d\nu(y).$$

On the other hand, let $\bar{\gamma} \in \Gamma(\mu,\nu)$ be optimal. Then Theorem 2.3.8 implies that $\mathrm{supp}(\bar{\gamma})$ is cyclically monotone. Thus by Rockafellar's Theorem 2.4.2, there exists a convex map $\varphi\colon \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ such that $\mathrm{supp}(\bar{\gamma}) \subset \partial\varphi$, that is, for any $(\bar{x},\bar{y}) \in \mathrm{supp}(\bar{\gamma})$ we have $\bar{y} \in \partial\varphi(\bar{x})$. Thanks to Proposition 2.4.4, we then have that $\varphi(\bar{x}) + \varphi^*(\bar{y}) = \bar{x}\cdot\bar{y}$. Therefore, this proves that $\varphi(x)+\varphi(y) = x\cdot y$ for $\bar{\gamma}$ a.e. $(x,y)$, and thus

$$\begin{aligned}\int_{\mathbb{R}^n\times\mathbb{R}^n} -x\cdot y d\bar{\gamma}(x,y) &= \int_{\mathbb{R}^n} -\varphi(x) d\bar{\gamma}(x,y) + \int_{\mathbb{R}^n} -\varphi^*(y) d\bar{\gamma}(x,y)\\ &= \int_{\mathbb{R}^n} -\varphi(x) d\mu(x) + \int_{\mathbb{R}^n} -\varphi^*(y) d\nu(y).\end{aligned}$$

Hence $(\bar{\gamma},\varphi,\varphi^*)$ gives equality in equation $(\star)$. □

*Remark* 2.4.6. The proof shows that if $\mathrm{supp}(\bar{\gamma})$ is cyclically monotone, then

$$\int -x\cdot y d\bar{\gamma} = \int -\varphi d\mu + \int -\varphi^* d\nu \le \inf_{\gamma\in\Gamma(\mu,\nu)} \int -x\cdot y d\gamma.$$

This implies that $\bar{\gamma}$ is optimal. So we proved the implication:

$$\mathrm{supp}(\bar{\gamma}) \text{ is cyclically monotone} \quad\Rightarrow\quad \bar{\gamma} \text{ is optimal.}$$

As a consequence of this remark, together with Theorems 2.3.8 and 2.4.2, we obtain the following:

**Corollary 2.4.7.** *The following are equivalent:*

- $\bar{\gamma}$ *is optimal;*
- $\mathrm{supp}(\bar{\gamma})$ *is cyclically monotone;*
- *There exists a convex map* $\varphi$ *such that* $\mathrm{supp}(\bar{\gamma}) \subset \partial\varphi$.

*Remark* 2.4.8. This last equivalence is often used to prove that some coupling of a transport map is optimal. Given a transport map $T$, if one can construct a convex $\varphi$ such that $\mathrm{supp}((\mathrm{id}\times T)_{\#}\mu) \subset \partial\varphi$, then $(\mathrm{id}\times T)_{\#}\mu$ is optimal.

## 2.5 Kantorovich theorem for general cost functions

General cost: $c \in C^0(X\times Y)$, for $X$ and $Y$ metric spaces. The goal of this section is to obtain what we did in the previous section (for the case $c(x,y) = -x\cdot y$ on $X = Y = \mathbb{R}^n$) but for general costs. As we shall see, the proofs are essentially identical, provided that one introduces the correct definitions.

First, we need a suitable analogue of the notion of convex function. Note that a possible way to define convex functions is as supremum of affine functions. Namely, a function $\phi$ is convex if it can be written as

$$\phi(x) = \sup_{y\in Y}\{x\cdot y + \lambda_y \mid \lambda_y \in \mathbb{R}\}.$$

Having in mind that before $x\cdot y = -c(x,y)$, this suggests the following general definition:

**Definition 2.5.1.** Given $c\colon X\times Y \to \mathbb{R}$ and $\varphi\colon X\to\mathbb{R}\cup\{+\infty\}$, we say $\varphi$ is **$c$-convex** if it can be written as

$$\varphi(x) = \sup_{y\in Y}\{-c(x,y) + \lambda_y \mid \lambda_y \in \mathbb{R}\}.$$

Then, for any $x\in X$, we define the **$c$-subdifferential** as

$$\partial_c\varphi(x) := \{y\in Y \mid \forall z\in X\colon \varphi(z) \ge -c(z,y) + c(x,y) + \varphi(x)\}.$$

Also, we define $\partial_c\varphi = \bigcup_{x\in X}\{x\}\times\partial_c\varphi(x) \subset X\times Y$.

The following is the analogue of Rockafellar's Theorem.

**Theorem 2.5.2.** *$S \subset X \times Y$ is c-cyclically monotone if and only if there exists a c-convex function $\varphi$ such that $S \subset \partial_c \varphi$.*

*Proof.* $\Leftarrow$ Let $(x_i, y_i)_{i=1,\dots,N} \subset S \subset \partial_c \varphi$, then

$$\varphi(z) \geq \varphi(x_i) - c(z, y_i) + c(x_i, y_i)$$

for every $z \in X$. Take $z = x_{i+1}$, sum over $i$, and we obtain

$$\sum_{i=1}^{N} -c(x_{i+1}, y_i) + c(x_i, y_i) \leq 0$$

(recall that $x_{N+1} = x_1$ by convention).

$\Rightarrow$ Define

$$\varphi(x) \coloneqq \sup\{-c(x, y_N) + c(x_N, y_N) - c(x_N, y_{N-1}) + \cdots + c(x_0, y_0) \mid (x_i, y_i)_{i=1,\dots,N} \subset S\}$$

and repeat the proof of Rockafellar's Theorem.

□

**Definition 2.5.3.** Given $\varphi \colon X \to \mathbb{R} \cup \{+\infty\}$ $c$-convex, we define the "$c$-Legendre transform" $\varphi^c \colon Y \to \mathbb{R} \cup \{+\infty\}$ as

$$\varphi^c(y) \coloneqq \sup_{x \in X} \{-c(x, y) - \varphi(x)\}.$$

**Proposition 2.5.4.** (a) $\varphi(x) + \varphi^c(y) + c(x, y) \geq 0$;

(b) $\varphi(x) + \varphi^c(y) + c(x, y) = 0$ *if and only if* $y \in \partial_c \varphi(x)$.

*Proof.* Identical to the proof in the previous section. □

**Theorem 2.5.5.** (Kantorovich duality: General case) *Let $c \in C^0(X \times Y)$ and assume that $\inf_{\gamma \in \Gamma(\mu,\nu)} \int c d\gamma < +\infty$. Then*

$$\min_{\gamma \in \Gamma(\mu,\nu)} \int c d\gamma = \max_{\varphi(x)+\psi(y)+c(x,y) \geq 0} \int -\varphi d\mu + \int -\psi d\nu.$$

*Proof.* The proof is left to the reader. The steps are essentially the same as in Theorem 2.4.5, just replace convexity with $c$-convexity and subdifferential with $c$-subdifferential, etc... □

### 2.5.1 Alternative proof of Kantorovich duality

A popular alternative approach to Kantorovich duality is based on some general abstract results in convex analysis, and goes as follows:

$$\begin{aligned}
\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{X\times Y} d\gamma(x,y) &\overset{\heartsuit}{=} \inf_{\gamma\geq 0}\sup_{\varphi,\psi}\Bigg\{\int c(x,y)d\gamma(x,y) + \overbrace{\left[\int\varphi(x)d\gamma(x,y)-\int\varphi(x)d\mu(x)\right]}^{\text{Lagrange multiplier for } (\pi_X)_{\#}\gamma=\mu} \\
&\qquad + \overbrace{\left[\int\psi(y)d\gamma(x,y)-\int\psi(y)d\nu(y)\right]}^{\text{Lagrange multiplier for } (\pi_Y)_{\#}\gamma=\nu}\Bigg\} \\
&\overset{\clubsuit}{=} \sup_{\varphi,\psi}\inf_{\gamma\geq 0}\left\{\int -\varphi d\mu + \int -\psi d\nu + \int \big[c(x,y)+\varphi(x)+\psi(y)\big]d\gamma\right\} \\
&= \sup_{\varphi,\psi}\left\{\left(\int -\varphi d\mu + \int -\psi d\nu\right) + \inf_{\gamma\geq 0}\int \big[c(x,y)+\varphi(x)+\psi(y)\big]d\gamma\right\} \\
&\overset{\diamondsuit}{=} \sup_{\varphi(x)+\psi(y)+c(x,y)\geq 0}\int -\varphi d\mu + \int -\psi d\nu.
\end{aligned}$$

where:

- $\heartsuit$ one should note that we do not require $\gamma$ to be a probability anymore, and we also drop the coupling constraint, only the sign constraint $\gamma \geq 0$ remains. The other constraints are "hidden" in the Lagrange multipliers. Indeed the supremum over $\varphi$ is $+\infty$ if $(\pi_X)_{\#}\gamma \neq \mu$ (resp. the supremum over $\psi$ is $+\infty$ if $(\pi_Y)_{\#}\gamma \neq \nu$). Note also that once $(\pi_X)_{\#}\gamma = \mu$ (or $(\pi_Y)_{\#}\gamma = \nu$) then $\int 1 d\gamma = \int 1 d\mu = 1$ which implies that $\gamma \in \mathcal{P}(X\times Y)$.
- $\clubsuit$ we used a Theorem of chapter 1.1. in [Vil03] to exchange inf and sup.
- $\diamondsuit$ we have the two following possible situations:
  - If $c(x,y)+\varphi(x)+\psi(y) \geq 0$ for any $(x,y)$, then $\inf_{\gamma\geq 0}\int[\dots]d\gamma = 0$ (take $\gamma \equiv 0$).
  - If there exists $(\bar{x},\bar{y})$ such that $c(\bar{x},\bar{y})+\varphi(\bar{x})+\psi(\bar{x}) < 0$, then take $\gamma = M\delta_{(\bar{x},\bar{y})}$ and let $M \to +\infty$. So, unless $c(x,y)+\varphi(x)+\psi(y) \geq 0$ for any $(x,y)$, the infimum over $\gamma$ is $-\infty$.

## 2.6 Brenier's Theorem

**Theorem 2.6.1.** (Brenier '87) *Let $X = Y = \mathbb{R}^n$ and $c(x,y) = \frac{|x-y|^2}{2}$ (or equivalently $-x\cdot y$). Suppose that*

$$\int |x|^2 d\mu + \int |y|^2 d\nu < +\infty$$

*and that $\mu \ll dx$ (i.e. $\mu$ is absolutely continuous with respect to the Lebesgue measure). Then there exists a unique optimal plan $\gamma$, where $\gamma = (\mathrm{id}\times T)_{\#}\mu$ and $T = \nabla\varphi$ for some convex function $\varphi$.*

*Proof.* The proof takes four steps, 1.-3. for existence, and 4. for uniqueness.

Step 1. There exists a non-trivial optimal transport plan $\gamma$: the cost $c(x,y) = \frac{|x-y|^2}{2}$ is non-negative and continuous. Taking $\mu\otimes\nu \in \Gamma(\mu,\nu)$ as plan, we obtain

$$\begin{aligned}
\int_{\mathbb{R}^n\times\mathbb{R}^n} |x-y|^2 d(\mu\otimes\nu) &\leq 2\int_{\mathbb{R}^n\times\mathbb{R}^n}\big(|x|^2+|y|^2\big)d(\mu\otimes\nu) \\
&= 2\int_{\mathbb{R}^n}|x|^2 d\mu + 2\int_{\mathbb{R}^n}|y|^2 d\nu < +\infty.
\end{aligned}$$

Thus Theorem 2.3.2 ensures the existence of a non-trivial optimal transport plan $\gamma$.

Step 2. Let $\gamma$ be optimal. Therefore $\operatorname{supp}(\gamma) \subset \partial\varphi$ for some convex $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ (see for instance Corollary 2.4.7). Also, by Proposition 2.4.4,

$$\varphi(x) + \varphi^*(y) = x \cdot y \qquad \text{on } \partial\varphi,$$

where $\varphi^*(y) \coloneqq \sup_{z\in\mathbb{R}^n}\{z \cdot y - \varphi(z)\}$. Therefore this condition holds almost everywhere on $\operatorname{supp}(\gamma)$. Hence $(\varphi(x), \varphi^*(y))$ is finite $\gamma$-a.e. and thus $\varphi$ is finite $\mu$-a.e. Since $\mu \ll dx$, a result from convex analysis tells us that convex functions are differentiable a.e. on the region where they are finite, hence $\varphi$ is differentiable $\mu$-a.e.

Step 3. Now let $A \subset \mathbb{R}^n$ with $\mu(A) = 0$ be such that $\varphi$ is differentiable everywhere in $\mathbb{R}^n \setminus A$. Let $\bar{x} \in \mathbb{R}^n \setminus A$ and suppose that $(\bar{x}, \bar{y}) \in \operatorname{supp}(\gamma) \subset \partial\varphi$. Then

$$\begin{aligned} \varphi(\bar{x}) + \varphi^*(\bar{y}) &= \bar{x} \cdot \bar{y} \\ \forall z \in \mathbb{R}^n\colon \varphi(z) + \varphi^*(\bar{y}) &\geq z \cdot \bar{y}. \end{aligned}$$

Thus

$$\Phi_{\bar{x}}(z) \coloneqq \varphi(z) - \varphi(\bar{x}) - \langle \bar{y}, z - \bar{x}\rangle \geq 0,$$

with equality at $z = \bar{x}$.
Since $\varphi$ is differentiable at $\bar{x}$, so is $\Phi_{\bar{x}}$ and it has a minimum at $\bar{x}$. Hence,

$$0 = \nabla\Phi_{\bar{x}}(\bar{x}) = \nabla\varphi(\bar{x}) - \bar{y}.$$

Therefore $\bar{y} = \nabla\varphi(\bar{x})$ for all $\bar{x} \in \mathbb{R}^n \setminus A$ and $(\bar{x}, \bar{y}) \in \operatorname{supp}(\gamma)$. This implies that $\operatorname{supp}(\gamma) \cap [(\mathbb{R}^n \setminus A) \times \mathbb{R}^n] \subset \operatorname{graph}(\nabla\varphi)$.
On the other hand, $\gamma(A \times \mathbb{R}^n) = \mu(A) = 0$. Taking this into account, we obtain that

$$(x, y) = (x, \nabla\varphi(x)) \ \gamma\text{-a.e.}$$

Thus for any test function $F\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, we have

$$\begin{aligned} \int_{\mathbb{R}^n\times\mathbb{R}^n} F(x,y)d\gamma(x,y) &= \int_{\mathbb{R}^n\times\mathbb{R}^n} F(x, \nabla\varphi(x))d\gamma(x,y) \\ &= \int_{\mathbb{R}^n} F(x, \nabla\varphi(x))d\mu(x) \\ &= \int_{\mathbb{R}^n\times\mathbb{R}^n} F(x,y)d((\mathrm{id}\times\nabla\varphi)_{\#}\mu)(x,y), \end{aligned}$$

hence $\gamma = (\mathrm{id} \times \nabla\varphi)_{\#}\mu$.

Step 4. If $\gamma_1$ and $\gamma_2$ are optimal, then so is $\frac{\gamma_1+\gamma_2}{2}$. Indeed,

$$\int |x-y|^2 d\left(\frac{\gamma_1+\gamma_2}{2}\right) = \frac{1}{2}\int |x-y|^2 d\gamma_1 + \frac{1}{2}\int |x-y|^2 d\gamma_2$$

and for any $\psi \in C_b(\mathbb{R}^n)$ it holds

$$\int \psi(x) d\left(\frac{\gamma_1+\gamma_2}{2}\right) = \frac{1}{2}\int \psi(x)d\gamma_1 + \frac{1}{2}\int \psi(x)d\gamma_2 = \int \psi d\mu,$$

thus $(\pi_X)_{\#}(\frac{\gamma_1+\gamma_2}{2}) = \mu$. Analogously $(\pi_Y)_{\#}(\frac{\gamma_1+\gamma_2}{2}) = \nu$.
By step 2. and 3. applied to $\gamma_1, \gamma_2, \frac{\gamma_1+\gamma_2}{2}$, we have

- $\gamma_1 = (\mathrm{id} \times \nabla\varphi_1)_{\#}\mu$, hence $(x, y) = (x, \nabla\varphi_1(x))$ $\gamma_1$-a.e. ;

- $\gamma_2 = (\mathrm{id} \times \nabla\varphi_2)_{\#}\mu$, hence $(x, y) = (x, \nabla\varphi_2(x))$ $\gamma_2$-a.e. ;
- $\frac{\gamma_1+\gamma_2}{2} = (\mathrm{id} \times \nabla\bar{\varphi})_{\#}\mu$, hence $(x, y) = (x, \nabla\bar{\varphi}(x))$ $(\gamma_1 + \gamma_2)$-a.e.,

hence $(x, \nabla\varphi_1(x)) = (x, \nabla\bar{\varphi}(x))$ $\gamma_1$-a.e., and therefore $\mu$-a.e. (since there is no dependence on $y$). Analogously $(x, \nabla\varphi_2(x)) = (x, \nabla\bar{\varphi}(x))$ $\mu$-a.e. So $\nabla\varphi_1 = \nabla\varphi_2$ $\mu$-a.e., thus $\gamma_1 = \gamma_2$.

□

**Corollary 2.6.2.** *Under the assumptions of Brenier's Theorem:*
*(1) There exists a unique optimal transport map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T_{\#}\mu = \nu$ and $T = \nabla\varphi$, with $\varphi$ convex.*
*(2) If $S_{\#}\mu = \nu$ and $S = \nabla\phi$ $\mu$-a.e. for some $\phi$ convex, then $S$ is the unique optimal transport map.*

*Proof.* (1) First of all, recall that the infimum in Monge is bounded below by the infimum in Kantorovich:

$$\inf_{T_{\#}\mu=\nu} \int |x - T(x)|^2 d\mu \geq \min_{\gamma\in\Gamma(\mu,\nu)} \int |x - y|^2 d\gamma(x, y).$$

Also, choosing $T = \nabla\varphi$ and $\gamma = (\mathrm{id} \times \nabla\varphi)_{\#}\mu$ from the previous proof gives equality:

$$\int |x - \nabla\varphi(x)|^2 d\mu = \min_{\gamma\in\Gamma(\mu,\nu)} \int |x - y|^2 d\gamma(x, y),$$

so $T = \nabla\varphi$ is optimal.

We now show that the solution to the Monge problem is unique. Let $T_1$ and $T_2$ be optimal for Monge. Since we have equality in the equation above, we deduce that $\gamma_1 = (\mathrm{id} \times T_1)_{\#}\mu$ and $\gamma_2 = (\mathrm{id} \times T_2)_{\#}\mu$ are optimal for Kantorovich. Because $\gamma_1 = \gamma_2$, we conclude that $T_1 = T_2$ $\mu$-a.e.

(2) Let $S_{\#}\mu = \nu$ and $S = \nabla\phi$ $\mu$-a.e. for some $\phi$ convex. Then it follows from Remark 2.4.8 that $(\mathrm{id} \times S)_{\#}\mu$ is optimal for Kantorovich, hence $S$ is optimal for Monge. □

**Corollary 2.6.3.** *Take the assumptions of Brenier's Theorem and assume also that $\nu \ll dx$. Let $\nabla\varphi$ be the optimal transport map from $\mu$ to $\nu$, and let $\nabla\psi$ be the optimal transport map from $\nu$ to $\mu$. Then $\nabla\varphi$ is invertible $\mu$-a.e., and its inverse is unique $\nu$-a.e. and given by $\nabla\psi$.*

*Proof.* By Brenier's Theorem, we have two convex maps $\varphi$ and $\psi$ such that

- $\nabla\varphi$ is an optimal transport map from $\mu$ to $\nu$;
- $\nabla\psi$ is an optimal transport map from $\nu$ to $\mu$.

Hence,

$$\int |x - \nabla\varphi(x)|^2 d\mu = \int |x - y|^2 d((\mathrm{id} \times \nabla\varphi)_{\#}\mu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int |x - y|^2 d\gamma$$

and (since the cost is symmetric in $x$ and $y$)

$$\int |\nabla\psi(y) - y|^2 d\nu = \int |x - y|^2 d((\nabla\psi \times \mathrm{id})_{\#}\nu) = \inf_{\tilde{\gamma}\in\Gamma(\mu,\nu)} \int |x - y|^2 d\tilde{\gamma}.$$

So $(\mathrm{id}\times\nabla\varphi)_{\#}\mu$ and $(\nabla\psi\times\mathrm{id})_{\#}\nu$ are both optimal, so they are equal. Thus, for any test-function $F\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we have

$$\begin{aligned}\int F(x, \nabla\varphi(x)) d\mu(x) &= \int F(x, y) d((\mathrm{id} \times \nabla\varphi)_{\#}\mu)(x, y) \\ &= \int F(x, y) d((\nabla\psi \times \mathrm{id})_{\#}\nu)(x, y) = \int F(\nabla\psi(y), y) d\nu(y).\end{aligned}$$

Choose $F(x,y) = |x - \nabla\psi(y)|^2$, then

$$\int |x - \nabla\psi(\nabla\varphi(x))|^2 d\mu(x) = \int |\nabla\psi(y) - \nabla\psi(y)|^2 d\nu(y) = 0,$$

thus $\nabla\psi \circ \nabla\varphi = \mathrm{id}$ $\mu$-a.e. By choosing $F(x,y) = |\nabla\varphi(x) - y|^2$, we get $\nabla\varphi \circ \nabla\psi = \mathrm{id}$ $\nu$-a.e. □

### 2.6.1 Application: Euler equations

Let $M$ be a manifold (that is, for simplicity, embedded in $\mathbb{R}^N$). To find minimising geodesics, one can solve the following minimization problem:

$$\min_{\gamma(0)=x,\gamma(1)=y} \int_0^1 |\dot\gamma|^2 dt, \tag{2.5}$$

where $\gamma : [0,1] \to M$ is a curve. It is well known that constant speed minimising geodesic satisfy the condition that $\ddot\gamma \perp T_\gamma M$. Viceversa, if a curve satisfies $\ddot\gamma \perp T_\gamma M$, then it is locally a minimizing constant speed geodesic.[4]

Hence, a possible equivalent definition of geodesic (here a geodesic should be understood as a curve that it is minimizing on small time intervals) is given by:

**Definition 2.6.4.** A curve $\gamma$ is a **geodesic** if $\ddot\gamma \perp T_\gamma M$.

**Application to incompressible Euler equations** Let $\Omega \subset \mathbb{R}^d$ be a bounded smooth set, $\nu$ an outward oriented unit vector field on the boundary $\partial\Omega$. Then the Euler equations are given by the following system:

- $v \cdot \nu = 0$ on $\partial\Omega$; (No-flux condition)
- $\mathrm{div}(v) = 0$; (Incompressibility condition)
- $\partial_t v + (v \cdot \nabla)v + \nabla p = 0$. (Euler equations)

The notation $v \cdot \nabla$ denotes the differential operator $\sum_{j=1}^d v^j \partial_j$. Hence, in coordinates $v = (v^1, \cdots, v^d)$, one reads the last equation as

$$\partial_t v^i + \sum_{j=1}^d v^j \partial_j v^i + \partial_i p = 0 \qquad \forall\, i = 1, \dots, n.$$

If $v$ is smooth, then

$$\begin{aligned}
\frac{d}{dt}\int_\Omega |v(t)|^2 &= \frac{d}{dt}\int_\Omega \sum_i v^i(t)^2 = 2\int_\Omega \sum_i v^i \partial_t v^i \\
&= -2\int_\Omega \sum_{ij} v^i v^j \partial_j v^i - 2\int_\Omega \sum_i v^i \partial_i p \\
&= -\int_\Omega \sum_j v^j \partial_j \Big(\sum_i (v^i)^2\Big) - 2\int_\Omega \sum_i v^i \partial_i p \\
&= -\int_\Omega v \cdot \nabla |v|^2 - 2\int_\Omega v \cdot \nabla p \\
&= -\int_{\partial\Omega} v \cdot \nu\ |v|^2 + \int_\Omega \mathrm{div}(v)\ |v|^2 - 2\int_{\partial\Omega} v \cdot \nu\ p + 2\int_\Omega \mathrm{div}(v)\ p \\
&= 0,
\end{aligned}$$

[4] Note that, if $\ddot\gamma \perp T_\gamma M$ then $|\dot\gamma| = const.$ Indeed, differentiating $|\dot\gamma|^2$ in time one gets

$$\frac{d}{dt}|\dot\gamma(2)|^2 = 2\langle \ddot\gamma(t), \dot\gamma(t)\rangle = 0,$$

where the last equality follows from the fact that $\dot\gamma \in T_\gamma M$, hence $\dot\gamma$ and $\ddot\gamma$ are orthogonal.

where we used the no-flux and incompressibility conditions.

Also, if $v$ is smooth, we can define its flow $g\colon I \times \Omega \to \Omega$ as

$$\begin{cases} \partial_t g(t,x) = v(t, g(t,x)); \\ g(0,x) = x. \end{cases}$$

Note that $g(t,\cdot)$ is a map from $\Omega$ to $\Omega$, since (thanks to the no-flux condition) the curve $t \mapsto g(t,x)$ never exits $\Omega$. Then

$$\partial_t \nabla_x g = \nabla_x [v(t, g(t,x))] = \nabla_x v(t, g(t,x)) \nabla_x g(t,x),$$

(note that $\nabla_x v$ and $\nabla_x g$ are $n \times n$ matrices, and $\nabla_x v(t, g(t,x))\nabla_x g(t,x)$ denotes their product (which is still a $n \times n$ matrix). This implies that

$$\nabla_x g(t+\varepsilon, x) = \nabla_x g(t,x) + \varepsilon \nabla_x v(t, g(t,x)) \nabla_x g(t,x) + o(\varepsilon).$$

Then, since $\det(AB) = \det(A)\det(B)$ and $\det(\mathrm{Id} + \varepsilon A) = 1 + \varepsilon \mathrm{tr}(A) + o(\varepsilon)$,

$$\begin{aligned} \frac{d}{dt} \det(\nabla_x g(t,x)) &= \lim_{\varepsilon \to 0} \frac{\det(\nabla_x g(t+\varepsilon, x)) - \det(\nabla_x g(t,x))}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{\det\big(\nabla_x g(t,x) + \varepsilon \nabla_x v(t, g(t,x)) \nabla_x g(t,x) + o(\varepsilon)\big) - \det(\nabla_x g(t,x))}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{\det\big(\nabla_x g(t,x) + \varepsilon \nabla_x v(t, g(t,x)) \nabla_x g(t,x)\big) - \det(\nabla_x g(t,x)) + o(\varepsilon)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{\det\big(\mathrm{Id} + \varepsilon \nabla_x v(t, g(t,x))\big) \det(\nabla_x g(t,x)) - \det(\nabla_x g(t,x))}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{[1 + \varepsilon \mathrm{tr}(\nabla_x v)(t, g(t,x)) - 1] \det(\nabla_x g(t,x))}{\varepsilon} \\ &= \mathrm{tr}(\nabla_x v)(t, g(t,x)) \det(\nabla_x g(t,x)) \\ &= \mathrm{div}(v)(t, g(t,v)) \det(\nabla_x g(t,x)) = 0, \end{aligned}$$

where the last equality follows from the incompressibility condition. Hence, since $\nabla_x g(0,x) = \mathrm{Id}$, we deduce that $\det(\nabla_x g(t,x)) \equiv 1$.

Now, if we differentiate in time the equation for $g$, using the Euler equations we get

$$\begin{aligned} \partial_{tt} g^i(t,x) &= \partial_t (v^i(t, g(t,x))) = \partial_t v^i(t,g) + \nabla v^i(t,g) \partial_t g \\ &= \partial_t v^i(t,g) + v(t,g) \cdot \nabla v^i(t,g) = -\partial_i p(t,g). \end{aligned}$$

Thus the equations

$$\begin{cases} \partial_{tt} g = -\nabla p(t,g) \text{ (Euler equation)} \\ \det \nabla_x g \equiv 1 \text{ (Incompressibility)} \\ g(0,x) = x \text{ (Initial conditions)} \end{cases}$$

are equivalent to the Euler equations.

**Theorem 2.6.5.** (Arnold '60) *The Euler equations are equivalent to the geodesic equation on the manifold* $SDiff(\Omega) := \{h\colon \Omega \to \Omega \mid h \text{ smooth measure preserving diffeomorphism}\} \subset L^2(\Omega; \mathbb{R}^d)$.

*Proof.* First we need to identify the tangent space of *SDiff*:
let $t \mapsto h(t) \in SDiff(\Omega)$ be a smooth "curve" of maps in *Sdiff*, and set $w(t) := \partial_t h(t)$. By definition of tangent space, $w(t) \in T_{h(t)} SDiff(\Omega)$.

Since $h(t)$ is a diffeomorphism of $\Omega$ it maps $\partial\Omega$ onto itself, and therefore $w(t) = \partial_t h(t)$ must be tangent to the boundary. Then define $\tilde{w}(t) := w \circ h^{-1}(t)$, so that $\partial_t h(t) = \tilde{w}(t, h(t))$,

and note that also $\tilde{w}(t)$ is tangent to $\partial\Omega$. Since $\det \nabla_x h(t,x) = 1$ (since $h(t) \in \mathit{SDiff}$), by the computations performed before we have

$$\begin{aligned} 0 &= \frac{d}{dt} \det \nabla_x h(t,x) \\ &= \operatorname{div}(\tilde{w})(t, h(t,x)) \underbrace{\det \nabla_x h(t,x)}_{=1}. \end{aligned}$$

Hence, $\operatorname{div}(\tilde{w}) = 0$. Thus

$$T_{h(t)}\mathit{SDiff} \subset \{w \mid \operatorname{div}(w \circ h^{-1}(t)) = 0,\ w \cdot \nu|_{\partial\Omega} = 0\} = \{\tilde{w} \circ h(t) \mid \operatorname{div}(\tilde{w}) = 0,\ \tilde{w} \cdot \nu|_{\partial\Omega} = 0\}.$$

Viceversa, given $h \in \mathit{SDiff}(\Omega)$ and a vector field $\tilde{w} : \Omega \to \mathbb{R}^d$ with $\operatorname{div}(\tilde{w}) = 0$ and $\tilde{w} \cdot \nu|_{\partial\Omega} = 0$, one can solve

$$\begin{cases} \partial_t h(t,x) = \tilde{w}(h(t,x)); \\ h(0,x) = h(x), \end{cases}$$

then using the same computation as above, one finds $\frac{d}{dt} \det \nabla h = 0$ and deduced that $h(t) : \Omega \to \Omega$ is a curve in $\mathit{SDiff}(\Omega)$. In particular, $\partial_t h(0) = \tilde{w} \circ h$ is an element of the tangent space of $\mathit{SDiff}(\Omega)$ at $h$.

Hence, we proved that, for any $h \in \mathit{SDiff}(\Omega)$,

$$T_h\mathit{SDiff} = \{w \mid \operatorname{div}(w \circ h^{-1}) = 0,\ w \cdot \nu|_{\partial\Omega} = 0\} = \{\tilde{w} \circ h \mid \operatorname{div}(\tilde{w}) = 0,\ \tilde{w} \cdot \nu|_{\partial\Omega} = 0\}.$$

Observation:

(a) For any measure preserving map $h \in \mathit{SDiff}(\Omega)$, for any $f_1, f_2 \colon \Omega \to \mathbb{R}^d$, we have

$$\langle f_1 \circ h, f_2 \circ h \rangle_{L^2} = \int_\Omega \langle f_1 \circ h, f_2 \circ h \rangle dx = \int_\Omega \langle f_1, f_2 \rangle dx = \langle f_1, f_2 \rangle_{L^2},$$

where the second equality follows from the fact that $h \in \mathit{SDiff}(\Omega)$, and therefore $h_\# dx = dx$.

(b) *(Helmholtz decomposition)*

$$L^2(\Omega, \mathbb{R}^d) \coloneqq \{w \colon \Omega \to \mathbb{R}^d \mid \operatorname{div}(w) = 0 \text{ and } w \cdot \nu|_{\partial\Omega} = 0\} \oplus \{\nabla q \mid q \colon \Omega \to \mathbb{R}^d\}.$$

Note that this decomposition is orthogonal.

Combining (a) and (b) yields

$$L^2(\Omega, \mathbb{R}^d) \coloneqq \underbrace{\{w \circ h \colon \Omega \to \mathbb{R}^d \mid \operatorname{div}(w) = 0 \text{ and } w \cdot \nu|_{\partial\Omega} = 0\}}_{=(T_h \mathit{SDiff})} \oplus \underbrace{\{\nabla q \circ h \mid q \colon \Omega \to \mathbb{R}^d\}}_{=(T_h \mathit{SDiff})^\perp}.$$

Hence, thanks to this characterization of $(T_h\mathit{SDiff})^\perp$, given a curve $t \to g(t) \in \mathit{SDiff}$, the following are equivalent:

- $t \to g(t)$ is a geodesic;
- $\partial_{tt} g \perp T_g \mathit{SDiff}$;
- $\partial_{tt} g(t,x) = \nabla q(t, g(t,x))$, for some $q(t) \colon \Omega \to \mathbb{R}^d$.

□

Thanks to Arnold's Theorem, we know that the incompressible Euler equations correspond to the geodesic equations in the space $SDiff(\Omega)$. We now recall that minimizing geodesics on manifolds can be found by considering the minimization problem (2.5). Thus, to find minimizing geodesics in *SDiff* one could consider the minimization problem

$$\int_0^1 \int_\Omega |\partial_t g(t,x)|^2 dx dt,$$

where $g(t) \in SDiff$ for any $t$, and $g(0) = g_0 \in SDiff$ and $g(1) = g_1 \in SDiff$ are prescribed.

This minimization problem is highly challenging and actually minimizers may fail to exist. So, a "weaker" version of this problem is to find, instead of the geodesic in *SDiff* from $g_0$ to $g_1$, the mid-point between them:

$$\begin{aligned} \mathrm{proj}_{SDiff} \colon L^2(\Omega, \mathbb{R}^d) &\to SDiff \\ \frac{g_0 + g_1}{2} &\mapsto \mathrm{proj}_{SDiff}\left(\frac{g_0 + g_1}{2}\right). \end{aligned}$$

Even this simpler problem is far from easy. One of the many issues is that *SDiff* is not well-behaved: it is neither convex, nor closed. So, as a first "relaxation" of the problem, one might want to consider its $L^2$-closure. This is characterized in the next (non-trivial) result due to Brenier and Gangbo.

**Theorem 2.6.6.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded set with Lipschitz boundary, and let $d \geq 2$. Then*

$$\overline{SDiff(\Omega)}^{L^2} = S(\Omega) := \{s \colon \Omega \to \Omega \mid s_\# dx = dx\}.$$

The next result gives a sufficient condition for the existence and uniqueness of a projection on $S(\Omega)$.

**Theorem 2.6.7.** (Brenier) *Let $h \in L^2(\Omega; \mathbb{R}^d)$ be such that $h_\# dx|_\Omega \ll dx$. Then there exists a unique projection $\bar{s}$ onto $S(\Omega)$ (i.e. for any $s \in S(\Omega)$ it holds $|h - \bar{s}|_{L^2(\Omega)} \leq |h - s|_{L^2(\Omega)}$).*

*Proof.* Existence: Take $h \colon \Omega \to \mathbb{R}^d$ and define $\mu := h_\#(dx|_\Omega) \ll dx$. Note that $\int_{\mathbb{R}^d} d\mu = \int_{h(\Omega)} d\mu = \int_\Omega dx = |\Omega|$. So, although Brenier's Theorem holds for probability measures, up to multiplying both $\mu$ and $dx|_\Omega$ by $\frac{1}{|\Omega|}$, we can apply it also in this context. Thus, by Brenier's Theorem applied both from $\mu$ to $dx|_\Omega$ and viceversa, there exist $\varphi \colon \Omega \to \mathbb{R}$ and $\psi \colon \mathbb{R}^d \to \mathbb{R}$ such that $\nabla\varphi$ and $\nabla\psi$ are optimal from $\mu$ to $dx|_\Omega$ and vice versa. Let $\bar{s} := \nabla\varphi \circ h \colon \Omega \to \Omega$, then

$$\begin{aligned} \int_\Omega |h(x) - \bar{s}(x)|^2 dx &= \int_\Omega |\nabla\varphi \circ h - h|^2 dx \\ &= \int_{\mathbb{R}^d} |\nabla\varphi - \mathrm{id}|^2 d\mu \\ &= \min_{\gamma \in \Gamma(\mu, dx|_\Omega)} \int |x - y|^2 d\gamma \\ &\leq \min_{s \in S(\Omega)} \int |x - y|^2 d\gamma_s = \min_{s \in S(\Omega)} \int_\Omega |h(x) - s(x)|^2 dx, \end{aligned}$$

where we set $\gamma_s = (h \times s)_\#(dx|_\Omega)$, and we used that $(\pi_1)_\#\gamma_s = h_\#(dx|_\Omega) = \mu$ and $(\pi_2)_\#\gamma_s = s_\#(dx|_\Omega) = dx|_\Omega$ (hence $\gamma_s \in \Gamma(\mu, dx|_\Omega)$). Thus $\bar{s}$ is a projection.
Uniqueness: Suppose that $s_1$ is another projection. Then by the previous computation, it follows that $\gamma_{\bar{s}}$ and $\gamma_{s_1}$ are optimal in Kantorovich. Thus, by uniqueness, the transport plans are equal. Hence, for any test-function $F$, we have

$$\int_\Omega F(h(x), s_1(x)) dx = \int_\Omega F(h(x), \bar{s}(x)) dx.$$

Take $F(x,y) = |\nabla\varphi(x) - y|^2$. Then, since $\bar{s} = \nabla\varphi \circ h$,

$$0 = \int_\Omega |\nabla\varphi \circ h - \bar{s}|^2 dx = \int_\Omega |\bar{s} - s_1|^2 dx.$$

Hence $s_1 = \bar{s}$ a.e. □

**Corollary 2.6.8.** (Polar Decomposition) *Let $h \in L^2(\Omega;\mathbb{R}^d)$ be such that $h_\# dx|_\Omega \ll dx$. Then $h = \nabla\psi \circ s$, with $s$ measure preserving and $\psi$ convex.*

*Remark* 2.6.9. Polar decomposition can be seen (at least formally) as a generalization of some well-known results:

(a) Any matrix $M \in \mathbb{R}^{n\times n}$ can be decomposed into $S\cdot O$, with $S$ symmetric and $O$ orthogonal. To see this, take $h(x) = Mx$. Then $h = \nabla\varphi \circ S$, with $\varphi(x) = \frac{1}{2}\langle x, Sx\rangle$ and $S(x) = Ox$.

(b) Consider a smooth vector field $w : \mathbb{R}^d \to \mathbb{R}^d$. Let $h_t = h(t,\cdot)$ be the flow of $w$:

$$\begin{cases} \partial_t h(t,x) = w(h(t,x)); \\ h(0,x) = x. \end{cases}$$

"Look" at the "linearisation" of $h_t := h(t)$:

$$\begin{aligned} h_\varepsilon(x) &= h_0(x) + \partial_t h_t(x)|_{t=0} \cdot \varepsilon + o(\varepsilon) \\ &= x + \varepsilon\, w(x) + o(\varepsilon). \end{aligned}$$

Then by using the polar decomposition of Brenier, we obtain

$$h_\varepsilon = \nabla\psi_\varepsilon \circ s_\varepsilon,$$

where $\psi_\varepsilon(x) = \frac{|x|^2}{2} + \varepsilon\, q(x) + o(\varepsilon)$ and $s_\varepsilon(x) = x + \varepsilon\, u(x) + o(\varepsilon)$. Since, for any $\varepsilon$, $1 \equiv \det\nabla s_\varepsilon = 1 + \varepsilon \mathrm{div}(u) + o(\varepsilon)$, we have that $\mathrm{div}(u) = 0$. Hence, by combining everything, we get

$$h_\varepsilon = \nabla\psi_\varepsilon \circ s_\varepsilon = (x + \varepsilon\nabla q(x)) \circ (x + \varepsilon u(x)) + o(\varepsilon) = x + \varepsilon(u(x) + \nabla q(x)) + o(\varepsilon).$$

By comparing the terms in order of $\varepsilon$, we deduce that $w = u + \nabla q$. Thus, morally speaking, the Helmholtz decomposition is the infinitesimal version of the polar decomposition.

## 2.7 General cost functions: existence/uniqueness of optimal transport maps

As we have seen in Section 2.5, if $c \in C^0(X \times Y)$ and $c \geq 0$, then the following are equivalent:

- $\bar{\gamma}$ is optimal;
- $\mathrm{supp}(\bar{\gamma})$ is $c$-cyclically monotone;
- There exists a pair $(\varphi, \varphi^c)$ such that

$$\varphi(x) = \sup_{y\in Y}\{-c(x,y) - \lambda_y\}$$

is $c$-convex,

$$\varphi^c(y) = \sup_{x\in X}\{-c(x,y) - \phi(x)\}$$

so that

$$\varphi(x) + \varphi^c(y) + c(x,y) \geq 0, \qquad (\star)$$

and equality holds for every $(x,y) \in \mathrm{supp}(\bar{\gamma})$.

**Theorem 2.7.1.** *Let $X = Y = \mathbb{R}^n$, $\mu \ll dx$ and $\mathrm{supp}(\nu)$ compact. Let $c \in C^0(X \times Y)$, $c \geq 0$ and $\inf_{\gamma \in \Gamma(\mu,\nu)} \int c d\gamma < +\infty$. Assume that:*
*- for every $y \in \mathrm{supp}(\nu)$, the map $\mathbb{R}^n \ni x \mapsto c(x,y)$ is differentiable;*
*- for every $x \in \mathbb{R}^n$, the map $\mathrm{supp}(\nu) \ni y \mapsto \nabla_x c(x,y) \in \mathbb{R}^n$ is injective.*
*Furthermore, assume that $|\nabla_x c(x,y)| \leq C$ for every $y \in \mathrm{supp}(\nu)$ and every $x \in \mathbb{R}^n$. Then there is a unique optimal $\bar{\gamma}$, with $\bar{\gamma} = (\mathrm{id} \times T)_{\#}\mu$ and $T$ satisfying*

$$\nabla_x c(x,y)|_{y=T(x)} + \nabla\varphi(x) = \nabla_x c(x, T(x)) + \nabla\varphi(x) = 0,$$

*for some $\varphi$ c-convex.*

*Remark* 2.7.2. For $c(x,y) = -x \cdot y$, we have $\nabla_x c(x,y) = -y$, thus the map $y \mapsto \nabla_x c(x,y) = -y$ is injective. Also

$$\nabla\varphi(x) + \nabla_x c(x, T(x)) = 0$$

implies that

$$\nabla\varphi(x) - T(x) = 0,$$

thus $T = \nabla\varphi$.

*Proof.* Let $\bar{\gamma}$ be optimal and $(\varphi, \varphi^c)$ be as in $(\star)$. We define

$$\tilde{\varphi}(x) = \sup_{y \in \mathrm{supp}(\bar{\gamma})} \{-c(x,y) - \varphi^c(y)\}.$$

Then

$$\tilde{\varphi}(x) + \varphi^c(y) + c(x,y) \geq 0$$

for all $(x,y) \in \mathbb{R}^n \times \mathrm{supp}(\nu)$. Therefore $\tilde{\varphi}(x) \leq \varphi(x)$. Besides, on $\mathrm{supp}(\bar{\gamma}) \subset \mathbb{R}^n \times \mathrm{supp}(\nu)$, we have

$$0 \leq \tilde{\varphi}(x) + \varphi^c(x) + c(x,y) \leq \varphi(x) + \varphi^c(x) + c(x,y) = 0.$$

Thus $\tilde{\varphi}(x) + \varphi^c(x) + c(x,y) = 0$ on $\mathrm{supp}(\bar{\gamma})$.
Now, for each $y \in \mathrm{supp}(\nu)$, look at the map

$$x \mapsto -c(x,y) - \varphi^c(y).$$

The gradient is given by $-\nabla_x c(x,y)$, which is uniformly bounded by $C$. Thus, the maps

$$x \mapsto -c(x,y) - \varphi^c(y)$$

are $C$-Lipschitz, and therefore so also the map $\tilde{\varphi}$ (being their supremum) is $C$-Lipschitz.

Since Lipschitz maps are differentiable a.e., this means that there exists a set $A$, with $|A| = 0$, such that $\tilde{\varphi}$ is differentiable on $\mathbb{R} \setminus A$. Since $\mu \ll dx$, we have $\mu(A) = 0$.
Fix $(x,y) \in \mathrm{supp}(\bar{\gamma})$ with $x \notin A$. Then

$$z \mapsto \tilde{\varphi}(z) + \varphi^c(y) + c(z,y) \geq 0$$

with equality at $z = x$. Thus

$$\nabla\tilde{\varphi}(x) + \nabla_x c(x,y) = 0.$$

Since $\nabla_x c(x,y)$ is injective, the equation above has at most one solution. Thus $y$ is uniquely determined in terms of $x$, and we call it $T(x)$. So $\mathrm{supp}(\bar{\gamma}) \cap [(\mathbb{R}^n \setminus A) \times \mathrm{supp}(\nu)] \subset \mathrm{graph}(T)$. As in the proof of Brenier's Theorem, since $\bar{\gamma}(A \times \mathrm{supp}(\nu)) \subset \bar{\gamma}(A \times \mathbb{R}^n) = \mu(A) = 0$, we deduce that $\bar{\gamma} = (\mathrm{id} \times T)_{\#}\mu$.
Uniqueness is the same argument as in step 4. of the proof of Brenier's Theorem, i.e. if $\gamma_1$ and $\gamma_2$ are optimal then so is $\frac{\gamma_1+\gamma_2}{2}$. Then $\mathrm{supp}(\frac{\gamma_1+\gamma_2}{2}) \overset{a.e.}{=} \mathrm{graph}(T_1) \cup \mathrm{graph}(T_2)$ must be a graph, and this is only possible if $T_1 = T_2$ $\mu$-a.e. □

*Remark* 2.7.3. The same proof is valid if we assume that for every $R > 0$ there exists a constant $C_R$ such that $|\nabla_x c(x,y)| \leq C_R$ for every $x \in B_R$ and every $y \in \text{supp}(\nu)$. Indeed, this assumption tells us that $x \mapsto -c(x,y) - \varphi^c(y)$ is $C_R$-Lipschitz continuous on $B_R$, and thus $\tilde{\varphi}$ is $C_R$ Lipschitz on $B_R$. This implies that $\tilde{\varphi}$ is differentiable almost everywhere, which is what is needed for our proof.

**Example 2.7.4.** Let $c(x,y) = |x-y|^p$, where $p > 1$. We claim that the map $y \mapsto \nabla_x c(x,y)$ is injective. To prove this, fix $x, v \in \mathbb{R}^n$ and assume that $v = \nabla_x c(x,y)$. We need to prove that $y$ is unique.
We have $v = p|x-y|^{p-2}(x-y)$. Since $|x-y|^{p-2}$ is positive, we deduce that the vectors $v$ and $(x-y)$ are parallel and point in the same direction. Thus

$$\frac{x-y}{|x-y|} = \frac{v}{|v|}.$$

We also know that

$$|v| = p|x-y|^p \Leftrightarrow |x-y| = \left(\frac{|v|}{p}\right)^{\frac{1}{p-1}}.$$

Combining this, we obtain that

$$x - y = \frac{v}{|v|}|x-y| = \frac{v}{|v|}\left(\frac{|v|}{p}\right)^{\frac{1}{p-1}}$$

and therefore $y = x - \frac{v}{|v|}\left(\frac{|v|}{p}\right)^{\frac{1}{p-1}}$, which proves that $y$ is unique.

For $p = 1$, the reasoning fails since $v = \nabla_x c(x,y) = \frac{x-y}{|x-y|}$. Thus $v = \frac{x-y}{|x-y|}$ for any $y \in x - \mathbb{R}_+ v$.

In fact the previous theorem does not apply: if $\mu = dx|_{[0,1]}$ and $\nu = dx|_{[1,2]}$, then $T_1(x) = x+1$ and $T_2(x) = 2 - x$ are both optimal, so we definitely have no uniqueness. In addition, if we define $\gamma_i := (\text{id} \times T_i)_{\#}\mu$ $(i = 1, 2)$, then $\frac{\gamma_1 + \gamma_2}{2}$ is optimal and it is not induced by a graph.

## 2.8 The $p$-Wasserstein metric and geodesics

**Definition 2.8.1.** Let $X$ be a metric space. Given $p \geq 1$, we define

$$\mathcal{P}_p(X) := \left\{\sigma \in \mathcal{P}(X) \mid \int_X d(x,x_0)^p d\sigma(x) < +\infty \text{ for some } x_0 \in X\right\},$$

the **set of probability measures with finite $p$-moment**.

*Remark* 2.8.2. Given $x_1 \in X$, since

$$d(x,x_1)^p \leq [d(x,x_0) + d(x_0,x_1)]^p \leq 2^{p-1}[d(x,x_0)^p + d(x_0,x_1)^p],$$

if $\int_X d(x,x_0)^p d\sigma(x) < +\infty$, then also $\int_X d(x,x_1)^p d\sigma(x)$ is finite. This means that the definition above is independent of the basepoint $x_0$.

**Definition 2.8.3.** Given $\mu, \nu \in \mathcal{P}(X)$, we call

$$W_p(\mu,\nu) := \left[\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X\times X} d(x,y)^p d\gamma(x,y)\right]^{\frac{1}{p}}$$

the **$p$-Wasserstein distance**.

*Remark* 2.8.4. For all $\gamma \in \Gamma(\mu, \nu)$:

$$\int d(x,y)^p d\gamma \leq 2^{p-1} \int [d(x,x_0)^p + d(x_0,y)^p] d\gamma = 2^{p-1} \left[ \int_X d(x,x_0)^p d\mu + \int_X d(y,x_0)^p d\nu \right] < \infty$$

if $\mu, \nu \in \mathcal{P}_p(X)$.

**Theorem 2.8.5.** *$W_p$ is a distance.*

*Proof.*

- If $W_p(\mu,\nu) = 0$, then there exists $\bar{\gamma}$ such that

$$\int_{X\times X} d(x,y)^p d\bar{\gamma}(x,y) = 0.$$

  Thus $x = y$ $\bar{\gamma}$-a.e., but the set $\{(x,y) \in X \times X \mid x = y\}$ is the graph of the identity. Therefore $\bar{\gamma} = (\mathrm{id} \times \mathrm{id})_\# \mu$ and thus $\nu = (\pi_2)_\# \bar{\gamma} = \mu$.

- $W_p(\mu,\nu) = W_p(\nu,\mu)$: For all $\gamma \in \Gamma(\mu,\nu)$, define $\tilde{\gamma} = S_\# \gamma$, with $S(x,y) = (y,x)$. Then $\tilde{\gamma} \in \Gamma(\nu,\mu)$ and

$$\int_{X\times X} d(x,y)^p d\tilde{\gamma} = \int_{X\times X} d(x,y)^p d\gamma.$$

- Triangle inequality: Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(X)$. Let $\gamma_{12} \in \Gamma_0(\mu_1,\mu_2) \coloneqq \{\text{optimal plans in } \Gamma(\mu_1,\mu_2)\}$ and let $\gamma_{23} \in \Gamma_0(\mu_2,\mu_3) \coloneqq \{\text{optimal plans in } \Gamma(\mu_2,\mu_3)\}$, i.e. transport plans which realise the Wasserstein distance. Note that

$$\left| \int_{X\times X} d(x,y)^p d\gamma_{12}(x,y) \right|^{\frac{1}{p}} = \|d(\cdot,\cdot)\|_{L^p(\gamma_{1,2})}.$$

  By disintegration with respect to the middle space $X_2 = X$, we have

$$\gamma_{12}(dx_1 dx_2) = \gamma_{12,x_2}(dx_1) \otimes d\mu_2(x_2)$$

  and

$$\gamma_{23}(dx_2 dx_3) = \gamma_{23,x_2}(dx_3) \otimes d\mu_2(x_2).$$

  We define

$$\tilde{\gamma}(dx_1 dx_2 dx_3) = \gamma_{12,x_2}(dx_2) \otimes \gamma_{23,x_2}(dx_3) \otimes d\mu_2(x_2).$$

  This measure has the property that

$$\int \varphi(x_1,x_2) d\tilde{\gamma}(x_1,x_2,x_3) = \iint \varphi(x_1,x_2) \gamma_{12,x_2}(dx_1) \left[ \int \gamma_{23}(dx_3) \right] d\mu_2(x_2) = \int \varphi(x_1,x_2) \gamma_{12}(dx_1 dx_2).$$

  Similarly

$$\int \varphi(x_2,x_3) d\tilde{\gamma}(x_1,x_2,x_3) = \int \varphi(x_2,x_3) \gamma_{23}(dx_2 dx_3).$$

  What we have managed to do is to embed the $\gamma$'s into a common space. Note that we also have that

$$\int \psi(x_1) d\tilde{\gamma} = \int \psi(x_1) d\gamma_{12} = \int \psi(x_1) d\mu_1$$

and

$$\int \psi(x_3)d\tilde{\gamma} = \int \psi(x_3)d\mu_3.$$

Set $\bar{\gamma}_{13} = \int_X \tilde{\gamma}(x_1, dx_2, x_3)$, i.e. integrate $x_2$ out. Then

$$\int \varphi(x_1, x_3)d\bar{\gamma}_{13} = \int \varphi(x_1, x_3)d\tilde{\gamma}(x_1, x_2, x_3).$$

This whole construction above is sometimes called *gluing Lemma*.
By triangle inequality in the $L^p$ space obtained by considering $X \times X \times X$ with the probability measure $\tilde{\gamma}$ we have

$$\begin{aligned} W_p(\mu_1, \mu_3) &\leq \left[\int_{X\times X} d(x_1, x_3)^p d\bar{\gamma}_{13}(x_1, x_3)\right]^{\frac{1}{p}} = \|d(x_1, x_3)\|_{L^p(d\bar{\gamma}_{13})} \\ &= \|d(x_1, x_3)\|_{L^p(d\tilde{\gamma})} \leq \|d(x_1, x_2) + d(x_2, x_3)\|_{L^p(d\tilde{\gamma})} \\ &\leq \|d(x_1, x_2)\|_{L^p(d\tilde{\gamma})} + \|d(x_2, x_3)\|_{L^p(d\tilde{\gamma})} \\ &= \|d(x_1, x_2)\|_{L^p(d\gamma_{12})} + \|d(x_2, x_3)\|_{L^p(d\gamma_{23})} \\ &= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3), \end{aligned}$$

which concludes the proof.

□

### 2.8.1 Construction of geodesics

Let $X = \mathbb{R}^n$ and $\gamma \in \Gamma_0(\mu, \nu)$ (i.e., $\gamma$ is optimal) for $W_p$. Set $\pi_t(x, y) = (1 - t)x + ty$. Then

$$\begin{cases} (\pi_0)_{\#}\gamma = \mu \\ (\pi_1)_{\#}\gamma = \nu \end{cases}.$$

Define $\mu_t = (\pi_t)_{\#}\gamma$ and let $\gamma_{s,t} \coloneqq (\pi_s, \pi_t)_{\#}\gamma \in \Gamma(\mu_s, \mu_t)$. Then

$$\begin{aligned} W_p(\mu_s, \mu_t) &\leq \left(\int |z - z'|^p d\gamma_{s,t}(z, z')\right)^{\frac{1}{p}} \\ &= \left(\int |\pi_s(x, y) - \pi_t(x, y)|^p d\gamma(x, y)\right)^{\frac{1}{p}} \\ &= |t - s|\left(\int |x - y|^p d\gamma\right)^{\frac{1}{p}} \\ &= |t - s|\, W_p(\mu_0, \mu_1). \end{aligned}$$

Hence, for $0 \leq s \leq t \leq 1$,

$$W_p(\mu_0, \mu_s) + W_p(\mu_s, \mu_t) + W_p(\mu_t, \mu_1) \leq [s + (t - s) + 1 - t]W_p(\mu_0, \mu_1) = W_p(\mu_0, \mu_1).$$

Note that the converse inequality always holds, by triangle inequality. Hence, all inequalities are equalities and we know that

$$W_p(\mu_s, \mu_t) = |t - s|W_p(\mu_0, \mu_1) \qquad \forall\, 0 \leq s, t \leq 1. \tag{$\dagger$}$$

**Definition 2.8.6.** A **constant speed geodesic curve** $(\mu_t)_{t\in[0,1]}$ is a curve such that ($\dagger$) holds.

# 3 Gradient Flows

## 3.1 Informal introduction

Let $\phi\colon \mathcal{H} \to \mathbb{R}$, where $\mathcal{H}$ is a Hilbert space, be $C^1$ (think for instance $\mathcal{H} = \mathbb{R}^d$). Given $x_0 \in \mathcal{H}$, the gradient flow (GF) of $\phi$ starting at $x_0$ is given by

$$\begin{cases} \dot{x}(t) &= -\nabla\phi(x(t)) \\ x(0) &= x_0 \end{cases}.$$

Note that

$$\frac{d}{dt}\phi(x(t)) = \nabla\phi(x(t)) \cdot \dot{x}(t) = -|\nabla\phi|^2(x(t)) \leq 0.$$

Thus $\phi$ decreases along the curve $x(t)$ and we have equality if and only if $|\nabla\phi|(x(t)) = 0$, i.e. $x(t)$ is a stationary point. In particular, if $\phi$ has a unique stationary point which coincides with the global minimizer (this is for instance the case if $\phi$ is strictly convex), then one expect to have that $x(t)$ converges to the minimizer as $t \to +\infty$.

*Remark* 3.1.1. To define a gradient flow, one needs a scalar product. Indeed, given a function $f\colon \mathcal{H} \to \mathbb{R}$, one defines $df(x)$ as follows:

$$\forall v \in \mathcal{H}\colon \quad df(x)[v] = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon v) - f(x)}{\varepsilon}.$$

Thus $df(x)\colon \mathcal{H} \to \mathbb{R}$ is linear, so $df(x) \in \mathcal{H}^*$. On the other hand, if $t \mapsto x(t) \in \mathcal{H}$ is a curve, then

$$\dot{x}(t) = \lim_{\varepsilon \to 0} \frac{x(t+\varepsilon) - x(t)}{\varepsilon} \in \mathcal{H}.$$

So $\dot{x}(t)$ and $df(x(t))$ live in different spaces.

However, if $\langle \cdot, \cdot \rangle$ is a scalar product on $\mathcal{H} \times \mathcal{H}$, we can define the gradient of $f$ at $x$ as

$$\langle \nabla f(x), v \rangle \coloneqq df(x)[v],$$

for all $v \in \mathcal{H}$. "We identify the gradient and the differential via a scalar product". Thanks to this identification, we can now make sense of $\dot{x}(t) = -\nabla f(x(t))$.

How does one constructs a (GF)? By using an analogue of the *implicit Euler scheme*. We discretise time, i.e. fix $\tau > 0$ a time step and set $x_0^\tau = x_0$. Given $x_k^\tau$, we want to find $x_{k+1}^\tau$. Assuming $\phi$ is $C^1$, then we proceed in the following way: We want to solve

$$\frac{x_{k+1}^\tau - x_k^\tau}{\tau} = -\nabla\phi(x_{k+1}^\tau).$$

This is equivalent to solving

$$\frac{x_{k+1}^\tau - x_k^\tau}{\tau} + \nabla\phi(x_{k+1}^\tau) = 0,$$

which again is equivalent to solving

$$\nabla_x \left( \frac{|x - x_k^\tau|^2}{2\tau} + \phi(x) \right) |_{x = x_{k+1}^\tau} = 0.$$

Hence, to find $x_{k+1}^\tau$, a possible way is finding a minimizer of $\psi_k^\tau(x) := \frac{|x - x_k^\tau|^2}{2\tau} + \phi(x)$.

If $\phi$ is not $C^1$, we want to still given a meaning to the notion of gradient flow. So, assume that $\phi\colon \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is convex and lower semi-continuous, and recall the notion of subdifferential introduced in Definition 2.4.1. Then we define a generalised gradient flow in the following way:

**Definition 3.1.2.**

$$(\text{GF}) := \begin{cases} \dot{x}(t) \in -\partial\phi(x(t)) \\ x(0) = x_0 \end{cases}.$$

Proceeding by analogy, for $\phi$ convex and lower semi-continuous we can still repeat the construction of discrete solutions via the implicit Euler scheme, and the first equation reads as:

$$\frac{x_{k+1}^\tau - x_k^\tau}{\tau} \in -\partial\phi(x_{k+1}^\tau).$$

This equation is then equivalent to

$$0 \in \frac{x_{k+1}^\tau - x_k^\tau}{\tau} + \partial\phi(x_{k+1}^\tau) =: \partial\psi_k^\tau(x_{k+1}^\tau).$$

Note that saying that 0 is in the subdifferential of $\psi_k^\tau$ at $x_{k+1}^\tau$ is equivalent to saying that $x_{k+1}^\tau$ is a minimizer of $\psi_k^\tau$ (see Definition 2.4.1). Hence, given $x_k^\tau$, we find $x_{k+1}^\tau$ by minimizing

$$x \mapsto \frac{|x - x_k^\tau|^2}{2\tau} + \phi(x).$$

Once the sequence $\{x_k^\tau\}_{k\geq 0}$ is constructed, we define a discrete solution as $x^\tau(t) = x_k^\tau$ for $t \in [k\tau, (k+1)\tau]$. Then, we send $\tau \to 0$.

*Remark* 3.1.3. On uniqueness: for $\phi$ a $C^1$-map and solutions $x(t), y(t)$ of (GF) with initial conditions $x_0$ and $y_0$ respectively, we have

$$\begin{aligned} \frac{d}{dt}\frac{|x(t) - y(t)|^2}{2} &= \langle x(t) - y(t), \dot{x}(t) - \dot{y}(t)\rangle \\ &= -\langle x(t) - y(t), \nabla\phi(x(t)) - \nabla\phi(y(t))\rangle. \end{aligned}$$

This last line is non-positive if $\phi$ is convex and thus yields uniqueness and stability of the solutions (if the initial data are close, then the solutions remain close). The above argument still works with $\phi$ convex and not necessarily $C^1$. Indeed, if $x(t)$ and $y(t)$ are (GF), then

$$\dot{x}(t) = -p(t) \quad \text{and} \quad \dot{y}(t) = -q(t), \qquad p(t) \in \partial\phi(x(t)), \quad q(t) \in \partial\phi(y(t)),$$

and we have

$$\begin{aligned} \frac{d}{dt}\frac{|x(t) - y(t)|^2}{2} &= \langle x(t) - y(t), \dot{x}(t) - \dot{y}(t)\rangle \\ &= -\langle x(t) - y(t), p(t) - q(t)\rangle \leq 0. \end{aligned}$$

**Example 3.1.4.** Let $\mathcal{H} = L^2(\mathbb{R}^d)$ and

$$\phi(u) = \begin{cases} \frac{1}{2}\int_{\mathbb{R}^d} |\nabla u|^2, & \text{if } u \in W^{1,2}(\mathbb{R}^d) \\ \infty, & \text{otherwise} \end{cases}.$$

Claim:

$$\partial\phi(u) \neq \emptyset \quad \Leftrightarrow \quad \Delta u \in L^2(\mathbb{R}^d),$$

and in that case $\partial\phi(u) = \{-\Delta u\}$.

*Proof.* $\boxed{\Rightarrow}$ Let $p \in L^2(\mathbb{R}^d)$ with $p \in \partial\phi(u)$. Then, by definition, for any $v \in L^2(\mathbb{R}^d)$ we have

$$\phi(v) \geq \phi(u) + \langle p, v - u \rangle_{L^2}.$$

Take $v = u + \varepsilon w$ with $w \in W^{1,2}(\mathbb{R}^d)$ and $\varepsilon > 0$. Then the equation above takes the form

$$\int \frac{|\nabla(u + \varepsilon w)|^2}{2} - \int \frac{|\nabla u|^2}{2} \geq \varepsilon \int pw.$$

Rearranging the terms and dividing by $\varepsilon$ yields

$$\int \nabla u \cdot \nabla w + \frac{\varepsilon}{2} \int |\nabla w|^2 \geq \int pw.$$

Then by letting $\varepsilon \to 0$, we obtain

$$\forall w \in W^{1,2}(\mathbb{R}^d) \colon \int \nabla u \cdot \nabla w \geq \int pw.$$

By taking the inequality above with $w$ and $-w$, we obtain that

$$\int \underbrace{-\Delta u}_{\text{as a distribution}} w = \int \nabla u \cdot \nabla w = \int pw,$$

i.e. $-\Delta u = p \in L^2(\mathbb{R}^d)$.

$\boxed{\Leftarrow}$ Assume that $\Delta u \in L^2$. By definition of $\phi$, we have

$$\phi(u + w) - \phi(u) = \int \nabla u \cdot \nabla w + \frac{1}{2} \int |\nabla w|^2 \geq \int \nabla u \cdot \nabla w = \int -\Delta u \; w,$$

for any $w \in W^{1,2}(\mathbb{R}^d)$. If, on the other hand, $w \notin W^{1,2}(\mathbb{R}^d)$, then

$$\phi(u + w) = \infty \geq \phi(u) + \int -\Delta u \; w.$$

Thus $-\Delta u \in \partial\phi(u)$.

□

**Corollary 3.1.5.** (Heat equation as gradient flow) *Let $\mathcal{H} = L^2(\mathbb{R}^d)$ and*

$$\phi(u) = \begin{cases} \frac{1}{2} \int_{\mathbb{R}^d} |\nabla u|^2, & \text{if } u \in W^{1,2}(\mathbb{R}^d) \\ \infty, & \text{otherwise} \end{cases}.$$

*Then the GF of $\phi$ with respect to the $L^2$ scalar product is*

$$\partial_t u(t) \in -\partial\phi(u(t)) \Leftrightarrow \partial_t u(t, x) = \Delta u(t, x)$$

## 3.2 Heat equation and optimal transport

In the previous section we stated that the heat equation can be solved first by solving the discrete problems

$$u^\tau_{k+1} \text{ is the minimum of } \frac{\|u - u^\tau_k\|^2_{L^2}}{2\tau} + \phi(u)$$

and then letting $\tau \to 0$ (where $\phi$ is the Dirichlet energy).

In [JKO98], the authors found a new construction based on optimal transport: more precisely, the above scheme to construct discrete solutions to the heat equation can be replaced by the following one:

$\rho_{k+1}^\tau$ is the minimizer of $\frac{W_2^2(\rho,\rho_k^\tau)}{2\tau} + \int \rho\log(\rho)dx$, where $\rho$ is a probability density.

Note that, given $\rho$ and $\tilde{\rho}$ probability densities, we identify them with the probability measures $\rho dx$ and $\tilde{\rho}dx$, thus

$$W_2(\rho,\tilde{\rho}) \coloneqq W_2(\rho dx, \tilde{\rho}dx) = \inf_{\gamma\in\Gamma(\rho dx,\tilde{\rho}dx)} \left(\int |x-y|^2 d\gamma\right)^{\frac{1}{2}}.$$

*Remark* 3.2.1. We will work with probability densities, but up to normalisation one can always reduce to this case whenever one take as initial datum a nonnegative function $\rho_0 \in L^1$.

In the paper [JKO98], the authors consider solutions in the whole space $\mathbb{R}^d$. Here instead we consider $\Omega \subset \mathbb{R}^d$ a bounded convex domain. We take $\rho_0$ to be a probability density in $\Omega$ such that

$$\underbrace{\int_\Omega \rho_0\log(\rho_0)dx}_{\text{Entropy}} < \infty.$$

We fix $\tau > 0$, we set $\rho_0^\tau := \rho_0$, and for every $k \in \mathbb{N}$, given $\rho_k^\tau$, define $\rho_{k+1}^\tau$ as the minimizer of

$$\frac{W_2^2(\rho,\rho_k^\tau)}{2\tau} + \int_\Omega \rho\log(\rho)dx. \tag{‡}$$

Claim: $\rho_{k+1}^\tau$ exists.

*Proof.* Take $\{\rho_m\}_{m\in\mathbb{N}}$ a minimising sequence, that is

$$\frac{W_2^2(\rho_m,\rho_k^\tau)}{2\tau} + \int_\Omega \rho_m\log(\rho_m)dx \to \inf_\rho \frac{W_2^2(\rho,\rho_k^\tau)}{2\tau} + \int_\Omega \rho\log(\rho)dx,$$

with $\int_\Omega \rho_m dx = 1$. For all $M \geq 1$, the sequence $\{\rho_m \wedge M\}$ is bounded, thus $\rho_m \wedge M \underset{L^\infty}{\rightharpoonup}{}^* \rho_M$ and

$$\int_\Omega (\rho_m - \rho_m\wedge M) = \int_{\{\rho_m\geq M\}} (\rho_m - M) \leq \int_{\{\rho_m\geq M\}} \rho_m \frac{\log(\rho_m)}{\log(M)}.$$

Thus, since $s\log(s) + 1 \geq 0$ for all $s \geq 0$,

$$\begin{aligned}
\int_\Omega (\rho_m - \rho_m\wedge M) &\leq \frac{1}{\log(M)} \int_{\Omega\cap\{\rho_m\geq M\}} \rho_m\log(\rho_m) \\
&\leq \frac{1}{\log(M)} \int_{\Omega\cap\{\rho_m\geq M\}} (\rho_m\log(\rho_m) + 1) \\
&\leq \frac{1}{\log(M)} \int_\Omega (\rho_m\log(\rho_m) + 1) \\
&\leq \frac{C}{\log(M)},
\end{aligned}$$

where the last bound follows from the fact that $\rho_m$ is a minimising sequence, hence $\int \rho_m\log(\rho_m)$ is uniformly bounded.

Let $\rho_\infty \coloneqq \sup_M \rho_M$. We know that

$$\begin{aligned}
&\rho_m \wedge M \; dx \overset{*}{\rightharpoonup} \rho_M \; dx, \\
&\rho_M \overset{L^1}{\to} \rho_\infty, \\
&\|\rho_m \wedge M - \rho_m\|_{L^1} \leq \frac{C}{\log(M)}.
\end{aligned}$$

Hence, by a diagonal argument we can find a sequence of indices $\{m_M\}_{M\in\mathbb{N}}$, with $m_M \to \infty$, such that $\rho_{m_M} \wedge M \overset{*}{\rightharpoonup} \rho_\infty$. Hence, it holds $\rho_{m_M} \overset{*}{\rightharpoonup} \rho_\infty$. Similarly, we can prove that any subsequence of $\rho_m$ admits a subsequence that converges (in the weak topology of measures) to $\rho_\infty$, thus

$$\rho_m dx \rightharpoonup^* \rho_\infty dx.$$

We now want to show that $\rho_\infty$ is still a probability density (this is not obvious, since some mass may have "escaped" from $\Omega$). For this, we set $N_\varepsilon := \{x \in \Omega \mid \mathrm{dist}(x, \partial\Omega) < \varepsilon\}$. Since $|N_\varepsilon| \le C\varepsilon$, setting $M := \frac{1}{\varepsilon|\log(\varepsilon)|}$ we have

$$\begin{aligned}\int_{N_\varepsilon} \rho_m &\le \int_{N_\varepsilon \cap \{\rho_m \le M\}} \rho_m + \int_{N_\varepsilon \cap \{\rho_m \ge M\}} \rho_m \frac{\log(\rho_m)}{\log(M)} \\ &\le M|N_\varepsilon| + \frac{C}{\log(M)} \le C\left(\varepsilon M + \frac{1}{\log(M)}\right) \le \frac{C}{|\log(\varepsilon)|}.\end{aligned}$$

Then

$$\int_{\Omega\setminus N_\varepsilon} \rho_m \ge 1 - \frac{C}{|\log(\varepsilon)|}$$

and thus

$$\int_{\Omega\setminus N_\varepsilon} \rho_\infty \ge 1 - \frac{C}{|\log(\varepsilon)|}.$$

Letting $\varepsilon \to 0$, we conclude that $\rho_\infty$ is a probability measure. Note also that the above bounds imply that the densities $\rho_m$ are tight (considering the sequence of compact sets $K_\varepsilon := \Omega \setminus N_\varepsilon$). Hence the convergence of $\rho_m$ to $\rho_\infty$ is also narrow (see the end of the proof of Lemma 2.1.12).

We now observe that, since $[0,\infty) \ni s \mapsto s\log(s)$ is convex, by [AFP00, Theorem 5.2] we have

$$\int_\Omega \rho_\infty \log(\rho_\infty) \le \liminf_{m\to\infty} \int_\Omega \rho_m \log(\rho_m).$$

We now want to study the behaviour of $W_2^2(\rho_m, \rho_k^\tau)$ as $m \to \infty$.

Let $\gamma_m \in \Gamma(\rho_m, \rho_k^\tau)$. Then, since $\rho_m$ is tight (see the discussion above), the proof of Lemma 2.3.1 shows that also $\gamma_m$ is tight. Hence, up to taking a subsequence, $\gamma_m \rightharpoonup \gamma_\infty$ with

$$\begin{aligned}(\pi_2)_\# \gamma_m &= \rho_k^\tau; \\ (\pi_1)_\# \gamma_m &= \rho_m \rightharpoonup \rho_\infty.\end{aligned}$$

Note also that, since $|x-y|^2$ is continuous and bounded on $\Omega\times\Omega$,

$$W_2^2(\rho_m, \rho_k^\tau) = \int_{\Omega\times\Omega} |x-y|^2 d\gamma_m \to \int |x-y|^2 d\gamma_\infty \ge W_2^2(\rho_\infty, \rho_k^\tau).$$

Thus $\gamma_\infty \in \Gamma(\rho_\infty, \rho_k^\tau)$, and combining together the lower semicontinuity of $\int \rho_m \log(\rho_m)$ with the equation above, we have

$$\liminf_m \frac{W_2^2(\rho_m, \rho_k^\tau)}{2\tau} + \int_\Omega \rho_m \log(\rho_m) \ge \frac{W_2^2(\rho_\infty, \rho_k^\tau)}{2\tau} + \int_\Omega \rho_\infty \log(\rho_\infty).$$

Hence $\rho_\infty$ is a minimizer, and we set $\rho_{k+1}^\tau := \rho_\infty$.

□

We now want find the optimality conditions associated of the functional (‡). To do this, let $\xi \in C_c^\infty(\Omega, \mathbb{R}^d)$ be a vector field (tangent to the boundary of $\Omega$). Then, for $x \in \Omega$ we consider the flow

$$\begin{cases} \dot{\Phi}(t,x) = \xi(\Phi(t,x)) \\ \Phi(0,x) = x \end{cases}.$$

Note that $\Phi(t) : \Omega \to \Omega$ is a diffeomorphism. Define

$$\rho_\varepsilon \coloneqq \Phi(\varepsilon)_\# \rho_{k+1}^\tau.$$

Then, it follows by Section 1.5 that

$$\rho_{k+1}^\tau(x) = \rho_\varepsilon(\Phi(\varepsilon, x)) \det \nabla \Phi(\varepsilon, x).$$

It holds

$$\begin{aligned} \int_\Omega \rho_\varepsilon(y) \log(\rho_\varepsilon(y)) dy &= \int_\Omega \rho_{k+1}^\tau(x) \log(\rho_\varepsilon(\Phi(\varepsilon, x))) dx \\ &= \int \rho_{k+1}^\tau(x) \log\left( \frac{\rho_{k+1}^\tau(x)}{\det \nabla \Phi(\varepsilon, x)} \right) dx. \end{aligned}$$

Then, if we Taylor-expand in $\varepsilon$, since $\log(1 + z + o(z)) = z + o(z)$ we obtain (recall the computations performed in Section 2.6.1)

$$\begin{aligned} \int \rho_\varepsilon \log(\rho_\varepsilon) &= \int \rho_{k+1}^\tau \log(\rho_{k+1}^\tau) - \int \rho_{k+1}^\tau \log(\underbrace{\det \nabla \Phi(\varepsilon, x)}_{1 + \varepsilon \mathrm{div} \xi + o(\varepsilon)}) dx \\ &= \int \rho_{k+1}^\tau \log(\rho_{k+1}^\tau) - \varepsilon \int \rho_{k+1}^\tau \mathrm{div} \xi dx. \end{aligned}$$

Now, given $\gamma \in \Gamma(\rho_{k+1}^\tau, \rho_k^\tau)$ optimal, define $\gamma_\varepsilon \coloneqq (\Phi(\varepsilon, \cdot) \times \mathrm{id})_\# \gamma$. We have

$$(\pi_1)_\# \gamma_\varepsilon = \Phi(\varepsilon, \cdot)_\# \rho_{k+1}^\tau = \rho_\varepsilon$$

and

$$(\pi_2)_\# \gamma_\varepsilon = (\pi_2)_\# \gamma = \rho_k^\tau.$$

Therefore,

$$\begin{aligned} W_2^2(\rho_\varepsilon, \rho_k^\tau) &\leq \int |x - y|^2 d\gamma_\varepsilon \\ &= \int | \underbrace{\Phi(\varepsilon, x)}_{x + \varepsilon \xi(x) + o(\varepsilon)} - y|^2 d\gamma \\ &= \int \left[ |x - y|^2 + 2\varepsilon \langle \xi(x), x - y \rangle + o(\varepsilon) \right] d\gamma \end{aligned}$$

Then, since $\gamma$ is optimal from $\rho_{k+1}^\tau$ to $\rho_k^\tau$, we get

$$W_2^2(\rho_\varepsilon, \rho_k^\tau) \leq W_2^2(\rho_{k+1}^\tau, \rho_k^\tau) + 2\varepsilon \int \langle \xi(x), x - y \rangle d\gamma + o(\varepsilon).$$

Combining everything together, we obtain

$$\begin{aligned}\frac{W_2^2(\rho_{k+1}^\tau,\rho_k^\tau)}{2\tau}+\int_\Omega \rho_{k+1}^\tau\log(\rho_{k+1}^\tau)dx &\le \frac{W_2^2(\rho_\varepsilon,\rho_k^\tau)}{2\tau}+\int_\Omega \rho_\varepsilon\log(\rho_\varepsilon)\\ &\le \frac{W_2^2(\rho_{k+1}^\tau,\rho_k^\tau)}{2\tau}+\int_\Omega \rho_{k+1}^\tau\log(\rho_{k+1}^\tau)dx\\ &\quad+\underbrace{\frac{\varepsilon}{\tau}\int_\Omega\langle\xi(x),x-y\rangle d\gamma-\varepsilon\int_\Omega\rho_{k+1}^\tau\mathrm{div}\xi dx}_{(\star)}+o(\varepsilon).\end{aligned}$$

Hence we see that the term $(\star)$ has to vanish (indeed it must be nonnegative, but replacing $\xi$ with $-\xi$ implies that it must be null). Therefore our optimality conditions for $\rho_{k+1}^\tau$ read as

$$\int_\Omega\rho_{k+1}^\tau\mathrm{div}\xi dx-\frac{1}{\tau}\int_{\Omega\times\Omega}\langle\xi(x),x-y\rangle d\gamma=0,$$

where $\gamma$ realises the 2-Wasserstein distance between $\rho_{k+1}^\tau$ and $\rho_k^\tau$.

To simplify the notation, we can use that by Brenier's Theorem the optimal plan $\gamma$ is unique and it is induced by an optimal map $T_{k+1}$ from $\rho_k^\tau$ to $\rho_{k+1}^\tau$, namely $\gamma=(T_{k+1}\times\mathrm{id})_\#\rho_k^\tau$. Thus

$$\int_{\Omega\times\Omega}\langle\xi(x),x-y\rangle d\gamma=\int_\Omega\langle\xi\circ T_{k+1}(x),T_{k+1}(x)-x\rangle\rho_k^\tau(x)dx,$$

and the optimality equations become [5]

$$\int_\Omega\rho_{k+1}^\tau\mathrm{div}\xi dx-\frac{1}{\tau}\int_\Omega\langle\xi\circ T_{k+1},T_{k+1}-x\rangle\rho_k^\tau=0,$$

In order to obtain the heat equation, take $\xi=\nabla\psi$ for $\psi\in C_c^\infty(\Omega)$. Then, by Taylor expansion (with integral reminder), one gets

$$\psi(x)-\psi(y)=\langle\nabla\psi(y),x-y\rangle+\frac{1}{2}\int_0^1 D^2\psi(tx+(1-t)y)[x-y,x-y]dt,$$

and thus

$$|\psi(x)-\psi(y)-\langle\nabla\psi(y),x-y\rangle|\le\|D^2\psi\|_\infty|x-y|^2.$$

Therefore

$$\begin{aligned}\int\big|\langle\nabla\psi\circ T_k,T_k-x\rangle+\psi(x)-\psi(T_k)\big|\rho_{k-1}^\tau &\le\|D^2\psi\|_\infty\int|T_k-x|^2\rho_{k-1}^\tau\\ &=\|D^2\psi\|_\infty W_2^2(\rho_k^\tau,\rho_{k-1}^\tau).\end{aligned}$$

[5] Alternatively, one could have proceeded as follows: let $T_{k+1}$ be an optimal transport map from $\rho_k^\tau$ to $\rho_{k+1}^\tau$, and note that $\Phi_\varepsilon\circ T_{k+1}$ transports $\rho_k^\tau$ to $\rho_\varepsilon$. Hence,

$$\begin{aligned}W_2^2(\rho_\varepsilon,\rho_k^\tau)&\le\int|\Phi_\varepsilon\circ T_{k+1}-x|^2\rho_k^\tau \qquad(\star\star\star)\\ &=\int|T_{k+1}+\varepsilon\xi\circ T_{k+1}-x|^2\rho_k^\tau+o(\varepsilon)\\ &=\underbrace{\int|T_{k+1}-x|^2\rho_k^\tau}_{W_2^2(\rho_{k+1}^\tau,\rho_k^\tau)}+2\varepsilon\int\langle\xi\circ T_{k+1},T_{k+1}-x\rangle\rho_k^\tau+o(\varepsilon).\end{aligned}$$

Using this formula, one gets the desired formula for the optimality conditions.

Then by the Euler-Lagrange equation:

$$\left| -\int \Delta\psi \rho_k^\tau + \frac{1}{\tau} \underbrace{\int [\psi(T_k) - \psi(x)]\rho_{k-1}^\tau}_{=\int \psi\rho_k^\tau - \int \psi\rho_{k-1}^\tau} \right| \leq \|D^2\psi\|_\infty \frac{W_2^2(\rho_k^\tau, \rho_{k-1}^\tau)}{\tau} \qquad (\star_k)$$

We would like to take the limit for $\tau \to 0$. Note that "morally", if we think that $W_2(\rho_k^\tau, \rho_{k-1}^\tau)$ should be of order $\tau$,

$$\frac{W_2^2(\rho_k^\tau, \rho_{k-1}^\tau)}{\tau} \propto \frac{\tau^2}{\tau} \to 0 \ (\tau \to 0).$$

To rigorously control this term note that, by minimality,

$$\frac{W_2^2(\rho_k^\tau, \rho_{k-1}^\tau)}{2\tau} + \int \rho_k^\tau \log(\rho_k^\tau) \leq \left( \frac{W_2^2(\rho, \rho_{k-1}^\tau)}{2\tau} + \int \rho \log(\rho) \right)|_{\rho=\rho_{k-1}^\tau}.$$

Thus, by taking the telescopic sum over $k = 1, \ldots, \infty$, one gets

$$\sum_{k=1}^{\infty} \frac{W_2^2(\rho_k^\tau, \rho_{k-1}^\tau)}{2\tau} \leq \int \rho_0 \log(\rho_0).$$

Also, by summing over finitely many terms, one obtains the following result

$$\underbrace{\sum_{k=1}^{K_0} \frac{W_2^2(\rho_k^\tau, \rho_{k-1}^\tau)}{2\tau}}_{\geq 0} + \int \rho_{K_0} \log(\rho_{K_0}) \leq \int \rho_0 \log(\rho_0),$$

i.e. "$\rho \log(\rho)$" decreases over $k$.
Now set $\rho^\tau(t, x) \coloneqq \rho_k^\tau(x)$ for $t \in [(k-1)\tau, k\tau]$. Then $\{\rho^\tau(t, x)\}_{t \geq 0} \subset \mathcal{P}(\Omega)$,

$$\int_\Omega \rho^\tau(t, x) \log(\rho^\tau(t, x)) \leq \int \rho_0 \log(\rho_0)$$

and

$$\sum_{k=1}^{\infty} \frac{W_2^2(\rho^\tau((k-1)\tau), \rho^\tau(k\tau))}{2\tau} \leq \int \rho_0 \log(\rho_0). \qquad (3.1)$$

Thanks to the above bound on $\int \rho^\tau(t) \log(\rho^\tau(t))$ and the fact $\int \rho^\tau(t) = 1$, on each interval $[0, T]$ one has:

$$\int_{t_1}^{t_2} \int_\Omega \rho^\tau(t, x) dx dt = t_2 - t_1 \qquad \forall\, 0 \leq t_1 \leq t_2 \leq T,$$

and

$$\int_0^T \int_\Omega \rho^\tau(t, x) \log(\rho^\tau(t, x)) dx dt \leq T \int \rho_0 \log(\rho_0).$$

As shown in the proof of the existence of $\rho_{k+1}^\tau$, these bounds imply that the measures $\rho^\tau$ cannot concentrate, nor escape to the boundary of $\Omega$. Hence, up to a subsequence $\rho^\tau$ converges weakly in $L^1([0, T] \times \Omega)$ to a density $\rho(t, x)$ with $\int_\Omega \rho(t, x) dx = 1$ for all $t \in [0, T]$.

Take $\zeta \in C_c^\infty([0,+\infty))$. We want to test the heat equation for $\rho^\tau$ against $\psi(x)\zeta(t)$. Take $(\star_k)$ and multiply it by $\tau\,\zeta((k-1)\tau)$:

$$\left|\int \psi(x)\rho^\tau(k\tau)\zeta((k-1)\tau) - \int \psi(x)\rho^\tau((k-1)\tau)\zeta((k-1)\tau) - \tau\int \Delta\psi(x)\rho^\tau(k\tau)\zeta((k-1)\tau)\right| \\ \leq \|D^2\psi\|_\infty \|\zeta\|_\infty W_2^2(\rho^\tau((k-1)\tau),\rho^\tau(k\tau)).$$

Summing over $k = 1,\dots,\infty$ yields

$$\Bigg|-\int_\Omega \psi(x)\rho_0(x)dx\zeta(0) + \overbrace{\sum_{k=2}^\infty \int \psi(x)\rho^\tau(kt)\zeta((k-1)\tau)) - \sum_{k=1}^\infty \int \psi(x)\rho^\tau(k\tau)\zeta(k\tau)}^{(1)} \\ \underbrace{- \sum_{k=1}^\infty \tau\int_\Omega \Delta\psi\rho^\tau(k\tau)\zeta((k-1)\tau)\Bigg|}_{(2)} \\ \leq C(\psi,\zeta)\sum_{k=1}^\infty W_2^2(\rho^\tau((k-1)\tau),\rho^\tau(k\tau)) \\ \leq C\,\tau,$$

where the last bound follows from (3.1).

We can rewrite the terms (1) and (2):

$$(1) = \sum_{k=1}^\infty \int \psi(x) \overbrace{\rho^\tau(k\tau)}^{\rho^\tau(t)\ ;\ t\in[(k-1)\tau,k\tau]} \overbrace{[\zeta((k-1)\tau) - \zeta(k\tau)]}^{-\int_{(k-1)\tau}^{k\tau}\partial_t\zeta(t)dt} \\ = -\sum_{k=1}^\infty \int_{(k-1)\tau}^{k\tau}\int_\Omega \psi(x)\rho^\tau(t,x)\partial_t\zeta(t)dxdt = -\int_0^\infty\int_\Omega \psi(x)\rho^\tau(t,x)\partial_t\zeta dxdt,$$

and since

$$\tau\zeta((k-1)\tau) = \int_{(k-1)\tau}^{k\tau}\zeta((k-1)\tau)dt = \int_{(k-1)\tau}^{k\tau}\zeta(t)dt + \underbrace{\int_{(k-1)\tau}^{k\tau}(\zeta((k-1)\tau)-\zeta(t))dt}_{\leq\|\partial_t\zeta\|_\infty\tau},$$

we have

$$(2) = \sum_{k=1}^\infty \int_\Omega \Delta\psi\rho^\tau(t)\tau\zeta((k-1)\tau) \\ = \sum_{k=1}^\infty \int_{(k-1)\tau}^{k\tau}\int_\Omega \Delta\psi(x)\rho^\tau(t,x)\zeta(t)dtdx + O\Big(\tau\sum_{k=1}^\infty\int_{(k-1)\tau}^{k\tau}\Delta\psi(x)\rho^\tau(t,x)\|\partial_t\zeta\|_\infty dtdx\Big).$$

Therefore

$$\left|-\int_\Omega \psi(x)\rho_0(x)dx\zeta(0) - \int_0^\infty\int_\Omega \psi(x)\rho^\tau(t,x)\partial_t\zeta(t)dtdx - \int_0^\infty\int_\Omega \Delta\psi(x)\rho^\tau(t,x)\zeta(t)dtdx\right| \\ \leq C\tau + \tau\int_0^T\int_\Omega |\Delta\psi(x)|\rho^\tau(t,x)\|\partial_t\zeta\|_\infty \to 0\ (\tau\to 0).$$

Hence

$$-\int_\Omega \psi(x)\rho_0(x)dx\zeta(0) - \int_0^\infty\int_\Omega \psi(x)\rho(t,x)\partial_t\zeta(t)dtdx - \int_0^\infty\int_\Omega \Delta\psi(x)\rho(t,x)\zeta(t)dtdx = 0,$$

which shows that $\rho$ solves the heat equation in its weak form.

To conclude the proof one should find the boundary conditions for $\rho$. Since the mass of $\rho$ cannot enter nor leave $\Omega$ (by the way the solutions was constructed), the natural boundary conditions are the Neumann boundary conditions $\partial_\nu \rho(t)|_{\partial\Omega} = 0$. However, we shall not discuss this here.

# 4 Differential viewpoint of optimal transport

## 4.1 Riemannian structure of $(\mathcal{P}(\Omega), W_2)$

In this section, most computations will be formal and not rigorous, we just want to get an idea.

Let $\bar{\rho}_0$ and $\bar{\rho}_1$ be two probability distributions and $v(t,x) \in \mathbb{R}^d$ a vector field. Let $X(t,x)$ solve

$$\begin{cases} \dot{X}(t,x) = v(t, X(t,x)); \\ X(0,x) = x. \end{cases}$$

Set $\rho_t = (X(t))_{\#}\bar{\rho}_0$.
Claim: $\partial_t \rho_t + \operatorname{div}(v_t \rho_t) = 0$ (Continuity equation), where $v_t := v(t,\cdot)$.

*Proof.* (In the sense of distribution) Consider $\int_\Omega \rho_t \psi(x)dx$, then

$$\begin{aligned} \int_\Omega \partial_t \rho_t \psi dx &= \frac{d}{dt}\int_\Omega \rho_t \psi(x)dx = \frac{d}{dt}\int_\Omega \psi(X(t,x))\bar{\rho}_0(x)dx \\ &= \int_\Omega \nabla\psi(X(t,x)) \cdot \dot{X}(t,x)\bar{\rho}_0(x)dx = \int_\Omega \nabla\psi(X(t,x)) \cdot v_t(X(t,x))\bar{\rho}_0(x)dx \\ &= \int_\Omega \nabla\psi(x) \cdot v_t(x)\rho_t(x)dx = -\int_\Omega \psi(x)\operatorname{div}(v_t\rho_t)dx. \end{aligned}$$

□

**Definition 4.1.1.** $A[\rho_t, v_t] = \int_0^1 \int_\Omega |v_t(x)|^2 \rho_t(x)dxdt$.

**Theorem 4.1.2.** (Benamou-Brenier)

$$W_2^2(\bar{\rho}_0, \bar{\rho}_1) = \inf\{A[\rho_t, v_t] \mid \rho_0 = \bar{\rho}_0,\ \rho_1 = \bar{\rho}_1,\ \partial_t\rho_t + \operatorname{div}(v_t\rho_t) = 0\}.$$

*Proof. (Formal)* Let $(\rho_t, v_t)$ be a competitor. Define $X(t,x)$ as the flow of $v_t$. , by the uniqueness for the continuity equation, the unique solution $\rho_t$ is given by the formula proved above, namely $\rho_t = (X(t))_{\#}\bar{\rho}_0$. Then, by Hölder inequality,

$$\begin{aligned} A[\rho_t, v_t] &= \int_0^1 \int_\Omega |v_t|^2 \rho_t dxdt = \int_0^1 \int_\Omega |v_t(X(t,x))|^2 \bar{\rho}_0(x)dxdt \\ &= \int_0^1 \int_\Omega |\dot{X}(t,x)|^2 \bar{\rho}_0(x)dtdx = \int_\Omega \bar{\rho}_0(x) \int_0^1 |\dot{X}(t,x)|^2 dtdx \\ &\geq \int_\Omega \bar{\rho}_0(x)\left|\int_0^1 \dot{X}(t,x)dt\right|^2 dx = \int_\Omega \bar{\rho}_0(x)|X(1,x) - x|^2 dx \\ &\geq W_2^2(\bar{\rho}_0, \bar{\rho}_1) \end{aligned}$$

(since $X(1)_{\#}\bar{\rho}_0 = \bar{\rho}_1$).

To show equality, take $X(1,x) = T(x)$, where $T = \nabla\varphi$ is optimal from $\bar{\rho}_0$ to $\bar{\rho}_1$. Define $X(t,x) = x + t(T(x) - x)$, $\rho_t := X(t)_{\#}\bar{\rho}_0$, and $v_t$ such that $\dot{X}(t) = v_t \circ X(t)$. With this choice we have $(T(x) - x) = \dot{X}(t,x) = v_t(X(t,x))$, and looking at the computations above one can easily check what, with this choice,

$$A[\rho_t, v_t] = W_2^2(\bar{\rho}_0, \bar{\rho}_1).$$

So, to conclude the proof, we need just to ensure that $v_t = \dot{X} \circ X^{-1}(t)$ is well defined, and for this we need to show that $X(t)^{-1}$ exists. Note that

$$\begin{aligned}|X(t,x) - X(t,\tilde{x})||x-\tilde{x}| &\geq \langle X(t,x) - X(t,\tilde{x}), x-\tilde{x}\rangle \\ &= (1-t)\langle x-\tilde{x}, x-\tilde{x}\rangle + t\underbrace{\langle \nabla\varphi(x) - \nabla\varphi(\tilde{x}), x-\tilde{x}\rangle}_{\geq 0 \ (\varphi \text{ convex})} \\ &\geq (1-t)|x-\tilde{x}|^2.\end{aligned}$$

Thus, for $t \in [0,1)$, $X(t)$ is injective and $X(t)^{-1}$ exists, concluding the proof. □

*Remark* 4.1.3. For $t=1$, $X(1) = \nabla\varphi(x)$. If $\bar{\rho}_1 = \delta_{\bar{x}}$, then $\nabla\varphi(x) = \bar{x}$ is constant and obviously not injective.

Otto's idea

$$\begin{aligned}W_2^2(\bar{\rho}_0,\bar{\rho}_1) &= \inf_{\rho_t, v_t}\left\{\int_0^1 \left(\int |v_t|^2 \rho_t dx\right)dt \mid \partial_t\rho + \operatorname{div}(v_t\rho_t) = 0,\ \rho_0 = \bar{\rho}_0,\ \rho_1 = \bar{\rho}_1\right\} \\ &= \inf_{\rho_t}\left\{\inf_{v_t}\int_0^1 |v_t|^2\rho_t dx dt \mid \partial_t\rho + \operatorname{div}(v_t\rho_t) = 0,\ \rho_0 = \bar{\rho}_0,\ \rho_1 = \bar{\rho}_1\right\} \\ &= \inf_{\rho_t}\left\{\int_0^1 \underbrace{\left\{\inf_{v_t}\int |v_t|^2\rho_t dx \mid \operatorname{div}(v_t\rho_t) = -\partial_t\rho_t\right\}}_{=:\|\partial_t\rho_t\|_{\rho_t}^2} dt \mid \rho_0 = \bar{\rho}_0,\ \rho_1 = \bar{\rho}_1\right\}.\end{aligned}$$

We call $\|\partial_t\rho_t\|_{\rho_t}$ the **Wasserstein-norm** of $\partial_t\rho_t$ at $\rho_t$.

Now, for $t$ and $\rho_t$ fixed (hence $\partial_t\rho_t$ is given), let $v_t$ be a minimizer of $\int |v_t|^2\rho_t dx$ among all $v_t$ such that $\operatorname{div}(v_t\rho_t) = -\partial_t\rho_t$. Fix $w$ a vector field such that $\operatorname{div}(w) \equiv 0$. Then for every $\varepsilon > 0$, we have

$$\operatorname{div}\left(v_t + \varepsilon\frac{w}{\rho_t}\rho_t\right) = -\partial_t\rho_t.$$

Thus $v_t + \varepsilon\frac{w}{\rho_t}$ is an admissible vector field with

$$\begin{aligned}\int |v_t|^2\rho_t &\leq \int \left|v_t + \varepsilon\frac{w}{\rho_t}\right|^2\rho_t \\ &= \int |v_t|^2\rho_t + 2\varepsilon\int\langle v_t, w\rangle + \varepsilon^2\int\frac{|w|^2}{\rho_t}.\end{aligned}$$

Dividing by $\varepsilon$ and letting it go to zero yields

$$\int\langle v_t, w\rangle = 0$$

for every $w$ such that $\operatorname{div}(w) \equiv 0$. By the Helmholtz decomposition, we know that

$$v_t \in \{w \mid \operatorname{div}(w) = 0\}^\perp = \{\nabla q \mid q \in C^\infty(\Omega)\}.$$

Therefore, there exists a $\psi_t$ such that $v_t = \nabla\psi_t$. Also, $\psi_t$ solves

$$\operatorname{div}(\rho_t\nabla\psi_t) = -\partial_t\rho_t \ (\star).$$

Note that, if $\rho_t$ is nice (positive and smooth) then $(\star)$ is a nice elliptic equation for $\psi_t$, and therefore the solution $\psi_t$ is unique up to a constant. So one can define

$$\|\partial_t\rho_t\|_{\rho_t}^2 = \int |\nabla\psi_t|^2\rho_t,$$

where $\psi_t$ solves $(\star)$. So, we found a nice expression for the Wasserstein norm.

Once a norm is defined, we can construct the scalar product: given $h_1, h_2$, one can define their **Wasserstein scalar product** at $\rho$ as

$$\langle h_1, h_2 \rangle_\rho \coloneqq \int \nabla\psi_1 \cdot \nabla\psi_2 \rho,$$

where $\mathrm{div}(\nabla\psi_i\rho) = -h_i$.

We can now define gradients. Given $F\colon \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ a functional, define the **gradient with respect to the Wasserstein scalar product** at $\bar\rho$ as the unique function $\mathrm{grad}_{W_2}F[\bar\rho]$ such that the following identity holds:

$$\left\langle \mathrm{grad}_{W_2}F[\bar\rho], \left.\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0} \right\rangle_{\bar\rho} = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} F[\rho_\varepsilon]$$

for any smooth curve $\rho_\varepsilon$ such that $\rho_0 = \bar\rho$.

Then, given $F$, the **gradient flow of F with respect to** $W_2$ is given by

$$\partial_t\rho_t = -\mathrm{grad}_{W_2}F[\rho_t].$$

**Example 4.1.4.** Let $F[\rho] = \int \rho\log(\rho)dx + \int V\rho dx$ for $V$ a vector field. Fix $\bar\rho$ and let $\varepsilon \mapsto \rho_\varepsilon$ be a smooth curve such that $\rho_0 = \bar\rho$. Also, let $\psi$ solve $\mathrm{div}(\nabla\psi\bar\rho) = -\frac{\partial\rho_\varepsilon}{\partial\varepsilon}|_{\varepsilon=0}$. Then

$$\begin{aligned}
\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} F[\rho_\varepsilon] &= \int \left.\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0} \log(\bar\rho) + \int \bar\rho \left.\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0} \frac{1}{\bar\rho} + \int V \left.\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0} \\
&= -\int [\log(\bar\rho) + 1 + V]\mathrm{div}(\nabla\psi\bar\rho) \\
&= \int \nabla[\log(\bar\rho) + 1 + V]\nabla\psi\bar\rho.
\end{aligned}$$

Thus, recalling the definition of the Wasserstein scalar product, we see

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} F[\rho_\varepsilon] = \left\langle \left.\frac{\partial_\varepsilon\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0}, \bar h \right\rangle_{\bar\rho},$$

where $\bar h \coloneqq -\mathrm{div}(\nabla[\log\bar\rho + 1 + V]\bar\rho)$. Thus, by the definition of gradient in Wasserstein, we proved that

$$\mathrm{grad}_{W_2}F[\bar\rho] = -\mathrm{div}(\nabla[\log\bar\rho + 1 + V]\bar\rho).$$

Hence the gradient flow of $F$ with respect to $W_2$ is

$$\begin{aligned}
\partial_t\rho_t &= \mathrm{div}(\nabla[\log(\rho_t) + 1 + V]\rho_t) \\
&= \mathrm{div}(\nabla\rho_t + (\nabla V)\rho_t).
\end{aligned}$$

General rule: Given $F$ a functional, then

$$\left\langle \mathrm{grad}_{W_2}F[\bar\rho], \left.\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0} \right\rangle_{\bar\rho} = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} F[\rho_\varepsilon] = \int \underbrace{\frac{\delta F[\bar\rho]}{\delta\rho}}_{\substack{\text{first } L^2-\text{variation}}} \left.\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0}.$$

Let $\psi$ solve $\mathrm{div}(\nabla\psi\bar\rho) = -\frac{\partial\rho_\varepsilon}{\partial\varepsilon}|_{\varepsilon=0}$. Thus

$$\left\langle \mathrm{grad}_{W_2}F[\bar\rho], \left.\frac{\partial\rho_\varepsilon}{\partial\varepsilon}\right|_{\varepsilon=0} \right\rangle_{\bar\rho} = -\int \frac{\delta F[\bar\rho]}{\delta\rho}\,\mathrm{div}(\nabla\psi\bar\rho) = \int \nabla\frac{\delta F[\bar\rho]}{\delta\rho}\cdot\nabla\psi\bar\rho.$$

Hence, by the definition of Wasserstein gradient,

$$\mathrm{grad}_{W_2}F[\bar\rho] = -\mathrm{div}\Big(\nabla\Big(\frac{\delta F[\bar\rho]}{\delta\rho}\Big)\bar\rho\Big).$$

**Example 4.1.5.** • Given $F[\rho] = \frac{1}{m-1}\int \rho^m dx$ $(m \neq 1)$, then

$$\frac{\delta F[\bar\rho]}{\delta \rho} = \frac{m}{m-1}\bar\rho^{m-1} \quad \Rightarrow \quad \nabla\Big(\frac{\delta F[\bar\rho]}{\delta \rho}\Big)\bar\rho = m\bar\rho^{m-2}\nabla\bar\rho\,\bar\rho = \nabla(\bar\rho^m),$$

hence

$$\partial_t \rho_t = -\mathrm{grad}_{W_2} F[\rho_t] = \Delta[\rho_t^m].$$

• Given $F[\rho] = \frac{1}{2}\iint \rho(x)\rho(y)W(x-y)dxdy$ with $W(z) = W(-z)$, then

$$\frac{\delta F[\bar\rho]}{\delta \rho} = W * \bar\rho \quad \Rightarrow \quad \nabla\Big(\frac{\delta F[\bar\rho]}{\delta \rho}\Big)\bar\rho = (\nabla W * \bar\rho)\bar\rho,$$

thus

$$\partial_t \rho_t = -\mathrm{grad}_{W_2} F[\rho_t] = \mathrm{div}((\nabla W * \rho_t)\rho_t).$$

## 4.2 Wasserstein convexity

*Displacement convexity - McCann '95*

Convexity along geodesics: let $\rho_t$ be a geodesic given by $(T_t)_\#\bar\rho_0$ where $T_t(x) = x + t(\nabla\varphi(x) - x)$, where $\nabla\varphi$ is the optimal transport from $\bar\rho_0$ to $\bar\rho_1$ (see Section 2.8.1). Let $F[\rho] = \int G(\rho(x))dx$ and consider the function

$$[0,1] \ni t \mapsto F[\rho_t] = \int G(\rho_t(x))dx.$$

We want to understand under which assumptions on $G$ this function is convex.

Assume that $T$ is smooth. Recalling Section 1.5, we have $\rho_t \circ T_t = \frac{\bar\rho_0}{\det \nabla T_t}$. Therefore, since $\nabla T_t = (1-t)\mathrm{Id} + t\, D^2\varphi$,

$$\begin{aligned} F[\rho_t] &= \int \frac{G(\rho_t(x))}{\rho_t(x)}\rho_t(x)dx = \int \frac{G(\rho_t \circ T_t)}{\rho_t \circ T_t}\bar\rho_0 dx \\ &= \int G\Big(\frac{\bar\rho_0}{\det \nabla T_t}\Big)\det \nabla T_t dx \\ &= \int G\Big(\frac{\bar\rho_0}{\det((1-t)\mathrm{Id} + tD^2\varphi)}\Big)\det\big((1-t)\mathrm{Id} + tD^2\varphi(x)\big)dx. \end{aligned}$$

*Remark* 4.2.1. $D^2\varphi \geq 0$ since $\varphi$ is convex.

Fix $x$. Then, in some suitable basis that makes $D^2\varphi(x)$ diagonal,

$$\begin{aligned} \det((1-t)\mathrm{id} + tD^2\varphi) &= \det\begin{pmatrix} (1-t)+t\lambda_1(x) & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & (1-t)+\lambda_n(x) \end{pmatrix}. \\ &= \Pi_{i=1}^n((1-t)+t\lambda_i(x)) \\ &= \Big[\underbrace{\Pi_{i=1}^n((1-t)+t\lambda_i(x))^{\frac{1}{n}}}_{:=\det(t)^{\frac{1}{n}}}\Big]^n. \end{aligned}$$

Fact: $t \mapsto \det(t)^{1/n}$ is concave. Then

$$F[\rho_t] = \int G\Big(\frac{\bar\rho_0(x)}{(\det(t)^{\frac{1}{n}})^n}\Big)\Big[\det(t)^{\frac{1}{n}}\Big]^n dx.$$

When is this composition convex? Since $t \mapsto \det(t)^{1/n}$ is concave, the function

$$t \mapsto G\bigg(\frac{\bar\rho_0(x)}{(\det(t)^{\frac1n})^n}\bigg)\Big[\det(t)^{\frac1n}\Big]^n$$

is convex for any $x$ if the function $s \mapsto G(\frac{a}{s^n}s^n)$ is convex and decreasing for any $a \geq 0$.

**Example 4.2.2.** Let $F[\rho] = \int G(\rho(x))dx$. The following functionals are convex in the Wasserstein sense:

$$G(\tau) \coloneqq \begin{cases} \tau\log(\tau) & \to \partial_t\rho_t = \Delta\rho \text{ (Heat eq.)} \\ \frac{1}{m-1}\tau^m \quad \text{for } m > 1 & \to \partial_t\rho_t = \Delta(\rho^m) \text{ (Porous medium eq.)} \\ -\frac{1}{1-m}\tau^m \quad \text{for } m \in [1-\frac1n, 1)) & \to \partial_t\rho_t = \Delta(\rho^m) \text{ (Fast diffusion eq.).} \end{cases}$$

**Example 4.2.3.** The map $t \mapsto \int V\rho_t$ is convex if $V$ is.
The map $t \mapsto \frac12 \iint W(x-y)\rho_t(x)\rho_t(y)$ is convex if $W$ is.
*Exercise.* Prove it.
*Hint.* Write

$$\int V(x)\rho_t(x) = \int V(T_t(x))\bar\rho_0(x) = \int V((1-t)x + tT(x))\bar\rho_0(x)$$

and

$$\iint W(x-y)\rho_t(x)\rho_t(y) = \iint W((1-t)x - (1-t)y + tT(x) - tT(y))\bar\rho_0(x)\bar\rho_0(y),$$

and compute the second derivative with respect to $t$.

## 4.3 Generalizations/extensions

### 4.3.1 $\lambda$-convexity

A smooth function $\phi : \mathbb{R}^d \to \mathbb{R}$ is said to be $\lambda$-convex for some $\lambda \in \mathbb{R}$ if

$$D^2\phi(x) \geq \lambda\,\mathrm{Id} \qquad \forall\, x \in \mathbb{R}^d.$$

Equivalently, since $\frac{d^2}{d\varepsilon^2}\phi(x+\varepsilon v)|_{\varepsilon=0} = D^2\phi(x)[y,y]$, $F$ is said to be $\lambda$-convex if

$$\frac{d^2}{d\varepsilon^2}F(x+\varepsilon y)\bigg|_{\varepsilon=0} \geq \lambda|y|^2 \qquad \forall\, x, y.$$

*Remark* 4.3.1. The definition of $\lambda$-convexity can also be given for non-smooth functions (in the same way as one can introduce the concept of convex functions without smoothness). More precisely, since by Taylor

$$\phi(x+w) = \phi(x) + \langle\nabla F(x), w\rangle + \frac12\int_0^1 D^2\phi(x+tw)[w,w]\,dt$$

then if $\psi$ is $\lambda$-convex we obtain

$$\begin{aligned}
&(1-s)\phi(x+sy) + s\phi(x-(1-s)y) \\
&= \phi(x) + \frac{(1-s)s^2}{2}\int_0^1 D^2F(x+tsy)[y,y]\,dt + \frac{s(1-s)^2}{2}\int_0^1 D^2\phi(x-(1-s)ty)[y,y]\,dt \\
&\geq \phi(x) + \frac{(1-s)s^2 + s(1-s)^2}{2}\lambda|y|^2 = \phi(x) + \frac{(1-s)s}{2}\lambda|y|^2.
\end{aligned}$$

Thus, setting $x + sy = z$ and $x - (1-s)y = z'$ we have $x = (1-s)z + sz'$ and $|y| = |z - z'|$, hence

$$(1-s)\phi(z) + s\phi(z') \geq \phi((1-s)z + sz') + \frac{(1-s)s}{2}\lambda|z - z'|^2 \qquad \forall\, 0 \leq s \leq 1,\ z, z' \in \mathbb{R}^n.$$

Then, one takes the above formula as definition of $\lambda$-convexity for general functions $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$.

Assume for simplicity that $F : \mathbb{R}^d \to \mathbb{R}$ is a smooth $\lambda$-convex function, and let $\dot{x} = -\nabla\phi(x)$ and $\dot{y} = -\nabla\phi(y)$ be two GF of $\phi$. Then, since $\nabla\phi(x) - \nabla\phi(y) = \int_0^1 D^2\phi(tx + (1-t)y) \cdot (x-y)\, dt$,

$$\begin{aligned}\frac{d}{dt}\frac{|x-y|^2}{2} &= \langle \dot{x} - \dot{y}, x - y\rangle = -\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \\ &= -\int_0^1 \langle D^2\phi(tx + (1-t)y) \cdot (x-y), x - y\rangle dt \leq -\lambda|x-y|^2,\end{aligned}$$

thus

$$t \mapsto e^{2\lambda t}|x(t) - y(t)|^2 \quad \text{ is decreasing.}$$

In particular

$$|x(t) - y(t)| \leq e^{-\lambda t}|x(0) - y(0)|.$$

If $\lambda > 0$, this proves that $x(t)$ and $y(t)$ approach each other exponentially fast, while for $\lambda < 0$ this estimate provides a quantitative closeness between $x(t)$ and $y(t)$ on any finite time interval $[0, T]$. One can extend this analogy to the Wasserstein space. More precisely, one says that $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ is $\lambda$-convex if, for any $\{\rho_s\}_{0 \leq s \leq 1}$ 2-Wasserstein geodesic,

$$(1-s)F[\rho_0] + sF[\rho_1] \geq F[\rho_s] + \frac{(1-s)s}{2}\lambda W_2(\rho_0, \rho_1)^2 \qquad \forall\, 0 \leq s \leq 1.$$

If $F$ is sufficiently smooth, this is equivalent to saying that

$$\frac{d^2}{ds^2}F[\rho_s] \geq \lambda\|\partial_s\rho_s\|_{\rho_s}^2,$$

where $\|\partial_s\rho_s\|_{\rho_s}$ is the Wasserstein-norm of $\partial_s\rho_s$. Then, if $\partial_t\rho_i = -\mathrm{grad}_{W_2}F[\rho_i]$, $i = 1, 2$, and $F$ is $\lambda$-convex then

$$\frac{d}{dt}W_2^2(\rho_1(t), \rho_2(t)) \leq 2\lambda W_2^2(\rho_1(t), \rho_2(t)).$$

**Example 4.3.2.** Let $F$ be defined on the space of densities as

$$F[\rho] := \int G(\rho(x))dx + \int V(x)\rho(x)dx + \frac{1}{2}\iint W(x-y)\rho(x)\rho(y)dxdy.$$

Assume that:
- $(0, +\infty) \ni s \mapsto G(\frac{a}{s^n})s^n$ is convex and decreasing for any $a \geq 0$.
- $V : \mathbb{R}^d \to \mathbb{R}$ is $\lambda$-convex.
- $W : \mathbb{R}^d \to \mathbb{R}$ is even and convex.

Then $F$ is $\lambda$-convex. (More precisely, the first and the third terms are convex in Wasserstein, while the second term is $\lambda$-convex, so the sum is $\lambda$-convex).

### 4.3.2 From $\mathbb{R}^d$ to Riemannian manifolds

What if one considers a Riemannian manifold $(M, g)$ and the space $\mathcal{P}(M)$? One can repeat essentially verbatim all the construction of Otto calculus. However, when computing the convexity properties of functionals along Wasserstein geodesics, the geometric of the manifold $M$ plays a crucial role. More precisely, the convexity of the functionals is affected by the Ricci curvature of $(M, g)$. One can actually prove the following characterization:

$$\rho \mapsto F[\rho] = \int \rho(x) \log(\rho(x)) d\mathrm{vol}(x) \text{ is Wasserstein convex}$$
$$\text{if and only if } M \text{ has nonnegative Ricci curvature.}$$

This is the starting point of a still very active area of research concerning the study of spaces with Ricci curvature bounded from below.

# 5 Exercises on optimal transport (with solutions)

In this set of exercises, an *optimal transport map* should be understood as an optimal transport plan that is also a map (so, it must be optimal among all possible transport plans). The linear cost is $c(x,y) = |x-y|$, whereas the quadratic cost is $c(x,y) = \frac{1}{2}|x-y|^2$.

**Exercise 1.1** (Translations are optimal)**.** Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be the translation map $T(x) \coloneqq x + x_0$. For any probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$, show that $T$ is an optimal transport map from $\mu$ to $T_\#\mu$ with respect to the quadratic cost.

*Solution.* Let us recall that the gradient of a convex function is always an optimal map (with respect to the quadratic cost) from a probability measure to its push-forward through such map (see Corollary 2.6.2). Hence, to prove the optimality of the translation map $T$, it is sufficient to check that it is the gradient of a convex function.

Let $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be the function $\varphi(x) \coloneqq \frac{1}{2}|x + x_0|^2$. Since $\varphi$ is convex and it holds $T = \nabla\varphi$, the optimality of $T$ follows. □

**Exercise 1.2** (Homotheties are optimal)**.** Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be the homothety $T(x) \coloneqq \lambda x$ where $\lambda > 0$. For any compactly supported probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$, show that $T$ is an optimal transport map from $\mu$ to $T_\#\mu$ with respect to the quadratic cost.

*Solution.* As explained in the solution of Exercise 1.1, it is sufficient to show that the homothety $T$ is the gradient of a convex function.

Let $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be the function $\varphi(x) \coloneqq \frac{\lambda}{2}|x|^2$. Since $\varphi$ is convex and it holds $T = \nabla\varphi$, the optimality of $T$ follows. □

**Exercise 1.3.** Let $\mu \coloneqq \frac{1}{\pi}\chi_{B(0,1)}\mathcal{L}^2$ be the uniform probability measure on $B(0,1) \subset \mathbb{R}^2$ and let $p_1 \coloneqq (1,0)$, $p_2 \coloneqq (2,0)$ be two fixed points in $\mathbb{R}^2$. Describe the optimal transport map between $\mu$ and $\frac{1}{2}(\delta_{p_1} + \delta_{p_2})$ in the following two cases:

(a) when the cost is the quadratic cost $\frac{1}{2}|x-y|^2$;

(b) when the cost is the linear cost $|x-y|$.

*Solution.* Let us work in a more general setting. Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu = \frac{1}{2}\delta_{p_1} + \frac{1}{2}\delta_{p_2}$, where $p_1, p_2$ are two distinct points in $\mathbb{R}^d$. Let $\gamma \in \Gamma(\mu,\nu)$ be an *optimal* plan with respect to a certain cost $c : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ that we assume to be continuous.

From the fact that $\gamma$ is an admissible plan, we can deduce that it is supported on $\mathbb{R}^d \times \{p_1, p_2\}$. Moreover, being optimal, it must also be supported on a $c$-cyclically monotone set. Thus, there must be two measurable subsets $A_1, A_2 \subseteq \mathbb{R}^d$ such that $\gamma$ is supported on $A \coloneqq A_1 \times \{p_1\} \cup A_2 \times \{p_2\}$ and $A$ is a $c$-cyclically monotone set. Since $\gamma$ is an admissible transport plan, for any measurable $S \subseteq \mathbb{R}^d$ it holds

$$\mu(S) = \gamma((A_1 \cap S) \times \{p_1\}) + \gamma((A_2 \cap S) \times \{p_2\})\,. \tag{5.1}$$

Take $a_1 \in A_1$ and $a_2 \in A_2$. By the $c$-cyclical monotonicity of $A$, we deduce

$$c(a_1,p_1) + c(a_2,p_2) \le c(a_1,p_2) + c(a_2,p_1) \iff c(a_1,p_1) - c(a_1,p_2) \le c(a_2,p_1) - c(a_2,p_2)\,.$$

Therefore, denoting $w(x) \coloneqq c(x,p_1) - c(x,p_2)$, we have shown $w(a_1) \le w(a_2)$. Since this holds for any $a_1 \in A_1$ and $a_2 \in A_2$, there must be a value $t_0 \in \mathbb{R}$ such that $A_1 \subseteq \{w \le t_0\}$ and $A_2 \subseteq \{w \ge t_0\}$.

In order to get some additional information on the value of $t_0$, let us apply (5.1). Setting $S = \{w \le t_0\}$ we obtain

$$\mu(\{w \le t_0\}) = \gamma(A_1 \times \{p_1\}) + \gamma((A_2 \cap \{w \le t_0\}) \times \{p_1\}) \ge \gamma(A_1 \times \{p_1\}) = \nu(\{p_1\}) = \frac{1}{2}\,.$$

Differently, setting $S = \{w < t_0\}$, we get

$$\begin{aligned}\mu(\{w < t_0\}) &= \gamma((A_1 \cap \{w < t_0\}) \times \{p_1\}) + \gamma(\emptyset \times \{p_1\}) = \gamma((A_1 \cap \{w < t_0\}) \times \{p_1\}) \\ &\leq \gamma(A_1 \times \{p_1\}) = \nu(\{p_1\}) = \frac{1}{2}\,.\end{aligned}$$

Joining the last two inequalities, we get

$$\mu(\{w < t_0\}) \leq \frac{1}{2} \leq \mu(\{w \leq t_0\})\,. \tag{5.2}$$

If we assume that $\mu(\{w = t\}) = 0$ for every $t \in \mathbb{R}$, the condition (5.2) identifies uniquely $t_0$. Moreover, under this additional assumption, it follows directly from our observations that $\gamma$ is unique and is induced by the map $T : \mathbb{R}^d \to \mathbb{R}^d$ defined as

$$T(x) \coloneqq \begin{cases} p_1 & \text{if } w(x) \leq t_0, \\ p_2 & \text{if } w(x) > t_0. \end{cases} \tag{5.3}$$

Now we go back to solving the statement of the exercise.

(a) Since $c(x, y) = \frac{1}{2}|x - y|^2$, it holds

$$w(x) = c(x, p_1) - c(x, p_2) = \langle x, p_2 - p_1\rangle + \frac{1}{2}(p_1^2 - p_2^2) = \langle x, e_1\rangle + \frac{3}{2}\,.$$

Notice that, for any $t \in \mathbb{R}$, $\{w = t\}$ is a vertical line and therefore $\mu(\{w = t\}) = 0$. Hence the assumption necessary to deduce that the optimal transport map is (5.3) is satisfied. To conclude, it remains only to determine the value of $t_0$. The condition (5.2) tells us that the line $\{w = t_0\}$ must split in two parts with equal mass the measure $\mu$ and therefore it must split in two parts with the same area the ball $B(0, 1)$. Therefore the optimal transport map between $\mu$ and $\nu$ is

$$T(x) \coloneqq \begin{cases} p_1 & \text{if } x \cdot e_1 \leq 0, \\ p_2 & \text{if } x \cdot e_1 > 0. \end{cases}$$

(b) In this case it holds $c(x, y) = |x - y|$ and therefore

$$w(x) = c(x, p_1) - c(x, p_2) = |x - p_1| - |x - p_2|\,.$$

Notice that, for any $t \in \mathbb{R}$, $\{w = t\}$ is a hyperbola with foci $p_1$ and $p_2$ (or the empty set) and therefore $\mu(\{w = t\}) = 0$. Reasoning exactly as in part (a), we deduce that the optimal map from $\mu$ to $\nu$ is

$$T(x) \coloneqq \begin{cases} p_1 & \text{if } |x - e_1| - |x - 2e_1| \leq t_0, \\ p_2 & \text{if } |x - e_1| - |x - 2e_1| > t_0; \end{cases}$$

where $t_0 \in \mathbb{R}$ is the only value such that (5.2) is satisfied (so the optimal map sends the interior of an hyperbola into $p_1$ and the exterior into $p_2$). We content ourselves with this description of the optimal map, without trying to determine explicitly the value of $t_0$.

□

**Exercise 1.4** (Counterexamples)**.** For any of the following statements, find two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with compact support such that the statement holds (you can choose also the dimension $d \in \mathbb{N}$). Each of the statements should be treated independently.

(a) There is more than one[6] optimal transport map from $\mu$ to $\nu$ with respect to the linear cost $|x-y|$.

(b) There is more than one optimal transport map from $\mu$ to $\nu$ with respect to the quadratic cost $\frac{1}{2}|x-y|^2$.

(c) There is not an optimal transport plan between $\mu$ and $\nu$ with respect to the cost $c(x,y) = \lfloor |x-y| \rfloor$ (the floor function[7] of the distance).

(d) There is an optimal transport map from $\mu$ to $\nu$ with respect to the linear cost, but there is none with respect to the quadratic cost.

(e) There is an optimal transport map from $\mu$ to $\nu$ with respect to the quadratic cost, but there is none with respect to the linear cost.

*Hint:* To solve (c), show that the infimum of the Kantorovich problem for $\mu = \chi_{[0,1]}\mathcal{L}^1$, $\nu = \chi_{[1,2]}\mathcal{L}^1$ is 0 but any transport plan has *strictly* positive cost.
To solve (e), it might be useful to first solve Exercise 1.3.

*Solution.*

(a) Let $d = 1$ and $\mu := \frac{1}{2}(\delta_0 + \delta_1)$, $\nu := \frac{1}{2}(\delta_1 + \delta_2)$. Every transport plan from $\mu$ to $\nu$ has (linear) cost 2, thus the two maps $T_1, T_2 : \{0,1\} \to \{1,2\}$ given by

$$T_1(0) = 1, \quad T_1(1) = 2, \qquad T_2(0) = 2, \quad T_2(1) = 1,$$

both send $\mu$ to $\nu$ and are both optimal.

(b) Let $d = 2$ and let $p_1, p_2, p_3 \subseteq \mathbb{R}^2$ be the three vertices of an equilateral triangle. Let $\mu := \frac{1}{2}(\delta_{p_1} + \delta_{p_2})$ and $\nu := \frac{1}{2}(\delta_{p_3} + \delta_{\frac{p_1+p_2+p_3}{3}})$. Every transport plan from $\mu$ to $\nu$ has the same (quadratic) cost. Thus the two maps from $\mu$ to $\nu$ (namely, $T_1(p_1) = p_3$, $T_1(p_2) = \frac{p_1+p_2+p_3}{3}$, and $T_2(p_1) = \frac{p_1+p_2+p_3}{3}$, $T_2(p_2) = p_3$) are both optimal.

(c) Let $d = 1$ and $\mu := \mathcal{L}^1|_{[0,1]}$, $\nu := \mathcal{L}^1|_{[1,2]}$. We will show that the infimum of the Kantorovich problem is 0, but any transport plan has *strictly* positive cost.

Given $\varepsilon > 0$, consider the map $T_\varepsilon : [0,1] \to [1,2]$ defined as

$$T_\varepsilon(x) := \begin{cases} x + 2 - \varepsilon & \text{if } 0 \le x \le \varepsilon, \\ x + 1 - \varepsilon & \text{if } \varepsilon < x \le 1. \end{cases}$$

One can check that $(T_\varepsilon)_\# \mu = \nu$. Also, it holds

$$\int_{[0,1]} c(x, T_\varepsilon(x)) \, \mathrm{d}\mu(x) = \int_0^\varepsilon \lfloor 2 - \varepsilon \rfloor \, \mathrm{d}x + \int_\varepsilon^1 \lfloor 1 - \varepsilon \rfloor = \varepsilon \,.$$

Since $\varepsilon$ can be chosen arbitrarily small, we have proven that the infimum of the Kantorovich problem is 0.

Let us assume by contradiction that $\gamma \in \Gamma(\mu,\nu)$ achieves cost 0. Hence $\gamma$ must be supported on the set

$$\{(x,y) \in \mathbb{R}^2 : \ c(x,y) = 0\} = \{(x,y) \in \mathbb{R}^2 : \ |x-y| < 1\} \,. \tag{5.4}$$

[6] Uniqueness should be understood in the $\mu$-almost everywhere sense.
[7] Given $t \ge 0$, $\lfloor t \rfloor$ is the largest integer $n$ such that $n \le t$.

Moreover, since $\gamma \in \Gamma(\mu,\nu)$, the plan is supported on $[0,\,1]\times[1,\,2]$. Therefore, for any $0<l<1$, it holds

$$\begin{aligned} l &= \nu([2-l,\,2]) = \gamma(\mathbb{R}\times[2-l,\,2]) = \gamma([0,\,1]\times[2-l,\,2]) \\ &\overset{(5.4)}{=} \gamma([1-l,\,1]\times[2-l,\,2]) = \mu([1-l,\,1]) - \gamma([1-l,\,1]\times[1,\,2-l)) \\ &= l - \gamma([1-l,\,1]\times[1,\,2-l)) \end{aligned}$$

and therefore $\gamma([1-l,\,1]\times[1,\,2-l))=0$. Since it holds

$$([0,\,1]\times[1,\,2])\cap\{(x,y)\in\mathbb{R}^2:\ |x-y|<1\} \subseteq \bigcup_{l\in\mathbb{Q}\cap(0,\,1)} [1-l,\,1]\times[1,\,2-l)\ ,$$

we reach a contradiction as we have proven that $\gamma\equiv 0$.

(d) Let $d=1$ and $\mu := \mathcal{L}^1|_{[0,\,\frac12]} + \frac12\delta_1$, $\nu := \mathcal{L}^1|_{[\frac52,\,3]} + \frac12\delta_2$. Since $\mu$ is supported on $\{x<2\}$, whereas $\nu$ is supported on $\{x\geq 2\}$, any admissible plan has the same cost with respect to the linear cost, and this cost is

$$\int x\,\mathrm{d}\nu - \int x\,\mathrm{d}\mu = \frac74\,.$$

In particular, any admissible map is optimal with respect to the linear cost (and it is clear that there is at least one).

On the other hand, the optimal map with respect to the quadratic cost (if it exists) must be nondecreasing (since, in one dimension, begin the gradient of a convex function is equivalent to being nondecreasing). Let us assume by contradiction that there is an admissible transport map $T:\mathbb{R}\to\mathbb{R}$ that is nondecreasing. Since it is an admissible map, it must hold $T(1)=2$. The monotonicity then implies that $T(x)\leq 2$ for any $x\leq 1$ and thus $\nu = T_\#\mu$ is supported on $x\leq 2$, that is a contradiction.

(e) Let $d=1$ and $\nu := \frac12(\delta_{p_1}+\delta_{p_2})$, where $p_1=(1,0)$ and $p_2=(2,0)$. Let $A=\{(x,y)\in\mathbb{R}^2 : x<0\}$ and $B=\{p\in\mathbb{R}^2 : |p-p_1|-|p-p_2|<-\frac12\}$ (let us remark that the exact form of $A$ and $B$ is not important, it is only important that they are respectively a half-plane and the interior of an hyperbola).

Take $p_3\in A\cap\partial B$, $p_4\in \bar A^{\mathsf{c}}\cap B$, $p_5\in \bar A^{\mathsf{c}}\cap \bar B^{\mathsf{c}}$ and define $\mu := \frac12\delta_{p_3}+\frac14\delta_{p_4}+\frac14\delta_{p_5}$.

Recalling the observations contained in the solution of Exercise 1.3, we can say that (we identify the mass that $\mu$ and $\nu$ give to a point with the point itself):

- Any optimal transport plan with respect to the quadratic cost from $\mu$ to $\nu$ sends $p_3$ into $p_1$ and $\{p_4,p_5\}$ into $p_2$. In particular, there is an optimal transport map (because the map that does exactly as described is admissible).
- Any optimal transport plan with respect to the linear cost from $\mu$ to $\nu$ sends $p_4$ to $p_1$ and $p_5$ to $p_2$. This is an obstruction to the existence of an optimal transport map, as the mass of $p_3$ (that is $\frac12$) should be split into two equal parts (but a map cannot split a delta).

In particular, there is an optimal transport map from $\mu$ to $\nu$ with respect to the quadratic cost, but there is known with respect to the linear cost.

□

**Exercise 1.5** (Birkhoff-von Neumann theorem)**.** An $n\times n$ matrix $A\in\mathscr{M}(n,\mathbb{R})$ with nonnegative entries is:

- *a doubly-stochastic matrix* if $\sum_{i=1}^n A_{ij} = 1$ for any $j = 1, \dots, n$ and $\sum_{j=1}^n A_{ij} = 1$ for any $i = 1, \dots, n$.
- *a permutation matrix* if there is a permutation $\sigma \in S^n$ such that $A_{i\sigma(i)} = 1$ and $A_{ij} = 0$ if $j \neq \sigma(i)$.

Prove that any doubly-stochastic matrix can be written as a finite convex combination of permutation matrices.
*Hint:* Use Hall's marriage theorem to show that for any doubly-stochastic matrix $A$, there is a permutation matrix $P$ and a number $\varepsilon > 0$ such that $A_{ij} \geq \varepsilon P_{ij}$ for any $1 \leq i, j \leq n$.

*Solution.* Let us begin with the following lemma.

**Lemma.** *Given a doubly-stochastic matrix $A$, there is a permutation $\sigma \in S_n$ such that $A_{i\sigma(i)} > 0$ for any $i = 1, \dots, n$.*

*Proof.* Let us construct a bipartite graph as follows. There are $2n$ vertices labeled $\{1_r, \dots, n_r\}$ and $\{1_c, \dots, n_c\}$; the indexes $r, c$ stand for row and column. There is an edge between $i_r$ and $j_c$ if and only if $A_{ij} > 0$. We denote the presence of an edge with $i_r \sim j_c$. The first step of the proof consists in showing that such a bipartite graph admits a perfect matching (i.e. there is a permutation $\sigma \in S_n$ such that $i_r \sim \sigma(i)_c$ for any $i = 1, \dots, n$). In order to do so, we want to apply Hall's marriage theorem. Given a subset $S \subseteq \{1, \dots, n\}$, let $T$ be the subset defined as

$$T = \{t \in \{1, \dots, n\} : \ s_r \sim t_c \text{ for at least one } s \in S\}.$$

Exploiting the fact that the matrix $A$ is doubly-stochastic and the definition of $T$, we obtain

$$|S| = \sum_{s \in S} \sum_{j=1}^n A_{sj} = \sum_{s \in S} \sum_{t \in T} A_{st} \leq \sum_{i=1}^n \sum_{t \in T} A_{it} = \sum_{t \in T} \sum_{i=1}^n A_{it} = |T|.$$

Since we can choose $S$ arbitrarily, the inequality $|S| \leq |T|$ is exactly the hypothesis necessary to apply Hall's marriage theorem and deduce the existence of a perfect matching. Hence, by definition of perfect matching, there is a permutation $\sigma \in S_n$ such that $i_r \sim \sigma(i)_c$ for any $i = 1, \dots, n$. This last fact is equivalent to the desired statement. □

We prove the statement of the theorem by induction on the number of nonzero entries of the matrix $A$.

Since $A$ is doubly-stochastic, it is easy to see that it must have at least $n$ nonzero entries. Moreover, if it has exactly $n$ nonzero entries, then it must be already a permutation matrix.

Let us assume that the number of nonzero entries of $A$ is $k > n$. Let $P^\sigma$ be the permutation matrix induced by the permutation $\sigma$ whose existence is provided by the lemma. Let $\lambda > 0$ be the maximum value such that $\lambda P^\sigma \leq A$ (the inequality must be understood entry-wise). Notice that $\lambda < 1$ as if $\lambda \geq 1$, then $A$ would have exactly $n$ nonzero entries. Let $A' := \frac{1}{1-\lambda}(A - \lambda P^\sigma)$. Since $\lambda P^\sigma \leq A$, all entries of $A'$ are nonnegative. Moreover, thanks to the choice of $\lambda$, the matrix $A'$ has at most $k - 1$ nonzero entries. Finally, $A'$ is doubly-stochastic. By the inductive hypothesis, we are able to write $A'$ as a convex combination of permutation matrices

$$A' = \sum_{i \in I} \lambda_i P^{\sigma_i},$$

where $I$ is a finite set of indices, $\sum_{i \in I} \lambda_i = 1$ and $P^{\sigma_i}$ are permutation matrices (induced by the permutations $\sigma_i$). From the definition of $A'$, it follows

$$A = \lambda P^\sigma + \sum_{i \in I} \lambda_i (1 - \lambda) P^{\sigma_i},$$

that is the sought expression of $A$ as a convex combination of permutation matrices. □

**Exercise 1.6** (Discrete optimal transport)**.** Given two families $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$ of points in $\mathbb{R}^d$, let $\mu := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. Prove that, for *any* choice of a finite cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, there exists an optimal transport map from $\mu$ to $\nu$.
*Hint:* Use Exercise 1.5.

*Solution.* Let $\gamma \in \Gamma_{opt}(\mu, \nu)$ be an optimal transport plan. Let $A_\gamma$ be the $n \times n$ matrix with $A_{ij} = \gamma(\{x_i, y_j\})$. From the admissibility of $\gamma$ it follows that $nA$ is a doubly-stochastic matrix. Hence, applying Exercise 1.5, we can express $nA$ as a convex combination of permutation matrices

$$nA = \sum_{i \in I} \lambda_i P^{\sigma_i} ,$$

where $I$ is a finite set of indices, $\sum_{i\in I} \lambda_i = 1$ and $P^{\sigma_i}$ are permutation matrices (induced by the permutations $\sigma_i$).

Let us define the cost of an $n \times n$ matrix $B$ as

$$\mathcal{C}(B) = \sum_{i=1}^n \sum_{i=1}^n B_{ij} c(x_i, y_j) .$$

By definition, the cost $\mathcal{C}$ is linear and it holds

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \, \mathrm{d}\gamma(x, y) = \mathcal{C}(A) = \frac{1}{n} \sum_{i \in I} \lambda_i \mathcal{C}(P^{\sigma_i}) \geq \frac{1}{n} \min_{i \in I} \mathcal{C}(P^{\sigma_i}) .$$

Hence, there is a permutation $\sigma$ ($= \sigma_i$ for a certain $i \in I$) such that

$$\frac{1}{n} \sum_{i=1}^n c(x_i, y_{\sigma_i}) = \frac{1}{n} \mathcal{C}(P^\sigma) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \, \mathrm{d}\gamma(x, y)$$

and therefore any map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T(x_i) = y_{\sigma_i}$ is optimal. □

**Exercise 1.7** (Prokhorov's theorem)**.** Let $X$ be a complete and separable metric space, a family of probability measures $\mathcal{F} \subset \mathcal{P}(X)$ is sequentially relatively compact with respect to the narrow topology if $\mathcal{F}$ is equi-tight, i.e. for every $\varepsilon > 0$ there exists $K \subset X$ compact such that $\mu(X \setminus K) < \varepsilon$ for all $\mu \in \mathcal{F}$.
*Hint:*

1. Prove the result in the case of $X$ compact using Banach-Alaoglu-Bourbaki theorem.
2. Deduce the result in its full generality taking an exhaustion of $X$ by compact sets.

*Remark.* Actually Prokhorov's theorem states the "if and only if", namely $\mathcal{F}$ is sequentially relatively compact with respect to the narrow topology *if and only if* $\mathcal{F}$ is equi-tight. However, for our aims, we are interested only in the aforementioned implication.

*Solution.* For the solution, see the last part of the proof of [Bog07, Theorem 8.6.2] (the first part of the proof concerns the converse direction, that is more demanding). □

**Exercise 1.8.** Let $S : \mathbb{R}^d \to \mathbb{R}^d$ be the function $S(x) := -x$. Characterize the probabilities $\mu \in \mathcal{P}(\mathbb{R}^d)$ with compact support such that $S$ is an optimal transport map between $\mu$ and $S_\# \mu$ with respect to the quadratic cost.

*Solution.* Assume that $S$ is optimal from $\mu$ to $S_\#\mu$. Let $\gamma := (\mathbb{1} \times S)_\# \mu$ be the associated plan. There exists a subset $A \subseteq \mathbb{R}^d \times \mathbb{R}^d$ that is $c$-cyclically monotone (here $c$ is the quadratic cost) and $\gamma$ is supported on $A$. Since $\gamma$ is supported on graph$(S)$, we can assume without loss of generality that $A \subseteq$ graph$(S)$.

Take two points $x, y \in \mathbb{R}^d$ such that $(x, S(x)), (y, S(y)) \in A$. By $c$-cyclical monotonicity, it holds

$$\begin{gathered}\frac{1}{2}|x - S(x)|^2 + \frac{1}{2}|y - S(y)|^2 \leq \frac{1}{2}|x - S(y)|^2 + \frac{1}{2}|y - S(x)|^2 \\ \Updownarrow \\ 2|x|^2 + 2|y|^2 \leq |x + y|^2 \\ \Updownarrow \\ |x - y|^2 \leq 0 \\ \Updownarrow \\ x = y\,.\end{gathered}$$

Thus, $A$ contains only one point $(x_0, S(x_0))$ and therefore it must hold $\mu = \delta_{x_0}$.

On the other hand, if $\mu$ is $\delta_{x_0}$ for some $x_0 \in \mathbb{R}^d$, then it is straight-forward to check that $S$ is optimal from $\mu$ to $S_\#\mu = \delta_{S(x_0)}$. □

**Exercise 1.9.** Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two compactly supported probability measures invariant under rotations (that is $\mu(L(E)) = \mu(E)$ and $\nu(L(E)) = \nu(E)$ for any Borel set $E \in \mathscr{B}(\mathbb{R}^d)$ and any orthogonal transformation $L \in O(d)$). Show that, if $\mu \ll \mathcal{L}^d$, the optimal transport map from $\mu$ to $\nu$ with respect to the quadratic cost can be written as $x \to \lambda(|x|)\frac{x}{|x|}$ where $\lambda : [0, \infty) \to [0, \infty)$ is a suitable nondecreasing function.
*Hint:* The function $\lambda$ is the monotone transport map between two suitable 1-dimensional measures.

*Solution.* Let $\Phi : \mathbb{R}^d \to [0, \infty)$ be the norm $\Phi(x) \coloneqq |x|$. Let $\tilde{\mu} \coloneqq \Phi_\#\mu$ and $\tilde{\nu} \coloneqq \Phi_\#\nu$. From the identity $\Phi_\#\mathcal{L}^d = \omega_d t^{d-1}\mathcal{L}^1|_{[0,\,\infty)}$, where $\omega_d$ is the measure of the unit sphere in $\mathbb{R}^d$, it follows $\tilde{\mu} \ll \Phi_\#\mathcal{L}^d \ll \mathcal{L}^1$. Hence there is a convex function $\varphi : [0, \infty) \to \mathbb{R}$ such that $\lambda \coloneqq \varphi'$ is the optimal transport map from $\tilde{\mu}$ to $\tilde{\nu}$. Notice that $\varphi$ is nondecreasing.

Let $T(x) \coloneqq \lambda(|x|)\frac{x}{|x|}$. We will prove that $T$ is the optimal transport map from $\mu$ to $\nu$. Let us begin by checking that $T$ is the gradient of a convex function. Consider the function $\bar{\varphi}(x) : \mathbb{R}^d \to \mathbb{R}$ defined as $\bar{\varphi}(x) \coloneqq \varphi(|x|)$. It holds

$$\nabla\bar{\varphi}(x) = \varphi'(|x|)\frac{x}{|x|} = \lambda(|x|)\frac{x}{|x|} = T(x)\,.$$

Moreover, for any $a, b \geq 0$ with $a + b = 1$ and $x, y \in \mathbb{R}^d$ it holds

$$\bar{\varphi}(ax + by) = \varphi(|ax + by|) \leq \varphi(a|x| + b|y|) \leq a\varphi(|x|) + b\varphi(|y|) = a\bar{\varphi}(x) + b\bar{\varphi}(y)\,,$$

where we have used that $\varphi$ is convex and nondecreasing. Hence we have shown that $T$ is the gradient of a convex function.

It remains only to prove $T_\#\mu = \nu$, indeed the optimality of $T$ follows directly from the fact that it is the gradient of a convex function. Let us start by showing that $T_\#\mu$ is rotationally invariant and $\Phi_\#T_\#\mu = \Phi_\#\nu$. Given $L \in O(d)$, it holds

$$L_\#T_\#\mu = (L \circ T)_\#\mu = (T \circ L)_\#\mu = T_\#L_\#\mu = T_\#\mu\,,$$

where we have used that $\mu$ is rotationally invariant. Hence $T_\#\mu$ is rotationally invariant. From the identity $\Phi \circ T = \lambda \circ \Phi$, we deduce

$$\Phi_\#T_\#\mu = (\Phi \circ T)_\#\mu = (\lambda \circ \Phi)_\#\mu = \lambda_\#\tilde{\mu} = \tilde{\nu} = \Phi_\#\nu\,.$$

To conclude that $T_\# = \nu$ we apply the following lemma (to the probability measures $T_\#\mu$ and $\nu$).

**Lemma.** *Let $\mu^0, \mu^1 \in \mathcal{P}(\mathbb{R}^d)$ be two rotationally invariant probability measures. If $\Phi_\# \mu^0 = \Phi_\# \mu^1$ then $\mu^0 = \mu^1$.*

*Proof.* The statement is very intuitive, but its proof is rather technical. We give only a sketch.

For $i = 0, 1$, thanks to the disintegration theorem (see Exercise 1.16), there is a family of probability measures $(\mu^i_r)_{r \geq 0}$, with $\mu^i_r$ supported on $\{|x| = r\} \subseteq \mathbb{R}^d$, such that $\mu^i = \int_0^\infty \mu^i_r \, \mathrm{d}\Phi_\# \mu^i(r)$ (in the sense of the statement of Exercise 1.16) and the family $(\mu^i_r)_{r \geq 0}$ is unique up to a $\Phi_\# \mu^i$-negligible set.

Given a rotation $L \in O(d)$, it holds

$$\mu^i = L_\# \mu^i = \int_0^\infty L_\# \mu^i_r \, \mathrm{d}\Phi_\# \mu^i(r) \,,$$

therefore the uniqueness of the decomposition implies $L_\# \mu^i_r = \mu^i_r$ for $\Phi_\# \mu^i$-almost every $r$. Choosing $L$ in a countable family of rotations that is dense in $O(d)$, we deduce that, for $\Phi_\# \mu^i$-almost every $r$, the probability measure $\mu^i_r$ is rotationally invariant. Since it is well-known that the only rotationally invariant probability measure on the sphere is the appropriate rescaling of $\mathscr{H}^{d-1}|_{\mathbb{S}^{d-1}}$ (the uniform measure on the sphere), we deduce that for $\Phi_\# \mu^i$-almost every $r$ it holds

$$\mu^i_r = \frac{\mathscr{H}^{d-1}|_{\{|x|=r\}}}{\mathscr{H}^{d-1}(\{|x| = r\})} \,,$$

in particular $\mu^0_r = \mu^1_r$ and therefore $\mu^0 = \mu^1$. □

□

**Exercise 1.10** (Middle point)**.** Given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ be the infimum of the Kantorovich problem with respect to the quadratic cost

$$\mathcal{C}(\mu, \nu) \coloneqq \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{|x - y|^2}{2} \, \mathrm{d}\gamma(x, y) \,.$$

Let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures with compact support. A probability measure $\mu_{\frac{1}{2}}$ is a middle point of $\mu_0$ and $\mu_1$ if $\mathcal{C}(\mu_0, \mu_{\frac{1}{2}}) = \mathcal{C}(\mu_1, \mu_{\frac{1}{2}}) = \frac{1}{4}\mathcal{C}(\mu_0, \mu_1)$.

(a) If $\mu_0 = \delta_{p_0}$ and $\mu_1 = \delta_{p_1}$, show that the middle point is unique and $\mu_{\frac{1}{2}} = \delta_{\frac{p_1 + p_2}{2}}$.

(b) Prove that there is always at least one middle point.

(c) Find two probability measures $\mu_0, \mu_1$ such that they have more than one middle point.

(d) Show that if the optimal transport plan between $\mu_0$ and $\mu_1$ is unique, then there is a unique middle point.

(e) Prove that if $\mu_0, \mu_1 \ll \mathcal{L}^d$, then the middle point is unique and $\mu_{\frac{1}{2}} \ll \mathcal{L}^d$.

*Hint:* To solve (d), glue the plans from $\mu_0$ to $\mu_{\frac{1}{2}}$ and from $\mu_{\frac{1}{2}}$ to $\mu_1$ to obtain a plan from $\mu_0$ to $\mu_1$.

To solve (e), use the fact that the gradient of a strongly convex function is bi-Lipschitz (both the gradient and its inverse are Lipschitz-continuous).

*Solution.* Instead of attacking directly the various parts of the exercise, let us spend some time to understand a bit better the properties of middle points. Let us begin with the following useful fact.

**Lemma 1.11.** *For any $\rho \in \mathcal{P}(\mathbb{R}^d)$, it holds*

$$\mathcal{C}(\mu_0, \rho) + \mathcal{C}(\rho, \mu_1) \geq \frac{1}{2}\mathcal{C}(\mu_0, \mu_1)\,.$$

*Moreover, if equality holds, then there is an optimal plan $\gamma \in \Gamma_{opt}(\mu_0, \mu_1)$ such that $(\frac{x+z}{2})_\#\gamma = \rho$ (here, $x, z$ denote the first and second coordinate of $\mathbb{R}^d \times \mathbb{R}^d$).*

*Proof.* Let $\gamma_0 \in \Gamma_{opt}(\mu_0, \rho)$, $\gamma_1 \in \Gamma_{opt}(\rho, \mu_1)$ be two optimal plans from $\mu_0$ to $\rho$ and from $\rho$ to $\mu_1$. The gluing lemma (that is, ultimately, a consequence of the disintegration theorem) ensures the existence of $\tilde{\gamma} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ such that ($x, y, z$ denote the coordinates of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$)

$$(x, y)_\#\tilde{\gamma} = \gamma_0 \quad \text{and} \quad (y, z)_\#\tilde{\gamma} = \gamma_1\,.$$

Let $\gamma \coloneqq (x, z)_\#\tilde{\gamma}$. It follows directly from the properties of $\tilde{\gamma}$ that $\gamma$ is an admissible plan from $\mu_0$ to $\mu_1$. Therefore it holds

$$\begin{aligned}
\mathcal{C}(\mu_0, \mu_1) &\leq \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d} |x - z|^2 \,\mathrm{d}\gamma(x, z) = \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |x - z|^2 \,\mathrm{d}\tilde{\gamma}(x, y, z) \\
&\leq \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} 2|x - y|^2 + 2|y - z|^2 \,\mathrm{d}\tilde{\gamma}(x, y, z) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} |x - y|^2 \,\mathrm{d}\gamma_0(x, y) + \int_{\mathbb{R}^d\times\mathbb{R}^d} |y - z|^2 \,\mathrm{d}\gamma_1(y, z) = 2\left(\mathcal{C}(\mu_0, \rho) + \mathcal{C}(\rho, \mu_1)\right)\,.
\end{aligned}$$

If equality holds, then all the inequalities we have applied must be equalities. Hence (consider the first inequality of the chain) $\gamma$ has to be an optimal plan and (consider the second inequality) $|x - z|^2 = 2|x - y|^2 + 2|y - z|^2$ has to be true $\tilde{\gamma}$-almost everywhere. The latter identity implies that $\tilde{\gamma}$-almost everywhere it holds $y = \frac{x+z}{2}$, thus

$$\rho = y_\#\tilde{\gamma} = (\tfrac{x+z}{2})_\#\tilde{\gamma} = (\tfrac{x+z}{2})_\#\gamma\,.$$

□

Thanks to the lemma, we know that a measure $\mu_{\frac{1}{2}}$ is a middle-point if and only if $\mathcal{C}(\mu_0, \mu_{\frac{1}{2}}) \leq \frac{1}{4}\mathcal{C}(\mu_0, \mu_1)$ and $\mathcal{C}(\mu_1, \mu_{\frac{1}{2}}) \leq \frac{1}{4}\mathcal{C}(\mu_0, \mu_1)$. Indeed these two inequalities, together with $\mathcal{C}(\mu_0, \mu_{\frac{1}{2}}) + \mathcal{C}(\mu_{\frac{1}{2}}, \mu_1) \geq \frac{1}{2}\mathcal{C}(\mu_0, \mu_1)$ (this inequality follows from the triangle inequality for $W_2$, since $\mathcal{C} = W_2^2$) imply $\mathcal{C}(\mu_0, \mu_{\frac{1}{2}}) = \mathcal{C}(\mu_1, \mu_{\frac{1}{2}}) = \frac{1}{4}\mathcal{C}(\mu_0, \mu_1)$.

Let us consider an optimal plan $\gamma \in \Gamma_{opt}(\mu_0, \mu_1)$. We claim that $\mu_{\frac{1}{2}} \coloneqq (\frac{x+z}{2})_\#\gamma$ is a middle-point. Since $(x, \frac{x+z}{2})_\#\gamma$ is an admissible plan from $\mu_0$ to $\mu_{\frac{1}{2}}$, it holds

$$\mathcal{C}(\mu_0, \mu_{\frac{1}{2}}) \leq \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d} \left|x - \frac{x+z}{2}\right|^2 \mathrm{d}\gamma(x, z) = \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{4}|x - z|^2 \,\mathrm{d}\gamma(x, z) = \frac{1}{4}\mathcal{C}(\mu_0, \mu_1)\,.$$

The same reasoning yields also $\mathcal{C}(\mu_1, \mu_{\frac{1}{2}}) \leq \frac{1}{4}\mathcal{C}(\mu_0, \mu_1)$ and therefore, as observed above, $\mu_{\frac{1}{2}}$ is a middle-point.

Hence, given an optimal plan $\gamma$ we can produce the middle point $(\frac{x+z}{2})_\#\gamma$ and (thanks to the lemma above) given a middle point $\mu_{\frac{1}{2}}$ there is an optimal plan $\gamma$ such that $(\frac{x+z}{2})_\#\gamma = \mu_{\frac{1}{2}}$.[8] Now we are ready to tackle the statement of the exercise.

[8] During the exercise class it was erroneously deduced from these observations that the map between optimal plans and middle points is an isomorphism. In order to show it, it remains to check that two optimal plans $\gamma, \gamma' \in \Gamma_{opt}(\mu_0, \mu_1)$ such that $(\frac{x+z}{2})_\#\gamma = (\frac{x+z}{2})_\#\gamma'$ must be equal. Such a statement is true, but not straightforward, and luckily we do not need it in this exercise.

(a) Since there is a unique optimal plan (that is $\delta_{p_0} \times \delta_{p_1}$) there can be only one middle point and it must be $(\frac{x+z}{2})_\#(\delta_{p_0} \times \delta_{p_1}) = \delta_{\frac{p_0+p_1}{2}}$.

(b) The existence of a middle point follows directly from the existence of an optimal plan.

(c) Consider the two probability measures constructed in the solution of Exercise 1.4 (b). Since we are able to produce a middle point given a transport plan, it is straight-forward to check that the two mentioned probability measures are a good example.

(d) Let $\gamma \in \Gamma_{opt}(\mu_0, \mu_1)$ be the unique optimal transport plan. Then, thanks to our observations, $\mu_{\frac{1}{2}} = (\frac{x+z}{2})_\#\gamma$ has to be the unique middle point.

(e) Brenier's Theorem asserts that there is a unique optimal map between $\mu_0$ and $\mu_1$, that we denote $T : \mathbb{R}^d \to \mathbb{R}^d$. Thus, our observations imply that $\mu_{\frac{1}{2}} \coloneqq (\frac{x+T(x)}{2})_\#\mu_0$ is the unique middle point. Recall that, once again by Brenier's Theorem, it holds $T = \nabla\varphi$ where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a convex function. Hence

$$\frac{x + T(x)}{2} = \frac{1}{2}\nabla\left(\frac{|x|^2}{2} + \varphi\right)$$

and therefore, since the gradient of a strongly convex function is bi-Lipschitz, we deduce that $\frac{x+T(x)}{2}$ is a bi-Lipschitz map $\mathbb{R}^d \to \mathbb{R}^d$. In particular we conclude

$$\mu_{\frac{1}{2}} = \left(\frac{x + T(x)}{2}\right)_\# \mu_0 \ll \left(\frac{x + T(x)}{2}\right)_\# \mathcal{L}^d \ll \mathcal{L}^d\,.$$

□

**Exercise 1.12.** Consider $n$ red points $P_1, \dots, P_n$ and $n$ blue points $Q_1, \dots, Q_n$ on the plane. Assume that these $2n$ points are distinct and there are no 3 collinear points.

Show that it is possible to connect each red point to a distinct blue point with a segment in such a way that these segments do not intersect each other. Namely, there exists a permutation $\sigma \in S_n$ such that the segment $P_iQ_{\sigma(i)}$ does not intersect the segment $P_jQ_{\sigma(j)}$ for any $i \neq j$.

*Solution.* Consider an optimal map $T$ (that exists[9] thanks to Exercise 1.6) with respect to the linear cost between $\frac{1}{n}\sum_{i=1}^n \delta_{P_i}$ and $\frac{1}{n}\sum_{i=1}^n \delta_{Q_i}$. Let $\sigma \in S_n$ be the permutation induced by the map $T$, that is $T(P_i) = Q_{\sigma_i}$. We claim that $\sigma$ satisfies the requirements of the exercise. Let us assume by contradiction that this is not the case. Hence there exist $i \neq j$ such that $P_iQ_{\sigma(i)}$ intersects $P_jQ_{\sigma(j)}$. For notational simplicity, let us denote $P \coloneqq P_i$, $P' \coloneqq P_j$, $Q \coloneqq Q_i$, $Q' \coloneqq Q_j$. Without loss of generality (since we can always translate all the points) we can assume that the intersection of the two segments is the origin $O = (0, 0)$. Hence, there are $\lambda, \lambda' > 0$ such that $Q = -\lambda P, Q' = -\lambda' P'$.

As a consequence of the optimality of the map $T$ with respect to the linear cost, it holds

$$|P - Q| + |P' - Q'| \leq |P - Q'| + |P' - Q|$$

and therefore the triangle inequality (that is strict, as we are assuming that there are no three collinear points) implies

$$(1 + \lambda)|P| + (1 + \lambda')|P'| \leq |P + \lambda' P'| + |P' + \lambda P| < |P| + \lambda'|P'| + |P'| + \lambda|P|\,,$$

that is the sought contradiction. □

[9]For this exercise it would be sufficient to consider a map $T$ that is optimal in the family of maps. Showing the existence of such a *restricted optimizer* is easy in this setting as there are finitely many admissible maps.

**Exercise 1.13** (Easy disintegration)**.** Let $\mu \in \mathscr{M}(\mathbb{R}^2)$ be a finite measure on $\mathbb{R}^2$ that is absolutely continuous with respect to the Lebesgue measure with density $\rho : \mathbb{R}^2 \to \mathbb{R}$. Let $\nu \in \mathscr{M}(\mathbb{R})$ be the measure with density $\eta(x) \coloneqq \int_{\mathbb{R}} \rho(x,y)dy$. For any $x \in \mathbb{R}$ such that $\eta(x) \neq 0$, let $\mu_x$ be the measure with density $\rho_x(y) \coloneqq \frac{\rho(x,y)}{\eta(x)}$. If $\eta(x) = 0$, then simply set $\mu_x \coloneqq 0$.

Show that for any $g \in L^1(\mu)$ it holds

$$\int_{\mathbb{R}^2} g(x,y)d\mu(x,y) = \int_{\mathbb{R}} \int_{\mathbb{R}} g(x,y)d\mu_x(y)d\nu(x).$$

*Solution.* The desired identity is a direct consequence of Fubini's Theorem:

$$\begin{aligned}\int_{\mathbb{R}} \int_{\mathbb{R}} g(x,y)d\mu_x(y)d\nu(x) &= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} g(x,y)\rho_x(y)dy \right) \eta(x)dx = \int_{\mathbb{R}} \int_{\mathbb{R}} g(x,y)\rho(x,y)dydx \\ &= \int_{\mathbb{R}^2} g(x,y)d\mu(x,y)\,.\end{aligned}$$

□

For the remaining exercises concerning the disintegration theorem, we suggest the interested students to take a look at the proof of [AFP00, Theorem 2.28] (let us recall once again that this is absolutely facultative). The mentioned reference follows the sketch described in the hints to prove the theorem.

**Exercise 1.14** (Disintegration for product of compact spaces)**.** Let $X, Y$ be two compact spaces and let $\mu \in \mathscr{M}(X \times Y)$ be a finite measure on the product $X \times Y$. Let us denote $\nu \coloneqq (\pi_1)_{\#}\mu$ where $\pi_1 : X \times Y \to X$ is the projection on the first coordinate. Prove that there exists a family of probabilities $(\mu_x)_{x \in X} \subseteq \mathcal{P}(Y)$ such that:

(a) For any Borel set $E \in \mathscr{B}(Y)$ the map $x \mapsto \mu_x(E)$ is Borel.

(b) For any $g \in L^1(\mu)$ it holds

$$\int_{X \times Y} g(x,y)d\mu(x,y) = \int_X \int_Y g(x,y)d\mu_x(y)d\nu(x).$$

Moreover, if $(\mu_x)_{x \in X}$ and $(\tilde{\mu}_x)_{x \in X}$ are two families with the mentioned properties, then $\mu_x = \tilde{\mu}_x$ for $\nu$-almost every $x \in X$.
*Hint:*

1. Given $\psi \in C^0(Y)$, consider the map $A_\psi : L^1(X,\nu) \to \mathbb{R}$ given by the formula $A_\psi(\phi) \coloneqq \int_{X \times Y} \phi(x)\psi(y)d\mu(x,y)$. Prove that the said map is linear continuous and therefore $A_\psi \in L^\infty(X,\nu)$.

2. Fix a countable dense subset $S \subseteq C^0(Y)$. Prove that for $\nu$-almost every $x \in X$ the map $\mu_x : S \to \mathbb{R}$ given by $\mu_x(\psi) \coloneqq A_\psi(x)$ is linear continuous and therefore $\mu_x \in \mathcal{P}(Y)$. Show that the said family $(\mu_x)_{x \in X}$ satisfies *(a)*.

3. Show that *(b)* holds when $g \in L^1(X,\nu) \times S$. Show that this implies that it holds also when $g \in L^1(X,\nu) \times C^0(Y)$. Finally show that this implies *(b)* for any $g \in L^1(\mu)$.

**Exercise 1.15** (Disintegration for product of Polish spaces)**.** Show the statement of the previous exercise when $X$ and $Y$ are Polish spaces, i.e. they are complete and separable.
*Hint:* Use Prokhorov's theorem to find a suitable exhaustion in compact sets that allows to apply the previous exercise.

**Exercise 1.16** (Disintegration for fibers of a map)**.** Let $X, Y$ be two Polish spaces (complete and separable), let $f : Y \to X$ be a Borel map and let $\mu \in \mathscr{M}(Y)$ be a finite measure on $Y$. Let us denote $\nu := f_{\#}\mu$. Show that there exists a family of probabilities $(\mu_x)_{x\in X} \subseteq \mathcal{P}(Y)$ such that:

- For any Borel set $E \in \mathscr{B}(Y)$ the map $x \mapsto \mu_x(E)$ is Borel.
- For $\nu$-almost every $x \in X$ the measure $\mu_x$ is supported on the fiber $f^{-1}(x)$.
- For any $g \in L^1(\mu)$ it holds

$$\int_Y g(y)d\mu(y) = \int_X \int_{f^{-1}(x)} g(y)d\mu_x(y)d\nu(x).$$

Moreover, if $(\mu_x)_{x\in X}$ and $(\tilde{\mu}_x)_{x\in X}$ are two families with the mentioned properties, then $\mu_x = \tilde{\mu}_x$ for $\nu$-almost every $x \in X$.
*Hint:* Apply the previous exercise on the measure $(f \times \mathrm{id})_{\#}\mu$.

## References


[AFP00] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000, pp. xviii+434. ISBN: 0-19-850245-1.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Second. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008, pp. x+334. ISBN: 978-3-7643-8721-1.

[Bog07] V. I. Bogachev. *Measure theory. Vol. I, II*. Springer-Verlag, Berlin, 2007, Vol. I: xviii+500 pp., Vol. II: xiv+575. ISBN: 978-3-540-34513-8; 3-540-34513-2. DOI: 10.1007/978-3-540-34514-5. URL: https://doi.org/10.1007/978-3-540-34514-5.

[JKO98] Richard. Jordan, David. Kinderlehrer, and Felix. Otto. "The Variational Formulation of the Fokker–Planck Equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17. DOI: 10.1137/S0036141096303359. eprint: https://doi.org/10.1137/S0036141096303359. URL: https://doi.org/10.1137/S0036141096303359.

[Vil03] Cédric Villani. *Topics in Optimal Transportation*. Vol. 58. Graduate Studies in Mathematics. Amer. Math. Soc., 2003. Chap. 1.