# Chapter 1
# Optimal Transport


In this preliminary chapter, we introduce the problem of optimal transport, which is the main concept behind Wasserstein spaces. General references on this topic are the books by Rachev and Rüschendorf [107], Villani [124, 125], Ambrosio et al. [12], Ambrosio and Gigli [10], and Santambrogio [119]. This chapter includes only few proofs, when they are simple, informative, or are not easily found in one of the cited references.

## 1.1 The Monge and the Kantorovich Problems

In 1781, Monge [95] asked the following question: given a pile of sand and a pit of equal volume, how can one optimally transport the sand into the pit? In modern mathematical terms, the problem can be formulated as follows. There is a sand space $\mathscr{X}$, a pit space $\mathscr{Y}$, and a cost function $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ that encapsulates how costly it is to move a unit of sand at $x \in \mathscr{X}$ to a location $y \in \mathscr{Y}$ in the pit. The sand distribution is represented by a measure $\mu$ on $\mathscr{X}$, and the shape of the pit is described by a measure $\nu$ on $\mathscr{Y}$. Our decision where to put each unit of sand can be thought of as a function $T : \mathscr{X} \to \mathscr{Y}$, and it incurs a total transport cost of

$$C(T) = \int_{\mathscr{X}} c(x, T(x)) \, \mathrm{d}\mu(x).$$

Moreover, one cannot put all the sand at a single point $y$ in the pit; it is not allowed to shrink or expand the sand. The map $T$ must be mass-preserving: for any subset $B \subseteq$

**Electronic Supplementary Material** The online version of this chapter (https://doi.org/10.1007/978-3-030-38438-8_1) contains supplementary material.



V. M. Panaretos, Y. Zemel, *An Invitation to Statistics in Wasserstein Space*,
SpringerBriefs in Probability and Mathematical Statistics,
https://doi.org/10.1007/978-3-030-38438-8_1

$\mathscr{Y}$ representing a region of the pit of volume $v(B)$, exactly that same volume of sand must go into $B$. The amount of sand allocated to $B$ is $\{x \in \mathscr{X} : T(x) \in B\} = T^{-1}(B)$, so the mass preservation requirement is that $\mu(T^{-1}(B)) = v(B)$ for all $B \subseteq \mathscr{Y}$. This condition will be denoted by $T\#\mu = v$ and in words: $v$ is the push-forward of $\mu$ under $T$, or $T$ pushes $\mu$ forward to $v$. To make the discussion mathematically rigorous, we must assume that $c$ and $T$ are measurable maps, and that $\mu(T^{-1}(B)) = v(B)$ for all measurable subsets of $\mathscr{Y}$. When the underlying measures are understood from the context, we call $T$ a *transport map*. Specifying $B = \mathscr{Y}$, we see that no such $T$ can exist unless $\mu(\mathscr{X}) = v(\mathscr{Y})$; we shall assume that this quantity is finite, and by means of normalisation, that $\mu$ and $v$ are probability measures. In this setting, the Monge problem is to find the optimal transport map, that is, to solve

$$\inf_{T:T\#\mu=v} C(T).$$

We assume throughout this book that $\mathscr{X}$ and $\mathscr{Y}$ are complete and separable metric spaces,[1] endowed with their *Borel $\sigma$-algebra*, which, we recall, is defined as the smallest $\sigma$-algebra containing the open sets. Measures defined on the Borel $\sigma$-algebra of $\mathscr{X}$ are called *Borel measures*. Thus, if $\mu$ is a Borel measure on $\mathscr{X}$, then $\mu(A)$ is defined for any $A$ that is open, or closed, or a countable union of closed sets, etc., and any continuous map on $\mathscr{X}$ is measurable. Similarly, we endow $\mathscr{Y}$ with its Borel $\sigma$-algebra. The product space $\mathscr{X} \times \mathscr{Y}$ is also complete and separable when endowed with its product topology; its Borel $\sigma$-algebra is generated by the product $\sigma$-algebra of those of $\mathscr{X}$ and $\mathscr{Y}$; thus, any continuous cost function $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ is measurable. It will henceforth always be assumed, without explicit further notice, that $\mu$ and $v$ are Borel measures on $\mathscr{X}$ and $\mathscr{Y}$, respectively, and that the cost function is continuous and nonnegative.

It is quite natural to assume that the cost is an increasing function of the distance between $x$ and $y$, such as a power function. More precisely, that $\mathscr{Y} = \mathscr{X}$ is a complete and separable metric space with metric $d$, and

$$c(x,y) = d^p(x,y), \qquad p \geq 0, \quad x,y \in \mathscr{X}. \tag{1.1}$$

In particular, $c$ is continuous, hence measurable, if $p > 0$. The limit case $p = 0$ yields the discontinuous function $c(x,y) = \mathbf{1}\{x = y\}$, which nevertheless remains measurable because the diagonal $\{(x,x) : x \in \mathscr{X}\}$ is measurable in $\mathscr{X} \times \mathscr{X}$. Particular focus will be put on the quadratic case $p = 2$ (Sect. 1.6) and the linear case $p = 1$ (Sect. 1.8.2).

The problem introduced by Monge [95] is very difficult, mainly because the set of transport maps $\{T : T\#\mu = v\}$ is intractable. And, it may very well be empty: this will be the case if $\mu$ is a Dirac measure at some $x_0 \in \mathscr{X}$ (meaning that $\mu(A) = 1$ if $x_0 \in A$ and 0 otherwise) but $v$ is not. Indeed, in that case the set $B = \{T(x_0)\}$ satisfies $\mu(T^{-1}(B)) = 1 > v(B)$, so no such $T$ can exist. This also shows that the problem is asymmetric in $\mu$ and $v$: in the Dirac example, there always exists a map $T$ such that $T\#v = \mu$—the constant map $T(x) = x_0$ for all $x$ is the unique such map. A less

[1] But see the bibliographical notes for some literature on more general spaces.

extreme situation occurs in the case of absolutely continuous measures. If $\mu$ and $\nu$ have densities $f$ and $g$ on $\mathbb{R}^d$ and $T$ is continuously differentiable, then $T\#\mu = \nu$ if and only if for $\mu$-almost all $x$

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

This is a highly non-linear equation in $T$, nowadays known as a particular case of a family of partial differential equations called *Monge–Ampère equations*. More than two centuries after the work of Monge, Caffarelli [32] cleverly used the theory of Monge–Ampère equations to show smoothness of transport maps (see Sect. 1.6.4).

As mentioned above, if $\mu = \delta\{x_0\}$ is a Dirac measure and $\nu$ is not, then no transport maps from $\mu$ to $\nu$ can exist, because the mass at $x_0$ must be sent to a unique point $x_0$. In 1942, Kantorovich [77] proposed a relaxation of Monge's problem in which mass can be split. In other words, for each point $x \in \mathscr{X}$ one constructs a probability measure $\mu_x$ that describes how the mass at $x$ is split among different destinations. If $\mu_x$ is a Dirac measure at some $y$, then all the mass at $x$ is sent to $y$. The formal mathematical object to represent this idea is a probability measure $\pi$ on the product space $\mathscr{X} \times \mathscr{Y}$ (which is $\mathscr{X}^2$ in our particular setting). Here $\pi(A \times B)$ is the amount of sand transported from the subset $A \subseteq \mathscr{X}$ into the part of the pit represented by $B \subseteq \mathscr{Y}$. The total mass sent from $A$ is $\pi(A \times \mathscr{Y})$, and the total mass sent into $B$ is $\pi(\mathscr{X} \times B)$. Thus, $\pi$ is mass-preserving if and only if

$$\begin{aligned} \pi(A \times \mathscr{Y}) &= \mu(A), \qquad A \subseteq \mathscr{X} \quad \text{Borel;} \\ \pi(\mathscr{X} \times B) &= \nu(B), \qquad B \subseteq \mathscr{Y} \quad \text{Borel.} \end{aligned} \tag{1.2}$$

Probability measures satisfying (1.2) will be called *transference plans*, and the set of those will be denoted by $\Pi(\mu, \nu)$. We also say that $\pi$ is a *coupling* of $\mu$ and $\nu$, and that $\mu$ and $\nu$ are the first and second *marginal distributions*, or simply *marginals*, of $\pi$. The total cost associated with $\pi \in \Pi(\mu, \nu)$ is

$$C(\pi) = \int_{\mathscr{X} \times \mathscr{Y}} c(x,y)\, \mathrm{d}\pi(x,y).$$

In our setting of a complete separable metric space $\mathscr{X}$, one can represent $\pi$ as a collection of probability measures $\{\pi_x\}_{x \in \mathscr{X}}$ on $\mathscr{Y}$, in the sense that for all measurable nonnegative $g$

$$\int_{\mathscr{X} \times \mathscr{Y}} g(x,y)\, \mathrm{d}\pi(x,y) = \int_{\mathscr{X}} \left[ \int_{\mathscr{Y}} g(x,y)\, \mathrm{d}\pi_x(y) \right] \mathrm{d}\mu(x).$$

The collection $\{\pi_x\}$ is that of the *conditional distributions*, and the iteration of integrals is called *disintegration*. For proofs of existence of conditional distributions, one can consult Dudley [47, Section 10.2] or Kallenberg [76, Chapter 5]. Conversely, the measure $\mu$ and the collection $\{\pi_x\}$ determine $\pi$ uniquely by choosing $g$ to be indicator functions. An interpretation of these notions in terms of random variables will be given in Sect. 1.2.

The Kantorovich problem is to find the best transference plan, that is, to solve

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi).$$

The Kantorovich problem is a relaxation of the Monge problem, because to each transport map $T$ one can associate a transference plan $\pi = \pi_T$ of the same total cost. To see this, choose the conditional distribution $\pi_x$ to be a Dirac at $T(x)$. Disintegration then yields

$$C(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,\mathrm{d}\pi(x,y) = \int_{\mathscr{X}} \left[\int_{\mathscr{Y}} c(x,y)\,\mathrm{d}\pi_x(y)\right] \mathrm{d}\mu(x) = \int_{\mathscr{X}} c(x,T(x))\,\mathrm{d}\mu(x) = C(T).$$

This choice of $\pi$ satisfies (1.2) because $\pi(A\times B) = \mu(A\cap T^{-1}(B))$ and $\nu(B) = \mu(T^{-1}(B))$ for all Borel $A\subseteq\mathscr{X}$ and $B\subseteq\mathscr{Y}$.

Compared to the Monge problem, the relaxed problem has considerable advantages. Firstly, the set of transference plans is never empty: it always contains the product measure $\mu\otimes\nu$ defined by $[\mu\otimes\nu](A) = \mu(A)\nu(B)$. Secondly, both the objective function $C(\pi)$ and the constraints (1.2) are linear in $\pi$, so the problem can be seen as infinite-dimensional linear programming. To be precise, we need to endow the space of measures with a linear structure, and this is done in the standard way: define the space $M(\mathscr{X})$ of all finite signed Borel measures on $\mathscr{X}$. This is a vector space with $(\mu_1+\alpha\mu_2)(A) = \mu_1(A)+\alpha\mu_2(A)$ for $\alpha\in\mathbb{R}$, $\mu_1,\mu_2\in M(\mathscr{X})$ and $A\subseteq\mathscr{X}$ Borel. The set of probability measures on $\mathscr{X}$ is denoted by $P(\mathscr{X})$, and is a convex subset of $M(\mathscr{X})$. The set $\Pi(\mu,\nu)$ is then a convex subset of $P(\mathscr{X}\times Y)$, and as $C(\pi)$ is linear in $\pi$, the set of minimisers is a convex subset of $\Pi(\mu,\nu)$. Thirdly, there is a natural symmetry between $\Pi(\mu,\nu)$ and $\Pi(\nu,\mu)$. If $\pi$ belongs to the former and we define $\tilde{\pi}(B\times A) = \pi(A\times B)$, then $\tilde{\pi}\in\Pi(\nu,\mu)$. If we set $\tilde{c}(y,x) = c(x,y)$, then

$$C(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\,\mathrm{d}\pi(x,y) = \int_{\mathscr{Y}\times\mathscr{X}} \tilde{c}(y,x)\,\mathrm{d}\tilde{\pi}(y,x) = \tilde{C}(\tilde{\pi}).$$

In particular, when $\mathscr{X} = \mathscr{Y}$ and $c = \tilde{c}$ is symmetric (as in (1.1)),

$$\inf_{\pi\in\Pi(\mu,\nu)} C(\pi) = \inf_{\tilde{\pi}\in\Pi(\nu,\mu)} \tilde{C}(\tilde{\pi}),$$

and $\pi\in\Pi(\mu,\nu)$ is optimal if and only if its natural counterpart $\tilde{\pi}$ is optimal in $\Pi(\nu,\mu)$. This symmetry will be fundamental in the definition of the Wasserstein distances in Chap. 2.

Perhaps most importantly, a minimiser for the Kantorovich problem exists under weak conditions. In order to show this, we first recall some definitions. Let $C_b(\mathscr{X})$ be the space of real-valued, continuous bounded functions on $\mathscr{X}$. A sequence of probability measures $\{\mu_n\}\in M(\mathscr{X})$ is said to converge *weakly*[2] to $\mu\in M(\mathscr{X})$ if for all $f\in C_b(\mathscr{X})$, $\int f\,\mathrm{d}\mu_n \to \int f\,\mathrm{d}\mu$. To avoid confusion with other types of convergence, we will usually write $\mu_n\to\mu$ weakly; in the rare cases where a symbol

[2] Weak convergence is sometimes called narrow convergence, weak* convergence, or convergence in distribution.

is needed we shall use the notation $\mu_n \xrightarrow{w} \mu$. Of course, if $\mu_n \to \mu$ weakly and $\mu_n \in P(\mathscr{X})$, then $\mu$ must be in $P(\mathscr{X})$ too (this is seen by taking $f \equiv 1$ and by observing that $\int f \, d\mu \geq 0$ if $f \geq 0$).

A collection of probability measures $\mathscr{K}$ is *tight* if for all $\varepsilon > 0$ there exists a compact set $K$ such that $\inf_{\mu \in \mathscr{K}} \mu(K) > 1 - \varepsilon$. If $\mathscr{K}$ is represented by a sequence $\{\mu_n\}$, then Prokhorov's theorem (Billingsley [24, Theorem 5.1]) states that a subsequence of $\{\mu_n\}$ must converge weakly to some probability measure $\mu$.

We are now ready to show that the Kantorovich problem admits a solution when $c$ is continuous and nonnegative and $\mathscr{X}$ and $\mathscr{Y}$ are complete separable metric spaces. Let $\{\pi_n\}$ be a minimising sequence for $C$. Then, according to [24, Theorem 1.3], $\mu$ and $\nu$ must be tight. If $K_1$ and $K_2$ are compact with $\mu(K_1), \nu(K_2) > 1 - \varepsilon$, then $K_1 \times K_2$ is compact and for all $\pi \in \Pi(\mu, \nu)$, $\pi(K_1 \times K_2) > 1 - 2\varepsilon$. It follows that the entire collection $\Pi(\mu, \nu)$ is tight, and by Prokhorov's theorem $\pi_n$ has a weak limit $\pi$ after extraction of a subsequence. For any integer $K$, $c_K(x,y) = \min(c(x,y), K)$ is a continuous bounded function, and

$$C(\pi_n) = \int c(x,y) \, d\pi_n(x,y) \geq \int c_K(x,y) \, d\pi_n(x,y) \to \int c_K(x,y) \, d\pi(x,y), \qquad n \to \infty.$$

By the monotone convergence theorem

$$\liminf_{n \to \infty} C(\pi_n) \geq \lim_{K \to \infty} \int c_K(x,y) \, d\pi(x,y) = C(\pi) \qquad \text{if } \pi_n \to \pi \text{ weakly.} \tag{1.3}$$

Since $\{\pi_n\}$ was chosen as a minimising sequence for $C$, $\pi$ must be a minimiser, and existence is established.

As we have seen, the Kantorovich problem is a relaxation of the Monge problem, in the sense that

$$\inf_{T : T\#\mu = \nu} C(T) = \inf_{\pi_T : T\#\mu = \nu} C(\pi) \geq \inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = C(\pi^*),$$

for some optimal $\pi^*$. If $\pi^* = \pi_T$ for some transport map $T$, then we say that the solution is induced from a transport map. This will happen in two different and important cases that are discussed in Sects. 1.3 and 1.6.1.

A remark about terminology is in order. Many authors talk about the *Monge–Kantorovich problem* or the *optimal transport(ation) problem*. More often than not, they refer to what we call here the Kantorovich problem. When one of the scenarios presented in Sects. 1.3 and 1.6.1 is considered, this does not result in ambiguity.

## 1.2 Probabilistic Interpretation

The preceding section was an analytic presentation of the Monge and the Kantorovich problems. It is illuminating, however, to also recast things in probabilistic terms, and this is the topic of this section.

A *random element* on a complete separable metric space (or any topological space) $\mathscr{X}$ is simply a measurable function $X$ from some (generic) probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $\mathscr{X}$ (with its Borel $\sigma$-algebra). The *probability law* (or *probability distribution*) is the probability measure $\mu_X = X\#\mathbb{P}$ defined on the space $\mathscr{X}$; this is the Borel measure satisfying $\mu_X(A) = \mathbb{P}(X \in A)$ for all Borel sets $A$.

Suppose that one is given two random elements $X$ and $Y$ taking values in $\mathscr{X}$ and $\mathscr{Y}$, respectively, and a cost function $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$. The Monge problem is to find a measurable function $T$ such that $T(X)$ has the same distribution as $Y$, and such that the expectation

$$C(T) = \int_{\mathscr{X}} c(x, T(x))\, \mathrm{d}\mu(x) = \int_{\Omega} c[X(\omega), T(X(\omega))]\, \mathrm{d}\mathbb{P}(\omega) = \mathbb{E}c(X, T(X))$$

is minimised.

The Kantorovich problem is to find a joint distribution for the pair $(X,Y)$ whose marginals are the original distributions of $X$ and $Y$, respectively, and such that the probability law $\pi = (X,Y)\#\mathbb{P}$ minimises the expectation

$$C(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\, \mathrm{d}\pi(x,y) = \int_{\Omega} c[X(\omega), Y(\omega))]\, \mathrm{d}\mathbb{P}(\omega) = \mathbb{E}_{\pi} c(X,Y).$$

Any such joint distribution is called a coupling of $X$ and $Y$. Of course, $(X, T(X))$ is a coupling when $T(X)$ has the same distribution as $Y$. The measures $\pi_x$ in the previous section are then interpreted as the conditional distribution of $Y$ given $X = x$.

Consider now the important case where $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$, $c(x,y) = \|x-y\|^2$, and $X$ and $Y$ are square integrable random vectors ($\mathbb{E}\|X\|^2 + \mathbb{E}\|Y\|^2 < \infty$). Let $A$ and $B$ be the covariance matrices of $X$ and $Y$, respectively, and notice that the covariance matrix of a coupling $\pi$ must have the form $C = \begin{pmatrix} A & V \\ V^t & B \end{pmatrix}$ for a $d \times d$ matrix $V$. The covariance matrix of the difference $X - Y$ is

$$\begin{pmatrix} I_d & -I_d \end{pmatrix} \begin{pmatrix} A & V \\ V^t & B \end{pmatrix} \begin{pmatrix} I_d \\ -I_d \end{pmatrix} = A + B - V^t - V$$

so that

$$\mathbb{E}_{\pi} c(X,Y) = \mathbb{E}_{\pi}\|X-Y\|^2 = \|\mathbb{E}X - \mathbb{E}Y\|^2 + \mathrm{tr}_{\pi}[A + B - V^t - V].$$

Since only $V$ depends on the coupling $\pi$, the problem is equivalent to that of maximising the trace of $V$, the cross-covariance matrix between $X$ and $Y$. This must be done subject to the constraint that a coupling $\pi$ with covariance matrix $C$ exists; in particular, $C$ has to be positive semidefinite.

## 1.3 The Discrete Uniform Case

There is a special case in which the Monge–Kantorovich problem reduces to a finite combinatorial problem. Although it may seem at first hand as an oversimplification of the original problem, it is of importance in practice because arbitrary measures can be approximated by discrete measures by means of the strong law of large numbers. Moreover, the discrete case is important in theory as well, as a motivating example for the Kantorovich duality (Sect. 1.4) and the property of cyclical monotonicity (Sect. 1.7).

Suppose that $\mu$ and $\nu$ are each uniform on $n$ distinct points:

$$\mu = \frac{1}{n}\left(\delta\{x_1\} + \cdots + \delta\{x_n\}\right), \qquad \nu = \frac{1}{n}\left(\delta\{y_1\} + \cdots + \delta\{y_n\}\right).$$

The only relevant costs are $c_{ij} = c(x_i, y_j)$, the collection of which can be represented by an $n \times n$ matrix $\mathbf{C}$. Transport maps $T$ are associated with *permutations* in $S_n$, the set of all bijective functions from $\{1, \ldots, n\}$ to itself: given $\sigma \in S_n$, a transport map can be constructed by defining $T(x_i) = y_{\sigma(i)}$. If $\sigma$ is not a permutation, then $T$ will not be a transport map from $\mu$ to $\nu$. Transference plans $\pi$ are equivalent to $n \times n$ matrices $M$ with coordinates $M_{ij} = \pi(\{(x_i, y_j)\}) = M_{ij}$; this is the amount of mass sent from $x_i$ to $y_j$. In order for $\pi$ to a be a transference plan, it must be that $\sum_j M_{ij} = 1/n$ for all $i$ and $\sum_i M_{ij} = 1/n$ for all $j$, and in addition $M$ must be nonnegative. In other words, the matrix $M' = nM$ belongs to $B_n$, the set of bistochastic matrices of order $n$, defined as $n \times n$ matrices $M'$ satisfying

$$\sum_{j=1}^{n} M'_{ij} = 1, \quad i = 1, \ldots, n; \qquad \sum_{i=1}^{n} M'_{ij} = 1, \quad j = 1, \ldots, n; \qquad M'_{ij} \geq 0.$$

The Monge problem is the combinatorial optimisation problem over permutations

$$\inf_{\sigma \in S_n} C(\sigma) = \frac{1}{n} \inf_{\sigma \in S_n} \sum_{i=1}^{n} c_{i,\sigma(i)},$$

and the Kantorovich problem is the linear program

$$\inf_{nM \in B_n} \sum_{i,j=1}^{n} c_{ij} M_{ij} = \inf_{M \in B_n/n} \sum_{i,j=1}^{n} c_{ij} M_{ij} = \inf_{M \in B_n/n} C(M).$$

If $\sigma$ is a permutation, then one can define $M = M(\sigma)$ by $M_{ij} = 1/n$ if $j = \sigma(i)$ and 0 otherwise. Then $M \in B_n/n$ and $C(M) = C(\sigma)$. Such $M$ (or, more precisely, $nM$) is called a *permutation matrix*.

The Kantorovich problem is a linear program with $n^2$ variables and $2n$ constraints. It must have a solution because $B_n$ (hence $B_n/n$) is a compact (nonempty) set in $\mathbb{R}^{n^2}$ and the objective function is linear in the matrix elements, hence continuous. (This property is independent of the possibly infinite-dimensional spaces $\mathscr{X}$

and $\mathscr{Y}$ in which the points lie.) The Monge problem also admits a solution because $S_n$ is a finite set. To see that the two problems are essentially the same, we need to introduce the following notion. If $B$ is a convex set, then $x \in B$ is an *extremal point* of $B$ if it cannot be written as a convex combination $tz + (1-t)y$ for some distinct points $y, z \in B$. It is well known (Luenberger and Ye [89, Section 2.5]) that there exists an optimal solution that is extremal, so that it becomes relevant to identify the extremal points of $B_n$. It is fairly clear that each permutation matrix is extremal in $B_n$; the less obvious converse is known as Birkhoff's theorem, a proof of which can be found, for instance, at the end of the introduction in Villani [124] or (in a different terminology) in Luenberger and Ye [89, Section 6.5]. Thus, we have:

**Proposition 1.3.1 (Solution of Discrete Problem)** *There exists $\sigma \in S_n$ such that $M(\sigma)$ minimises $C(M)$ over $B_n/n$. Furthermore, if $\{\sigma_1, \ldots, \sigma_k\}$ is the set of optimal permutations, then the set of optimal matrices is the convex hull of $\{M(\sigma_1), \ldots, M(\sigma_k)\}$. In particular, if $\sigma$ is the unique optimal permutation, then $M(\sigma)$ is the unique optimal matrix.*

Thus, in the discrete case, the Monge and the Kantorovich problems coincide. One can of course use the simplex method [89, Chapter 3] to solve the linear program, but there are $n!$ vertices, and there is in principle no guarantee that the simplex method solves the problem efficiently. However, the constraints matrix has a very specific form (it contains only zeroes and ones, and is totally unimodular), so specialised algorithms for this problem exist. One of them is the Hungarian algorithm of Kuhn [85] or its variant of Munkres [96] that has a worst-case computational complexity of at most $O(n^4)$. Another alternative is the class of net flow algorithms described in [89, Chapter 6]. In particular, the algorithm of Edmonds and Karp [50] has a complexity of at most $O(n^3 \log n)$. This monograph does not focus on computational aspects for optimal transport. This is a fascinating and very active area of contemporary research, and readers are directed to Peyré and Cuturi [103].

**Remark 1.3.2** *The special case described here could have been more precisely called "the discrete uniform case on the same number of points", as "the discrete case" could refer to any two finitely supported measures $\mu$ and $\nu$. In the Monge context, the setup discussed here is the most interesting case, see page 8 in the supplement for more details.*

## 1.4 Kantorovich Duality

The discrete case of Sect. 1.3 is an example of a linear program and thus enjoys a rich duality theory (Luenberger and Ye [89, Chapter 4]). The general Kantorovich problem is an infinite-dimensional linear program, and under mild assumptions admits similar duality.

### 1.4.1 Duality in the Discrete Uniform Case

We can represent any matrix $M$ as a vector in $\mathbb{R}^{n^2}$, say $\mathbf{M}$, by enumeration of the elements row by row. If $nM$ is bistochastic, i.e., $M \in B_n/n$, then the $2n$ constraints can be represented in a $(2n) \times n^2$ matrix $A$. For instance, if $n = 3$, then

$$A = \begin{pmatrix} 1 & 1 & 1 & & & & & & \\ & & & 1 & 1 & 1 & & & \\ & & & & & & 1 & 1 & 1 \\ 1 & & & 1 & & & 1 & & \\ & 1 & & & 1 & & & 1 & \\ & & 1 & & & 1 & & & 1 \end{pmatrix} \in \mathbb{R}^{6\times 9}.$$

For general $n$, the constraints read $A\mathbf{M} = n^{-1}(1,\dots,1) \in \mathbb{R}^{2n}$ and $A$ takes the form

$$A = \begin{pmatrix} \mathbf{1}_n & & & \\ & \mathbf{1}_n & & \\ & & \ddots & \\ & & & \mathbf{1}_n \\ I_n & I_n & \dots & I_n \end{pmatrix} \in \mathbb{R}^{2n\times n^2}, \qquad \mathbf{1}_n = (1,\dots,1) \in \mathbb{R}^n,$$

with $I_n$ the $n \times n$ identity matrix. Thus, the problem can be written

$$\min_{\mathbf{M}} \mathbf{C}^t\mathbf{M} \qquad \text{subject to} \qquad A\mathbf{M} = \frac{1}{n}(1,\dots,1) \in \mathbb{R}^{2n}; \quad \mathbf{M} \geq 0.$$

The last constraint is to be interpreted coordinate-wise; all the elements of $M$ must be nonnegative. The *dual problem* is constructed by introducing one variable for each row of $A$, transposing the constraint matrix and interchanging the roles of the objective vector $\mathbf{C}$ and the constraints vector $b = n^{-1}(1,\dots,1)$. Call the new variables $p_1,\dots,p_n$ and $q_1,\dots,q_n$, and notice that each column of $A$ corresponds to exactly one $p_i$ and one $q_j$, and that the $n^2$ columns exhaust all possibilities. Hence, the dual problem is

$$\max_{p,q\in\mathbb{R}^n} b^t \begin{pmatrix} p \\ q \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n p_i + \frac{1}{n}\sum_{j=1}^n q_j \qquad \text{subject to} \quad p_i + q_j \leq c_{ij}, \quad i,j = 1,\dots,n. \tag{1.4}$$

In the context of duality, one uses the terminology *primal problem* for the original optimisation problem. *Weak duality* states that if $\mathbf{M}$ and $(p,q)$ satisfy the respective constraints, then

$$b^t \begin{pmatrix} p \\ q \end{pmatrix} = \sum_i p_i \frac{1}{n} + \sum_j q_j \frac{1}{n} = \sum_{i,j}(p_i + q_j)M_{ij} \leq \sum_{i,j} C_{ij}M_{ij} = \mathbf{C}^t\mathbf{M}.$$

In particular, if equality holds, then $\mathbf{M}$ is primal optimal and $(p,q)$ is dual optimal. *Strong duality* is the nontrivial assertion that there exist $\mathbf{M}^*$ and $(p^*,q^*)$ satisfying $\mathbf{C}^t\mathbf{M}^* = b^t\binom{p^*}{q^*}$.

### 1.4.2 Duality in the General Case

The vectors $\mathbf{C}$ and $\mathbf{M}$ were obtained from the cost function $c$ and the transference plan $\pi$ as $C_{ij} = c(x_i,y_j)$ and $M_{ij} = \pi(\{(x_i,y_j)\})$. Similarly, we can view the vectors $p$ and $q$ as restrictions of functions $\varphi : \mathscr{X} \to \mathbb{R}$ and $\psi : \mathscr{Y} \to \mathbb{R}$ of the form $p_i = \varphi(x_i)$ and $q_j = \psi(y_j)$. The constraint vector $b = (\mathbf{1}_n, \mathbf{1}_n)$ can be written as $b_i = \mu(\{x_i\})$ and $b_{n+j} = \nu(\{y_j\})$. In this formulation, the constraint $p_i + q_j \le c_{ij}$ writes $(\varphi,\psi) \in \Phi_c$ with

$$\Phi_c = \big\{(\varphi,\psi) \in L_1(\mu) \times L_1(\nu) : \varphi(x) + \psi(y) \le c(x,y) \text{ for all } x,y\big\},$$

and the dual problem (1.4) becomes

$$\sup_{(\varphi,\psi)\in L_1(\mu)\times L_1(\nu)} \left[\int_{\mathscr{X}} \varphi(x)\,\mathrm{d}\mu(x) + \int_{\mathscr{Y}} \psi(y)\,\mathrm{d}\nu(y)\right] \qquad \text{subject to} \quad (\varphi,\psi) \in \Phi_c.$$

Simple measure theory shows that the set constraints (1.2) defining the transference plans set $\Pi(\mu,\nu)$ are equivalent to functional constraints. For future reference, we state this formally as:

**Lemma 1.4.1 (Functional Constraints)** *Let $\mu$ and $\nu$ be probability measures. Then $\pi \in \Pi(\mu,\nu)$ if and only if for all integrable functions $\varphi \in L_1(\mu)$, $\psi \in L_1(\nu)$,*

$$\int_{\mathscr{X}\times\mathscr{Y}} [\varphi(x) + \psi(y)]\,\mathrm{d}\pi(x,y) = \int_{\mathscr{X}} \varphi(x)\,\mathrm{d}\mu(x) + \int_{\mathscr{Y}} \psi(y)\,\mathrm{d}\nu(y).$$

The proof follows from the fact that (1.2) yields the above equality when $\varphi$ and $\psi$ are indicator functions. One then uses linearity and approximations to deduce the result.

Weak duality follows immediately from Lemma 1.4.1. For if $\pi \in \Pi(\mu,\nu)$ and $(\varphi,\psi) \in \Phi_c$, then

$$\int_{\mathscr{X}} \varphi(x)\,\mathrm{d}\mu(x) + \int_{\mathscr{Y}} \psi(y)\,\mathrm{d}\nu(y) = \int_{\mathscr{X}\times\mathscr{Y}} [\varphi(x) + \psi(y)]\,\mathrm{d}\pi(x,y) \le C(\pi).$$

Strong duality can be stated in the following form:

**Theorem 1.4.2 (Kantorovich Duality)** *Let $\mu$ and $\nu$ be probability measures on complete separable metric spaces $\mathscr{X}$ and $\mathscr{Y}$, respectively, and let $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}_+$ be a measurable function. Then*

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathscr{X}\times\mathscr{Y}} c\,\mathrm{d}\pi = \sup_{(\varphi,\psi)\in\Phi_c}\left[\int_{\mathscr{X}}\varphi\,\mathrm{d}\mu + \int_{\mathscr{Y}}\psi\,\mathrm{d}\nu\right].$$

See the Bibliographical Notes for other versions of the duality.

When the cost function is continuous, or more generally, a countable supremum of continuous functions, the infimum is attained (see (1.3)). The existence of maximisers $(\varphi,\psi)$ is more delicate and requires a finiteness condition, as formulated in Proposition 1.8.1 below.

The next sections are dedicated to more concrete examples that will be used through the rest of the book.

## 1.5 The One-Dimensional Case

When $\mathscr{X} = \mathscr{Y} = \mathbb{R}$, the Monge–Kantorovich problem has a particularly simple structure, because the class of "nice" transport maps contains at most a single element. Identify $\mu,\nu \in P(\mathbb{R})$ with their cumulative distribution functions $F$ and $G$ defined by

$$F(t) = \mu((-\infty,t]), \qquad G(t) = \nu((-\infty,t]), \qquad t\in\mathbb{R}.$$

Let the cost function be (momentarily) quadratic: $c(x,y) = |x-y|^2/2$. Since for $x_1 \le x_2$, $y_1 \le y_2$

$$c(y_2,x_1) + c(y_1,x_2) - c(y_1,x_1) - c(y_2,x_2) = (x_2-x_1)(y_2-y_1) \ge 0,$$

it seems natural to expect the optimal transport map to be monotonically increasing. It turns out that, on the real line, there is at most one such transport map: if $T$ is increasing and $T\#\mu = \nu$, then for all $t\in\mathbb{R}$

$$G(t) = \nu((-\infty,t]) = \mu((-\infty,T^{-1}(t)]) = F(T^{-1}(t)).$$

If $t = T(x)$, then the above equation reduces to $T(x) = G^{-1}(F(x))$. This formula determines $T$ uniquely, and has an interesting probabilistic interpretation: it is well-known that if $X$ is a random variable with *continuous* distribution function $F$, then $F(X)$ follows a uniform distribution on $(0,1)$. Conversely, if $U$ follows a uniform distribution, $G$ is any distribution function, and

$$G^{-1}(u) = \inf G^{-1}([u,1]) = \inf\{x\in\mathbb{R} : G(x)\ge u\}, \qquad 0<u<1,$$

is the *quantile function* of $X$, then the random variable $G^{-1}(U)$ has distribution function $G$. We say that $G^{-1}$ is the *left-continuous inverse* of $G$. In terms of push-forward maps, we can write $F\#\mu = \mathrm{Leb}|_{[0,1]}$ and $G^{-1}\#\mathrm{Leb}|_{[0,1]} = \nu$, with Leb standing for Lebesgue measure, and it is restricted to the interval $[0,1]$. Consequently, if $F$ is continuous and $G$ is arbitrary, then $T\#\mu = \nu$; we can view $T$ as pushing $\mu$ forward

to $\nu$ in two steps: firstly, $\mu$ is pushed forward to $\mathrm{Leb}|_{[0,1]}$ and secondly, $\mathrm{Leb}|_{[0,1]}$ is pushed forward to $\nu$.

Using the change of variables formula, we see that the total cost of $T$ is

$$C(T) = \int_{\mathbb{R}} |G^{-1}(F(x)) - x|^2 \, \mathrm{d}\mu(x) = \int_0^1 |G^{-1}(u) - F^{-1}(u))|^2 \, \mathrm{d}u.$$

If $F$ is discontinuous, then $F\#\mu$ is not Lebesgue measure, and $T$ is not necessarily defined. But there will exist an optimal transference plan $\pi \in \Pi(\mu,\nu)$ that is monotone in the following sense: there exists a set $\Gamma \subset \mathbb{R}^2$ such that $\pi(\Gamma) = 1$ and whenever $(x_i, y_i) \in \Gamma$,

$$|y_2 - x_1|^2 + |y_1 - x_2|^2 - |y_1 - x_1|^2 - |y_2 - x_2|^2 \geq 0.$$

Thus, mass at $x_1$ and $x_2$ can be split if need be, but in a monotone way. For example, if $\mu$ puts mass $1/2$ at $x_1 = -1$ and at $x_2 = 1$ and $\nu$ is uniform on $[-1,1]$. Then the transference plan spreads the mass of $x_1$ uniformly on $[-1,0]$, and the mass of $x_2$ uniformly on $[0,1]$. This is a particular case of the cyclical monotonicity that will be discussed in Sect. 1.7.

Elementary calculations show that the inequality

$$c(y_2,x_1) + c(y_1,x_2) - c(y_1,x_1) - c(y_2,x_2) \geq 0, \qquad x_1 \leq x_2; \quad y_1 \leq y_2$$

holds more generally than the quadratic cost $c(x,y) = |x-y|^2$. Specifically, it suffices that $c(x,y) = h(|x-y|)$ with $h$ convex on $\mathbb{R}_+$.

Since any distribution can be approximated by continuous distributions, in view of the above discussion, the following result from Villani [124, Theorem 2.18] should not be too surprising.

**Theorem 1.5.1 (Optimal Transport in $\mathbb{R}$)** *Let $\mu, \nu \in P(\mathbb{R})$ with distribution functions $F$ and $G$, respectively, and let the cost function be of the form $c(x,y) = h(|x-y|)$ with $h$ convex and nonnegative. Then*

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = \int_0^1 h(G^{-1}(u) - F^{-1}(u)) \, \mathrm{d}u.$$

*If the infimum is finite and $h$ is strictly convex, then the optimal transference plan is unique. Furthermore, if $F$ is continuous, then the infimum is attained by the transport map $T = G^{-1} \circ F$.*

The prototypical choice for $h$ is $h(z) = |z|^p$ with $p > 1$. This result allows in particular a direct evaluation of the Wasserstein distances for measures on the real line (see Chap. 2).

Note that no regularity is needed in order that the optimal transference plan be unique, unlike in higher dimensions (compare Theorem 1.8.2). The structure of solutions in the concave case ($0 < p < 1$) is more complicated, see McCann [94].

When $p = 1$, the cost function is convex but not strictly so, and solutions will not be unique. However, the total cost in Theorem 1.5.1 admits another representation that is often more convenient.

**Proposition 1.5.2 (Quantiles and Distribution Functions)** *If $F$ and $G$ are distribution functions, then*

$$\int_0^1 |G^{-1}(u) - F^{-1}(u)|\,\mathrm{d}u = \int_{\mathbb{R}} |G(x) - F(x)|\,\mathrm{d}x.$$

The proof is a simple application of Fubini's theorem; see page 13 in the supplement.

**Corollary 1.5.3** *If $c(x,y) = |x-y|$, then under the conditions of Theorem 1.5.1*

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = \int_{\mathbb{R}} |G(x) - F(x)|\,\mathrm{d}x.$$

## 1.6 Quadratic Cost

This section is devoted to the specific cost function

$$c(x,y) = \frac{\|x-y\|^2}{2}, \qquad x, y \in \mathscr{X},$$

where $\mathscr{X}$ is a separable Hilbert space. This cost is popular in applications, and leads to a lucid and elegant theory. The factor of $1/2$ does not affect the minimising coupling $\pi$ and leads to cleaner expressions. (It does affect the optimal dual pair, but in an obvious way.)

### *1.6.1 The Absolutely Continuous Case*

We begin with the Euclidean case, where $\mathscr{X} = \mathscr{Y} = (\mathbb{R}^d, \|\cdot\|)$ is endowed with the Euclidean metric, and use the Kantorovich duality to obtain characterisations of optimal maps.

Since the dual objective function to be maximised

$$\int_{\mathbb{R}^d} \varphi\,\mathrm{d}\mu + \int_{\mathbb{R}^d} \psi\,\mathrm{d}\nu$$

is increasing in $\varphi$ and $\psi$, one should seek functions that take values as large as possible subject to the constraint $\varphi(x) + \psi(y) \le \|x-y\|^2/2$. Suppose that an oracle tells us that some $\varphi \in L_1(\mu)$ is a good candidate. Then the largest possible $\psi$ satisfying $(\varphi, \psi) \in \Phi_c$ is

$$\psi(y) = \inf_{x\in\mathbb{R}^d}\left[\frac{\|x-y\|^2}{2} - \varphi(x)\right] = \frac{\|y\|^2}{2} + \inf_{x\in\mathbb{R}^d}\left[\frac{\|x\|^2}{2} - \varphi(x) - \langle x,y\rangle\right].$$

In other words,

$$\widetilde{\psi}(y) := \frac{\|y\|^2}{2} - \psi(y) = \sup_{x\in\mathbb{R}^d}\left[\langle x,y\rangle - \widetilde{\varphi}(x)\right], \qquad \widetilde{\varphi}(x) = \frac{\|x\|^2}{2} - \varphi(x).$$

As a supremum over affine functions (in $y$), $\widetilde{\psi}$ enjoys some useful properties. We remind the reader that a function $f : \mathscr{X} \to \mathbb{R}\cup\{\infty\}$ is *convex* if $f(tx+(1-t)y) \le tf(x)+(1-t)f(y)$ for all $x,y \in \mathscr{X}$ and $t \in [0,1]$. It is *lower semicontinuous* if for all $x \in \mathscr{X}$, $f(x) \le \liminf_{y\to x} f(y)$. Affine functions are convex and lower semicontinuous, and it straightforward from the definitions that both convexity and lower semicontinuity are preserved under the supremum operation. Thus, the function $\widetilde{\psi}$ is convex and lower semicontinuous. In particular, it is Borel measurable due to the following characterisation: $f$ is lower semicontinuous if and only if $\{x : f(x) \le \alpha\}$ is a closed set for all $\alpha \in \mathbb{R}$.

From the preceding subsection, we now know that optimal dual functions $\varphi$ and $\psi$ must take the form of the difference between $\|\cdot\|^2/2$ and a convex function. Given the vast wealth of knowledge on convex functions (Rockafellar [113]), it will be convenient to work with $\widetilde{\varphi}$ and $\widetilde{\psi}$, and to assume that $\widetilde{\psi} = (\widetilde{\varphi})^*$, where

$$f^*(y) = \sup_{x\in\mathbb{R}^d}[\langle x,y\rangle - f(x)], \qquad y \in \mathbb{R}^d$$

is the *Legendre transform* of $f$ ([113, Chapter 26]; [124, Chapter 2]), and is of fundamental importance in convex analysis. Now by symmetry, one can also replace $\widetilde{\varphi}$ by $(\widetilde{\psi})^* = (\widetilde{\varphi})^{**}$, so it is reasonable to expect that an optimal dual pair should take the form $(\|\cdot\|^2/2 - \widetilde{\varphi}, \|\cdot\|^2/2 - (\widetilde{\varphi})^*)$, with $\widetilde{\varphi}$ convex and lower semicontinuous.

The alternative representation of the dual objective value as

$$\int_{\mathbb{R}^d}\varphi\,\mathrm{d}\mu + \int_{\mathbb{R}^d}\psi\,\mathrm{d}\nu = \frac{1}{2}\int_{\mathbb{R}^d}\|x\|^2\,\mathrm{d}\mu(x) + \frac{1}{2}\int_{\mathbb{R}^d}\|y\|^2\,\mathrm{d}\nu(y) - \int_{\mathbb{R}^d}\widetilde{\varphi}\,\mathrm{d}\mu - \int_{\mathbb{R}^d}\widetilde{\psi}\,\mathrm{d}\nu$$

is valid under the integrability condition

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

that $\mu$ and $\nu$ have finite second moments. This condition also guarantees that an optimal $\varphi$ exists, as the conditions of Proposition 1.8.1 are satisfied. An alternative direct proof for the quadratic case can be found in Villani [124, Theorem 2.9].

Suppose that an optimal $\varphi$ is found. What can we say about optimal transference plans $\pi$? According to the duality, a necessary and sufficient condition is that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}\frac{\|x-y\|^2}{2}\,\mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d}\varphi\,\mathrm{d}\mu + \int_{\mathbb{R}^d}\psi\,\mathrm{d}\nu,$$

where $\psi = \|\cdot\|^2/2 - (\|\cdot\|^2/2 - \varphi)^*$. Equivalently (using Lemma 1.4.1),

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} [\widetilde{\varphi}(x) + (\widetilde{\varphi})^*(y) - \langle x, y\rangle] \, \mathrm{d}\pi(x,y) = 0. \tag{1.5}$$

Since we have $\widetilde{\varphi}(x) + (\widetilde{\varphi})^*(y) \geq \langle x, y\rangle$ everywhere, the integrand is nonnegative. Hence, the integral vanishes if and only if $\pi$ is concentrated on the set of $(x,y)$ such that $\widetilde{\varphi}(x) + \widetilde{\varphi}^*(y) = \langle x, y\rangle$. By definition of the Legendre transform as a supremum, this happens if and only if the supremum defining $\widetilde{\varphi}^*(y)$ is attained at $x$; equivalently

$$\widetilde{\varphi}(z) - \widetilde{\varphi}(x) \geq \langle z - x, y\rangle, \qquad z \in \mathscr{X}.$$

This condition is precisely the definition of $y$ being a *subgradient* of $\widetilde{\varphi}$ at $x$ [113, Chapter 23]. When $\widetilde{\varphi}$ is differentiable at $x$, its unique subgradient is the gradient $y = \nabla\widetilde{\varphi}(x)$ [113, Theorem 25.1]. If we are fortunate and $\widetilde{\varphi}$ is differentiable everywhere, or even $\mu$-almost everywhere, then the optimal transference plan $\pi$ is unique, and in fact induced from the transport map $\nabla\widetilde{\varphi}$. The problem, of course, is that $\widetilde{\varphi}$ may fail to be differentiable $\mu$-almost surely. This is remedied by assuming some regularity on the source measure $\mu$ in order to make sure that *any* convex function be differentiable $\mu$-almost surely, and is done via the following regularity result, which, roughly speaking, states that convex functions are differentiable almost surely. A stronger version is given in Rockafellar [113, Theorem 2.25], with an alternative proof in Alberti and Ambrosio [6, Chapter 2]. One could also combine the local Lipschitz property of convex functions [113, Chapter 10] with Rademacher's theorem (Villani [125, Theorem 10.8]).

**Theorem 1.6.1 (Differentiability of Convex Functions)** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex function with domain $\mathrm{dom} f = \{x \in \mathbb{R}^d : f(x) < \infty\}$ and let $\mathscr{N}$ be the set of points at which $f$ is not differentiable. Then $\mathscr{N} \cap \overline{\mathrm{dom} f}$ has Lebesgue measure 0.*

Theorem 1.6.1 is usually stated for the interior of $\mathrm{dom} f$, denoted $\mathrm{int}(\mathrm{dom} f)$, rather than the closure. But, since $A = \mathrm{dom} f$ is convex, its boundary has Lebesgue measure zero. To see this assume first that $A$ is bounded. If $\mathrm{int} A$ is empty, then $A$ lies in a lower dimensional subspace [113, Theorem 2.4]. Otherwise, without loss of generality $0 \in \mathrm{int} A$, and then by convexity of $A$, $\partial A \subseteq (1+\varepsilon)A$ for all $\varepsilon > 0$. When $A$ is unbounded, write it as $\cup_n A \cap [-n,n]^d$.

Another issue that might arise is that optimal $\varphi$'s might not exist. This is easily dealt with using Proposition 1.8.1. If we assume that $\mu$ and $\nu$ have finite second moments:

$$\int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x) < \infty \qquad \text{and} \qquad \int_{\mathbb{R}^d} \|y\|^2 \, \mathrm{d}\nu(y) < \infty,$$

then any transference plan $\pi \in \Pi(\mu, \nu)$ has a finite cost, as is seen from integrating the elementary inequality $\|x - y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and using Lemma 1.4.1:

$$C(\pi) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} [\|x\|^2 + \|y\|^2] \, \mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} \|y\|^2 \, \mathrm{d}\nu(y) < \infty.$$

With these tools, we can now prove a fundamental existence and uniqueness result for the Monge–Kantorovich problem. It has been proven independently by several authors, including Brenier [31], Cuesta-Albertos and Matrán [37], Knott and Smith [83], and Rachev and Rüschendorf [117].

**Theorem 1.6.2 (Quadratic Cost in Euclidean Spaces)** *Let $\mu$ and $\nu$ be probability measures on $\mathbb{R}^d$ with finite second moments, and suppose that $\mu$ is absolutely continuous with respect to Lebesgue measure. Then the solution to the Kantorovich problem is unique, and is induced from a transport map $T$ that equals $\mu$-almost surely the gradient of a convex function $\phi$. Furthermore, the pair $(\|x\|^2/2-\phi, \|y\|^2/2-\phi^*)$ is optimal for the dual problem.*

*Proof.* To alleviate the notation we write $\phi$ instead of $\widetilde{\varphi}$. By Proposition 1.8.1, there exists an optimal dual pair $(\varphi,\psi)$ such that $\phi(x)=\|x\|^2/2-\varphi(x)$ is convex and lower semicontinuous, and by the discussion in Sect. 1.1, there exists an optimal $\pi$. Since $\phi$ is $\mu$-integrable, it must be finite almost everywhere, i.e., $\mu(\mathrm{dom}\phi)=1$. By Theorem 1.6.1, if we define $\mathscr{N}$ as the set of nondifferentiability points of $\phi$, then $\mathrm{Leb}(\mathscr{N}\cap\mathrm{dom}\phi)=0$; as $\mu$ is absolutely continuous, the same holds for $\mu$. (Here Leb denotes Lebesgue measure.)

We conclude that $\mu(\mathrm{int}(\mathrm{dom}\phi)\setminus\mathscr{N})=1$. In other words, $\phi$ is differentiable $\mu$-almost everywhere, and so for $\mu$-almost any $x$, there exists a unique $y$ such that $\phi(x)+\phi^*(y)=\langle x,y\rangle$, and $y=\nabla\phi(x)$. This shows that $\pi$ is unique and induced from the transport map $\nabla\phi(x)$. The gradient $\nabla\phi$ is Borel measurable, since each of its coordinates can be written as $\limsup_{q\to 0,q\in\mathbb{Q}} q^{-1}(\phi(x+qv)-\phi(x))$ for some vector $v$ (the canonical basis of $\mathbb{R}^d$), which is Borel measurable because the limit superior is taken on countably many functions (and $\phi$ is measurable because it is lower semicontinuous).

### *1.6.2 Separable Hilbert Spaces*

The finite-dimensionality of $\mathbb{R}^d$ in the previous subsection was only used in order to apply Theorem 1.6.1, so one could hope to extend the results to infinite-dimensional separable Hilbert spaces.

Although there is no obvious parallel for Lebesgue measure (i.e., translation invariant) on infinite-dimensional Banach spaces, one can still define absolute continuity via Gaussian measures. Indeed, $\mu\in P(\mathbb{R}^d)$ is absolutely continuous with respect to Lebesgue measure if and only if the following holds: if $\mathscr{N}\subset\mathbb{R}^d$ is such that $\nu(\mathscr{N})=0$ for any nondegenerate Gaussian measure $\nu$, then $\mu(\mathscr{N})=0$. This definition can be extended to any separable Banach space $\mathscr{X}$ via projections, as follows. Let $\mathscr{X}^*$ be the (topological) dual of $\mathscr{X}$, consisting of all real-valued, continuous linear functionals on $\mathscr{X}$.

**Definition 1.6.3 (Gaussian Measures)** *A probability measure $\mu\in P(\mathscr{X})$ is a nondegenerate Gaussian measure if for any $\ell\in\mathscr{X}^*\setminus\{0\}$, $\ell\#\mu\in P(\mathbb{R})$ is a Gaussian measure with positive variance.*

**Definition 1.6.4 (Gaussian Null Sets and Absolutely Continuous Measures)** *A subset $\mathcal{N} \subset \mathcal{X}$ is a Gaussian null set if whenever $\nu$ is a nondegenerate Gaussian measure, $\nu(\mathcal{N}) = 0$. A probability measure $\mu \in P(\mathcal{X})$ is absolutely continuous if $\mu$ vanishes on all Gaussian null sets.*

Clearly, if $\nu$ is a nondegenerate Gaussian measure, then it is absolutely continuous.

As explained in Ambrosio et al. [12, Section 6.2], a version of Rademacher's theorem holds in separable Hilbert spaces: a locally Lipschitz function is Gâteaux differentiable except on a Gaussian null set of $\mathcal{X}$. Theorem 1.6.2 (and more generally, Theorem 1.8.2) extend to infinite dimensions; see [12, Theorem 6.2.10].

### *1.6.3 The Gaussian Case*

Apart from the one-dimensional case of Sect. 1.5, there is another special case in which there is a unique *and* explicit solution to the Monge–Kantorovich problem.

Suppose that $\mu$ and $\nu$ are Gaussian measures on $\mathbb{R}^d$ with zero means and nonsingular covariance matrices $A$ and $B$. By Theorem 1.6.2, we know that there exists a unique optimal map $T$ such that $T\#\mu = \nu$. Since linear push-forwards of Gaussians are Gaussian, it seems natural to guess that $T$ should be linear, and this is indeed the case.

Since $T$ is a linear map that should be the gradient of a convex function $\phi$, it must be that $\phi$ is quadratic, i.e., $\phi(x) - \phi(0) = \langle x, Qx\rangle$ for $x \in \mathbb{R}^d$ and some matrix $Q$. The gradient of $\phi$ at $x$ is $(Q + Q^t)x$ and the Hessian matrix is $Q + Q^t$. Thus, $T = Q + Q^t$ and since $\phi$ is convex, $T$ must be positive semidefinite.

Viewing $T$ as a matrix leads to the *Riccati equation* $TAT = B$ (since $T$ is symmetric). This is a quadratic equation in $T$, and so we wish to take square roots in a way that would isolate $T$. This is done by multiplying the equation from both sides with $A^{1/2}$:

$$[A^{1/2}TA^{1/2}][A^{1/2}TA^{1/2}] = A^{1/2}TATA^{1/2} = A^{1/2}BA^{1/2} = [A^{1/2}B^{1/2}][B^{1/2}A^{1/2}].$$

All matrices in brackets are positive semidefinite. By taking square roots and multiplying with $A^{-1/2}$, we finally find

$$T = A^{-1/2}[A^{1/2}BA^{1/2}]^{1/2}A^{-1/2}.$$

A straightforward calculation shows that $TAT = B$ indeed, and $T$ is positive definite, hence optimal. To calculate the transport cost $C(T)$, observe that $(T - I)\#\mu$ is a centred Gaussian measure with covariance matrix

$$TAT - TA - AT + A = A + B - A^{1/2}[A^{1/2}BA^{1/2}]^{1/2}A^{-1/2} - A^{-1/2}[A^{1/2}BA^{1/2}]^{1/2}A^{1/2}.$$

If $Y \sim \mathcal{N}(0, C)$, then $\mathbb{E}\|Y\|^2$ equals the trace of $C$, denoted tr$C$. Hence, by properties of the trace,

$$C(T) = \mathrm{tr}\left[A + B - 2(A^{1/2}BA^{1/2})^{1/2}\right]. \tag{1.6}$$

By continuity arguments, (1.6) is the total transport cost between any two Gaussian distributions with zero means, even if $A$ is singular.

If $AB = BA$, the above formulae simplify to

$$T = B^{1/2}A^{-1/2}, \qquad C(T) = \mathrm{tr}\left[A + B - 2A^{1/2}B^{1/2}\right] = \|A^{1/2} - B^{1/2}\|_F^2,$$

with $F$ the Frobenius norm.

If the means of $\mu$ and $\nu$ are $m$ and $n$, one simply needs to translate the measures. The optimal map and the total cost are then

$$Tx = n + A^{-1/2}[A^{1/2}BA^{1/2}]^{1/2}A^{-1/2}(x - m);$$

$$C(T) = \|n - m\|^2 + \mathrm{tr}[A + B - 2(A^{1/2}BA^{1/2})^{1/2}].$$

From this, we can deduce a lower bound on the total cost between *any* two measures in $\mathbb{R}^d$ in terms of their second order structure. This is worth mentioning, because such lower bounds are not very common (the Monge–Kantorovich problem is defined by an infimum, and thus typically easier to bound from above).

**Proposition 1.6.5 (Lower Bound for Quadratic Cost)** *Let $\mu, \nu \in P(\mathbb{R}^d)$ have means $m$ and $n$ and covariance matrices $A$ and $B$ and let $\pi$ be the optimal map. Then*

$$C(\pi) \geq \|n - m\|^2 + \mathrm{tr}[A + B - 2(A^{1/2}BA^{1/2})^{1/2}].$$

*Proof.* It will be convenient here to use the probabilistic terminology of Sect. 1.2. Let $X$ and $Y$ be random variables with distributions $\mu$ and $\nu$. Any coupling of $X$ and $Y$ will have covariance matrix of the form $C = \begin{pmatrix} A & V \\ V^t & B \end{pmatrix} \in \mathbb{R}^{2d \times 2d}$ for some matrix $V \in \mathbb{R}^{d \times d}$, constrained so that $C$ is positive semidefinite. This gives the lower bound

$$\inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi \|X - Y\|^2 = \|m - n\|^2 + \inf_{\pi \in \Pi(\mu,\nu)} \mathrm{tr}_\pi[A + B - 2V] \geq \|m - n\|^2 + \inf_{V : C \geq 0} \mathrm{tr}[A + B - 2V].$$

As we know from the Gaussian case, the last infimum is given by (1.6).

### *1.6.4 Regularity of the Transport Maps*

The optimal transport map $T$ between Gaussian measures on $\mathbb{R}^d$ is linear, so it is of course very smooth (analytic). The densities of Gaussian measures are analytic too, so that $T$ inherits the regularity of $\mu$ and $\nu$. Using the formula for $T$, one can show that a similar phenomenon takes place in the one-dimensional case. Though we do not have a formula for $T$ at our disposal when $\mu$ and $\nu$ are general absolutely continuous measures on $\mathbb{R}^d$, $d \geq 2$, it turns out that even in that case, $T$ inherits the regularity of $\mu$ and $\nu$ if some convexity conditions are satisfied.

To guess what kind of results can be hoped for, let us first examine the case $d = 1$. Let $F$ and $G$ denote the distribution functions of $\mu$ and $\nu$, respectively. Suppose that $G$ is continuously differentiable and that $G' > 0$ on some open interval (finite or not) $I$ such that $\nu(I) = 1$. Then the inverse function theorem says that $G^{-1}$ is also continuously differentiable. Recall that the *support* of a (Borel) probability measure $\mu$ (denoted $\text{supp}\mu$) is the smallest closed set $K$ such that $\mu(K) = 1$. A simple application of the chain rule (see page 19 in the supplement) gives:

**Theorem 1.6.6 (Regularity in $\mathbb{R}$)** *Let $\mu, \nu \in P(\mathbb{R})$ possess distribution functions $F$ and $G$ of class $C^k$, $k \geq 1$. Suppose further that $\text{supp}\nu$ is an interval $I$ (possibly unbounded) and that $G' > 0$ on the interior of $I$. Then the optimal map is of class $C^k$ as well. If $F, G \in C^0$ are merely continuous, then so is the optimal map.*

The assumption on the support of $\nu$ is important: if $\mu$ is Lebesgue measure on $[0,1]$ and the support of $\nu$ is disconnected, then $T$ cannot even be continuous, no matter how smooth $\nu$ is.

The argument above cannot be easily extended to measures on $\mathbb{R}^d$, $d \geq 2$, because there is no explicit formula available for the optimal maps. As before, we cannot expect the optimal map to be continuous if the support of $\nu$ is disconnected. It turns out that the condition on the support of $\nu$ is not connectedness, but rather convexity. This was shown by Caffarelli, who was able to prove ([32] and the references within) that the optimal maps have the same smoothness as the measures. To state the result, we recall the following notation for an open $\Omega \subseteq \mathbb{R}^d$, $k \geq 0$ and $\alpha \in (0,1]$. We say that $f \in C^{k,\alpha}(\Omega)$ if all the partial derivatives of order $k$ of $f$ are locally $\alpha$-Hölder on $\Omega$. For example, if $k = 1$, this means that for any $x \in \Omega$ there exists a constant $L$ and an open ball $B$ containing $x$ such that

$$\|\nabla f(z) - \nabla f(y)\| \leq L\|y - z\|^{\alpha}, \qquad y, z \in B.$$

Note that $f \in C^{k+1} \Longrightarrow f \in C^{k,\beta} \Longrightarrow f \in C^{k,\alpha} \Longrightarrow f \in C^k$, for $0 \leq \alpha \leq \beta \leq 1$ so $\alpha$ gives a "fractional" degree of smoothness for $f$. Moreover, $C^{k,0} = C^k$ and $C^{k,1}$ is quite close to $C^{k+1}$, since Lipschitz functions are almost surely differentiable.

**Theorem 1.6.7 (Regularity of Transport Maps)** *Fix open sets $\Omega_1, \Omega_2 \subseteq \mathbb{R}^d$, with $\Omega_2$ convex, and absolutely continuous measures $\mu, \nu \in P(\mathbb{R}^d)$ with finite second moments and bounded, strictly positive densities $f, g$, respectively, such that $\mu(\Omega_1) = 1 = \nu(\Omega_2)$. Let $\phi$ be such that $\nabla\phi \# \mu = \nu$.*

1. *If $\Omega_1$ and $\Omega_2$ are bounded and $f, g$ are bounded below, then $\phi$ is strictly convex and of class $C^{1,\alpha}(\Omega_1)$ for some $\alpha > 0$.*
2. *If $\Omega_1 = \Omega_2 = \mathbb{R}^d$ and $f, g \in C^{0,\alpha}$, then $\phi \in C^{2,\alpha}(\Omega_1)$.*

*If in addition $f, g \in C^{k,\alpha}$, then $\phi \in C^{k+2,\alpha}(\Omega_1)$.*

In other words, the optimal map $T = \nabla\phi \in C^{k+1,\alpha}(\Omega_1)$ is one derivative smoother than the densities, so has the same smoothness as the measures $\mu, \nu$.

Theorem 1.6.7 will be used in two ways in this book. Firstly, it is used to derive criteria for a Karcher mean of a collection of measures to be the Fréchet mean of that collection (Theorem 3.1.15). Secondly, it allows one to obtain very smooth estimates

for the transport maps. Indeed, any two measures $\mu$ and $\nu$ can be approximated by measures satisfying the second condition: one can approximate them by discrete measures using the law of large numbers and then employ a convolution with, e.g., a Gaussian measure (see, for instance, Theorem 2.2.7). It is not obvious that the transport maps between the approximations converge to the transport maps between the original measures, but we will see this to be true in the next section.

## 1.7 Stability of Solutions Under Weak Convergence

In this section, we discuss the behaviour of the solution to the Monge–Kantorovich problem when the measures $\mu$ and $\nu$ are replaced by approximations $\mu_n$ and $\nu_n$. Since any measure can be approximated by discrete measures *or* by smooth measures, this allows us to benefit from both worlds. On the one hand, approximating $\mu$ and $\nu$ with discrete measures leads to the finite discrete problem of Sect. 1.3 that can be solved exactly. On the other hand, approximating $\mu$ and $\nu$ with Gaussian convolutions thereof leads to very smooth measures (at least on $\mathbb{R}^d$), and so the regularity results of the previous section imply that the respective optimal maps will also be smooth. Finally, in applications, one would almost always observe the measures of interest $\mu$ and $\nu$ with a certain amount of noise, and it is therefore of interest to control the error introduced by the noise. In image analysis, $\mu$ can represent an image that has undergone blurring, or some other perturbation (Amit et al. [13]). In other applications, the noise could be due to sampling variation, where instead of $\mu$ one observes a discrete measure $\mu_N$ obtained from realisations $X_1, \dots, X_N$ of random elements with distribution $\mu$ as $\mu_N = N^{-1} \sum_{i=1}^{N} \delta\{X_i\}$ (see Chap. 4).

In Sect. 1.7.1, we will see that the optimal transference plan $\pi$ depends continuously on $\mu$ and $\nu$. With this result under one's belt, one can then deduce an analogous property for the optimal map $T$ from $\mu$ to $\nu$ given some regularity of $\mu$, as will be seen in Sect. 1.7.2.

We shall assume throughout this section that $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly, which, we recall, means that $\int_{\mathscr{X}} f \, \mathrm{d}\mu_n \to \int_{\mathscr{X}} f \, \mathrm{d}\mu$ for all continuous bounded $f : \mathscr{X} \to \mathbb{R}$. The following equivalent definitions for weak convergence will be used not only in this section, but elsewhere as well.

**Lemma 1.7.1 (Portmanteau)** *Let $\mathscr{X}$ be a complete separable metric space and let $\mu, \mu_n \in P(\mathscr{X})$. Then the following are equivalent:*

- *$\mu_n \to \mu$ weakly;*
- *$F_n(x) \to F(x)$ for any continuity point $x$ of $F$. Here $\mathscr{X} = \mathbb{R}^d$, $F_n$ is the distribution function of $\mu_n$ and $F$ is that of $\mu$;*
- *for any open $G \subseteq \mathscr{X}$, $\liminf \mu_n(G) \geq \mu(G)$;*
- *for any closed $F \subseteq \mathscr{X}$, $\limsup \mu_n(F) \leq \mu(F)$;*
- *$\int h \, \mathrm{d}\mu_n \to \int h \, \mathrm{d}\mu$ for any bounded measurable $h$ whose set of discontinuity points is a $\mu$-null set.*

For a proof, see, for instance, Billingsley [24, Theorem 2.1]. The equivalence with the last condition can be found in Pollard [104, Section III.2].

### 1.7.1 Stability of Transference Plans and Cyclical Monotonicity

In this subsection, we state and sketch the proof of the fact that if $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly, then the optimal transference plans $\pi_n \in \Pi(\mu_n, \nu_n)$ converge to an optimal $\pi \in \Pi(\mu, \nu)$. The result, as stated in Villani [125, Theorem 5.20], is valid on complete separate metric spaces with general cost functions, and reads as follows.

**Theorem 1.7.2 (Weak Convergence and Optimal Plans)** *Let $\mu_n$ and $\nu_n$ converge weakly to $\mu$ and $\nu$, respectively, in $P(\mathscr{X})$ and let $c : \mathscr{X}^2 \to \mathbb{R}_+$ be continuous. If $\pi_n \in \Pi(\mu_n, \nu_n)$ are optimal transference plans and*

$$\limsup_{n\to\infty} \int_{\mathscr{X}^2} c(x,y)\, \mathrm{d}\pi_n(x,y) < \infty.$$

*then $(\pi_n)$ is a tight sequence and each of its weak limits $\pi \in \Pi(\mu, \nu)$ is optimal.*

One can even let $c$ vary with $n$ under some conditions.

Let $c(x,y) = \|x-y\|^2/2$. We prefer to keep the notation $c(\cdot,\cdot)$ in order to stress the generality of the arguments. A key idea in the proof is the replacement of optimality by another property called *cyclical monotonicity*, which behaves nicely with respect to weak convergence. To motivate this property, we recall the discrete case of Sect. 1.3 where $\mu = N^{-1}\sum_{i=1}^N \delta\{x_i\}$ and $\nu = N^{-1}\sum_{i=1}^N \delta\{y_i\}$. There exists an optimal transference plan $\pi$ induced from a permutation $\sigma_0 \in S_N$. Since the ordering of $\{x_i\}$ and $\{y_i\}$ is irrelevant in the representations of $\mu$ and $\nu$, we may assume without loss of generality that $\sigma_0$ is the identity permutation. Then, by definition of optimality,

$$\sum_{i=1}^N c(x_i, y_i) \le \sum_{i=1}^N c(x_i, y_{\sigma(i)}), \qquad \sigma \in S_N. \tag{1.7}$$

If $\sigma$ is the identity except for a subset $i_1, \dots, i_n$, $n \le N$, then in particular

$$\sum_{k=1}^n c(x_{i_k}, y_{i_k}) \le \sum_{k=1}^n c(x_{i_k}, y_{\sigma(i_k)}), \qquad \sigma \in S_n,$$

and if we choose $\sigma(i_k) = i_{k-1}$ with $i_0 = i_n$, this writes

$$\sum_{k=1}^n c(x_{i_k}, y_{i_k}) \le \sum_{k=1}^n c(x_{i_k}, y_{i_{k-1}}). \tag{1.8}$$

By decomposing a permutation $\sigma \in S_N$ to disjoint cycles, one can verify that (1.8) implies (1.7). This will be useful since, as it turns out, a variant of (1.8) holds for arbitrary measures $\mu$ and $\nu$ for which there is no relevant finite $N$ as in (1.7).

**Definition 1.7.3 (Cyclically Monotone Sets and Measures)** *A set $\Gamma \subseteq \mathscr{X}^2$ is cyclically monotone if for any $n$ and any $(x_1,y_1),\dots,(x_n,y_n) \in \Gamma$,*

$$\sum_{i=1}^{n} c(x_i,y_i) \le \sum_{i=1}^{n} c(x_i,y_{i-1}), \qquad (y_0 = y_n). \tag{1.9}$$

*A probability measure $\pi$ on $\mathscr{X}^2$ is cyclically monotone if there exists a monotone Borel set $\Gamma$ such that $\pi(\Gamma) = 1$.*

The relevance of cyclical monotonicity becomes clear from the following observation. If $\mu$ and $\nu$ are discrete uniform measures on $N$ points and $\sigma$ is an optimal permutation for the Monge–Kantorovich problem, then the coupling $\pi = (1/N)\sum_{i=1}^{N} \delta\{(x_i,y_{\sigma(i)})\}$ is cyclically monotone. In fact, even if the optimal permutation is not unique, the set

$$\Gamma = \{(x_i,y_{\sigma(i)}) : i = 1,\dots,N, \sigma \in S_N \text{ optimal}\}$$

is cyclically monotone. Furthermore, $\pi \in \Pi(\mu,\nu)$ is optimal if and only if it is cyclically monotone, if and only if $\pi(\Gamma) = 1$. It is heuristically easy to see that cyclical monotonicity is a necessary condition for optimality:

**Proposition 1.7.4 (Optimal Plans Are Cyclically Monotone)** *Let $\mu,\nu \in P(\mathscr{X})$ and suppose that the cost function $c$ is nonnegative and continuous. Assume that the optimal $\pi \in \Pi(\mu,\nu)$ has a finite total cost. Then* supp$\pi$ *is cyclically monotone. In particular, $\pi$ is cyclically monotone.*

The idea of the proof is that if for some $(x_1,y_1),\dots,(x_n,y_n)$ in the support of $\pi$,

$$\sum_{i=1}^{n} c(x_i,y_i) > \sum_{i=1}^{n} c(x_i,y_{i-1}),$$

then by continuity of $c$, the same inequality holds on some balls of positive measure. One can then replace $\pi$ by a measure having $(x_i,y_{i-1})$ rather than $(x_i,y_i)$ in its support, and this measure will incur a lower cost than $\pi$. A rigorous proof can be found in Gangbo and McCann [59, Theorem 2.3].

Thus, optimal transference plans $\pi$ solve infinitely many discrete Monge–Kantorovich problems emanating from their support. More precisely, for any finite collection $(x_i,y_i) \in$ supp$\pi$, $i = 1,\dots,N$ and any permutation $\sigma \in S_N$, (1.7) is satisfied. Therefore, the identity permutation is optimal between the measures $(1/N)\sum \delta\{x_i\}$ and $(1/N)\sum \delta\{y_j\}$.

In the same spirit as $\Gamma$ defined above for the discrete case, one can strengthen Proposition 1.7.4 and prove existence of a cyclically monotone set $\Gamma$ that includes the support of *any* optimal transference plan $\pi$: take $\Gamma = \cup$supp$(\pi)$ for $\pi$ optimal.

The converse of Proposition 1.7.4 also holds.

**Proposition 1.7.5 (Cyclically Monotone Plans Are Optimal)** *Let $\mu,\nu \in P(\mathscr{X})$, $c : \mathscr{X}^2 \to \mathbb{R}_+$ continuous and $\pi \in \Pi(\mu,\nu)$ a cyclically monotone measure with $C(\pi)$ finite. Then $\pi$ is optimal in $\Pi(\mu,\nu)$.*

Let us sketch the proof in the quadratic case $c(x,y) = \|x-y\|^2/2$ and see how convexity comes into play. Straightforward algebra shows that (1.9) is equivalent, in the quadratic case, to

$$\sum_{i=1}^{n} \langle y_i, x_{i+1} - x_i \rangle \leq 0, \qquad (x_{n+1} = x_1). \tag{1.10}$$

Fix $(x_0, y_0) \in \Gamma = \mathrm{supp}\pi$ and define $\phi : \mathscr{X} \to \mathbb{R} \cup \{\infty\}$ by

$$\phi(x) = \sup \{ \langle y_0, x_1 - x_0 \rangle + \cdots + \langle y_{m-1}, x_m - x_{m-1} \rangle + \langle y_m, x - x_m \rangle : m \in \mathbb{N}, \quad (x_i, y_i) \in \Gamma \}.$$

This function is defined as a supremum of affine functions, and is therefore convex and lower semicontinuous. Cyclical monotonicity of $\Gamma$ implies that $\phi(x_0) = 0$, so $\phi$ is not identically infinite (it would have been so if $\Gamma$ were not cyclically monotone). Straightforward computations show that $\Gamma$ is included in the subdifferential of $\phi$: $y$ is a subgradient of $\phi$ at $x$ when $(x,y) \in \Gamma$. Optimality of $\pi$ then follows by weak duality, since $\pi$ assigns full measure to the set of $(x,y)$ such that $\phi(x) + \phi^*(y) = \langle x, y \rangle$; see (1.5) and the discussion around it.

The argument for more general costs follows similar lines and is sketched at the end of this subsection.

Given these intermediary results, it is now instructive to prove Theorem 1.7.2.

*Proof (Proof of Theorem 1.7.2).* Since $\mu_n \to \mu$ weakly, it is a tight sequence, and similarly for $\nu_n$. Consequently, the entire set of plans $\cup_n \Pi(\mu_n, \nu_n)$ is tight too (see the discussion before deriving (1.3)). Therefore, up to a subsequence, $(\pi_n)$ has a weak limit $\pi$. We need to show that $\pi$ is cyclically monotone and that $C(\pi)$ is finite. The latter is easy, since $c_M(x,y) = \min(M, c(x,y))$ is continuous and bounded:

$$C(\pi) = \lim_{M\to\infty} \int_{\mathscr{X}^2} c_M \, \mathrm{d}\pi = \lim_{M\to\infty} \lim_{n\to\infty} \int_{\mathscr{X}^2} c_M \, \mathrm{d}\pi_n \leq \liminf_{n\to\infty} \int_{\mathscr{X}^2} c \, \mathrm{d}\pi_n < \infty.$$

To show that $\pi$ is cyclically monotone, fix $(x_1, y_1), \ldots, (x_N, y_N) \in \mathrm{supp}\pi$. We show that there exist $(x_k^n, y_k^n) \in \mathrm{supp}\pi_n$ that converge to $(x_k, y_k)$. Once this is established, we conclude from the cyclical monotonicity of $\mathrm{supp}\pi_n$ and the continuity of $c$ that

$$\sum_{k=1}^{N} c(x_k, y_k) = \lim_{n\to\infty} \sum_{k=1}^{N} c(x_k^n, y_k^n) \leq \lim_{n\to\infty} \sum_{k=1}^{N} c(x_k^n, y_{k-1}^n) = \sum_{k=1}^{N} c(x_k, y_{k-1}).$$

The existence proof for the sequence is standard. For $\varepsilon > 0$, let $B = B_\varepsilon(x_k, y_k)$ be an open ball around $(x_k, y_k)$. Then $\pi(B) > 0$ and by the portmanteau Lemma 1.7.1, $\pi_n(B) > 0$ for sufficiently large $n$. It follows that there exist $(x_k^n, y_k^n) \in B \cap \mathrm{supp}\pi_n$. Let $\varepsilon = 1/m$, say, then for all $n \geq N_m$ we can find $(x_k^n, y_k^n) \in \mathrm{supp}\mu_n$ of distance $2/m$ from $(x_k, y_k)$. We can choose $N_{m+1} > N_m$ without loss of generality in order to complete the proof.

A few remarks are in order. Firstly, quadratic cyclically monotone sets (with respect to $\|x-y\|^2/2$) are included in the subdifferential of convex functions. The converse is also true, as can be easily deduced from summing up the subgradient inequalities

$$\phi(x_{i+1}) \geq \phi(x_i) + \langle y_i, x_{i+1} - x_i \rangle, \qquad i = 1, \ldots, N,$$

where $y_i$ is a subgradient of $\phi$ at $x_i$. For future reference, we state this characterisation as a theorem (which is valid in infinite dimensions too).

**Theorem 1.7.6 (Rockafellar [112])** *A nonempty $\Gamma \subseteq \mathscr{X}^2$ is quadratic cyclically monotone if and only if it is included in the graph of the subdifferential of a lower semicontinuous convex function that is not identically infinite.*

Secondly, we have not used at all the Kantorovich duality, merely its weak form. The machinery of cyclical monotonicity can be used in order to prove the duality Theorem 1.4.2. This is indeed the strategy of Villani [125, Chapter 5], who explains its advantage with respect to Hahn–Banach-type duality proofs.

Lastly, the idea of the proof of Proposition 1.7.5 generalises to other costs in a natural way. Given a cyclically monotone (with respect to a cost function $c$) set $\Gamma$ and a fixed pair $(x_0, y_0) \in \Gamma$, define (Rüschendorf [116])

$$\varphi(x) = \inf\{c(x_1, y_0) - c(x_0, y_0) + c(x_m, y_{m-1}) - c(x_{m-1}, y_{m-1}) + c(x, y_m) - c(x_m, y_m)\}.$$

Then under some conditions, $(\varphi, \psi)$ is dual optimal for some $\psi$. As explained in Sect. 1.8, $\psi$ can be chosen to be essentially $\varphi^c$ (as defined in that section).

### *1.7.2 Stability of Transport Maps*

We now extend the weak convergence of $\pi_n$ to $\pi$ of the previous subsection to convergence of optimal maps. Because of the applications we have in mind, we shall work exclusively in the Euclidean space $\mathscr{X} = \mathbb{R}^d$ with the quadratic cost function; our results can most likely be extended to more general situations.

In this setting, we know that optimal plans are supported on graphs of subdifferentials of convex functions. Suppose that $\pi_n$ is induced by $T_n$ and $\pi$ is induced by $T$. Then in some sense, the weak convergence of $\pi_n$ to $\pi$ yields convergence of the graphs of $T_n$ to the graph of $T$. Our goal is to strengthen this to uniform convergence of $T_n$ to $T$. Roughly speaking, we show the following: there exists a set $A$ with $\mu(A) = 1$ and such that $T_n$ converge uniformly to $T$ on every compact subset of $A$. For the reader's convenience, we give a user-friendly version here; a more general statement is given in Proposition 1.7.11 below.

**Theorem 1.7.7 (Uniform Convergence of Optimal Maps)** *Let $\mu_n, \mu$ be absolutely continuous measures with finite second moments on an open convex set $U \subseteq \mathbb{R}^d$ such that $\mu_n \to \mu$ weakly, and let $\nu_n \to \nu$ weakly with $\nu_n, \nu \in P(\mathbb{R}^d)$ with finite second moments. If $T_n$ and $T$ are continuous on $U$ and $C(T_n)$ is bounded uniformly in $n$, then*

$$\sup_{x\in\Omega} \|T_n(x) - T(x)\| \to 0, \qquad n \to \infty,$$

*for any compact* $\Omega \subseteq U$.

Since $T_n$ and $T$ are only defined up to Lebesgue null sets, it will be more convenient to work directly with the subgradients. That is, we view $T_n$ and $T$ as *set-valued* functions that to each $x \in \mathbb{R}^d$ assign a (possibly empty) subset of $\mathbb{R}^d$. In other words, $T_n$ and $T$ take values in the power set of $\mathbb{R}^d$, denoted by $2^{\mathbb{R}^d}$.

Let $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex, $y_1 \in \partial\phi(x_1)$ and $y_2 \in \partial\phi(x_2)$. Putting $n = 2$ in the definition of cyclical monotonicity (1.10) gives

$$\langle y_2 - y_1, x_2 - x_1 \rangle \geq 0.$$

This property (which is weaker than cyclical monotonicity) is important enough to have its own name. Following the notation of Alberti and Ambrosio [6], we call a set-valued function (or multifunction) $u : \mathbb{R}^d \to 2^{\mathbb{R}^d}$ *monotone* if whenever $y_i \in u(x_i)$, $i = 1, 2$,

$$\langle y_2 - y_1, x_2 - x_1 \rangle \geq 0.$$

If $d = 1$, this simply means that $u$ is a nondecreasing (set-valued) function. For example, one can define $u(x) = \{0\}$ for $x \in [0, 1)$, $u(1) = [0, 1]$ and $u(x) = \emptyset$ if $x \notin [0, 1]$. Next, $u$ is said to be *maximally monotone* if no points can be added to its graph while preserving monotonicity:

$$\big\{\langle y' - y, x' - x\rangle \geq 0 \quad \text{whenever } y \in u(x)\big\} \quad \Longrightarrow \quad y' \in u(x').$$

It will be convenient to identify $u$ with its graph; we will often write $(x, y) \in u$ to mean $y \in u(x)$. Note that $u(x)$ can be empty, even when $u$ is maximally monotone. The previous example for $u$ is not maximally monotone, but it will be if we modify $u(0)$ to be $(-\infty, 0]$ and $u(1)$ to be $[0, \infty)$.

Of course, if $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex, then $u = \partial\phi$ is monotone. It follows from Theorem 1.7.6 that $u$ is maximally cyclically monotone (no points can be added to its graph while preserving cyclical monotonicity). It can actually be shown that $u$ is maximally monotone [6, Section 7]. In what follows, we will always work with subdifferentials of convex functions, so unless stated otherwise, $u$ will always be assumed maximally monotone.

Maximally monotone functions enjoy the following very useful continuity property. It is proven in [6, Corollary 1.3] and will be used extensively below.

**Proposition 1.7.8 (Continuity at Singletons)** *Let* $x \in \mathbb{R}^d$ *such that* $u(x) = \{y\}$ *is a singleton. Then* $u$ *is nonempty on some neighbourhood of* $x$ *and it is continuous at* $x$*: if* $x_n \to x$ *and* $y_n \in u(x_n)$*, then* $y_n \to y$.

Notice that this result implies that if a convex function $\phi$ is differentiable on some open set $E \subseteq \mathbb{R}^d$, then it is continuously differentiable there (Rockafellar [113, Corollary 25.5.1]).

If $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is any function, one can define its subgradient at $x$ locally as

$$\begin{aligned}\partial f(x) &= \{y : f(z) \geq f(x) + \langle y, z-x\rangle + o(\|z-x\|)\} \\ &= \left\{y : \liminf_{z\to x} \frac{f(z)-f(x)+\langle y, z-x\rangle}{\|z-x\|} \geq 0\right\}.\end{aligned}$$

(See the discussion after Theorem 1.8.2.) When $f$ is convex, one can remove the $o(\|z-x\|)$ term and the inequality holds for all $z$, i.e., globally and not locally. Since monotonicity is related to convexity, it should not be surprising that monotonicity is in some sense a local property. Suppose that $u(x_0) = \{y_0\}$ is a singleton and that for some $y^* \in \mathbb{R}^d$,

$$\langle y - y^*, x - x_0\rangle \geq 0$$

for all $x \in \mathbb{R}^d$ and $y \in u(x)$. Then by maximality, $y^*$ must equal $y_0$. By "local property", we mean that the conclusion $y^* = y_0$ holds if the above inequality holds for $x$ in a small neighbourhood of $x_0$ (an open set that includes $x_0$). We will need a more general version of this result, replacing neighbourhoods by a weaker condition that can be related to Lebesgue points. The strengthening is somewhat technical; the reader can skip directly to Lemma 1.7.10 and assume that $G$ is open without losing much intuition.

Let $B_r(x_0) = \{x : \|x - x_0\| < r\}$ for $r \geq 0$ and $x_0 \in \mathbb{R}^d$. The interior of a set $G \subseteq \mathbb{R}^d$ is denoted by $\mathrm{int}G$ and the closure by $\overline{G}$. If $G$ is measurable, then $\mathrm{Leb}G$ denotes the Lebesgue measure of $G$. Finally, $\mathrm{conv}G$ denotes the convex hull of $G$.

A point $x_0$ is a *Lebesgue point* (or of *Lebesgue density*) of a measurable set $G \subseteq \mathbb{R}^d$ if for any $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that

$$\frac{\mathrm{Leb}(B_t(x_0) \cap G)}{\mathrm{Leb}(B_t(x_0))} > 1 - \varepsilon, \qquad 0 < t < t_\varepsilon.$$

An illuminating example is the set $\{y \leq \sqrt{|x|}\}$ in $\mathbb{R}^2$ (see Fig. 1.1). Since the "slope" of the square root is infinite, $x_0 = (0,0)$ is a Lebesgue point, but the fraction above is strictly smaller than one, for all $t > 0$.

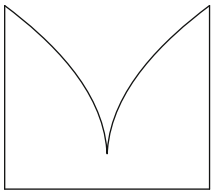

Fig. 1.1: The set $G = \{(x,y) : |x| \leq 1,\ -0.2 \leq y \leq \sqrt{|x|}\}$

We denote the set of points of Lebesgue density of $G$ by $G^{\mathrm{den}}$. Here are some facts about $G^{\mathrm{den}}$: clearly, $\mathrm{int}G \subseteq G^{\mathrm{den}} \subseteq \overline{G}$. Stein and Shakarchi [121, Chapter 3, Corollary 1.5] show that $\mathrm{Leb}(G \setminus G^{\mathrm{den}}) = 0$ (and $\mathrm{Leb}(G^{\mathrm{den}} \setminus G) = 0$, so $G^{\mathrm{den}}$ is very

close to $G$). By the Hahn–Banach theorem, $G^{\text{den}} \subseteq \text{int}(\text{conv}(G))$: indeed, if $x$ is not in $\text{int}(\text{conv}G)$, then there is a separating hyperplane between $x$ and $\text{conv}G \supseteq G$, so the fraction above is at most $1/2$ for all $t > 0$.

The "denseness" of Lebesgue points is materialised in the following result. It is given as exercise in [121] when $d = 1$, and the proof can be found on page 27 in the supplement.

**Lemma 1.7.9 (Density Points and Distance)** *Let $x_0$ be a point of Lebesgue density of a measurable set $G \subseteq \mathbb{R}^d$. Then*

$$\delta(z) = \delta_G(z) = \inf_{x \in G} \|z - x\| = o(\|z - x_0\|), \qquad \text{as } z \to x_0.$$

Of course, this result holds for any $x_0 \in \overline{G}$ if the little $o$ is replaced by big $O$, since $\delta$ is Lipschitz. When $x_0 \in \text{int}G$, this is trivial because $\delta$ vanishes on $\text{int}G$.

The important part here is the following corollary: for almost all $x \in G$, $\delta(z) = o(\|z - x\|)$ as $z \to x$. This can be seen in other ways: since $\delta$ is Lipschitz, it is differentiable almost everywhere. If $x \in \overline{G}$ and $\delta$ is differentiable at $x$, then $\nabla\delta(x)$ must be 0 (because $\delta$ is minimised there), and then $\delta(z) = o(\|z - x\|)$. We just showed that $\delta$ is differentiable with vanishing derivative at all Lebesgue points of $x$. The converse is not true: $G = \{\pm 1/n\}_{n=1}^{\infty}$ has no Lebesgue points, but $\delta(y) \leq 4y^2$ as $y \to 0$.

The locality of monotone functions can now be stated as follows. It is proven on page 27 of the supplement.

**Lemma 1.7.10 (Local Monotonicity)** *Let $x_0 \in \mathbb{R}^d$ such that $u(x_0) = \{y_0\}$ and $x_0$ is a Lebesgue point of a set $G$ satisfying*

$$\langle y - y^*, x - x_0 \rangle \geq 0 \qquad \forall x \in G \, \forall y \in u(x).$$

*Then $y^* = y_0$. In particular, the result is true if the inequality holds on $G = O \setminus \mathcal{N}$ with $\emptyset \neq O$ open and $\mathcal{N}$ Lebesgue negligible.*

These continuity properties cannot be of much use unless $u(x)$ is a singleton for reasonably many values of $x$. Fortunately, this is indeed the case: the set of points $x$ such that $u(x)$ contains more than one element has Lebesgue measure 0 (see Alberti and Ambrosio [6, Remark 2.3] for a stronger result). Another issue is that $u$ may be empty, and convexity comes into play here again. Let $\text{dom}u = \{x : u(x) \neq \emptyset\}$. Then there exists a convex closed set $K$ such that

$$\text{int}K \subseteq \text{dom}u \subseteq K.$$

[6, Corollary 1.3(2)]. Although $\text{dom}u$ itself may fail to be convex, it is almost convex in the above sense. By convexity, $K \setminus \text{int}K$ has Lebesgue measure 0 (see the discussion after Theorem 1.6.1) and so the set of points in $K$ where $u$ is not a singleton,

$$\{x \in K : u(x) = \emptyset\} \cup \{x \in K : u(x) \text{ contains more than one point}\},$$

has Lebesgue measure 0, and $u(x)$ is empty for all $x \notin K$. (It is in fact not difficult to show that if $x \in \partial K$, then $u(x)$ cannot be a singleton, by the Hahn–Banach theorem.)

With this background on monotone functions at our disposal, we are now ready to state the stability result for the optimal maps. We assume the following.

**Assumptions 1** *Let $\mu_n, \mu, \nu_n, \nu \in P(\mathbb{R}^d)$ with optimal couplings (with respect to quadratic cost) $\pi_n \in \Pi(\mu_n, \nu_n)$, $\pi \in \Pi(\mu, \nu)$ and convex potentials $\phi_n$ and $\phi$, respectively, such that*

- *(*convergence*) $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly;*
- *(*finiteness*) the optimal couplings $\pi_n \in \Pi(\mu_n, \nu_n)$ satisfy*

$$\limsup_{n\to\infty} \int_{\mathscr{X}^2} \frac{1}{2}\|x-y\|^2 \, \mathrm{d}\pi_n(x,y) < \infty;$$

- *(*unique limit*) the optimal $\pi \in \Pi(\mu, \nu)$ is unique.*

*We further denote the subgradients $\partial\phi_n$ and $\partial\phi$ by $u_n$ and $u$, respectively.*

These assumptions imply that $\pi$ has a finite total cost. This can be shown by the lim inf argument in the proof of Theorem 1.7.2 but also from the uniqueness of $\pi$. As a corollary of the uniqueness of $\pi$, it follows that $\pi_n \to \pi$ weakly; notice that this holds even if $\pi_n$ is not unique for any $n$. We will now translate this weak convergence to convergence of the maximal monotone maps $u_n$ to $u$, in the following form.

**Proposition 1.7.11 (Uniform Convergence of Optimal Maps)** *Let Assumptions 1 hold true and denote $E = \mathrm{supp}\mu$ and $E^{\mathrm{den}}$ the set of its Lebesgue points. Let $\Omega$ be a compact subset of $E^{\mathrm{den}}$ on which $u$ is univalued (i.e., $u(x)$ is a singleton for all $x \in \Omega$). Then $u_n$ converges to $u$ uniformly on $\Omega$: $u_n(x)$ is nonempty for all $x \in \Omega$ and all $n > N_\Omega$, and*

$$\sup_{x\in\Omega} \sup_{y\in u_n(x)} \|y - u(x)\| \to 0, \qquad n \to \infty.$$

*In particular, if $u$ is univalued throughout $\mathrm{int}(E)$ (so that $\phi \in C^1$ there), then uniform convergence holds for any compact $\Omega \subset \mathrm{int}(E)$.*

The proof of Proposition 1.7.11, given on page 28 of the supplement, follows two separate steps:

- if a sequence in the graph of $u_n$ converges, then the limit is in the graph of $u$;
- sequences in the graph of $u_n$ are bounded if the domain is bounded.

**Corollary 1.7.12 (Pointwise Convergence $\mu$-Almost Surely)** *If in addition $\mu$ is absolutely continuous, then $u_n(x) \to u(x)$ $\mu$-almost surely.*

*Proof.* We first claim that $E \subseteq \overline{\mathrm{dom}u}$. Indeed, for any $x \in E$ and any $\varepsilon > 0$, the ball $B = B_\varepsilon(x)$ has positive measure. Consequently, $u$ cannot be empty on the entire ball, because otherwise $\mu(B) = \pi(B \times \mathbb{R}^d)$ would be 0. Since dom$u$ is almost convex (see the discussion before Assumptions 1), this implies that actually $\mathrm{int}(\mathrm{conv}E) \subseteq \mathrm{dom}u$.

The rest is now easy: the set of points $x \in E$ for which $\Omega = \{x\}$ fails to satisfy the conditions of Proposition 1.7.11 is included in

$$(E \setminus E^{\text{den}}) \cup \{x \in \text{int}(\text{conv}(E)) : u(x) \text{ contains more than one point}\},$$

which is $\mu$-negligible because $\mu$ is absolutely continuous and both sets have Lebesgue measure 0.

## 1.8 Complementary Slackness and More General Cost Functions

It is well-known (Luenberger and Ye [89, Section 4.4]) that the solutions to the primal and dual problems are related to each other via *complementary slackness*. In other words, solution of one problem provides a lot of information about the solution of the other problem. Here, we show that this idea remains true for the Kantorovich primal and dual problems, extending the discussion in Sect. 1.6.1 to more general cost functions.

Let $\mathscr{X}$ and $\mathscr{Y}$ be complete separable metric spaces, $\mu \in P(\mathscr{X})$, $\nu \in P(\mathscr{Y})$, and $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}_+$ be a measurable cost function.

If one finds functions $(\varphi, \psi) \in \Phi_c$ and a transference plan $\pi \in \Pi(\mu, \nu)$ having the same objective values, then by weak duality $(\varphi, \psi)$ is optimal in $\Phi_c$ and $\pi$ is optimal in $\Pi(\mu, \nu)$. Having the same objective values is equivalent to

$$\int_{\mathscr{X} \times \mathscr{Y}} [c(x,y) - \varphi(x) - \psi(y)] \, \mathrm{d}\pi(x,y) = 0$$

which is in turn equivalent to

$$\varphi(x) + \psi(y) = c(x,y), \qquad \pi\text{-almost surely.}$$

It has already been established that there exists an optimal transference plan $\pi^*$. Assuming that $C(\pi^*) < \infty$ (otherwise all transference plans are optimal), a pair $(\varphi, \psi) \in \Phi_c$ is optimal if and only if

$$\varphi(x) + \psi(y) = c(x,y), \qquad \pi^*\text{-almost surely.}$$

Conversely, if $(\varphi_0, \psi_0)$ is an optimal pair, then $\pi$ is optimal if and only if it is concentrated on the set

$$\{(x,y) : \varphi_0(x) + \psi_0(y) = c(x,y)\}.$$

In particular, if for a given $x$ there exists a unique $y$ such that $\varphi_0(x) + \psi_0(y) = c(x,y)$, then the mass at $x$ must be sent entirely to $y$ and not be split; if this is the case for $\mu$-almost all $x$, then this relation defines $y$ as a function of $x$ and the resulting optimal $\pi$ is in fact induced from a transport map. This idea provides a criterion for solvability of the Monge problem (Villani [125, Theorem 5.30]).

### 1.8.1 Unconstrained Dual Kantorovich Problem

It turns out that the dual Kantorovich problem can be recast as an unconstrained optimisation problem of only one function $\varphi$. The new formulation is not only conceptually simpler than the original one, but also sheds light on the properties of the optimal dual variables. Since the dual objective function to be maximised,

$$\int_{\mathscr{X}} \varphi \, \mathrm{d}\mu + \int_{\mathscr{Y}} \psi \, \mathrm{d}\nu,$$

is increasing in $\varphi$ and $\psi$, one should seek functions that take values as large as possible subject to the constraint $\varphi(x) + \psi(y) \leq c(x,y)$. Suppose that an oracle tells us that some $\varphi \in L_1(\mu)$ is a good candidate. Then the largest possible $\psi$ satisfying $(\varphi, \psi) \in \Phi_c$ is defined as

$$\psi(y) = \inf_{x \in \mathscr{X}} [c(x,y) - \varphi(x)] := \varphi^c(y).$$

A function taking this form is called *c-concave* [124, Chapter 2]; we say that $\psi$ is the *c-transform* of $\varphi$. It is not necessarily true that $\varphi^c$ is integrable or even measurable, but if we neglect this difficulty, then it is obvious that

$$\sup_{\psi \in L_1(\nu) : (\varphi,\psi) \in \Phi_c} \left[ \int_{\mathscr{X}} \varphi \, \mathrm{d}\mu + \int_{\mathscr{Y}} \psi \, \mathrm{d}\nu \right] = \int_{\mathscr{X}} \varphi \, \mathrm{d}\mu + \int_{\mathscr{Y}} \varphi^c \, \mathrm{d}\nu.$$

The dual problem can thus be formulated as the unconstrained problem

$$\sup_{\varphi \in L_1(\mu)} \left[ \int_{\mathscr{X}} \varphi \, \mathrm{d}\mu + \int_{\mathscr{Y}} \varphi^c \, \mathrm{d}\nu \right].$$

One can apply this $c$-transform again and replace $\varphi$ by

$$\varphi^{cc}(x) = (\varphi^c)^c(x) = \inf_{y \in \mathscr{Y}} [c(x,y) - \varphi^c(y)] \geq \varphi(x),$$

so that $\varphi^{cc}$ has a better objective value yet still $(\varphi^{cc}, \varphi^c) \in \Phi_c$ (modulo measurability issues). An elementary calculation shows that in general $\varphi^{ccc} = \varphi^c$. Thus, for any function $\varphi_1$, the pair of functions $(\varphi, \psi) = (\varphi_1^{cc}, \varphi_1^c)$ has a better objective value than $(\varphi_1, \psi_1)$, and satisfies $(\varphi, \psi) \in \Phi_c$. Moreover, $\varphi^c = \psi$ and $\psi^c = \varphi$; in words, $\varphi$ and $\psi$ are *c-conjugate*. An optimal dual pair $(\varphi, \psi)$ can be expected to be $c$-conjugate; this is indeed true almost surely:

**Proposition 1.8.1 (Existence of an Optimal Pair)** *Let $\mu$ and $\nu$ be probability measures on $\mathscr{X}$ and $\mathscr{Y}$ such that the independent coupling with respect to the nonnegative and lower semicontinuous cost function is finite: $\int_{\mathscr{X} \times \mathscr{Y}} c(x,y) \, \mathrm{d}\mu(x) \mathrm{d}\nu(y) < \infty$. Then there exists an optimal pair $(\varphi, \psi)$ for the dual Kantorovich problem. Furthermore, the pair can be chosen in a way that $\mu$-almost surely, $\varphi = \psi^c$ and $\nu$-almost surely, $\psi = \varphi^c$.*

Proposition 1.8.1 is established (under weaker conditions) by Ambrosio and Pratelli [11, Theorem 3.2]. It is clear from the discussion above that once existence of an

optimal pair $(\varphi_1, \psi_1)$ is established, the pair $(\varphi, \psi) = (\varphi_1^{cc}, \varphi_1^c)$ should be optimal. Combining Proposition 1.8.1 with the preceding subsection, we see that if $\varphi$ is optimal (for the unconstrained dual problem), then any optimal transference plan $\pi^*$ must be concentrated on the set

$$\{(x,y) : \varphi(x) + \varphi^c(y) = c(x,y)\}.$$

If for $\mu$-almost every $x$ this equation defines $y$ uniquely as a (measurable) function of $x$, then $\pi^*$ is induced by a transport map. Indeed, we have seen how this is the case, in the quadratic case $c(x,y) = \|x-y\|^2/2$, when $\mu$ is absolutely continuous. An extension to $p > 1$ (instead of $p = 2$) is sketched in Sect. 1.8.3.

We remark that at the level of generality of Proposition 1.8.1, the function $\varphi^c$ may fail to be Borel measurable; Ambrosio and Pratelli show that this pair can be modified up to null sets in order to be Borel measurable. If $c$ is continuous, however, then $\varphi^c$ is an infimum of a collection of continuous functions (in $y$). Hence $-\varphi^c$ is lower semicontinuous, which yields that $\varphi^c$ is measurable. When $c$ is *uniformly* continuous, measurability of $\varphi^c$ is established in a more lucid way, as exemplified in the next subsection.

### *1.8.2 The Kantorovich–Rubinstein Theorem*

Whether $\varphi^c(y)$ is tractable to evaluate depends on the structure of $c$. We have seen an example where $c$ was the quadratic Euclidean distance. Here, we shall consider another useful case, where $c$ is a metric. Assume that $\mathscr{X} = \mathscr{Y}$, denote their metric by $d$, and let $c(x,y) = d(x,y)$. If $\varphi = \psi^c$ is $c$-concave, then it is 1-Lipschitz. Indeed, by definition and the triangle inequality

$$\varphi(z) = \inf_{y\in\mathscr{Y}} [d(z,y) - \psi(y)] \le \inf_{y\in\mathscr{Y}} [d(x,y) + d(x,z) - \psi(y)] = \varphi(x) + d(x,z).$$

Interchanging $x$ and $z$ yields $|\varphi(x) - \varphi(z)| \le d(x,z)$.[3]

Next, we claim that if $\varphi$ is Lipschitz, then $\varphi^c(y) = -\varphi(y)$. Indeed, choosing $x = y$ in the infimum shows that $\varphi^c(y) \le d(y,y) - \varphi(y) = -\varphi(y)$. But the Lipschitz condition on $\varphi$ implies that for all $x$, $d(x,y) - \varphi(x) \ge -\varphi(y)$. In view of that, we can take in the dual problem $\varphi$ to be Lipschitz and $\psi = -\varphi$, and the duality formula (Theorem 1.4.2) takes the form

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}^2} d(x,y)\,\mathrm{d}\pi(x,y) = \sup_{\|\varphi\|_{\mathrm{Lip}}\le 1} \left| \int_{\mathscr{X}} \varphi\,\mathrm{d}\mu - \int_{\mathscr{X}} \varphi\,\mathrm{d}\nu \right|,$$

$$\|\varphi\|_{\mathrm{Lip}} = \sup_{x\ne y} \frac{|\varphi(x) - \varphi(y)|}{d(x,y)}. \tag{1.11}$$

[3] In general, $\psi^c$ inherits the modulus of continuity of $c$, see Santambrogio [119, page 11].

This is known as the *Kantorovich–Rubinstein theorem* [124, Theorem 1.14]. (We have been a bit sloppy because $\varphi$ may not be integrable. But if for some $x_0 \in \mathscr{X}$, $x \mapsto d(x,x_0)$ is in $L_1(\mu)$, then any Lipschitz function is $\mu$-integrable. Otherwise one needs to restrict the supremum to, e.g., bounded Lipschitz $\varphi$.)

### 1.8.3 Strictly Convex Cost Functions on Euclidean Spaces

We now return to the Euclidean case $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$ and explore the structure of $c$-transforms. When $c$ is different than $\|x-y\|^2/2$, we can no longer "open up the square" and relate the Monge–Kantorovich problem to convexity. However, we can still apply the idea that $\varphi(x) + \varphi^c(y) = c(x,y)$ if and only if the infimum is attained at $x$. Indeed, recall that

$$\varphi^c(y) = \inf_{x \in \mathscr{X}} [c(x,y) - \varphi(x)],$$

so that $\varphi(x) + \varphi^c(y) = c(x,y)$ if and only if

$$\varphi(z) - \varphi(x) \le c(z,y) - c(x,y), \qquad z \in \mathscr{X}.$$

Notice the similarity to the subgradient inequality for convex functions, with the sign being reversed. In analogy, we call the collection of $y$'s satisfying the above in equality the *$c$-superdifferential* of $\varphi$ at $x$, and we denote it by $\partial^c \varphi(x)$. Of course, if $c(x,y) = \|x-y\|^2/2$, then $y \in \partial^c(x)$ if and only if $y$ is a subgradient of $(\|\cdot\|^2/2 - \varphi)$ at $x$.

The following result generalises Theorem 1.6.2 to other powers $p > 1$ of the Euclidean norm. These cost functions define the Wasserstein distances of the next chapter.

**Theorem 1.8.2 (Strictly Convex Costs on $\mathbb{R}^d$)** *Let $c(x,y) = h(x-y)$ with $h(v) = \|v\|^p/p$ for some $p > 1$ and let $\mu$ and $\nu$ be probability measures on $\mathbb{R}^d$ with finite $p$-th moments such that $\mu$ is absolutely continuous with respect to Lebesgue measure. Then the solution to the Kantorovich problem with cost function $c$ is unique and induced from a transport map $T$. Furthermore, there exists an optimal pair $(\varphi, \varphi^c)$ of the dual problem, with $\varphi$ $c$-concave. The solutions are related by*

$$T(x) = x - \nabla\varphi(x)\|\nabla\varphi(x)\|^{1/(p-1)-1} \qquad (\mu\text{-almost surely}).$$

*Proof (Assuming $\nu$ has Compact Support).* The existence of the optimal pair $(\varphi, \varphi^c)$ with the desired properties follows from Proposition 1.8.1 (they are Borel measurable because $c$ is continuous). We shall now show that $\varphi$ has a unique $c$-supergradient $\mu$-almost surely.

**Step 1: $\varphi$ is $c$-superdifferentiable.** Let $\pi^*$ be an optimal coupling. By duality arguments, $\pi$ is concentrated on the set of $(x,y)$ such that $y \in \partial^c \varphi(x)$. Consequently, for $\mu$-almost any $x$, the $c$-superdifferential of $\varphi$ at $x$ is nonempty.

**Step 2: $\varphi$ is differentiable.** Here, we impose the additional condition that $\nu$ is compactly supported. Then $\varphi$ can be taken as a $c$-transform on the compact support of $\nu$. Since $h$ is locally Lipschitz (it is $C^1$ because $p > 1$) this implies that $\varphi$ is locally Lipschitz. Hence, it is differentiable Lebesgue almost surely, and consequently $\mu$-almost surely.

**Step 3: Conclusion.** For $\mu$-almost every $x$ there exists $y \in \partial^c \varphi(x)$ and a gradient $u = \nabla\varphi(x)$. In particular, $u$ is a subgradient of $\varphi$:

$$\varphi(z) - \varphi(x) \geq \langle u, z - x\rangle + o(\|z - x\|).$$

Here and more generally, $o(\|z-x\|)$ denotes a function $r(z)$ (defined in a neighbourhood of $x$) such that $r(z)/\|z-x\| \to 0$ as $z \to x$. (If $\varphi$ were convex, then we could take $r \equiv 0$, so the definition for convex functions is equivalent, and then the inequality holds globally and not only locally.) But $y \in \partial^c \varphi(x)$ means that as $z \to x$,

$$h(z-y) - h(x-y) = c(z,y) - c(x,y) \geq \varphi(z) - \varphi(x) \geq \langle u, z-x\rangle + o(\|z-x\|).$$

In other words, $u$ is a subgradient of $h$ at $x - y$. But $h$ is differentiable with gradient $\nabla h(v) = v\|v\|^{p-2}$ (zero if $v = 0$). We obtain $\nabla\varphi(x) = u = \nabla h(x-y)$ and since the gradient of $h$ is invertible, we conclude

$$y = T(x) := x - (\nabla h)^{-1}[\nabla\varphi(x)],$$

which defines $y$ as a (measurable) function of $x$.[4] Hence, the optimal transference plan $\pi$ is unique and induced from the transport map $T$.

The general result, without assuming compact support for $\nu$, can be found in Gangbo and McCann [59]. It holds for a larger class of functions $h$, those that are strictly convex on $\mathbb{R}^d$ (this yields that $\nabla h$ is invertible), have superlinear growth $(h(v)/\|v\| \to \infty$ as $v \to \infty)$ and satisfying a technical geometric condition (which $\|v\|^p/p$ does when $p > 1$). Furthermore, if $h$ is sufficiently smooth, namely $h \in C^{1,1}$ locally (it is if $p \geq 2$, but not if $p \in (1,2)$), then $\mu$ does not need to be absolutely continuous; it suffices that it not give positive measure to any set of Hausdorff dimension smaller or equal than $d-1$. When $d = 1$ this means that Theorem 1.8.2 is still valid as long as $\mu$ has no atoms ($\mu(\{x\}) = 0$ for all $x \in \mathbb{R}$), which is a weaker condition than $\mu$ being absolutely continuous.

It is also noteworthy that for strictly concave cost functions (e.g., $p \in (0,1)$), the situation is similar *when the supports of $\mu$ and $\nu$ are disjoint*. The reason is that $h$ may fail to be differentiable at 0, but it only needs to be differentiated at $x - y$ with $x \in \text{supp}\mu$ and $y \in \text{supp}\nu$. If the supports are not disjoint, then one needs to leave all common mass in place until the supports become disjoint (Villani [124, Chapter 2]) and then the result of [59] applies. As a simple example, let $\mu$ be uniform on $[0,1]$ and $\nu$ be uniform on $[0,2]$. After leaving common mass in place, we are left with uniforms on $[0,1]$ and $[1,2]$ (with total mass $1/2$) with essentially disjoint supports,

[4] Gradients of Borel functions are measurable, as the limit can be taken on a countable set. The inverse $(\nabla h)^{-1}$ equals the gradient of the Legendre transform $h^*$ and is therefore Borel measurable.

for which the optimal transport map is the *decreasing* map $T(x) = 2 - x$. Thus, the unique optimal $\pi$ is not induced from a map, but rather from an equal weight mixture of $T$ and the identity. Informally, each point $x$ in the support of $\mu$ needs to be split; half stays and $x$ and the other half transported to $2 - x$. The optimal coupling from $\nu$ to $\mu$ is unique and induced from the map $S(x) = x$ if $x \leq 1$ and $2 - x$ if $x \geq 1$, which is neither increasing nor decreasing.

## 1.9 Bibliographical Notes

Many authors, including Villani [124, Theorem 1.3]; [125, Theorem 5.10], give the duality Theorem 1.4.2 for lower semicontinuous cost functions. The version given here is a simplification of Beiglböck and Schachermayer [17, Theorem 1]. The duality holds for functions that take values in $[-\infty, \infty]$ provided that they are finite on a sufficiently large subset of $\mathscr{X} \times \mathscr{Y}$, but there are simple counterexamples if $c$ is infinite too often [17, Example 4.1]. For results outside the Polish space setup, see Kellerer [80] and Rachev and Rüschendorf [107, Chapter 4].

Theorem 1.5.1 for the one-dimensional case is taken from [124], where it is proven using the general duality theorem. For direct proofs and the history of this result, one may consult Rachev [106] or Rachev and Rüschendorf [107, Section 3.1]. The concave case is carefully treated by McCann [94].

The results in the Gaussian case were obtained independently by Olkin and Pukelsheim [98] and Givens and Shortt [65]. The proof given here is from Bhatia [20, Exercise 1.2.13]. An extension to separable Hilbert spaces can be found in Gelbrich [62] or Cuesta-Albertos et al. [39].

The regularity theory of Sect. 1.6.4 is very delicate. Caffarelli [32] showed the first part of Theorem 1.6.7; the proof can also be found in Figalli's book [52, Theorem 4.23]. Villani [124, Theorem 4.14] states the result without proof and refers to Alesker et al. [7] for a sketch of the second part of Theorem 1.6.7. Other regularity results exist, Villani [125, Chapter 12]; Santambrogio [119, Section 1.7.6]; Figalli [52].

Cuesta-Albertos et al. [40, Theorem 3.2] prove Theorem 1.7.2 for the quadratic case; the form given here is from Schachermayer and Teichmann [120, Theorem 3].

The definition of cyclical monotonicity depends on the cost function. It is typically referred to as $c$-cyclical monotonicity, with "cyclical monotonicity" reserved to the special case of quadratic cost. Since we focus on the quadratic case and for readability, we slightly deviate from the standard jargon. That cyclical monotonicity implies optimality (Proposition 1.7.5) was shown independently by Pratelli [105] (finite lower semicontinuous cost) and Schachermayer and Teichmann [120] (possibly infinite continuous cost). A joint generalisation is given by Beiglböck et al. [18].

Section 1.7.2 is taken from Zemel and Panaretos [134, Section 7.5]; a slightly weaker version was shown independently by Chernozhukov et al. [35]. Heinich and Lootgieter [68] establish almost sure pointwise convergence. If $\mu_n = \mu$, then the optimal maps converge in $\mu$-measure [125, Corollary 5.23] in a very general setup,

but there are simple examples where this fails if $\mu_n \neq \mu$ [125, Remark 5.25]. In the quadratic case, further stability results of a weaker flavour (focussing on the convex potential $\phi$, rather than its derivative, which is the optimal map) can be found in del Barrio and Loubes [42].

The idea of using the $c$-transform (Sect. 1.8) is from Rüschendorf [116].