An invitation to Optimal Transport, Wasserstein Distances and Gradient Flows

Alessio Figalli and Federico Glaudo

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Abstract

This monograph is intended as a self-contained introduction to optimal transport.

After a brief introductory section, we discuss the optimal transport problem and prove the existence of optimal transport maps, both for the quadratic cost and for more general costs on Euclidean spaces. We then introduce Wasserstein distances and gradient flows, and we show their connection via the so-called JKO scheme. Then, we develop the Otto's calculus and its application to Wasserstein gradient flows. To conclude, we briefly present a list of references concerning some of the several applications of this beautiful theory.

In the appendix we collect a series of exercises (with solutions) on optimal transport that may be useful to the reader in order to get more familiar with the topic. A guided proof, in the form of a series of exercises, of the disintegration theorem is contained in a short second appendix.

Contents

1 Introduction

In this introductory section we first give a brief historical review about optimal transport. Then we recall some basic definitions and facts from measure theory and Riemannian geometry, and finally we present three examples of (non-necessarily optimal) transport maps, with an application to the Euclidean isoperimetric inequality.

1.1 Historical Overview

1781 - Monge. In his celebrated work, Gaspard Monge introduced the concept of transport maps starting from the following practical question: "Assume one extract soil from the ground to build fortifications. How to transport the soil int he cheapest possible way?" To rigorously formulate this question, one needs to specify the transportation cost, namely how much one pays to move a unit of mass from a point x to a point y. In Monge's case, the ambient space was $\mathbb{R}^3$, and the cost was the Euclidean distance $c(x,y) := |x-y|$.

1940's - Kantorovich. After 150 years, Leonid Kantorovich revisited Monge's problem from a different view point. To explain this, consider N bakeries located at positions $(x_i)_{i=1,\dots,N}$, and M coffee shops located at $(y_j)_{j=1,\dots,M}$. Assume that the i-th bakery produces an amount $\alpha_i \geq 0$ of bread, and that the j-th coffee shop needs an amount $\beta_j \geq 0$. Also, assume that demand=request, and normalize them to be equal to 1: in other words $\sum_i \alpha_i = \sum_j \beta_j = 1$.

In Monge's formulation, the transport is deterministic: the mass located at x can be sent to a unique destination $T(x)$. Unfortunately this formulation is incompatible with the problem above, since one bakery may supply bread to multiple coffee shops, and one coffee shop may buy bread from multiple bakeries. For this reason Kantorovich introduced a new formulation: given $c(x_i, y_j)$ the cost to move one unit of mass from x_i to y_j, he looked for matrices $(\gamma_{ij})_{\substack{i=1,\dots,N\\ j=1,\dots,M}}$ such that:

(i) $\gamma_{ij} \geq 0$ (the amount of bread going from x_i to y_j is a nonnegative quantity);

(ii) $\forall\, i\colon \alpha_i = \sum_{j=1}^{M} \gamma_{ij}$ (the total amount of bread sent to the different coffee shops is equal to the production);

(iii) $\forall\, j\colon \beta_j = \sum_{i=1}^{N} \gamma_{ij}$ (the total amount of bread bought from the different bakeries is equal to the demand);

(iv) γ_{ij} minimize the cost $\sum_{i,j=1}^{d} \gamma_{ij} c(x_i, y_j)$ (the total transportation cost is minimized).

It is interesting to observe that constraint (i) is convex, constraints (ii) and (iii) are linear, and the objective function in (iv) is also linear (all with respect to γ_{ij}). In other words, Kantorovich's formulation corresponds to minimizing a linear function with convex/linear constraints.

Applications. Optimal transport has been a topic of high interest in the last 30 years due to its connection to several areas of mathematics. The properties and the applications of optimal transport depend heavily of the choice of the cost function $c(x,y)$, representing the cost of moving a unit of mass from x to y. Let us mention some important choices:

- $c(x,y) = |x-y|^2$ in $\mathbb{R}^d$ – It is connected to: Euler equations; Isoperimetric and Sobolev inequalities; evolution PDEs such as $\partial_t u = \Delta u$, $\partial_t u = \Delta(u^m)$, and $\partial_t u = \operatorname{div}(\nabla W * u\, u)$.

- $c(x,y) = |x-y|$ in $\mathbb{R}^d$ – Appears in probability and kinetic theory.

- $c(x,y) = d(x,y)^2$ on a Riemannian manifold, with $d(\,\cdot\,,\,\cdot\,)$ denoting the Riemannian distance – Has connections and applications to the study of Ricci curvature.

In this monograph we mostly focus on the Euclidean quadratic cost $|x-y|^2$, and we shall give references for further applications in Section 5.

1.2 Basics of measure theory

For simplicity, throughout this book we shall always work on locally compact, separable and complete metric spaces, that will be usually denoted by X (the space where the source measure lives) and Y (the space where the target measure lives). These assumptions are not optimal but simplify some of the proofs in the next chapter (see also Remark 2.1.1). Still, the reader who is not interested in such a level of generality can always think that $X = Y = \mathbb{R}^d$.

Remark 1.2.1. All measures under consideration are Borel measures, and all maps are Borel (i.e., if $S\colon X \to Y$, then $S^{-1}(A)$ is Borel for all $A \subset Y$ Borel). The set of probability measures over a space X will be denoted by $\mathcal{P}(X)$, and the class of Borel-measurable sets by $\mathcal{B}(X)$. Also, $\mathbb{1}_A$ denotes the indicator function of a set:

$$\mathbb{1}_A(x) := \begin{cases} 1 & \text{if } x \in A\,, \\ 0 & \text{if } x \notin A\,. \end{cases}$$

Definition 1.2.2. Take a map $T\colon X \to Y$ and a probability measure $\mu \in \mathcal{P}(X)$. We define the *image measure* (or *push-forward measure*) $T_\#\mu \in \mathcal{P}(Y)$ as

$$(T_\#\mu)(A) := \mu(T^{-1}(A)) \quad \text{for any } A \in \mathcal{B}(Y).$$

Lemma 1.2.3. *$T_\#\mu$ is a probability measure on Y.*

Proof. The proof consists in checking that $T_\#\mu$ is nonnegative, has total mass 1, gives no mass to the empty set, and is σ-additive on disjoint sets.

1. $(T_\#\mu)(\emptyset) = \mu(T^{-1}(\emptyset)) = \mu(\emptyset) = 0$;
2. $(T_\#\mu)(Y) = \mu(T^{-1}(Y)) = \mu(X) = 1$;
3. $(T_\#\mu)(A) = \mu(T^{-1}(A)) \geq 0$ for all $A \in \mathcal{B}(X)$;
4. 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 was not 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 fact, using that μ is a measure (and thus σ-additive on disjoint sets) we get

$$T_\#\mu\Big(\bigcup_{i\in I} A_i\Big) = \mu\Big(T^{-1}\big(\bigcup_{i\in I} A_i\big)\Big) = \mu\Big(\bigcup_{i\in I} T^{-1}(A_i)\Big) = \sum_{i\in I} \mu(T^{-1}(A_i)) = \sum_{i\in I} T_\#\mu(A_i)\,.$$

□

Remark 1.2.4. One might also be tempted to define the "pull-back measure" $S^\#\nu(E) := \nu(S(E))$ for $S\colon X \to Y$ and $\nu \in \mathcal{P}(Y)$. However, this construction does not work in general. Indeed, since the image of two disjoint sets might coincide (consider for instance the case when S is a constant map), $S^\#\nu$ may not be additive on disjoint sets.

Lemma 1.2.5. *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 $A \in \mathcal{B}(Y)$. We now focus on the other implication.

For any Borel subset $A \subset Y$, it holds

$$\int_Y \mathbb{1}_A \, d\nu = \nu(A) = \mu(T^{-1}(A)) = \int_X \mathbb{1}_{T^{-1}(A)} \, d\mu = \int_X \mathbb{1}_A \circ T \, d\mu \, .$$

Thus, by linearity of the integral, we immediately deduce

$$\int_Y \varphi \, d\nu = \int_X \varphi \circ T \, d\mu$$

for any simple function $\varphi : Y \to \mathbb{R}$, i.e., for any φ of the form $\sum_{i \in I} \lambda_i \mathbb{1}_{A_i}$ where I is a finite set, $(A_i)_{i \in I}$ are Borel subsets, and $(\lambda_i)_{i \in I}$ are real values.

In order to deduce the desired result, fix a bounded Borel function $\varphi : Y \to \mathbb{R}$. Since any bounded Borel function can be approximated uniformly by simple functions[1], there is a sequence of simple functions $(\varphi_k)_{k \in \mathbb{N}}$ such that $\|\varphi_k - \varphi\|_\infty \to 0$ as $k \to \infty$. Therefore we have

$$\int_Y \varphi \, d\nu = \lim_{k \to \infty} \int_Y \varphi_k \, d\nu = \lim_{k \to \infty} \int_X \varphi_k \circ T \, d\mu = \int_X \varphi \, d\mu \, ,$$

that is the desired identity. □

An immediate consequence of the previous lemma is the following:

Corollary 1.2.6. *For any function $\varphi : Y \to \mathbb{R}$ Borel and bounded it holds*

$$\int_Y \varphi \; d(T_{\#}\mu) = \int_X \varphi \circ T \; d\mu.$$

Then next lemma shows the relation between composition and push-forward.

Lemma 1.2.7. *Let $T \colon X \to Y$ and $S \colon Y \to Z$ be measurable, then*

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

Proof. Thanks to Corollary 1.2.6, for any $\varphi : Z \to \mathbb{R}$ Borel and bounded we have

$$\begin{aligned} \int_Z \varphi \, d(S \circ T)_{\#}\mu &= \int_X \varphi \circ (S \circ T) \, d\mu = \int_X (\varphi \circ S) \circ T \, d\mu \\ &= \int_Y \varphi \circ S \, dT_{\#}\mu = \int_Z \varphi \, dS_{\#}(T_{\#}\mu). \end{aligned}$$

The result follows from Lemma 1.2.5. □

[1] To prove this, given $\varphi : Y \to \mathbb{R}$ a bounded Borel function, fix $\epsilon > 0$ and for any $i \in \mathbb{Z}$ consider the set $A_i := \{\epsilon i \le \varphi < \epsilon(i+1)\}$. Then, define $\varphi_\epsilon := \sum_{i \in \mathbb{Z}} \epsilon i \mathbb{1}_{A_i}$. Since φ is bounded we have $A_i = \emptyset$ for $|i| \gg 1$, hence φ_ϵ is a simple function. Also

$$\|\varphi - \varphi_\epsilon\|_{L^\infty} = \max_{i \in \mathbb{Z}} \|\varphi - \varphi_\epsilon\|_{L^\infty(A_i)} \le \epsilon.$$

1.3 Basics of Riemannian geometry

Even though we are not going to work with Riemannian manifolds, some of the results we present (namely Arnold's Theorem, geodesics in the Wasserstein space, and the differential structure of the Wasserstein space) are heavily inspired by classical concepts in Riemannian geometry. Hence, we provide a very short introduction to the subject, with an emphasis on those facts and structures that may help the reader to fully appreciate the content of this book.

First, for embedded submanifolds, we recall the definitions of tangent space, Riemannian distance, (minimizing) geodesic, and gradient. Then we briefly explain how these definitions can be generalized to the (more abstract) case of a (non necessarily embedded) Riemannian manifold.

Our presentation of the subject is quick and superficial, but should be sufficient to understand the related topics in this book. This material, and much more, may be found in any introductory text on Riemannian Geometry (see, for example, [Cha93; GHL04; Lee97; Pet06]). The reader who has already some experience with the subject may skip this section.

Embedded Submanifolds. Let M be a compact d-dimensional smooth manifold embedded in $\mathbb{R}^D$. We are going to show how the Euclidean scalar product of the ambient $\mathbb{R}^D$ induces a distance –the Riemannian distance– on M, and how this gives rise to a number of related concepts (gradients, minimizing geodesics, and geodesics).

In what follows, we implicitly assume that all curves are C^1.

Let us begin with the definition of tangent space. Notice that, for its definition, we are not going to use the Euclidean scalar product of the ambient.

Definition 1.3.1 (Tangent space)**.** Given a point $p \in M$, the tangent space $T_pM \subset \mathbb{R}^D$ of M at p is defined as

$$T_pM \coloneqq \{\dot\gamma(0) \mid \gamma : (-1,1) \to M,\, \gamma(0) = p\}\,.$$

Intuitively, the tangent space contains all the directions tangent to M at p. One can show that T_pM is a d-dimensional subspace of $\mathbb{R}^D$.

We now give the definition of gradient of a function, which is a convenient representation of its differential.

Definition 1.3.2 (Gradient)**.** Let $F : M \to \mathbb{R}$ be a smooth function. Its gradient $\nabla F : M \to \mathbb{R}^D$ is defined as the unique tangent vector field on M, that is $\nabla F(x) \in T_xM$ for all $x \in M$, such that the following holds: for any curve $\gamma : (-1,1) \to M$,

$$\langle \nabla F(\gamma(0)), \dot\gamma(0)\rangle = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} F(\gamma(t))\,.$$

For the definition of the gradient we are using that the Euclidean scalar product endows the tangent spaces of a scalar product (i.e., the restriction of the ambient scalar product).

Given a curve $\gamma : [a,b] \to M$, its length is given by the formula

$$\int_a^b |\dot\gamma(t)|\, dt\,.$$

Notice that the length of a curve is invariant under reparametrization. Notice also that, to define the length of a curve, we need to compute the Euclidean norm only of vectors tangent to M.

Once one knows how to measure the length of a curve, the following definition of (Riemannian) distance is fairly natural.

Definition 1.3.3 (Riemannian distance)**.** Given two points $x, y \in M$, their Riemannian distance $d_M(x, y)$ is defined as

$$d_M(x,y) := \inf \left\{ \int_a^b |\dot\gamma(t)|\, dt \mid \gamma : [a,b] \to M,\, \gamma(a) = x,\, \gamma(b) = y \right\}.$$

The Riemannian distance is indeed a distance on M, that is, it satisfies the triangle inequality (besides $d_M(x, y) = d_M(y, x)$, and $d_M(x, y) = 0$ if and only if $x = y$).

Since any curve can be reparametrized to have constant speed, one can show that an equivalent definition of the Riemannian distance is given by

$$d_M(x,y)^2 = \inf \left\{ \int_0^1 |\dot\gamma(t)|^2\, dt \mid \gamma : [0,1] \to M,\, \gamma(0) = x,\, \gamma(1) = y \right\}. \tag{1.2}$$

It turns out that there is always a (non necessarily unique) curve achieving the infimum in the definition of the Riemannian distance (this follows from the compactness of M or, more in general, from its completeness).

Definition 1.3.4 (Minimizing geodesic)**.** A curve $\gamma : [a, b] \to M$ with constant speed (i.e., $|\dot\gamma|$ is constant) such that $\gamma(a) = x, \gamma(b) = y$, and whose length is equal to $d_M(x, y)$, is called a *minimizing geodesic*.

The restriction of a minimizing geodesic on a smaller interval is still a minimizing geodesic. Moreover any minimizing geodesic is smooth.

One may think of minimizing geodesics as "straight lines in a curved space". Indeed, since a minimizing geodesic has constant speed and achieves the minimum also in (1.2), it can be proven (with a variational argument, as a consequence of the minimality) that

$$\ddot\gamma(t) \perp T_{\gamma(t)}M \tag{1.3}$$

for all $t \in [0, 1]$. In other words, apart from the distortion induced by M, minimizing geodesics go "as straight as possible".

Definition 1.3.5 (Geodesic)**.** A (non necessarily minimizing) *geodesic* is a curve $\gamma : [a, b] \to M$ that satisfies (1.3).

It can be readily checked that a geodesic has constant speed; indeed

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

where we have used that $\ddot\gamma \perp T_\gamma M \ni \dot\gamma$.

Moreover, any geodesic is locally minimizing. More precisely, if $\gamma : [a, b] \to M$ satisfies (1.3), then for any $t_0 \in (a, b)$ there is $\varepsilon > 0$ such that γ restricted on $[t_0 - \varepsilon, t_0 + \varepsilon]$ is a minimizing geodesic.

Abstract Riemannian Manifolds. In the previous paragraph we described how a submanifold of $\mathbb{R}^D$ inherits a number of structures (tangent space, gradient, distance, geodesics) from the ambient. Let us briefly explain what is necessary for an abstract manifold to have such structures.

Given a compact d-dimensional smooth manifold M, there is an intrinsic definition of tangent space T_pM (as an appropriate quotient of the curves through p, where two curves are identified if "they have the same derivative at p"). To proceed further and talk about gradients, lengths, etc., we need to endow our manifold M of an additional structure, that is, a Riemannian metric. A *Riemannian metric* is a (symmetric and positive definite) scalar product $g_x : T_xM \times T_xM \to \mathbb{R}$ defined on each tangent space, that varies continuously with respect to $x \in M$. If M is

endowed with a Riemannian metric $g = (g_x)_{x\in M}$, we say that (M, g) is a *Riemannian manifold*. On a Riemannian manifold, all the definitions given previously (gradient, length, Riemannian distance, and minimizing geodesic) make perfect sense (for example, the length of a curve is $\int_a^b g_\gamma(\dot\gamma,\dot\gamma)^{\frac12}$), and all the facts we have stated remain true.

It is more delicate to generalize (1.3) to this more abstract setting, and thus to define what a (non necessarily minimizing) geodesic is. We prefer not to delve into this topic, as it goes beyond the basic understanding of Riemannian geometry that is necessary to appreciate the rest of this book.

1.4 Transport maps

Definition 1.4.1. 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 μ to ν if $T_\#\mu = \nu$.

Remark 1.4.2. Given μ and ν, the set $\{T \mid T_\#\mu = \nu\}$ may be empty. For instance, given $\mu = \delta_{x_0}$ with $x_0 \in X$ and a map $T\colon X \to Y$, we have

$$\int_Y \varphi(y)\, d(T_\#\mu)(y) = \int_Y \varphi\circ T(x)\, d\mu(x) = \varphi(T(x_0)) \quad \forall\, \varphi : Y \to \mathbb{R} \qquad \Rightarrow \qquad T_\#\mu = \delta_{T(x_0)}.$$

Hence, unless ν is a Dirac delta, for any map T we have $T_\#\mu \neq \nu$ and the set $\{T \mid T_\#\mu = \nu\}$ is empty.

Definition 1.4.3. We call $\gamma \in \mathcal{P}(X \times Y)$ a *coupling*[2] of μ and ν 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

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

and

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

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

Remark 1.4.4. Given μ and ν, the set $\Gamma(\mu, \nu)$ is always nonempty. Indeed the product measure $\gamma = \mu \otimes \nu$ (defined by $\int \phi(x, y)\, d\gamma(x, y) = \int \phi(x, y)\, d\mu(x)\, d\nu(y)$ for every $\phi : X \times Y \to \mathbb{R}$) is a coupling:

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

[2]The terminology "coupling" is common in probability. However, in optimal transport theory, one often uses the expression *transport plan* in place of coupling.

Remark 1.4.5 (Transport map vs. Coupling)**.** Let $T\colon X \to Y$ satisfy $T_\#\mu = \nu$. Consider the map $\mathrm{Id} \times T\colon X \to X \times Y$, i.e., $x \mapsto (x, T(x))$, and define

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

We claim that $\gamma_T \in \Gamma(\mu, \nu)$. Indeed, recalling Lemma 1.2.7, we have

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

This proves that any transport map T induces a coupling γ_T.

1.4.1 Examples of transport maps

We now discuss three examples of transport maps: the measurable transport, the one-dimensional monotone rearrangement, and the Knothe's map.

Measurable transport. The following result can be found in [BBP16, Theorem 11.25]:

Theorem 1.4.6. *Let $\mu \in \mathcal{P}(X)$ be a probability measure such that μ has no atoms (i.e., $\mu(\{x\}) = 0$ for any $x \in X$). Then there exists $T_\mu\colon X \to \mathbb{R}$ such that T_μ is injective μ-a.e. and*

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

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

In other words, given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ without atoms, this abstract theorem tells us that we can always transport one onto the other by simply considering $T_\nu^{-1} \circ T_\mu$ (this is a transport map from μ to ν) or $T_\mu^{-1} \circ T_\nu = (T_\nu^{-1} \circ T_\mu)^{-1}$ (this is a transport map from ν to μ). Unfortunately these maps have no structure, so they have little interest in concrete applications in analysis/geometry. Indeed, as we shall see in this book, a very important feature of optimal transport maps are their structural properties (for instance, optimal maps for the quadratic cost are gradients of convex functions, see Theorem 2.5.9).

Monotone rearrangement. Given $\mu, \nu \in \mathcal{P}(\mathbb{R})$, set

$$F(x) \coloneqq \int_{-\infty}^{x} d\mu(t), \qquad G(y) \coloneqq \int_{-\infty}^{y} d\nu(t).$$

Note that these maps are not well defined at points where measures have atoms, since one needs to decide whether the mass of the atom is included in the value of the integral or not. We adopt the convention that the mass of the atoms are included, so that both maps are continuous from the right. More precisely, we set

$$F(x) \coloneqq \lim_{\varepsilon \to 0^+} \int_{-\infty}^{x+\varepsilon} d\mu(t) = \mu\big((-\infty, x]\big), \qquad G(y) \coloneqq \lim_{\varepsilon \to 0^+} \int_{-\infty}^{y+\varepsilon} d\nu(t) = \nu\big((-\infty, y]\big).$$

Note that F and G are nondecreasing. If G was strictly increasing, it would be injective and we could naturally consider its inverse G^{-1}. However, G may be constant in some regions, so we need to define a "pseudo-inverse" as follows:

$$G^{-1}(y) \coloneqq \inf\{t \in \mathbb{R} \mid G(t) > y\}.$$

Note that also G^{-1} is continuous from the right.

With these definitions, we define the nondecreasing map $T \coloneqq G^{-1} \circ F\colon \mathbb{R} \to \mathbb{R}$ and we want to prove that it transports μ to ν. Of course this cannot be true in general, since the set of transport maps may be empty (recall Remark 1.4.2). The following result shows that this is the case if μ has no atoms:

Theorem 1.4.7. *If μ has no atoms, then $T_\#\mu = \nu$.*

To prove this theorem, we need some preliminary results.

Lemma 1.4.8. *If μ has no atoms, then for all $t \in [0,1]$ we have*

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

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

Also, since μ has no atoms,

$$|F(t_k) - F(t)| = \left| \int_t^{t_k} d\mu \right| \xrightarrow[t_k \to t]{} 0 \qquad \forall\, t \in \mathbb{R},$$

thus $F \in C^0(\mathbb{R}, \mathbb{R})$. Since $F(t) \to 0$ as $t \to -\infty$ and $F(t) \to 1$ as $t \to +\infty$, by the intermediate value theorem it follows that F is surjective on $(0,1)$.

Given $t \in (0,1)$, consider the largest value $x \in \mathbb{R}$ such that $F(x) = t$ (this point exists by the continuity of F). With this choice of x, we have

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

as desired. □

Corollary 1.4.9. *If μ has no atoms, then for all $t \in [0,1]$ we have*

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

Proof. We apply Lemma 1.4.8 to the intervals $[0,t]$ and $[0,t-\varepsilon]$ with $\varepsilon > 0$:

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

□

Proof of Theorem 1.4.7. We split the proof into five steps.

1. Let $A = (-\infty, a]$ with $a \in \mathbb{R}$. Applying Corollary 1.4.9, we have

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

2. Let $A = (a, b] = (-\infty, b] \setminus (-\infty, a]$. Applying Step 1, we have

$$T_\#\mu(A) = T_\#\mu((-\infty, b]) - T_\#\mu((-\infty, a]) = \nu((-\infty, b]) - \nu((-\infty, a]) = \nu(A).$$

3. Let $A = (a, b)$, and consider $A_\varepsilon := (a, b - \varepsilon]$. Thanks to Step 2 and monotone convergence, we have

$$\nu(A) \nwarrow \nu(A_\varepsilon) = T_\#\mu(A_\varepsilon) \nearrow T_\#\mu(A) \qquad \text{as } \varepsilon \to 0^+.$$

4. Let $A \subset \mathbb{R}$ be an open set. We can write $A = \bigcup_{i \in I} (a_i, b_i)$ with $\big((a_i, b_i)\big)_{i \in I}$ disjoint and countable. Thus, by Step 3, we get

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

5. Since open sets are generators of the Borel σ-algebra, Step 4 proves that $T_\#\mu = \nu$.

□

Knothe's map. We are going to build a transport map, known as the Knothe's map [Kno57], that is a multidimensional generalization of monotone rearrangement. First, we need to state the disintegration theorem (for a proof of this result, see Appendix B at the end of these lectures).

Theorem 1.4.10 (Disintegration Theorem)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^2)$ and set $\mu_1 := (\pi_1)_{\#}\mu \in \mathcal{P}(\mathbb{R})$, 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 $(\mu_{x_1})_{x_1 \in \mathbb{R}} \subset \mathcal{P}(\mathbb{R})$ 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

$$\int_{\mathbb{R}^2} \varphi(x_1, x_2)\, d\mu(x_1, x_2) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \varphi(x_1, x_2)\, d\mu_{x_1}(x_2) \right) d\mu_1(x_1).$$

Moreover, the measures μ_{x_1} are unique μ_1-a.e.

Example 1.4.11. Let $\mu = f(x_1, x_2)\, dx_1\, dx_2$ with $\int_{\mathbb{R}^2} f\, dx_1\, dx_2 = 1$, and set

$$\mu_1 := (\pi_1)_{\#}\mu, \qquad F_1(x_1) := \int_{\mathbb{R}} f(x_1, x_2)\, dx_2.$$

We claim that $\mu_1 = F_1 dx_1$. Indeed, given any test function $\varphi : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned} \int_{\mathbb{R}} \varphi(x_1)\, d\mu_1(x_1) = \int_{\mathbb{R}^2} \varphi(x_1)\, d\mu(x_1, x_2) &= \int_{\mathbb{R}^2} \varphi(x_1) f(x_1, x_2)\, dx_1, dx_2 \\ \underbrace{=}_{\text{Fubini}} \int_{\mathbb{R}} \varphi(x_1) \left(\int_{\mathbb{R}} f(x_1, x_2)\, dx_2 \right) dx_1 &= \int_{\mathbb{R}} \varphi(x_1) F_1(x_1)\, dx_1, \end{aligned}$$

as desired.

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

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

Hence, by uniqueness of the disintegration, we deduce that

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

Note that μ_{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 1.4.12 (An absolutely continuous measure lives where its density is positive)**.** Note that $F_1 > 0$ μ_1-a.e. Indeed

$$\int_{\{F_1 = 0\}} d\mu_1 = \int_{\{F_1 = 0\}} F_1\, dx_1 = \int_{\{F_1 = 0\}} 0\, dx_1 = 0.$$

Construction of Knothe's map. Take two *absolutely continuous* measures on $\mathbb{R}^2$, namely

$$\mu(x_1,x_2) = f(x_1,x_2)\,dx_1dx_2 = \frac{f(x_1,x_2)}{F_1(x_1)}\,dx_2 \otimes F_1(x_1)\,dx_1,$$
$$\nu(y_1,y_2) = g(y_1,y_2)dy_1dy_2 = \frac{g(y_1,y_2)}{G_1(y_1)}dy_2 \otimes G_1(y_1)dy_1,$$

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 1.4.7, the monotone rearrangement provides us with a map $T_1\colon \mathbb{R}\to\mathbb{R}$ such that $T_{1\#}(F_1dx_1) = G_1dy_1$. Then, for F_1dx_1-a.e. $x_1\in\mathbb{R}$, we consider the monotone rearrangement $T_2(x_1,\cdot)\colon\mathbb{R}\to\mathbb{R}$ such that

$$T_2(x_1,\cdot)_\#\left(\frac{f(x_1,\cdot)}{F_1(x_1)}\,dx_2\right) = \frac{g(T_1(x_1),\cdot)}{G_1(T_1(x_1))}\,dy_2. \tag{1.4}$$

In other words, for each fixed x_1, $F(x_1,\cdot)$ is a map that sends the disintegration of μ at the point x_1 onto the disintegration of ν and the point $T(x_1)$.

Theorem 1.4.13. *The Knothe's map $T(x_1,x_2) \coloneqq (T_1(x_1), T_2(x_1,x_2))$ transports μ to ν.*

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

$$\begin{aligned}
\int_{\mathbb{R}^2}\varphi(y_1,y_2)g(y_1,y_2)\,dy_1dy_2 &= \int_{\mathbb{R}}\underbrace{\left(\int_{\mathbb{R}}\varphi(y_1,y_2)\frac{g(y_1,y_2)}{G_1(y_1)}dy_2\right)}_{\Psi(y_1)}G(y_1)\,dy_1\\
&\overset{(T_1)_\#(F_1dx_1)=G_1dy_1}{=} \int_{\mathbb{R}}\Psi(T_1(x_1))F_1(x_1)\,dx_1\\
&= \int_{\mathbb{R}}\left(\int_{\mathbb{R}}\varphi(T_1(x_1),y_2)\frac{g(T_1(x_1),y_2)}{G_1(T_1(x_1))}dy_2\right)F_1(x_1)\,dx_1\\
&\overset{(1.4)}{=} \int_{\mathbb{R}}\left(\int_{\mathbb{R}}\varphi(T_1(x_1),T_2(x_1,x_2))\frac{f(x_1,x_2)}{F_1(x_1)}\,dx_2\right)F_1(x_1)\,dx_1\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\varphi(T_1(x_1),T_2(x_1,x_2))f(x_1,x_2)\,dx_2\,dx_1\\
&= \int_{\mathbb{R}^2}(\varphi\circ T)(x_1,x_2)\,d\mu(x_1,x_2).
\end{aligned}$$

□

Remark 1.4.14. 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)

$$\nabla T = \begin{pmatrix} \partial_1T_1\geq 0 & * \\ 0 & \partial_2T_2\geq 0\end{pmatrix}$$

One can use the previous construction of the Knothe's map in $\mathbb{R}^2$ and iterate it to obtain a Knothe's map on $\mathbb{R}^d$. Let

$$\mu(x_1,\dots,x_d) = f(x_1,\dots,x_d)\,dx_1\cdots dx_d, \qquad \nu(y_1,\dots,y_d) = g(y_1,\dots,y_d)\,dy_1\cdots dy_d$$

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 $F_1(x_1) = \int f\,dx_2\dots dx_d$ and $G_1(y_1) = \int g\,dy_2\dots dy_d$.

Also, the analogues of Theorem 1.4.10 and Example 1.4.11 in $\mathbb{R}^d$, yield probability measures on $\mathbb{R}^{d-1}$ given by

$$\mu_{x_1}(x_2,\dots,x_d) = \frac{f(x_1,x_2,\dots,x_d)}{F_1(x_1)}dx_2\cdots dx_d$$

and

$$\nu_{y_1}(y_2,\dots,y_d) = \frac{g(y_1,y_2,\dots,y_d)}{G_1(y_1)}dy_2\cdots dy_d,$$

such that $\mu = \mu_{x_1}\otimes F_1 dx_1$ and $\nu = \nu_{y_1}\otimes G_1 dy_1$.

By induction on the dimension, there exists a Knothe map $T_{x_1}:\mathbb{R}^{d-1}\to\mathbb{R}^{d-1}$ sending μ_{x_1} onto $\nu_{T_1(x_1)}$, and then we obtain a Knothe's map in $\mathbb{R}^d$ as

$$T(x_1,\dots,x_d) \coloneqq (T_1(x_1), T_{x_1}(x_2,\dots,x_d)).$$

Remark 1.4.15. Suppose again that the map T is smooth. Then

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

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

Remark 1.4.16. Although we call it *the* Knothe's map, the map itself is by no means unique. Indeed, by fixing a basis in $\mathbb{R}^d$ but changing the order of integration, one obtains a different Knothe's map. Even more, changing the basis of $\mathbb{R}^d$ yields in general a different map.

1.5 An application to isoperimetric inequalities

The following is the classical (sharp) isoperimetric inequality in $\mathbb{R}^d$.

Theorem 1.5.1. *Let $E\subset\mathbb{R}^d$ be a bounded set with smooth boundary. Then*

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

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 1.5.2. *Let T be a Knothe's map from μ to ν, and assume it to be smooth*[3]*. Then:*

1. *For any $x\in E$, it holds $|T(x)|\leq 1$.*
2. *$\det\nabla T = \frac{|B_1|}{|E|}$ in E.*
3. *$\operatorname{div} T \geq d\,(\det\nabla T)^{\frac{1}{d}}$.*

Proof. We prove the three properties.

[3]The smoothness assumption can be dropped with some fine analytic arguments. To obtain a rigorous proof one can also work with the optimal transport map (instead of the Knothe's map) and use the theory of functions with bounded variation, as done in [FMP10].

1. If $x \in E$, then $T(x) \in B_1$ and thus $|T(x)| \le 1$.

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

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

On the other hand, by the change of variable formulas, setting $y = T(x)$ we have $dy = |\det \nabla T| dx$, therefore

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

Furthermore, since ∇T is upper triangular and its diagonal elements are nonnegative (see Remark 1.4.15), it follows that $\det \nabla T \ge 0$, hence

$$\int_{T^{-1}(A)} \frac{dx}{|E|} = \nu(A) = \int_{T^{-1}(A)} \frac{1}{|B_1|} \det \nabla T(x) \, dx.$$

Since $A \subset B_1$ is arbitrary, we obtain

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

3. Note that, since the matrix ∇T is upper-triangular (see Remark 1.4.15), its determinant is given by the product of its diagonal elements. Hence

$$\operatorname{div} T(x) = \sum_{i=1}^d \partial_i T_i(x) = d \left(\frac{1}{d} \sum_{i=1}^d \partial_i T_i(x) \right) \ge d \left(\prod_{i=1}^d \partial_i T_i(x) \right)^{\frac{1}{d}} = d \left(\det \nabla T(x) \right)^{\frac{1}{d}},$$

where the inequality follows from the fact that the arithmetic mean of the nonnegative numbers $\partial_i T_i(x)$ is greater than the geometric one.

□

Proof of Theorem 1.5.1. Thanks to the properties in Proposition 1.5.2, denoting by ν_E the outer unit normal to ∂E, and by $d\sigma$ the surface measure on ∂E, we have

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

where the equality in † follows from Stokes' Theorem. □

1.6 A Jacobian equation for transport maps

Let $T : \mathbb{R}^d \to \mathbb{R}^d$ 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 measure, for any bounded Borel function $\zeta : \mathbb{R}^d \to \mathbb{R}$ we have

$$\int_{\mathbb{R}^d} \zeta(y) g(y) \, dy = \int_{\mathbb{R}^d} \zeta(T(x)) f(x) \, dx.$$

On the other hand, using the change of variables $y = T(x)$ we have $dy = \det \nabla T(x)\, dx$, therefore

$$\int_{\mathbb{R}^d} \zeta(y) g(y)\, dy = \int_{\mathbb{R}^d} \zeta(T(x)) g(T(x)) \det \nabla T(x)\, dx.$$

Comparing the two equations above, since ζ is arbitrary we deduce that T satisfies

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

Note that the transport maps we are going to construct in the next sections (and also the Knothe's map we have just studied) are not smooth diffeomorphisms in general, thus proving that the validity (in a suitable sense) of this Jacobian equation would require some additional work.

2 Optimal Transport

This section contains what is usually considered to be the core of optimal transport theory: the solution of Kantorovich's problem for general costs (i.e., the existence of an optimal transport plan), the duality theory, and the solution of Monge's problem (i.e., the existence of an optimal transport map) for suitable costs. We will also present a couple of classical applications of the theory: the polar decomposition, and an application to the Euler equations of fluid-dynamics.

In order to pursue our plan we will need some preliminaries in measure theory; hence we shall devote the first subsection to these preliminaries.

2.1 Preliminaries in measure theory

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

Remark 2.1.1. The assumptions in this book are far from being sharp, as our goal is to emphasize the main ideas of the theory. In particular, the existence of optimal transport plans (Theorem 2.3.2) and the duality theorem (Theorem 2.6.6) hold in arbitrary separable metric spaces. The interested reader may look at [AGS08, Chapters 5.1-5.4 and 6.1].

Remark 2.1.2. By the *Riesz representation Theorem* (see [BBP16, Theorem 7.7]) we have the following equalities (recall that, given a Banach space $\mathcal{E}$, the notation $\mathcal{E}^*$ denotes its dual):

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

Remark 2.1.3. Note that $C_c(X)$ is not closed if X is not compact. E.g., for $X = \mathbb{R}$, if $\psi_n : \mathbb{R} \to [0,1]$ are continuous functions such that $\psi_n(x) = 1$ for $x \in [-n, n]$ and $\psi_n(x) = 0$ for $x \notin [-n-1, n+1]$, then the sequence of functions

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

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

Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence of probability measures. Then $\mu_k(X) = 1$ and therefore the whole sequence $(\mu_k)_{k\in\mathbb{N}}$ is uniformly bounded in $\mathcal{M}(X)$. Thus, thanks to Banach-Alaoglu's Theorem, there exists a subsequence $(\mu_{k_j})_{j\in\mathbb{N}}$ that weakly-$*$ converges to a measure $\mu \in \mathcal{M}(X)$:

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

i.e.,

$$\int_X \varphi\, d\mu_{k_j} \to \int_X \varphi\, d\mu \quad \text{for any } \varphi \in C_c(X).$$

Note that, since $\mu_k \geq 0$ (by assumption) we have that $\mu \geq 0$. On the other hand, even if μ_k are all probability measures, μ may not be a probability measure, as shown in the following example.

Example 2.1.4. Let $X = \mathbb{R}$ and $\mu_k = \delta_k$ for $k \in \mathbb{Z}$. Then, for any $\varphi \in C_c(\mathbb{R})$,

$$\int_{\mathbb{R}} \varphi\, d\mu_k = \varphi(k) \xrightarrow{k\to\infty} 0.$$

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 2.1.5. Let $C_b(X)$ be the set of continuous bounded functions. We say that μ_k converges to μ *narrowly* if

$$\int_X \varphi \, d\mu_k \to \int_X \varphi \, d\mu \quad \text{for any } \varphi \in C_b(X).$$

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

Remark 2.1.6. The narrow convergence is particularly useful in our context, as it guarantees that limits of probability measures are still probability measures. Indeed, assume that $\mu_k \in \mathcal{P}(X)$ 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 2.1.7. Take $X = \mathbb{R}^d$ and $\mu_k = (1-\frac{1}{k})\delta_0 + \frac{1}{k}\delta_{x_k}$ for some $x_k \in \mathbb{R}^d$. Then, if $\varphi \in C_b(\mathbb{R}^d)$, we have

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

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

The difference with respect to the case when $\mu_k \overset{*}{\rightharpoonup} \mu$ is that, in the weak-$*$ convergence, some mass of μ_k may escape to ∞. To avoid this, one needs to guarantee that almost all the mass of μ_k remain in a fixed compact set. This motivates the following:

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 $\mu \in \mathcal{A}$.

We are going to see that the tightness of a family is equivalent to its compactness with respect to the narrow topology. But before proving such a result, let us present the following fundamental lemma regarding the exhaustion of a measure by compact sets and compactly supported functions.

Lemma 2.1.9. *Given a probability measure $\mu \in \mathcal{P}(X)$, the following statements hold:*

(a) For any $\varepsilon > 0$, there is a compact set $K_\varepsilon \subset X$ such that $\mu(K_\varepsilon) \geq 1 - \varepsilon$.

(b) For any $\varepsilon > 0$, there is $\eta_\varepsilon \in C_c(X)$ with $0 \leq \eta_\varepsilon \leq 1$ such that $\int \eta_\varepsilon \, d\mu \geq 1 - \varepsilon$.

Proof. (a) Since X is separable, there is a countable sequence of points $(x_n)_{n\in\mathbb{N}}$ that is dense in X. Hence, for any $r > 0$, we have

$$\bigcup_{n\in\mathbb{N}} \overline{B(x_n, r)} = X \, .$$

Therefore, given $\varepsilon > 0$, for any $k \in \mathbb{N}$ there exists $n_{k,\varepsilon} \in \mathbb{N}$ such that

$$\mu\Big(\bigcup_{1\leq n\leq n_{k,\varepsilon}} \overline{B(x_n, k^{-1})} \Big) \geq 1 - \frac{\varepsilon}{2^k} \, . \tag{2.1}$$

Let us consider the subset $K_\varepsilon \subset X$ defined as

$$K_\varepsilon \coloneqq \bigcap_{k\in\mathbb{N}} \bigcup_{1\leq n\leq n_{k,\varepsilon}} \overline{B(x_n, k^{-1})} \, . \tag{2.2}$$

Being the intersection of finite unions of closed balls, K_ε is closed. Also, by construction, the set K_ε is also totally bounded. Hence, since X is complete, we deduce that K_ε is compact [Wil70, Theorem 39.9]. Finally, (2.1) implies that

$$\mu(X \setminus K_\varepsilon) \leq \sum_{k\in\mathbb{N}} \mu\left(X \setminus \bigcup_{1\leq n\leq n_{k,\varepsilon}} \overline{B(x_n, k^{-1})}\right) \leq \sum_{k\in\mathbb{N}} \frac{\varepsilon}{2^k} = \varepsilon,$$

thus $\mu(K_\varepsilon) \geq 1-\varepsilon$, as desired.

(b) Let K_ε be the compact set provided by the previous step. Since X is locally compact, there exists a compact set H_ε such that $K_\varepsilon \subset \mathring{H}_\varepsilon$. Thus, Tietze's extension Theorem [Wil70, p. 15.8] guarantees the existence of a function $\eta_\varepsilon \in C(X)$ such that $\eta_\varepsilon \equiv 1$ in K_ε, $\eta_\varepsilon \equiv 0$ in $X \setminus H_\varepsilon$, and $0 \leq \eta_\varepsilon \leq 1$. This function satisfies all the requirements in the statement.

□

Remark 2.1.10. Notice that (a) of Lemma 2.1.9 implies that the singleton $\{\mu\}$ constitutes a tight family.

We are now ready to prove that tightness is a necessary and sufficient condition for compactness with respect to the narrow convergence.

Theorem 2.1.11 (Prokhorov)**.** *A family $\mathcal{A} \subset \mathcal{P}(X)$ is tight if and only if $\mathcal{A}$ is relatively compact for the narrow convergence, i.e., for any sequence $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{A}$ there exists a subsequence $(\mu_{k_j})_{j\in\mathbb{N}}$ and a probability measure $\mu \in \mathcal{P}(X)$ such that*

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

Proof. We prove only the implication "tightness implies compactness" (see Remark 2.1.12 below); for the other implication, we refer the interested reader to the proof of [Bog07, Theorem 8.6.2].

Since the family is tight, there is a sequence of compact sets $(K_n)_{n\in\mathbb{N}}$ such that

$$\mu(X \setminus K_n) \leq n^{-1} \qquad \forall\, \mu \in \mathcal{A}. \tag{2.3}$$

Since the space X is locally compact, up to enlarging inductively each compact set, we may assume $K_n \subset \mathring{K}_{n+1}$ for any $n \in \mathbb{N}$.

Given a sequence $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{A}$, by Banach-Alaoglu's Theorem the restricted measures $\mu_k|_{K_n}$[4] converge weakly-$*$, up to subsequence, to a measure $\mu^{(n)} \in \mathscr{M}(X)$. Therefore, by a diagonal argument, there exists a subsequence $\{k_j\}_{j\in\mathbb{N}}$ such that

$$\mu_{k_j}|_{K_n} \overset{*}{\rightharpoonup} \mu^{(n)} \in \mathscr{M}(X) \qquad \forall\, n \in \mathbb{N}. \tag{2.4}$$

Note that $\mu^{(n)}$ vanishes outside K_n and $\mu^{(n)}(X \setminus K_m) \leq m^{-1}$ for any $m \in \mathbb{N}$ (recall (2.3)).

Testing (2.4) against functions compactly supported in $\mathring{K}_n$, we deduce that $\mu^{(n+1)}|_{\mathring{K}_n} = \mu^{(n)}|_{\mathring{K}_n}$. By construction we have $\mu^{(n+1)} \geq \mu^{(n)}$ and thus

$$\hat{\mu} := \sup_{n\in\mathbb{N}} \mu^{(n)}$$

is a well-defined measure satisfying $\hat{\mu}(X \setminus K_n) \leq n^{-1}$ and $\hat{\mu}|_{\mathring{K}_n} = \mu^{(n)}|_{\mathring{K}_n}$ for every $n \in \mathbb{N}$. Therefore, recalling (2.4), we have

$$\mu_{k_j}|_{\mathring{K}_n} \overset{*}{\rightharpoonup} \hat{\mu}|_{\mathring{K}_n} \qquad \forall\, n \in \mathbb{N}.$$

[4] That is, $\mu_k|_{K_n}(E) := \mu_k(E \cap K_n)$ for any $E \subset X$ Borel.

Since $\mathring{K}_n$ is the subset of a compact set (that is K_n), the latter weak-$*$ convergence is equivalent to the narrow convergence

$$\mu_{k_j}|_{\mathring{K}_n} \rightharpoonup \hat{\mu}|_{\mathring{K}_n} \qquad \forall\, n \in \mathbb{N}. \tag{2.5}$$

Thus, recalling that $\mu^{(n)}(X \setminus K_n) \leq n^{-1}$ and $\hat{\mu}(X \setminus K_n) \leq n^{-1}$, and $K_{n-1} \subset \mathring{K}_n$, it follows from (2.5) that, for any $\varphi \in C_b(X)$,

$$\begin{aligned}
&\limsup_{j\to\infty} \left| \int_X \varphi \, d\mu_{k_j} - \int_X \varphi \, d\hat{\mu} \right| \\
&\quad \leq \limsup_{n\to\infty} \limsup_{j\to\infty} \left| \int_{X\setminus\mathring{K}_n} \varphi \, d\mu_{k_j} \right| + \left| \int_{X\setminus\mathring{K}_n} \varphi \, d\hat{\mu} \right| + \left| \int_X \varphi \, d\mu_{k_j}|_{\mathring{K}_n} - \int_X \varphi \, d\mu|_{\mathring{K}_n} \right| \\
&\quad = \limsup_{n\to\infty} \|\varphi\|_\infty (n-1)^{-1} + \|\varphi\|_\infty (n-1)^{-1} = 0\,.
\end{aligned}$$

Since φ is arbitrary, we have shown that $\mu_k \rightharpoonup \hat{\mu}$ narrowly and in particular $\hat{\mu} \in \mathcal{P}(X)$.

□

Remark 2.1.12. For us, the important implication will be that a tight family is relatively compact with respect to the narrow convergence.

In the next lemma we show that if a sequence of probability measures converges weakly-$*$ to a probability measure (so, we are assuming that the limit has still mass 1), then in fact the convergence is narrow.

Lemma 2.1.13 (weak-$*$ convergence + mass conservation = narrow convergence)**.** *Let $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{P}(X)$, and assume that $\mu_k \overset{*}{\rightharpoonup} \mu$ for some $\mu \in \mathcal{P}(X)$. Then the family $\{\mu_k : k \in \mathbb{N}\}$ is tight and $\mu_k \rightharpoonup \mu$.*

Proof. Choose $\varepsilon > 0$. Thanks to Lemma 2.1.9 there is a compactly supported function $\eta_\varepsilon \in C_c(X)$ such that $0 \leq \eta_\varepsilon \leq 1$ and

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

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

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

Therefore there exists k_ε 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 $k < k_\varepsilon$, applying Lemma 2.1.9 again, we can find a compact set $K_{\varepsilon,k}$ such that

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

Set $\hat{K}_\varepsilon := \mathrm{supp}(\eta_\varepsilon) \cup \bigcup_{k=1}^{k_\varepsilon} K_{\varepsilon,k}$. 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 all $k \in \mathbb{N}$ (or equivalently $\mu_k(X \setminus \hat{K}_\varepsilon) \leq 2\varepsilon$), thus the family $\{\mu_k : k \in \mathbb{N}\}$ is tight.

Hence, given any subsequence μ_{k_j}, Theorem 2.1.11 implies the existence of a subsequence $\mu_{k_{j_\ell}}$ such that $\mu_{k_{j_\ell}} \rightharpoonup \nu \in \mathcal{P}(X)$. On the other hand, since $\mu_k \overset{*}{\rightharpoonup} \mu$, we also have $\mu_{k_{j_\ell}} \overset{*}{\rightharpoonup} \mu$. Therefore, for any $\varphi \in C_c(X)$ we have

$$\int_X \varphi \, d\nu \leftarrow \int_X \varphi \, d\mu_{k_{j_\ell}} \to \int_X \varphi \, d\mu.$$

The arbitrariness of φ implies that $\mu = \nu$.

In other words, we proved that for any narrowly converging subsequence of μ_k, the limit is independent of the choice of the subsequence and coincides with μ. This implies that the whole sequence μ_k narrowly converges to μ, as desired. □

The next result shows the lower semicontinuity of the map $\mu \mapsto \int \varphi \, d\mu$ under weak-$*$ convergence, whenever the integrand φ is nonnegative and lower semicontinuous. Since the narrow topology is stronger than the weak-$*$ topology (i.e., narrow convergence implies weak-$*$ convergence), this result implies also the lower semicontinuity of the map $\mu \mapsto \int \varphi \, d\mu$ under narrow convergence.

Lemma 2.1.14 (lower semicontinuity of integrals)**.** *Let $\mu_k \overset{*}{\rightharpoonup} \mu$, and let $\varphi\colon X \to [0, +\infty]$ be a lower semicontinuous function.*[5] *Then*

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

Proof. If $\varphi \equiv +\infty$ then the statement is trivial, hence we assume that this is not the case. Given $\lambda \geq 0$, define

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

where $d : X \times X \to \mathbb{R}$ denotes the distance function on X.

We claim that the functions φ_λ satisfy the following properties:

(a) If $\lambda < \lambda'$ then $\varphi_\lambda \leq \varphi_{\lambda'} \leq \varphi$;

(b) φ_λ is λ-Lipschitz;

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

Let us prove the mentioned properties.

(a) For any $y \in X$ we have $\varphi_\lambda(x) \leq \varphi(y) + \lambda \, d(x,y) \leq \varphi(y) + \lambda' \, d(x,y)$. Taking the infimum over $y \in X$ shows that $\varphi_\lambda \leq \varphi_{\lambda'}$. Also, taking $y = x$ in the definition of $\varphi_{\lambda'}$ proves that $\varphi_{\lambda'} \leq \varphi$.

(b) Let $x, x' \in X$. Then, by the triangle inequality,

$$\varphi_\lambda(x') \leq \varphi(y) + \lambda \, d(x', y) \leq \varphi(y) + \lambda \, d(x,y) + \lambda \, d(x,x') \qquad \forall \, y \in X.$$

Taking the infimum over y yields $\varphi_\lambda(x') \leq \varphi_\lambda(x) + \lambda \, d(x,x')$. Since the argument is symmetric in x and x', this proves that

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

(c) Fix $x \in X$. Since φ is lower semicontinuous, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that $\varphi(y) \geq \min\big\{\varphi(x) - \varepsilon, \frac{1}{\varepsilon}\big\}$ for all $y \in X$ with $d(x,y) \leq \delta$.[6] Thus, recalling that $\varphi \geq 0$, we have

$$\begin{cases} \varphi(y) + \lambda \, d(x,y) \geq \min\big\{\varphi(x) - \varepsilon, \frac{1}{\varepsilon}\big\} & \text{if } d(x,y) \leq \delta \\ \varphi(y) + \lambda \, d(x,y) \geq \lambda \, \delta & \text{if } d(x,y) > \delta \end{cases}$$

[5] Recall that a function φ is lower semicontinuous if $\liminf_{k\to\infty} \varphi(x_k) \geq \varphi(x)$ as $x_k \to x$.

[6] Indeed, 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 $\delta > 0$ such that $\varphi(y) \geq \varphi(x) - \varepsilon$ for $d(x,y) \leq \delta$.
- If $\varphi(x) = +\infty$, then for any $\varepsilon > 0$ there exists $\delta > 0$ such that $\varphi(y) \geq \frac{1}{\varepsilon}$ for $d(x,y) \leq \delta$.

from which it follows that

$$\varphi_\lambda(x) \geq \min\left\{\varphi(x) - \varepsilon, \frac{1}{\varepsilon}, \lambda\,\delta\right\}.$$

Letting $\lambda \to \infty$, this implies that $\liminf_{\lambda\to\infty} \varphi_\lambda(x) \geq \min\{\varphi(x) - \varepsilon, \frac{1}{\varepsilon}\}$. Since ε is arbitrary, this proves that $\liminf_{\lambda\to\infty} \varphi_\lambda(x) \geq \varphi(x)$. Because (a) yields the converse inequality, we conclude that

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

Note that since $\varphi \geq 0$, we have $\varphi_\lambda \geq 0$. Consider the family of compactly supported functions $(\eta_\varepsilon)_{\varepsilon>0}$ constructed in Lemma 2.1.9. We may assume that $\eta_{\frac{1}{i}} \leq \eta_{\frac{1}{i+1}}$ for any $i \in \mathbb{N}$, and therefore $\eta_{\frac{1}{i}} \nearrow 1$ μ-almost everywhere as $i \to \infty$. Then, we define

$$\psi_i(x) \coloneqq \varphi_i(x)\,\eta_{\frac{1}{i}}(x).$$

Note that ψ_i is continuous and compactly supported. Thus, given that $\psi_i \leq \varphi$ (by property (a) above, since $\psi_i \leq \varphi_i$), by the weak-$*$ convergence of μ_k to μ we get

$$\int_X \psi_i\,d\mu = \lim_{k\to\infty} \int_X \psi_i\,d\mu_k \leq \liminf_{k\to\infty} \int_X \varphi\,d\mu_k.$$

Since $\psi_i \nearrow \varphi$ μ-almost everywhere, we conclude by monotone convergence:

$$\int_X \varphi\,d\mu = \lim_{i\to\infty} \int_X \psi_i\,d\mu \leq \liminf_{k\to\infty} \int_X \varphi\,d\mu_k.$$

□

2.2 Monge vs. Kantorovich

Fix $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c\colon X \times Y \to [0, +\infty]$ lower semicontinuous. The Monge and the Kantorovich's problems can be stated as follows (recall Definition 1.4.3):

$$C_M(\mu,\nu) \coloneqq \inf\left\{\int_X c(x,T(x))\,d\mu(x) \mid T_\#\mu = \nu\right\}. \tag{M}$$

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

In other words, Monge's problem (M) consists in minimizing the transportation cost among all transport maps, while Kantorovich's problem (K) consists in minimizing the transportation cost among all couplings.

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

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

In other words, any transport map T induces a coupling γ_T with the same cost. Thanks to this fact, we deduce that

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

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 from μ to ν. In other words, if we have a coupling with the structure of a graph, this yields a transport map.

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. We split the proof in two steps: we first prove tightness and then closedness.

- $\Gamma(\mu,\nu)$ *is tight.* Thanks to Lemma 2.1.9, for all $\varepsilon > 0$, there exists a set $K_\varepsilon \subset X$ such that $\mu(X \setminus K_\varepsilon) \le \frac{\varepsilon}{2}$. Analogously for ν, there exists a set $\tilde{K}_\varepsilon \subset Y$ such that $\nu(Y \setminus \tilde{K}_\varepsilon) \le \frac{\varepsilon}{2}$.

  Define the compact set $\bar{K}_\varepsilon \coloneqq K_\varepsilon \times \tilde{K}_\varepsilon \subset X \times Y$. Then, for any $\gamma \in \Gamma(\mu,\nu)$, we have

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

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

- $\Gamma(\mu,\nu)$ *is closed under narrow convergence.* Take a sequence $(\gamma_k)_{k\in\mathbb{N}} \subset \Gamma(\mu,\nu)$, and assume that $\gamma_k \rightharpoonup \gamma \in \mathcal{P}(X \times Y)$. Then, for any $\varphi \in C_b(X)$ we have

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

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

□

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

Proof. Without loss of generality $\alpha \coloneqq \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma < +\infty$ (if $\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma = +\infty$, then the statement is trivial since every $\gamma \in \Gamma(\mu,\nu)$ is a minimizer).

Let $(\gamma_k)_{k\in\mathbb{N}} \subset \Gamma(\mu,\nu)$ be a minimizing sequence, namely

$$\int_{X\times Y} c\, d\gamma_k \to \alpha \qquad \text{as } k \to \infty.$$

Since $\{\gamma_k\} \subset \Gamma(\mu,\nu)$ is tight, by Theorem 2.1.11 there exists a subsequence $(\gamma_{k_j})_{j\in\mathbb{N}}$ such that $\gamma_{k_j} \rightharpoonup \bar{\gamma}$. Since c is nonnegative and lower semicontinuous, it follows from Lemma 2.1.14 that

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

Since $\bar{\gamma} \in \Gamma(\mu,\nu)$ (thanks to Lemma 2.3.1), we clearly have $\int_{X\times Y} c\, d\bar{\gamma} \ge \alpha$. This proves that $\int_{X\times Y} c\, d\bar{\gamma} = \alpha$, thus $\bar{\gamma}$ is a minimizer. □

In other words, under very general assumptions on the cost function, an optimal coupling always exists.

Remark 2.3.3. Note that, up to replacing $c(x,y)$ with $c(x,y) + C$ for some constant $C \in \mathbb{R}$, all results proved in this book still hold for costs c bounded from below.

The natural questions that now arise are the following:

1. Is the minimizer γ unique?
2. Is it given by a transport map?

In order to get an intuition on these two important questions, let us consider 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 there exists a unique element in $\Gamma(\mu,\nu)$, given by the coupling $\gamma := \frac{1}{2}\delta_{(x_0,y_0)} + \frac{1}{2}\delta_{(x_0,y_1)}$. So the minimizer is unique (for every cost), but it is not induced by a transport map.

Example 2.3.5. Let $X = Y = \mathbb{R}^2$, let $c(x,y) = |x-y|^2$, consider the points in $\mathbb{R}^2$ given by

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

and define the measures

$$\mu = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}, \qquad \nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}.$$

In this case the set of all couplings from μ to ν is obtained by sending an amount $\alpha \in \left[0, \frac{1}{2}\right]$ from x_1 to y_1, the remaining amount $\frac{1}{2} - \alpha$ from x_1 to y_2, then an amount $\beta \in \left[0, \frac{1}{2}\right]$ from x_2 to y_2, and finally the remaining amount $\frac{1}{2} - \beta$ from x_2 to y_1.

In other words, the set $\Gamma(\mu,\nu)$ is given by:

$$\gamma_{\alpha,\beta} = \alpha\delta_{(x_1,y_1)} + \left(\frac{1}{2} - \alpha\right)\delta_{(x_1,y_2)} + \left(\frac{1}{2} - \beta\right)\delta_{(x_2,y_1)} + \beta\delta_{(x_2,y_2)}, \qquad \alpha,\beta \in \left[0, \frac{1}{2}\right].$$

Note that, for all $\alpha, \beta \in \left[0, \frac{1}{2}\right]$,

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

Hence all couplings $\gamma_{\alpha,\beta}$ are optimal, ruling out the uniqueness of the optimal plan without further assumptions.

2.4 c-cyclical monotonicity

Let us recall the definition of support of a measure.

Definition 2.4.1. Given a measure $\mu \in \mathcal{M}(X)$, its *support* is defined as

$$\operatorname{supp}(\mu) := \{x \in X \mid \forall \varepsilon > 0 \colon \mu(B_\varepsilon(x)) > 0\}.$$

We want to investigate the properties of the support of an optimal coupling.

Given an optimal coupling $\bar\gamma \in \Gamma(\mu,\nu)$ with finite cost (i.e., $\int_{X\times Y} c\, d\bar\gamma = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma < +\infty$), its support is

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

So, morally speaking, a pair of points (x,y) belongs to the support if some mass goes from x to y.

To understand how to exploit the optimality of $\bar{\gamma}$, suppose for instance that $(x_i, y_i)_{i=1,2,3} \in \mathrm{supp}(\bar{\gamma})$. This means that $\bar{\gamma}$ sends mass from x_i to y_i. Now, consider another transport plan which 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, this "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 we set $x_4 \equiv x_1$.[7] Since this property needs to hold for any collection of points in the support of $\bar{\gamma}$, this motivates the following definition:

Definition 2.4.2. A set $\Lambda \subset X \times Y$ is said to be *c-cyclically monotone* if for any finite sequence $(x_i, y_i)_{i=1,\dots,N} \subset \Lambda$, the following holds:

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

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

The above discussion suggests that optimality implies c-cyclically monotonicity. This is indeed the case:

Theorem 2.4.3. *Let $\bar{\gamma}$ be optimal and $c\colon X \times Y \to \mathbb{R}$ continuous. Then* $\mathrm{supp}(\bar{\gamma})$ *is c-cyclically monotone.*

Remark 2.4.4. We will show later that the above statement is an *if and only if*, see Theorem 2.6.3.

Proof. By contradiction, suppose $\mathrm{supp}(\bar{\gamma})$ is not c-cyclically monotone. Then there exist $\eta > 0$ and N pairs of points $(x_1, y_1), \dots, (x_N, y_N) \in \mathrm{supp}(\bar{\gamma})$ such that

$$\sum_{i=1}^{N} c(x_i, y_i) \geq \sum_{i=1}^{N} c(x_{i+1}, y_i) + \eta. \tag{2.6}$$

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

$$|c(x, y) - c(x_i, y_i)| \leq \frac{\eta}{4N} \qquad \forall\, (x, y) \in U_i \times V_i \tag{2.7}$$

and

$$|c(x, y) - c(x_{i+1}, y_i)| \leq \frac{\eta}{4N} \qquad \forall\, (x, y) \in U_{i+1} \times V_i\,. \tag{2.8}$$

Set $\varepsilon_i \coloneqq \bar{\gamma}(U_i \times V_i)$. Note that all ε_i are positive, since (x_i, y_i) belong to the support of $\bar{\gamma}$. Now set $\varepsilon \coloneqq \min_{i=1,\dots,N} \varepsilon_i$ and[8] $\gamma_i \coloneqq \frac{\bar{\gamma}|_{U_i \times V_i}}{\varepsilon_i} \in \mathcal{P}(X \times Y)$. Then we define the measures $\mu_i \coloneqq (\pi_X)_{\#}\gamma_i \in \mathcal{P}(X)$ and $\nu_i \coloneqq (\pi_Y)_{\#}\gamma_i \in \mathcal{P}(Y)$, and we set

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

[7] Of course this argument is not rigorous, since the points (x_i, y_i) may have zero mass for $\bar{\gamma}$. However, as we shall see later, one can make this argument rigorous by considering some small neighborhoods of (x_i, y_i).

[8] 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 $E \subset X$, $\bar{\gamma}|_A(E) \coloneqq \bar{\gamma}(A \cap E)$).

Let us show that $\gamma' \geq 0$. Since $\varepsilon \leq \varepsilon_i$, we have

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

Let us also check that $\gamma' \in \Gamma(\mu,\nu)$. 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

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

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

It remains only to prove $\int_{X\times Y} c\,d\gamma' < \int_{X\times Y} c\,d\bar{\gamma}$ because this yields the sought contradiction, since $\bar{\gamma}$ was assumed to be optimal. Note that, since $\mu_i \in \mathcal{P}(X)$ is supported inside U_i and $\nu_i \in \mathcal{P}(Y)$ is supported inside V_i, it follows from (2.8) that

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

Analogously, since $\gamma_i \in \mathcal{P}(X\times Y)$ is supported inside $U_i\times V_i$,

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

Then, recalling (2.6), we get

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

a contradiction that concludes the proof. □

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

Let $X = Y = \mathbb{R}^d$ and $c(x,y) = \frac{|x-y|^2}{2}$. Also, assume that $\int_{\mathbb{R}^d}\frac{|x|^2}{2}\,d\mu + \int_{\mathbb{R}^d}\frac{|y|^2}{2}\,d\nu < +\infty$. Let $\gamma \in \Gamma(\mu,\nu)$, then

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^d}\frac{|x-y|^2}{2}\,d\gamma(x,y) &= \int_{\mathbb{R}^d\times\mathbb{R}^d}\left(\frac{|x|^2}{2} + \frac{|y|^2}{2} - x\cdot y\right)d\gamma \\ &= \int_{\mathbb{R}^d}\frac{|x|^2}{2}\,d\mu + \int_{\mathbb{R}^d}\frac{|y|^2}{2}\,d\nu + \int_{\mathbb{R}^d\times\mathbb{R}^d} -x\cdot y\,d\gamma.\end{aligned} \tag{2.9}$$

Since the first two terms in the last expression are independent of γ, we deduce that γ is optimal for the cost $c(x,y) = \frac{|x-y|^2}{2}$ if and only if it is optimal for the cost $c(x,y) = -x \cdot y$.

Hence, in the next section we shall work with the cost function $c(x,y) = -x \cdot y$, as it simplifies several definitions and computations.

2.5.1 Cyclical monotonicity and Rockafellar's Theorem

In the case $c(x,y) = -x \cdot y$, the condition

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

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[9] on $\mathbb{R}^d$, and by convention $x_{N+1} \equiv x_1$. Any subset of $\mathbb{R}^d \times \mathbb{R}^d$ satisfying this last property (for any family of points $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ contained in such set) is called a *cyclically monotone set*.

The goal of this section is to characterize cyclical monotonicity in terms of subdifferential of convex functions. We first recall the definition.

Definition 2.5.1. Given $\varphi \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ convex, we define the *subdifferential* of φ as

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

Also, we define $\partial\varphi \coloneqq \bigcup_{x \in \mathbb{R}^d} \{x\} \times \partial\varphi(x) \subset \mathbb{R}^d \times \mathbb{R}^d$.

Theorem 2.5.2 (Rockafellar)**.** *A set $S \subset \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone if and only if there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that $S \subset \partial\varphi$.*

Proof. First we show that if such a convex function exists then the set S has to be cyclically monotone. The converse implication will then by proved constructing φ explicitly.

$\Leftarrow$ Assume that $S \subset \partial\varphi$, and take a finite set of points $(x_i, y_i)_{i=1,\ldots,N} \subset S \subset \partial\varphi$. Then, for each i we have that $y_i \in \partial\varphi(x_i)$, and therefore

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

In particular, choosing $z = x_{i+1}$ we obtain

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

and summing over i (where we adopt the convention $N + 1 \equiv 1$) yields

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

Since the two summands containing φ are equal, this implies that

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

[9] Throughout this book, we shall use both notation $\langle \cdot, \cdot \rangle$ and $\cdot$ indistinguishably.

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

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

Note that:

(i) φ is a supremum of affine functions, thus it is convex.

(ii) Choosing $N = 1$ and $(x_1, y_1) = (x_0, x_0)$ yields

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

and in particular $\varphi(x_0) \geq 0$.

(iii) For any $(x_i, y_i)_{i=1,\dots,N} \subset S$, because of cyclic monotonicity, we have

$$\langle y_N, x_0 - x_N\rangle + \cdots + \langle y_0, x_1 - x_0\rangle \leq 0.$$

Hence $\varphi(x_0) \leq 0$, that combined with (ii) implies that $\varphi(x_0) = 0$. In particular $\varphi \not\equiv +\infty$.

We now prove that $S \subset \partial\varphi$.

Take $(\bar{x}, \bar{y}) \in S$ and let $\alpha < \varphi(\bar{x})$. Then, by the definition of φ, there exist $N \geq 1$ and a sequence $(x_i, y_i)_{i=1,\dots,N}$ such that

$$\langle y_N, \bar{x} - x_N\rangle + \cdots + \langle y_0, x_1 - x_0\rangle \geq \alpha. \tag{2.10}$$

Consider now the sequence $(x_i, y_i)_{i=1,\dots,N+1}$ obtained by taking $(x_{N+1}, y_{N+1}) = (\bar{x}, \bar{y})$. Since this new sequence is admissible in the definition of φ, using (2.10) we deduce that, for any $z \in \mathbb{R}^d$,

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

Letting $\alpha \to \varphi(\bar{x})$, this shows that $\varphi(z) \geq \langle \bar{y}, z - \bar{x}\rangle + \varphi(\bar{x})$ for all $z \in \mathbb{R}^d$, thus $\bar{y} \in \partial\varphi(\bar{x})$ (or equivalently $(\bar{x}, \bar{y}) \in \partial\varphi$), as desired.

□

2.5.2 Kantorovich Duality

With the use of the Legendre transform, we now want to find a *dual problem* to the Kantorovich's problem. We shall do this in a constructive way. However, the reader familiar with convex optimization will not be surprised: since Kantorovich's problem is a linear minimization with convex constraints, it admits a dual problem by "abstract convex analysis" (see Remark 2.6.8).

Definition 2.5.3. Given $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ convex (with $\varphi \not\equiv +\infty$), one defines the *Legendre transform* of φ,

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

as

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

Proposition 2.5.4. *The following properties hold:*

(a) $\varphi(x) + \varphi^*(y) \geq x \cdot y$ *for all* $x, y \in \mathbb{R}^d$*;*

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

Proof. As we shall see, both properties follow easily from our definitions.

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

(b) $\Rightarrow$ Assume that $\varphi(x) + \varphi^*(y) = x \cdot y$. By (a) we have

$$\varphi^*(y) \geq z \cdot y - \varphi(z) \qquad \forall\, z \in \mathbb{R}^d.$$

Since $x \cdot y - \varphi(x) = \varphi^*(y)$, this implies that

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

thus $y \in \partial\varphi(x)$.

$\Leftarrow$ If $y \in \partial\varphi(x)$, then for any $z \in \mathbb{R}^d$ we have $\varphi(z) \geq \varphi(x) + \langle y, z - x \rangle$, or equivalently

$$x \cdot y - \varphi(x) \geq z \cdot y - \varphi(z) \qquad \forall\, z \in \mathbb{R}^d.$$

By taking the supremum over $z \in \mathbb{R}^d$ 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 $c(x, y) = -x \cdot y$ does not immediately follow from our previous results, since we only proved existence of an optimal coupling for nonnegative cost functions. However, we can use that the cost $c(x, y) = -x \cdot y$ is equivalent to the cost $c'(x, y) \coloneqq \frac{|x-y|^2}{2}$ provided that $\int \frac{|x|^2}{2}\, d\mu + \int \frac{|y|^2}{2}\, d\nu < +\infty$ (see (2.9)). In addition, noticing that

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

it follows that $\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times Y} c'\, d\gamma < +\infty$ (recall that $\mu \otimes \nu \in \Gamma(\mu, \nu)$).

Hence, 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.

Theorem 2.5.5 (Kantorovich duality)**.** *Assume that*

$$\int_{\mathbb{R}^d} |x|^2\, d\mu(x) + \int_{\mathbb{R}^d} |y|^2\, d\nu(y) < +\infty.$$

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

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

Proof. Consider $\varphi, \psi \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that

$$\varphi(x) + \psi(y) \geq x \cdot y \qquad \forall\, x, y \in \mathbb{R}^d.$$

Integrating this inequality with respect to an arbitrary coupling $\gamma \in \Gamma(\mu, \nu)$ yields[10]

$$\begin{aligned} \int_{\mathbb{R}^d \times \mathbb{R}^d} -x \cdot y \, d\gamma(x,y) &\geq \int_{\mathbb{R}^d \times \mathbb{R}^d} -\varphi(x) \, d\gamma(x,y) + \int_{\mathbb{R}^d \times \mathbb{R}^d} -\psi(y) \, d\gamma(x,y) \\ &= \int_{\mathbb{R}^d} -\varphi(x) \, d\mu(x) + \int_{\mathbb{R}^d} -\psi(y) \, d\nu(y). \end{aligned} \tag{2.11}$$

Note that the left-hand side does not depend on φ and ψ, and the right-hand side does not depend on γ. Thus

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} -x \cdot y \, d\gamma(x,y) \geq \sup_{\varphi(x)+\psi(y) \geq x \cdot y} \int_{\mathbb{R}^d} -\varphi(x) \, d\mu(x) + \int_{\mathbb{R}^d} -\psi(y) \, d\nu(y). \tag{2.12}$$

On the other hand, let $\bar{\gamma} \in \Gamma(\mu, \nu)$ be optimal. Theorem 2.4.3 implies that $\operatorname{supp}(\bar{\gamma})$ is cyclically monotone, and so Theorem 2.5.2 yields the existence of a convex map $\varphi \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that $\operatorname{supp}(\bar{\gamma}) \subset \partial\varphi$, that is, $y \in \partial\varphi(x)$ for any $(x, y) \in \operatorname{supp}(\bar{\gamma})$. Thanks to Proposition 2.5.4, this implies that $\varphi(x) + \varphi^*(y) = x \cdot y$ for $\bar{\gamma}$-almost every (x, y). Thus we have

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

Hence the triple $(\bar{\gamma}, \varphi, \varphi^*)$ gives equality in equation (2.11). □

Remark 2.5.6. In the proof above, the optimality of $\bar{\gamma}$ is only used to deduce that $\operatorname{supp}(\bar{\gamma}) \subset \partial\varphi$ for some convex function φ. Hence, the proof actually shows that if $\operatorname{supp}(\bar{\gamma}) \subset \partial\varphi$ with φ convex, then

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} -x \cdot y \, d\bar{\gamma} = \int_{\mathbb{R}^d} -\varphi \, d\mu + \int_{\mathbb{R}^d} -\varphi^* \, d\nu.$$

[10] Here there is a subtle point: to apply Fubini's Theorem and say that $\int_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(x) \, d\gamma(x,y) = \int_{\mathbb{R}^d} \varphi(x) \, d\mu(x)$, one would need to make sure that φ is integrable (and analogously for ψ). Hence, to justify this identity, we argue as follows: since $\varphi(x) + \psi(y) \geq x \cdot y \geq -\frac{|x|^2}{2} - \frac{|y|^2}{2}$, it means that

$$\varphi(x) + \frac{|x|^2}{2} + \psi(y) + \frac{|y|^2}{2} \geq 0 \quad \forall\, x, y \in \mathbb{R}^d.$$

Choosing two points $x_0, y_0 \in \mathbb{R}^d$ where φ and ψ are respectively finite (if these points do not exist it means that $\varphi \equiv +\infty$ or $\psi \equiv +\infty$, and then (2.11) is trivially true), this implies that

$$\Phi(x) := \varphi(x) + \frac{|x|^2}{2} \geq -C_0 := -\psi(y_0) - \frac{|y_0|^2}{2} \quad \forall\, x, \qquad \Psi(y) := \psi(y) + \frac{|y|^2}{2} \geq -C_1 := -\varphi(x_0) - \frac{|x_0|^2}{2} \quad \forall\, y.$$

Now, since $\Phi + C_0$ and $\Psi + C_1$ are nonnegative, we can monotonically approximate them with the Borel and bounded functions $\Phi_k := \min\{\Phi + C_0, k\}$ and $\Psi_k := \min\{\Psi + C_1, k\}$, $k \in \mathbb{N}$. Then, applying the definition of coupling to Φ_k and Ψ_k (see Definition 1.4.3) and letting $k \to \infty$, by monotone convergence we get

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \big(\Phi(x) + C_0\big) \, d\gamma(x,y) = \int_{\mathbb{R}^d} \big(\Phi(x) + C_0\big) \, d\mu(x), \qquad \int_{\mathbb{R}^d \times \mathbb{R}^d} \big(\Psi(y) + C_1\big) \, d\gamma(x,y) = \int_{\mathbb{R}^d} \big(\Psi(y) + C_1\big) \, d\nu(y).$$

Finally, since $\int |x|^2 d\gamma = \int |x|^2 d\mu < +\infty$ and $\int |y|^2 d\gamma = \int |y|^2 d\nu < +\infty$, we can subtract $\frac{|x|^2}{2} + C_0$ (resp. $\frac{|y|^2}{2} + C_1$) from the equation above to deduce that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \varphi(x) \, d\gamma(x,y) = \int_{\mathbb{R}^d} \varphi(x) \, d\mu(x), \qquad \int_{\mathbb{R}^d \times \mathbb{R}^d} \psi(y) \, d\gamma(x,y) = \int_{\mathbb{R}^d} \psi(y) \, d\nu(y).$$

Since the right-hand side is bounded from above by $\inf_{\gamma\in\Gamma(\mu,\nu)}\int -x\cdot y\,d\gamma$ (thanks to (2.11)), we conclude that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} -x\cdot y\,d\bar{\gamma} \le \inf_{\gamma\in\Gamma(\mu,\nu)}\int -x\cdot y\,d\gamma,$$

thus $\bar{\gamma}$ is optimal. So we proved the implication

$$\mathrm{supp}(\bar{\gamma})\subset\partial\varphi \text{ with } \varphi \text{ convex} \quad\Rightarrow\quad \bar{\gamma} \text{ is optimal.}$$

As a consequence of this remark, together with Theorems 2.4.3 and 2.5.2, we obtain the following:

Corollary 2.5.7. *Let $c(x,y)=\frac{|x-y|^2}{2}$ (or equivalently $c(x,y)=-x\cdot y$). The following are equivalent:*

- *$\bar{\gamma}$ is optimal;*
- *$\mathrm{supp}(\bar{\gamma})$ is cyclically monotone;*
- *there exists a convex map $\varphi:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ such that $\mathrm{supp}(\bar{\gamma})\subset\partial\varphi$.*

Remark 2.5.8. These equivalences are particularly useful when proving that a certain transport map is optimal. Indeed, given a transport map T from μ to ν, $\gamma_T := (\mathrm{Id}\times T)_{\#}\mu$ is optimal if and only if there exists a convex map φ such that $\mathrm{supp}(\gamma_T)\subset\partial\varphi$ (recall Remark 2.2.1). Recalling the definition of $\partial\varphi$, this is equivalent to asking that

$$T(x)\in\partial\varphi(x) \qquad \text{for } \mu\text{-a.e. } x. \tag{2.13}$$

In particular, given a convex function ϕ and a measure $\mu\in\mathcal{P}(\mathbb{R}^d)$, assume that ϕ is differentiable μ-a.e. Then the map $T:=\nabla\phi$ is well defined μ-a.e., and we can consider the measure $\nu:=(\nabla\phi)_{\#}\mu$. Since $\partial\phi(x)=\{\nabla\phi(x)\}$ at every differentiability point of ϕ, the above optimality condition (2.13) is trivially satisfied and therefore

$$\nabla\phi \text{ is an optimal map from } \mu \text{ onto } \nu=(\nabla\phi)_{\#}\mu.$$

2.5.3 Brenier's Theorem

We are now ready to state and prove a cornerstone of the optimal transport theory [Bre87].

Theorem 2.5.9 (Brenier's Theorem)**.** *Let $X=Y=\mathbb{R}^d$ and $c(x,y)=\frac{|x-y|^2}{2}$ (or equivalently $c(x,y)=-x\cdot y$). Suppose that*

$$\int_{\mathbb{R}^d}|x|^2d\mu+\int_{\mathbb{R}^d}|y|^2d\nu<+\infty$$

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

Proof. The proof takes four steps: Steps 1-3 for the existence, and Step 4 for the uniqueness.

1. Note that the cost $c(x,y)=\frac{|x-y|^2}{2}$ is nonnegative and continuous. Also, taking $\mu\otimes\nu\in\Gamma(\mu,\nu)$ as coupling, we obtain

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

   Thus Theorem 2.3.2 ensures the existence of a nontrivial optimal transport plan $\bar{\gamma}$.

2. Since $\bar\gamma$ is optimal, we know from Corollary 2.5.7 that $\mathrm{supp}(\bar\gamma) \subset \partial\varphi$ for some convex $\varphi\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$. Also, by Proposition 2.5.4,

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

where $\varphi^*(y) \coloneqq \sup_{z\in\mathbb{R}^d}\{z\cdot y - \varphi(z)\}$. Therefore

$$\varphi(x) + \varphi^*(y) = x\cdot y \qquad \text{on } \mathrm{supp}(\bar\gamma).$$

In particular $(\varphi(x), \varphi^*(y))$ is finite for $\bar\gamma$-a.e. (x,y), and thus $\varphi(x)$ is finite at μ-a.e. point x. Since $\mu \ll dx$, and convex functions are differentiable a.e. on the region where they are finite (this follows by Alexandrov's Theorem, see [Vil09, Theorem 14.25]), we deduce that φ is differentiable μ-a.e.

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

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

which implies that

$$\Phi_{\bar x}(z) \coloneqq \varphi(z)-\varphi(\bar x)-\langle \bar y, z-\bar x\rangle \geq 0, \qquad \text{with equality at } z=\bar x.$$

Since φ is differentiable at $\bar x$, so is $\Phi_{\bar x}$. Hence, since $\Phi_{\bar x}$ has a minimum at $\bar x$, we deduce that

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

Therefore, we proved that

$$\bar y = \nabla\varphi(\bar x) \qquad \text{for all } \bar x\in\mathbb{R}^d\setminus A \text{ and } (\bar x,\bar y)\in\mathrm{supp}(\bar\gamma),$$

or, in other words,

$$\mathrm{supp}(\bar\gamma)\cap[(\mathbb{R}^d\setminus A)\times\mathbb{R}^d] \subset \mathrm{graph}(\nabla\varphi).$$

Since $\bar\gamma(A\times\mathbb{R}^d)=\mu(A)=0$, this proves that

$$(x,y) = (x,\nabla\varphi(x)) \qquad \bar\gamma\text{-a.e.} \tag{2.14}$$

Thus, for any function $F\in C_b(\mathbb{R}^d\times\mathbb{R}^d)$ we have

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^d} F(x,y)\,d\bar\gamma(x,y) &\overset{(2.14)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} F(x,\nabla\varphi(x))\,d\bar\gamma(x,y)\\ &= \int_{\mathbb{R}^d} F(x,\nabla\varphi(x))\,d\mu(x)\\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} F(x,y)\,d((\mathrm{Id}\times\nabla\varphi)_\#\mu)(x,y),\end{aligned}$$

hence $\bar\gamma = (\mathrm{Id}\times\nabla\varphi)_\#\mu$, as desired.

4. We now prove uniqueness. Assume that $\bar\gamma_1$ and $\bar\gamma_2$ are optimal. By linearity of the problem (and convexity of the constraints) also $\frac{\bar\gamma_1+\bar\gamma_2}{2}$ is optimal; indeed

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|^2 d\left(\frac{\bar\gamma_1+\bar\gamma_2}{2}\right) = \frac12\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|^2 d\bar\gamma_1 + \frac12\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|^2 d\bar\gamma_2$$

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

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \psi(x)\, d\left(\frac{\bar\gamma_1+\bar\gamma_2}{2}\right) = \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d} \psi(x)\, d\bar\gamma_1 + \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d} \psi(x)\, d\bar\gamma_2 = \int_{\mathbb{R}^d} \psi\, d\mu,$$

thus $(\pi_X)_\#(\frac{\bar\gamma_1+\bar\gamma_2}{2}) = \mu$. Analogously $(\pi_Y)_\#(\frac{\bar\gamma_1+\bar\gamma_2}{2}) = \nu$.

Hence, by Steps 2 and 3 applied to $\bar\gamma_1, \bar\gamma_2, \frac{\bar\gamma_1+\bar\gamma_2}{2}$, there exist three convex functions $\varphi_1, \varphi_2, \bar\varphi$ such that:

(i) $\bar\gamma_1 = (\mathrm{Id}\times\nabla\varphi_1)_\#\mu$, thus $(x,y) = (x, \nabla\varphi_1(x))$ $\bar\gamma_1$-a.e.;
(ii) $\bar\gamma_2 = (\mathrm{Id}\times\nabla\varphi_2)_\#\mu$, thus $(x,y) = (x, \nabla\varphi_2(x))$ $\bar\gamma_2$-a.e.;
(iii) $\frac{\bar\gamma_1+\bar\gamma_2}{2} = (\mathrm{Id}\times\nabla\bar\varphi)_\#\mu$, thus $(x,y) = (x, \nabla\bar\varphi(x))$ $\frac{\bar\gamma_1+\bar\gamma_2}{2}$-a.e.

In particular, it follows by (iii) that $(x,y) = (x, \nabla\bar\varphi(x))$ holds $\bar\gamma_1$-a.e., that combined with (i) yields

$$(x, \nabla\varphi_1(x)) = (x, \nabla\bar\varphi(x)) \quad \bar\gamma_1\text{-a.e.} \quad\Longrightarrow\quad \nabla\varphi_1(x) = \nabla\bar\varphi(x) \quad \mu\text{-a.e.},$$

where the implication follows from the fact that there is no dependence on y (so a relation true $\bar\gamma_1$-a.e. is also true μ-a.e.). Analogously, combining (ii) and (iii), we deduce that $\nabla\varphi_2(x) = \nabla\bar\varphi(x)$ holds for μ-a.e $x \in \mathbb{R}^d$. Thus $\nabla\varphi_1 = \nabla\varphi_2$ μ-a.e., and therefore $\bar\gamma_1 = \bar\gamma_2$, as desired.

□

Corollary 2.5.10. *Under the assumptions of Brenier's Theorem (Theorem 2.5.9):*

1. *There exists a unique optimal transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\#\mu = \nu$. Also, $T = \nabla\varphi$ with $\varphi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ convex.*
2. *If $S_\#\mu = \nu$ and $S = \nabla\phi$ μ-a.e. for some $\phi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ convex, then S is the unique optimal transport map.*

Proof. As we shall see, the proof is an immediate consequence of the previous results.

1. First of all, recall that the infimum in Monge's problem is bounded from below by the infimum in Kantorovich's problem (see Remark 2.2.1):

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

   Let φ be the convex function provided by Theorem 2.5.9, and set $\bar\gamma := (\mathrm{Id}\times\nabla\varphi)_\#\mu$. With this choice we have

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

   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. Then, it follows by the discussion above that $\gamma_1 = (\mathrm{Id}\times T_1)_\#\mu$ and $\gamma_2 = (\mathrm{Id}\times T_2)_\#\mu$ are optimal for Kantorovich. Because $\gamma_1 = \gamma_2$ (by Theorem 2.5.9), we conclude that $T_1 = T_2$ μ-a.e.

2. The optimality of S follows directly from Remark 2.5.8, while the uniqueness follows from the first part of this corollary.

□

We conclude this section by proving that, whenever μ and ν are both absolutely continuous, the optimal transport from μ to ν is invertible, and its inverse is given by the optimal transport map from ν to μ.

Corollary 2.5.11. *Under the assumptions of Brenier's Theorem (Theorem 2.5.9), assume also that $\nu \ll dx$. Let $\nabla\varphi$ be the optimal transport map from μ to ν, and let $\nabla\psi$ be the optimal transport map from ν to μ. Then $\nabla\varphi$ is invertible μ-a.e., and its inverse is unique ν-a.e. and given by $\nabla\psi$.*

Proof. By Brenier's Theorem (Theorem 2.5.9), we have two convex maps φ and ψ such that

- $\nabla\varphi$ is an optimal transport map from μ to ν;
- $\nabla\psi$ is an optimal transport map from ν to μ.

Hence

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

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

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

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

$$\begin{aligned}\int_{\mathbb{R}^d} F(x,\nabla\varphi(x))\,d\mu(x) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} F(x,y)\,d((\mathrm{Id}\times\nabla\varphi)_\#\mu)(x,y)\\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} F(x,y)\,d((\nabla\psi\times\mathrm{Id})_\#\nu)(x,y) = \int_{\mathbb{R}^d} F(\nabla\psi(y),y)\,d\nu(y).\end{aligned}$$

Choosing $F(x,y) = |x-\nabla\psi(y)|^2$, this gives

$$\int_{\mathbb{R}^d} |x-\nabla\psi(\nabla\varphi(x))|^2 d\mu(x) = \int_{\mathbb{R}^d} |\nabla\psi(y) - \nabla\psi(y)|^2 d\nu(y) = 0,$$

thus $\nabla\psi\circ\nabla\varphi = \mathrm{Id}$ μ-a.e. Similarly, choosing $F(x,y) = |\nabla\varphi(x)-y|^2$, we get $\nabla\varphi\circ\nabla\psi = \mathrm{Id}$ ν-a.e. □

Remark 2.5.12. Using the definition of Legendre transform, one can show that

$$y\in\partial\varphi(x) \quad\Leftrightarrow\quad x\in\partial\varphi^*(y)$$

and that the map ψ provided by the previous corollary actually coincides with φ^*. Since this will never be used in this book, we shall not prove it, but the interested reader is encouraged to try to prove this fact.

2.5.4 An application to Euler equations

Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with smooth boundary, and let ν be the outer unit normal to $\partial\Omega$. The Euler equations describe the evolution on a time interval $[0,T]$ of the velocity $v = v(t,x) \in \mathbb{R}^d$ of an incompressible fluid. They are given by the following system:

$$\begin{cases} \partial_t v + (v\cdot\nabla)v + \nabla p = 0 \text{ in } \Omega & \text{(Euler equation)} \\ \operatorname{div}(v) = 0 \text{ in } \Omega & \text{(Incompressibility condition)} \\ v\cdot\nu = 0 \text{ on } \partial\Omega & \text{(No-flux condition).} \end{cases}$$

Here $p = p(t,x) \in \mathbb{R}$ denotes the pressure of the fluid at time t and position x.

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

$$\partial_t v^i + \sum_{j=1}^d v^j \partial_{x_j} v^i + \partial_{x_i} p = 0 \qquad \forall\, i = 1, \dots, d.$$

If v is smooth, then

$$\begin{aligned} \frac{d}{dt}\int_\Omega |v(t)|^2 &= \frac{d}{dt}\int_\Omega \sum_{i=1}^d v^i(t)^2 = 2\int_\Omega \sum_{i=1}^d v^i \partial_t v^i \\ &= -2\int_\Omega \sum_{i,j=1}^d v^i v^j \partial_{x_j} v^i - 2\int_\Omega \sum_{i=1}^d v^i \partial_{x_i} p \\ &= -\int_\Omega \sum_j v^j \partial_{x_j}\left(\sum_{i=1}^d (v^i)^2\right) - 2\int_\Omega \sum_{i=1}^d v^i \partial_{x_i} p \\ &= -\int_{\partial\Omega} \sum_j v^j \nu^j \left(\sum_{i=1}^d (v^i)^2\right) + \int_\Omega \sum_j \partial_{x_j} v^j \left(\sum_{i=1}^d (v^i)^2\right) \\ &\quad - 2\int_{\partial\Omega} \sum_{i=1}^d v^i \nu^i p + 2\int_\Omega \sum_{i=1}^d \partial_{x_i} v_i\, p \\ &= -\int_{\partial\Omega} v\cdot\nu\ |v|^2 + \int_\Omega \operatorname{div}(v)\ |v|^2 - 2\int_{\partial\Omega} v\cdot\nu\ p + 2\int_\Omega \operatorname{div}(v)\ p \\ &= 0, \end{aligned}$$

where we used the no-flux and incompressibility conditions.

Also, if v is smooth, we can define its flow $g\colon [0,T]\times\Omega \to \Omega$ 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 Ω to Ω, since (thanks to the no-flux condition) the curve $t \mapsto g(t,x)$ never exits Ω. Also, differentiating the ODE for g with respect to x, we get

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

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

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

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

$$\begin{aligned}
\frac{d}{dt}\det(\nabla_x g(t,x)) &= \lim_{\varepsilon\to 0} \frac{\det(\nabla_x g(t+\varepsilon,x)) - \det(\nabla_x g(t,x))}{\varepsilon} \\
&= \lim_{\varepsilon\to 0} \frac{\det\big(\nabla_x g(t,x) + \varepsilon \nabla_x v(t,g(t,x))\cdot\nabla_x g(t,x) + o(\varepsilon)\big) - \det(\nabla_x g(t,x))}{\varepsilon} \\
&= \lim_{\varepsilon\to 0} \frac{\det\big(\nabla_x g(t,x) + \varepsilon \nabla_x v(t,g(t,x))\cdot\nabla_x g(t,x)\big) - \det(\nabla_x g(t,x)) + o(\varepsilon)}{\varepsilon} \\
&= \lim_{\varepsilon\to 0} \frac{\det\big(\mathrm{Id} + \varepsilon \nabla_x v(t,g(t,x))\big)\det(\nabla_x g(t,x)) - \det(\nabla_x g(t,x))}{\varepsilon} \\
&= \lim_{\varepsilon\to 0} \frac{[1 + \varepsilon \operatorname{tr}(\nabla_x v)(t,g(t,x)) - 1]\det(\nabla_x g(t,x))}{\varepsilon} \\
&= \operatorname{tr}(\nabla_x v)(t,g(t,x))\det(\nabla_x g(t,x)) \\
&= \operatorname{div}(v)(t,g(t,v))\det(\nabla_x g(t,x)) = 0,
\end{aligned} \tag{2.15}$$

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 $g = (g^1, \dots, g^d)$, using the Euler equations we get

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

Thus the Euler equations are equivalent to the following system for a curve $t \mapsto g(t)$ of smooth diffeomorphisms of Ω:

$$\begin{cases}
\partial_{tt} g = -\nabla p(t,g) & \text{(Euler equation } \to \text{ 2nd order ODE for } g) \\
\det \nabla_x g = 1 & \text{(Incompressibility } \to g \text{ preserves the Lebesgue measure)} \\
g(0,x) = x & \text{(Initial condition),}
\end{cases} \tag{2.16}$$

where $p(t) : \Omega \to \mathbb{R}$ is some function that represents the pressure.

Arnold's Theorem. It was observed by Arnold in the 1960's [Arn66] that, at least formally, the Euler equations for fluids can be seen as a geodesic curve in an appropriate infinite-dimensional manifold. That the reader can find the definition of geodesic, and a brief presentation of the concepts necessary to appreciate the following theorem, in Section 1.3.

Theorem 2.5.13 (Arnold's Theorem)**.** *The Euler equations are equivalent to the geodesic equation on the manifold* $\mathrm{SDiff}(\Omega) \subset L^2(\Omega;\mathbb{R}^d)$ *defined as*

$$\mathrm{SDiff}(\Omega) := \{h\colon \Omega \to \Omega \mid h \text{ measure preserving and orientation preserving diffeomorphism}\}.$$

Proof. First of all, we need to identify the tangent space of $\mathrm{SDiff}(\Omega)$.

Given $\bar h \in \mathrm{SDiff}(\Omega)$, let $t \mapsto h(t) \in \mathrm{SDiff}(\Omega)$ be a smooth curve of maps in $\mathrm{SDiff}(\Omega)$ with $h(0) = \bar h$, and set $w(t) := \partial_t h(t)$. By definition of tangent space, $w(t) \in T_{h(t)}\mathrm{SDiff}(\Omega)$.

Since $h(t)$ is a diffeomorphism of Ω, it maps $\partial\Omega$ onto itself, and therefore $w(t) = \partial_t h(t)$ must be tangent to the boundary. Define $\tilde w(t) := w(t)\circ h^{-1}(t)$ so that $\partial_t h(t) = \tilde w(t, h(t))$, and note that $\tilde w(t)$ is also tangent to $\partial\Omega$. Since $\det\nabla_x h(t,x) \equiv 1$ (because $h(t) \in \mathrm{SDiff}(\Omega)$), by the computations in (2.15) we have

$$0 = \frac{d}{dt}\det\nabla_x h(t,x) = \operatorname{div}(\tilde w)(t,h(t,x))\underbrace{\det\nabla_x h(t,x)}_{\equiv 1} \quad\Longrightarrow\quad \operatorname{div}(\tilde w) = 0.$$

Thus, taking $t = 0$, we deduce that

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

Viceversa, given a vector field $\tilde w : \Omega \to \mathbb{R}^d$ with $\mathrm{div}(\tilde w) = 0$ and $\tilde w \cdot \nu|_{\partial\Omega} = 0$, we solve

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

and using the same computation as in (2.15) we find that $\frac{d}{dt} \det \nabla h = 0$. Thus $h(t) : \Omega \to \Omega$ is a curve in $\mathrm{SDiff}(\Omega)$, and in particular $\partial_t h(0) = \tilde w \circ \bar h$ is an element of the tangent space of $\mathrm{SDiff}(\Omega)$ at $\bar h$.

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

$$T_{\bar h}\mathrm{SDiff}(\Omega) = \{\tilde w \circ \bar h \mid \mathrm{div}(\tilde w) = 0,\ \tilde w \cdot \nu|_{\partial\Omega} = 0\}.$$

Let us observe that:

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

$$\langle f_1 \circ h, f_2 \circ h\rangle_{L^2} = \int_\Omega f_1 \circ h(x) \cdot f_2 \circ h(x)\, dx = \int_\Omega f_1(x) \cdot f_2(x)\, dx = \langle f_1, f_2\rangle_{L^2},$$

where in the second equality we used that $h \in \mathrm{SDiff}(\Omega)$ (and therefore $h_\# dx = dx$).

(b) Every vector field in $L^2(\Omega, \mathbb{R}^d)$ can be written as the sum of a gradient and a divergence-free vector field, that is

$$L^2(\Omega, \mathbb{R}^d) := \{w \colon \Omega \to \mathbb{R}^d \mid \mathrm{div}(w) = 0 \text{ and } w \cdot \nu|_{\partial\Omega} = 0\} \oplus \{\nabla q \mid q \colon \Omega \to \mathbb{R}\}. \tag{2.17}$$

Note that this decomposition is orthogonal. Indeed

$$\langle w, \nabla q\rangle_{L^2} = \int_\Omega w \cdot \nabla q\, dx = -\int_{\partial\Omega} \underbrace{w \cdot \nu}_{=0}\, q - \int_\Omega \underbrace{\mathrm{div}(w)}_{=0}\, q\, dx = 0.$$

This is known as *Helmholtz decomposition.*

Combining (a) and (b) yields that, for any $h \in \mathrm{SDiff}(\Omega)$, we can decompose $L^2(\Omega, \mathbb{R}^d)$ as

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

Since this decomposition is orthogonal in $L^2(\Omega, \mathbb{R}^d)$, we conclude that, given $h \in \mathrm{SDiff}(\Omega)$,

$$\big(T_h\mathrm{SDiff}(\Omega)\big)^\perp = \{\nabla q \circ h \mid q \colon \Omega \to \mathbb{R}^d\}.$$

Hence, thanks to this characterization and recalling Definition 1.3.5, given a curve $t \to g(t) \in \mathrm{SDiff}(\Omega)$, the following are equivalent:

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

Recalling (2.16), this proves the result taking $p(t) := -q(t)$. □

A connection between Arnold's and Brenier's Theorems. Thanks to Arnold's Theorem, we know that the incompressible Euler equations correspond to the geodesic equations in the space $\mathrm{SDiff}(\Omega)$. We now recall that minimizing geodesics on manifolds can be found by considering the minimization problem (1.2). Thus, to find minimizing geodesics in $\mathrm{SDiff}(\Omega)$, one could consider the minimization problem

$$\inf\left\{\int_0^1 \int_\Omega |\partial_t g(t,x)|^2 \, dx \, dt \mid g(t) \in \mathrm{SDiff},\, g(0) = g_0,\, g(1) = g_1\right\},$$

where $g_0, g_1 \in \mathrm{SDiff}(\Omega)$ are prescribed.

This minimization problem is very challenging and actually minimizers may fail to exist (see for instance [DF13]). Thus, we consider a simpler version of the problem. Namely, instead of searching the minimizing geodesic from g_0 to g_1, we look for an *approximate midpoint* between them.

Recall that, given a d-dimensional manifold $M \subset \mathbb{R}^D$, and two points $x_0, x_1 \in M$, a good approximation of the midpoint between them is found by considering the Euclidean midpoint $\frac{x_0+x_1}{2}$ (note that this point may not belong to M) and then finding the closest point to $\frac{x_0+x_1}{2}$ on M, that is $\mathrm{proj}_M\left(\frac{x_0+x_1}{2}\right)$. By analogy, given $g_0, g_1 \in \mathrm{SDiff}(\Omega)$, one looks for the closest function (with respect to the L^2 norm) in $\mathrm{SDiff}(\Omega)$ to $\frac{g_0+g_1}{2}$. Thus, we want to study the map

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

Even this simpler problem is far from trivial, the main difficulty being that $\mathrm{SDiff}(\Omega)$ is neither convex nor closed in $L^2(\Omega, \mathbb{R}^d)$. So, as a first relaxation of the problem, one might want to consider the L^2-closure of $\mathrm{SDiff}(\Omega)$. This closure is characterized in the next (nontrivial) result due to Brenier and Gangbo [BG03].

Theorem 2.5.14. *Let $\Omega \subset \mathbb{R}^d$ be a bounded set with Lipschitz boundary, and let $d \geq 2$. Then*

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

The next result gives a sufficient condition for the existence and uniqueness of the projection of a map $h \in L^2(\Omega, \mathbb{R}^d)$ in $S(\Omega)$.

Theorem 2.5.15 ([Bre87])**.** *Let $h \in L^2(\Omega; \mathbb{R}^d)$ satisfy $h_\#(dx|_\Omega) \ll dx$. Then:*

(i) There exists a unique projection $\bar{s}$ onto $S(\Omega)$ (i.e., for any $s \in S(\Omega)$ it holds $\|h - \bar{s}\|_{L^2(\Omega)} \leq \|h - s\|_{L^2(\Omega)}$).

(ii) There exists a convex function ψ such that $h = \nabla\psi \circ \bar{s}$ (this formula is called polar decomposition*, see Remark 2.5.16).*

Proof. We split the proof in three steps: in Steps 1 and 2 we prove (i) (with Step 1 for the existence, and Step 2 for the uniqueness), and in Step 3 we prove (ii).

1. 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 (Theorem 2.5.9) holds for probability measures, up to multiplying both μ and $dx|_\Omega$ by $\frac{1}{|\Omega|}$, we can apply it also in this context. Thus, by Corollary 2.5.11, there exist convex functions $\varphi, \psi \colon \mathbb{R}^d \to \mathbb{R}$ such that $\nabla\varphi$ and $\nabla\psi$ are optimal from μ to $dx|_\Omega$ and viceversa.

Let $\bar{s} \coloneqq \nabla\varphi \circ h \colon \Omega \to \Omega$. Then, by the optimality of $\nabla\varphi$,

$$\begin{aligned} \int_\Omega |\bar{s}(x) - h(x)|^2 dx = \int_\Omega |\nabla\varphi \circ h - h|^2 \, dx = \int_{\mathbb{R}^d} |\nabla\varphi - \mathrm{Id}|^2 d\mu \\ = \min_{\gamma \in \Gamma(\mu, dx|_\Omega)} \int_{\mathbb{R}^d \times \Omega} |x - y|^2 d\gamma. \end{aligned} \tag{2.18}$$

We now observe that if $s \in S(\Omega)$ then $\gamma_s \coloneqq (h \times s)_\#(dx|_\Omega)$ belongs to $\Gamma(\mu, dx|_\Omega)$. Indeed, $(\pi_X)_\#\gamma_s = h_\#(dx|_\Omega) = \mu$ and $(\pi_Y)_\#\gamma_s = s_\#(dx|_\Omega) = dx|_\Omega$. This implies that

$$\min_{\gamma \in \Gamma(\mu, dx|_\Omega)} \int_{\mathbb{R}^d \times \Omega} |x - y|^2 d\gamma \le \min_{s \in S(\Omega)} \int_{\mathbb{R}^d \times \Omega} |x - y|^2 d\gamma_s = \min_{s \in S(\Omega)} \int_\Omega |h(x) - s(x)|^2 \, dx,$$

that combined with (2.18) yields

$$\int_\Omega |h(x) - \bar{s}(x)|^2 dx \le \min_{s \in S(\Omega)} \int_\Omega |h(x) - s(x)|^2 \, dx,$$

thus $\bar{s}$ is a projection.

2. Suppose that $\hat{s}$ is another projection. Then by the previous step it follows that $\gamma_{\bar{s}}$ and $\gamma_{\hat{s}}$ are both optimal couplings. Thus, by uniqueness (see Theorem 2.5.9) the transport plans are equal, therefore

$$\int_\Omega F(h(x), \hat{s}(x)) \, dx = \int_\Omega F(h(x), \bar{s}(x)) \, dx \qquad \forall \, F \in C_b(\mathbb{R}^d \times \mathbb{R}^d).$$

Choosing $F(x, y) = |\nabla\varphi(x) - y|^2$ and recalling that $\bar{s} = \nabla\varphi \circ h$, we conclude that

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

hence $\hat{s} = \bar{s}$, as desired.

3. This follows from the fact that $\bar{s} = \nabla\varphi \circ h$ and $\nabla\psi = (\nabla\varphi)^{-1}$ (see Corollary 2.5.11).

□

Remark 2.5.16. The polar decomposition can be seen (at least formally) as a generalization of some well-known results:

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

(b) Consider a smooth vector field $w : \mathbb{R}^d \to \mathbb{R}^d$, and let $h_t(x) \coloneqq h(t, x)$ be the flow of w:

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

Then

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

Also, the polar decomposition of h_ε yields

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

At least formally, since $h_\varepsilon(x)$ is a perturbation of x, it looks natural to assume that also $\nabla\psi_\varepsilon$ and s_ϵ are perturbations of the identity map. More precisely, we suppose that

$$\psi_\varepsilon(x) = \frac{|x|^2}{2} + \varepsilon\, q(x) + o(\varepsilon), \qquad s_\varepsilon(x) = x + \varepsilon\, u(x) + o(\varepsilon).$$

Also, since

$$\det \nabla s_\varepsilon = \det(\mathrm{Id} + \varepsilon \nabla u + o(\varepsilon)) = 1 + \varepsilon \operatorname{div}(u) + o(\varepsilon)$$

and s_ε is measure preserving (hence $1 \equiv \det \nabla s_\varepsilon$), we deduce that $\operatorname{div}(u) = 0$. Hence, combining all these equations, we get

$$\begin{aligned} x + \varepsilon\, w(x) + o(\varepsilon) = h_\varepsilon &= \nabla\psi_\varepsilon \circ s_\varepsilon \\ &= (x + \varepsilon \nabla q(x)) \circ (x + \varepsilon u(x)) + o(\varepsilon) \\ &= x + \varepsilon(u(x) + \nabla q(x)) + o(\varepsilon), \end{aligned}$$

therefore $w = u + \nabla q$. In other words, this formally shows that any vector field w can be written as the sum of a divergence free vector field and a gradient, which is nothing but the Helmholtz decomposition (2.17). Thus, morally speaking, Helmholtz decomposition is the infinitesimal version of the polar decomposition.

2.6 General cost functions: Kantorovich duality

The goal of this section is to repeat, in the case of general costs, what we did in the previous sections for the case $c(x, y) = -x \cdot y$ on $X = Y = \mathbb{R}^d$. As we shall see, some proofs are essentially identical provided that one introduces the correct definitions.

2.6.1 c-convexity and c-cyclical monotonicity

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 $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex if

$$\phi(x) = \sup_{y \in \mathbb{R}^d} \{x \cdot y + \lambda_y\}$$

for some choice of values $\{\lambda_y\}_{y\in\mathbb{R}^d}$ with $\lambda_y \in \mathbb{R} \cup \{-\infty\}$.[11] Having in mind that before $x \cdot y = -c(x, y)$, this suggests the following general definition (cp. Definition 2.5.1):

Definition 2.6.1. Given X and Y metric spaces, $c\colon X \times Y \to \mathbb{R}$, and $\varphi\colon X \to \mathbb{R} \cup \{+\infty\}$, we say that φ is *c-convex* if

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

for some $\{\lambda_y\}_{y\in Y} \subset \mathbb{R} \cup \{-\infty\}$.

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

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

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

[11] If one wants to avoid setting $\lambda_y = -\infty$ for some y, one can instead say that a function φ is convex if there exist $A \subset \mathbb{R}^d$ and a family $\{\lambda_y\}_{y\in A} \subset \mathbb{R}$ such that

$$\varphi(x) = \sup_{y\in A}\{x \cdot y + \lambda_y\}.$$

In other words, A corresponds to the set $\{\lambda_y > -\infty\}$.

Remark 2.6.2. When $X = Y$ is a metric space and the cost is the distance $c(x, y) = d(x, y)$, it turns out that a function $\varphi : X :\to \mathbb{R} \cup \{+\infty\}$ is c-convex if and only if it is 1-Lipschitz, i.e., it holds

$$|\varphi(x) - \varphi(y)| \leq d(x, y) \qquad \forall\, x, y \in X.$$

Indeed, if φ is 1-Lipschitz, then $\varphi(x) \geq \varphi(y) - d(x, y)$ with equality for $y = x$, hence

$$\varphi(x) = \sup_{y \in X} \{\varphi(y) - d(x, y)\} \tag{2.19}$$

which shows that φ is c-convex (with the choice $\lambda_y = \varphi(y)$).

Viceversa, assume that φ is c-convex and fix $x, y \in X$. By definition of c-convexity, For any $\varepsilon > 0$ there exists $z \in X$ such that

$$\varphi(x) \leq -d(x, z) + \lambda_z + \varepsilon\,.$$

Moreover, once again by definition of c-convexity, we have

$$\varphi(y) \geq -d(y, z) + \lambda_z\,.$$

Combining theSE two inequalities and applying the triangle inequality, we obtain

$$\varphi(x) - \varphi(y) \leq d(y, z) - d(x, z) + \varepsilon \leq d(x, y) + \varepsilon.$$

Since $\varepsilon > 0$ can be chosen arbitrarily small, and the role of x and y can be exchanged, this latter inequality implies that φ is 1-Lipschitz.

The following is the analogue of Theorem 2.5.2.

Theorem 2.6.3. *A set $S \subset X \times Y$ is c-cyclically monotone if and only if there exists a c-convex function φ such that $S \subset \partial_c \varphi$.*

Proof. The proof is essentially the same as the one of Theorem 2.5.2, provided one replaces $-x \cdot y$ with $c(x, y)$. We write the details just for one implication.

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

$$\varphi(z) \geq \varphi(x_i) - c(z, y_i) + c(x_i, y_i) \qquad \forall\, z \in X.$$

Choosing $z = x_{i+1}$ (with the convention $x_{N+1} = x_1$) and summing over i, we obtain

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

$\Rightarrow$ It suffices to define

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

and repeat the proof of Theorem 2.5.2 (see [Vil09, pp. 77–78] for the details).

□

2.6.2 A general Kantorovich duality

In this section, X and Y are locally compact, separable and complete metric spaces.

Definition 2.6.4. Given a c-convex function $\varphi\colon X \to \mathbb{R} \cup \{+\infty\}$, we define its *$c$-Legendre transform* $\varphi^c\colon Y \to \mathbb{R} \cup \{+\infty\}$ as

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

As in the case of the classical Legendre transform, φ and φ^c are related by several interesting properties.

Proposition 2.6.5. *The following properties hold:*

(a) $\varphi(x) + \varphi^c(y) + c(x,y) \geq 0$ *for all* $x \in X$, $y \in Y$;

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

Proof. The proof is identical to the one of Proposition 2.5.4 and is left to the interested reader. □

One then obtains the following general duality result (recall also Remark 2.3.3).

Theorem 2.6.6 (Kantorovich dualiy: General case)**.** *Let $c \in C^0(X \times Y)$ be bounded from below, and assume that $\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma < +\infty$. Then*

$$\min_{\gamma \in \Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma = \max_{\varphi(x)+\psi(y)+c(x,y)\geq 0} \int_X -\varphi\, d\mu + \int_Y -\psi\, d\nu.$$

Proof. Again, the steps are the same as in the proof of Theorem 2.5.5, just replacing convexity with c-convexity, subdifferential with c-subdifferential, etc. We refer the reader to [Vil09, pp. 78–79] for the details. □

Remark 2.6.7. When $X = Y$ is a metric space, and $c(x,y) = d(x,y)$ is the distance function, then it follows from Remark 2.6.2 that φ is c-convex if and only if it is 1-Lipschitz. Also, using (2.19) with $-\varphi$ in place of φ, we deduce that $\varphi^c(y) = -\varphi(y)$. Combining these facts with Theorem 2.6.6, we obtain the following the following important duality result relating 1-Lipschitz functions and the Kantorovich problem for the case "cost=distance":
Let $X = Y$ be a metric space, and let $c(x,y) = d(x,y)$ be the distance function. Assume that $\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X\times X} d(x,y)\, d\gamma < +\infty$. Then

$$\min_{\gamma \in \Gamma(\mu,\nu)} \int_{X\times X} d(x,y)\, d\gamma = \max_{\varphi\ 1\text{-}Lipschitz} \int_X \varphi\, d\mu - \int_X \varphi\, d\nu.$$

Remark 2.6.8. A popular alternative approach to Kantorovich duality is based on general abstract results in convex analysis, and goes as follows:

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

where:

- $\heartsuit$ one should note that we do not require γ 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 φ is $+\infty$ if $(\pi_X)_\#\gamma \neq \mu$ (resp. the supremum over ψ 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 [Vil03, Theorem 1.9] to exchange inf and sup.
- $\diamondsuit$ we have the two following possible situations:
  - (i) 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 $\gamma \equiv 0$).
  - (ii) 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 γ is $-\infty$.

We refer the reader to [Vil03, Chapter 1] for a detailed discussion of this approach.

We conclude this section by noticing that, as a consequence of the previous results, we obtain the following corollary (cf. Corollary 2.5.7):

Corollary 2.6.9. *Let $c \in C^0(X\times Y)$ be bounded from below, and assume $\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{X\times Y} c\,d\gamma < +\infty$. For a coupling $\bar{\gamma}\in\Gamma(\mu,\nu)$, the following statements are equivalent:*

(i) $\bar{\gamma}$ is optimal;

(ii) $\mathrm{supp}(\bar{\gamma})$ is c-cyclically monotone;

(iii) there exists a c-convex function $\varphi : X \to \mathbb{R}\cup\{+\infty\}$ such that $\mathrm{supp}(\bar{\gamma}) \subset \partial^c\varphi$.

2.7 General cost functions: existence and uniqueness of optimal transport maps

Thanks to the previous results, we can now mimic the proof of Brenier's Theorem (Theorem 2.5.9) to prove the existence and uniqueness of optimal transport maps. Since the proof of Theorem 2.5.9 involves taking derivatives, one needs X to have a differentiable structure. For simplicity we shall prove the result when $X = Y = \mathbb{R}^d$, but the same argument can be generalized to the case when X is an arbitrary Riemannian manifold and Y has no differentiable structure. Also, to simplify the argument, we shall assume that $\mathrm{supp}(\nu)$ is compact. For more general statements, we refer to [Vil09, Chapters 9-10].

Theorem 2.7.1. *Let $X = Y = \mathbb{R}^d$, $\mu \ll dx$, and $\mathrm{supp}(\nu)$ compact. Let $c \in C^0(X \times Y)$ be bounded from below, and assume that $\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma < +\infty$. Also, suppose that:*

- *for every $y \in \mathrm{supp}(\nu)$, the map $\mathbb{R}^d \ni x \mapsto c(x,y)$ is differentiable;*
- *for every $x \in \mathbb{R}^d$, the map $\mathrm{supp}(\nu) \ni y \mapsto \nabla_x c(x,y) \in \mathbb{R}^d$ is injective;*
- *for every $y \in \mathrm{supp}(\nu)$ and $R > 0$, $|\nabla_x c(x,y)| \le C_R$ for every $x \in B_R$.*

Then there exists a unique optimal coupling $\bar\gamma$, 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 c-convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$.

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, c-convex functions are the same as convex functions, and

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

implies that

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

therefore $T = \nabla\varphi$. Hence, Theorem 2.7.1 covers Theorem 2.5.9 (the only extra assumption is that now $\mathrm{supp}(\nu)$ is assumed to be compact).

Proof. Let $\bar\gamma$ be optimal, let φ be as in Corollary 2.6.9(iii), and define φ^c as in Definition 2.6.4. Note that, as a consequence of Corollary 2.6.9(iii) and Proposition 2.6.5(b), it follows that

$$\varphi(x) + \varphi^c(x) + c(x,y) = 0 \qquad \forall\, (x,y) \in \mathrm{supp}(\bar\gamma). \tag{2.20}$$

In the proof of Theorem 2.5.9 we used that convex functions are differentiable a.e. Here, in order to obtain the a.e. differentiability of φ, we would like to show that it is locally Lipschitz. In general this is not clear for φ itself, but we can show that we can replace it with another function $\tilde\varphi$ which is locally Lipschitz.

Indeed, define

$$\tilde\varphi(x) := \sup_{y\in\mathrm{supp}(\nu)} \{-c(x,y) - \varphi^c(y)\} = \sup_{y\in Y}\{-c(x,y) + \lambda_y\}, \qquad \lambda_y := \begin{cases} -\varphi_c(y) & \text{if } y \in \mathrm{supp}(\nu), \\ -\infty & \text{if } y \notin \mathrm{supp}(\nu). \end{cases}$$

Note that $\tilde\varphi$ is c-convex. Also, since $-c(x,y) - \varphi^c(y) \le \varphi(x)$ for every x, y (see Proposition 2.6.5(a)), we have $\tilde\varphi \le \varphi$.

On the other hand, it follows immediately from the definition of $\tilde\varphi$ that

$$\tilde\varphi(x) + \varphi^c(y) + c(x,y) \ge 0 \qquad \forall\, (x,y) \in \mathbb{R}^d \times \mathrm{supp}(\nu). \tag{2.21}$$

Hence, since $\mathrm{supp}(\bar\gamma) \subset \mathbb{R}^d \times \mathrm{supp}(\nu)$ (because $(\pi_Y)_\#\gamma = \nu$) and using (2.20), it follows that

$$0 \le \tilde\varphi(x) + \varphi^c(x) + c(x,y) \le \varphi(x) + \varphi^c(x) + c(x,y) = 0 \qquad \forall\, (x,y) \in \mathrm{supp}(\bar\gamma),$$

thus

$$\tilde\varphi(x) + \varphi^c(x) + c(x,y) = 0 \qquad \forall\, (x,y) \in \mathrm{supp}(\bar\gamma). \tag{2.22}$$

We now claim that $\tilde\varphi$ is locally Lipschitz. Indeed, for each $y \in \mathrm{supp}(\nu)$, consider the map

$$\mathbb{R}^d \ni x \mapsto -c(x,y) - \varphi^c(y).$$

The gradient is given by $-\nabla_x c(x,y)$, which (by our assumption) is uniformly bounded by C_R for $x \in B_R$. Thus, for any $R > 0$ the maps

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

are C_R-Lipschitz, and therefore also the map $\tilde{\varphi}$ (being their supremum) is C_R-Lipschitz inside B_R for any $R > 0$. This proves the claim.

Since locally Lipschitz maps are differentiable a.e., there exists a set A, with $|A| = 0$, such that $\tilde{\varphi}$ is differentiable on $\mathbb{R} \setminus A$. Also, since $\mu \ll dx$, we have $\mu(A) = 0$.

Now, fix $(x,y) \in \mathrm{supp}(\bar{\gamma})$ with $x \notin A$. Then it follows from (2.21) and (2.22) that the function

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

attains its minimum at $z = x$, therefore

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

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 this unique point $T(x)$. Hence, we proved that $\mathrm{supp}(\bar{\gamma}) \cap [(\mathbb{R}^d \setminus A) \times \mathrm{supp}(\nu)] \subset \mathrm{graph}(T)$. As in the proof of Theorem 2.5.9, since $\bar{\gamma}(A \times \mathrm{supp}(\nu)) \subset \bar{\gamma}(A \times \mathbb{R}^d) = \mu(A) = 0$ we conclude that $\bar{\gamma} = (\mathrm{Id} \times T)_{\#}\mu$.

Finally, uniqueness follows by the same argument as in Step 4 of the proof of Theorem 2.5.9. More precisely if γ_1 and γ_2 are optimal then so is $\frac{\gamma_1+\gamma_2}{2}$. Then

$$\mathrm{graph}(T_1) \cup \mathrm{graph}(T_2) \overset{a.e.}{=} \mathrm{supp}(\gamma_1) \cup \mathrm{supp}(\gamma_2) = \mathrm{supp}\left(\frac{\gamma_1+\gamma_2}{2}\right) \overset{a.e.}{=} \mathrm{graph}(\bar{T})$$

for some map $\bar{T}$, and this is only possible if $T_1 = T_2$ μ-a.e. □

Example 2.7.3. Let $c(x,y) = |x-y|^p$, with $p > 1$. We want to show that Theorem 2.7.1 applies. We only prove that the map $y \mapsto \nabla_x c(x,y)$ is injective, since the other assumptions on c are easily checked.

To show this, fix $x, v \in \mathbb{R}^d$ and assume that $v = \nabla_x c(x,y) = p|x-y|^{p-2}(x-y)$. Since $|x-y|^{p-2}$ is positive, we deduce that the vectors v and $(x-y)$ are parallel and point in the same direction, hence

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

We also know that

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

Combining these two facts, we deduce that 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.

Remark 2.7.4. We note that, for $p = 1$, the reasoning above fails. Indeed, given $x, v \in \mathbb{R}^d$, the relation

$$v = \nabla_x c(x,y) = \frac{x-y}{|x-y|}$$

implies that necessarily $|v| = 1$, and under this condition the relation is satisfied by every $y = x - tv$ with $t > 0$, which shows that y is not unique.

In fact, the previous theorem is false for the cost $c(x, y) = |x - y|$. To see this, consider $d = 1$ and take the measures $\mu = dx|_{[0,1]}$ and $\nu = dx|_{[1,2]}$. As shown in [Vil03, Remark 2.19(iii)] or [San15, Proposition 2.7], the optimal transportation cost between these two densities is given by

$$\int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|\, dx, \qquad \text{where } F_\sigma(x) := \sigma\big((-\infty, x]\big).$$

In this particular case, one can check that $\int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|\, dx = 1$.

Noticing that $T_1(x) := x + 1$ and $T_2(x) := 2 - x$ satisfy

$$(T_i)_\# \mu = \nu, \qquad \int_{\mathbb{R}} |T_i(x) - x|\, d\mu(x) = 1, \qquad i = 1, 2,$$

we deduce that both maps T_1 and T_2 are optimal, so we have no uniqueness. In addition, if we define $\gamma_i := (\mathrm{Id} \times T_i)_\# \mu$ $(i = 1, 2)$, then $\frac{\gamma_1 + \gamma_2}{2}$ is optimal and it is not induced by a graph.

Remark 2.7.5. Consider again the cost $c(x, y) = |x - y|$ on $\mathbb{R} \times \mathbb{R}$, and let $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Assume that

$$x \le y \qquad \text{for all } x \in \mathrm{supp}(\mu),\ y \in \mathrm{supp}(\nu).$$

Then for any coupling γ (resp., any transport map T) we have

$$x \le y \quad \forall\, (x, y) \in \mathrm{supp}(\gamma) \qquad (\text{resp. } x \le T(x) \quad \forall\, x \in \mathrm{supp}(\mu)).$$

Hence

$$\int_{\mathbb{R}\times\mathbb{R}} |y - x|\, d\gamma = \int_{\mathbb{R}\times\mathbb{R}} (y - x)\, d\gamma = \int_{\mathbb{R}\times\mathbb{R}} y\, d\gamma - \int_{\mathbb{R}\times\mathbb{R}} x\, d\gamma = \int_{\mathbb{R}} y\, d\nu - \int_{\mathbb{R}} x\, d\mu\,,$$

and analogously

$$\int_{\mathbb{R}} |T(x) - x|\, d\mu = \int_{\mathbb{R}} (T(x) - x)\, d\mu = \int_{\mathbb{R}} T(x)\, d\mu - \int_{\mathbb{R}} x\, d\mu = \int_{\mathbb{R}} y\, d\nu - \int_{\mathbb{R}} x\, d\mu\,.$$

In other words the cost is independent of the coupling or of the transport map, and therefore every coupling/map is optimal.

3 Wasserstein Distances and Gradient Flows

The goal of this section is to show a surprising connection between optimal transport, gradient flows, and PDEs. More precisely, after introducing the Wasserstein distances, we will first give a brief general introduction to gradient flows in Hilbert spaces. Then, following the seminal approach of Jordan, Kinderlehrer, and Otto [JKO98], we are going to prove that the gradient flow of the entropy functional in the Wasserstein space coincides with the heat equation.

Our general treatment of gradient flows gives only a glimpse of the theory. We suggest the expository paper [San17a] for an introduction to the subject and the monograph [AGS08] for a thorough study of gradient flows, both in the Hilbertian setting and in the vastly more general metric setting.

3.1 p-Wasserstein distances and geodesics

We are going to introduce the space of measures with finite p-moment and then a distance on this space induced by optimal transport.

Definition 3.1.1. Let (X, d) be a locally compact and separable metric space. Given $1 \leq p < \infty$, let

$$\mathcal{P}_p(X) \coloneqq \left\{ \sigma \in \mathcal{P}(X) \mid \int_X d(x, x_0)^p \, d\sigma(x) < +\infty \text{ for some } x_0 \in X \right\}, \tag{3.1}$$

be the set of probability measures with finite p-moment.

Remark 3.1.2. Given $x_1 \in X$, by the triangle inequality and the convexity of $\mathbb{R}^+ \ni s \mapsto s^p$ we get[12]

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

Hence, if $\sigma \in \mathcal{P}(X)$ satisfies $\int_X d(x, x_0)^p \, d\sigma(x) < +\infty$, then also $\int_X d(x, x_1)^p \, d\sigma(x)$ is finite. This means that the definition of $\mathcal{P}_p(X)$ is independent of the basepoint x_0.

Definition 3.1.3. Given $\mu, \nu \in \mathcal{P}_p(X)$, we define their *$p$-Wasserstein distance* as

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

Remark 3.1.4. If $\mu, \nu \in \mathcal{P}_p(X)$ then for all $\gamma \in \Gamma(\mu, \nu)$ it holds

$$\begin{aligned} \int_{X \times X} d(x, y)^p \, d\gamma &\leq 2^{p-1} \int_{X \times X} [d(x, x_0)^p + d(x_0, y)^p] d\gamma \\ &= 2^{p-1} \left[\int_X d(x, x_0)^p \, d\mu + \int_X d(y, x_0)^p \, d\nu \right] < \infty. \end{aligned}$$

Hence W_p is finite on $\mathcal{P}_p(X) \times \mathcal{P}_p(X)$.

To justify the terminology "p-Wasserstein distance", we now prove the following:

Theorem 3.1.5. *W_p is a distance on the space $\mathcal{P}_p(X)$.*

Proof. As we shall see, the most delicate part of the proof consists in proving the triangle inequality.

[12] By convexity, given $a, b \geq 0$ we have $\left(\frac{a+b}{2}\right)^p \leq \frac{a^p+b^p}{2}$, or equivalently $(a+b)^p \leq 2^{p-1}(a^p + b^p)$.

1. If $W_p(\mu,\nu)=0$, then (thanks to Theorem 2.3.2) 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., which means that γ is concentrated on the graph of the identity map. Therefore $\bar\gamma=(\mathrm{Id}\times\mathrm{Id})_\#\mu$, which yields $\nu=(\pi_2)_\#\bar\gamma=\mu$.

2. We now prove that W_p is symmetric. Indeed, given $\gamma\in\Gamma(\mu,\nu)$ optimal, define $\tilde\gamma:=S_\#\gamma$, with $S(x,y):=(y,x)$. Then $\tilde\gamma\in\Gamma(\nu,\mu)$ and therefore, since $d(x,y)=d(y,x)$, we get
$$W_p(\nu,\mu)\le\int_{X\times X} d(x,y)^p\,d\tilde\gamma=\int_{X\times X} d(x,y)^p\,d\gamma=W_p(\mu,\nu).$$
Exchanging the roles of μ and ν proves that $W_p(\nu,\mu)=W_p(\mu,\nu)$, as desired.

3. We now prove the triangle inequality. Let $\mu_1,\mu_2,\mu_3\in\mathcal{P}_p(X)$, and let $\gamma_{12}\in\Gamma(\mu_1,\mu_2)$ and $\gamma_{23}\in\Gamma(\mu_2,\mu_3)$ be optimal couplings. Applying the disintegration theorem (recall Theorem 1.4.10) with respect to the variable x_2, we can write
$$\gamma_{12}(dx_1,dx_2)=\gamma_{12,x_2}(dx_1)\otimes\mu_2(dx_2)$$
and
$$\gamma_{23}(dx_2,dx_3)=\gamma_{23,x_2}(dx_3)\otimes\mu_2(dx_2).$$
Consider the measure $\tilde\gamma\in\mathcal{P}(X\times X\times X)$ given by
$$\tilde\gamma(dx_1,dx_2,dx_3):=\gamma_{12,x_2}(dx_2)\otimes\gamma_{23,x_2}(dx_3)\otimes\mu_2(dx_2).$$
This measure has the property that
$$\begin{aligned}\int_{X\times X\times X}\varphi(x_1,x_2)\,d\tilde\gamma(x_1,x_2,x_3)&=\int_{X\times X}\varphi(x_1,x_2)\gamma_{12,x_2}(dx_1)\left[\int_X d\gamma_{23}(x_3)\right]d\mu_2(x_2)\\&=\int_{X\times X}\varphi(x_1,x_2)\,d\gamma_{12}(x_1,x_2).\end{aligned}$$
Similarly
$$\int_{X\times X\times X}\varphi(x_2,x_3)\,d\tilde\gamma(x_1,x_2,x_3)=\int_{X\times X}\varphi(x_2,x_3)\,d\gamma_{23}(x_2,x_3).$$
In other words, the measure $\tilde\gamma$ allows us to think of the couplings γ_{12} and γ_{23} as if they lived in a common space $X\times X\times X$, with γ_{12} that does not depend on the third variable, and γ_{23} that does not depend on the first variable.[13]

Note that we have
$$\int_{X\times X\times X}\psi(x_1)\,d\tilde\gamma(x_1,x_2,x_3)=\int_{X\times X}\psi(x_1)\,d\gamma_{12}(x_1,x_2)=\int_X\psi(x_1)\,d\mu_1(x_1)\tag{3.2}$$
and similarly
$$\int_{X\times X\times X}\psi(x_3)\,d\tilde\gamma(x_1,x_2,x_3)=\int_X\psi(x_3)\,d\mu_3(x_3).\tag{3.3}$$

[13] This whole construction above is sometimes called *gluing Lemma*.

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

$$\int_{X\times X} \varphi(x_1, x_3)\, d\bar{\gamma}_{13} = \int_{X\times X\times X} \varphi(x_1, x_3)\, d\tilde{\gamma}(x_1, x_2, x_3),$$

it follows from (3.2) and (3.3) that $\bar{\gamma}_{13} \in \Gamma(\mu_1, \mu_3)$.

Thus, by the triangle inequality in $L^p(X \times X \times X, \tilde{\gamma})$ we have

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

where the last equality follows from the optimality of γ_{12} and γ_{23}. This concludes the proof.

□

Being a distance, W_p induces a topology on $\mathcal{P}_p(X)$. In the next theorem we show the connection between the Wasserstein topology and the weak-$*$ topology.

Theorem 3.1.6. *Fix an exponent $1 \le p < \infty$ and a base point $x_0 \in X$. Let $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}_p(X)$ be a sequence of probability measures, and let $\mu \in \mathcal{P}_p(X)$. The following statements are equivalent:*

1. *$\mu_n \overset{*}{\rightharpoonup} \mu$ and $\int_X d(x_0, x)^p\, d\mu_n \to \int_X d(x_0, x)^p\, d\mu$.*
2. *$W_p(\mu_n, \mu) \to 0$.*

Proof. We prove the two implications independently.

- 1. $\Rightarrow$ 2. Fix $\delta > 0$, and define

$$M_n \coloneqq \int_X \big(1 + d(x_0, x)^p\big)\, d\mu_n(x), \qquad M \coloneqq \int_X \big(1 + d(x_0, x)^p\big)\, d\mu(x).$$

Since μ_n and μ are probability measures and $\int_X d(x_0, x)^p\, d\mu_n \to \int_X d(x_0, x)^p\, d\mu$, it follows that $M_n \to M$ as $n \to \infty$. Define the probability measures

$$\nu_n \coloneqq \frac{1}{M_n}\big(1 + d(x_0, x)^p\big)\mu_n, \qquad \nu \coloneqq \frac{1}{M}\big(1 + d(x_0, x)^p\big)\mu.$$

Since $\mu_n \overset{*}{\rightharpoonup} \mu$ and $x \mapsto \big(1 + d(x_0, x)^p\big)$ is a continuous function, we easily deduce that $\nu_n \overset{*}{\rightharpoonup} \nu$.[14] Therefore, Lemma 2.1.13 implies that ν_n converges narrowly to ν as $n \to \infty$. Thus, by Theorem 2.1.11, we can find a compact set $K \subseteq X$ such that, for all $n \in \mathbb{N}$,

$$\frac{1}{M_n}\int_{X\setminus K} \big(1 + d(x_0, x)^p\big)\, d\mu_n(x) \le \delta, \qquad \frac{1}{M}\int_{X\setminus K} \big(1 + d(x_0, x)^p\big)\, d\mu(x) \le \delta.$$

[14] Indeed, given $\varphi \in C_c(X)$, the function $X \ni x \mapsto \big(1 + d(x_0, x)^p\big)\varphi(x)$ is continuous and compactly supported. Hence, since $\mu_n \overset{*}{\rightharpoonup} \mu$, it follows that

$$\int_X \varphi\, d\nu_n = \int_X \big(1 + d(x_0, x)^p\big)\varphi(x)\, d\mu_n(x) \to \int_X \big(1 + d(x_0, x)^p\big)\varphi(x)\, d\mu(x) = \int_X \varphi\, d\nu \qquad \text{as } n \to \infty.$$

Since $\varphi \in C_c(X)$ is arbitrary, this proves that $\nu_n \overset{*}{\rightharpoonup} \nu$.

Recalling that $M_n \to M$ as $n \to \infty$, this implies in particular that

$$\int_{X\setminus K} \big(1 + d(x_0,x)^p\big)\, d\mu_n(x) + \int_{X\setminus K} \big(1 + d(x_0,x)^p\big)\, d\mu(x) \le 3M\delta \qquad \forall\, n \gg 1. \tag{3.4}$$

Now, since by assumption the space X is locally compact, we can find a finite family of nonnegative functions $(\varphi_i)_{i\in I} \subset C_c(X)$ such that

$$\sum_{i\in I} \varphi_i \le 1 \text{ in } X, \quad \sum_{i\in I} \varphi_i(x) = 1 \text{ for all } x \in K, \quad \operatorname{diam}(\operatorname{supp}(\varphi_i)) \le \delta \text{ for all } i \in I. \tag{3.5}$$

Set

$$\Lambda_{n,i} := \int_X \varphi_i \, d\mu_n, \qquad \Lambda_i := \int_X \varphi_i \, d\mu, \qquad \lambda_{n,i} := \min\big\{\Lambda_{n,i}, \Lambda_i\big\},$$

and define the measures

$$\alpha_{n,i} := \frac{\lambda_{n,i}}{\Lambda_{n,i}} \varphi_i \mu_n, \qquad \beta_{n,i} := \frac{\lambda_{n,i}}{\Lambda_i} \varphi_i \mu,$$

$$\alpha_n := \mu_n - \sum_{i\in I} \alpha_{n,i}, \qquad \beta_n := \mu - \sum_{i\in I} \beta_{n,i}$$

Note that $\alpha_{n,i}(X) = \beta_{n,i}(X) = \lambda_{n,i}$, and $\alpha_n(X) = \beta_n(X) = 1 - \sum_{i\in I} \lambda_{n,i}$. Then, we define $\gamma_n \in \mathcal{P}(X \times X)$ as

$$\gamma_n := \sum_{i\in I} \frac{\alpha_{n,i} \otimes \beta_{n,i}}{\lambda_{n,i}} + \frac{\alpha_n \otimes \beta_n}{1 - \sum_{i\in I} \lambda_{n,i}}\,.$$

One can easily check that γ_n is a transport plan from μ_n to μ, i.e., $\gamma_n \in \Gamma(\mu_n, \mu)$. Also, since $\operatorname{diam}(\operatorname{supp}(\varphi_i)) \le \delta$,

$$\int_{X\times X} d(x,y)^p \, d\frac{\alpha_{n,i} \otimes \beta_{n,i}}{\lambda_{n,i}} = \int_{\operatorname{supp}(\varphi_i)\times\operatorname{supp}(\varphi_i)} d(x,y)^p \, d\frac{\alpha_{n,i} \otimes \beta_{n,i}}{\lambda_{n,i}} \le \lambda_{n,i}\delta^p\,. \tag{3.6}$$

Recalling that $\mu_n \overset{*}{\rightharpoonup} \mu$, we also have $\lambda_{n,i} \to \Lambda_i = \int \varphi_i \, d\mu$ as $n \to \infty$. Therefore, recalling (3.5), it follows from (3.4) that

$$\left|1 - \sum_{i\in I} \lambda_{n,i}\right| \le 4M\delta \quad \text{for } n \gg 1, \qquad \alpha_n(K) + \beta_n(K) \to 0 \quad \text{as } n \to \infty.$$

We now observe that

$$\begin{aligned}\int_{X\times X} d(x,y)^p \, d\frac{\alpha_n \otimes \beta_n}{1 - \sum_{i\in I} \lambda_{n,i}} &\le 2^p \int_{X\times X} \big[d(x_0,x)^p + d(x_0,y)^p\big]\, d\frac{\alpha_n \otimes \beta_n}{1 - \sum_{i\in I} \lambda_{n,i}} \\ &\le 2^p \Big(\int_X d(x_0,x)^p \, d\alpha_n(x) + \int_X d(x_0,x)^p \, d\beta_n(x) \Big)\,.\end{aligned} \tag{3.7}$$

Since $\alpha_n \le \mu_n$, $\beta_n \le \mu$, and $\alpha_n(K) \to 0$, $\beta_n(K) \to 0$, (3.4) implies that

$$\int_X d(x_0,x)^p \, d\alpha_n(x) + \int_X d(x_0,x)^p \, d\beta_n(x) \le 4M\delta \qquad \text{for } n \gg 1. \tag{3.8}$$

Hence, combining (3.6), (3.7) and (3.8), we finally deduce that

$$W_p(\mu_n,\mu)^p \le \int_{X\times X} d(x,y)^p \, d\gamma_n(x,y) \le \delta^p + 4M2^p\delta \qquad \forall\, n \gg 1.$$

Since $\delta > 0$ can be chosen arbitrarily small, this proves that $W_p(\mu_n,\mu) \to 0$ as $n \to \infty$.

- 2. $\Rightarrow$ 1. Let $\gamma_n \in \Gamma(\mu_n, \mu)$ be an optimal transport plan with respect to the cost $c(x,y) = d(x,y)^p$. Applying the triangle inequality for the Wasserstein distance (recall Theorem 3.1.5) and using that $W_p(\mu_n, \mu) \to 0$, we have

$$\int_X d(x_0,x)^p \, d\mu_n = W_p(\delta_{x_0}, \mu_n)^p \to W_p(\delta_{x_0}, \mu)^p = \int_X d(x_0,x)^p \, d\mu \,.$$

It remains to show that $\mu_n \overset{*}{\rightharpoonup} \mu$. Let $\varphi \in C_c(X)$ be a compactly supported function, and let $\omega : [0,\infty) \to [0,\infty)$ be its modulus of continuity (i.e., $|\varphi(x) - \varphi(y)| \le \omega(d(x,y))$ for all $x, y \in X$). Given $\delta > 0$, we have

$$\begin{aligned}
\left| \int_X \varphi \, d\mu_n - \int_X \varphi \, d\mu \right| &\le \int_{X\times X} |\varphi(x) - \varphi(y)| \, d\gamma_n(x,y) \\
&\le \int_{\{d(x,y)\le\delta\}} \omega(\delta) \, d\gamma_n(x,y) + \int_{\{d(x,y)>\delta\}} 2\|\varphi\|_\infty \, d\gamma_n(x,y) \\
&\le \omega(\delta) + 2\|\varphi\|_\infty \int_{\{d(x,y)>\delta\}} \frac{d(x,y)^p}{\delta^p} \, d\gamma_n(x,y) \\
&\le \omega(\delta) + \frac{2\|\varphi\|_\infty}{\delta^p} \int_{X\times X} d(x,y)^p \, d\gamma_n(x,y) \\
&= \omega(\delta) + \frac{2\|\varphi\|_\infty}{\delta^p} W_p(\mu_n, \mu)^p.
\end{aligned}$$

By first letting $n \to \infty$ and then $\delta \to 0$, the last inequality implies that $\int_X \varphi \, d\mu_n \to \int_X \varphi \, d\mu$, concluding the proof.

□

Theorem 3.1.6 is particularly useful when the ambient space X is compact (or, equivalently, when all measures $\mu_n \in \mathcal{P}(X)$ live inside a fixed compact set). Indeed, since in this case the function $d(x_0,\cdot)^p$ has compact support (because the whole space is compact), the convergence $\int_X d(x_0,x)^p \, d\mu_n \to \int_X d(x_0,x)^p \, d\mu$ is a consequence of the weak-$*$ convergence of μ_n to μ. Hence we immediately deduce that, on compact sets, Wasserstein convergence is equivalent to weak-$*$ convergence.

Corollary 3.1.7. *Let X be compact, $p \ge 1$, $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}_p(X)$ a sequence of probability measures, and $\mu \in \mathcal{P}_p(X)$. Then*

$$\mu_n \overset{*}{\rightharpoonup} \mu \qquad \Leftrightarrow \qquad W_p(\mu_n, \mu) \to 0.$$

3.1.1 Construction of geodesics

Let $X = \mathbb{R}^d$ and $\gamma \in \Gamma(\mu,\nu)$ be an optimal coupling for W_p. Set $\pi_t(x,y) \coloneqq (1-t)x + ty$, so that

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

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

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

Applying this bound on the intervals $[0,s]$, $[s,t]$, and $[t,1]$, we get

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

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

$$W_p(\mu_s,\mu_t)=|t-s|W_p(\mu_0,\mu_1)\qquad \forall\, 0\le s,t\le 1. \tag{3.9}$$

Definition 3.1.8. A curve of measure $(\mu_t)_{t\in[0,1]}\subset W_p(\mathbb{R}^d)$ is said to be a *constant speed geodesic* if (3.9) holds.

Remark 3.1.9. Notice that, on a Riemannian manifold, a minimizing geodesic (as defined in Section 1.3) satisfy (3.9) with W_p replaced by the Riemannian distance. Also the converse implication is true, if a curve on a Riemannian manifold satisfies (3.9) (with W_p replaced by the Riemannian distance) then the curve is a minimizing geodesic.

It follows from the discussion above that any optimal coupling γ induces a geodesic via the formula $\mu_t := (\pi_t)_\#\gamma$. Note that, in the particular case when the coupling $\gamma=(\mathrm{Id}\times T)_\#\mu$ is induced by a map, the geodesic μ_t takes the form

$$\mu_t=(\pi_t)_\#(\mathrm{Id}\times T)_\#\mu=\big(\pi_t\circ(\mathrm{Id}\times T)\big)_\#\mu=(T_t)_\#\mu,$$

where $T_t(x):=(1-t)x+tT(x)$ is the linear interpolation between the identity map and the transport map T.

3.2 An informal introduction to gradient flows in Hilbert spaces

Let $\mathcal{H}$ be a Hilbert space (think, as a first example, $\mathcal{H}=\mathbb{R}^d$) and let $\phi\colon\mathcal{H}\to\mathbb{R}$ be of class C^1. Given $x_0\in\mathcal{H}$, the gradient flow (GF) of ϕ starting at x_0 is given by the ordinary differential equation

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

Note that, for a solution $x(t)$ of the gradient flow, it holds

$$\frac{d}{dt}\phi(x(t))=\nabla\phi(x(t))\cdot\dot x(t)=-|\nabla\phi|^2(x(t))\le 0. \tag{3.10}$$

Thus:

- ϕ decreases along the curve $x(t)$;
- we have $\frac{d}{dt}\phi(x(t))=0$ if and only if $|\nabla\phi|(x(t))=0$, i.e., $x(t)$ is a critical point of ϕ.

In particular, if ϕ has a unique stationary point which coincides with the global minimizer (this is for instance the case if ϕ is strictly convex), then one expect $x(t)$ to converge to the minimizer as $t\to+\infty$.

Remark 3.2.1. To define a gradient flow, one needs a scalar product (exactly as in the definition of gradient of a function on a manifold, see Definition 1.3.2). Indeed, as a general fact, given a function $f\colon\mathcal{H}\to\mathbb{R}$ one defines its differential $df(x):\mathcal{H}\to\mathbb{R}$ as

$$df(x)[v]=\lim_{\varepsilon\to0}\frac{f(x+\varepsilon v)-f(x)}{\varepsilon}\,.$$

If f is sufficiently regular, the map $df(x)\colon\mathcal{H}\to\mathbb{R}$ is linear and continuous, which means that $df(x)\in\mathcal{H}^*$ (the dual space of $\mathcal{H}$). On the other hand, if $t\mapsto x(t)\in\mathcal{H}$ is a curve, then

$$\dot x(t)=\lim_{\varepsilon\to0}\frac{x(t+\varepsilon)-x(t)}{\varepsilon}\in\mathcal{H}.$$

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

To define a (GF), we need a way to identify $\mathcal{H}$ and $\mathcal{H}^*$. This can be done if we introduce a scalar product. Indeed, 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 the unique element of $\mathcal{H}$ such that

$$\langle \nabla f(x), v \rangle := df(x)[v] \qquad \forall\, v \in \mathcal{H}.$$

In other words, the scalar product allows us to identify the gradient and the differential, and thanks to this identification we can now make sense of $\dot{x}(t) = -\nabla f(x(t))$.

Now the first question is: how does one constructs a solution to (GF)? If $\nabla\phi$ is Lipschitz continuous, one can simply rely on the Picard–Lindelöf Theorem (see [Tes12, Theorem 2.2]). Actually, even if $\nabla\phi$ is only continuous, one could rely on Peano Theorem (see [Tes12, Theorem 2.19]) to get existence of a solution. Unfortunately, as we shall see, in most situations of interest $\nabla\phi$ is not continuous. So, even the assumption of C^1 regularity is too strong; for the time being we keep this assumption just to emphasize the ideas, but later we shall remove it.

A classical way to construct solutions of (GF) is by discretizing the ODE in time, via the so-called *implicit Euler scheme*. More precisely, fixed a small time step $\tau > 0$, we discretize the time derivative $\dot{x}(t)$ as $\frac{x(t+\tau)-x(t)}{\tau}$, so that our ODE becomes

$$\frac{x(t+\tau) - x(t)}{\tau} = -\nabla\phi(y)$$

for a suitable choice of the point y. A natural idea would be to choose $y = x(t)$ (as in the *explicit* Euler scheme), but for our purposes the choice $y = x(t+\tau)$ (as in the *implicit* Euler scheme) works better. Thus, given $x(t)$, one looks for a point $x(t+\tau) \in \mathcal{H}$ solving the relation

$$\frac{x(t+\tau) - x(t)}{\tau} = -\nabla\phi(x(t+\tau)).$$

With this idea in mind, we set $x_0^\tau = x_0$. Then, given $k \geq 0$ and x_k^τ, we want to find x_{k+1}^τ by solving

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

or equivalently

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

where $\|\cdot\|$ denotes the norm induced by the scalar product introduced before. In other words, x_{k+1}^τ is a critical point of the function $\psi_k^\tau(x) := \frac{\|x-x_k^\tau\|^2}{2\tau} + \phi(x)$. Therefore, a natural way to construct x_{k+1}^τ is by looking for a global minimizer of ψ_k^τ.

As mentioned above, the C^1 assumption on ϕ is generally to strong. So, let us assume instead that $\phi\colon \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is convex and lower semicontinuous, and recall the notion of subdifferential introduced in Definition 2.5.1. Then we define a generalized gradient flow in the following way:

Definition 3.2.2. An absolutely continuous curve[15] $x : [0, +\infty) \to \mathcal{H}$ is a gradient flow for the

[15] An absolutely continuous curve is a continuous curve which is differentiable a.e., its derivative satisfies $|\dot{x}(t)| \in L^1_{\mathrm{loc}}([0, +\infty))$, and the fundamental theorem of calculus holds:

$$x(t) - x(s) = \int_s^t \dot{x}(\tau)\, d\tau \qquad \forall\, s, t \in [0, +\infty).$$

We refer to [AGS08, Section 1.1] for a general introduction to absolutely continuous curves.

convex and lower semicontinuous function ϕ with initial point $x_0 \in \mathcal{H}$ if

$$(\mathrm{GF}) := \begin{cases} x(0) = x_0\,, \\ \dot{x}(t) \in -\partial\phi(x(t)) \quad \text{for a.e. } t > 0. \end{cases}$$

Proceeding by analogy with what we did before, for ϕ convex and lower semicontinuous we can still repeat the construction of discrete solutions via the implicit Euler scheme: we set $x_0^\tau = x_0$, and given $k \geq 0$ and x_k^τ we look for a point x_{k+1}^τ satisfying

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

One can check that this relation is equivalent to

$$0 \in \frac{x_{k+1}^\tau - x_k^\tau}{\tau} + \partial\phi(x_{k+1}^\tau) =: \partial\psi_k^\tau(x_{k+1}^\tau), \qquad \psi_k^\tau(x) := \frac{\|x - x_k^\tau\|^2}{2\tau} + \phi(x).$$

Note that $0 \in \partial\psi_k^\tau(x_{k+1}^\tau)$ is equivalent to saying that x_{k+1}^τ is a global minimizer of ψ_k^τ (this follows immediately from Definition 2.5.1). Hence, given x_k^τ, one finds x_{k+1}^τ by minimizing $x \mapsto \psi_k^\tau(x)$.

It is not difficult to prove that a minimizer exists[16], so we can construct the sequence $(x_k^\tau)_{k\geq 0}$. Then, setting $x^\tau(0) := x_0$ and $x^\tau(t) := x_k^\tau$ for $t \in ((k-1)\tau, k\tau]$, one obtains a curve $t \mapsto x^\tau(t)$ that should be almost a solution to the (GF).

Then, the main challenge is to let $\tau \to 0$ and prove that there exists a limit curve $x(t)$ that indeed solves (GF). We shall not discuss this here, and we refer to [AG13, Section 3.1] and the references therein for more details.

Remark 3.2.3 (Uniqueness and stability)**.** Let ϕ be a convex function, and let $x(t), y(t)$ be solutions of (GF) with initial conditions x_0 and y_0 respectively. If ϕ is of class C^1 then

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

where the last inequality follows from the convexity of ϕ.

[16] Actually, in this case one can prove that there exists a unique minimizer. Indeed, to prove this, fix $x_0 \in \mathcal{H}$ a point where the subdifferential of ϕ is nonempty, and fix $p_0 \in \partial\phi(x_0)$. Then

$$\phi(x) \geq \phi(x_0) - \langle p_0, x - x_0\rangle \geq \phi(x_0) - \|p_0\|\big(\|x\| + \|x_0\|\big) = -A\|x\| - B \qquad \forall\, x \in \mathcal{H},$$

where $A := \|p_0\|$ and $B := \|p_0\|\,\|x_0\| - \phi(x_0)$. Hence, recalling the definition of ψ_k^τ, this proves that

$$\lim_{\|x\|\to\infty} \psi_k^\tau(x) \geq \lim_{\|x\|\to\infty} \frac{\|x - x_k^\tau\|^2}{2\tau} - A\|x\| - B = +\infty.$$

Thus, if x_j is a minimizing sequence of ψ_k^τ (i.e., $\psi_k^\tau(x_j) \to \inf_{\mathcal{H}} \psi_k^\tau$ as $j \to \infty$), it follows from the equation above that $\|x_j\|$ cannot go to infinity. This means that x_j is a bounded sequence in the Hilbert space $\mathcal{H}$, so by Banach-Alaoglu's Theorem it has a subsequence x_{j_ℓ} that converges weakly to some point $\bar{x}$. Note now that ψ_k^τ is a lower semicontinuous convex function. Also, for convex functions, lower semicontinuity with respect to the strong convergence is equivalent to lower semicontinuity with respect to the weak convergence (see for instance [Bre11, Corollary 3.9]). Hence

$$\psi_k^\tau(\bar{x}) \leq \liminf_{\ell\to\infty} \psi_k^\tau(x_{j_\ell}) = \inf_{\mathcal{H}} \psi_k^\tau,$$

which proves that $\bar{x}$ is a minimizer. Note also that, since ϕ_k^τ is uniformly convex (being the sum of the convex function ϕ and a uniformly convex function), the minimizer is unique: indeed, if $\bar{x}_1$ and $\bar{x}_2$ are minimizers then

$$\psi_k^\tau\left(\frac{\bar{x}_1 + \bar{x}_2}{2}\right) \leq \frac{\psi_k^\tau(\bar{x}_1) + \psi_k^\tau(\bar{x}_2)}{2} = \frac{\inf_{\mathcal{H}} \psi_k^\tau + \inf_{\mathcal{H}} \psi_k^\tau}{2} = \inf_{\mathcal{H}} \psi_k^\tau,$$

so equality holds in the first inequality, and therefore $\bar{x}_1 = \bar{x}_2$.

More in general, if ϕ convex but not necessarily C^1, we have

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

and therefore

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

where the last inequality follows from the monotonicity of the subdifferential of convex functions (this is just a particular case of the cyclical monotonicity of the subdifferential of a convex function in the case $N = 2$, see Theorem 2.5.2).

In particular, in both cases the (GF) is unique. Even more, if the initial conditions x_0 and y_0 are close, then $x(t)$ and $y(t)$ remain uniformly close for all times.

Example 3.2.4. Let $\mathcal{H} = L^2(\mathbb{R}^d)$ and

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

We claim that

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

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

Proof. Even though the proofs are quite similar, we prove the two implications separately.

$\Rightarrow$ 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

$$\int_{\mathbb{R}^d} \frac{|\nabla(u+\varepsilon w)|^2}{2}\, dx - \int_{\mathbb{R}^d} \frac{|\nabla u|^2}{2}\, dx \geq \varepsilon \int_{\mathbb{R}^d} p\, w\, dx.$$

Rearranging the terms and dividing by ε yields

$$\int_{\mathbb{R}^d} \nabla u \cdot \nabla w\, dx + \frac{\varepsilon}{2}\int_{\mathbb{R}^d} |\nabla w|^2 dx \geq \int_{\mathbb{R}^d} p\, w\, dx,$$

so by letting $\varepsilon \to 0$ we obtain

$$\int_{\mathbb{R}^d} \nabla u \cdot \nabla w \geq \int_{\mathbb{R}^d} p\, w\, dx \qquad \forall\, w \in W^{1,2}(\mathbb{R}^d).$$

Replacing w with $-w$ in the inequality above, we conclude that

$$\int_{\mathbb{R}^d} \underbrace{-\Delta u}_{\text{as a distribution}}\ w = \int_{\mathbb{R}^d} \nabla u \cdot \nabla w\, dx = \int_{\mathbb{R}^d} p\, w\, dx \qquad \forall\, w \in W^{1,2}(\mathbb{R}^d),$$

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

$\Leftarrow$ Assume that the distributional Laplacian Δu belongs to $L^2(\mathbb{R}^d)$. By definition of ϕ, for any $w \in W^{1,2}(\mathbb{R}^d)$ we have

$$\phi(u+w) - \phi(u) = \int_{\mathbb{R}^d} \nabla u \cdot \nabla w \, dx + \frac{1}{2}\int_{\mathbb{R}^d} |\nabla w|^2 dx \geq \int_{\mathbb{R}^d} \nabla u \cdot \nabla w \, dx = \int_{\mathbb{R}^d} -\Delta u \; w \, dx.$$

On the other hand, if $w \notin W^{1,2}(\mathbb{R}^d)$ then trivially

$$\phi(u+w) = +\infty \geq \phi(u) + \int_{\mathbb{R}^d} -\Delta u \; w \, dx.$$

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

□

As a consequence of this discussion, we obtain the following:

Corollary 3.2.5 (Heat equation as gradient flow)**.** *Let $\mathcal{H} = L^2(\mathbb{R}^d)$ and consider the* Dirichlet energy functional

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

Then the (GF) of ϕ with respect to the L^2-scalar product is the heat equation*, i.e.,*

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

3.3 Heat equation and optimal transport: the JKO scheme

In the previous paragraph we saw that the heat equation is the L^2-gradient flow of the Dirichlet energy functional (see Corollary 3.2.5). Hence, by the discussion done in the previous section, a way to find solutions to the heat equation is by solving the (GF) by means of the implicit Euler scheme

$$u^\tau_{k+1} \quad \text{is the minimizer in } L^2(\mathbb{R}^d) \text{ of } \quad u \mapsto \frac{\|u - u^\tau_k\|^2_{L^2(\mathbb{R}^d)}}{2\tau} + \phi(u),$$

and then letting $\tau \to 0$.

In [JKO98], the authors discovered a completely new and surprising way of constructing solutions of the heat equations as gradient flows. More precisely, the authors discovered that by replacing the Dirichlet energy functional with the so-called "entropy functional" $\int \rho \log(\rho)$, and by replacing the L^2-norm with the 2-Wasserstein distance, one obtains again the heat equation. In other words, the scheme above can be replaced by the following one:[17]

$$\rho^\tau_{k+1} \quad \text{is the minimizer in } \mathcal{P}(\mathbb{R}^d) \text{ of } \quad \rho \mapsto \frac{W_2^2(\rho, \rho^\tau_k)}{2\tau} + \int_{\mathbb{R}^d} \rho \log(\rho) \, dx.$$

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

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

Remark 3.3.1. In all this section we will work with probability densities. However, up to multiplying the initial datum ρ_0 by a positive constant, one can always reduce to this setting whenever $\rho_0 \in L^1(\mathbb{R}^d)$ is nonnegative.

[17] We adopt the convention that $\int_{\mathbb{R}^d} \rho \log(\rho) \, dx := +\infty$ if $\rho \in \mathcal{P}(\mathbb{R}^d)$ is not absolutely continuous with respect to the Lebesgue measure.

In the paper [JKO98], the authors consider solutions in the whole space $\mathbb{R}^d$. Here, instead, we consider the setting of a bounded convex domain $\Omega \subset \mathbb{R}^d$.

More precisely, we take ρ_0 to be a probability density in Ω such that

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

Fix $\tau > 0$, set $\rho_0^\tau := \rho_0$, and given ρ_k^τ we define ρ_{k+1}^τ as the minimizer of

$$\rho \mapsto \frac{W_2^2(\rho, \rho_k^\tau)}{2\tau} + \int_\Omega \rho \log(\rho)\, dx. \tag{3.11}$$

The goal of this section is to show that, as $\tau \to 0$, the scheme converges to the solution of the heat equation. We begin by proving the existence of discrete solutions.[18]

Lemma 3.3.2. *For any $k \geq 0$, ρ_{k+1}^τ exists (i.e., the functional in (3.11) has a minimum).*

Proof. Fix $k \geq 0$, and take $(\rho_m)_{m\in\mathbb{N}} \subset \mathcal{P}(\Omega)$ a minimizing sequence, that is

$$\frac{W_2^2(\rho_m, \rho_k^\tau)}{2\tau} + \int_\Omega \rho_m \log(\rho_m)\, dx \to \inf_{\rho\in\mathcal{P}(\Omega)} \left\{ \frac{W_2^2(\rho, \rho_k^\tau)}{2\tau} + \int_\Omega \rho \log(\rho)\, dx \right\}.$$

For all $M \in \mathbb{N}$ the sequence $\{\rho_m \wedge M\}_{m\in\mathbb{N}}$ is bounded in $L^\infty(\Omega)$, thus by Banach-Alaoglu's Theorem it is weakly-$*$ compact in L^∞. Hence, by a diagonal argument, we can find a subsequence m_ℓ independent of M such that $\rho_{m_\ell} \wedge M \overset{*}{\rightharpoonup} \rho_M$ in $L^\infty(\Omega)$ for each $M \in \mathbb{N}$. Also, since $s\log(s) + 1 \geq 0$ for all $s \geq 0$, we can bound

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

where the last bound follows from the fact that ρ_{m_ℓ} is a minimizing sequence (hence $\int_\Omega \rho_{m_\ell} \log(\rho_{m_\ell})$ is uniformly bounded) and Ω is bounded (hence it has finite volume).

Set $\rho_\infty := \sup_M \rho_M$. We know that

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

Hence, thanks to the first two properties, we can find a sequence of indices $(m_{\ell_M})_{M\in\mathbb{N}}$, with $m_{\ell_M} \to \infty$, such that $\rho_{m_{\ell_M}} \wedge M \rightharpoonup \rho_\infty$ in $L^1(\Omega)$ as $M \to \infty$. Also, thanks to the third property, $\rho_{m_{\ell_M}} \rightharpoonup \rho_\infty$ in $L^1(\Omega)$.

We now want to show that ρ_∞ is still a probability density. Note that this is not obvious, since some mass may have "escaped" from Ω. To prove this, set $N_\varepsilon := \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) < \varepsilon\}$.

[18] The reader familiar with Dunford-Pettis' Theorem will find the proof longer than needed. However, we have decided to present a more elementary proof based only on the weak-$*$ compactness of L^∞.

Since $|N_\varepsilon| \le C\varepsilon$, for $L := \frac{1}{\varepsilon|\log(\varepsilon)|}$ we have

$$\begin{aligned}\int_{N_\varepsilon} \rho_m &\le \int_{N_\varepsilon \cap \{\rho_m \le L\}} \rho_m + \int_{N_\varepsilon \cap \{\rho_m \ge L\}} \rho_m \frac{\log(\rho_m)}{\log(L)} \\ &\le L|N_\varepsilon| + \frac{C}{\log(L)} \le C\left(\varepsilon L + \frac{1}{\log(L)}\right) \le \frac{C}{|\log(\varepsilon)|} \qquad \forall\, m \in \mathbb{N},\end{aligned}$$

so in particular

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

and therefore

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

Letting $\varepsilon \to 0$, we conclude that ρ_∞ is a probability density. In particular, it follows by Lemma 2.1.13 that the family $\{\rho_{m_{\ell_M}}\}_{m\in\mathbb{N}}$ is tight and the convergence of $\rho_{m_{\ell_M}}$ to ρ_∞ is also narrow.

We now observe that, since $[0,\infty) \ni s \mapsto s\log(s)$ is convex, [AFP00, Theorem 5.2] implies that[19]

$$\int_\Omega \rho_\infty \log(\rho_\infty) \le \liminf_{M\to\infty} \int_\Omega \rho_{m_{\ell_M}} \log(\rho_{m_{\ell_M}}). \tag{3.12}$$

We now want to study the behaviour of $W_2^2(\rho_{m_{\ell_M}}, \rho_k^\tau)$ as $M \to \infty$.

Let $\gamma_M \in \Gamma(\rho_{m_{\ell_M}}, \rho_k^\tau)$. Then, since the family $\{\rho_{m_{\ell_M}}\}_{m\in\mathbb{N}}$ is tight (by the previous discussion), the proof of Lemma 2.3.1 shows that also γ_M is tight. Hence, up to taking a subsequence, $\gamma_M \rightharpoonup \gamma_\infty$ with

$$(\pi_1)_\#\gamma_\infty = \rho_\infty, \qquad (\pi_2)_\#\gamma_\infty = \rho_k^\tau,$$

thus $\gamma_\infty \in \Gamma(\rho_\infty, \rho_k^\tau)$. Note also that, since $|x-y|^2$ is continuous and bounded on $\Omega\times\Omega$,

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

Hence, combining together the lower semicontinuity of $\int_\Omega \rho_{m_{\ell_M}} \log(\rho_{m_{\ell_M}})$ with the equation above, we get

$$\liminf_{M\to\infty} \frac{W_2^2(\rho_{m_{\ell_M}}, \rho_k^\tau)}{2\tau} + \int_\Omega \rho_{m_{\ell_M}} \log(\rho_{m_{\ell_M}}) \ge \frac{W_2^2(\rho_\infty, \rho_k^\tau)}{2\tau} + \int_\Omega \rho_\infty \log(\rho_\infty).$$

Since ρ_m was a minimizing sequence, this proves that ρ_∞ is a minimizer. Hence, we define $\rho_{k+1}^\tau := \rho_\infty$. □

[19] As simple way to prove (3.12) is the following: note that, for each $s \ge 0$, it holds

$$s\log s \ge s(\sigma+1) - e^\sigma \quad \forall\, \sigma \in \mathbb{R}, \qquad \text{with equality for } \sigma = \log(s).$$

Hence, given any continuous function $\phi(x)$, we have

$$\liminf_{M\to\infty} \int_\Omega \rho_{m_{\ell_M}} \log(\rho_{m_{\ell_M}}) \ge \liminf_{M\to\infty} \int_\Omega \left(\rho_{m_{\ell_M}}(x)(\phi(x)+1) - e^{\phi(x)}\right) dx = \int_\Omega \left(\rho_\infty(x)(\phi(x)+1) - e^{\phi(x)}\right) dx,$$

where we applied the previous formula with $s = \rho(x)$ and $\sigma = \phi(x)$, and the final equality follows from the narrow convergence of $\rho_{m_{\ell_M}}$ to ρ_∞. Choosing $\{\phi_k\}_{k\in\mathbb{N}}$ a sequence of functions converging to $\log(\rho_\infty)$, the result follows by applying the above formula to $\phi = \phi_k$ and letting $k \to \infty$.

Next, since ρ^τ_{k+1} minimizes the functional (3.11), we expect it to satisfy some kind of *minimality equation*. This is the purpose of the next:

Lemma 3.3.3. *For any vector field $\xi \in C^\infty(\Omega, \mathbb{R}^d)$ tangent to the boundary of Ω, it holds*

$$\int_\Omega \rho^\tau_{k+1} \operatorname{div}(\xi)\, dx = \frac{1}{\tau} \int_\Omega \langle \xi \circ T_{k+1}, T_{k+1} - x \rangle \rho^\tau_k \, dx\,,$$

where $T_{k+1} : \Omega \to \Omega$ is the optimal map from ρ^τ_k to ρ^τ_{k+1}.

Proof. In order to exploit the minimality of ρ^τ_{k+1}, we want to perturb it. We do it as follows.

Consider the flow of ξ:

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

Since ξ is tangent to $\partial\Omega$, it follows that $\Phi(t) : \Omega \to \Omega$ is a diffeomorphism. So we can define

$$\rho_\varepsilon \coloneqq \Phi(\varepsilon)_\# \rho^\tau_{k+1} \in \mathcal{P}(\Omega).$$

It follows by Section 1.6 that

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

therefore

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

Then, a Taylor expansion gives (cp. (2.15))

$$\begin{aligned} \int_\Omega \rho_\varepsilon \log(\rho_\varepsilon) &= \int_\Omega \rho^\tau_{k+1} \log(\rho^\tau_{k+1}) - \int_\Omega \rho^\tau_{k+1} \log(\underbrace{\det \nabla \Phi(\varepsilon, x)}_{1+\varepsilon \operatorname{div} \xi + o(\varepsilon)})\, dx \\ &= \int_\Omega \rho^\tau_{k+1} \log(\rho^\tau_{k+1}) - \varepsilon \int_\Omega \rho^\tau_{k+1} \operatorname{div} \xi \, dx + o(\varepsilon). \end{aligned}$$

Now, given a coupling $\gamma \in \Gamma(\rho^\tau_{k+1}, \rho^\tau_k)$ optimal for the W_2-distance, define $\gamma_\varepsilon \coloneqq (\Phi(\varepsilon, \cdot) \times \mathrm{Id})_\# \gamma$. Since

$$(\pi_1)_\# \gamma_\varepsilon = \Phi(\varepsilon, \cdot)_\# \rho^\tau_{k+1} = \rho_\varepsilon, \qquad (\pi_2)_\# \gamma_\varepsilon = (\pi_2)_\# \gamma = \rho^\tau_k,$$

and $\Phi(\varepsilon, x) = x + \varepsilon \xi(x) + o(\varepsilon)$, we get

$$\begin{aligned} W_2^2(\rho_\varepsilon, \rho^\tau_k) &\le \int_{\Omega\times\Omega} |x-y|^2 d\gamma_\varepsilon = \int_{\Omega\times\Omega} |\Phi(\varepsilon, x) - y|^2 d\gamma \\ &= \int_{\Omega\times\Omega} \left[|x-y|^2 + 2\varepsilon \langle \xi(x), x - y\rangle + o(\varepsilon) \right] d\gamma\,. \end{aligned}$$

Therefore, recalling that γ is optimal from ρ^τ_{k+1} to ρ^τ_k, we obtain

$$W_2^2(\rho_\varepsilon, \rho^\tau_k) \le W_2^2(\rho^\tau_{k+1}, \rho^\tau_k) + 2\varepsilon \int_{\Omega\times\Omega} \langle \xi(x), x - y\rangle \, d\gamma + o(\varepsilon).$$

Combining everything together, we proved that

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

Hence, since ε can be chosen both positive and negative, we see that the term $(\star)$ has to vanish. Therefore our optimality condition for ρ_{k+1}^τ reads as

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

where γ realizes the 2-Wasserstein distance between ρ_{k+1}^τ and ρ_k^τ.

To simplify the formula we apply Theorem 2.5.9 to deduce that the optimal plan γ is unique and is induced by an optimal map T_{k+1} from ρ_k^τ to ρ_{k+1}^τ, 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 equation becomes[20]

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

as wanted. □

We are now ready to state and prove the main result of this section.

Theorem 3.3.4. *Given $\tau>0$, let $\rho^\tau:[0,\infty)\to\mathcal{P}(\Omega)$ be the curve of probability densities given by*

$$\rho^\tau(t)\coloneqq\begin{cases}\rho_0 & \text{for } t=0,\\ \rho_k^\tau & \text{for } t\in((k-1)\tau,k\tau],\ k\ge1.\end{cases}\tag{3.13}$$

Then there exists a curve of probability measures $\rho\in L^1_{\mathrm{loc}}([0,\infty)\times\Omega)$ such that, up to a subsequence in τ, $\rho^\tau\rightharpoonup\rho$ weakly in $L^1_{\mathrm{loc}}([0,\infty)\times\Omega)$. Furthermore, ρ satisfies the heat equation (in the distributional sense) with initial datum ρ_0 and zero Neumann boundary conditions.

Proof. By the minimality of ρ_k^τ, we have

$$\begin{aligned}\frac{W_2^2(\rho_k^\tau,\rho_{k-1}^\tau)}{2\tau}+\int_\Omega\rho_k^\tau\log(\rho_k^\tau)\,dx &\le\left(\frac{W_2^2(\rho,\rho_{k-1}^\tau)}{2\tau}+\int_\Omega\rho\log(\rho)\,dx\right)\Big|_{\rho=\rho_{k-1}^\tau}\\ &=\int_\Omega\rho_{k-1}^\tau\log(\rho_{k-1}^\tau)\,dx.\end{aligned}$$

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

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

Using this expression, one gets the desired formula for the optimality condition.

Thus, by taking the telescopic sum over $k = 1, \dots, k_0$, one gets

$$\underbrace{\sum_{k=1}^{k_0} \frac{W_2^2(\rho_k^\tau, \rho_{k-1}^\tau)}{2\tau}}_{\geq 0} + \int_\Omega \rho_{k_0}^\tau \log(\rho_{k_0}^\tau)\, dx \leq \int_\Omega \rho_0 \log(\rho_0)\, dx.$$

In particular, we deduce that the entropy $\int_\Omega \rho_k \log(\rho_k)$ decreases in k. Therefore, recalling (3.13), we have

$$\int_\Omega \rho^\tau(t,x) \log(\rho^\tau(t,x))\, dx \leq \int_\Omega \rho_0 \log(\rho_0)\, dx \qquad \forall\, \tau > 0,\ t \geq 0. \tag{3.14}$$

Also, since $s \log s \geq -1$ on $[0, +\infty)$, for any $k_0 \geq 1$ it holds

$$\sum_{k=1}^{k_0} \frac{W_2^2(\rho^\tau(k\tau), \rho^\tau((k-1)\tau))}{2\tau} \leq \int_\Omega \rho_0 \log(\rho_0)\, dx - \int_\Omega \rho_{k_0} \log(\rho_{k_0})\, dx \leq \int_\Omega \big(\rho_0 \log(\rho_0) + 1\big)\, dx. \tag{3.15}$$

Furthermore, since $\int_\Omega \rho^\tau(t) = 1$, we have

$$\int_{t_1}^{t_2} \int_\Omega \rho^\tau(t,x)\, dx\, dt = t_2 - t_1 \qquad \forall\, 0 \leq t_1 \leq t_2. \tag{3.16}$$

As shown in the proof of Lemma 3.3.2, the bound (3.14) implies that the measures $\rho^\tau(t)$ cannot concentrate nor escape to the boundary of Ω, uniformly in t. Hence, up to a subsequence, ρ^τ converges weakly in $L^1_{\rm loc}([0,\infty) \times \Omega)$ to a density $\rho(t,x)$, and by passing to the limit in (3.16) we deduce that $\int_\Omega \rho(t,x)\, dx = 1$ for a.e. $t \in [0,T]$.

Now that we have shown the convergence of ρ^τ, we want to show that ρ satisfies the heat equation. The idea is to test the heat equation against a test function of the form $\psi(x)\zeta(t)$. So, first we fix $\psi \in C^\infty(\Omega)$ such that $\frac{\partial \psi}{\partial \nu}|_{\partial\Omega} = 0$. Note that, by a Taylor expansion with integral reminder, one has

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

In particular

$$|\psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle| \leq \frac{1}{2}\|D^2\psi\|_\infty |x-y|^2,$$

from which it follows that

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

Then, applying Lemma 3.3.3 with $\xi = \nabla\psi$ (note that, since $\frac{\partial\psi}{\partial\nu}|_{\partial\Omega} = 0$, $\nabla\psi$ is tangent to the boundary) we obtain

$$\left| -\int_\Omega \Delta\psi\, \rho_k^\tau\, dx + \frac{1}{\tau} \underbrace{\int_\Omega [\psi(T_k) - \psi(x)]\, \rho_{k-1}^\tau\, dx}_{= \int \psi\rho_k^\tau - \int \psi\rho_{k-1}^\tau} \right| \leq \frac{1}{2}\|D^2\psi\|_\infty \frac{W_2^2(\rho_k^\tau, \rho_{k-1}^\tau)}{\tau} \tag{3.17}$$

We now take $\zeta \in C_c^\infty([0,+\infty))$, and we multiply (3.17) against $\tau\,\zeta((k-1)\tau)$. Then, recalling (3.13), we get

$$\begin{aligned}\Big|\int_\Omega \psi(x)\,\rho^\tau(k\tau,x)\,\zeta((k-1)\tau)\,dx - \int_\Omega \psi(x)\,\rho^\tau((k-1)\tau,x)\,\zeta((k-1)\tau)\,dx \\ - \tau\int_\Omega \Delta\psi(x)\,\rho^\tau(k\tau,x)\,\zeta((k-1)\tau)\,dx\Big| \\ \le \frac{1}{2}\|D^2\psi\|_\infty\|\zeta\|_\infty W_2^2(\rho^\tau(k\tau),\rho^\tau((k-1)\tau)).\end{aligned}$$

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

$$\begin{aligned}\Big|-\zeta(0)\int_\Omega \psi(x)\,\rho_0(x)\,dx + \overbrace{\sum_{k=2}^\infty \int_\Omega \psi(x)\,\rho^\tau(kt)\,\zeta((k-1)\tau))\,dx - \sum_{k=1}^\infty \int_\Omega \psi(x)\,\rho^\tau(k\tau,x)\,\zeta(k\tau)\,dx}^{(I)} \\ \underbrace{- \sum_{k=1}^\infty \tau\int_\Omega \Delta\psi(x)\,\rho^\tau(k\tau,x)\,\zeta((k-1)\tau)\;dx}_{(II)}\Big| \\ \le C\sum_{k=1}^\infty W_2^2(\rho^\tau(k\tau),\rho^\tau((k-1)\tau)) \le C\,\tau,\end{aligned}$$

where C depends on ψ and ζ, and the last bound follows from (3.15). We now rewrite the terms (I) and (II) as follows. For the term (I), we have

$$\begin{aligned}(I) &= \sum_{k=1}^\infty \int \psi(x)\,\overbrace{\rho^\tau(k\tau)}^{\rho^\tau(t)\ ,\ t\in[(k-1)\tau,k\tau]}\,\overbrace{[\zeta((k-1)\tau)-\zeta(k\tau)]}^{-\int_{(k-1)\tau}^{k\tau}\partial_t\zeta(t)dt}\,dx \\ &= -\sum_{k=1}^\infty \int_{(k-1)\tau}^{k\tau}\int_\Omega \psi(x)\,\rho^\tau(t,x)\,\partial_t\zeta(t)\,dx\,dt = -\int_0^\infty\int_\Omega \psi(x)\,\rho^\tau(t,x)\,\partial_t\zeta\,dx\,dt.\end{aligned}$$

For the term (II), since

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

we have

$$\begin{aligned}(II) &= \sum_{k=1}^\infty \int_\Omega \Delta\psi(x)\,\rho^\tau(t,x)\,\tau\,\zeta((k-1)\tau)\,dx \\ &= \sum_{k=1}^\infty \int_{(k-1)\tau}^{k\tau}\int_\Omega \Delta\psi(x)\,\rho^\tau(t,x)\,\zeta(t)\,dx\,dt + O\Big(\tau\sum_{k=1}^\infty\int_{(k-1)\tau}^{k\tau}\int_\Omega |\Delta\psi(x)|\,\rho^\tau(t,x)|\partial_t\zeta(t)|\,dx\,dt\Big) \\ &= \sum_{k=1}^\infty \int_{(k-1)\tau}^{k\tau}\int_\Omega \Delta\psi(x)\,\rho^\tau(t,x)\,\zeta(t)\,dx\,dt + O\Big(\tau\sum_{k=1}^\infty\int_0^\infty\int_\Omega |\Delta\psi(x)|\,\rho^\tau(t,x)|\partial_t\zeta(t)|\,dx\,dt\Big).\end{aligned}$$

Therefore, choosing $T>0$ such that $\mathrm{supp}(\zeta)\subset[0,T]$, we proved that

$$\begin{aligned}\Big|-\zeta(0)\int_\Omega \psi(x)\,\rho_0(x)\,dx - \int_0^\infty\int_\Omega \psi(x)\,\rho^\tau(t,x)\,\partial_t\zeta(t)\,dt\,dx - \int_0^\infty\int_\Omega \Delta\psi(x)\,\rho^\tau(t,x)\,\zeta(t)\,dx\,dt\Big| \\ \le C\tau + \tau\|\Delta\psi\|_\infty\|\partial_t\zeta\|_\infty\int_0^T\int_\Omega \rho^\tau(t,x)\,dx\,dt \to 0 \quad \text{as } \tau\to 0.\end{aligned}$$

Hence, since $\rho^\tau \rightharpoonup \rho$ in $L^1_{\mathrm{loc}}([0,\infty)\times\Omega)$, we conclude that

$$-\zeta(0)\int_\Omega \psi(x)\,\rho_0(x)\,dx - \int_0^\infty\int_\Omega \psi(x)\,\rho(t,x)\,\partial_t\zeta(t)\,dx\,dt - \int_0^\infty\int_\Omega \Delta\psi(x)\,\rho(t,x)\,\zeta(t)\,dx\,dt = 0 \tag{3.18}$$

for any smooth ψ satisfying $\frac{\partial\psi}{\partial\nu}|_{\partial\Omega}=0$.

We claim that (3.18) corresponds to saying that ρ solves, in the distributional sense, the heat equation with Neumann boundary conditions[21] $\frac{\partial\rho(t)}{\partial\nu}|_{\partial\Omega}=0$ and with initial datum ρ_0.

To prove the claim we first note that, integrating by parts in time,

$$\begin{aligned}
&-\zeta(0)\int_\Omega \psi(x)\,\rho_0(x)\,dx - \int_0^\infty\int_\Omega \psi(x)\,\rho(t,x)\,\partial_t\zeta(t)\,dx\,dt\\
&= -\zeta(0)\int_\Omega \psi(x)\,\rho_0(x)\,dx - \int_\Omega \psi(x)\,\rho(t,x)\,\zeta(t)\,dx\Big|_{t=0}^{t=\infty} + \int_\Omega \psi(x)\,\underbrace{\partial_t\rho}_{\text{as a distribution}}\,\zeta(t)\,dx\\
&= -\zeta(0)\int_\Omega \psi(x)\Big[\rho_0(x) - \underbrace{\rho(0)}_{\text{in the trace sense}}\Big]\,dx + \int_\Omega \psi(x)\,\underbrace{\partial_t\rho}_{\text{as a distribution}}\,\zeta(t)\,dx.
\end{aligned}$$

On the other hand, integrating by parts in space and using that $\frac{\partial\psi}{\partial\nu}|_{\partial\Omega}=0$, we have

$$\begin{aligned}
&\int_0^\infty\int_\Omega \Delta\psi(x)\,\rho(t,x)\,\zeta(t)\,dx\,dt\\
&= \int_0^\infty\left(\int_{\partial\Omega}\underbrace{\frac{\partial\psi}{\partial\nu}(x)}_{=0}\,\rho(t)\right)\zeta(t)\,dt - \int_0^\infty\int_\Omega \nabla\psi(x)\cdot\underbrace{\nabla\rho}_{\text{as a distribution}}\,\zeta(t)\,dx\,dt\\
&= -\int_0^\infty\left(\int_{\partial\Omega}\psi(x)\,\underbrace{\frac{\partial\rho(t)}{\partial\nu}}_{\text{as a distribution}}\right)\zeta(t)\,dt + \int_0^\infty\int_\Omega \psi(x)\,\underbrace{\Delta\rho}_{\text{as a distribution}}\,\zeta(t)\,dx\,dt.
\end{aligned}$$

Hence, in the sense of distributions, (3.18) is equivalent to

$$\begin{aligned}
0 = -\zeta(0)\int_\Omega \psi(x)\,[\rho_0(x)-\rho(0)]\,dx - \int_0^\infty\left(\int_{\partial\Omega}\psi(x)\frac{\partial\rho(t)}{\partial\nu}\right)\zeta(t)\,dt\\
+\int_0^\infty\int_\Omega \psi(x)\left[\partial_t\rho-\Delta\rho\right]\zeta(t)\,dx\,dt.
\end{aligned}\tag{3.19}$$

Choosing first $\psi\in C_c^\infty(\Omega)$ and $\zeta\in C_c^\infty((0,+\infty))$, we get

$$\int_0^\infty\int_\Omega \psi(x)\left[\partial_t\rho-\Delta\rho\right]\zeta(t)\,dx\,dt = 0,$$

and so by the arbitrariness of ψ and ζ we deduce that $\partial_t\rho-\Delta\rho=0$ in the sense of distributions.

Hence (3.19) becomes

$$0 = -\zeta(0)\int_\Omega \psi(x)\,[\rho_0(x)-\rho(0)]\,dx - \int_0^\infty\left(\int_{\partial\Omega}\psi(x)\frac{\partial\rho(t)}{\partial\nu}\right)\zeta(t)\,dt. \tag{3.20}$$

We now use (3.20) with $\zeta\in C_c^\infty((0,+\infty))$ and $\psi\in C^\infty(\Omega)$ such that $\frac{\partial\psi}{\partial\nu}|_{\partial\Omega}=0$ to get

$$0 = \int_0^\infty\left(\int_{\partial\Omega}\psi(x)\frac{\partial\rho(t)}{\partial\nu}\right)\zeta(t)\,dt.$$

[21] It is natural to expect that ρ satisfies the Neumann boundary condition $\frac{\partial\rho(t)}{\partial\nu}|_{\partial\Omega}=0$. Indeed, this condition corresponds to saying that the mass of ρ cannot enter nor leave Ω, and this is coherent with the way the solution was constructed.

Note that the constraint $\frac{\partial \psi}{\partial \nu}|_{\partial\Omega} = 0$ plays no role on the possible values of ψ on $\partial\Omega$. In other words, $\psi|_{\partial\Omega}$ can be chosen arbitrarily, and so the equation above implies that $\frac{\partial \rho(t)}{\partial \nu}|_{\partial\Omega} = 0$.

Finally, combining this information with (3.20), we deduce that

$$0 = -\zeta(0) \int_\Omega \psi(x) \left[\rho_0(x) - \rho(0)\right] dx \qquad \forall\, \psi \in C_c^\infty(\Omega),$$

hence $\rho(0) = \rho_0$, as desired. □

4 Differential viewpoint of optimal transport

The goal of this section is to introduce a differential structure on the space of probability measure, starting from the Benamou-Brenier formula and then introducing Otto's formalism. This will allow us to interpret several important PDEs as gradient flows with respect to the 2-Wasserstein distance. In order to focus on the main ideas behind this important theory, most computations of this section will be formal.

In the previous section we have seen that, considering the entropy functional $\rho \to \int \rho \log(\rho)$ in the 2-Wasserstein space, the discrete Euler scheme for gradient flows produces solutions to the heat equation. It is natural to wonder whether we can say that, in some sense:

"The heat equation is the gradient flow of the entropy with respect to the W_2 metric."

Moreover, one might ask whether a similar strategy could handle other evolution equations. In this section we give an answer to these questions by endowing the Wasserstein space with a differential structure. This makes it much easier to guess (and, with the right toolbox, prove) that the gradient flow of the entropy is the heat equation. Also, this will allow us to repeat the same strategy for other functionals/evolution equations.

4.1 The continuity equation and Benamou-Brenier formula

Let $\Omega \subset \mathbb{R}^d$ be convex set ($\Omega = \mathbb{R}^d$ is admissible), let $\bar\rho_0 \in \mathcal{P}_2(\Omega)$ be a probability measure with finite second moments (recall (3.1)), and let $v : [0,T] \times \Omega \to \mathbb{R}^d$ be a smooth bounded vector field tangent to the boundary of Ω. Let $X(t,x)$ denote the flow of v, namely

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

and set $\rho_t = (X(t))_\# \bar\rho_0$. Note that, since v is tangent to the boundary, the flow remains inside Ω, hence $\rho_t \in \mathcal{P}_2(\Omega)$.[22]

Lemma 4.1.1. *Let $v_t(\cdot) \coloneqq v(t,\cdot)$. The* continuity equation

$$\partial_t \rho_t + \operatorname{div}(v_t \rho_t) = 0 \tag{4.1}$$

holds in the distributional sense.

Proof. Let $\psi \in C_c^\infty(\Omega)$, and consider the function $t \mapsto \int_\Omega \rho_t(x)\,\psi(x)\,dx$. Then, using the definition of X and ρ_t, we get

$$\begin{aligned} \int_\Omega \partial_t \rho_t(x)\,\psi(x)\,dx &= \frac{d}{dt}\int_\Omega \rho_t(x)\,\psi(x)\,dx = \frac{d}{dt}\int_\Omega \psi(X(t,x))\,\bar\rho_0(x)\,dx \\ &= \int_\Omega \nabla\psi(X(t,x))\cdot \dot X(t,x)\bar\rho_0(x)\,dx = \int_\Omega \nabla\psi(X(t,x))\cdot v_t(X(t,x))\,\bar\rho_0(x)\,dx \\ &= \int_\Omega \nabla\psi(x)\cdot v_t(x)\,\rho_t(x)\,dx = -\int_\Omega \psi(x)\,\operatorname{div}(v_t\rho_t)\,dx. \end{aligned}$$

□

[22] Since v_t is bounded, $|X(t,x)-x| \le Ct$, therefore

$$\int_\Omega |x|^2 \rho_t(x)\,dx = \int_\Omega |X(t,x)|^2 \bar\rho_0(x)\,dx \le 2\int_\Omega \left(|x|^2 + (Ct)^2\right)\bar\rho_0(x)\,dx.$$

Thus, if ρ_0 has finite second moments, the same holds for ρ_t.

Definition 4.1.2. Given a pair (ρ_t, v_t) solving the continuity equation (4.1), with $v_t \cdot \nu|_{\partial\Omega} = 0$ (namely, v_t is tangent to the boundary of Ω), we define its action as

$$A[\rho_t, v_t] \coloneqq \int_0^1 \int_\Omega |v_t(x)|^2 \rho_t(x)\, dx\, dt\,.$$

The following remarkable formula, due to Benamou and Brenier [BB00], shows a link between the continuity equation and the W_2-distance.

Theorem 4.1.3 (Benamou-Brenier formula)**.** *Given two probability measures $\bar\rho_0, \bar\rho_1 \in \mathcal{P}_2(\Omega)$, it holds*

$$W_2^2(\bar\rho_0, \bar\rho_1) = \inf\left\{A[\rho_t, v_t] \mid \rho_0 = \bar\rho_0,\ \rho_1 = \bar\rho_1,\ \partial_t\rho_t + \operatorname{div}(v_t\rho_t) = 0,\ v_t \cdot \nu|_{\partial\Omega} = 0\right\}.$$

Proof. We give only a formal proof.

Let (ρ_t, v_t) be a couple "probability measure/smooth vector field" satisfying $\rho_0 = \bar\rho_0$, $\rho_1 = \bar\rho_1$, and $\partial_t\rho_t + \operatorname{div}(v_t\rho_t) = 0$,. Let $X(t,x)$ denote the flow of v_t. By the uniqueness for the continuity equation[23], the unique solution ρ_t is the one constructed in Lemma 4.1.1, namely $\rho_t = (X(t))_\#\bar\rho_0$. In particular $X(1)_\#\bar\rho_0 = \bar\rho_1$, which implies that $X(1)$ is a transport map from $\bar\rho_0$ to $\bar\rho_1$. Then, by the definition of X and Hölder inequality, we get

$$\begin{aligned} A[\rho_t, v_t] &= \int_0^1 \int_\Omega |v_t|^2 \rho_t\, dxdt = \int_0^1 \int_\Omega |v_t(X(t,x))|^2 \bar\rho_0(x)\, dxdt \\ &= \int_0^1 \int_\Omega |\dot X(t,x)|^2 \bar\rho_0(x) dtdx = \int_\Omega \bar\rho_0(x) \int_0^1 |\dot X(t,x)|^2 dtdx \\ &\geq \int_\Omega \bar\rho_0(x) \left|\int_0^1 \dot X(t,x) dt\right|^2 dx = \int_\Omega \bar\rho_0(x) |X(1,x) - x|^2\, dx \geq W_2^2(\bar\rho_0, \bar\rho_1). \end{aligned} \tag{4.2}$$

Hence, this proves that $W_2^2(\bar\rho_0, \bar\rho_1)$ is always less than or equal to the infimum appearing in the statement.

To show equality, take $X(t,x) = x + t(T(x) - x)$, where $T = \nabla\varphi$ is optimal from $\bar\rho_0$ to $\bar\rho_1$, set $\rho_t \coloneqq X(t)_\#\bar\rho_0$, and let v_t be such that $\dot X(t) = v_t \circ X(t)$.[24] 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 that all inequalities in (4.2) become equalities, therefore

$$A[\rho_t, v_t] = W_2^2(\bar\rho_0, \bar\rho_1).$$

□

[23] The uniqueness for the continuity equation, at least for smooth vector fields, can be obtained exploiting the duality with the transport equation, as sketched in [Amb08, pg. 3, (2)].

[24] This definition corresponds to saying that $v_t \coloneqq \dot X \circ X^{-1}(t)$. To show that this makes sense, we need to show that $X(t)^{-1}$ exists. Note that

$$\begin{aligned} |X(t,x) - X(t,\tilde x)||x - \tilde x| &\geq \langle X(t,x) - X(t,\tilde x), x - \tilde x\rangle \\ &= (1-t)\langle x - \tilde x, x - \tilde x\rangle + t\underbrace{\langle \nabla\varphi(x) - \nabla\varphi(\tilde x), x - \tilde x\rangle}_{\geq 0\ (\varphi\ \text{convex})} \geq (1-t)|x - \tilde x|^2, \end{aligned} \tag{4.3}$$

thus $|X(t,x) - X(t,\tilde x)| \geq (1-t)|x - \tilde x|$. This implies that, for $t \in [0,1)$, $X(t)$ is injective and therefore $X(t)^{-1}$ exists. In addition, this also proves that

$$|X(t)^{-1}(y) - X(t)^{-1}(\tilde y)| \leq \frac{1}{1-t}|y - \tilde y|,$$

so $X(t)^{-1}$ is also Lipschitz.

Note that the injectivity may be false for $t = 1$. Indeed, if $\bar\rho_1 = \delta_{\bar x}$ then $T(x) = \bar x$ is constant and obviously not injective.

4.2 Otto's calculus: from Benamou-Brenier to a Riemannian structure

In [Ott01], Otto generalized classical notions from Riemannian geometry (recall Section 1.3) to the Wasserstein space: the norm, the scalar product, and the gradient. We will not follow precisely the line of thought of the mentioned paper. Instead, we will use the Benamou-Brenier formula as a starting point for our reasoning.

Thanks to the Benamou-Brenier formula (Theorem 4.1.3), we have

$$\begin{aligned} W_2^2(\bar\rho_0,\bar\rho_1) &= \inf_{\rho_t,v_t}\left\{\int_0^1\left(\int_\Omega |v_t|^2\rho_t\,dx\right)dt \mid \partial_t\rho+\operatorname{div}(v_t\rho_t)=0,\ v_t\cdot\nu|_{\partial\Omega}=0,\ \rho_0=\bar\rho_0,\ \rho_1=\bar\rho_1\right\}\\ &= \inf_{\rho_t}\left\{\inf_{v_t}\int_0^1\int_\Omega |v_t|^2\rho_t\,dx\,dt \mid \partial_t\rho+\operatorname{div}(v_t\rho_t)=0,\ v_t\cdot\nu|_{\partial\Omega}=0,\ \rho_0=\bar\rho_0,\ \rho_1=\bar\rho_1\right\}\\ &= \inf_{\rho_t}\Bigg\{\int_0^1 \underbrace{\inf_{v_t}\left\{\int_\Omega |v_t|^2\rho_t\,dx \mid \operatorname{div}(v_t\rho_t)=-\partial_t\rho_t,\ v_t\cdot\nu|_{\partial\Omega}=0\right\}}_{=:\|\partial_t\rho_t\|^2_{\rho_t}} dt \mid \rho_0=\bar\rho_0,\ \rho_1=\bar\rho_1\Bigg\}, \end{aligned}$$

where in the last equality we used that, for each time t, given ρ_t and $\partial_t\rho_t$, one can minimize with respect to all vector fields v_t satisfying the constraint $\operatorname{div}(v_t\rho_t)=-\partial_t\rho_t$. In analogy with the formula for the Riemannian distance on manifold (see Definition 1.3.3), it is natural to define the *Wasserstein-norm* of the derivative $\partial_t\rho_t$ at ρ_t as

$$\|\partial_t\rho_t\|^2_{\rho_t} := \inf_{v_t}\left\{\int_\Omega |v_t|^2\rho_t\,dx \mid \operatorname{div}(v_t\rho_t)=-\partial_t\rho_t,\ v_t\cdot\nu|_{\partial\Omega}=0\right\}. \tag{4.4}$$

In other words the continuity equation gives, at each time t, a constraints on the divergence of $v_t\rho_t$, and we got the formula

$$W_2^2(\bar\rho_0,\bar\rho_1)=\inf_{\rho_t}\left\{\int_0^1\|\partial_t\rho_t\|^2_{\rho_t}\,dt \mid \rho_0=\bar\rho_0,\ \rho_1=\bar\rho_1\right\}.$$

To find a better formula for the Wasserstein-norm of $\partial_t\rho_t$ we want to understand the properties of the vector field v_t that realizes the infimum (in the definition (4.4)). Hence, given ρ_t and $\partial_t\rho_t$, let v_t be a minimizer, and let w be a vector field such that $\operatorname{div}(w)\equiv 0$. Then for every $\varepsilon>0$, we have

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

Thus $v_t+\varepsilon\frac{w}{\rho_t}$ is an admissible vector field in the minimization problem (4.4), and so by minimality of v_t we get

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

Dividing by ε and letting it go to zero yields

$$\int_\Omega\langle v_t,w\rangle=0$$

for every w such that $\operatorname{div}(w)\equiv 0$. By the Helmholtz decomposition (2.17), this implies that

$$v_t\in\{w\mid\operatorname{div}(w)=0\}^\perp=\{\nabla q\mid q:\Omega\to\mathbb{R}\}.$$

Therefore, there exists a function ψ_t such that $v_t = \nabla\psi_t$. Also, since $\operatorname{div}(v_t\rho_t) = -\partial_t\rho_t$ and $v_t \cdot \nu|_{\partial\Omega} = 0$, then ψ_t is a solution of

$$\begin{cases} \operatorname{div}(\rho_t \nabla\psi_t) = -\partial_t\rho_t & \text{in } \Omega, \\ \frac{\partial\psi_t}{\partial\nu} = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.5}$$

Note that, if ρ_t is "nice" (say, positive and smooth) then (4.5) is a uniformly elliptic equation with Neumann boundary conditions for ψ_t, and the solution ψ_t is unique up to a constant. So one can define

$$\|\partial_t\rho_t\|^2_{\rho_t} = \int_\Omega |\nabla\psi_t|^2 \rho_t \, dx,$$

where ψ_t solves (4.5). More in general, given $\rho \in \mathcal{P}_2(\Omega)$ and $h : \Omega \to \mathbb{R}$ such that $\int_\Omega h = 0$,[25] we can define

$$\|h\|^2_\rho := \int_\Omega |\nabla\psi|^2 \rho \, dx, \qquad \text{where} \qquad \begin{cases} \operatorname{div}(\rho\nabla\psi) = -h & \text{in } \Omega, \\ \frac{\partial\psi}{\partial\nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

Hence, we obtained a nice expression for the Wasserstein norm of the derivative of a curve.

Once a norm is defined, we can canonically construct the scalar product.

Definition 4.2.1. Given two functions $h_1, h_2 : \Omega \to \mathbb{R}$ with $\int_\Omega h_1 = \int_\Omega h_2 = 0$, one can define their *Wasserstein scalar product* at ρ as

$$\langle h_1, h_2\rangle_\rho := \int_\Omega \nabla\psi_1 \cdot \nabla\psi_2 \, \rho \, dx, \qquad \text{where} \qquad \begin{cases} \operatorname{div}(\rho\nabla\psi_i) = -h_i & \text{in } \Omega, \\ \frac{\partial\psi_i}{\partial\nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

Now that we have a scalar product, we can define the gradient of a functional in the Wasserstein space (cf. Definition 1.3.2).

Definition 4.2.2. Given a functional $F\colon \mathcal{P}_2(\Omega) \to \mathbb{R}$, its *gradient with respect to the Wasserstein scalar product* at $\bar\rho \in \mathcal{P}_2(\Omega)$ is the unique function $\operatorname{grad}_{W_2}\mathcal{F}[\bar\rho]$ (if it exists) such that

$$\Big\langle \operatorname{grad}_{W_2}\mathcal{F}[\bar\rho], \frac{\partial\rho_\varepsilon}{\partial\varepsilon}\Big|_{\varepsilon=0}\Big\rangle_{\bar\rho} = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{F}[\rho_\varepsilon]$$

for any smooth curve $\rho_\varepsilon : (-\varepsilon_0, \varepsilon_0) \to \mathcal{P}(\Omega)$ with $\rho_0 = \bar\rho$.

Before going on, let us spend some time to obtain a more explicit formula for the Wasserstein gradient of a functional. Given a functional $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ and a probability measure $\bar\rho \in \mathcal{P}_2(\Omega)$, let us denote with $\frac{\delta\mathcal{F}[\bar\rho]}{\delta\rho}$ its first L^2-variation[26], that is the function in $L^2(\Omega)$ such that

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{F}[\rho_\varepsilon] = \int_\Omega \frac{\delta\mathcal{F}[\bar\rho]}{\delta\rho}(x) \, \frac{\partial\rho_\varepsilon(x)}{\partial\varepsilon}\Big|_{\varepsilon=0} dx$$

[25] The condition $\int_\Omega h = 0$ is needed for the solvability of the elliptic equation

$$\begin{cases} \operatorname{div}(\rho\nabla\psi) = -h & \text{in } \Omega, \\ \frac{\partial\psi}{\partial\nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

since

$$\int_\Omega h \, dx = \int_\Omega \operatorname{div}(\rho\nabla\psi) \, dx = \int_{\partial\Omega} \frac{\partial\psi}{\partial\nu} \, \rho = 0.$$

Also, it is a classical fact that this is sufficient for solvability.

Note that, whenever ρ_t is a curve of probability measures, then

$$\int_\Omega \partial_t\rho_t \, dx = \frac{d}{dt}\int_\Omega \rho_t \, dx = \frac{d}{dt} 1 = 0.$$

[26] The first L^2-variation does not exist for all functionals, but it does for the ones we will consider.

for any (smooth) curve $\rho : (-\varepsilon_0, \varepsilon_0) \to \mathcal{P}_2(\Omega)$ such that $\rho_0 = \bar\rho$.

Then, by the definition of Wasserstein gradient,

$$\big\langle \mathrm{grad}_{W_2}\mathcal{F}[\bar\rho], \frac{\partial \rho_\varepsilon}{\partial \varepsilon}\Big|_{\varepsilon=0}\big\rangle_{\bar\rho} = \int_\Omega \frac{\delta \mathcal{F}[\bar\rho]}{\delta \rho}\frac{\partial \rho_\varepsilon}{\partial \varepsilon}\Big|_{\varepsilon=0} dx.$$

Thus, denoting by ψ the solution of $\mathrm{div}(\nabla\psi\,\bar\rho) = -\frac{\partial \rho_\varepsilon}{\partial \varepsilon}\big|_{\varepsilon=0}$ with zero Neumann boundary conditions, we have

$$\big\langle \mathrm{grad}_{W_2}\mathcal{F}[\bar\rho], \frac{\partial \rho_\varepsilon}{\partial \varepsilon}\Big|_{\varepsilon=0}\big\rangle_{\bar\rho} = -\int_\Omega \frac{\delta \mathcal{F}[\bar\rho]}{\delta \rho}\,\mathrm{div}(\nabla\psi\,\bar\rho)\,dx = \int_\Omega \nabla\frac{\delta \mathcal{F}[\bar\rho]}{\delta \rho}\cdot\nabla\psi\,\bar\rho\,dx$$

and therefore, by definition of Wasserstein scalar product, we deduce that

$$\mathrm{grad}_{W_2}\mathcal{F}[\bar\rho] = -\,\mathrm{div}\Big(\nabla\Big(\frac{\delta \mathcal{F}[\bar\rho]}{\delta \rho}\Big)\bar\rho\Big). \tag{4.6}$$

Example 4.2.3. If $\mathcal{F}[\rho] = \int_\Omega U(\rho(x))\,dx$ with $U : \mathbb{R} \to \mathbb{R}$, then for any smooth variation $\varepsilon \mapsto \rho_\varepsilon$ it holds

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\int_\Omega U(\rho_\varepsilon(x))\,dx = \int_\Omega U'(\bar\rho(x))\,\frac{\partial \rho_\varepsilon(x)}{\partial \varepsilon}\Big|_{\varepsilon=0} dx,$$

therefore the first L^2-variation of $\mathcal{F}[\rho]$ at $\bar\rho \in \mathcal{P}_2(\Omega)$ is given by

$$\frac{\delta \mathcal{F}[\bar\rho]}{\delta \rho}(x) = U'(\bar\rho(x)).$$

Using (4.6), this implies that the Wasserstein gradient of $\mathcal{F}$ is

$$\mathrm{grad}_{W_2}\mathcal{F}[\bar\rho] = -\,\mathrm{div}\left(\bar\rho\,\nabla[U'(\bar\rho)]\right) = -\,\mathrm{div}\left(\bar\rho\,U''(\bar\rho)\nabla\bar\rho\right).$$

In the special case $U(s) = s\log(s)$ (hence $\mathcal{F}$ is the entropy) one has $U''(s) = \frac{1}{s}$, thus

$$\mathrm{grad}_{W_2}\mathcal{F}[\bar\rho] = -\Delta\bar\rho\,.$$

If instead $U(s) = \frac{s^m}{m-1}$ for some $m \neq 1$, then we get

$$\mathrm{grad}_{W_2}\mathcal{F}[\bar\rho] = -\,\mathrm{div}\left(\bar\rho\,m\,\bar\rho^{m-2}\nabla\bar\rho\right) = -\Delta(\bar\rho^m)\,.$$

Example 4.2.4. If $\mathcal{F}[\rho] = \int_\Omega \rho(x)V(x)\,dx$ with $V : \Omega \to \mathbb{R}$, then its first L^2-variation at $\bar\rho \in \mathcal{P}_2(\Omega)$ is

$$\frac{\delta \mathcal{F}[\bar\rho]}{\delta \rho}(x) = V(x)\,,$$

therefore the Wasserstein gradient of $\mathcal{F}$ is

$$\mathrm{grad}_{W_2}\mathcal{F}[\bar\rho] = -\,\mathrm{div}\left(\nabla V\,\bar\rho\right).$$

Example 4.2.5. If $\mathcal{F}[\rho] = \frac{1}{2}\iint \rho(x)\,\rho(y)\,W(x-y)\,dx\,dy$ with $W : \mathbb{R}^d \to \mathbb{R}$ such that $W(z) = W(-z)$, then its first L^2 variation at $\bar\rho$ is

$$\frac{\delta \mathcal{F}[\bar\rho]}{\delta \rho}(x) = W * \bar\rho(x) = \int_\Omega W(x-y)\,\rho(y)\,dy\,,$$

where $*$ denotes the convolution, and therefore the Wasserstein gradient of $\mathcal{F}$ is

$$\mathrm{grad}_{W_2}\mathcal{F}[\bar\rho] = -\,\mathrm{div}\left((\nabla W * \bar\rho)\bar\rho\right).$$

The definition of gradient flow in the Wasserstein space is (at least on a purely formal level) exactly the expected one.

Definition 4.2.6. Given a functional $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$, a curve of probability measure $\rho : [0,T) \to \mathcal{P}_2(\Omega)$ is a gradient flow of $\mathcal{F}$ with respect to W_2 and with starting point $\bar\rho_0$ if

$$\begin{cases} \partial_t \rho_t = -\mathrm{grad}_{W_2}\mathcal{F}[\rho_t] \\ \rho_0 = \bar\rho_0. \end{cases}$$

By the computation in Example 4.2.3, the Wasserstein gradient flow of the entropy functional $\mathcal{F}[\rho] = \int_\Omega \rho \log(\rho)\, dx$ is the heat equation

$$\partial_t \rho = -\mathrm{grad}_{W_2}\mathcal{F}[\rho] = \Delta\rho,$$

as expected from what we proved in Section 3.3.

On the other hand, if $\mathcal{F}[\rho] = \frac{1}{m-1}\int_\Omega \rho^m$ for $m \neq 1$ with $m > 0$, then the gradient flow is (cf. Example 4.2.3)

$$\partial_t \rho = \Delta(\rho^m)\,,$$

that is, the porous medium equation (if $m > 1$) or the fast diffusion equation (if $m \in (0,1)$) .

4.3 Displacement convexity

We move our attention to the convexity properties of functionals in the Wasserstein space. As we observed in Section 3.2, convexity of a functional is extremely useful: indeed, not only it allows one to define the concept of gradient flow via the notion of subdifferential, but more importantly it implies existence and uniqueness of the gradient flow (at least in the Hilbertian setting).[27]

The first author to introduce and investigate the convexity of functionals in the Wasserstein space (i.e., convexity along W_2-geodesics) was McCann in [McC97].

Definition 4.3.1. We say that a functional $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ is *W_2-convex*, or *displacement convex*, if the 1-dimensional function

$$[0,1] \ni t \mapsto \mathcal{F}[\rho_t]$$

is convex for any W_2-geodesic $\rho : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$.

Let us focus first on the special case $\mathcal{F}[\rho] = \int_\Omega U(\rho)\, dx$, with $U : \mathbb{R} \to \mathbb{R}$. We want to understand under which assumption on U the functional $\mathcal{F}$ is W_2-convex.

Given $\rho_0 \in \mathcal{P}_2(\Omega)$, let $\varphi : \Omega \to \mathbb{R}$ be a smooth convex function, and set $T := \nabla\varphi$ and $\rho_1 := T_\#\rho_0$. It follows by Remark 2.5.8 that T is the optimal map from ρ_0 to ρ_1, and the Wasserstein geodesic connecting these two measures is given by $\rho_t := (T_t)_\#\rho_0$ with $T_t(x) = x + t(T(x) - x)$ (see Section 3.1.1).

Recalling Section 1.6, we have that $\rho_t \circ T_t = \frac{\rho_0}{\det \nabla T_t}$. Therefore, since $\nabla T_t = (1-t)\mathrm{Id} + t\, D^2\varphi$,

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

[27] It is however important to remark that convexity is not needed to define gradient flows (see [AGS08]), and it is actually possible to prove existence under very weak assumption. More challenging is usually to prove uniqueness, but there are several cases of interests where uniqueness can still be proved without relying on convexity (see, for example, [FG10; FGY11]).

Since φ is convex, $D^2\varphi$ is a nonnegative symmetric matrix. For any $x \in \Omega$, let $\lambda_1(x), \dots, \lambda_d(x) \geq 0$ be the eigenvalues of $D^2\varphi(x)$. It holds

$$D(x,t) \coloneqq \det\big((1-t)\mathrm{Id} + tD^2\varphi(x)\big)^{1/d} = \det \begin{pmatrix} (1-t)+t\lambda_1(x) & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & (1-t)+t\lambda_d(x) \end{pmatrix}^{1/d}$$
$$= \prod_{i=1}^{d} \big((1-t)+t\lambda_i(x)\big)^{1/d}.$$

We leave as an exercise to the reader to prove, using the identity above involving the eigenvalues, that $t \mapsto D(x,t)$ is concave (for a proof, see [Vil03, Lemma 5.21]). Then we can rewrite

$$\mathcal{F}[\rho_t] = \int_\Omega U\left(\frac{\rho_0(x)}{D(x,t)^d}\right) D(x,t)^d \, dx.$$

We now ask ourselves:
When is the map $t \mapsto U\left(\frac{\rho_0(x)}{D(x,t)^d}\right) D(x,t)^d$ convex for every x?

Since $t \mapsto D(x,t)$ is concave, a sufficient condition is that the function $(0,\infty) \ni s \mapsto U\left(\frac{1}{s^d}\right) s^d$ is convex and nonincreasing.[28] Hence, we have proven the following important:

Proposition 4.3.2. *Let $U : [0,\infty) \to \mathbb{R}$ satisfy*

$$(0,\infty) \ni s \mapsto U\Big(\frac{1}{s^d}\Big) s^d \quad \textit{is convex and nonincreasing.}$$

Then the functional $\mathcal{F}[\rho] \coloneqq \int_\Omega U(\rho)\,dx$ is W_2-convex.

Example 4.3.3. Let $\mathcal{F}[\rho] = \int U(\rho(x))\,dx$. Using the above mentioned criterion, it is not hard to show that the following choices yield W_2-convex functionals:

$$U(s) \coloneqq \begin{cases} s\log(s) & & \Rightarrow & \partial_t\rho_t = \Delta\rho & \text{(Heat eq.)} \\ \frac{1}{m-1}s^m & \text{for } m > 1 & \Rightarrow & \partial_t\rho_t = \Delta(\rho^m) & \text{(Porous medium eq.)} \\ \frac{1}{m-1}s^m & \text{for } m \in [1-\frac{1}{d}, 1) & \Rightarrow & \partial_t\rho_t = \Delta(\rho^m) & \text{(Fast diffusion eq.)} \end{cases}$$

Let us conclude this section with two criteria useful to establish the convexity of other kind of functionals.

1. If $V : \mathbb{R}^d \to \mathbb{R}$ is convex, then the functional $\mathcal{F}[\rho] \coloneqq \int V\rho\,dx$ is W_2-convex.
2. If $W : \mathbb{R}^d \to \mathbb{R}$ is convex, then the functional $\mathcal{F}[\rho] \coloneqq \iint W(x-y)\rho(x)\rho(y)\,dx\,dy$ is W_2-convex.

Let us prove the two criteria together.

Consider the geodesic $\rho_t = (T_t)_\#\rho_0$, with $T_t(x) = (1-t)x + tT(x)$ where T is the optimal map between ρ_0 and ρ_1. We have that the two functionals can be expressed as

$$\int_\Omega V(x)\,\rho_t(x)\,dx = \int_\Omega V(T_t(x))\rho_0(x)\,dx\,,$$
$$\iint_{\Omega\times\Omega} W(x-y)\,\rho_t(x)\,\rho_t(y)\,dx\,dy = \iint_{\Omega\times\Omega} W(T_t(x)-T_t(y))\,\rho_0(x)\rho_0(y)\,dx\,dy\,.$$

[28] Indeed, fixed $x \in \Omega$, set $G(s) \coloneqq U(\frac{1}{s^d})s^d$ and $H(t) \coloneqq \frac{D(x,t)}{\rho_0(x)^{1/d}}$. Then

$$U\left(\frac{\rho_0(x)}{D(x,t)^d}\right) D(x,t)^d = \rho_0(x)\,G\big(H(t)\big).$$

Hence, since $G' \leq 0$, $G'' \geq 0$, and $H'' \leq 0$, we get

$$\frac{d^2}{dt^2}G(H(t)) = G''(H(t))\,[H'(t)]^2 + G'(H(t))\,H''(t) \geq 0\,.$$

Since the map $t \mapsto T_t(x)$ is affine, the two functions $t \mapsto V(T_t(x))$ and $t \mapsto W(T_t(x) - T_t(y))$ are convex for any $x, y \in \mathbb{R}^d$ (here we are using the convexity of V and W), and thus the two functionals are convex along the geodesic ρ_t, as desired.

4.4 An excursion into the linear Fokker-Planck equation

We are going to apply all the tools that we developed so far to the case of the linear Fokker-Planck equation, that is

$$\partial_t \rho = \Delta \rho + \operatorname{div}(\nabla V\, \rho),$$

where $\rho : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}^+$ is a nonnegative function, and $V : \mathbb{R}^d \to \mathbb{R}$ is a C^2 convex function. Here, $\rho(t)$ usually represents a density of particles, while V plays the role of an external confining potential.

As we will see, this equation is a gradient flow in the Wasserstein space, and this perspective allows us to obtain quantitative convergence rates to the equilibrium for the solution $\rho(t)$ as $t \to +\infty$. A nice perk of this strategy is that, as a byproduct, it yields a proof of the logarithmic Sobolev inequality.

The functional and its gradient. Let us consider the functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ on the Wasserstein space (as usual, $\mathcal{F}[\rho] := +\infty$ if $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ is not absolutely continuous with respect to the Lebesgue measure)

$$\mathcal{F}[\rho] := \int_{\mathbb{R}^d} \big(\rho \log(\rho) + \rho V\big)\, dx.$$

Notice that we can rewrite

$$\mathcal{F}[\rho] = \int_{\mathbb{R}^d} \eta \log(\eta) e^{-V}\, dx, \qquad \eta := e^V \rho, \tag{4.7}$$

therefore the functional $\mathcal{F}$ can be seen as the relative entropy with respect to the measure $e^{-V} dx$.

Assuming that $\int_{\mathbb{R}^d} e^{-V}\, dx < \infty$, up to adding a constant to V we can assume that e^{-V} is a probability measure, i.e., $\int_{\mathbb{R}^d} e^{-V}\, dx = 1$. Since $[0, \infty) \ni s \mapsto s \log(s)$ is a convex function, Jensen's inequality implies

$$\begin{aligned}\mathcal{F}[\rho] = \int_{\mathbb{R}^d} \eta \log(\eta) e^{-V}\, dx &\geq \left(\int_{\mathbb{R}^d} \eta e^{-V}\, dx\right) \log\left(\int_{\mathbb{R}^d} \eta e^{-V}\, dx\right) \\ &= \left(\int_{\mathbb{R}^d} \rho\, dx\right) \log\left(\int_{\mathbb{R}^d} \rho\, dx\right) = 0 \end{aligned} \tag{4.8}$$

(recall that ρ is a probability density), thus the functional $\mathcal{F}$ is nonnegative. Also one can check that equality holds if and only if $\eta = 1$, thus $\mathcal{F}[e^{-V}] = 0$ is the only minimum.

Recalling (4.6), computing the Wasserstein gradient of $\mathcal{F}$ is routine:

$$\operatorname{grad}_{W_2} \mathcal{F}[\rho] = -\operatorname{div}\big(\nabla \rho + \rho \nabla V\big), \tag{4.9}$$

and its Wasserstein norm is given by (recall Definition 4.2.1, and that $\rho = e^{-V}\eta$)

$$\begin{aligned}\langle \operatorname{grad}_{W_2} \mathcal{F}[\rho], \operatorname{grad}_{W_2} \mathcal{F}[\rho] \rangle_\rho &= \int_{\mathbb{R}^d} \left| \frac{\nabla \rho + \rho \nabla V}{\rho} \right|^2 \rho\, dx \\ &= \int_{\mathbb{R}^d} \frac{|e^{-V} \nabla \eta - \eta \nabla V e^{-V} + \eta e^{-V} \nabla V|^2}{e^{-V} \eta}\, dx = \int_{\mathbb{R}^d} \frac{|\nabla \eta|^2}{\eta} e^{-V}\, dx. \end{aligned} \tag{4.10}$$

λ-convexity. We want to investigate the convexity of the functional $\mathcal{F}$ defined above. Exploiting the criteria described in Section 4.3, one easily checks that $\mathcal{F}$ is W_2-convex (the entropy is convex, and $\int \rho V\,dx$ is convex since V is convex by assumption). Unfortunately, for our purposes, this *nonquantitative* form of convexity is not sufficient. Hence we introduce the notion of λ-convexity.

Definition 4.4.1. Let $I \subset \mathbb{R}$ be an interval, and let $\varphi : I \to \mathbb{R}\cup\{+\infty\}$ be a lower semicontinuous function. Given $\lambda \in \mathbb{R}$, the function φ is said to be λ-convex if

$$(1-s)\varphi(x) + s\varphi(y) \geq \varphi((1-s)x + sy) + \frac{\lambda s(1-s)}{2}|x-y|^2 \qquad \forall\, x, y \in I,\ 0 \leq s \leq 1.$$

A lower semicontinuous function $\varphi : X \to \mathbb{R} \cup \{+\infty\}$ on a geodesic metric space (X, d) is said λ-convex if, given any geodesic $\gamma : [0,1] \to X$, the function $\varphi\circ\gamma : [0,1] \to \mathbb{R}\cup\{+\infty\}$ is λ-convex.

Notice that the notion of λ-convexity is stronger than convexity for $\lambda > 0$, and weaker for $\lambda < 0$ (and equivalent for $\lambda = 0$).

Also, λ-convexity behaves well under addition:

Lemma 4.4.2. *Let $i = 1, 2$, and let $\varphi_i : X \to \mathbb{R} \cup \{+\infty\}$ be λ_i-convex. Then $\varphi_1 + \varphi_2$ is $(\lambda_1 + \lambda_2)$-convex.*

Proof. Let $\gamma : [0,1] \to X$ be a geodesic. By the λ_i-convexity of φ_i it holds

$$(1-s)\varphi_i(\gamma(\tau)) + s\varphi_i(\gamma(\sigma)) \geq \varphi_i((1-s)\gamma(\tau) + s\gamma(\sigma)) + \frac{\lambda_i s(1-s)}{2}|\gamma_i(\tau) - \gamma_i(\sigma)|^2$$

for all $\sigma, \tau \in [0,1]$, $0 \leq s \leq 1$, $i = 1, 2$. Adding the two inequalities over $i = 1, 2$, we get the result. □

In order to understand the meaning of this definition, note that a function $\varphi : I \to \mathbb{R}$ is λ-convex if and only if the map

$$[0,1] \ni x \to \varphi(x) - \frac{\lambda}{2}|x|^2$$

is convex. In particular, if φ is of class C^2, this is true if and only if $\varphi'' \geq \lambda$.

In the case $X = \mathbb{R}^d$, one can easily check that the same characterization holds: $\varphi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is λ-convex if and only if

$$\mathbb{R}^d \ni x \mapsto \varphi(x) - \frac{\lambda}{2}|x|^2$$

is convex. Hence, given a λ-convex function $\varphi \in C^1(\mathbb{R}^d)$, applying the inequality

$$\psi(y) \geq \psi(x) + \langle \nabla\psi(x), y - x\rangle, \qquad \text{with} \quad \psi(z) \coloneqq \varphi(z) - \frac{\lambda}{2}|z|^2,$$

we get

$$\varphi(y) \geq \varphi(x) + \langle \nabla\varphi(x), y - x\rangle + \frac{\lambda}{2}|y-x|^2 \qquad \forall\, x, y \in \mathbb{R}^d. \tag{4.11}$$

Exchanging the role of x, y in (4.11) and adding the two inequalities, we also get

$$\langle x - y, \nabla\varphi(x) - \nabla\varphi(y)\rangle \geq \lambda|x-y|^2\ . \tag{4.12}$$

Exploiting these inequalities, we now prove two useful properties of λ-convex functions.

Consider a λ-convex function $\varphi \in C^1(\mathbb{R}^d, \mathbb{R})$ with $\lambda > 0$, and let x_0 be the unique minimum of φ.

(i) Applying (4.11), since $\nabla\varphi(x_0) = 0$, we deduce

$$\varphi(x) \geq \varphi(x_0) + \frac{\lambda}{2}|x - x_0|^2 \quad \Longrightarrow \quad \sqrt{\frac{2}{\lambda}\big(\varphi(x) - \varphi(x_0)\big)} \geq |x - x_0|. \tag{4.13}$$

(ii) Applying (4.11) again, we deduce

$$\varphi(x_0) \geq \varphi(x) + \langle \nabla\varphi(x), x_0 - x\rangle + \frac{\lambda}{2}|x - x_0|^2$$
$$\Longrightarrow \quad \langle \nabla\varphi(x), x - x_0\rangle \geq \varphi(x) - \varphi(x_0) + \frac{\lambda}{2}|x - x_0|^2$$

therefore

$$|\nabla\varphi(x)| \geq \frac{\varphi(x) - \varphi(x_0)}{|x - x_0|} + \frac{\lambda}{2}|x - x_0| \geq \sqrt{2\lambda\big(\varphi(x) - \varphi(x_0)\big)}, \tag{4.14}$$

where we used the inequality

$$2\sqrt{ab} \leq a + b \qquad \forall\, a, b \geq 0$$

with $a = \frac{\varphi(x)-\varphi(x_0)}{|x-x_0|}$ and $b = \frac{\lambda}{2}|x - x_0|$.

Note that the proofs above work also on a general geodesic metric space, dropping also the assumption of C^1-regularity, provided that we can give a meaning to $|\nabla\varphi|$. This can be done in a very general setting (see [AGS08]), but for us it is sufficient to notice that the results above are true in the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. More precisely, the following holds:

Lemma 4.4.3. *Given a λ-convex lower semicontinuous functional $\mathcal{F} : P_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$, with $\lambda > 0$, let $\min \mathcal{F} = \mathcal{F}[\bar\rho]$. The natural generalizations of (4.13) and (4.14) read as*

$$W_2^2(\rho, \bar\rho) \leq \frac{2}{\lambda}\big(\mathcal{F}[\rho] - \mathcal{F}[\bar\rho]\big), \tag{4.15}$$

$$\mathcal{F}[\rho] - \mathcal{F}[\bar\rho] \leq \frac{1}{2\lambda}\langle \mathrm{grad}_{W_2}\mathcal{F}[\rho], \mathrm{grad}_{W_2}\mathcal{F}[\rho]\rangle_\rho, \tag{4.16}$$

for all $\rho \in \mathcal{P}_2(\mathbb{R}^d)$.

Proof. These bounds can be shown by mimicking the 1-dimensional proof on a geodesic between $\bar\rho$ and ρ. To show the method, we prove (4.15) and leave (4.16) to the interested reader.

Let $L := W_2(\rho, \bar\rho)$, let $\gamma \in \Gamma(\bar\rho, \rho)$ be a optimal plan, and set

$$\pi_t(x, y) := (L - t)x + ty, \quad t \in [0, L], \qquad \hat\rho_t := (\pi_t)_\# \gamma.$$

Repeating the argument in Section 3.1.1 one can show that $\hat\rho_t$ is a unit-speed W_2-geodesic.

Let $\hat\varphi : [0, L] \to \mathbb{R}$ be the composition $\hat\varphi(t) := \mathcal{F}[\hat\rho(t)]$. Since φ is λ-convex it follows (by definition) that $\hat\varphi$ is λ-convex, hence we can repeat the argument above to get

$$\sqrt{\frac{2}{\lambda}\big(\hat\varphi(L) - \hat\varphi(0)\big)} \geq L,$$

which is precisely (4.15). □

Let us move back to the study of the functional $\mathcal{F}[\rho] = \int_{\mathbb{R}^d} (\rho\log(\rho) + V\rho)\, dx$. Recall that $\int_{\mathbb{R}^d} \rho\log(\rho)$ is convex (see Example 4.3.3). Also, we proved that if $V : \mathbb{R}^d \to \mathbb{R}$ is convex then the functional $\rho \mapsto \int V\rho\, dx$ is convex, and the same proof shows that if $V : \mathbb{R}^d \to \mathbb{R}$ is λ-convex then $\rho \mapsto \int V\rho\, dx$ is λ-convex. Hence, thanks to Lemma 4.4.2, we obtain the following:

Proposition 4.4.4. *Consider the functional $\mathcal{F}[\rho] = \int_{\mathbb{R}^d} (\rho\log(\rho) + V\rho)\, dx$, and assume that $V : \mathbb{R}^d \to \mathbb{R}$ is λ-convex. Then $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ is λ-convex. In particular, if $\lambda > 0$ then (4.15) and (4.16) hold.*

Log-Sobolev inequality. Consider a λ-convex function $V : \mathbb{R}^d \to \mathbb{R}$ with $\lambda > 0$, and assume that $\int_{\mathbb{R}^d} e^{-V} = 1$. Given a function $\eta : \mathbb{R}^d \to [0, \infty)$ such that $\int_{\mathbb{R}^d} \eta e^{-V}\, dx = 1$, set $\rho := \eta e^{-V} \in \mathcal{P}(\mathbb{R}^d)$. Thanks to Proposition 4.4.4 we can apply (4.16), (4.8), and (4.10), to get to get

$$\int_{\mathbb{R}^d} \eta \log(\eta) e^{-V}\, dx = \mathcal{F}[\rho] \le \frac{1}{2\lambda} \langle \mathrm{grad}_{W_2} \mathcal{F}[\rho], \mathrm{grad}_{W_2} \mathcal{F}[\rho] \rangle_\rho = \frac{1}{2\lambda} \int_{\mathbb{R}^d} \frac{|\nabla \eta|^2}{\eta} e^{-V}\, dx.$$

Hence, this proves that

$$\int_{\mathbb{R}^d} \eta \log(\eta) e^{-V}\, dx \le \frac{1}{2\lambda} \int_{\mathbb{R}^d} \frac{|\nabla \eta|^2}{\eta} e^{-V}\, dx \qquad \forall\, \eta : \mathbb{R}^d \to [0, \infty) \text{ s.t. } \int_{\mathbb{R}^d} \eta e^{-V}\, dx = 1.$$

This inequality is known in the literature as the log-Sobolev inequality (see [Led01, Section 5.1]), and it is well-known that the constant $\frac{1}{2\lambda}$ is sharp.

Convergence to the equilibrium. Again, given a λ-convex function $V : \mathbb{R}^d \to \mathbb{R}$ with $\lambda > 0$ and $\int_{\mathbb{R}^d} e^{-V} = 1$, we consider the functional $\mathcal{F}[\rho] = \int_{\mathbb{R}^d} (\rho \log(\rho) + V\rho)\, dx$. Thanks to the formula (4.9), we know that the Wasserstein gradient flow associated to $\mathcal{F}$ is the linear Fokker-Planck equation

$$\begin{cases} \partial_t \rho = \Delta \rho + \mathrm{div}(\nabla V\, \rho), \\ \rho(0) = \bar\rho \in \mathcal{P}_2(\mathbb{R}^d). \end{cases} \tag{4.17}$$

For simplicity, we assume that $\mathcal{F}[\bar\rho] < +\infty$. (This assumption is not strictly necessary, as one can prove that $\mathcal{F}[\rho_t] < +\infty$ for any positive time $t > 0$.) Our goal is to understand the behavior of $\rho_t := \rho(t)$ as $t \to \infty$. We have already shown that e^{-V} is a minimum point of $\mathcal{F}$, hence we expect ρ_t to converge to e^{-V}, as remarked Section 3.2. Let us state (3.10) in the form we need here:

$$\frac{d}{dt} \mathcal{F}[\rho_t] = \langle \mathrm{grad}_{W_2} \mathcal{F}[\rho_t], \partial_t \rho_t \rangle_{W_2} = -\langle \mathrm{grad}_{W_2} \mathcal{F}[\rho_t], \mathrm{grad}_{W_2} \mathcal{F}[\rho_t] \rangle_{\rho_t} \le 0\ . \tag{4.18}$$

The functional $\mathcal{F}$ shall be interpreted as the energy of the system, that tends to its minimum as time evolves. Let us prove

$$\mathcal{F}[\rho_t] \to \mathcal{F}[e^{-V}] = 0 \quad \text{as } t \to \infty.$$

Combining (4.18) with (4.16) we get (recall that $\min \mathcal{F} = 0$)

$$\frac{d}{dt} \mathcal{F}[\rho_t] \le -2\lambda\, \mathcal{F}[\rho_t]$$

and therefore

$$\frac{d}{dt} \Big(\mathcal{F}[\rho_t] e^{2\lambda t} \Big) \le 0 \quad \Longrightarrow \quad \mathcal{F}[\rho_t] \le e^{-2\lambda t} \mathcal{F}[\bar\rho]. \tag{4.19}$$

Hence we have shown that the energy converges to 0 exponentially fast. Since the energy controls the Wasserstein distance to the equilibrium e^{-V} (recall (4.15)), we deduce immediately

$$W_2^2(\rho_t, e^{-V}) \le \frac{2}{\lambda} \mathcal{F}[\bar\rho] e^{-2\lambda t} \qquad \forall\, t \ge 0.$$

Remark 4.4.5. Thanks to the λ-convexity of $\mathcal{F}$, a stronger form of the latter inequality holds, namely

$$W_2^2(\rho_t, e^{-V}) \le e^{-2\lambda t} W_2^2(\bar\rho, e^{-V})\ .$$

In fact, this last inequality holds even if we replace e^{-V} with any curve of probability measures $\tilde\rho : [0, +\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ that is a gradient-flow with respect to the functional $\mathcal{F}$; more precisely, it holds

$$W_2^2(\rho_t, \tilde\rho_t) \le e^{-2\lambda t} W_2^2(\rho_0, \tilde\rho_0) \qquad \forall\, t \ge 0. \tag{4.20}$$

This property of the gradient-flow is usually called *contractivity*. The proof is rather technical, and we refer to [AGS08] for a proof. Here, we show it only in the simpler case of gradient-flows on $\mathbb{R}^d$.

Take a smooth λ-convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$, and consider two curves $x, y : [0, +\infty) \to \mathbb{R}^d$ that solve the gradient-flow equation

$$\dot{x}(t) = -\nabla\varphi(x(t)), \qquad \dot{y}(t) = -\nabla\varphi(y(t)).$$

Then

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

where in the last inequality we applied (4.12). Hence we have

$$\frac{d}{dt}\Big(\|x(t)-y(t)\|^2 e^{2\lambda t}\Big) \le 0 \implies \|x(t)-y(t)\|^2 \le e^{-2\lambda t}\|x(0)-y(0)\|^2 \,,$$

which is the Euclidean analogue of (4.20).

What we have proved until now is very robust and can be applied verbatim to several other functionals on $\mathcal{P}_2(\mathbb{R}^d)$. We want to conclude with a convergence result that cannot be easily adapted to other functionals: ρ_t converges to e^{-V} in the L^1-norm. More precisely, we claim that if ρ_t solves (4.17) then

$$\|\rho_t - e^{-V}\|_{L^1}^2 \le 2e^{-2\lambda t}\mathcal{F}[\bar{\rho}] \,.$$

This convergence follows from (4.19) together with the following inequality (that is valid for *any* $V : \mathbb{R}^d \to \mathbb{R}$ with $e^{-V} \in \mathcal{P}(\mathbb{R}^d)$):

$$\frac{1}{2}\|\rho - e^{-V}\|_{L^1}^2 \le \mathcal{F}[\rho] \qquad \forall\, \rho \in \mathcal{P}(\mathbb{R}^d). \tag{4.21}$$

This inequality is known in literature as Csiszár-Kullback-Pinsker inequality, see the introduction of [BV05].

Before proving (4.21), let us observe that the coefficient $\frac{1}{2}$ is optimal. Indeed, if $\rho = \eta e^{-V} \in \mathcal{P}(\mathbb{R}^d)$ with $\eta : \mathbb{R}^d \to \{1 \pm \varepsilon\}$ everywhere, then

$$\begin{aligned}\mathcal{F}[\rho] &= \int_{\mathbb{R}^d} \eta \log \eta e^{-V} = \int_{\mathbb{R}^d} \Big((\eta-1) + \frac{1}{2}(\eta-1)^2 + \mathcal{O}(|\eta-1|^3)\Big)e^{-V} \\ &= \frac{\varepsilon^2}{2} + \mathcal{O}(\varepsilon^3) = \frac{1+\mathcal{O}(\varepsilon)}{2}\|\rho - e^{-V}\|_{L_1}^2.\end{aligned}$$

We now prove (4.21).

Proof of (4.21). In this proof we denote $\mu = e^{-V}$ and $\rho = \eta e^{-V}$. Let us remark that the convexity of V is not necessary, hence the result holds for any probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$.

Applying Jensen's inequality, we have

$$\mathcal{F}[\rho] = \int_{\{\eta>1\}} \eta\log\eta\, d\mu + \int_{\{\eta\le 1\}} \eta\log\eta\, d\mu \ge \alpha(1+x)\log(1+x) + \beta(1-y)\log(1-y)\,, \tag{4.22}$$

where $\alpha := \mu(\{\eta > 1\})$, $\beta := \mu(\{\eta \le 1\})$, and x, y are defined so that

$$1 + x := \alpha^{-1}\int_{\{\eta>1\}} \eta\, d\mu \,, \quad 1 - y := \beta^{-1}\int_{\{\eta\le 1\}} \eta\, d\mu \,.$$

As a direct consequence of the definitions, it holds $\alpha, \beta \geq 0$, $\alpha + \beta = 1$, $x \geq 0$ and $0 \leq y \leq 1$. Also, since $\int_{\mathbb{R}^d} \eta \, d\mu = 1$, we have $\alpha x = \beta y$, therefore

$$\begin{cases} \alpha + \beta = 1, \\ \alpha x = \beta y \end{cases} \qquad \Longrightarrow \qquad \begin{cases} \alpha = \frac{y}{x+y} \\ \beta = \frac{x}{x+y}. \end{cases} \tag{4.23}$$

Furthermore the following identity holds:

$$\|\rho - e^{-V}\|_{L^1} = \int_{\mathbb{R}^d} |\eta - 1| \, d\mu = \alpha x + \beta y = \frac{2xy}{x+y} \ . \tag{4.24}$$

Combining (4.22), (4.23), and (4.24), the sought inequality (4.21) boils down to proving that

$$Z(x,y) \coloneqq \frac{1+x}{x} \log(1+x) + \frac{1-y}{y} \log(1-y) - \frac{2xy}{x+y} \geq 0 \qquad \forall \, x \geq 0, \, 0 \leq y \leq 1. \tag{4.25}$$

One can easily check that the function Z is continuous on $[0, +\infty) \times [0,1]$, and that

$$Z \geq 0 \qquad \text{on } \{x = 0\} \cup \{y = 0\} \cup \{y = 1\}.$$

Also, $Z(x,y) \to +\infty$ as $x \to +\infty$, hence there exists $R \gg 1$ such that $Z > 0$ for $x \geq R$. So it remains to check that $Z \geq 0$ inside $(0,R) \times (0,1)$.

If Z has no minimum point in $(0,R) \times (0,1)$, then $Z \geq 0$ everywhere (since it must attain its minimum on the boundary, where we know that Z is nonnegative). Hence, it suffices to show that $Z \geq 0$ at all critical points in $(0,R) \times (0,1)$ (notice that Z is smooth inside $(0,R) \times (0,1)$). Note that the critical point condition $\nabla Z(x,y) = (0,0)$ yields

$$\begin{aligned} 0 = \partial_x Z(x,y) \quad &\Longleftrightarrow \quad \frac{\log(1+x)}{x} = 1 - \frac{2xy^2}{(x+y)^2} \ , \\ 0 = \partial_y Z(x,y) \quad &\Longleftrightarrow \quad \frac{\log(1-y)}{y} = -1 - \frac{2x^2 y}{(x+y)^2} \ . \end{aligned} \tag{4.26}$$

Hence, using (4.26), (4.25) simplifies to

$$Z(x,y) = (1+x)\Big(1 - \frac{2xy^2}{(x+y)^2}\Big) + (1-y)\Big(-1 - \frac{2x^2 y}{(x+y)^2}\Big) - \frac{2xy}{x+y} = \frac{(x-y)^2}{x+y} \ ,$$

which is clearly nonnegative, concluding the proof. □

5 Further readings

Now that the reader has a basic knowledge of optimal transport, we present here a list of possible references in order to learn about some of the several applications of this beautiful theory. Our list of references is far from complete and shall be seen as a starting point for the study of these topics.

5.1 Functional and geometric inequalities

As we have seen in Section 1.5, transport theory can be used to prove the isoperimetric inequality. We also observed in Section 4.4 how the formalism of Otto's calculus yields a neat proof of the logarithmic Sobolev inequality. These two results are only the tip of the iceberg about the deep connection between optimal transport and functional/geometric inequalities. In recent years, optimal transport methods have been used to prove already known inequalities with new proofs, to establish new ones, and to obtain quantitative version of well-known ones. Let us mention some of the most remarkable examples.

The paper [CENV04] exploits optimal transport to prove the Sobolev inequality and some Gagliardo-Niremberg inequalities with sharp constants. In a similar fashion, the logarithimic Sobolev inequality, and several other related inequalities, are obtained in [CE02]. The Brunn-Minkowsky inequality (which is a geometric inequality that has the isoperimetric inequality as a byproduct) and several generalizations are proven in [McC97; CEMS01]. In [FMP10], the authors (starting from a variant of the proof of the isoperimetric that is contained in Section 1.5) manage to prove a sharp quantitative version of the anisotropic isoperimetric inequality. We remark that all these results rely on the structure/properties of optimal transport maps for the quadratic cost $c(x,y) = |x-y|^2$ (cf. Theorem 2.5.9).

Very recently, optimal transport with the distance cost $c(x,y) = d(x,y)$ has been used in [Kla17] to give a beautiful alternative proof of the isoperimetric inequality of Lévy–Gromov (and also of many other important inequalities) on weighted Riemannian manifolds with lower Ricci curvature bounds. The idea is to use an optimal transport plan γ for the cost $c = d$ to construct a foliation of the ambient space M, by considering the union of the geodesics connecting x to y for $(x,y) \in \mathrm{supp}(\gamma)$. This idea is a generalization of the so-called "needle decomposition", a deep localization method used in convex geometry to reduce the inequality from M to a 1-dimensional version along geodesics (called needles). It is worth noticing that the argument in [Kla17] does not rely on the deep regularity theory for isoperimetric minimizers. In particular, among several important applications, this method has been used to prove the Lévy–Gromov inequality on metric measure spaces satisfying a curvature-dimension condition in [CM17].

5.2 Probability

There are many directions of research that have been investigated regarding the applications of optimal transport in probability. Here we mention just three of them: martingale optimal transport, the quantitative central limit theorem, and the random matching problem.

The martingale optimal transport problem (which finds one of its motivations in mathematical finance) is a slight modification of the Kantorovich problem. We state it in purely probabilistic language, but of course it can be framed in a measure-theoretic setting using the disintegration theorem. Given two distributions $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and a cost function $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the goal is to minimize $\mathbb{E}[c(X,Y)]$ among all pairs (X,Y) of random variables with $X \sim \mu$, $Y \sim \nu$ satisfying the martingale condition $\mathbb{E}[Y|X] = X$. Without entering in technical details, let us only say that, even if it has various similarities with the Kantorovich problem (i.e., a dual formulation), many new ideas are necessary to tackle the martingale optimal transport problem. The interested reader might explore this beautiful problem starting from the two recent papers [BJ16; BNT17].

The central limit theorem states that, if $(X_i)_{i\in\mathbb{N}}$ is a sequence of independent random variables with the same law, mean equal to 0, and variance equal to 1, then the average $\bar{X}_k = \frac{X_1+\cdots+X_k}{k}$ converges in law, as $k \to \infty$, to a standard Gaussian. One may desire to make this statement quantitative: how far is the law of $\bar{X}_k$ from a Gaussian? Of course, optimal transport can be a tool to measure their distance. This is exactly the problem considered, and solved, by the authors in [BGM99].

Let us conclude with the random matching problem. Let $X_1, \dots, X_k$ be k independent points uniformly distributed on the interval $[0, 1]$. One would expect, intuitively, that the distribution of this k points approximates, as k grows, the uniform measure on $[0, 1]$. To make precise this intuition, let $\mu^k = \frac{1}{k}\sum_{i=1}^k \delta_{X_k}$ be the empirical measure associated to the k points; how close is it to the uniform measure on $[0, 1]$? What is the expected p-Wasserstein distance between μ^k and the uniform measure on $[0, 1]$?

This is the simplest possible instance of the random matching problem. There are many parameters that can be changed: one could replace the uniform measure with another distribution, or replace the interval $[0, 1]$ with the square $[0, 1]^2$. The problem changes quite drastically when varying this parameter (i.e., moving from a compactly supported measure to a noncompactly supported one or changing the dimension) and the literature is large and scattered through different fields (combinatorics, probability, theoretical computer science). Let us give only two important references: in the book [BL19] the 1-dimensional case is treated in detail, while in the recent paper [AST19] the 2-dimensional case is settled with PDE-techniques. The two mentioned references contain detailed historical reports on the problem.

5.3 Multi-marginal Optimal Transport

The multi-marginal optimal transport problem is a natural generalization of the classical optimal transport problem. It has attracted a lot of research in the latest years, both for the natural applications to economics, but also because it arises naturally in the subfield of quantum mechanics known as Density Functional Theory (see [BDPGG12]).

Given $k \geq 2$ measures $\mu_1, \dots, \mu_k \in \mathcal{P}(X)$ and a cost function $c : \underbrace{X \times \ldots \times X}_{k \text{ times}} \to \mathbb{R}$, we are interested in the minimization problem,

$$\min_{\gamma\in\Gamma(\mu_1,\dots,\mu_k)} \int_{X^k} c(x_1, \dots, x_k)\, d\mu_1(x_1) \cdots d\mu_k(x_k)\,, \tag{5.1}$$

where $\Gamma(\mu_1, \dots, \mu_k)$ is the set of all plans $\gamma \in \mathcal{P}(X^k)$ such that $(\pi_i)_\#(\gamma) = \mu_i$ for all $i = 1, \dots, k$ ($\pi_i : X^k \to X$ denotes the projection on the i-th coordinate). Notice that when $k = 2$ this is the usual Kantorovich problem.

There is also an analogue of the Monge problem that reads as follows. We are interested in finding $k - 1$ maps $T_2, \dots, T_k$ such that $(T_i)_\#(\mu_1) = \mu_i$ and which minimize the transportation cost

$$\int_X c(x, T_2(x), \dots, T_k(x))\, d\mu_1(x)\,.$$

Let us remark that an analogue of Brenier's Theorem (Theorem 2.5.9) holds for the multi-marginal case (as proven in [GS98]).

Even if the multi-marginal problem has many features in common with the Kantorovich problem (such as a duality theory, at least for some choices of the cost), many natural questions (mainly regarding repulsive costs, which are important for the applications) are still open. We suggest the surveys [Pas15; DMGN17] as introductions to this active topic of research.

5.4 Gradient Flows

In this book, we have treated the theory of gradient flows only formally (for example, we have never properly defined what is a gradient-flow in a metric space and we have not shown that,

in a general Hilbert space, the implicit Euler scheme converges to a proper solution). The main reason for this choice is that it would require an immense amount of work.

This remarkable goal is achieved in the book [AGS08] which, from the ground up, studies in depth the properties of gradient flows in metric space, with a particular attention to the Wasserstein space. Let us also mention the pioneering work by Brezis [Bre71], where the author defines and constructs gradient flows in Hilbert spaces. The recent survey [San17b] is an alternative, more compact, introduction to the subject.

The theory of gradient flows in the Wasserstein space has found also numerous applications in the realm of applied mathematics: many real-world phenomena can be modeled with appropriate gradient flows. The interested reader may find many useful references about this topic in the slides [Car14].

5.5 Regularity theory

Whenever $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are absolutely continuous with respect to the Lebesgue measure, the optimal transport map with respect to the quadratic cost is the gradient of a convex function (recall Theorem 2.5.9) and therefore it is differentiable a.e. Nonetheless, it can be discontinuous (for instance, if the support of μ is connected whereas the support of ν is not).

It is therefore natural to wonder whether one can get some regularity under suitable assumptions on the two measures μ, ν. Moreover, as we observed in Section 1.6, a (smooth) transport map must satisfy a Jacobian equation. It turns out that, for the case of the Brenier transport map (i.e., $T = \nabla\phi$ with ϕ convex), it is rather easy to check the validity of such equation (called Monge-Ampère equation) in the a.e. sense, whereas it is hard to understand whether it holds in a suitable "distributional" sense.

The regularity of the Brenier map and the validity of the Monge-Ampère equation are, as can be expected, tightly linked. In fact, one way to obtain the regularity of the transport maps (under suitable assumptions on the measures) is to show that it satisfies the Monge-Ampère equation in a suitable weak sense (i.e., it is an Alexandrov's solution), and then prove that solutions to the Monge-Ampère equation are regular. Since the topic is rich and vast, we point the reader to the book [Fig17] and to the survey paper [DPF14] as possible points of departure for a study of the subject. In particular, [DPF14, Section 4] discusses also the regularity theory of optimal maps on Riemannian manifolds, as well as the connection of this theory to the structure of the cut-locus of the underlying space.

5.6 Computational aspects

There is a vast range of applications of optimal transport to real-world problems, and this has generated a huge interest in the data-science community (i.e. statistics, machine-learning, image-processing, etc..). In a very broad sense, optimal transport provides a quantitative and robust way to decide whether two distributions are close (as a substitute of the more classical Kullback–Leibler divergence).

In the applications of the optimal transport theory, a fundamental issue to overcome is the actual computational cost of finding the optimal transport map. In the discrete setting (as in Exercise A.0.7) the problem, known as the assignment problem or minimum cost matching, has been studied greatly in the computer-science literature and the first efficient algorithm was found by Kuhn in [Kuh55]. On the other hand, the continuous (i.e., finding the transport cost between two absolutely continuous measures) and semi-discrete (i.e., finding the transport cost between a discrete measure and a density) version of the problem present various difficulties that do not arise in the discrete version. We suggest the monograph [PC19] for an in-depth account of these kind of questions.

Let us mention only one key idea that is often useful in the explicit computation of optimal transport maps: the entropic regularization. Given two measures $\mu, \nu \in \mathcal{P}(X)$ and a small

$\varepsilon > 0$, we consider the minimization problem

$$\min_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} c(x,y) + \varepsilon \Big(\log\Big(\frac{d\gamma}{d\mu \otimes \nu}\Big) - 1 \Big)\, d\gamma(x,y)\,,$$

where $c : X \to X \to [0,\infty)$ is the transport cost and $\frac{d\gamma}{d\mu\otimes\nu}$ represents the density of γ with respect to the product measure $\mu \otimes \nu$. This new formulation converges, as $\varepsilon \to 0$, to the classical Kantorovich formulation (both the cost and the plan converge). At the same time, the problem (for positive $\varepsilon > 0$) is strictly convex (and enjoys many other regularity properties, notice for example that the optimal plan is always absolutely continuous with respect to $\mu \otimes \nu$) and therefore an iterative algorithm (the Sinkhorn's algorithm) can be applied to get an approximation of the optimal plan. We suggest [PC19, Chapter 4] and the references therein for further details on entropic regularization.

5.7 From $\mathbb{R}^d$ to Riemannian manifolds and beyond: RCD spaces

The more advanced part of our study, i.e., the differential viewpoint of optimal transport and Otto's calculus, took place entirely in (an open set of) the Euclidean space. What if one considers probability measures on a Riemannian manifold (M, g)?

One can repeat essentially verbatim all the construction of Otto's calculus on $\mathcal{P}_2(M)$. However, when computing the convexity properties of functionals along Wasserstein geodesics, the geometry of the manifold M plays a crucial role.

More precisely, the convexity of the functionals is affected by the Ricci curvature of (M, g), and one can prove the following characterization (see [RS05]):

$$\rho \mapsto \mathcal{F}[\rho] = \int_M \rho(x) \log(\rho(x))\, d\mathrm{vol}(x) \text{ is } W_2\text{-convex}$$

if and only if M has nonnegative Ricci curvature.

In some sense, the geometric properties (i.e., the Ricci curvature) of a space are encoded in the convexity properties of the entropy functional. Since it is possible to define the entropy functional on a general metric measure space (i.e., a metric space endowed with a reference measure), one can say that, by definition:
A metric measure space is positively Ricci curved if the entropy functional is W_2-convex.

This is the starting point of a still very active area of research (begun with the fundamental papers [Stu06a; Stu06b; LV09]) concerning the study of spaces with Ricci curvature bounded from below. We refer the interested reader to the lecture notes [FV11] (see also [Vil09; AG13; Amb18] for an exhaustive discussion of this topic) and to the papers [Gig15; AGS14] where the notion of RCD spaces is introduced (this is a subclass of the spaces considered in [LV09], similarly to how Riemannian manifolds are a subclass of Finsler manifolds).

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.

[AG13] Luigi Ambrosio and Nicola Gigli. "A user's guide to optimal transport". In: *Modelling and optimisation of flows on networks*. Vol. 2062. Lecture Notes in Math. Springer, Heidelberg, 2013, pp. 1–155. DOI: 10.1007/978-3-642-32160-3_1. URL: https://doi.org/10.1007/978-3-642-32160-3_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.

[AGS14] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. "Metric measure spaces with Riemannian Ricci curvature bounded from below". In: *Duke Math. J.* 163.7 (2014), pp. 1405–1490. ISSN: 0012-7094. DOI: 10.1215/00127094-2681605. URL: https://doi.org/10.1215/00127094-2681605.

[Amb08] Luigi Ambrosio. "Transport equation and Cauchy problem for non-smooth vector fields". In: *Calculus of variations and nonlinear partial differential equations*. Vol. 1927. Lecture Notes in Math. Springer, Berlin, 2008, pp. 1–41. DOI: 10.1007/978-3-540-75914-0_1. URL: https://doi.org/10.1007/978-3-540-75914-0_1.

[Amb18] Luigi Ambrosio. "Calculus, heat flow and curvature-dimension bounds in metric measure spaces". In: *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. I. Plenary lectures*. World Sci. Publ., Hackensack, NJ, 2018, pp. 301–340.

[Arn66] V. Arnold. "Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits". In: *Ann. Inst. Fourier (Grenoble)* 16.fasc., fasc. 1 (1966), pp. 319–361. ISSN: 0373-0956. URL: http://www.numdam.org/item?id=AIF_1966__16_1_319_0.

[AST19] Luigi Ambrosio, Federico Stra, and Dario Trevisan. "A PDE approach to a 2-dimensional matching problem". In: *Probab. Theory Related Fields* 173.1-2 (2019), pp. 433–477. ISSN: 0178-8051. DOI: 10.1007/s00440-018-0837-x. URL: https://doi.org/10.1007/s00440-018-0837-x.

[BB00] Jean-David Benamou and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". In: *Numer. Math.* 84.3 (2000), pp. 375–393. ISSN: 0029-599X. DOI: 10.1007/s002110050002. URL: https://doi.org/10.1007/s002110050002.

[BBP16] Hari Bercovici, Arlen Brown, and Carl Pearcy. *Measure and integration*. Springer, Cham, 2016, pp. xi+300. ISBN: 978-3-319-29044-7; 978-3-319-29046-1. DOI: 10.1007/978-3-319-29046-1. URL: https://doi.org/10.1007/978-3-319-29046-1.

[BDPGG12] Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi. "Optimal-transport formulation of electronic density-functional theory". In: *Phys. Rev. A* 85 (6 2012), p. 062502. DOI: 10.1103/PhysRevA.85.062502. URL: https://link.aps.org/doi/10.1103/PhysRevA.85.062502.

[BG03] Yann Brenier and Wilfrid Gangbo. "L^p approximation of maps by diffeomorphisms". In: *Calc. Var. Partial Differential Equations* 16.2 (2003), pp. 147–164. ISSN: 0944-2669. DOI: 10.1007/s005260100144. URL: https://doi.org/10.1007/s005260100144.

[BGM99] Eustasio del Barrio, Evarist Giné, and Carlos Matrán. "Central limit theorems for the Wasserstein distance between the empirical and the true distributions". In: *Ann. Probab.* 27.2 (1999), pp. 1009–1071. ISSN: 0091-1798. DOI: 10.1214/aop/1022677394. URL: https://doi.org/10.1214/aop/1022677394.

[BJ16] Mathias Beiglböck and Nicolas Juillet. "On a problem of optimal transport under marginal martingale constraints". In: *Ann. Probab.* 44.1 (2016), pp. 42–106. ISSN: 0091-1798. DOI: 10.1214/14-AOP966. URL: https://doi.org/10.1214/14-AOP966.

[BL19] Sergey Bobkov and Michel Ledoux. "One-dimensional empirical measures, order statistics, and Kantorovich transport distances". In: *Mem. Amer. Math. Soc.* 261.1259 (2019), pp. v+126. ISSN: 0065-9266. DOI: 10.1090/memo/1259. URL: https://doi.org/10.1090/memo/1259.

[BNT17] Mathias Beiglböck, Marcel Nutz, and Nizar Touzi. "Complete duality for martingale optimal transport on the line". In: *Ann. Probab.* 45.5 (2017), pp. 3038–3074. ISSN: 0091-1798. DOI: 10.1214/16-AOP1131. URL: https://doi.org/10.1214/16-AOP1131.

[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.

[Bre11] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011, pp. xiv+599. ISBN: 978-0-387-70913-0.

[Bre18] Haïm Brezis. "Remarks on the Monge-Kantorovich problem in the discrete setting". In: *C. R. Math. Acad. Sci. Paris* 356.2 (2018), pp. 207–213. ISSN: 1631-073X. DOI: 10.1016/j.crma.2017.12.008. URL: https://doi.org/10.1016/j.crma.2017.12.008.

[Bre71] Haïm Brezis. "Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations". In: *Contributions to nonlinear functional analysis (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1971)*. 1971, pp. 101–156.

[Bre87] Yann Brenier. "Décomposition polaire et réarrangement monotone des champs de vecteurs". In: *C. R. Acad. Sci. Paris Sér. I Math.* 305.19 (1987), pp. 805–808. ISSN: 0249-6291.

[BV05] François Bolley and Cédric Villani. "Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities". In: *Ann. Fac. Sci. Toulouse Math. (6)* 14.3 (2005), pp. 331–352. ISSN: 0240-2963. URL: http://afst.cedram.org/item?id=AFST_2005_6_14_3_331_0.

[Car14] Josè-Antonio Carrillo. *Gradient Flows: Qualitative Properties & Numerical Schemes*. Slides of workshop. 2014. URL: https://www.ricam.oeaw.ac.at/specsem/specsem2014/school2/CharlaRICAM2014-1.pdf.

[CE02] Dario Cordero-Erausquin. “Some applications of mass transport to Gaussian-type inequalities”. In: *Arch. Ration. Mech. Anal.* 161.3 (2002), pp. 257–269. ISSN: 0003-9527. DOI: 10.1007/s002050100185. URL: https://doi.org/10.1007/s002050100185.

[CEMS01] Dario Cordero-Erausquin, Robert J. McCann, and Michael Schmuckenschläger. “A Riemannian interpolation inequality à la Borell, Brascamp and Lieb”. In: *Invent. Math.* 146.2 (2001), pp. 219–257. ISSN: 0020-9910. DOI: 10.1007/s002220100160. URL: https://doi.org/10.1007/s002220100160.

[CENV04] D. Cordero-Erausquin, B. Nazaret, and C. Villani. “A mass-transportation approach to sharp Sobolev and Gagliardo-Nirenberg inequalities”. In: *Adv. Math.* 182.2 (2004), pp. 307–332. ISSN: 0001-8708. DOI: 10.1016/S0001-8708(03)00080-X. URL: https://doi.org/10.1016/S0001-8708(03)00080-X.

[Cha93] Isaac Chavel. *Riemannian geometry—a modern introduction*. Vol. 108. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1993, pp. xii+386. ISBN: 0-521-43201-4; 0-521-48578-9.

[CM17] Fabio Cavalletti and Andrea Mondino. “Sharp and rigid isoperimetric inequalities in metric-measure spaces with lower Ricci curvature bounds”. In: *Invent. Math.* 208.3 (2017), pp. 803–849. ISSN: 0020-9910. DOI: 10.1007/s00222-016-0700-6. URL: https://doi.org/10.1007/s00222-016-0700-6.

[DF13] Sara Daneri and Alessio Figalli. “Variational models for the incompressible Euler equations”. In: *HCDTE lecture notes. Part II. Nonlinear hyperbolic PDEs, dispersive and transport equations*. Vol. 7. AIMS Ser. Appl. Math. Am. Inst. Math. Sci. (AIMS), Springfield, MO, 2013, p. 51.

[DMGN17] Simone Di Marino, Augusto Gerolin, and Luca Nenna. “Optimal transportation theory with repulsive costs”. In: *Topological optimization and optimal transport*. Vol. 17. Radon Ser. Comput. Appl. Math. De Gruyter, Berlin, 2017, pp. 204–256.

[DPF14] Guido De Philippis and Alessio Figalli. “The Monge-Ampère equation and its link to optimal transportation”. In: *Bull. Amer. Math. Soc. (N.S.)* 51.4 (2014), pp. 527–580. ISSN: 0273-0979. DOI: 10.1090/S0273-0979-2014-01459-4. URL: https://doi.org/10.1090/S0273-0979-2014-01459-4.

[FG10] Alessio Figalli and Nicola Gigli. “A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions”. In: *J. Math. Pures Appl. (9)* 94.2 (2010), pp. 107–130. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2009.11.005. URL: https://doi.org/10.1016/j.matpur.2009.11.005.

[FGY11] Alessio Figalli, Wilfrid Gangbo, and Türkay Yolcu. “A variational method for a class of parabolic PDEs”. In: *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* 10.1 (2011), pp. 207–252. ISSN: 0391-173X.

[Fig17] Alessio Figalli. *The Monge-Ampère equation and its applications*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2017, pp. x+200. ISBN: 978-3-03719-170-5. DOI: 10.4171/170. URL: https://doi.org/10.4171/170.

[FMP10] A. Figalli, F. Maggi, and A. Pratelli. “A mass transportation approach to quantitative isoperimetric inequalities”. In: *Invent. Math.* 182.1 (2010), pp. 167–211. ISSN: 0020-9910. DOI: 10.1007/s00222-010-0261-z. URL: https://doi.org/10.1007/s00222-010-0261-z.

[FV11] Alessio Figalli and Cédric Villani. “Optimal transport and curvature”. In: *Nonlinear PDE's and applications*. Vol. 2028. Lecture Notes in Math. Springer, Heidelberg, 2011, pp. 171–217. DOI: 10.1007/978-3-642-21861-3_4. URL: https://doi.org/10.1007/978-3-642-21861-3_4.

[GHL04] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry*. Third. Universitext. Springer-Verlag, Berlin, 2004, pp. xvi+322. ISBN: 3-540-20493-8. DOI: 10.1007/978-3-642-18855-8. URL: https://doi.org/10.1007/978-3-642-18855-8.

[Gig15] Nicola Gigli. “On the differential structure of metric measure spaces and applications”. In: *Mem. Amer. Math. Soc.* 236.1113 (2015), pp. vi+91. ISSN: 0065-9266. DOI: 10.1090/memo/1113. URL: https://doi.org/10.1090/memo/1113.

[GS98] Wilfrid Gangbo and Andrzej Święch. “Optimal maps for the multidimensional Monge-Kantorovich problem”. In: *Comm. Pure Appl. Math.* 51.1 (1998), pp. 23–45. ISSN: 0010-3640. DOI: 10.1002/(SICI)1097-0312(199801)51:1<23::AID-CPA2>3.0.CO;2-H. URL: https://doi.org/10.1002/(SICI)1097-0312(199801)51:1<23::AID-CPA2>3.0.CO;2-H.

[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.

[Kla17] Bo'az Klartag. “Needle decompositions in Riemannian geometry”. In: *Mem. Amer. Math. Soc.* 249.1180 (2017), pp. v+77. ISSN: 0065-9266. DOI: 10.1090/memo/1180. URL: https://doi.org/10.1090/memo/1180.

[Kno57] Herbert Knothe. “Contributions to the theory of convex bodies”. In: *Michigan Math. J.* 4 (1957), pp. 39–52. ISSN: 0026-2285. URL: http://projecteuclid.org/euclid.mmj/1028990175.

[Kuh55] H. W. Kuhn. “The Hungarian method for the assignment problem”. In: *Naval Res. Logist. Quart.* 2 (1955), pp. 83–97. ISSN: 0028-1441. DOI: 10.1002/nav.3800020109. URL: https://doi.org/10.1002/nav.3800020109.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. Vol. 89. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001, pp. x+181. ISBN: 0-8218-2864-9.

[Lee97] John M. Lee. *Riemannian manifolds*. Vol. 176. Graduate Texts in Mathematics. An introduction to curvature. Springer-Verlag, New York, 1997, pp. xvi+224. ISBN: 0-387-98271-X. DOI: 10.1007/b98852. URL: https://doi.org/10.1007/b98852.

[LV09] John Lott and Cédric Villani. “Ricci curvature for metric-measure spaces via optimal transport”. In: *Ann. of Math. (2)* 169.3 (2009), pp. 903–991. ISSN: 0003-486X. DOI: 10.4007/annals.2009.169.903. URL: https://doi.org/10.4007/annals.2009.169.903.

[McC97] Robert J. McCann. “A convexity principle for interacting gases”. In: *Adv. Math.* 128.1 (1997), pp. 153–179. ISSN: 0001-8708. DOI: 10.1006/aima.1997.1634. URL: https://doi.org/10.1006/aima.1997.1634.

[Ott01] Felix Otto. “The geometry of dissipative evolution equations: the porous medium equation”. In: *Comm. Partial Differential Equations* 26.1-2 (2001), pp. 101–174. ISSN: 0360-5302. DOI: 10.1081/PDE-100002243. URL: https://doi.org/10.1081/PDE-100002243.

[Pas15] Brendan Pass. "Multi-marginal optimal transport: theory and applications". In: *ESAIM Math. Model. Numer. Anal.* 49.6 (2015), pp. 1771–1790. ISSN: 0764-583X. DOI: 10.1051/m2an/2015020. URL: https://doi.org/10.1051/m2an/2015020.

[PC19] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11 (5-6) (2019), pp. 355–602. URL: https://arxiv.org/abs/1803.00567.

[Pet06] Peter Petersen. *Riemannian geometry*. Second. Vol. 171. Graduate Texts in Mathematics. Springer, New York, 2006, pp. xvi+401. ISBN: 978-0387-29246-5; 0-387-29246-2.

[RS05] Max-K. von Renesse and Karl-Theodor Sturm. "Transport inequalities, gradient estimates, entropy, and Ricci curvature". In: *Comm. Pure Appl. Math.* 58.7 (2005), pp. 923–940. ISSN: 0010-3640. DOI: 10.1002/cpa.20060. URL: https://doi.org/10.1002/cpa.20060.

[San15] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Vol. 87. Progress in Nonlinear Differential Equations and their Applications. Calculus of variations, PDEs, and modeling. Birkhäuser/Springer, Cham, 2015, pp. xxvii+353. ISBN: 978-3-319-20827-5; 978-3-319-20828-2. DOI: 10.1007/978-3-319-20828-2. URL: https://doi.org/10.1007/978-3-319-20828-2.

[San17a] Filippo Santambrogio. "{Euclidean, metric, and Wasserstein} gradient flows: an overview". In: *Bull. Math. Sci.* 7.1 (2017), pp. 87–154. ISSN: 1664-3607. DOI: 10.1007/s13373-017-0101-1. URL: https://doi.org/10.1007/s13373-017-0101-1.

[San17b] Filippo Santambrogio. "{Euclidean, metric, and Wasserstein} gradient flows: an overview". In: *Bull. Math. Sci.* 7.1 (2017), pp. 87–154. ISSN: 1664-3607. DOI: 10.1007/s13373-017-0101-1. URL: https://doi.org/10.1007/s13373-017-0101-1.

[Stu06a] Karl-Theodor Sturm. "On the geometry of metric measure spaces. I". In: *Acta Math.* 196.1 (2006), pp. 65–131. ISSN: 0001-5962. DOI: 10.1007/s11511-006-0002-8. URL: https://doi.org/10.1007/s11511-006-0002-8.

[Stu06b] Karl-Theodor Sturm. "On the geometry of metric measure spaces. II". In: *Acta Math.* 196.1 (2006), pp. 133–177. ISSN: 0001-5962. DOI: 10.1007/s11511-006-0003-7. URL: https://doi.org/10.1007/s11511-006-0003-7.

[Tes12] Gerald Teschl. *Ordinary differential equations and dynamical systems*. Vol. 140. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2012, pp. xii+356. ISBN: 978-0-8218-8328-0. DOI: 10.1090/gsm/140. URL: https://doi.org/10.1090/gsm/140.

[Vil03] Cédric Villani. *Topics in Optimal Transportation*. Vol. 58. Graduate Studies in Mathematics. Amer. Math. Soc., 2003. Chap. 1.

[Vil09] Cédric Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973. ISBN: 978-3-540-71049-3. DOI: 10.1007/978-3-540-71050-9. URL: https://doi.org/10.1007/978-3-540-71050-9.

[Wil70] Stephen Willard. *General topology*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, 1970, pp. xii+369.