Chapter 1

Introduction

In this introductory chapter we first give a brief historical review of optimal transport, then we recall some basic definitions and facts from measure theory and Riemannian geometry, and finally we present three examples of (not 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 extracts soil from the ground to build fortifications. What is the cheapest possible way to transport the soil? To formulate this question rigorously, 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 $c(x, y) := |x - y|$.

1940s – Kantorovich. After 150 years, Leonid Kantorovich revisited Monge's problem from a different viewpoint. To explain this, consider N bakeries located at positions $(x_i)_{i=1,\ldots,N}$ and M coffee shops located at $(y_j)_{j=1,\ldots,M}$. Assume that the ith bakery produces an amount $\alpha_i \geq 0$ of bread and that the jth coffee shop needs an amount $\beta_j \geq 0$. Also, assume that demand=supply, 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 $(\gamma_{ij})_{\substack{i=1,\ldots,N\\ j=1,\ldots,M}}$ such that

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

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

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

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

It is interesting to observe that constraint (a) is convex, constraints (b) and (c) are linear, and the objective function in (d) 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 $\mathbb{R}^d$: 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.
- $c(x, y) = -\log(|x - y|)$ on the sphere $\mathbb{S}^2 \subset \mathbb{R}^3$: solving the optimal transport problem between two densities on the sphere produces a solution to the associated reflector antenna problem of how to construct an antenna (which is a reflecting surface) in such a way that a light coming from the origin with a given density (in the space of directions, which is parametrized by $\mathbb{S}^2$) is reflected into another given density (again, in the space of directions).

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

1.2 Push-forward of measures

For simplicity, throughout this book we will always work on locally compact, separable, and complete metric spaces, which 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, readers 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 $S: X \to Y$, 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 $\mathscr{P}(X)$, and the class of Borel-measurable sets by $\mathscr{B}(X)$. 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:

- $(T_\#\mu)(\emptyset) = \mu(T^{-1}(\emptyset)) = \mu(\emptyset) = 0$;
- $(T_\#\mu)(Y) = \mu(T^{-1}(Y)) = \mu(X) = 1$;
- $(T_\#\mu)(A) = \mu(T^{-1}(A)) \geq 0$ for all $A \in \mathcal{B}(X)$;
- 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

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

■

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) $\Rightarrow$ $\nu = T_\#\mu$ 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 that

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

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 $\varphi\colon Y \to \mathbb{R}$, 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 $\varphi\colon Y \to \mathbb{R}$. 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

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

which is the desired identity. ■

An immediate consequence of the previous lemma is the following:

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

$$\int_Y \varphi \, d(T_\# \mu) = \int_X \varphi \circ T \, d\mu.$$

The 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 $\varphi\colon Z \to \mathbb{R}$ Borel and bounded we have

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

The result follows from Lemma 1.2.5. ■

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

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

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 readers 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 (not 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, [31,47,56, 63]). Readers with some experience in the subject may skip this chapter.

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

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

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 $F\colon M \to \mathbb{R}$ be a smooth function. Its gradient $\nabla F\colon M \to \mathbb{R}^D$ 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 $\gamma\colon (-1,1) \to M$,

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

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 $\gamma\colon [a,b] \to M$, 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

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

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

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

It turns out that there is always a (not necessarily unique) curve achieving the infimum in the definition of the Riemannian distance (this follows from the compactness of M or, more generally, from its completeness).

Definition 1.3.4 (Minimizing geodesic). A curve $\gamma\colon [a,b] \to M$ 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 (not necessarily minimizing) *geodesic* is a curve $\gamma\colon [a,b] \to M$ 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 $\gamma\colon [a,b] \to M$ 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 with an additional structure, that is, a Riemannian metric. A *Riemannian metric* is a (symmetric and positive definite) scalar product $g_x\colon T_xM \times T_xM \to \mathbb{R}$, 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 $\int_a^b g_\gamma(\dot{\gamma},\dot{\gamma})^{\frac{1}{2}}$), 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 (not 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 $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(Y)$, 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

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

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 $\gamma \in \mathscr{P}(X \times Y)$ a *coupling*[2] of μ and ν if

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

where

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

This is equivalent to requiring that

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

for all $\varphi \colon X \to \mathbb{R}$ Borel and bounded, and

$$\int_{X\times Y} \psi(y)\, d\gamma(x, y) = \int_{X\times Y} \psi \circ \pi_Y(x, y)\, d\gamma(x, y) = \int_Y \psi(y)\, d\nu(y)$$

for all $\psi \colon Y \to \mathbb{R}$ 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 $\phi \colon X \times Y \to \mathbb{R}$) is a coupling:

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

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

$$\gamma_T := (\mathrm{Id} \times T)_\# \mu \in \mathscr{P}(X \times Y).$$

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

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

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

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

1.4.1 Examples of transport maps

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

Measurable transport. The following result can be found in [15, Thm. 11.25]:

Theorem 1.4.6. *Let $\mu \in \mathscr{P}(X)$ 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 $\mu \in \mathscr{P}(X)$ and $\nu \in \mathscr{P}(Y)$ 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 are of little interest in concrete applications in analysis/geometry. Indeed, as we will see in this book, a very important feature of optimal transport maps is their structural properties (for instance, optimal maps for the quadratic cost are gradients of convex functions; see Theorem 2.5.10).

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

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

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 masses of the atoms are included, so that both maps are continuous from the right. More precisely, we set

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

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:

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

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

With these definitions, we define the nondecreasing map $T := G^{-1} \circ F \colon \mathbb{R} \to \mathbb{R}$ 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,

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

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 $\varepsilon > 0$:

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

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

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

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

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

(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 map. We are going to build a transport map, known as the Knothe map [53], 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).

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

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

that is, for any $\varphi : \mathbb{R}^2 \to \mathbb{R}$ 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

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

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

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

as desired.

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

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

Hence, by uniqueness of the disintegration we deduce that

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

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

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

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, monotone rearrangement provides us with a map $T_1\colon \mathbb{R}\to\mathbb{R}$ such that $T_{1\#}(F_1\, dx_1) = G_1\, dy_1$. Then, for $F_1\, dx_1$-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 map $T(x_1, x_2) := (T_1(x_1), T_2(x_1, x_2))$ transports μ to ν.*

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

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

■

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)

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

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

$$\begin{aligned}
\mu(x_1, \dots, x_d) &= f(x_1, \dots, x_d)\, dx_1 \cdots dx_d, \\
\nu(y_1, \dots, y_d) &= g(y_1, \dots, y_d)\, dy_1 \cdots dy_d
\end{aligned}$$

be two absolutely continuous measures. Using monotone rearrangement we get a map $T_1\colon \mathbb{R} \to \mathbb{R}$ such that $T_{1\#}(F_1\, dx_1) = G_1\, dy_1$, 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 $T_{x_1}\colon \mathbb{R}^{d-1} \to \mathbb{R}^{d-1}$ sending μ_{x_1} onto $\nu_{T_1(x_1)}$, and then we obtain a Knothe map in $\mathbb{R}^d$ as

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

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 *nonnegative*. This will be important for the next section.

Remark 1.4.16. Although we call it *the* Knothe 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 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, let $|E|$ denote the Lebesgue measure of E and consider the probability measures $\mu := \frac{1}{|E|}\mathbb{1}_E$ and $\nu := \frac{1}{|B_1|}\mathbb{1}_{B_1}$. Notice that here, as often in the book, we identify a measure with its density with respect to Lebesgue.

Proposition 1.5.2. *Let T be a Knothe map from μ to ν, and assume it to be smooth.*[3] *Then,*

(a) *for any $x \in E$, it holds that $|T(x)| \leq 1$;*

(b) $\det \nabla T = \frac{|B_1|}{|E|}$ *in E;*

(c) $\operatorname{div} T \geq d\,(\det \nabla T)^{\frac{1}{d}}$.

[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 map) and use the theory of functions with bounded variation, as done in [44].

Proof. We prove the three properties.

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

(b) 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 $\det \nabla T \geq 0$, 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

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

(c) 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

$$\begin{aligned} \operatorname{div} T(x) = \sum_{i=1}^{d} \partial_i T_i(x) &= d\left(\frac{1}{d} \sum_{i=1}^{d} \partial_i T_i(x)\right) \\ &\geq d\left(\prod_{i=1}^{d} \partial_i T_i(x)\right)^{\frac{1}{d}} = d\big(\det \nabla T(x)\big)^{\frac{1}{d}}, \end{aligned}$$

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 properties (a), (b), (c) 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

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

where the equality marked with † follows from the Stokes theorem. ■

1.6 A Jacobian equation for transport maps

Let $T\colon \mathbb{R}^d \to \mathbb{R}^d$ 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 the push-forward measure, for any bounded Borel function $\zeta\colon \mathbb{R}^d \to \mathbb{R}$ 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 $dy=\det\nabla T(x)\,dx$, and 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 chapters (and also the Knothe 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.