2

Discrete Transport Problems

In this chapter we introduce the basic ideas of Kantorovich's approach to transport problems by working in the discrete setting. From a formal viewpoint, this is just a particular case of the theory developed in Chapter 3, so its discussion in a separate chapter is mainly motivated by pedagogical reasons. The notions of transport plan (Section 2.1) and c-cyclical monotonicity (Section 2.2) arise effortlessly in this context and without any real technical burden. We then move to consider c-cyclical monotonicity in the model case of the quadratic transport cost $c(x,y) = |x-y|^2$, thus establishing its link with convexity and the corresponding Kantorovich duality (Sections 2.3 and 2.4). Finally, in Section 2.5, we consider the discrete Monge problem.

2.1 The Discrete Kantorovich Problem

Discrete transport problems involve origin and final mass distributions μ and ν that are concentrated at finitely many points; that is to say, we consider

$$\mu = \sum_{i=1}^{N} \mu_i\, \delta_{x_i}, \qquad \nu = \sum_{j=1}^{M} \nu_j\, \delta_{y_j}, \tag{2.1}$$

where $X = \{x_i\}_{i=1}^N$ and $Y = \{y_j\}_{j=1}^M$ are collections of *distinct*[1] points in $\mathbb{R}^n$, and with μ_i and ν_j *positive numbers* such that

$$\mu(\mathbb{R}^n) = \sum_{i=1}^{N} \mu_i = 1, \qquad \nu(\mathbb{R}^n) = \sum_{j=1}^{M} \nu_j = 1. \tag{2.2}$$

The corresponding Monge problem $\mathbf{M}_c(\mu,\nu)$ may be ill-posed (independently from the choice of a transport cost c) for the basic reason that there may be no transport maps from μ to ν.

[1] By this we mean that X is a family of N distinct points in $\mathbb{R}^n$ and that Y is a family of M distinct points in $\mathbb{R}^n$, although we are not requiring $X \cap Y$ to be empty.

Remark 2.1 (Nonexistence of transport maps) Consider (2.1) and (2.2) with $N = 1$ (hence, $\mu_1 = 1$) and $M = 2$. Since μ is concentrated at x_1, every $\mathbb{R}^n$-valued map T defined at x_1 transports μ with $T_\# \mu = \delta_{T(x_1)}$. Thus, $T_\# \mu = \nu = \nu_1 \, \delta_{y_1} + \nu_2 \, \delta_{y_2}$ cannot hold, as both ν_1 and ν_2 are positive and $y_1 \neq y_2$. Hence, $\mathbf{M}_c(\mu, \nu) = +\infty$, with empty competition class.

In the situation of Remark 2.1, one would like to transport a ν_1-amount of the mass sitting at x_1 to y_1 and a $1 - \nu_1 = \nu_2$-amount to y_2: the only problem is that these simple "mass splitting instructions" cannot be described by maps. However, they can be efficiently and naturally described by using matrices,

$$\gamma = \{\gamma_{ij}\} \in \mathbb{R}^{N \times M},$$

so that $\gamma_{ij} \in [0,1]$ is the amount of mass sitting at x_i to be transported to y_j. Since the partial sum $\sum_{i=1}^N \gamma_{ij}$ represents the total mass received at the site y_j, and the partial sum $\sum_{j=1}^M \gamma_{ij}$ represents the total mass shipped from the site x_i, the conditions for γ to represent transport instructions from μ to ν are that

$$\mu_i = \sum_{j=1}^{M} \gamma_{ij}, \qquad \nu_j = \sum_{i=1}^{N} \gamma_{ij}. \tag{2.3}$$

Any $\gamma \in \mathbb{R}^{N \times M}$ with nonnegative entries and satisfying (2.3) for μ and ν as in (2.1) and (2.2) is called a **discrete transport plan** from μ to ν. The set of all transport plans from μ to ν is the convex[2] set $\Gamma \subset \mathbb{R}^{N \times M}$ defined by

$$\Gamma(\mu, \nu) = \Big\{ \gamma \in \mathbb{R}^{N \times M} : \gamma_{ij} \geq 0, \, \mu_i = \textstyle\sum_{k=1}^M \gamma_{ik}, \, \nu_j = \sum_{k=1}^N \gamma_j \text{ for every } i, j \Big\}. \tag{2.4}$$

Given a cost function $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, the total cost associated to the discrete transport plan γ is then given by

$$\mathrm{Cost}(\gamma) = \sum_{i,j} c(x_i, y_j) \, \gamma_{ij}, \tag{2.5}$$

and we obtain **the discrete Kantorovich problem**,

$$\mathbf{K}_c(\mu, \nu) = \inf \left\{ \sum_{i,j} c(x_i, y_j) \, \gamma_{ij} : \gamma \in \Gamma(\mu, \nu) \right\}. \tag{2.6}$$

In sharp contrast with the case of the Monge problem (see Chapter 1 and Remark 2.1), establishing the existence of minimizers in (2.6) is actually *trivial*. A minimizer γ in $\mathbf{K}_c(\mu, \nu)$ is called an **optimal discrete transport plan**.

Theorem 2.2 (Optimal discrete transport plans) *If $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ are discrete (i.e., if* (2.1) *and* (2.2) *hold), then for every function $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ there is a*

[2] A set $X \subset \mathbb{R}^n$ is convex if $t\, x + (1 - t)\, y \in X$ whenever $x, y \in X$ and $t \in (0, 1)$.

minimizer of the discrete Kantorovich problem $\mathbf{K}_c(\mu,\nu)$. *Moreover,* $\mathbf{M}_c(\mu,\nu) \ge \mathbf{K}_c(\mu,\nu)$.

Proof The function $\gamma \mapsto \mathrm{Cost}(\gamma)$ is linear on $\mathbb{R}^{N\times M}$, while $\Gamma(\mu,\nu)$ is a non-empty ($\gamma_{ij} = \mu_i\,\nu_j$ always belongs to $\Gamma(\mu,\nu)$), convex, compact set in $\mathbb{R}^{N\times M}$ (the constraints defining $\Gamma(\mu,\nu)$ are clearly convex and closed, and $\Gamma(\mu,\nu)$ is bounded since $\gamma \in \Gamma(\mu,\nu)$ implies $0 \le \gamma_{ij} \le 1$ for every i,j). Therefore, the existence of a minimizer of $\mathbf{K}_c(\mu,\nu)$ is trivially established by the Direct Method.

The inequality $\mathbf{M}_c(\mu,\nu) \ge \mathbf{K}_c(\mu,\nu)$ is trivial if there are no transport maps. Now, if T transports μ into ν, then $T : \{x_i\}_{i=1}^N \to \mathbb{R}^n$ is such that $T_\#\mu = \sum_{i=1}^N \mu_i\,\delta_{T(x_i)}$ equals $\nu = \sum_{j=1}^M \nu_j\,\delta_{y_j}$. This means, necessarily, that $N \ge M$ and that there is $\sigma : \{1,\dots,N\} \to \{1,\dots,M\}$ surjective such that, for every i and j, respectively,

$$T(x_i) = y_{\sigma(i)}, \qquad \nu_j = \sum_{\{i:\sigma(i)=j\}} \mu_i.$$

Correspondingly, the plan[3]

$$\gamma_{ij} = \mu_i\,\delta_{\sigma(i),j}$$

is such that $\gamma \in \Gamma(\mu,\nu)$, with

$$\begin{aligned}\mathrm{Cost}(\gamma) &= \sum_{i=1}^N \sum_{j=1}^M c(x_i,y_j)\,\mu_i\,\delta_{\sigma(i),j} \qquad (2.7)\\ &= \sum_{i=1}^N c(x_i,T(x_i))\,\mu_i = \int_{\mathbb{R}^n} c(x,T(x))\,d\mu(x).\end{aligned}$$

Hence, $\mathbf{K}_c(\mu,\nu) \le \mathrm{Cost}(\gamma) \le \int_{\mathbb{R}^n} c(x,T(x))\,d\mu(x)$ whenever T transports μ into ν, and $\mathbf{K}_c(\mu,\nu) \le \mathbf{M}_c(\mu,\nu)$ follows by arbitrariness of T. □

Two basic features of discrete Kantorovich problems are illustrated in the following remarks.

Remark 2.3 ($\Gamma(\mu,\nu)$ contains open segments if $N,M \ge 2$.) Let us notice, first of all, that $\Gamma(\mu,\nu)$ consists of a single element of $\mathbb{R}^{M\times N}$ unless $N \ge 2$ and $M \ge 2$. This is obvious if $N = M = 1$. If $N = 1$, $M \ge 1$, then (2.3) implies $\gamma_{1j} = \nu_j$ for every j, and similarly if $N \ge 2$, $M = 1$, then (2.3) gives $\gamma_{i1} = \mu_i$ for every i. This said, as soon as $N,M \ge 2$, for every $\gamma \in \Gamma(\mu,\nu)$ such that $\gamma_{ij} > 0$ for every i,j (one such element is always given by $\gamma_{ij} = \mu_i\,\nu_j$), there exists an open segment centered at γ which is entirely contained in $\Gamma(\mu,\nu)$. Indeed, if we define $\gamma^t = \{\gamma^t_{ij}\}$ by

[3] Here $\delta_{h,k}$ is the Kronecker symbol of h and k, not to be confused with δ_x, the Dirac mass concentrated at $x \in \mathbb{R}^n$.

$$\begin{aligned} \gamma^t_{11} &= \gamma_{11} + t, & \gamma^t_{12} &= \gamma_{12} - t, \\ \gamma^t_{21} &= \gamma_{21} - t, & \gamma^t_{21} &= \gamma_{21} + t, \end{aligned} \qquad \gamma^t_{ij} = \gamma_{ij} \text{ if either } i \geq 3 \text{ or } j \geq 3,$$

then $\gamma^t \in \Gamma(\mu, \nu)$ for every sufficiently small value of $|t|$.

Remark 2.4 (Nonuniqueness of minimizers) $\mathbf{K}_c(\mu, \nu)$ may possess multiple minimizers. This is always the case, for example, if c, X and Y are such that,[4] for some $\lambda > 0$, $c(x_i, y_j) = \lambda$ for every i, j. Indeed, in that case,

$$\mathrm{Cost}(\gamma) = \sum_{i,j} c(x_i, y_j)\, \gamma_{ij} = \lambda \sum_{i=1}^{N} \sum_{j=1}^{M} \gamma_{ij} = \lambda \sum_{i=1}^{N} \mu_i = \lambda,$$

i.e., cost is constant on $\Gamma(\mu, \nu)$, so every discrete transport plan is optimal.

2.2 c-Cyclical Monotonicity with Discrete Measures

We now further develop the remark made in Remark 2.3 to obtain a necessary optimality condition for minimizers γ in problem $\mathbf{K}_c(\mu, \nu)$. We start by noticing that we could well have $\gamma_{ij} = 0$ for some pair of indexes (i, j): when this happens, it means that the optimal plan γ has no convenience in sending any of the mass stored at x_i to the destination y_j. We thus look at those pairs (i, j) such that $\gamma_{ij} > 0$ and consider the set

$$S(\gamma) = \{(x_i, y_j) \in \mathbb{R}^n \times \mathbb{R}^n : \gamma_{ij} > 0\} \tag{2.8}$$

of those pairs of locations in the supports of μ and ν that are exchanging mass under the plan γ. We now formulate a necessary condition for the optimality of γ in terms of a geometric property of $S(\gamma)$.

Theorem 2.5 *If $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ are discrete (i.e., if* (2.1) *and* (2.2) *hold) and γ is a minimizer in the discrete Kantorovich problem $\mathbf{K}_c(\mu, \nu)$, then for every finite subset $\{(z_\ell, w_\ell)\}_{\ell=1}^L$ of $S(\gamma)$ we have*

$$\sum_{\ell=1}^{L} c(z_\ell, w_\ell) \leq \sum_{\ell=1}^{L} c(z_{\ell+1}, w_\ell), \tag{2.9}$$

where $z_{L+1} = z_1$.

Remark 2.6 Based on Theorem 2.5, we introduce the following crucial notion, to be discussed at length in the sequel: a set $S \subset \mathbb{R}^n \times \mathbb{R}^n$ is **c-cyclically monotone** if every finite subset $\{(z_\ell, w_\ell)\}_{\ell=1}^L$ of S satisfies (2.9).

[4] This situation can of course be achieved in many different ways: for example, we could work in $\mathbb{R}^2$, with $x_1 = (0, 0)$, $x_2 = (1, 1)$, $y_1 = (1, 0)$, $y_2 = (0, 1)$, and with $c(x, y)$ being any nonnegative function of the Euclidean distance $|x - y|$; see Figure 2.5.

Proof of Theorem 2.5 By construction, $z_\ell = x_{i(\ell)}$ and $w_\ell = y_{j(\ell)}$ for suitable functions $i : \{1,\dots,L\} \to \{1,\dots,N\}$ and $j : \{1,\dots,L\} \to \{1,\dots,M\}$, and $\alpha_\ell = \gamma_{i(\ell)j(\ell)} > 0$ for every ℓ. Given $\varepsilon > 0$ with $\varepsilon < \min_\ell \alpha_\ell$, we construct a family of transport plans γ^ε by making, first of all, the following changes:

$$\begin{aligned} \gamma_{i(1)j(1)} &\quad\to\quad \gamma^\varepsilon_{i(1)j(1)} = \gamma_{i(1)j(1)} - \varepsilon, \\ \gamma_{i(2)j(2)} &\quad\to\quad \gamma^\varepsilon_{i(2)j(2)} = \gamma_{i(2)j(2)} - \varepsilon, \\ \dots \\ \gamma_{i(L)j(L)} &\quad\to\quad \gamma^\varepsilon_{i(L)j(L)} = \gamma_{i(L)j(L)} - \varepsilon. \end{aligned}$$

i.e., we decrease by ε the amount of mass sent by γ from $z_\ell = x_{i(\ell)}$ to $w_\ell = y_{j(\ell)}$. Without further changes, the resulting plan γ^ε is not admissible: indeed, we have left unused an ε of mass at each origin site z_ℓ, while each of the destination sites w_ℓ is missing an ε of mass. To fix things, we transport the excess mass ε sitting at $z_{\ell+1}$ to w_ℓ and thus prescribe the following changes:

$$\begin{aligned} \gamma_{i(2)j(1)} &\quad\to\quad \gamma^\varepsilon_{i(2)j(1)} = \gamma_{i(2)j(1)} + \varepsilon, \\ \gamma_{i(3)j(2)} &\quad\to\quad \gamma^\varepsilon_{i(3)j(2)} = \gamma_{i(3)j(2)} + \varepsilon, \\ \dots \\ \gamma_{i(L+1)j(L)} &\quad\to\quad \gamma^\varepsilon_{i(L+1)j(L)} = \gamma_{i(1)j(L)} + \varepsilon, \end{aligned}$$

where $i(L+1) = 1$. Notice that, since $\varepsilon > 0$ by assumption, we do not need $\gamma_{i(\ell+1)j(\ell)}$ to be positive to prescribe the second round of changes. Finally, by setting $\gamma^\varepsilon_{ij} = \gamma_{ij}$ for every $(i,j) \notin \{(i(\ell), j(\ell)) : 1 \le \ell \le L\}$, we find that $\gamma^\varepsilon \in \Gamma(\mu,\nu)$ and therefore that

$$0 \le \mathrm{Cost}(\gamma^\varepsilon) - \mathrm{Cost}(\gamma) = \sum_{\ell=1}^{L} -\varepsilon\, c(z_\ell, w_\ell) + \varepsilon\, c(z_{\ell+1}, w_\ell).$$

Given that $\varepsilon > 0$, we deduce the validity of (2.9). □

When minimizing (as done in $\mathbf{K}_c(\mu,\nu)$) a linear function f on a compact convex set K, if f is nonconstant on K, then any minimum point x_0 will necessarily lie on ∂K. In particular, given a unit vector τ with $x_0 + t\,\tau \in K$ for every sufficiently small and positive t, by differentiating in t the inequality $f(x_0 + t\,\tau) \ge f(x_0)$, we find that $\nabla f(x_0) \cdot \tau \ge 0$. The family of inequalities $\nabla f(x_0) \cdot \tau \ge 0$ indexed over all the admissible directions τ is then a necessary and sufficient condition for x_0 to be a minimum point of f on K.

From this viewpoint, in Theorem 2.5 we have identified a family of “directions τ” that can be used to take admissible one-sided variations of an optimal transport plan γ. Understanding if c-cyclical monotonicity is not only a necessary condition for minimality but also a *sufficient* one is tantamount to prove that such one-sided variations exhaust all the admissible ones. The Kantorovich duality theorem (Theorem 3.13) provides an elegant way to prove that this is indeed the case – i.e., that c-cyclical monotonicity fully characterizes optimality in $\mathbf{K}_c(\mu, \nu)$. The main difficulties related to the Kantorovich duality theorem are not of a technical character – actually the theorem is deduced by somehow elementary considerations – but rather conceptual – for example, if one insists (as it would seem natural when working with transport problems) in having a clear geometric understanding of things. Indeed, while the combinatorial nature of c-cyclical monotonicity is evident, its geometric content is definitely less immediate.

Luckily, in the case of the quadratic cost $c(x, y) = |x - y|^2$, c-cyclical monotonicity is immediately related to convexity. By examining this relation in detail, and by moving in analogy with it in the case of general costs, we will develop geometric and analytical ways to approach c-cyclical monotonicity, as well as develop the Kantorovich duality theory. To explain the relation with convexity, it is sufficient to look at (2.9), with $c(x, y) = |x - y|^2$, to expand the squares $|z_\ell - w_\ell|^2$ and $|z_{\ell+1} - w_\ell|^2$, to cancel out the sums over ℓ of $|z_\ell|^2$, $|w_\ell|^2$ and $|z_{\ell+1}|^2$ (as we can thanks to $z_{L+1} = z_1$), and, finally, to obtain the equivalent condition

$$\sum_{\ell=1}^{L} w_\ell \cdot (z_{\ell+1} - z_\ell) \le 0, \qquad \text{for all } \{(z_\ell, w_\ell)\}_{\ell=1}^{L} \subset S. \tag{2.10}$$

This condition is (very well) known in convex geometry as **cyclical monotonicity** (of S). As proved in the next section, (2.10) is equivalent to require that S lies in the graph of the gradient of a convex function on $\mathbb{R}^n$. We can quickly anticipate this result by looking at the simple case when $n = 1$ and $L = 2$, and (2.10) just says

$$0 \le w_2(z_2 - z_1) - w_1(z_2 - z_1) = (w_2 - w_1)(z_2 - z_1),$$

that is,

$$\begin{cases} (z_1, w_1), (z_2, w_2) \in S, \\ z_1 \le z_2, \end{cases} \qquad \Rightarrow \qquad w_1 \le w_2. \tag{2.11}$$

The geometric meaning of (2.11) is absolutely clear (see Figure 2.1): S must be contained in the extended graph of a monotone increasing function from $\mathbb{R}$ to $\mathbb{R}$ (where the term "extended" indicates that vertical segments corresponding to jump points are included in the graph). Since monotone functions are the gradients of convex functions, the connection between convexity and OMT problems with quadratic transport cost is drawn.

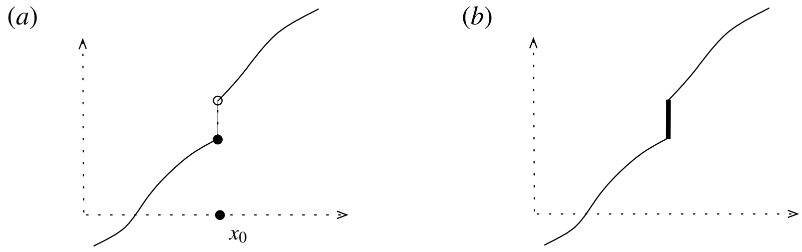

Figure 2.1 (a) The graph of an increasing function $f : \mathbb{R} \to \mathbb{R}$ with a discontinuity at a point x_0. The black dot indicates that the function takes the lowest possible value compatible with being increasing; (b) The extended graph of f is a subset of $\mathbb{R}^2$ which contains the graph of f and the whole vertical segment of values that f may take at x_0 without ceasing to be increasing. The property of being contained into the extended graph of an increasing function is easily seen to be equivalent to (2.11).

2.3 Basics about Convex Functions on $\mathbb{R}^n$

We now review some key concepts concerning convex functions on $\mathbb{R}^n$ that play a central role in our discussion. In OMT it is both natural and convenient to consider convex functions taking values in $\mathbb{R} \cup \{+\infty\}$. Since this setting may be unfamiliar to some readers, we offer here a review of the main results, including proofs of the less obvious ones.

Convex sets: A **convex set** in $\mathbb{R}^n$ is a set $K \subset \mathbb{R}^n$ such that $t\,x + (1-t)\,y \in K$ whenever $t \in (0,1)$ and $x, y \in K$. If $K \neq \emptyset$, the **(affine) dimension of** K is defined as the dimension of the smallest affine space containing K. The **relative interior** $\mathrm{Ri}(K)$ of a convex set K is its interior as a subset of the smallest affine space containing it; of course $\mathrm{Ri}(K) = \mathrm{Int}(K)$, where $\mathrm{Int}(K)$ is the set of interior points of K as a subset of $\mathbb{R}^n$, whenever K has dimension n. Given $E \subset \mathbb{R}^n$ we say that $z \in \mathbb{R}^n$ is a **convex combination in** E if $z = \sum_{i=1}^N t_i\, x_i$ for some coefficients $t_i \in [0,1]$ such that $\sum_{i=1}^N t_i = 1$ and some $\{x_i\}_{i=1}^N \subset E$. The **convex envelope** $\mathrm{conv}(E)$ of $E \subset \mathbb{R}^n$ is the collection of all the convex combinations in E. The convex envelope can be characterized as the intersection of all the convex sets containing E. Of course, K is convex if and only if $K = \mathrm{conv}(K)$.

Convex functions: A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a **convex function** if

$$f(t\,x + (1-t)\,y) \le t\,f(x) + (1-t)\,f(y), \qquad \forall t \in [0,1], x, y \in \mathbb{R}^n, \quad (2.12)$$

or, equivalently, if the **epigraph** of f, $\mathrm{Epi}(f) = \{(x,t) : t \ge f(x)\} \subset \mathbb{R}^{n+1}$, is a convex set in $\mathbb{R}^{n+1}$. The **domain of** f, $\mathrm{Dom}(f) = \{x \in \mathbb{R}^n : f(x) < \infty\} = \{f < \infty\}$, is a convex set in $\mathbb{R}^n$. Notice that, with this definition, whenever f

is not identically equal to $+\infty$, the dimension of $\mathrm{Dom}(f)$ could be any integer between 0 and n. Given a convex set K in $\mathbb{R}^n$ and a function $f : K \to \mathbb{R}$ satisfying (2.12) for $x, y \in K$, by extending $f = +\infty$ on $\mathbb{R}^n \setminus K$, we obtain a convex function with $K = \mathrm{Dom}(f)$; therefore, the point of view adopted here includes what is probably the more standard notion of "finite-valued, convex function defined on a convex set" that readers may be familiar with. The **indicator function** I_K of a convex set $K \subset \mathbb{R}^n$, defined by setting $I_K(x) = 0$ if $x \in K$ and $I_K(x) = +\infty$ if $x \notin K$, is a convex function. In particular, the basic optimization problem "minimize a finite-valued convex function g over a convex set K" can be simply recast as the minimization over $\mathbb{R}^n$ of the $\mathbb{R} \cup \{+\infty\}$-valued function $f = g + I_K$, that is

$$\inf_K g = \inf_{\mathbb{R}^n} \{ g + I_K \}. \tag{2.13}$$

Finally, a more practical reason for considering $\mathbb{R} \cup \{+\infty\}$-valued convex functions is that, as well shall see subsequently, many natural convex functions arise by taking suprema of families of affine functions, and thus they may very well take the value $+\infty$ outside of a convex set.

Lipschitz continuity and a.e. differentiability: We prove that *convex functions are always locally Lipschitz*[5] *in the relative interiors of their domains*. We give details in the case when $\mathrm{Dom}(f)$ has affine dimension n, since the general case is proved similarly. Let $\Omega = \mathrm{IntDom}(f)$, and let us first prove that f **is locally bounded in** Ω. To this end, given $B_r(x) \subset\subset \Omega$ we notice that, for every $z \in B_r(x)$, $y = 2x - z \in B_r(x)$ is such that $x = (y+z)/2$. Hence, $f(x) \le (f(y) + f(z))/2$, which gives

$$\inf_{B_r(x)} f \ge 2\, f(x) - \sup_{B_r(x)} f,$$

i.e., f is locally bounded in Ω if it is locally bounded *from above* in Ω. To show boundedness from above, let us fix $n+1$ unit vectors $\{v_i\}_{i=1}^{n+1}$ in $\mathbb{R}^n$ so that the simplex Σ with vertexes v_i – defined as the set of all the convex combinations $\sum_{i=1}^{n+1} t_i v_i$ corresponding to $0 < t_i < 1$ with $\sum_{i=1}^{n+1} t_i = 1$ – is an open set in $\mathbb{R}^n$ containing the origin in its interior. Now, for each $x \in \Omega$, we can find $r > 0$ such that $\Sigma_{x,r} = x + r\,\Sigma$ is contained in Ω, and since the convexity of f implies that

$$f\left(\sum_{i=1}^N t_i\, x_i\right) \le \sum_{i=1}^N t_i\, f(x_i), \qquad \begin{cases} \forall N \in \mathbb{N},\ \forall \{t_i\}_{i=1}^N \in [0,1] \text{ s.t. } \sum_{i=1}^N t_i = 1, \\ \forall \{x_i\}_{i=1}^N \subset \mathbb{R}^n, \end{cases} \tag{2.14}$$

[5] See Appendix A.10 for the basics on Lipschitz functions.

we conclude that

$$\sup_{\Sigma_{x,r}} f \le \max_{1\le i\le n+1} f(x + r\, v_i) < \infty.$$

By a covering argument, we conclude that f is locally bounded (from above and, thus, also from below) in Ω. We next exploit local boundedness to show that f **is locally Lipschitz in** Ω. Indeed, if B_{2r} is a ball of radius $2r$ compactly contained in Ω, and if B_r is concentric to B_{2r} with radius r, then

$$\mathrm{Lip}(f; B_r) \le \frac{1}{r}\Big(\sup_{B_{2r}} f - \inf_{B_{2r}} f\Big).$$

To show this, pick $x, y \in B_r$ and write $y = t\,x + (1-t)\,z$ for some $z \in \partial B_{2r}$. Then, $|x-y| = |1-t|\,|x-z| \ge r\,|1-t|$ so that

$$\begin{aligned} f(y) - f(x) &\le t\,f(x) + (1-t)\,f(z) - f(x) \le |1-t|\,|f(x) - f(z)| \\ &\le \frac{|x-y|}{r}\Big(\sup_{B_{2r}} f - \inf_{B_{2r}} f\Big). \end{aligned}$$

In particular, by Rademacher's theorem, f **is a.e. differentiable in** Ω, and a simple consequence of (2.12) shows that

$$f(y) \ge f(x) + \nabla f(x)\cdot(y-x), \qquad \forall y \in \mathbb{R}^n, \tag{2.15}$$

whenever f is differentiable at x with gradient $\nabla f(x)$. Condition (2.15) expresses the familiar property that convex functions lie above the tangent hyperplanes to their graphs whenever the latter are defined. In fact, inequality (2.15) points at a very fruitful way to think about convex functions, which we are now going to discuss.

Convex functions as suprema of affine functions: If $\mathcal{A}$ is any family of **affine functions** on $\mathbb{R}^n$ (i.e., if $\alpha \in \mathcal{A}$, then $\alpha(x) = a + y\cdot x$ for some $a \in \mathbb{R}$ and $y \in \mathbb{R}^n$), then it is trivial to check that

$$f = \sup_{\alpha\in\mathcal{A}} \alpha \tag{2.16}$$

defines a convex function on $\mathbb{R}^n$ with values in $\mathbb{R}\cup\{+\infty\}$. A convex function defined in this way is automatically lower semicontinuous on $\mathbb{R}^n$, as it is the supremum of continuous functions. Of course, not every convex function is going to be lower semicontinuous on $\mathbb{R}^n$ (e.g., $f(x) = I_{[0,1)}(x)$ is not lower semicontinuous at $x = 1$), so not every convex function will satisfy an identity like (2.16) on the whole $\mathbb{R}^n$. However, it is not hard to deduce from (2.15) that

$$f(z) = \sup\Big\{f(x) + \nabla f(x)\cdot(z-x) : f \text{ is differentiable at } x\Big\} \quad \forall z \in \mathrm{IntDom}(f), \tag{2.17}$$

so that (2.16) always holds on IntDom(f) if $\mathcal{A} = \{\alpha_x\}_x$ for x ranging among the points of differentiability of f, $\alpha_x(z) = a_x + y_x \cdot z$, $a_x = f(x) - \nabla f(x) \cdot x$, and $y_x = \nabla f(x)$. We now introduce the concepts of subdifferential and Fenchel–Legendre transform of a convex function. These concepts lead to a representation formula for convex functions similar to (but more robust than) (2.17).

Subdifferential at a point: Given a convex function f, a point $x \in \mathrm{Dom}(f)$, and a hyperplane L in $\mathbb{R}^{n+1}$, we say that L is a **supporting hyperplane of** f **at** x if L is the graph of an affine function α on $\mathbb{R}^n$ such that $\alpha \le f$ on $\mathbb{R}^n$ and $\alpha(x) = f(x)$. If $a \in \mathbb{R}$ and $y \in \mathbb{R}^n$ are such that $\alpha(z) = a + y \cdot z$ for all $z \in \mathbb{R}^n$, then y is called the **slope of** L. The **subdifferential** $\partial f(x)$ **of** f **at** x is defined as follows: if $x \in \mathrm{Dom}(f)$, then we set

$$\begin{aligned}\partial f(x) &= \Big\{\text{slopes of all the supporting hyperplanes of } f \text{ at } x\Big\}\\ &= \left\{y \in \mathbb{R}^n : \exists\, a \in \mathbb{R} \text{ s.t. } \begin{array}{ll} a + y \cdot z \le f(z) & \forall z \in \mathbb{R}^n \\ a + y \cdot x = f(x). & \end{array}\right\},\\ &= \Big\{y \in \mathbb{R}^n : f(z) \ge f(x) + y \cdot (z - x) \quad \forall z \in \mathbb{R}^n\Big\}; \end{aligned} \tag{2.18}$$

otherwise, i.e., if $f(x) = +\infty$, we set $\partial f(x) = \emptyset$. If f is differentiable at some $x \in \mathrm{Dom}(f)$, then $x \in \mathrm{IntDom}(f)$ (because f is finite in a neighborhood of x) and

$$\partial f(x) = \{\nabla f(x)\} \tag{2.19}$$

(see Proposition 2.7 for the proof). At a generic point $x \in \mathrm{Dom}(f)$, where f may not be differentiable, we always have that $\partial f(x)$ is a closed convex set in $\mathbb{R}^n$. For example, if $f(x) = |x|$, then $\partial f(0)$ is the closed unit ball in $\mathbb{R}^n$ centered at the origin (see Figure 2.2); if f is the maximum of finitely many affine functions α_i, with slope y_i, then $\partial f(x)$ is the convex envelope of those y_i such that x belongs to $\{f = \alpha_i\}$. In the following proposition we prove (2.19) together with a sort of continuity property of subdifferentials.

Proposition 2.7 (Continuity of subdifferentials) *If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. Moreover, for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\partial f(B_\delta(x)) \subset B_\varepsilon(\nabla f(x)). \tag{2.20}$$

Proof *Step one*: We prove (2.19). Given $y \in \partial f(x)$, set

$$F_{x,y}(z) = f(z) - f(x) - y \cdot (z - x), \qquad z \in \mathbb{R}^n.$$

Notice that $F_{x,y}$ has a minimum at x, since $F_{x,y}(z) \ge 0 = F_{x,y}(x)$ for every $z \in \mathbb{R}^n$. In particular, if f is differentiable at x, then $F_{x,y}$ is differentiable at x with $0 = \nabla F_{x,y}(x) = \nabla f(x) - y$ so that (2.19) is proved.

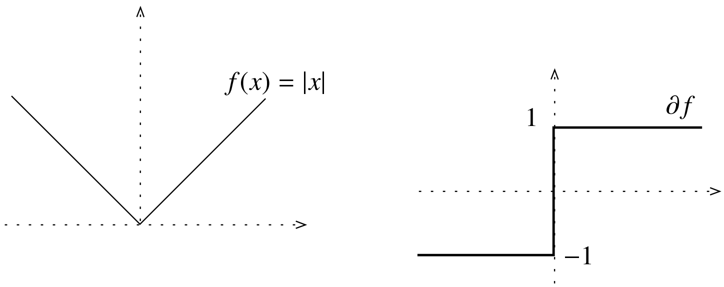

Figure 2.2 The subdifferential of $f(x) = |x|$ when $n = 1$.

Step two: If (2.20) fails, then there exist $\varepsilon > 0$ and $x_j \to x$ as $j \to \infty$ such that $|y_j - \nabla f(x)| \geq \varepsilon$ for every j and $y_j \in \partial f(x_j)$. It is easily seen that since f is bounded in a neighborhood of x, the sequence $\{y_j\}_j$ must be bounded in $\mathbb{R}^n$ and thus up to extracting subsequences, that $y_j \to y$ as $j \to \infty$. By taking limits as $j \to \infty$ in "$f(z) \geq f(x_j) + y_j \cdot (z - x_j)$ for every $z \in \mathbb{R}^n$," we deduce that $y \in \partial f(x) = \{\nabla f(x)\}$, in contradiction with $|y_j - \nabla f(x)| \geq \varepsilon$ for every j. □

Fundamental theorem of (convex) Calculus: It is well-known that a smooth function $g : \mathbb{R} \to \mathbb{R}$ is the derivative f' of a smooth *convex* function $f : \mathbb{R} \to \mathbb{R}$ if and only if g is increasing on $\mathbb{R}$. The notion of subdifferential and a proper generalization of monotonicity to $\mathbb{R}^n$ allow to extend this theorem to $\mathbb{R} \cup \{+\infty\}$-valued convex functions on $\mathbb{R}^n$. First of all, let us introduce the notion of **(total) subdifferential** of f, defined as

$$\partial f = \bigcup_{x \in \mathbb{R}^n} \{x\} \times \partial f(x)\,; \tag{2.21}$$

see Figure 2.2. Notice that ∂f is a subset of $\mathbb{R}^n \times \mathbb{R}^n$, which is closed as soon as f is lower semicontinuous. Recalling that, as set in (2.10), $S \subset \mathbb{R}^n \times \mathbb{R}^n$ is **cyclically monotone** if for every finite set $\{(x_i, y_i)\}_{i=1}^N \subset S$ one has

$$\sum_{i=1}^{N} y_i \cdot (x_{i+1} - x_i) \leq 0, \qquad \text{where } x_{N+1} = x_1. \tag{2.22}$$

Thus, we have the following theorem.

Theorem 2.8 (Rockafellar theorem) *Let $S \subset \mathbb{R}^n \times \mathbb{R}^n$ be a non-empty set: S is cyclically monotone if and only if there exists a convex and lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that*

$$S \subset \partial f. \tag{2.23}$$

Remark 2.9 Notice that $S \neq \emptyset$ and $S \subset \partial f$ imply $\mathrm{Dom}(f) \neq \emptyset$.

Proof *Proof that* (2.23) *implies cyclical monotonicity*: Let us consider a finite subset $\{(x_i, y_i)\}_{i=1}^N$ of S. Then, $y_i \in \partial f(x_i)$ implies $\partial f(x_i) \neq \emptyset$, and thus $f(x_i) < \infty$. For every $i = 1, \ldots, N$, we know that

$$f(x) \geq f(x_i) + y_i \cdot (x - x_i), \qquad \forall x \in \mathbb{R}^n,$$

so that testing the i-th inequality at $x = x_{i+1}$ and summing up over $i = 1, \ldots, N$ gives

$$\sum_{i=1}^N f(x_{i+1}) \geq \sum_{i=1}^N f(x_i) + y_i \cdot (x_{i+1} - x_i).$$

We find (2.22) since $\sum_{i=1}^N f(x_{i+1}) = \sum_{i=1}^N f(x_i)$ by the convention $x_{N+1} = x_1$.

Proof that cyclical monotonicity implies (2.23): We need to define a convex function f which contains S in its subdifferential. To this end, we fix[6] $(x_0, y_0) \in S$ and define

$$f(z) = \sup\left\{ y_N \cdot (z - x_N) + \sum_{i=1}^{N-1} y_i \cdot (x_{i+1} - x_i) + y_0 \cdot (x_1 - x_0) : \{(x_i, y_i)\}_{i=1}^N \subset S \right\} \tag{2.24}$$

for $z \in \mathbb{R}^n$. Clearly, f is a convex and lower semicontinuous function on $\mathbb{R}^n$ with values in $\mathbb{R} \cup \{+\infty\}$. We also notice that $f(x_0) \in \mathbb{R}$. Indeed, by applying (2.22) to $\{(x_i, y_i)\}_{i=0}^N \subset S$ we find

$$y_N \cdot (x_0 - x_N) + \sum_{i=1}^{N-1} y_i \cdot (x_{i+1} - x_i) + y_0 \cdot (x_1 - x_0) \leq 0 \qquad \forall \{(x_i, y_i)\}_{i=1}^N \subset S, \tag{2.25}$$

so that $f(x_0) \leq 0$. (Actually, we can even say that $f(x_0) = 0$, since $f(x_0) \geq 0$ by testing (2.24) with $\{(x_1, y_1)\} = \{(x_0, y_0)\}$ at $z = x_0$.) Interestingly, the proof that $f(x_0) < \infty$ is the only point of this argument where cyclical monotonicity plays a role.

We now prove that $S \subset \partial f$. Indeed, let $(x_*, y_*) \in S$ and let $t \in \mathbb{R}$ be such that $t < f(x_*)$. By definition of $f(x_*)$, we can find $\{(x_i, y_i)\}_{i=1}^N \subset S$ such that

$$y_N \cdot (x_* - x_N) + \sum_{i=0}^{N-1} y_i \cdot (x_{i+1} - x_i) \geq t. \tag{2.26}$$

If we now define $\{(x_i, y_i)\}_{i=1}^{N+1} \subset S$ by setting $x_{N+1} = x_*$ and $y_{N+1} = y_*$, then, by testing the definition of f with $\{(x_i, y_i)\}_{i=1}^{N+1} \subset S$, we find that, for every $z \in \mathbb{R}^n$,

[6] The choice of (x_0, y_0) is analogous to the choice of an arbitrary additive constant in the classical fundamental theorem of Calculus.

$$
\begin{aligned}
f(z) &\geq y_{N+1} \cdot (z - x_{N+1}) + \sum_{i=0}^{N} y_i \cdot (x_{i+1} - x_i) \\
&= y_* \cdot (z - x_*) + y_N \cdot (x_* - x_N) + \sum_{i=0}^{N-1} y_i \cdot (x_{i+1} - x_i) \\
&\geq t + y_* \cdot (z - x_*),
\end{aligned} \tag{2.27}
$$

where in the last inequality we have used (2.26). Since $f(z)$ is finite at $z = x_0$, by letting $t \to f(x_*)^-$ in (2.27) first with $z = x_0$, we see that $x_* \in \mathrm{Dom}(f)$, and then, by taking the same limit for an arbitrary z, we see that $y_* \in \partial f(x_*)$. □

Fenchel–Legendre transform: Given a convex function f, and the slope y of a supporting hyperplane to f, we know that there exists $a \in \mathbb{R}$ such that

$$
a + y \cdot x \leq f(x), \qquad \forall x \in \mathbb{R}^n.
$$

The largest value of $a \in \mathbb{R}$ such that this condition holds can be obviously characterized as $a = -f^*(y)$, where

$$
f^*(y) = \sup \left\{ y \cdot x - f(x) : x \in \mathbb{R}^n \right\}. \tag{2.28}
$$

The function f^* is called the **Fenchel–Legendre transform** of f. It is a convex function, and it is automatically lower semicontinuous on $\mathbb{R}^n$. Moreover, as it is easily seen, f^{**} is the lower semicontinuous envelope of f – i.e., the largest lower semicontinuous function lying below f: in particular, if f is convex and lower semicontinuous, then $f = f^{**}$, i.e.,

$$
f(x) = \sup \left\{ x \cdot y - f^*(y) : y \in \mathbb{R}^n \right\} \qquad \forall x \in \mathbb{R}^n. \tag{2.29}
$$

This is the "more robust" reformulation of (2.17). The last basic fact about convex functions that will be needed in the sequel is contained in the following two assertions:

$$
f(x) + f^*(y) \geq x \cdot y, \qquad \forall x, y \in \mathbb{R}^n, \tag{2.30}
$$

$$
f(x) + f^*(y) = x \cdot y, \qquad \text{iff } x \in \mathrm{Dom}(f) \text{ and } y \in \partial f(x). \tag{2.31}
$$

Notice that (2.30) is immediate from the definition (2.28). If $f(x) + f^*(y) = x \cdot y$, then $x \cdot y - f(x) = f^*(y) \geq y \cdot z - f(z)$, i.e., $f(z) \geq f(x) + y \cdot (z - x)$ for every $z \in \mathbb{R}^n$, i.e., $y \in \partial f(x)$; and, vice versa, if $y \in \partial f(x)$, then $x \cdot y - f(x) \geq y \cdot z - f(z)$ for every $z \in \mathbb{R}^n$ so that $f^*(y) \leq x \cdot y - f(x)$ – which combined with (2.30) gives $f^*(y) = x \cdot y - f(x)$.

Many common inequalities in analysis can be interpreted as instances of the **Fenchel–Legendre inequality** (2.30): for example, if $1 < p < \infty$ and

$f(x) = |x|^p/p$, one computes that $f^*(y) = |y|^{p'}/p'$ for $p' = p/(p-1)$, and thus finds[7] that (2.30) boils down to the classical **Young's inequality**.

Extremal points and the Choquet theorem:[8] Given a convex set K, we say that x_0 is an **extremal point** of K if $x_0 = (1-t)\,y + t\,z$ with $t \in [0,1]$ and $y, z \in K$ implies that either $t = 0$ or $t = 1$. We claim that

$$\begin{aligned}&\textit{if } K \subsetneq \mathbb{R}^n \textit{ is non-empty, closed, and convex,}\\ &\textit{then } K \textit{ has at least one extremal point.}\end{aligned} \tag{2.32}$$

To this end, we argue by induction on n, with the case $n = 1$ being trivial. If $n \geq 2$, since K is not empty and not equal to $\mathbb{R}^n$, there is a closed half-space H such that $K \subset H$ and $\partial H \cap \partial K \neq \emptyset$. In particular, $J = K \cap \partial H$ is a convex set with affine dimension $(n-1)$, and, by inductive hypothesis, there is an extremal point x_0 of J. We conclude by showing that x_0 is also an extremal point of K. Should this not be the case, we could find $t \in (0,1)$ and $x, y \in K$ such that $x_0 = (1-t)\,x + t\,y$. On the one hand, it must be $x, y \in \partial H$: otherwise, assuming, for example, that $x \in \mathrm{Int}(H)$, by $x_0 \in \partial H$ and $t \in (0,1)$ we would then find $y \notin H$, against $y \in K$; on the other hand, $x, y \in \partial H$ implies $x, y \in J$, and thus $x_0 = (1-t)\,x + t\,y$ with $t \in (0,1)$ would contradict the fact that x_0 is an extremal point of J. Having proved (2.32), we deduce from it the following statement (known as the **Choquet theorem**):

$$\begin{aligned}&\textit{if } K \subset \mathbb{R}^n \textit{ is convex and compact,}\\ &\textit{and } f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \textit{ is convex and lower semicontinuous,}\\ &\textit{then there is an extremal point } x_0 \textit{ of } K \textit{ such that } f(x_0) = \inf_K f.\end{aligned} \tag{2.33}$$

This is trivially true by (2.32) if $\mathrm{Dom}(f) = \emptyset$. Otherwise, since f is lower semicontinuous and K is compact, we can apply the Direct Method to show that the set J of the minimum points of f over K is non-empty and compact. Since f is convex, J is also convex. Hence, by (2.32), J admits an extremal point, and (2.33) is proved.

2.4 The Discrete Kantorovich Problem with Quadratic Cost

We now use the fundamental theorem of Calculus for convex functions proved in Section 2.3 to give a complete discussion of the discrete transport problem

[7] Of course, we have not just discovered an incredibly short proof of Young's inequality: indeed, showing that $f^*(y) = |y|^{p'}/p'$ is equivalent to prove Young's inequality! From this viewpoint, the importance of the Fenchel's inequality is more conceptual than practical.

[8] These results are only used in Section 2.5 and can be omitted on a first reading.

with quadratic transport cost $c(x,y) = |x-y|^2$. In particular, we make our first encounter with the Kantorovich duality formula; see (2.36), which comes into play as our means for proving that cyclical monotonicity is a sufficient condition for minimality in the transport problem.

Theorem 2.10 *If $\mu,\nu \in \mathcal{P}(\mathbb{R}^n)$ are discrete (i.e., if (2.1) and (2.2) hold) and $c(x,y) = |x-y|^2$, then, for every discrete transport plan $\gamma \in \Gamma(\mu,\nu)$, the following three statements are equivalent: (i) γ is a minimizer of $\mathbf{K}_c(\mu,\nu)$; (ii) $S(\gamma) = \{(x_i,y_j) : \gamma_{ij} > 0\} \subset \mathbb{R}^n \times \mathbb{R}^n$ is cyclically monotone; (iii) there exists a convex function f such that $S(\gamma) \subset \partial f$. Moreover, denoting $\mathcal{H}$ as the family of pairs (α,β) such that $\alpha,\beta : \mathbb{R}^n \to \mathbb{R}$ satisfy*

$$\alpha(x) + \beta(y) \le -x\cdot y \qquad \forall x,y \in \mathbb{R}^n \tag{2.34}$$

and defining $H : \mathcal{H} \to \mathbb{R}$ by setting

$$H(\alpha,\beta) = \sum_{i=1}^{N} \alpha(x_i)\,\mu_i + \sum_{j=1}^{M} \beta(y_j)\,\nu_j,$$

we have, for every $\gamma \in \Gamma(\mu,\nu)$ and $(\alpha,\beta) \in \mathcal{H}$,

$$\sum_{i,j} |x_i - y_j|^2\,\gamma_{ij} \ge \sum_{i=1}^{N} |x_i|^2\mu_i + \sum_{j=1}^{M} |y_j|^2\nu_j + 2\,H(\alpha,\beta). \tag{2.35}$$

Finally, if γ satisfies (iii), then (2.35) holds as an identity with $(\alpha,\beta) = (-f,-f^*)$. In particular,

$$\mathbf{K}_c(\mu,\nu) = \sum_{i=1}^{N} |x_i|^2\mu_i + \sum_{j=1}^{M} |y_j|^2\nu_j + 2\sup_{(\alpha,\beta)\in\mathcal{H}} H(\alpha,\beta). \tag{2.36}$$

Remark 2.11 Theorem 2.10 is, of course, a particular case of Theorem 3.20, in which the same assertions are proved without the discreteness assumption on μ and ν.

Proof of Theorem 2.10 *Step one*: We prove that (i) implies (ii) and that (ii) implies (iii). If (i) holds, then, by Theorem 2.5, we have

$$\sum_{\ell=1}^{L} |z_\ell - w_\ell|^2 \le \sum_{\ell=1}^{L} |z_{\ell+1} - w_\ell|^2, \tag{2.37}$$

whenever $\{(z_\ell,w_\ell)\}_{\ell=1}^L \subset S(\gamma)$ and $z_{L+1} = z_1$. By expanding the squares in (2.37),

$$\sum_{\ell=1}^{L} w_\ell\cdot(z_{\ell+1} - z_\ell) \le 0, \qquad \forall\{(z_\ell,w_\ell)\}_{\ell=1}^L \subset S(\gamma), \tag{2.38}$$

so that $S(\gamma)$ is cyclically monotone. In turn, if (ii) holds, then (iii) follows immediately by Rockafellar's theorem (Theorem 2.8).

Step two: For every $\gamma \in \Gamma(\mu,\nu)$ we have

$$\mathrm{Cost}(\gamma) = \sum_{i,j} |x_i - y_j|^2\, \gamma_{ij} = \sum_{i=1}^N |x_i|^2 \mu_i + \sum_{j=1}^M |y_j|^2 \nu_j + 2 \sum_{i,j} (-x_i \cdot y_j)\, \gamma_{ij},$$

so that (2.34) gives

$$\mathrm{Cost}(\gamma) - \sum_{i=1}^N |x_i|^2 \mu_i - \sum_{j=1}^M |y_j|^2 \nu_j = -2 \sum_{i,j} x_i \cdot y_j\, \gamma_{ij} \geq 2\, H(\alpha,\beta),$$

that is (2.35). Now, if γ satisfies (iii), then, by the Fenchel–Legendre inequality (2.30), we have $(-f,-f^*) \in \mathcal{H}$, while (2.31) and $S(\gamma) \subset \partial f$ give

$$f(x_i) + f^*(y_j) = x_i \cdot y_j, \qquad \text{if } \gamma_{ij} > 0, \tag{2.39}$$

which in turn implies that (2.35) holds as an identity if we choose $(\alpha,\beta) = (-f,-f^*)$. This shows at once that (2.36) holds and that γ is a minimizer of $\mathbf{K}_c(\mu,\nu)$. □

The following three remarks concern the lack of uniqueness in the discrete Kantorovich problem.

Remark 2.12 We already know that uniqueness does not hold in problem $\mathbf{K}_c(\mu,\nu)$ for arbitrary data; recall Remark 2.4. However, the following statement (which will be proved in full generality in Theorem 3.15) provides a "uniqueness statement of sorts" for the quadratic transport cost: *If $S = \bigcup_\gamma S(\gamma)$, where γ ranges over all the optimal plans in the quadratic-cost transport problem defined by two discrete measures μ and ν, then S is cyclically monotone; in particular, there exists a convex function f such that $S(\gamma) \subset \partial f$ for every such optimal plan γ.* We interpret this as a uniqueness statement since the subdifferential ∂f appearing in it has the property of "bundling together" all the optimal transport plans of problem $\mathbf{K}_c(\mu,\nu)$. As done in Remark 2.13, this property can be indeed exploited to prove uniqueness in special situations. Notice that the cyclical monotonicity of $S = \bigcup_\gamma S(\gamma)$ is not obvious, since, in general, the union of cyclically monotone sets is not cyclically monotone; see Figure 2.3. The reason why $S = \bigcup_\gamma S(\gamma)$ is, nevertheless, cyclically monotone lies in the linearity of Cost combined with the convexity of $\Gamma(\mu,\nu)$. Together, these two properties imply that the set $\Gamma_{\mathrm{opt}}(\mu,\nu)$ of all optimal transport plans for $\mathbf{K}_c(\mu,\nu)$ is convex: in particular, if γ^1 and γ^2 are optimal in $\mathbf{K}_c(\mu,\nu)$, then $(\gamma^1+\gamma^2)/2$ is an optimal plan, and thus $S((\gamma^1+\gamma^2)/2)$ is cyclically monotone, and since

$$S\left(\frac{\gamma^1+\gamma^2}{2}\right) = S(\gamma^1) \cup S(\gamma^2).$$

we conclude that $S(\gamma^1) \cup S(\gamma^2)$ is also cyclically monotone.

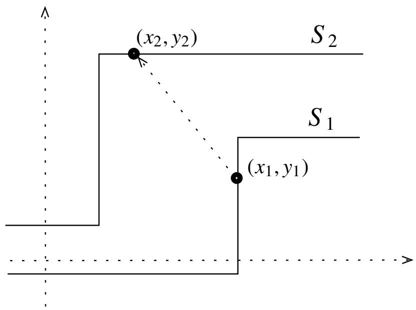

Figure 2.3 Two sets S_1 and S_2 that are cyclically monotone in $\mathbb{R}\times\mathbb{R}$, whose union $S_1 \cup S_2$ is not cyclically monotone. Indeed, by suitably picking points $(x_1, y_1) \in S_1$ and $(x_2, y_2) \in S_2$, we see that $(y_2 - y_1)(x_2 - x_1) < 0$, thus violating the cyclical monotonicity inequality on $\{(x_i, y_i)\}_{i=1}^2$.

Remark 2.13 (Uniqueness in dimension one) When $n = 1$, the statement in Remark 2.12 can be used to prove the uniqueness of minimizers for the discrete Kantorovich problem with quadratic cost. We only discuss this result informally. First of all, we construct a **monotone discrete transport plan** γ^* from μ to ν. Assuming without loss of generality that $\{x_i\}_{i=1}^N$ and $\{y_j\}_{j=1}^M$ are indexed so that $x_i < x_{i+1}$ and $y_j < y_{j+1}$, we define γ^*_{ij} as follows: if $\mu_1 \le \nu_1$, then set $\gamma^*_{11} = \mu_1$, and $\gamma^*_{1j} = 0$ for $j \ge 2$; otherwise, we let $j(1)$ be the largest index j such that $\mu_1 > \nu_1 + \ldots + \nu_j$ and set

$$\gamma^*_{1j} = \begin{cases} \nu_j, & \text{if } 1 \le j \le j(1), \\ \mu_1 - (\nu_1 + \ldots + \nu_{j(1)}), & \text{if } j = j(1) + 1, \\ 0, & \text{if } j(1) + 1 < j \le M. \end{cases}$$

In this way we have allocated all the mass μ_1 sitting at x_1 among the first $j(1)+1$ receiving sites, with the first $j(1)$ receiving sites completely filled. Next, we start distributing the mass μ_2 at site x_2, start moving the largest possible fraction of it to $y_{j(1)+1}$ (which can now receive a $\nu_{j(1)+1} - [\mu_1 - (\nu_1 + \ldots + \nu_{j(1)})]$ amount of mass), and keep moving any excess mass to the subsequent sites $y_{j(1)+k}$, $k \ge 2$, if needed. Evidently, the resulting transport plan γ^* is such that $S(\gamma^*)$ is contained in the extended graph of an increasing function so that γ^* is indeed optimal in $\mathbf{K}_c(\mu, \nu)$. A heuristic explanation of why this is the unique optimal transport plan is given in Figure 2.4a. For a proof, see Theorem 16.1-(i,ii).

Remark 2.14 We review Remark 2.4 in light of the results of this chapter. Denoting with superscripts the coordinates of points, so that $p = (p^1, p^2)$ is the generic point of $\mathbb{R}^2$, we take

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

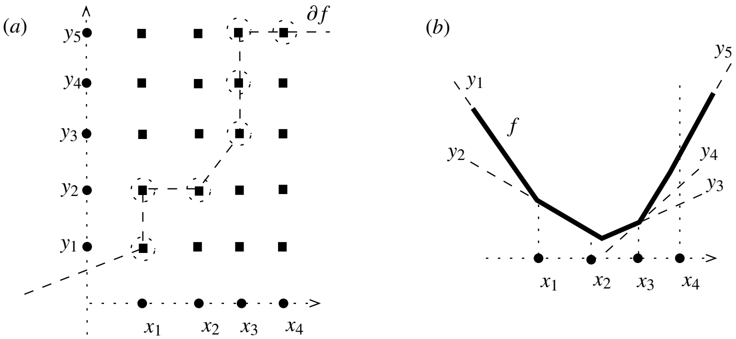

Figure 2.4 (a): A discrete transport problem with quadratic cost and $n = 1$. The black squares indicates all the possible interaction pairs (x_i, y_j). Weights μ_i and ν_j are such that $S(\gamma^*)$ consists of the circled black squares. If γ is another optimal transport plan, then, by the statement in Remark 2.12, $S(\gamma) \cup S(\gamma^*)$ is contained in the subdifferential of a convex function. This implies that $S(\gamma) \setminus S(\gamma^*)$ can only contain either (x_2, y_3) *or* (x_3, y_2). However, given the construction of γ^*, the fact that $S(\gamma^*)$ is "jumping diagonally" from (x_2, y_2) to (x_3, y_3) means that $\mu_1 + \mu_2 = \nu_1 + \nu_2$. So there is no mass left for γ to try something different: if γ activates (x_2, y_3) (i.e., if γ sends a fraction of μ_2 to y_3), then γ must activate (x_3, y_2) (sending a corresponding fraction of μ_3 to y_2 to compensate the first modification), and this violates cyclical monotonicity. Therefore, γ^* is a unique optimal plan for $\mathbf{K}_c(\mu, \nu)$. (b): A geometric representation of a potential f such that $S(\gamma^*) \subset \partial f$. Notice that $\partial f(x_1) = [y_1, y_2]$ (with $y_1 = f'(x_1^-)$ and $y_2 = f'(x_1^+)$), $\partial f(x_2) = \{f'(x_2) = y_2\}$, $\partial f(x_3) = [y_3, y_5]$ (with $y_3 = f'(x_3^-)$, $y_5 = f'(x_3^+)$ and y_4 in the interior of $\partial f(x_3)$), and $\partial f(x_4) = \{f'(x_4) = y_5\}$. Notice that we have large freedom in accommodating the y_js as elements of the subdifferentials $\partial f(x_i)$; in particular, we can find other convex functions g with $S(\gamma^*) \subset \partial g$ and such that $f - g$ is not constant.

No matter what values of μ_i and ν_j are chosen, all the admissible transport plans will have the same cost. If, say, $\mu_i = \nu_j = 1/2$ for all i and j, then all the plans

$$\gamma_{11}^t = t, \qquad \gamma_{12}^t = \frac{1}{2} - t, \qquad \gamma_{21}^t = \frac{1}{2} - t, \qquad \gamma_{22}^t = t,$$

corresponding to $t \in [0, 1/2]$ are optimal. If $t \in (0, 1/2)$, then $S(\gamma^t)$ contains all the four possible pairs $S = \{(x_i, y_j)\}_{i,j}$. How many convex functions (modulo additive constants) can contain S in their subdifferential? Just one. Indeed, the slopes $y_1 = (-1, 0)$ and $y_2 = (1, 0)$ correspond to the affine functions $\ell(p) = a - p^1$ and $m(p) = b + p^1$ for $a, b \in \mathbb{R}$. The only way for $y_1, y_2 \in \partial f(x_1) \cap \partial f(x_2)$ is that the set $\{\ell = m\}$ contains both $x_1 = (0, 1)$ and $x_2 = (0, -1)$. Hence, we must have $a = b$ and, modulo additive constants, there exists a unique convex potential; see Figure 2.5.

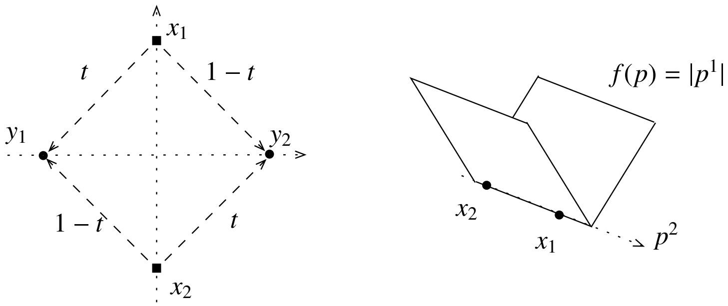

Figure 2.5 An example of discrete transport problem with quadratic cost where we have nonuniqueness of optimal transport plans, but where there is a unique (up to additive constants) convex potential such that the statement in Remark 2.12 holds.

2.5 The Discrete Monge Problem

We close this chapter with a brief discussion of the discrete Monge problem. The following theorem is our main result in this direction.

Theorem 2.15 *If $\{x_i\}_{i=1}^L$ and $\{y_j\}_{j=1}^L$ are families of L distinct points, and μ and ν denote the discrete measures*

$$\mu = \sum_{i=1}^{L} \frac{\delta_{x_i}}{L}, \qquad \nu = \sum_{j=1}^{L} \frac{\delta_{y_j}}{L}, \tag{2.40}$$

then for every optimal transport plan γ in $\mathbf{K}_c(\mu,\nu)$ there is a transport map T from μ to ν with the same transport cost as γ; in particular, T is optimal in $\mathbf{M}_c(\mu,\nu)$ and $\mathbf{M}_c(\mu,\nu) = \mathbf{K}_c(\mu,\nu)$.

Remark 2.16 It is interesting to notice that, by a perturbation argument, given an arbitrary pair of discrete probability measures (μ,ν) (i.e., μ and ν satisfy (2.1) and (2.2)), we can find a sequence $\{(\mu^\ell,\nu^\ell)\}_\ell$ of discrete probability measures such that $\mu^\ell \overset{*}{\rightharpoonup} \mu$ and $\nu^\ell \overset{*}{\rightharpoonup} \nu$ as $\ell \to \infty$, and each (μ^ℓ,ν^ℓ) satisfies the assumptions of Theorem 2.15 (with some $L = L_\ell$ in (2.40)). In a first approximation step, we can reduce to the case when all the weights μ_i and ν_j are rational numbers. Writing these weights with a common denominator L, we find $m_i, n_j \in \{1,\ldots,L\}$ such that

$$\mu_i = \frac{m_i}{L}, \qquad \nu_j = \frac{n_j}{L}, \qquad L = \sum_{i=1}^{N} m_i = \sum_{j=1}^{M} n_j .$$

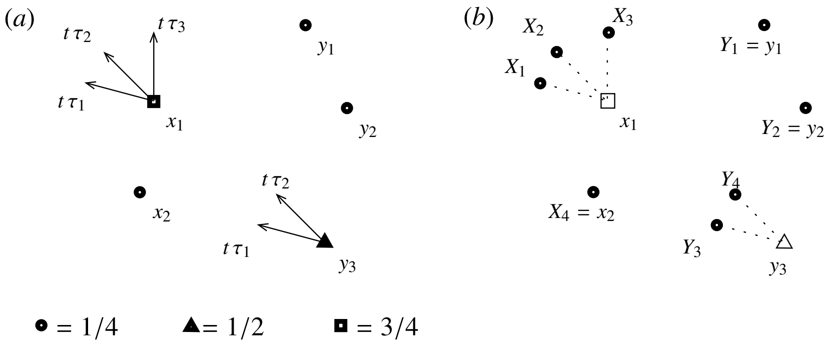

Figure 2.6 The second step in the approximation procedure of Remark 2.16: (a) The starting measures $\mu = (3/4)\delta_{x_1} + (1/4)\delta_{x_2}$ and $\nu = (1/4)(\delta_{y_1} + \delta_{y_2}) + (1/2)\,\delta_{y_3}$; notice that there is no transport map between these two measures; (b) The measures $\mu^t = (1/4)\sum_{h=1}^4 \delta_{X_h}$ and $\nu^t = (1/4)\sum_{h=1}^4 \delta_{Y_h}$ resulting from the approximation by splitting with $t > 0$. As $t \to 0^+$, these measure weak-star converge to μ and ν respectively. Notice that there is a transport map from μ^t to ν^t corresponding to every permutation of $\{1, \dots, 4\}$. One of them is optimal for the Monge problem with quadratic cost from μ^t to ν^t.

Then, in a second approximation step, we consider a set of L-distinct unit vectors $\{\tau_k\}_{k=1}^L$ in $\mathbb{R}^n$ and notice that, on the one hand,

$$\sum_{k=1}^{m_i} \frac{1}{L}\,\delta_{x_i + t\,\tau_k} \overset{*}{\rightharpoonup} \frac{m_i}{L}\,\delta_{x_i} \qquad \text{as } t \to 0^+,$$

while, on the other hand, for every sufficiently small but positive t, both

$$\{X_h\}_{h=1}^L = \{x_i + t\,\tau_k : 1 \le i \le N, 1 \le k \le m_i\},$$
$$\{Y_h\}_{h=1}^L = \{y_j + t\,\tau_k : 1 \le j \le M, 1 \le k \le n_j\},$$

consist of L-many distinct points. By combining these two approximation steps, we find a sequence $\{(\mu_j, \nu_j)\}_j$ with the required properties; see Figure 2.6.

Proof of Theorem 2.15 We notice that, by (2.40),

$$\Gamma(\mu,\nu) = \Big\{\{\gamma_{ij}\} : \gamma_{ij} = \frac{b_{ij}}{L}, \quad b = \{b_{ij}\} \in \mathcal{B}_L\Big\}, \tag{2.41}$$

where $\mathcal{B}_L \subset \mathbb{R}^{L\times L}$ is the set of the $L \times L$-**bistochastic matrices** $b = \{b_{ij}\}$, i.e., for every i, j, $b_{ij} \in [0,1]$ and $\sum_i b_{ij} = \sum_j b_{ij} = 1$. Therefore, by the Choquet theorem (2.33), if γ is an optimal plan in $\mathbf{K}_c(\mu,\nu)$, then $b = \{b_{ij} = L\,\gamma_{ij}\}$ is an extremal point of $\mathcal{B}_L$. The latter are characterized as follows (a result known as **the Birkhoff theorem**):

$$\textit{permutation matrices are the extremal points of } \mathcal{B}_L, \tag{2.42}$$

where $b = \{b_{ij}\}$ is a $L \times L$**-permutation matrix** if $b_{ij} = \delta_{j,\sigma(i)}$ for a permutation σ of $\{1,\dots,L\}$. The proof of Birkhoff's theorem is very similar to the proof of Theorem 2.5 and goes as follows. It is enough to prove that if $b \in \mathcal{B}_L$ and $b_{i(1)j(1)} \in (0,1)$ for some pair $(i(1), j(1))$, then b is not an extremal point of $\mathcal{B}_L$. Indeed, by $\sum_j b_{i(1)j} = 1$ we can find, on the $i(1)$-row of b, an entry $b_{i(1)j(2)} \in (0,1)$ with $j(2) \neq j(1)$. We can then find, in the $j(2)$-column of b, an entry $b_{i(2)j(2)} \in (0,1)$ with $i(2) \neq i(1)$. If we iterate this procedure, there is a first step $k \geq 3$ such that either $j(k) = j(1)$ or $i(k) = i(1)$. In the first case we have identified an *even* number of entries b_{ij} of b; in the second case, discarding the entry $b_{i(1)j(1)}$, we have also identified an *even* number of entries b_{ij} of b; in both cases, these entries are arranged into a closed loop in the matrix representation of b, and belong to $(0,1)$. We can exploit this cyclical structure to define a family of *variations* b^t of b: considering, for notational simplicity, the case when $j(3) = j(1)$, these variations take the form

$$b^t_{i(1)j(1)} = b_{i(1)j(1)} + t, \qquad b^t_{i(1)j(2)} = b_{i(1)j(2)} - t,$$
$$b^t_{i(2)j(1)} = b_{i(2)j(1)} - t, \qquad b^t_{i(2)j(2)} = b_{i(2)j(2)} + t,$$

and $b^t_{ij} = b_{ij}$ otherwise. In this way there is $t_0 \in (0,1)$ such that $b^t = \{b^t_{ij}\} \in \mathcal{B}_L$ whenever $|t| \leq t_0$. In particular, by $b = (b^{t_0} + b^{-t_0})/2$, we see that b is not an extremal point of $\mathcal{B}_L$ and conclude the proof of (2.42).

Having proved that for every discrete optimal transport plan γ in $\mathbf{K}_c(\mu,\nu)$ there is a permutation σ of $\{1,\dots,L\}$ such that $\gamma_{ij} = \delta_{\sigma(i),j}/L$, we can define a map $T : \{x_i\}_{i=1}^L \to \mathbb{R}^n$ by setting $T(x_i) = y_{\sigma(i)}$. By construction, $T_\# \mu = \nu$, with

$$\mathbf{M}_c(\mu,\nu) \leq \int_{\mathbb{R}^n} c(x,T(x))\, d\mu(x) = \mathrm{Cost}(\gamma) = \mathbf{K}_c(\mu,\nu) \leq \mathbf{M}_c(\mu,\nu),$$

where we have used (2.7) and the general inequality $\mathbf{K}_c(\mu,\nu) \leq \mathbf{M}_c(\mu,\nu)$. □